



3 9091 00299364 4

## Contents

Foreword by David Hendry	xi
Preface	xv
Acknowledgements	xx
List of symbols and abbreviations	xxi

### Part I      Introduction

<b>1      Econometric modelling, a preliminary view</b>	3
1.1    Econometrics – a brief historical overview	3
1.2    Econometric modelling – a sketch of a methodology	15
1.3    Looking ahead	22
<b>2      Descriptive study of data</b>	23
2.1    Histograms and their numerical characteristics	23
2.2    Frequency curves	27
2.3    Looking ahead	29

### Part II      Probability theory

<b>3      Probability</b>	33
3.1    The notion of probability	34
3.2    The axiomatic approach	37
3.3    Conditional probability	43

**Contents**

<b>4 Random variables and probability distributions</b>	47
4.1 The concept of a random variable	48
4.2 The distribution and density functions	55
4.3 The notion of a probability model	60
4.4 Some univariate distributions	62
4.5 Numerical characteristics of random variables	68
<b>5 Random vectors and their distributions</b>	78
5.1 Joint distribution and density functions	78
5.2 Some bivariate distributions	83
5.3 Marginal distributions	85
5.4 Conditional distributions	89
<b>6 Functions of random variables</b>	96
6.1 Functions of one random variable	96
6.2* Functions of several random variables	99
6.3 Functions of normally distributed random variables, a summary	108
6.4 Looking ahead	109
Appendix 6.1 – The normal and related distributions	110
<b>7 The general notion of expectation</b>	116
7.1 Expectation of a function of random variables	116
7.2 Conditional expectation	121
7.3 Looking ahead	127
<b>8* Stochastic processes</b>	130
8.1 The concept of a stochastic process	131
8.2 Restricting the time-heterogeneity of a stochastic process	137
8.3 Restricting the memory of a stochastic process	140
8.4 Some special stochastic processes	144
8.5 Summary	162
<b>9 Limit theorems</b>	165
9.1 The early limit theorems	165
9.2 The law of large numbers	168
9.3 The central limit theorem	173
9.4 Limit theorems for stochastic processes	178
Summary	180

<b>Contents</b>	vii
-----------------	-----

<b>10*</b>	<b>Introduction to asymptotic theory</b>	183
10.1	Introduction	183
10.2	Modes of convergence	185
10.3	Convergence of moments	192
10.4	The ‘big O’ and ‘little o’ notation	194
10.5	Extending the limit theorems	198
10.6	Error bounds and asymptotic expansions	202

## Part III Statistical inference

<b>11</b>	<b>The nature of statistical inference</b>	213
11.1	Introduction	213
11.2	The sampling model	215
11.3	The frequency approach	219
11.4	An overview of statistical inference	221
11.5	Statistics and their distributions	223
	Appendix 11.1 – The empirical distribution function	228
<b>12</b>	<b>Estimation I – properties of estimators</b>	231
12.1	Finite sample properties	232
12.2	Asymptotic properties	244
12.3	Predictors and their properties	247
<b>13</b>	<b>Estimation II – methods</b>	252
13.1	The method of least-squares	253
13.2	The method of moments	256
13.3	The maximum likelihood method	257
<b>14</b>	<b>Hypothesis testing and confidence regions</b>	285
14.1	Testing, definitions and concepts	285
14.2	Optimal tests	290
14.3	Constructing optimal tests	296
14.4	The likelihood ratio test procedure	299
14.5	Confidence estimation	303
14.6	Prediction	306
<b>15*</b>	<b>The multivariate normal distribution</b>	312
15.1	Multivariate distributions	312
15.2	The multivariate normal distribution	315
15.3	Quadratic forms related to the normal distribution	319

15.4	Estimation	320
15.5	Hypothesis testing and confidence regions	323
<b>16*</b>	<b>Asymptotic test procedures</b>	326
16.1	Asymptotic properties	326
16.2	The likelihood ratio and related test procedures	328
<b>Part IV</b>	<b>The linear regression and related statistical models</b>	
<b>17</b>	<b>Statistical models in econometrics</b>	339
17.1	Simple statistical models	339
17.2	Economic data and the sampling model	342
17.3	Economic data and the probability model	346
17.4	The statistical generating mechanism	349
17.5	Looking ahead	352
	Appendix 17.1 – Data	355
<b>18</b>	<b>The Gauss linear model</b>	357
18.1	Specification	357
18.2	Estimation	359
18.3	Hypothesis testing and confidence intervals	363
18.4	Experimental design	366
18.5	Looking ahead	367
<b>19</b>	<b>The linear regression model I – specification, estimation and testing</b>	369
19.1	Introduction	369
19.2	Specification	370
19.3	Discussion of the assumptions	375
19.4	Estimation	378
19.5	Specification testing	392
19.6	Prediction	402
19.7	The residuals	405
19.8	Summary and conclusion	408
	Appendix 19.1 – A note on measurement systems	409
<b>20</b>	<b>The linear regression model II – departures from the assumptions underlying the statistical GM</b>	412
20.1	The stochastic linear regression model	413
20.2	The statistical parameters of interest	418

Contents	ix
20.3 Weak exogeneity	421
20.4 Restrictions on the statistical parameters of interest	422
20.5 Collinearity	432
20.6 ‘Near’ collinearity	434
<b>21 The linear regression model III – departures from the assumptions underlying the probability model</b>	<b>443</b>
21.1 Misspecification testing and auxiliary regressions	443
21.2 Normality	447
21.3 Linearity	457
21.4 Homoskedasticity	463
21.5 Parameter time invariance	472
21.6 Parameter structural change	481
Appendix 21.1 – Variance stabilising transformations	487
<b>22 The linear regression model IV – departures from the sampling model assumption</b>	<b>493</b>
22.1 Implications of a non-random sample	494
22.2 Tackling temporal dependence	503
22.3 Testing the independent sample assumption	511
22.4 Looking back	521
Appendix 22.1 – Deriving the conditional expectation	523
<b>23 The dynamic linear regression model</b>	<b>526</b>
23.1 Specification	527
23.2 Estimation	533
23.3 Misspecification testing	539
23.4 Specification testing	548
23.5 Prediction	562
23.6 Looking back	567
<b>24 The multivariate linear regression model</b>	<b>571</b>
24.1 Introduction	571
24.2 Specification and estimation	574
24.3 A priori information	579
24.4 The Zellner and Malinvaud formulations	585
24.5 Specification testing	589
24.6 Misspecification testing	596

**Contents**

24.7	Prediction	599
24.8	The multivariate dynamic linear regression (MDLR) model	599
	Appendix 24.1 -- The Wishart distribution	602
	Appendix 24.2 – Kronecker products and matrix differentiation	603
<b>25</b>	<b>The simultaneous equations model</b>	608
25.1	Introduction	608
25.2	The multivariate linear regression and simultaneous equations models	610
25.3	Identification using linear homogeneous restrictions	614
25.4	Specification	619
25.5	Maximum likelihood estimation	621
25.6	Least-squares estimation	626
25.7	Instrumental variables	637
25.8	Misspecification testing	644
25.9	Specification testing	649
25.10	Prediction	654
<b>26</b>	<b>Epilogue: towards a methodology of econometric modelling</b>	659
26.1	A methodologist's critical eye	659
26.2	Econometric modelling, formalising a methodology	661
26.3	Conclusion	671
	References	673
	Index	689

\* Starred chapters and/or sections are typically more difficult and might be avoided at first reading.

## **PART I**

---

### **Introduction**

---

## CHAPTER 1

---

### Econometric modelling, a preliminary view

---

#### 1.1 Econometrics – a brief historical overview

It is customary to begin a textbook by defining its subject matter. In this case this brings us immediately up against the problem of defining ‘econometrics’. Such a definition, however, raises some very difficult methodological issues which could not be discussed at this stage. The epilogue might be a better place to give a proper definition. For the purposes of the discussion which follows it suffices to use a working definition which provides only broad guide-posts of its intended scope:

*Econometrics is concerned with the systematic study of economic phenomena using observed data.*

This definition is much broader than certain textbook definitions narrowing the subject matter of econometrics to the ‘measurement’ of theoretical relationships as suggested by economic theory. It is argued in the epilogue that the latter definition of econometrics constitutes a relic of an outdated methodology, that of the logical positivism (see Caldwell (1982)). The methodological position underlying the definition given above is largely hidden behind the word ‘systematic’. The term systematic is used to describe the use of observed data in a framework where economic theory as well as statistical inference play an important role, as yet undefined. The use of *observed data* is what distinguishes econometrics from other forms of studying economic phenomena.

Econometrics, defined as the study of the economy using observed data, can be traced as far back as 1676, predating economics as a separate discipline by a century. Sir William Petty could be credited with the first

'systematic' attempt to study economic phenomena using data in his *Political Arithmetik*. Systematic in this case is used relative to the state of the art in statistics and economics of the time.

Petty (1676) used the pioneering results in *descriptive statistics* developed by his friend John Graunt and certain rudimentary forms of economic theorising to produce the first 'systematic' attempt in studying economic phenomena using data. Petty might also be credited as the first to submit to a most serious temptation in econometric modelling. According to Hull, one of his main biographers and collector of his works:

Petty sometimes appears to be seeking figures that will support a conclusion he has already reached: Graunt uses his numerical data as a basis for conclusions, declining to go beyond them.

(See Hull (1899), p. xxv.)

Econometrics, since Petty's time, has developed alongside statistics and economic theory borrowing and lending to both subjects. In order to understand the development of econometrics we need to relate it to developments in these subjects.

- Graunt and Petty initiated three important developments in statistics:
- (i) the systematic collection of (numerical) data;
  - (ii) the mathematical *theory of probability* related to life-tables; and
  - (iii) the development of what we nowadays call *descriptive statistics* (see Chapter 2) into a coherent set of techniques for analysing numerical data.

It was rather unfortunate that the last two lines of thought developed largely independent of each other for the next two centuries. Their slow convergence during the second half of the nineteenth and early twentieth centuries in the hands of Galton, Edgeworth, Pearson and Yule, *inter alia*, culminated with the *Fisher paradigm* which was to dominate statistical theory to this day.

The development of the calculus of probability emanating from Graunt's work began with Halley [1656–1742] and continued with De Moivre [1667–1754], Daniel Bernoulli [1700–82], Bayes [1702–61], Lagrange [1736–1813], Laplace [1749–1827], Legendre [1752–1833], Gauss [1789–1857] *inter alia*. In the hands of De Moivre the main line of the calculus of probability emanating from Jacob Bernoulli [1654–1705] was joined up with Halley's life tables to begin a remarkable development of probability theory (see Hacking (1975), Maistrov (1974)).

The most important of these developments can be summarised under the following headings:

- (i) manipulation of probabilities (addition, multiplication);

- (ii) families of distribution functions (normal, binomial, Poisson, exponential);
- (iii) law of error, least-squares, least-absolute errors;
- (iv) limit theorems (law of large numbers, central limit theorem);
- (v) life-table probabilities and annuities;
- (vi) higher order approximations;
- (vii) probability generating functions.

Some of these topics will be considered in some detail in Part II because they form the foundation of statistical inference.

The tradition in *Political Arithmetik* originated by Petty was continued by Gregory King [1656–1714]. Davenant might be credited with publishing the first ‘empirical’ demand schedule (see Stigler (1954)), drawing freely from King’s unpublished work. For this reason his empirical demand for wheat schedule is credited to King and it has become known as ‘King’s law’. Using King’s data on the change in the price ( $p_t$ ) associated with a given change in quantity ( $q_t$ ) Yule (1915) derived the empirical equation explicitly as

$$p_t = -2.33q_t + 0.05q_t^2 - 0.0017q_t^3. \quad (1.1)$$

Apart from this demand schedule, King and Davenant extended the line of thought related to the population and death rates in various directions thus establishing a tradition in *Political Arithmetik*, ‘the art of reasoning by figures upon things, relating to government’. *Political Arithmetik* was to stagnate for almost a century without any major developments in the descriptive study of data apart from grouping and calculation of tendencies.

From the economic theory viewpoint *Political Arithmetik* played an important role in classical economics where numerical data on money stock, prices, wages, public finance, exports and imports were extensively used as important tools in their various controversies. The best example of the tradition established by Graunt and Petty is provided by Malthus’ ‘Essay on the Principles of Population’. In the bullionist and currency-banking schools controversies numerical data played an important role (see Schumpeter (1954)). During the same period the calculation of index numbers made its first appearance.

With the establishment of the statistical society in 1834 began a more coordinated activity in most European countries for more reliable and complete data. During the period 1850–90 a sequence of statistical congresses established a common tradition for collecting and publishing data on many economic and social variables making very rapid progress on this front. In relation to statistical techniques, however, the progress was much slower. Measures of central tendency (arithmetic mean, median,

## 6 A preliminary view

mode, geometric mean) and graphical techniques were developed. Measures of dispersion (standard deviation, interquartile range), correlation and relative frequencies made their appearance towards the end of this period. A leading figure of this period who can be considered as one of the few people to straddle all three lines of development emanating from Graunt and Petty was the Belgian statistician Quetelet [1796–1874].

From the econometric viewpoint the most notable example of empirical modelling was Engel's study of family budgets. The most important of his conclusions was:

The poorer a family, the greater the proportion of its total expenditure that must be devoted to the provision of food.

(Quoted by Stigler (1954).)

This has become known as Engel's law and the relationship between consumption and income as Engel's curve (see Deaton and Muellbauer (1980)).

The most important contribution during this period (early nineteenth century) from the modelling viewpoint was what we nowadays call the *Gauss linear model*. In modern notation this model takes the following general form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad (1.2)$$

where  $\mathbf{y}$  is a  $T \times 1$  vector of observations linearly related to the unknown  $k \times 1$  vector  $\boldsymbol{\beta}$  via a known fixed  $T \times k$  matrix  $\mathbf{X}$  ( $\text{rank}(\mathbf{X})=k$ ,  $T \geq k$ ) but subject to (observation) error  $\mathbf{u}$ . This formulation was used to model a situation such that:

it is suspected that for settings  $x_1, x_2, \dots, x_k$  there is a value  $y$  related by a linear relation:

$$y = \sum_{i=1}^k \beta_i x_i,$$

where  $\boldsymbol{\beta}=(\beta_i)$  is unknown. A number  $T$  of observations on  $y$  can be made, corresponding to  $T$  different sets of  $(x_1, \dots, x_k)$ , i.e. we obtain a data set  $(y_t, x_{t1}, x_{t2}, \dots, x_{tk})$ ,  $t = 1, 2, \dots, T$ , but the readings  $y_t$  on  $y$ , are subject to error.  
(See Heyde and Seneta (1977).)

The problem as seen at the time was one of interpolation (approximation), that is, to 'approximate' the value of  $\boldsymbol{\beta}$ . The solution proposed came in the form of the *least-squares* approximation of  $\boldsymbol{\beta}$  based on the minimisation of

$$\mathbf{u}'\mathbf{u} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \quad (1.3)$$

which lead to

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (1.4)$$

(see Seal (1967)). The problem, as well as the solution, had nothing to do with probability theory as such. The probabilistic arguments entered the problem as an afterthought in the attempts of Gauss and Laplace to justify the method of least-squares. If the error terms  $u_t, t = 1, 2, \dots, T$  are assumed to be independent and identically distributed (IID) according to the normal distribution, i.e.

$$u_t \sim N(0, \sigma^2), \quad t = 1, 2, \dots, T, \quad (1.5)$$

then  $\hat{\beta}$  in (4) can be justified as ‘the optimal solution’ from a probabilistic viewpoint (see Heyde and Seneta (1977), Seal (1967), Maistrov (1974)).

The Gauss linear model was later given a very different interpretation in the context of probability theory by Galton, Pearson, and Yule, which gave rise to what is nowadays called the *linear regression model*. The model given in (2) is now interpreted as based wholly on probabilistic arguments,  $y_t$  and  $\mathbf{X}_t$  are assumed to be jointly normally distributed random variables and  $\mathbf{X}\beta$  is viewed as the conditional expectations of  $y_t$  given that  $\mathbf{X}_t = \mathbf{x}_t$  ( $\mathbf{X}_t$  takes the value  $\mathbf{x}_t$ ) for  $t = 1, 2, \dots, T$ , i.e.

$$E(y_t / \mathbf{X}_t = \mathbf{x}_t) = \beta' \mathbf{x}_t, \quad t = 1, 2, \dots, T, \quad (1.6)$$

with the error term  $u_t$  defined by  $u_t = y_t - E(y_t / \mathbf{X}_t = \mathbf{x}_t)$  (see Chapter 19 for further details). The linear regression model

$$y_t = E(y_t / \mathbf{X}_t = \mathbf{x}_t) + u_t, \quad t = 1, 2, \dots, T \quad (1.7)$$

can be written in matrix form as in (2) and the two models become indistinguishable in terms of notation. From the modelling viewpoint, however, the two models are very different. The Gauss linear model describes a ‘law-like’ relationship where the  $x_{ti}$ s are known constants. On the other hand, the linear regression model refers to a ‘predictive-like’ relationship where  $y_t$  is related to the observed values of the random vector  $\mathbf{X}_t$  (for further discussion see Chapter 19). This important difference went largely unnoticed by Galton, Pearson and the early twentieth-century applied econometricians. Galton in particular used the linear regression model to establish ‘law-like’ causal relationships in support of his theories of heredity in the then newly established discipline of eugenics.

The Gauss linear model was initially developed by astronomers in their attempt to determine ‘law-like’ relationships for planetary orbits, using a large number of observations with less than totally accurate instruments. The nature of their problem was such as to enable them to assume that their theories could account for all the information in the data apart from a *white-noise* (see Chapter 8) error term  $u_t$ . The situation being modelled resembles an ‘experimental design’ situation because of the relative constancy of the phenomena in question with nature playing the role of the

experimenter. Later, Fisher extended the applicability of the Gauss linear model to 'experimental-like' phenomena using the idea of *randomisation* (see Fisher (1958)). Similarly, the linear regression model, firmly based on the idea of conditional expectation, was later extended by Pearson to the case of stochastic regressors (see Seal (1967)).

In the context of the Gauss linear and linear regression models the convergence of descriptive statistics and the calculus of probability became a reality, with Galton [1822–1911], Edgeworth [1815–1926], Pearson [1857–1936] and Yule [1871–1951] being the main protagonists. In the hands of Fisher [1890–1962] the convergence was completed and a new modelling paradigm was proposed. One of the most important contributing factors to these developments in the early twentieth century was the availability of more complete and reliable data towards the end of the nineteenth century. Another important development contributing to the convergence of the descriptive study of data and the calculus of probability came in the form of Pearson's family of frequency curves which provided the basis for the transition from histograms to probability density functions (see Chapter 2). Moreover, the various concepts and techniques developed in descriptive statistics were to be reinterpreted and provide the basis for the probability theory framework. The frequency curves as used in descriptive statistics provide convenient 'models' for the observed data at hand. On the other hand, probability density functions were postulated as 'models' of the population giving rise to the data with the latter viewed as a representative sample from the former. The change from the *descriptive statistics* to the *probability theory* approach in statistical modelling went almost unnoticed until the mid-1930s when the latter approach formalised by Fisher dominated the scene.

During the period 1890–1920 the distinction between the population from where the observed data constitute a sample and the sample itself was blurred by the applied statisticians. This was mainly because the paradigm tacitly used, as formulated by Pearson, was firmly rooted in the descriptive statistics tradition where the modelling proceeds from the observed data in hand to the frequency (probability) model and no distinction between the population and the sample is needed. In a sense the population consists of the data in hand. In the context of the Fisher paradigm, however, the modelling of a probability model is postulated as a generalised description of the actual data generation process (DGP), or the population and the observed data are viewed as a realisation of a sample from the process. The transition from the Pearson to the Fisher paradigm was rather slow and went largely unnoticed even by the protagonists. In the exchanges between Fisher and Pearson about the superiority of the maximum likelihood estimation over the method of moments on efficiency grounds, Pearson

never pointed out that his method of moments was developed for a different statistical paradigm where the probability model is not postulated a priori (see Chapter 13). The distinction between the population and the sample was initially raised during the last decade of the nineteenth century and early twentieth century in relation to *higher order approximations* of the central limit theorem (CLT) results emanating from Bernouli, De Moivre and Laplace. These limit theorems were sharpened considerably by the Russian school (Chebyshev [1821–94], Liapounov [1857–1922], Markov [1856–1922], Kolmogorov [1903– ] (see Maistov (1974)) and used extensively during this period. Edgeworth and Charlier, among others, proposed asymptotic expansions which could be used to improve the approximation offered by the CLT for a given sample size  $T$  (see Cramer (1972)). The development of a formal distribution theory based on a fixed sample size  $T$ , however, began with Gosset's (Student's)  $t$  and Fisher's  $F$ -distributions (see Kendal and Stuart (1969)). These results provided the basis of modern statistical theory based on the Fisher paradigm. The transition from the Pearson to the Fisher paradigm became apparent in the 1930s when the theory of estimation and testing as we know it today was formulated. It was also the time when probability theory itself was given its axiomatic foundations by Kolmogorov (1933) and firmly established as part of mathematics proper. By the late 1930s probability theory as well as statistical inference as we know them today were firmly established.

The Gauss linear and linear regression models were appropriate for modelling essentially static phenomena. Yule (1926) discussed the difficulties raised when time series data are used in the context of the linear regression model and gave an insightful discussion of ‘non-sense regressions’ (see Hendry and Morgan (1986)). In an attempt to circumvent these problems, Yule (1927) proposed the linear *autoregressive model* (AR( $m$ )) where the  $x_n$ s are replaced by the lagged  $y_t$ s, i.e.

$$y_t = \sum_{i=1}^m \alpha_i y_{t-i} + u_t. \quad (1.8)$$

An alternative model for time-series data was suggested by Slutsky (1927) in his discussion of the dangers in ‘smoothing’ such data using weighted averaging. He showed that by weighted averaging of a white-noise process  $u_t$  can produce a data series with periodicities. Hence, somebody looking for cyclic behaviour can be easily fooled when the data series have been smoothed. His discussion gave rise to the other important family of time-series models, subsequently called the *moving average model* (MA( $p$ )):

$$y_t = \sum_{i=1}^p b_i u_{t-i} + u_t. \quad (1.9)$$

Wold (1938) provided the foundations for time series modelling by relating the above models to the mathematical theory of probability established by Kolmogorov (1933). These developments in time series modelling were to have only a marginal effect on mainstream econometric modelling until the mid-70s when a slow but sure convergence of the two methodologies began. One of the main aims of the present book is to complete this convergence in the context of a reformulated methodology.

With the above developments in probability theory and statistical inference in mind, let us consider the history of econometric modelling in the early twentieth century. The marginalist revolution of the 1870s, with Walras and Jevons the protagonists, began to take root and with it a change of attitude towards mathematical and statistical techniques and their role in studying the economy. In classical economics observed data were used mainly to 'establish' tendencies in support of theoretical arguments or as 'facts' to be explained. The mathematisation of economic theory brought about by the marginalist revolution contributed towards a purposeful attempt to quantify theoretical relationships using observed data. The theoretical relationships formulated in terms of equations such as demand and supply functions seemed to offer themselves for quantification using the newly established techniques of correlation and regression.

The early literature in econometric modelling concentrated mostly on two general areas, *business cycles* and *demand curves* (see Stigler (1954)). This can be explained by the availability of data and the influence of the marginalist revolution. The statistical analysis of business cycles took the form of applying correlation as a tool to separate long-term secular movements, periodic movements and short-run oscillations (see Hooker (1905), Moore (1914) *inter alia*). The empirical studies in demand theory concentrated mostly on estimating demand curves using the Gauss linear model disguised as regression analysis. The estimation of such curves was treated as 'curve fitting' with any probabilistic arguments being coincidental. Numerous studies of empirical demand schedules, mostly of agricultural products, were published during the period 1910–30 (see Stigler (1954), Morgan (1982), Hendry and Morgan (1986)), seeking to establish an empirical foundation for the 'law of demand'. These studies purported to estimate demand schedules of the simple form

$$q_t^D = a_0 + a_1 p_t, \quad (1.10)$$

where  $q_t^D$  refers to quantities demanded at time  $t$  (intentions on behalf of economic agents to buy a certain quantity of a commodity) corresponding to a range of hypothetical prices  $p_t$ . By adopting the Gauss linear model these studies tried to 'approximate' (10) by fitting the 'best' line through the scatter diagram of  $(\tilde{q}_t, \tilde{p}_t)$ ,  $t = 1, 2, \dots, T$  where  $\tilde{q}_t$  usually referred to

quantities transacted (or produced) and the corresponding prices  $\tilde{p}_t$  at time  $t$ . That is, they would estimate

$$\tilde{q}_t = b_0 + b_1 \tilde{p}_t, \quad t = 1, 2, \dots, T, \quad (1.11)$$

using least-squares or some other interpolation method and interpret the estimated coefficients  $\tilde{b}_0$  and  $\tilde{b}_1$  as estimates of the theoretical parameters  $a_0$  and  $a_1$  respectively, if the signs and values were consistent with the 'law of demand'. This simplistic modelling approach, however, ran into difficulties immediately. Moore (1914) estimated (11) using data on pig-iron (raw steel) production and price and 'discovered' (or so he thought) a positively sloping demand schedule ( $\tilde{b}_1 > 0$ ). This result attracted considerable criticism from the applied econometricians of the time such as Lehfeldt and Wright (see Stigler (1962)) and raised the most important issue in econometric modelling; *the connection between the estimated equations using observed data and the theoretical relationships postulated by economic theory*. Lehfeldt (1915), commenting on Moore's 'discovery', argued that the estimated equation was not a demand but a supply curve. Several applied econometricians argued that Moore's estimated equation was a mixture of demand and supply. Others, taking a more extreme view, raised the issue of whether estimated equations represent statistical artifacts or genuine empirical demand or supply curves. It might surprise the reader to learn that the same issue remains largely unresolved to this day. Several 'solutions' have been suggested since then but no satisfactory answer has emerged.

During the next two decades (1910–30) the applied econometricians struggled with the problem and proposed several ingenious ways to 'resolve' some of the problems raised by the estimated versus theoretical relationships issue. Their attempts were mainly directed towards specifying more 'realistic' theoretical models and attempting to rid the observed data of 'irrelevant information'. For example, the scenario that the demand and supply curves simultaneously shifting allowing us to observe only their intersection points received considerable attention (see Working (1927) *inter alia*). The time dimension of time-series data proved particularly difficult to 'solve' given that the theoretical model was commonly static. Hence 'detrending' the data was a popular way to 'purify' the observed data in order to bring them closer to the theoretical concepts purporting to measure (see Morgan (1982)). As argued below the estimated-theoretical issue raises numerous problems which, given the state of the art as far as statistical inference is concerned, could not have been resolved in any satisfactory way. In modern terminology these problems can be summarised under the following headings:

- (i) theoretical variables versus observed data;

12      A preliminary view

- (ii) statistical model specification;
- (iii) statistical misspecification testing;
- (iv) specification testing, reparametrisation, identification;
- (v) empirical versus theoretical models.

By the late 1920s there was a deeply felt need for a more organised effort to face the problems raised by the early applied econometricians such as Moore, Mitchell, Schultz, Clark, Working, Wallace, Wright, *inter alia*. This led to the creation of the Econometric Society in 1930. Frisch, Tinbergen and Fisher (Irving) initiated the establishment of ‘an international society’ *for the advancement of economic theory in its relation to statistics and mathematics*. The decade immediately after the creation of the Econometric Society can be characterised as the period during which the foundations of modern econometrics were laid mainly by posing some important and insightful questions.

An important attempt to resolve some of the problems raised by the estimated theoretical distinction was made by Frisch (1928) (1934). Arguing from the Gauss linear model viewpoint Frisch suggested the so-called *errors-in-variables* formulation where the theoretical relationships defined in terms of theoretical variables  $\mu_t \equiv (\mu_{1t}, \dots, \mu_{kt})'$  are defined by the system of  $k$  linear equations:

$$\mathbf{A}'\boldsymbol{\mu}_t = \mathbf{0}, \quad (1.12)$$

and the observed data  $\mathbf{y}_t \equiv (y_{1t}, \dots, y_{kt})'$  are related to  $\boldsymbol{\mu}_t$  via

$$\mathbf{y}_t = \boldsymbol{\mu}_t + \boldsymbol{\varepsilon}_t, \quad (1.13)$$

where  $\boldsymbol{\varepsilon}_t$  are errors of measurement. This formulation emphasises the distinction between theoretical variables and observed data with the *measurement equations* (13) relating the two. The problem as seen by Frisch was one of approximation (interpolation) in the context of linear algebra in the same way as the Gauss linear model was viewed. Frisch, however, with his *confluence analysis* offered no proper solution to the problem. A complete solution to the simplest case was only recently provided, 50 years later, by Kalman (1982). It is fair to say that although Frisch understood the problems raised by the empirical theoretical distinction as the quotation below testifies, his formulation of the problem turned out to be rather unsuccessful in this respect. Commenting on Tinbergen’s ‘A statistical test of business cycle theories’, Frisch argued that:

The question of what connection there is between relations we work with in theory and those we get by fitting curves to actual statistical data is a very delicate one. Tinbergen in his work hardly mentions it. He more or less takes it for granted that the relations he has found are in their nature

the same as the theory . . . This is, in my opinion, unsatisfactory. In a work of this sort, the connection between *statistical* and *theoretical relations* must be thoroughly understood and the nature of the information which the statistical relations furnish – although they are not identical with the theoretical relations – should be clearly brought out.

(See Frisch (1938), pp. 2–3.)

As mentioned above, by the late 1930s the Fisher paradigm of statistical inference was formulated into a coherent body of knowledge with a firm foundation in probability theory. The first important attempt to introduce this paradigm into econometrics was made by Koopmans (1937). He proposed a resetting of Frisch's errors-in-variables formulation in the context of the Fisher paradigm and related the least-squares method to that of maximum likelihood, arguing that the latter paradigm provides us with additional insight as to the nature of the problem posed and its 'solution' (estimation). Seven years later Haavelmo (a student of Frisch) published his celebrated monograph on 'The probability approach in econometrics' (see Haavelmo (1944)) where he argued that the probability approach (the Fisher paradigm) was the most promising approach to econometric modelling (see Morgan (1984)). His argument in a nutshell was that if statistical inference (estimation, testing and prediction) are to be used systematically we need to accept the framework in the context of which these results become available. This entails formulating theoretical propositions in the context of a well-defined statistical model. In the same monograph Haavelmo exemplified a methodological awareness far ahead of his time. In relation to the above discussion of the appropriateness of the Gauss linear model in modelling economic phenomena he distinguished between observed data resulting from:

- (1) experiments that we should like to make to see if certain real economic phenomena – when artificially isolated from 'other influences' – would verify certain hypotheses, and
- (2) the stream of experiments that Nature is steadily turning out from her own enormous laboratory, and which we merely watch as passive observers

He went on to argue:

In the first case we can make the agreement or disagreement between theory and facts depend upon two things: the facts we choose to consider, as well as our theory about them . . . In the second case we can only try to adjust our theories to reality as it appears before us. And what is the meaning of a design of experiments in this case? It is this: We try to choose a theory and a design of experiments to go with it, in such a way that the

resulting data would be those which we get by passive observation of reality. And to the extent that we succeed in doing so, we become master of reality – by passive agreement.

Now if we examine current economic theories, we see that a great many of them, in particular the more profound ones, require experiments of the first type mentioned above. On the other hand, the kind of economic data that we actually have belong mostly to the second type.

(See Haavelmo (1944).)

Unfortunately for econometrics, Haavelmo's views on the methodology of econometric modelling had much lesser influence than his formulation of a statistical model thought to be tailor made for econometrics; the so-called *simultaneous equations model*.

In an attempt to capture the *interdependence* of economic relationships Haavelmo (1943) proposed an alternative to Frisch's errors-in-variables formulation where no distinction between theoretical variables and observed data is made. The simultaneous equation formulation was specified by the system

$$\Gamma'y_t + \Delta'x_t + \varepsilon_t = 0, \quad (1.14)$$

where  $y_t$  refers to the variables whose behaviour this system purports to explain (endogenous) and  $x_t$  to the explanatory (extraneous) variables whose behaviour lies outside the intended scope of the theory underlying (14) and  $\varepsilon_t$  is the error term (see Chapter 25). The statistical analysis of (14) provided the agenda for a group of distinguished statisticians and econometricians assembled in Chicago in 1945. This group known as the Cowles Foundation Group, introduced the newly developed techniques of estimation (maximum likelihood) and testing into econometrics via the simultaneous equation model. Their results, published in two monographs (see Koopmans (1950) and Hood and Koopmans (1953)) were to provide the main research agenda in econometrics for the next 25 years.

It is important to note that despite Haavelmo's stated intentions in his discussion of the methodology of econometric modelling (see Haavelmo (1944)), the simultaneous equation model was later viewed in the Gauss linear model tradition where the theory is assumed to account for all the information in the data apart from some non-systematic (white-noise) errors. Indeed the research in econometric theory for the next 25–30 years was dominated by the Gauss linear model and its misspecification analysis and the simultaneous equations model and its identification and estimation. The initial optimism about the potential of the simultaneous equations model and its appropriateness for econometric modelling was not fulfilled. The problems related to the issue of esimated versus

theoretical relationships mentioned above were largely ignored because of this initial optimism. By the late 1970s the experience with large macroeconometric models based on the simultaneous equations formulation called into question the whole approach to econometric modelling (see Sims (1980), Malinvaud (1982), *inter alia*).

The inability of large macroeconometric models to compete with Box-Jenkins ARIMA models, which have no economic theory content, on prediction grounds (see Cooper (1972)) renewed the interest of econometricians to the issue of *static theory versus dynamic time series data* raised in the 1920s and 30s. Granger and Newbold (1974) questioned the conventional econometric approach of paying little attention to the time series features of economic data; the result of specifying statistical models using only the information provided by economic theory. By the late 1970s it was clear that the simultaneous equations model, although very useful, was not a panacea for all econometric modelling problems. The whole econometric methodology needed a reconsideration in view of the experience of the three decades since the Cowles Foundation.

The purpose of the next section is to consider an outline of a particular approach to econometric modelling which takes account of some of the problems raised above. It is only an outline because in order to formulate the methodology in any detail we need to use concepts and results which are developed in the rest of the book. In particular an important feature of the proposed methodology is the recasting of statistical models of interest in econometrics in the Fisherian mould where the probabilistic assumptions are made directly in terms of the observable random variables giving rise to the observed data and not some unobservable error term. The concepts and ideas involved in this recasting are developed in Parts II–IV. Hence, a more detailed discussion of the proposed methodology is given in the epilogue.

## 1.2 Econometric modelling – a sketch of a methodology

In order to motivate the methodology of econometric modelling adopted below let us consider a simplistic view of a commonly propounded methodology as given in Fig. 1.1 (for similar diagrams see Intriligator (1978), Koutsoyiannis (1977), *inter alia*). In order to explain the procedure represented by the diagram let us consider an extensively researched theoretical relationship, the transactions demand for money.

There is a proliferation of theories related to the demand for money which are beyond the scope of the present discussion (for a survey see Fisher (1978)). For our purposes it suffices to consider a simple theory where the transactions demand for money depends on income, the price level and

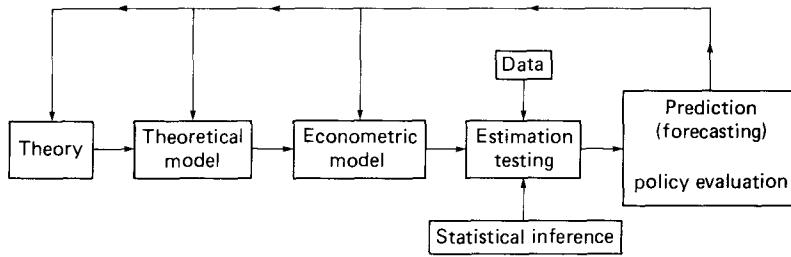


Fig. 1.1. The 'textbook' approach to econometric modelling.

interest rate, i.e.

$$M^D = f(Y, P, I). \quad (1.15)$$

Most theories of the demand for money can be accommodated in some variation of (15) by attributing different interpretations to  $Y$ . The theoretical model is a mathematical formulation of a theory. In the present case we expressed the theory directly in the functional form (15) in an attempt to keep the discussion to a minimum. Let the theoretical model be an explicit functional form for (15), say

$$M^D = AY^{\alpha_1}P^{\alpha_2}I^{\alpha_3} \quad (1.16)$$

or

$$\ln M^D = \alpha_0 + \alpha_1 \ln Y + \alpha_2 \ln P + \alpha_3 \ln I \quad (1.17)$$

in log-linear form with  $\alpha_0 = \ln A$  being a constant.

The next step in the methodological scheme represented by Fig. 1.1 is to transform the theoretical model (17) into an econometric model. This is commonly achieved in an interrelated sequence of steps which is rarely explicitly stated. Firstly, certain data series, assumed to represent measurements of the theoretical variables involved, are chosen. Secondly, the theoretical variables are assumed to coincide with the variables giving rise to the observed data chosen. This enables us to respecify (17) in terms of these observable variables, say,  $\tilde{M}_t$ ,  $\tilde{Y}_t$ ,  $\tilde{P}_t$  and  $\tilde{I}_t$ :

$$\ln \tilde{M}_t = \alpha_0 + \alpha_1 \ln \tilde{Y}_t + \alpha_2 \ln \tilde{P}_t + \alpha_3 \tilde{I}_t, \quad t \in \mathbb{T}. \quad (1.18)$$

The last step is to turn (18) into an econometric (statistical) model by attaching an error term  $u_t$  which is commonly assumed to be a normally distributed random variable of the form

$$u_t \sim N(0, \sigma^2), \quad t \in \mathbb{T}. \quad (1.19)$$

This error term is assumed to include either errors of measurement or/and

the effects of the excluded variables. Adding this error term onto (18) yields

$$\tilde{m}_t = \alpha_0 + \alpha_1 \tilde{y}_t + \alpha_2 \tilde{p}_t + \alpha_3 \tilde{l}_t + u_t, \quad t \in \mathbb{T}, \quad (1.20)$$

where small letters represent the logarithm of the corresponding capital letters. Equation (20) is now viewed as a Gauss linear model with the estimation, testing and prediction techniques related to this at our disposal to analyse ‘the transactions demand for money’.

The next stage is to estimate (20) using the statistical results related to the Gauss linear model and test the postulated assumptions for the error term. If any of the assumptions are invalid we correct by respecifying the error term and then we proceed to test the a priori restrictions suggested by the theory such as,  $\alpha_1 \simeq 1$ ,  $\alpha_2 \simeq 1$ ,  $-1 < \alpha_3 < 0$ , using the statistical techniques related to the linear model. When we satisfy ourselves that the theory is ‘correct’ we can proceed to use the estimated equation for prediction or/and policy evaluation.

In practice the above methodological procedure turns out to be much more difficult to apply and applied econometricians find themselves having to use ‘illegitimate’ procedures in order to get something worth publishing. Such procedures include estimating dozens of equations like (20) with various combinations of possibly relevant variables as well as including a few lagged variables among the explanatory variables. This is most graphically described in the introduction of Leamer (1978):

I began thinking about these problems when I was a graduate student in economics at the University of Michigan, 1966–1970. At that time there was a very active group building an econometric model of the United States. As it happens, the econometric modelling was done in the basement of the building and the econometric theory courses were taught on the top floor (the third). I was perplexed by the fact that the same language was used in both places. Even more amazing was the transmogrification of particular individuals who wantonly sinned in the basement and metamorphosed into the highest of high priests as they ascended to the third floor.

In the same book Leamer went on to attempt a systematisation of these ‘illegitimate’ procedures using Bayesian techniques. The approach proposed below will show that some of these ‘illegitimate’ procedures are indeed appropriate ways to ‘tackle’ certain problems in the context of an alternative methodology (see Chapter 22).

In an attempt to motivate the underlying logic of the methodology to be sketched below let us return to Fig. 1.1 in order to consider some of the possible ‘weak links’ in the textbook methodology.

The first possible weakness of the textbook methodology is that the

starting point of econometric modelling is some theory. This arises because the intended scope of econometrics is narrowly defined as the ‘measurement of theoretical relationships’. Such a definition was rejected at the outset of the present book as narrow and misleading. Theories are developed not for the sake of theorising but in order to understand some observable phenomenon of interest. Hence, defining the intended scope of econometrics as providing numbers for our own constructions and ignoring the original aim of explaining phenomena of interest, restricts its scope considerably by attaching ‘blinkers’ to the modeller. In a nutshell, it presupposes that the only ‘legitimate information’ contained in the observed data chosen is what the theory allows. This presents the modeller with insurmountable difficulties at the statistical model specification stage when the data do not fit the ‘straightjacket’ chosen for them without their nature being taken into consideration. The problem becomes more apparent when the theoretical model is turned into a statistical (econometric) model by attaching a white-noise error term to a reinterpreted equation in terms of observable variables. It is naive to suggest that the statistical model should be the same *whatever* the observed data chosen. In order to see this let us consider the demand schedule at time  $t$  referred to in Section 1.1:

$$q_t^D = \alpha_0 + \alpha_1 p_t. \quad (1.21)$$

If the data refer to intentions  $q_{it}^D = q_t^D(\tilde{p}_{it})$ ,  $i = 1, 2, \dots, n$  which correspond to the hypothetical range of prices  $\tilde{p}_{it}$ ,  $i = 1, 2, \dots, n$  then the most appropriate statistical model in the context of which (21) can be analysed is indeed the Gauss linear model. This is because the way the observed data were generated was under conditions which resemble an experimental situation; the hypothetical prices  $p_{it}$ ,  $i = 1, 2, \dots, n$  were called out and the economic agents considered their intentions to buy at time  $t$ . This suggests that  $\alpha_0$  and  $\alpha_1$  can be estimated using

$$\hat{q}_{it}^D = \alpha_0 + \alpha_1 \tilde{p}_{it} + u_{it}, \quad i = 1, 2, \dots, n. \quad (1.22)$$

In Haavelmo’s categorisation,  $(\hat{q}_{it}^D, \tilde{p}_{it})$ ,  $i = 1, 2, \dots, n$  constitute observed data of type one; experimental-like situations, see Section 1.1. On the other hand, if the observed data come in the form of time series  $(\tilde{q}_t, \tilde{p}_t)$ ,  $t = 1, 2, \dots, T$  where  $\tilde{q}_t$  refers to quantities transacted and  $\tilde{p}_t$  the corresponding prices at time  $t$  then the data are of type two; generated by nature. In this case the Gauss linear model seems wholly inappropriate unless there exists additional information ensuring that

$$q_t^D(\tilde{p}_t) = \tilde{q}_t \quad \text{for all } t. \quad (1.23)$$

Such a condition is highly unlikely to hold given that in practice other

factors such as supply-side, historical and institutional will influence the determination of  $\tilde{q}_t$  and  $\tilde{p}_t$ . It is highly likely that the data  $(\tilde{q}_t, \tilde{p}_t), t = 1, 2, \dots, T$  when used in the context of the Gauss linear model,

$$\hat{q}_t = \beta_0 + \beta_1 \tilde{p}_t + u_t \quad (1.24)$$

will give rise to very misleading estimates for the theoretical parameters of interest  $x_0$  and  $x_1$  (see Chapter 19 for the demand for money). This is because the GM represented by the Gauss linear model bears little, if any, resemblance to the actual DGP which gave rise to the observed data  $(\tilde{q}_t, \tilde{p}_t)$ ,  $t = 1, 2, \dots, T$ . In order to account for this some alternative statistical model should be specified in this case (see Part IV for several such models). Moreover, in this case the theoretical model (21) might not be *estimable*. A moment's reflection suggests that without any additional information the estimable form of the model is likely to be an *adjustment process* (price or/and quantity). If the observed data have a distinct time dimension this should be taken into consideration in deciding the estimable form of the model as well as in specifying the statistical model in the context of which the latter will be analysed. The estimable form of the model is directly related to the observable phenomenon of interest which gave rise to the data (the actual DGP). More often than not the intended scope of the theory in question is not the demand schedule itself but the explanation of changes in prices and quantities of interest. In such a case a demand or/and a supply schedule are used as a means to explain price and quantity changes not as the intended scope of the theory.

In the context of the textbook methodology distinguishing between the theoretical and estimable models in view of the observed data seems totally unnecessary for three interrelated reasons:

- (i) the observed data are treated as an afterthought;
- (ii) the actual DGP has no role to play; and
- (iii) theoretical variables are assumed to coincide (one-to-one) with the observed data chosen.

As in the case of (21) above the theoretical variables do not correspond directly to a particular observed data series unless we generate the data ourselves by 'artificially isolating the economic phenomenon of interest from other influences' (see the Haavelmo (1944) quotation in Section 1.1). We only have to think of theoretical variables such as aggregate demand for money, income, price level and interest rates and dozens of available data series become possible candidates for measuring these variables. Commonly, none of these data series measures what the theoretical variables refer to. Proceeding to assume that what is estimable coincides with the theoretical model and the statistical model differs from these by a

white-noise error term *regardless of the observed data chosen*, can only lead to misleading conclusions.

The question which naturally arises at this stage is whether we can tackle some of the problems raised above in the context of an alternative methodological framework. In view of the apparent limitations of the textbook methodology any alternative framework should be flexible enough so as to allow the modeller to ask some of the questions raised above even though readily available answers might not always be forthcoming. With this in mind such a methodological framework should attribute an important role to the actual DGP in order to widen the intended scope of econometric modelling. Indeed, the estimable model should be interpreted as an approximation to the actual DGP. This brings the nature of the observed data at the centre of the scene with the statistical model being defined directly in terms of the random variables giving rise to the data and not the error term. The statistical model should be specified as a generalised description of the mechanism giving rise to the data, in view of the estimable model, because the latter is going to be analysed in its context. A sketch of such a methodological framework is given in Fig. 1.2. An important feature of this framework is that it can include the textbook methodology as a special case under certain conditions. When the actual DGP is 'designed' to resemble the conditions assumed by the theory in question (Haavelmo type one observed data) then the theoretical and estimable models could coincide and the statistical model could differ from these by a white-noise error. In general, however, we need to distinguish between them even though the estimable model might not be readily available in some cases such as the case of the transactions demand for money (see Chapter 23).

In order to be able to turn the above skeleton of a methodology into a fully fleshed framework we need to formulate some of the concepts involved in more detail and discuss its implementation at length. Hence, a more detailed discussion of this methodology is considered in the epilogue where the various components shown in Fig. 1.2 are properly defined and their role explained. In the meantime the following working definitions will suffice for the discussion which follows:

*Actual DGP*: the mechanism underlying the observable phenomena of interest.

*Theory*: a conceptual construct providing an idealised description of the phenomena within its intended scope which will enable us to seek explanations and predictions related to the actual DGP.

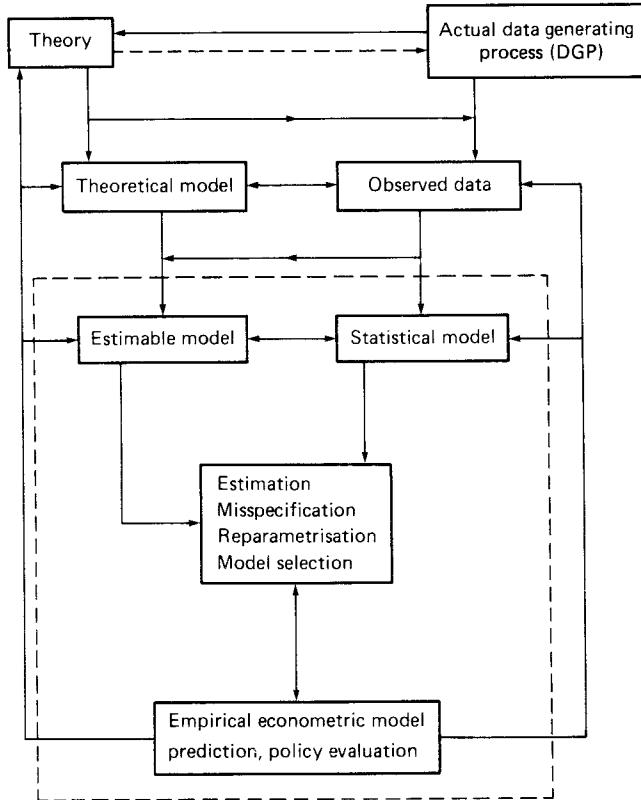


Fig. 1.2. An approach to econometric modelling.

*Theoretical model:* a mathematical formulation of the theory.

*Observed data:* observations on selected (by a theory) variables.

*Estimable model:* a particular form of the theoretical model which is potentially estimable in view of the actual DGP and the observed data chosen.

*Statistical model:* a probabilistic formulation purporting to provide a generalised description of the actual DGP with a view of analysing the estimable model in its context.

*Empirical econometric model:* a reformulation (reparametrisation/restriction) of a well-defined estimated statistical model in view of the estimable model which can be used for description, explanation or/and prediction.

### 1.3    Looking ahead

As the title of the book exemplifies, its main aim is the statistical foundations of econometric modelling. In relation to Fig. 1.2 the book concentrates mainly on the part within the dotted rectangle. The specification of a statistical model in terms of the variables giving rise to the observed data as well as the related statistical inference results will be the subject matter of Parts II and III. In Part IV various statistical models of interest in econometric modelling and the related statistical inference results will be considered in some detail. Special attention will be given to the procedure from the specification of the statistical model to the ‘design’ of the empirical econometric model. The transactions demand for money example considered above will be used throughout Part IV in an attempt to illustrate the ‘dangers’ awaiting the unaware in the context of the textbook methodology as well as compare this with the alternative methodology formalised in the present book.

Parts II and III form an integral part of econometric modelling and should not be viewed as providing a summary of the concepts and definitions to be used in Part IV. A sound background in probability theory and statistical inference is crucial for the implementation of the approach adopted in the present book. This is mainly because the modeller is required to specify the ‘appropriate’ statistical model taking into consideration the nature of the data in hand as well as the estimable model. This entails making decisions about characteristics of the random variables which gave rise to the observed data chosen such as normality, independence, stationarity, mixing, before any estimation is even attempted. This is one of the most crucial decisions in the context of econometric modelling because an inappropriate choice of the statistical model renders the related statistical inference conclusions invalid. Hence, the reader is advised to view Parts II and III as an integral part of econometric modelling and not as reference appendices. In Part IV the reader is encouraged to view econometric modelling as a thinking person’s activity and not as a sequence of technique recipes. Chapter 2 provides a very brief introduction to the Pearson paradigm in an attempt to motivate the Fisher paradigm which is the subject matter of Parts II and III.

#### **Additional references**

David (1962); Harter (1974–76); McAleer *et al.* (1985); Spanos (1985).

## CHAPTER 2

---

### Descriptive study of data

---

#### 2.1 Histograms and their numerical characteristics

By descriptive study of data we refer to the summarisation and exposition (tabulation, grouping, graphical representation) of observed data as well as the derivation of numerical characteristics such as measures of location, dispersion and shape.

Although the descriptive study of data is an important facet of modelling with real data in itself, in the present study it is mainly used to motivate the need for probability theory and statistical inference proper.

In order to make the discussion more specific let us consider the after-tax personal income data of 23 000 households for 1979–80 in the UK. These data in raw form constitute 23 000 numbers between £1000 and £50 000. This presents us with a formidable task in attempting to understand how income is distributed among the 23 000 households represented in the data. The purpose of descriptive statistics is to help us make some sense of such data. A natural way to proceed is to summarise the data by allocating the numbers into classes (intervals). The number of intervals is chosen a priori and it depends on the degree of summarisation needed. In the present case the income data are allocated into 15 intervals, as shown in Table 2.1 below (see *National Income and Expenditure* (1983)). The first column of the table shows the income intervals, the second column shows the number of incomes falling into each interval and the third column the relative frequency for each interval. The relative frequency is calculated by dividing the number of observations in each interval by the total number of observations. Summarising the data in Table 2.1 enables us to get some idea of how income is distributed among the various classes. If we plot the relative frequencies in a bar graph we get what is known as the *histogram*,

Table 2.1. Personal income in the UK, 1979–80

Income (£000)	No. of incomes	Relative frequency	Cumulative frequency
1.0–1.5	3440	0.150	0.150
1.5–2.0	2400	0.104	0.254
2.0–2.5	2320	0.101	0.355
2.5–3.0	2120	0.092	0.447
3.0–3.5	1990	0.086	0.533
3.5–4.0	1820	0.080	0.613
4.0–4.5	1500	0.065	0.678
4.5–5.0	2600	0.113	0.791
5.0–6.0	1890	0.082	0.873
6.0–7.0	1150	0.050	0.923
7.0–8.0	990	0.043	0.966
8.0–10.0	410	0.018	0.984
10.0–12.0	220	0.010	0.994
12.0–15.0	100	0.004	0.998
15.0–20.0	50	0.002	1.000
	23 000	1.000	

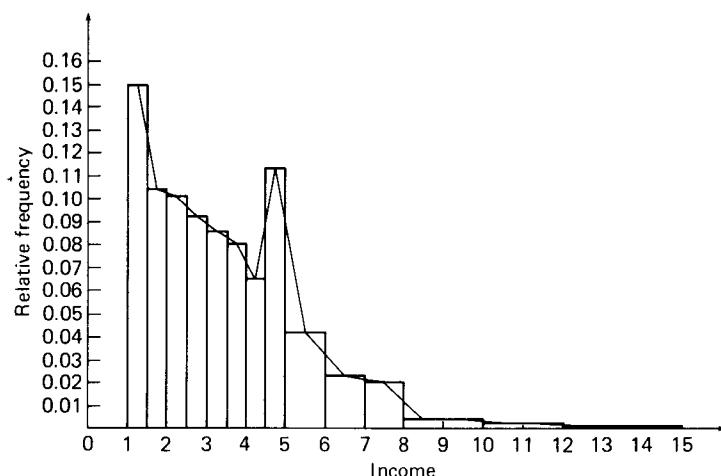


Fig. 2.1. The histogram and frequency polygon of the personal income data.

shown in Fig. 2.1. The pictorial representation of the relative frequencies gives us a more vivid impression of the distribution of income. Looking at the histogram we can see that most households earn less than £4500 and in some sense we can separate them into two larger groups: those earning between £1000 and £4500 and those above £4500. The first impression is

that the distribution of income inside these two larger groups appears to be rather similar.

For further information on the distribution of income we could calculate various numerical characteristics describing the histogram's location, dispersion and shape. Such measures can be calculated directly in terms of the raw data. However, in the present case it is more convenient for expositional purposes to use the grouped data. The main reason for this is to introduce various concepts which will be reinterpreted in the context of probability theory in Part II.

*The mean* as a measure of *location* takes the form

$$\bar{z} = \sum_{i=1}^{15} \phi_i z_i = 3.7, \quad (2.1)$$

where  $\phi_i$  and  $z_i$  refer to the relative frequency and the midpoint of interval  $i$ . *The mode* as a measure of location refers to the value of income that occurs most frequently in the data set. In the present case the mode belongs to the first interval £1.0–1.5. Another measure of location is the *median* referring to the value of income in the middle when incomes are arranged in an ascending (or descending) order according to the size of income. The best way to calculate the median is to plot the *cumulative frequency graph* which is more convenient for answering questions such as 'How many observations fall below a particular value of income?' (see Fig. 2.2). From the cumulative frequency graph we can see that the median belongs to the interval £3.0–3.5. Comparing the three measures of location we can see that

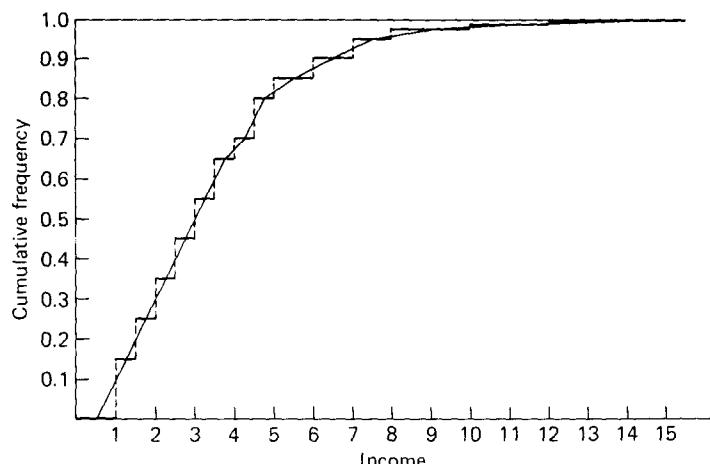


Fig. 2.2. The cumulative histogram and ogive of the personal income data.

## 26 Descriptive study of data

mode < median < mean, confirming the obvious asymmetry of the histogram.

Another important feature of the histogram is the dispersion of the relative frequencies around a measure of central tendency. The most frequently used measure of dispersion is the *variance* defined by

$$v^2 = \sum_{i=1}^{15} (z_i - \bar{z})^2 \phi_i = 4.85, \quad (2.2)$$

which is a measure of dispersion around the mean;  $v$  is known as the *standard deviation*.

We can extend the concept of the variance to

$$m_k = \sum_{i=1}^{15} (z_i - \bar{z})^k \phi_i, \quad k = 3, 4, \dots, \quad (2.3)$$

defining what are known as *higher central moments*. These higher moments can be used to get a better idea of the shape of the histogram. For example, the standardised form of the third and fourth moments defined by

$$SK = \frac{m_3}{v^3} \quad \text{and} \quad K = \frac{m_4}{v^4}, \quad (2.4)$$

known as the *skewness* and *kurtosis coefficients*, measure the asymmetry and peakedness of the histogram, respectively. In the case of a symmetric histogram,  $SK = 0$  and the less peaked the histogram the greater value of  $K$ . For the income data

$$SK = 1.43 \quad \text{and} \quad K = 7.33,$$

which confirms the asymmetry of the histogram (skewed to the right). The above numerical characteristics referring to the location, dispersion and shape were calculated for the data set as a whole. It was argued above, however, that it may be preferable to separate the data into two larger groups and study those separately. Let us consider the groups £1.0–4.5 and £4.0–20.0 separately. The numerical characteristics for the two groups are

$$\bar{z}_1 = 2.5, \quad v_1^2 = 0.996, \quad SK_1 = 0.252, \quad K_1 = 1.77$$

and

$$\bar{z}_2 = 6.18, \quad v_2^2 = 3.814, \quad SK_2 = 2.55, \quad K_2 = 11.93, \quad \text{respectively.}$$

Looking at these measures we can see that although the two subsets of the income data seemed qualitatively rather similar they actually differ substantially. The second group has much bigger dispersion, skewness and kurtosis coefficients.

Returning to the numerical characteristics of the data set as a whole we

can see that these seem to represent an uneasy compromise between the above two subsets. This confirms our first intuitive reaction based on the histogram that it might be more appropriate to study the two larger groups separately.

Another form of graphical representation for time-series data is the *time graph* ( $z_t, t$ ,  $t = 1, 2, \dots, T$ ). The temporal pattern of an economic time series is important not only in the context of descriptive statistics but also plays an important role in econometric modelling in the context of statistical inference proper; see Part IV.

## 2.2 Frequency curves

Although the histogram can be a very useful way to summarise and study observed data it is not a very convenient descriptor of data. This is because  $m - 1$  parameters  $\phi_1, \phi_2, \dots, \phi_{m-1}$  ( $m$  being the number of intervals) are needed to describe it. Moreover, analytically the histogram is a cumbersome step function of the form

$$h(z) = \sum_{i=1}^m \frac{\phi_i}{(z_{i+1} - z_i)} I([z_i, z_{i+1})), \quad z \in \mathbb{R}, \quad (2.5)$$

where  $[z_i, z_{i+1})$  represents the  $i$ th half-closed interval and  $I(\cdot)$  is the indicator function

$$I([z_i, z_{i+1})) = \begin{cases} 1 & \text{for } z \in [z_i, z_{i+1}) \\ 0 & \text{for } z \notin [z_i, z_{i+1}). \end{cases} \quad (2.6)$$

Hence, the histogram is not an ideal descriptor especially in relation to the modelling facet of observed data.

The first step towards a more convenient descriptor of observed data is the so-called *frequency polygon* which is a modified histogram. This is obtained by joining up the midpoints of the step function, as shown in Fig. 2.1, to get a continuous function.

An analogous graph for the cumulative frequency graph is known as the ogive (see Fig. 2.2). These two graphs can be interpreted as the histograms obtained by increasing the number of intervals. In summarising the data in the form of a histogram some information is lost. The greater the number of intervals the smaller the information lost. This suggests that increasing the number of intervals we might get more realistic descriptors for our data.

Intuition suggests that if we keep on increasing the number of intervals to infinity we should get a much smoother frequency curve. Moreover, with a smooth frequency curve we should be able to describe it in some functional form with fewer than  $m - 1$  parameters. For example, if we were to describe

the two subsets of the data separately we could conceivably be able to express a smoothed version of the frequency polygons in a polynomial form with one or two parameters. This line of reasoning led statisticians in the second part of the nineteenth century to suggest various such families of frequency curves with various shapes for describing observed data.

### ***The Pearson family of frequency curves***

In his attempt to derive a general family of frequency curves to describe observed data, Karl Pearson in the late 1890s suggested a family based on the differential equation

$$\frac{d(\log \phi(z))}{dz} = \frac{z+a}{b_0 + b_1 z + b_2 z^2}, \quad (2.7)$$

which satisfies the condition that the curve touches the  $z$ -axis at  $\phi(z)=0$  and has an optimum at  $z=-a$ , that is, the curve has one mode. Clearly, the solution of the above equation depends on the roots of the denominator. By imposing different conditions on these roots and choosing different values for  $a, b_0, b_1$  and  $b_2$  we can generate numerous frequency curves such as

$$(i) \quad \phi(z) = A_1(z - a_1)^{m_1}(z - a_2)^{m_2} \\ - \text{it can be bell-shaped, U-shaped or even J-shaped;} \quad (2.8)$$

$$(ii) \quad \phi(z) = A_2 \exp\left\{-\frac{z^2}{c}\right\} - \text{bell-shaped;} \quad (2.9)$$

$$(iii) \quad \phi(z) = A_3 \alpha z^{-\alpha+1} \quad - \text{J-shaped.} \quad (2.10)$$

In the case of the income data above we can see that the J-shaped (iii) frequency curve seems to be our best choice. As can be seen it has only one parameter  $\alpha$  and it is clearly a much more convenient descriptor (if appropriate) of the income data than the histogram. For  $A_3$  equal to the lowest income value this is known as the Pareto frequency curve. Looking at Fig. 2.1 we can see that for incomes greater than £4.5 the Pareto frequency curve seems a very reasonable descriptor.

An important property of the Pearson family of frequency curves is that the parameters  $\alpha, b_0, b_1$  and  $b_2$  are completely determined from knowledge of the first four moments. This implies that any frequency curve can be fitted to the data using these moments (see Kendall and Stuart (1969)). At this point, instead of considering how such frequency curves can be fitted to observed data we are going to leave the story unfinished to be taken up in Parts III and IV in order to look ahead to probability theory and statistical inference proper.

### 2.3 Looking ahead

The most important drawback of descriptive statistics is that the study of the observed data enables us to draw certain conclusions which relate *only* to the data in hand. The temptation in analysing the above income data is to attempt to make generalisations beyond the data in hand, in particular about the distribution of income in the UK. This, however, is not possible in the descriptive statistics framework. In order to be able to generalise beyond the data in hand we need ‘to model’ the distribution of income in the UK and not just ‘describe’ the observed data in hand. Such a general ‘model’ is provided by probability theory to be considered in Part II. It turns out that the model provided by probability theory owes a lot to the earlier developed descriptive statistics. In particular, most of the concepts which form the basis of the probability model were motivated by the descriptive statistics concepts considered above. The concepts of measures of location, dispersion and shape, as well as the frequency curve, were transplanted into probability theory with renewed interpretations. The frequency curve when reinterpreted becomes a density function purporting to model observable real world phenomena. In particular the Pearson family of frequency curves can be reinterpreted as a family of density functions. As for the various measures, they will now be reinterpreted in terms of the density function.

Equipped with the probability model to be developed in Part II we can go on to analyse observed data (now interpreted as generated by some assumed probability model) in the context of statistical inference proper; the subject matter of Part III. In such a context we can generalise beyond the observed data in hand. Probability theory and statistical inference will enable us to construct and analyse statistical models of particular interest in econometrics; the subject matter of Part IV.

In Chapter 2 we consider the axiomatic approach to probability which forms the foundation for the discussion in Part II. Chapter 3 introduces the concept of a random variable and related notions; arguably the most widely used concept in the present book. In Chapters 4–10 we develop the mathematical framework in the context of which the probability model could be analysed as a prelude to Part III.

#### Additional references

Bhattacharya and Johnson (1977); Haber and Runyon (1973); Johnson and Kotz (1970); Yeomans (1968).

## **PART II**

---

### **Probability theory**

---

## CHAPTER 3

---

### Probability

---

‘Why do we need probability theory in analysing observed data?’ In the descriptive study of data considered in the previous chapter it was emphasised that the results cannot be generalised outside the observed data under consideration. Any question relating to the population from which the observed data were drawn cannot be answered within the descriptive statistics framework. In order to be able to do that we need the theoretical framework offered by probability theory. In effect probability theory develops a *mathematical model* which provides the logical foundation of statistical inference procedures for analysing observed data.

In developing a mathematical model we must first identify the important features, relations and entities in the real world phenomena and then devise the concepts and choose the assumptions with which to project a generalised description of these phenomena; an idealised picture of these phenomena. The model as a consistent mathematical system has a ‘life of its own’ and can be analysed and studied without direct reference to real world phenomena. Moreover, by definition a model should not be judged as ‘true’ or ‘false’, because we have no means of making such judgments (see Chapter 26). A model can only be judged as a ‘good’ or ‘better’ approximation to the ‘reality’ it purports to explain if it enables us to come to grips with the phenomena in question. That is, whether in studying the model’s behaviour the patterns and results revealed can help us identify and understand the real phenomena within the theory’s intended scope.

The main aim of the present chapter is to construct a theoretical model for probability theory. In Section 3.1 we consider the notion of probability itself as a prelude to the axiomatisation of the concept in Section 3.2. The probability model developed comes in the form of a probability space ( $S, \mathcal{F}, P(\cdot)$ ). In Section 3.3 this is extended to a conditional probability space.

### 3.1    The notion of probability

The theory of probability had its origins in gambling and games of chance in the mid-seventeenth century and its early history is associated with the names of Huygens, Pascal, Fermat and Bernoulli. This early development of probability was rather sporadic and without any rigorous mathematical foundations. The first attempts at some mathematical rigour and a more sophisticated analytical apparatus than just combinatorial reasoning, are credited to Laplace, De Moivre, Gauss and Poisson (see Maistrov (1974)). Laplace proposed what is known today as the *classical* definition of probability:

*Definition 1*

*If a random experiment can result in  $N$  mutually exclusive and equally likely outcomes and if  $N_A$  of these outcomes result in the occurrence of the event  $A$ , then the **probability of A** is defined by*

$$P(A) = \frac{N_A}{N}. \quad (3.1)$$

To illustrate the definition let us consider the random experiment of tossing a fair coin twice and observing the face which shows up. The set of all equally likely outcomes is

$$S = \{(HT), (TH), (HH), (TT)\}, \quad \text{with } N = 4.$$

Let the event  $A$  be ‘observing at least one head ( $H$ )’, then

$$A = \{(HT), (TH), (HH)\}.$$

Since  $N_A = 3$ ,  $P(A) = \frac{3}{4}$ . Applying the classical definition in the above example is rather straightforward but in general it can be a tedious exercise in combinatorics (see Feller (1968)). Moreover, there are a number of serious shortcomings to this definition of probability, which render it totally inadequate as a foundation for probability theory. The obvious limitations of the classical approach are:

- (i)      it is applicable to situations where there is only a *finite* number of possible outcomes; and
- (ii)     the ‘equally likely’ condition renders the definition *circular*.

Some important random experiments, even in gambling games (in response to which the classical approach was developed) give rise to a set of infinite outcomes. For example, the game played by tossing a coin until it turns up heads gives rise to the infinite set of possible outcomes  $S = \{(H), (TH), (TTH), (TTTH), \dots\}$ ; it is conceivable that somebody could flip a coin indefinitely without ever turning up heads! The idea of ‘equally likely’ is

synonymous with ‘equally probable’, thus probability is defined using the idea of probability! Moreover, the definition is applicable to situations where an apparent ‘objective’ symmetry exists, which raises not only the question of circularity but also how this definition can be applied to the case of a biased coin or to consider the probability that next year’s rate of inflation in the UK will be 10%? Where are the ‘equally likely’ outcomes and which ones result in the occurrence of the event? These objections were well known even by the founders of this approach and since the 1850s several attempts have been made to resolve the problems related to the ‘equally likely’ presupposition and extend the area of applicability of probability theory.

The most influential of the approaches suggested in an attempt to tackle the problems posed by the classical approach are the so-called frequency and subjective approaches to probability. The *frequency approach* had its origins in the writings of Poisson but it was not until the late 1920s that Von Mises put forward a systematic account of the approach. The basic argument of the frequency approach is that probability does not have to be restricted to situations of apparent symmetry (equally likely) since the notion of probability should be interpreted as stemming from the observable stability of *empirical frequencies*. For example, in the case of a fair coin we say that the probability of  $A = \{H\}$  is  $\frac{1}{2}$ , not because there are two equally likely outcomes but because repeated series of large numbers of trials demonstrate that the empirical frequency of occurrence of  $A$  ‘converges’ to the limit  $\frac{1}{2}$  as the number of trials goes to infinity. If we denote by  $n_A$  the number of occurrences of an event  $A$  in  $n$  trials, then if

$$\lim_{n \rightarrow \infty} \left( \frac{n_A}{n} \right) = P_A, \quad (3.2)$$

we say that  $P(A) = P_A$ . Fig. 3.1 illustrates this notion for the case of  $A = \{H\}$  in a typical example of 100 trials. As can be seen, although there are some ‘wild fluctuations’ of the relative frequency for a small number of trials, as these increase the relative frequency tends to ‘settle’ (converge around  $\frac{1}{2}$ ).

Despite the fact that the frequency approach seems to be an improvement over the classical approach, giving objective status to the notion of probability by rendering it a property of real world phenomena, there are some obvious objections to it. ‘What is meant by “limit as  $n$  goes to infinity”?’ ‘How can we generate infinite sequences of trials?’ ‘What happens to phenomena where repeated trials are not possible?’

The *subjective approach* to probability renders the notion of probability a subjective status by regarding it as ‘degrees of belief’ on behalf of individuals assessing the uncertainty of a particular situation. The

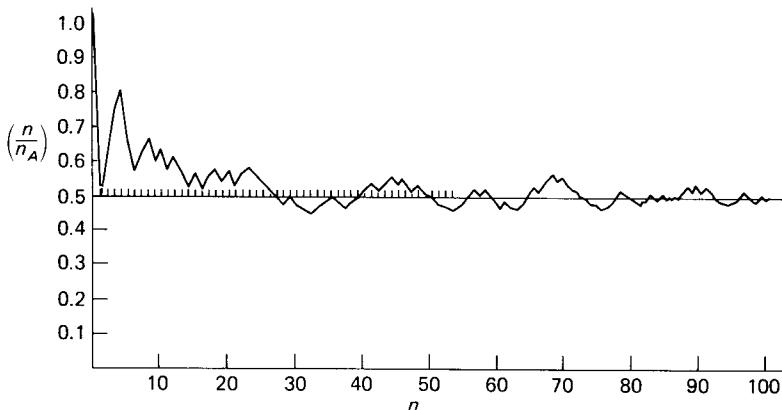


Fig. 3.1. Observed relative frequency of an experiment with 100 coin tossings.

protagonists of this approach are *inter alia* Ramsey (1926), de Finetti (1937), Savage (1954), Keynes (1921) and Jeffreys (1961); see Barnett (1973) and Leamer (1978) on the differences between the frequency and subjective approaches as well as the differences among the subjectivists.

Recent statistical controversies are mainly due to the attitudes adopted towards the frequency and subjective definitions of probability. Although these controversies are well beyond the material covered in this book, it is advisable to remember that the two approaches lead to alternative methods of statistical inference. The frequentists will conduct the discussion around what happens 'in the long-run' or 'on average', and attempt to develop 'objective' procedures which perform well according to these criteria. On the other hand, a subjectivist will be concerned with the question of revising prior beliefs in the light of the available information in the form of the observed data, and thus devise methods and techniques to answer such questions (see Barnett (1973)). Although the question of the meaning of probability was high on the agenda of probabilists from the mid-nineteenth century, this did not get in the way of impressive developments in the subject. In particular the systematic development of mathematical techniques related to what we nowadays call limit theorems (see Chapter 9). These developments were mainly the work of the Russian School (Chebyshev, Markov, Liapounov and Bernstein). By the 1920s there was a wealth of such results and probability began to grow into a systematic body of knowledge. Although various people attempted a systematisation of probability it was the work of the Russian mathematician Kolmogorov which proved to be the cornerstone for a systematic approach to

probability theory. Kolmogorov managed to relate the concept of probability to that of a measure in integration theory and exploited to the full the analogies between set theory and the theory of functions on the one hand and the concept of a random variable on the other. In a monumental monograph in 1933 he proposed an axiomatisation of probability theory establishing it once and for all as part of mathematics proper. There is no doubt that this monograph proved to be the watershed for the later development of probability theory growing enormously in importance and applicability. Probability theory today plays a very important role in many disciplines including physics, chemistry, biology, sociology and economics.

### 3.2 The axiomatic approach

The axiomatic approach to probability proceeds from a set of axioms (accepted without questioning as obvious), which are based on many centuries of human experience, and the subsequent development is built deductively using formal logical arguments, like any other part of mathematics such as geometry or linear algebra. In mathematics an axiomatic system is required to be complete, non-redundant and consistent. By complete we mean that the set of axioms postulated should enable us to prove every other theorem in the theory in question using the axioms and mathematical logic. The notion of non-redundancy refers to the impossibility of deriving any axiom of the system from the other axioms. Consistency refers to the non-contradictory nature of the axioms.

A probability model is by construction intended to be a description of a *chance mechanism* giving rise to observed data. The starting point of such a model is provided by the concept of a *random experiment* describing a simplistic and idealised process giving rise to observed data.

#### *Definition 2*

*A random experiment, denoted by  $\mathcal{E}$ , is an experiment which satisfies the following conditions:*

- (a) *all possible distinct outcomes are known a priori;*
- (b) *in any particular trial the outcome is not known a priori; and*
- (c) *it can be repeated under identical conditions.*

Although at first sight this might seem as very unrealistic, even as a model of a chance mechanism, it will be shown in the following chapters that it can be extended to provide the basis for much more realistic probability and statistical models.

The axiomatic approach to probability theory can be viewed as a formalisation of the concept of a *random experiment*. In an attempt to

formalise condition (a) all possible distinct outcomes are known a priori, Kolmogorov devised the set  $S$  which includes ‘all possible distinct outcomes’ and has to be postulated before the experiment is performed.

*Definition 3*

**The sample space**, denoted by  $S$ , is defined to be the set of all possible outcomes of the experiment  $\mathcal{E}$ . The elements of  $S$  are called **elementary events**.

*Example*

Consider the random experiment  $\mathcal{E}$  of tossing a fair coin twice and observing the faces turning up. The sample space of  $\mathcal{E}$  is

$$S = \{(HT), (TH), (HH), (TT)\},$$

with  $(HT)$ ,  $(TH)$ ,  $(HH)$ ,  $(TT)$  being the elementary events belonging to  $S$ .

The second ingredient of  $\mathcal{E}$  to be formulated relates to (b) and in particular to the various forms events can take. A moment’s reflection suggests that there is no particular reason why we should be interested in elementary outcomes only. For example, in the coin experiment we might be interested in such events as  $A_1$  – ‘at least one  $H$ ’,  $A_2$  – ‘at most one  $H$ ’ and these are not elementary events; in particular

$$A_1 = \{(HT), (TH), (HH)\}$$

and

$$A_2 = \{(HT), (TH), (TT)\}$$

are combinations of elementary events. All such outcomes are called *events* associated with the sample space  $S$  and they are defined by ‘combining’ elementary events. Understanding the concept of an event is crucial for the discussion which follows. Intuitively an event is any proposition associated with  $\mathcal{E}$  which may occur or not at each trial. We say that event  $A_1$  occurs when any one of the elementary events it comprises occurs. Thus, when a trial is made only one elementary event is observed but a large number of events may have occurred. For example, if the elementary event  $(HT)$  occurs in a particular trial,  $A_1$  and  $A_2$  have occurred as well.

Given that  $S$  is a set with members the elementary events this takes us immediately into the realm of set theory and events can be formally defined to be subsets of  $S$  formed by set theoretic operations (‘ $\cup$ ’ – union, ‘ $\cap$ ’ – intersection, ‘ ${}^c$ ’ – complementation) on the elementary events (see Binmore

(1980)). For example,

$$A_1 = \{(HT)\} \cup \{(TH)\} \cup \{(HH)\} = \{\overline{TT}\} \subset S,$$

i.e. ‘two tails’ does *not* occur,

$$A_2 = \{(HT)\} \cup \{(TH)\} \cup \{(TT)\} = \{\overline{HH}\} \subset S,$$

i.e. ‘two heads’ does *not* occur.

Two special events are  $S$  itself, called the *sure event* and the *impossible event*  $\emptyset$  defined to contain no elements of  $S$ , i.e.  $\emptyset = \{\}$ ; the latter is defined for completeness.

A third ingredient of  $\mathcal{E}$  associated with (b) which Kolmogorov had to formalise was the idea of uncertainty related to the outcome of any particular trial of  $\mathcal{E}$ . This he formalised in the notion of probabilities attributed to the various events associated with  $\mathcal{E}$ , such as  $P(A_1)$ ,  $P(A_2)$ , expressing the ‘likelihood’ of occurrence of these events. Although attributing probabilities to the elementary events presents no particular mathematical problems, doing the same for events in general is not as straightforward. The difficulty arises because if  $A_1$  and  $A_2$  are events  $\bar{A}_1 = S - A_1$ ,  $\bar{A}_2 = S - A_2$ ,  $A_1 \cup A_2$ ,  $A_1 \cap A_2$ ,  $A_1 - A_2$ , etc., are also events because the occurrence or non-occurrence of  $A_1$  and  $A_2$  implies the occurrence or not of these events. This implies that for the attribution of probabilities to make sense we have to impose some mathematical structure on the set of all events, say  $\mathcal{F}$ , which reflects the fact that whichever way we combine these events, the end result is always an event. The temptation at this stage is to define  $\mathcal{F}$  to be the set of all subsets of  $S$ , called the *power set*; surely this covers all possibilities! In the above example the power set of  $S$  takes the form

$$\begin{aligned} \mathcal{F} = \{S, \emptyset, & \{(HT)\}, \{(TH)\}, \{(HH)\}, \{(TT)\}, \{(TH), (HT)\}, \\ & \{(TH), (HH)\}, \{(TH), (TT)\}, \{(HT), (HH)\}, \{(HT), (TT)\}, \\ & \{(HH), (TT)\}, \{(HT), (TH), (HH)\}, \{(HT), (TH), (TT)\}, \\ & \{(HH), (TT), (TH)\}, \{(HH), (TT), (HT)\}\}. \end{aligned}$$

It can be easily checked that whichever way we combine any events in  $\mathcal{F}$  we end up with events in  $\mathcal{F}$ . For example,

$$\{(HH), (TT)\} \cap \{(TH), (HT)\} = \emptyset \in \mathcal{F},$$

$$\{(HH), (TH)\} \cup \{(TH), (HT)\} = \{(HH), (TH), (HT)\} \in \mathcal{F}, \text{ etc.}$$

It turns out that in most cases where the power set does not lead to any inconsistencies in attributing probabilities we define the set of events  $\mathcal{F}$  to be the power set of  $S$ . But when  $S$  is infinite or uncountable (it has as many

elements as there are real numbers) or we are interested in some but not all possible events, inconsistencies can arise. For example, if  $S = \{A_1, A_2, \dots\}$  such that  $A_i \cap A_j = \emptyset$  ( $i \neq j$ ),  $i, j = 1, 2, \dots$ ,  $\bigcup_{n=1}^{\infty} A_i = S$  and  $P(A_i) = a > 0, \forall i$ , where  $P(A)$  refers to the probability assigned to the event  $A$ . Then  $P(S) = \sum_{n=1}^{\infty} P(A_i) = \sum_{n=1}^{\infty} a > 1$  (see below), which is an absurd probability, being greater than one; similar inconsistencies arise when  $S$  is uncountable. Apart from these inconsistencies sometimes we are not interested in all the subsets of  $S$ . Hence, we need to define  $\mathcal{F}$  independently of the power set by endowing it with a mathematical structure which ensures that no inconsistencies arise. This is achieved by requiring that  $\mathcal{F}$  has a special mathematical structure, it is a  $\sigma$ -field related to  $S$ .

#### Definition 4

Let  $\mathcal{F}$  be a set of subsets of  $S$ .  $\mathcal{F}$  is called a  **$\sigma$ -field** if:

- (i)  $A \in \mathcal{F}$ , then  $\bar{A} \in \mathcal{F}$  – closure under complementation; and
- (ii)  $A_i \in \mathcal{F}, i = 1, 2, \dots$ , then  $(\bigcup_{i=1}^{\infty} A_i) \in \mathcal{F}$  – closure under countable union.

Note that (i) and (ii) taken together imply the following:

- (iii)  $S \in \mathcal{F}$ , because  $A \cup \bar{A} = S$ ;
- (iv)  $\emptyset \in \mathcal{F}$  (from (iii)  $\bar{S} = \emptyset \in \mathcal{F}$ ); and
- (v)  $A_i \in \mathcal{F}, i = 1, 2, \dots$ , then  $(\bigcap_{i=1}^{\infty} A_i) \in \mathcal{F}$ .

These suggest that a  $\sigma$ -field is a set of subsets of  $S$  which is closed under complementation, and countable unions and intersections. That is, any of these operations on the elements of  $\mathcal{F}$  will give rise to an element of  $\mathcal{F}$ . It can be checked that the power set of  $S$  is indeed a  $\sigma$ -field, and so is the set

$$\mathcal{F}_1 = \{\{(HT)\}, \{(HH), (TH), (TT)\}, \emptyset, S\},$$

but the set  $C = \{\{(HT), (TH)\}\}$  is not because  $\emptyset \notin C, S \notin C, \{(HT), (TH)\} \notin C$ . What we can do, however, in the latter case is to start from  $C$  and construct the *minimal  $\sigma$ -field* generated by its elements. This can be achieved by extending  $C$  to include all the events generated by set theoretic operations (unions, intersections, complementations) on the elements of  $C$ . Then the minimal  $\sigma$ -field generated by  $C$  is  $\mathcal{F}_c = \{S, \emptyset, \{(HT), (TH)\}, \{(HH), (TT)\}\}$  and we denote it by  $\mathcal{F}_c = \sigma(C)$ .

This way of constructing a  $\sigma$ -field can be very useful in cases where the events of interest are fewer than the ones given by the power set in the case of a finite  $S$ . For example, if we are interested in events with one of each  $H$  or  $T$  there is no point in defining the  $\sigma$ -field to be the power set, and  $\mathcal{F}_c$  can do as well with fewer events to attribute probabilities to. The usefulness of this method of constructing  $\sigma$ -fields is much greater in the case where  $S$  is either infinite or uncountable; in such cases this method is indispensable. Let us

consider an example where  $S$  is uncountable and discuss the construction of such a  $\sigma$ -field.

*Example*

Let  $S$  be the real line  $\mathbb{R} = \{x: -\infty < x < \infty\}$  and the set of events of interest be

$$J = \{B_x: x \in \mathbb{R}\} \quad \text{where } B_x = \{z: z \leq x\} = (-\infty, x]. \quad (3.3)$$

This is an educated choice, which will prove to be very useful in the sequel.

How can we construct a  $\sigma$ -field on  $\mathbb{R}$ ? The definition of a  $\sigma$ -field suggests that if we start from the events  $B_x$ ,  $x \in \mathbb{R}$  then extend this set to include  $\bar{B}_x$  and take countable unions of  $B_x$  and  $\bar{B}_x$  we should be able to define a  $\sigma$ -field on  $\mathbb{R}$ ,  $\sigma(J)$  – the *minimal  $\sigma$ -field generated by the events  $B_x$ ,  $x \in \mathbb{R}$* . By definition  $B_x \in \sigma(J)$ . If we take complements of  $B_x$ :  $\bar{B}_x = \{z: z \in \mathbb{R}, z > x\} = (x, \infty) \in \sigma(J)$ . Taking countable unions of  $B_x$ :  $\bigcup_{n=1}^{\infty} (-\infty, x - (1/n)] = (-\infty, x) \in \sigma(J)$ . These imply that  $\sigma(J)$  is indeed a  $\sigma$ -field. In order to see how large a collection  $\sigma(J)$  is we can show that events of the form  $(x, \infty)$ ,  $[x, \infty)$ ,  $(x, z)$  for  $x < z$ , and  $\{x\}$  also belong to  $\sigma(J)$ , using set theoretic operations as follows:

$$(x, \infty) = \overline{(-\infty, x]} \in \sigma(J), \quad (3.4)$$

$$[x, \infty) = \overline{(-\infty, x)} \in \sigma(J), \quad (3.5)$$

$$(x, z) = \overline{(-\infty, x] \cup [z, \infty)} \in \sigma(J), \quad (3.6)$$

$$\{x\} = \bigcap_{n=1}^{\infty} \left( x, x - \frac{1}{n} \right] \in \sigma(J).$$

This shows not only that  $\sigma(J)$  is a  $\sigma$ -field but it includes almost every conceivable subset (event) of  $\mathbb{R}$ , that is, it coincides with the  $\sigma$ -field generated by *any* set of subsets of  $\mathbb{R}$ , which we denote by  $\mathcal{B}$ , i.e.  $\sigma(J) = \mathcal{B}$ . The  $\sigma$ -field  $\mathcal{B}$  will play a very important role in the sequel; we call it the *Borel field* on  $\mathbb{R}$ .

Having solved the technical problem of possible inconsistencies in attributing probabilities to events by postulating the existence of a  $\sigma$ -field  $\mathcal{F}$  associated with the sample space  $S$ , Kolmogorov went on to formalise the concept of probability itself.

*Definition 5*

**Probability** is defined as a **set function** on  $\mathcal{F}$  satisfying the following axioms:

*Axiom 1:*  $P(A) \geq 0$  for every  $A \in \mathcal{F}$ ;

*Axiom 2:*  $P(S) = 1$ ; and

*Axiom 3:*  $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$  if  $\{A_i\}_{i=1}^{\infty}$  is a sequence of mutually exclusive events in  $\mathcal{F}$  (that is,  $A_i \cap A_j = \emptyset$  for  $i \neq j$ ) – (called countable additivity).

In other words, probability is defined to be a set function with  $\mathcal{F}$  as its domain and the closed real line interval  $[0, 1]$  as its range, so that

$$P(\cdot) : \mathcal{F} \rightarrow [0, 1].$$

The first two axioms seem rather self-evident and are satisfied by both the classical as well as frequency definitions of probability. Hence, in some sense, the axiomatic definition of probability ‘overcomes’ the deficiencies of the other definitions by making the *interpretation* of probability dispensable for the mathematical model to be built. The third axiom is less obvious, stating that the probability of the union of unrelated events must be equal to the addition of their separate probabilities. For example, since  $\{(HT)\} \cap \{(HH)\} = \emptyset$ ,

$$\begin{aligned} P(\{(HT)\} \cup \{(HH)\}) &= P(\{(HT)\}) + P(\{(HH)\}) \\ &= \frac{1}{4} + \frac{1}{4} = \frac{1}{2}. \end{aligned}$$

Again this coincides with the ‘frequency interpretation’ result. To summarise the argument so far, Kolmogorov formalised the conditions (a) and (b) of the random experiment  $\mathcal{E}$  in the form of the trinity  $(S, \mathcal{F}, P(\cdot))$  comprising the set of all outcomes  $S$  – the sample space, a  $\sigma$ -field  $\mathcal{F}$  of events related to  $S$  and a probability function  $P(\cdot)$  assigning probabilities to events in  $\mathcal{F}$ . For the coin example, if we choose  $\mathcal{F} = \{\{(HT)\}, \{(TH), (HH), (TT)\}, \emptyset, S\}$  to be the  $\sigma$ -field of interest,  $P(\cdot)$  is defined by

$$P(S) = 1, \quad P(\emptyset) = 0, \quad P(\{(HT)\}) = \frac{1}{4}, \quad P(\{(TH), (HH), (TT)\}) = \frac{3}{4}.$$

Because of its importance the trinity  $(S, \mathcal{F}, P(\cdot))$  is given a name.

*Definition 6*

A sample space  $S$  endowed with a  $\sigma$ -field  $\mathcal{F}$  and a probability function satisfying axioms 1–3 is called a **probability space**.

As far as condition (c) of  $\mathcal{E}$  is concerned, yet to be formalised, it will prove of paramount importance in the context of the limit theorems in Chapter 9, as well as in Part III.

Having defined the basic axioms of the theory we can now proceed to derive more properties for the probability set function using these axioms and mathematical logic. Although such properties will not be used directly in constructing what we called a probability model, they will be used indirectly. For this reason some of these properties will be listed here for references without any proofs:

- (P1)  $P(\bar{A}) = 1 - P(A)$ ,  $A \in \mathcal{F}$ .
- (P2)  $P(\emptyset) = 0$ .
- (P3) If  $A_1 \subset A_2$ ,  $P(A_1) \leq P(A_2)$ ,  $A_1, A_2 \in \mathcal{F}$ .
- (P4)  $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$ .
- (P5) If  $\{A_n\}_{n=1}^{\infty}$  is a monotone sequence of events in  $\mathcal{F}$  then  $P(\lim_{n \rightarrow \infty} A_n) = \lim_{n \rightarrow \infty} P(A_n)$ .

A monotone sequence of events in  $\mathcal{F}$  can be either increasing (expanding) or decreasing (contracting), i.e.  $A_1 \subset A_2 \subset \dots \subset A_{n-1} \subset A_n \dots$  or  $A_1 \supset \dots \supset A_{n-1} \supset A_n$ , respectively. For an increasing sequence  $\lim_{n \rightarrow \infty} A_n = \bigcup_{n=1}^{\infty} A_n$  and for a decreasing sequence  $\lim_{n \rightarrow \infty} A_n = \bigcap_{n=1}^{\infty} A_n$ . P5 is known as the *continuity property* of the set function  $P(\cdot)$  and plays an important role in probability theory. In particular it ensures that the distribution function (see Section 4.2) satisfies certain required conditions; see also Section 8.4 on martingales.

### 3.3 Conditional probability

One important extension of the above formalisation of the random experiment  $\mathcal{E}$  in the form of the probability space  $(S, \mathcal{F}, P(\cdot))$  is in the direction of conditional probabilities. So far we have considered probabilities of events on the assumption that no information is available relating to the outcome of a particular trial. Sometimes, however, additional information is available in the form of the *known* occurrence of some event  $A$ . For example, in the case of tossing a fair coin twice we might know that in the first trial it was heads. What difference does this information make to the original triple  $(S, \mathcal{F}, P(\cdot))$ ? Firstly, knowing that the first trial was a head, the set of all possible outcomes now becomes

$$S_A = \{(HT), (HH)\},$$

since  $(TH), (TT)$  are no longer possible. Secondly, the  $\sigma$ -field taken to be the power set now becomes

$$\mathcal{F}_A = \{S_A, \emptyset, \{(HT)\}, \{(HH)\}\}.$$

Thirdly the probability set function becomes

$$P_A(S_A) = 1, \quad P_A(\emptyset) = 0, \quad P_A(\{(HT)\}) = \frac{1}{2}, \quad P_A(\{(HH)\}) = \frac{1}{2}.$$

Thus, knowing that the event  $A$  – ‘at least one  $H$ ’ has occurred (in the first trial) transformed the original probability space  $(S, \mathcal{F}, P(\cdot))$  to the *conditional probability space*  $(S_A, \mathcal{F}_A, P_A(\cdot))$ . The question that naturally arises is to what extent we can derive the above conditional probabilities without having to transform the original probability space. The following formula provides us with a way to calculate the conditional probability.

$$\blacktriangleright \quad P_A(A_1) = P(A_1 | A) = \frac{P(A_1 \cap A)}{P(A)}. \quad (3.7)$$

In order to illustrate this formula let  $A_1 = \{(HT)\}$  and  $A = \{(HT), (HH)\}$ , then since  $P(A_1) = \frac{1}{4}$ ,  $P(A) = \frac{1}{2}$ ,  $P(A_1 \cap A) = P(\{(HT)\}) = \frac{1}{4}$ ,

$$P_A(A_1) = P(A_1 | A) = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2},$$

as above.

Note that  $P(A) > 0$  for the conditional probabilities to be defined.

Using the above rule of conditional probability we can deduce that

$$\begin{aligned} P(A_1 \cap A_2) &= P(A_1 | A_2) \cdot P(A_2) \\ &= P(A_2 | A_1) \cdot P(A_1), \quad \text{for } A_1, A_2 \in \mathcal{F}. \end{aligned} \quad (3.8)$$

This is called the *multiplication rule*. Moreover, when knowing that  $A_2$  has occurred does not change the original probability of  $A_2$ , i.e.

$$P(A_1 | A_2) = P(A_1), \quad (3.10)$$

we say that  $A_1$  and  $A_2$  are *independent*.

Independence is very different from *mutual exclusiveness* in the sense that  $A_1 \cap A_2 = \emptyset$  but  $P(A_1 | A_2) \neq P(A_1)$  and vice versa can both arise. Independence is a probabilistic statement which ensures that the occurrence of one event does not influence the occurrence (or non-occurrence) of the other event. On the other hand, mutual exclusiveness is a statement which refers to the events (sets) themselves not the associated probabilities. Two events are said to be mutually exclusive when they cannot occur together (see exercise 4).

The careful reader would have noticed that the axiomatic approach to probability does not provide us with ways to calculate probabilities for individual events unlike the classical or frequency approaches. What it provides us with are relationships between the probabilities of certain events when the events themselves are related in some way. This is a feature of the axiomatic approach which allows us to construct a probability model without knowing the numerical values of the probabilities but still lets us deduce them from empirical evidence.

**Important concepts**

random experiment;  
classical, frequency and subjective definitions of probability;  
sample space, elementary events;  
 $\sigma$ -field, minimal  $\sigma$ -field generated by events, Borel field;  
probability set function, probability space  $(S, \mathcal{F}, P(\cdot))$ ;  
conditional probability, independent events, mutually exclusive events.

**Questions**

1. Why do we need probability theory in analysing observed data?
2. What is the role of a mathematical model in attempting to explain real phenomena?
3. Compare and contrast the classical and frequency definitions of probability. How do they differ from the axiomatic definition?
4. Explain how the axiomatic approach formalises the concept of a random experiment  $\mathcal{E}$  to that of a probability space  $(S, \mathcal{F}, P(\cdot))$ .
5. Why do we need the concept of a  $\sigma$ -field in the axiomatisation of probability? Explain the concept intuitively.
6. Explain the concept of the minimal  $\sigma$ -field generated by some events using the half-closed intervals  $(-\infty, x]$ ,  $x \in \mathbb{R}$  on the real line as an example.
7. Explain intuitively the continuity property of the probability set function  $P(\cdot)$ .
8. Discuss the concept of conditional probability and show that  $P(\cdot | A)$  for some  $A \in \mathcal{F}$  is a proper probability set function.

**Exercises**

1. Consider the random experiment of throwing a dice and you stand to lose money if the number of dots is odd. Derive a  $\sigma$ -field which will enable you to consider your interests probabilistically. Explain your choice.
2. Consider the random experiment of tossing two *indistinguishable* fair coins and observing the faces turning up.
  - (i) Derive the sample space  $S$ , the  $\sigma$ -field of the power set  $\mathcal{F}$  and define the probability set function  $P(\cdot)$ .
  - (ii) Derive the  $\sigma$ -field generated by the events  $\{HH\}$  and  $\{TT\}$ .
  - (iii) If you stand to lose a pound every time a coin turns up ‘heads’ what is the  $\sigma$ -field of interest?

- (iv) Consider the effect on  $(S, \mathcal{F}, P(\cdot))$  when knowing that event  $A$  ‘at least one  $T$ ’ has occurred and define the new conditional probability space  $(S_A, \mathcal{F}_A, P_A(\cdot))$ . Confirm that for the event  $A_1$  – two tails,

$$P_A(A_1) = \frac{P(A \cap A_1)}{P(A)}.$$

- (v) Consider the events  $\{HH\}$  and  $\{TT\}$  and show whether they are mutually exclusive or/and independent.
3. Consider the random experiment of tossing a coin until it turns up ‘heads’. Define the sample space and discuss the question of defining a  $\sigma$ -field associated with it.
4. Consider the random experiment of selecting a card at random from an ordinary deck of 52 cards.
- (i) Find the probability of  
 $A_1$  – the card is an ace;  
 and  
 $A_2$  – the card is a diamond.
- (ii) Knowing that the card is a diamond show how the original  $(S, \mathcal{F}, P(\cdot))$  changes and calculate the probability of  
 $A_3$  – the card is the ace of diamonds.
- (iii) Find  $P(A_1 \cap A_2)$  and compare it with the probability of  $A_3$  derived in (ii).
- (iv) Define two events which are:  
 (a) mutually exclusive and independent;  
 (b) mutually exclusive but not independent;  
 (c) not mutually exclusive but independent; and  
 (d) not mutually exclusive and not independent.

#### **Additional references**

Barnett (1973); Giri (1974); Mood, Graybill and Boes (1974); Pfeiffer (1978); Rohatgi (1976).

## CHAPTER 4

---

### Random variables and probability distributions

---

In the previous chapter the axiomatic approach provided us with a mathematical model based on the triplet  $(S, \mathcal{F}, P(\cdot))$  which we called a probability space, comprising a sample space  $S$ , an event space  $\mathcal{F}$  ( $\sigma$ -field) and a probability set function  $P(\cdot)$ . The mathematical model was not developed much further than stating certain properties of  $P(\cdot)$  and introducing the idea of conditional probability. This is because the model based on  $(S, \mathcal{F}, P(\cdot))$  does not provide us with a flexible enough framework. The main purpose of this section is to change this probability space by mapping it into a much more flexible one using the concept of a random variable.

The basic idea underlying the construction of  $(S, \mathcal{F}, P(\cdot))$  was to set up a framework for studying probabilities of events as a prelude to analysing problems involving uncertainty. The probability space was proposed as a formalisation of the concept of a random experiment  $\mathcal{E}$ . One facet of  $\mathcal{E}$  which can help us suggest a more flexible probability space is the fact that when the experiment is performed the outcome is often considered in relation to some *quantifiable attribute*; i.e. an attribute which can be represented by numbers. Real world outcomes are more often than not expressed in numbers. It turns out that assigning numbers to qualitative outcomes makes possible a much more flexible formulation of probability theory. This suggests that if we could find a consistent way to assign numbers to outcomes we might be able to change  $(S, \mathcal{F}, P(\cdot))$  to something more easily handled. The concept of a *random variable* is designed to do just that without changing the underlying probabilistic structure of  $(S, \mathcal{F}, P(\cdot))$ .

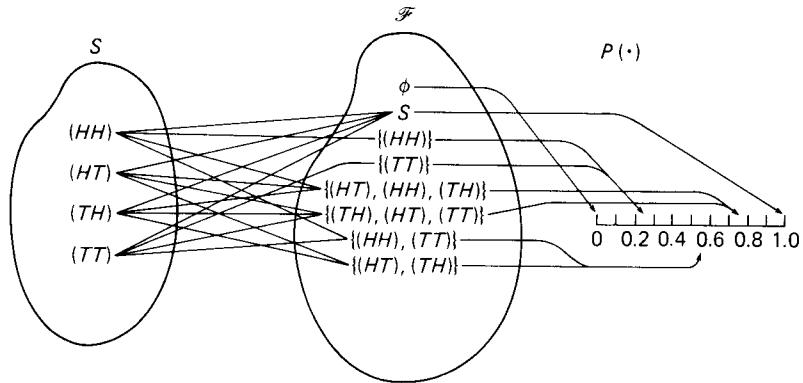


Fig. 4.1. The relationship between the sample space,  $\sigma$ -field and probability set function.

#### 4.1 The concept of a random variable

Fig. 4.1 illustrates the mathematical model \$(S, \mathcal{F}, P(\cdot))\$ for the coin-tossing example discussed in Chapter 3 with the  $\sigma$ -field of interest being \$\mathcal{F} = \{S, \emptyset, \{(HH)\}, \{(TT)\}, \{(HH, TT)\}, \{(HT), (HH), (TH)\}, \{(HT), (TH), (HH)\}, \{(HT), (TH), (TT)\}\}\$. The probability set function \$P(\cdot)\$ is defined on \$\mathcal{F}\$ and takes values in the interval \$[0, 1]\$, i.e. \$P(\cdot)\$ assigns probabilities to the events in \$\mathcal{F}\$. As can be seen, various combinations of the elementary events in \$S\$ define the  $\sigma$ -field \$\mathcal{F}\$ (ensure that it is a  $\sigma$ -field!) and the probability set function \$P(\cdot)\$ assigns probabilities to the elements of \$\mathcal{F}\$.

The main problem with the mathematical model \$(S, \mathcal{F}, P(\cdot))\$ is that the general nature of \$S\$ and \$\mathcal{F}\$ being defined as arbitrary sets makes the mathematical manipulation of \$P(\cdot)\$ very difficult; its domain being a  $\sigma$ -field of arbitrary sets. For example, in order to define \$P(\cdot)\$ we will often have to derive all the elements of \$\mathcal{F}\$ and tabulate it (a daunting task for large or infinite \$\mathcal{F}\$s), to say nothing about the differentiation or integration of such a set function.

Let us consider the possibility of defining a function \$X(\cdot)\$ which maps \$S\$ directly into the real line \$\mathbb{R}\$, that is,

$$X(\cdot): S \rightarrow \mathbb{R}_x, \quad (4.1)$$

assigning a real number \$x\_1\$ to each \$s\_1\$ in \$S\$ by \$x\_1 = X(s\_1)\$, \$x\_1 \in \mathbb{R}\$, \$s\_1 \in S\$. For example, in the coin-tossing experiment we could define the function \$X\$ – ‘the number of heads’. This maps all the elements of \$S\$ onto the set \$\mathbb{R}\_x = \{0, 1, 2\}\$, see Fig. 4.2.

The question arises as to whether every function from \$S\$ to \$\mathbb{R}\$ will provide

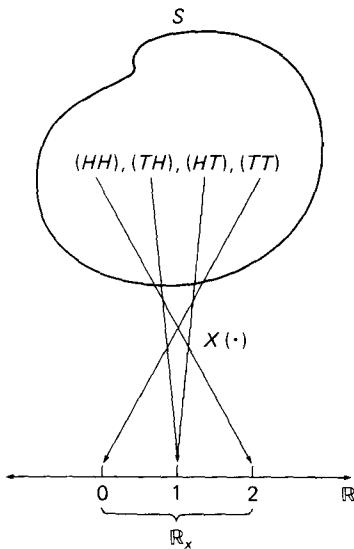


Fig. 4.2. The random variable  $X$ -number of 'heads' in the coin-tossing example.

us with a consistent way of attaching numbers to elementary events; consistent in the sense of preserving the event structure of the probability space  $(S, \mathcal{F}, P(\cdot))$ . The answer, unsurprisingly, is certainly not. This is because, although  $X$  is a function defined on  $S$ , probabilities are assigned to events in  $\mathcal{F}$  and the issue we have to face is how to define the values taken by  $X$  for the different elements of  $S$  in a way which preserves the event structure of  $\mathcal{F}$ . In order to illustrate this let us return to the earlier example. To each value of  $X$ , equal to 0, 1 and 2 there correspond some subset of  $S$ , i.e.

$$\begin{aligned} 0 &\rightarrow \{(TT)\}, \\ 1 &\rightarrow \{(TH), (HT)\}, \\ 2 &\rightarrow \{(HH)\}, \end{aligned}$$

and we denote it by

$$X^{-1}(0) = \{(TT)\}, \quad X^{-1}(1) = \{(TH), (HT)\}, \quad X^{-1}(2) = \{(HH)\},$$

using the inverse mapping  $X^{-1}(\cdot)$  ('inverse used by abuse of mathematical language'). What we require from  $X^{-1}(\cdot)$  (or  $X$ ) is to provide us with a correspondence between  $\mathbb{R}_x$  and  $S$  which reflects the event structure of  $\mathcal{F}$ , that is, it preserves unions, intersections and complements. In other words,

for each subset  $N$  of  $\mathbb{R}_x$  the inverse image  $X^{-1}(N)$  must be an event in  $\mathcal{F}$ . Looking at  $X$  as defined above we can see that  $X^{-1}(0) \in \mathcal{F}$ ,  $X^{-1}(1) \in \mathcal{F}$ ,  $X^{-1}(2) \in \mathcal{F}$ ,  $X^{-1}(\{0\} \cup \{1\}) \in \mathcal{F}$ ,  $X^{-1}(\{0\} \cup \{2\}) \in \mathcal{F}$ ,  $X^{-1}(\{1\} \cup \{2\}) \in \mathcal{F}$ , that is,  $X(\cdot)$  does indeed preserve the event structure of  $\mathcal{F}$ . On the other hand, the function  $Y(\cdot): S \rightarrow \mathbb{R}_y$  defined by  $Y(\{HT\}) = Y(\{HH\}) = 1$ ,  $Y(\{TH\}) = Y(\{TT\}) = 0$  does not preserve the event structure of  $\mathcal{F}$  since  $Y^{-1}(0) \notin \mathcal{F}$ ,  $Y^{-1}(1) \notin \mathcal{F}$ . This prompts us to define a random variable  $X$  to be any such function satisfying this *event preserving condition* in relation to some  $\sigma$ -field defined on  $\mathbb{R}_x$ ; for generality we always take the Borel field  $\mathcal{B}$  on  $\mathbb{R}$ .

### *Definition 1*

**A random variable**  $X$  is a real valued function from  $S$  to  $\mathbb{R}$  which satisfies the condition that for each Borel set  $B \in \mathcal{B}$  on  $\mathbb{R}$ , the set  $X^{-1}(B) = \{s: X(s) \in B, s \in S\}$  is an event in  $\mathcal{F}$ .

Three important features of this definition are worth emphasising.

- (i) A random variable is always defined relative to some specific  $\sigma$ -field  $\mathcal{F}$ .
- (ii) In deciding whether some function  $Y(\cdot): S \rightarrow \mathbb{R}$  is a random variable we proceed from the elements of the Borel field  $\mathcal{B}$  to those of the  $\sigma$ -field  $\mathcal{F}$  and not the other way around.
- (iii) A random variable is neither ‘random’ nor ‘a variable’.

Let us consider these important features in some more detail in order to enhance our understanding of the concept of a random variable; undoubtedly the most important concept in the present book.

The question ‘is  $X(\cdot): S \rightarrow \mathbb{R}$  a random variable?’ does not make any sense unless some  $\sigma$ -field  $\mathcal{F}$  is also specified. In the case of the function  $X$ —number of heads, in the coin-tossing example we see that it is a random variable relative to the  $\sigma$ -field  $\mathcal{F}$ , as defined in Fig. 4.1. On the other hand,  $Y$ , as defined above, is *not* a random variable relative to  $\mathcal{F}$ . This, however, does not preclude  $Y$  from being a random variable with respect to some other  $\sigma$ -field  $\mathcal{F}_y$ ; for instance  $\mathcal{F}_y = \{S, \emptyset, \{(HH), (HT)\}, \{(TH), (TT)\}\}$ . Intuition suggests that for any real valued function  $X(\cdot): S \rightarrow \mathbb{R}$  we should be able to define a  $\sigma$ -field  $\mathcal{F}_x$  on  $S$  such that  $X$  is a random variable. In the previous section we considered the  $\sigma$ -field generated by some set of events  $C$ . Similarly, we can generate  $\sigma$ -fields by functions  $X(\cdot): S \rightarrow \mathbb{R}$  which turn  $X(\cdot)$  into a random variable. Indeed  $\mathcal{F}_y$  above is the *minimal*  $\sigma$ -field generated by  $Y$ , denoted by  $\sigma(Y)$ . The way to generate such a minimal  $\sigma$ -field is to start from the set of events of the inverse mapping  $Y^{-1}(\cdot)$ , i.e.  $\{(HT), (HH)\} = Y^{-1}(1)$  and  $\{(TH), (TT)\} = Y^{-1}(0)$  and generate a  $\sigma$ -field by taking unions, intersections and complements. In the same way we can see

that the minimal  $\sigma$ -field generated by  $X$  – the number of heads,  $\sigma(X)$  coincides with the  $\sigma$ -field  $\mathcal{F}$  of Fig. 4.2; verify this assertion. In general, however, the  $\sigma$ -field  $\mathcal{F}$  associated with  $S$  on which a random variable  $X$  is defined does not necessarily coincide with  $\sigma(X)$ . Consider the function

$$\begin{aligned} X_1(\cdot) : S \rightarrow \mathbb{R} \\ X_1(\{(HH)\}) = X_1(\{(TH)\}) = X_1(\{(HT)\}) = 1, \quad X_1(\{(TT)\}) = 0 \end{aligned} \quad (4.2)$$

since  $X_1^{-1}(1) = \{(HH), (TH), (HT)\} \in \mathcal{F}$  (see Fig. 4.2),  $X_1^{-1}(0) = \{(TT)\} \in \mathcal{F}$ ,  $X_1^{-1}(\{0\}, \{1\}) = S \in \mathcal{F}$ .  $X_1$  is a random variable on  $S$  with respect to the  $\sigma$ -field  $\mathcal{F} = \sigma(X)$ . But  $\sigma(X_1) = \{S, \emptyset, \{(HH), (HT), (TH)\}\} \neq \mathcal{F}$ , indeed

$$\sigma(X_1) \subset \mathcal{F} = \sigma(X). \quad (4.3)$$

The above example is a special case of an important general result where  $X_1, X_2, \dots, X_n$  are random variables on the same probability space  $(S, \mathcal{F}, P(\cdot))$  and we define the new random variables

$$\begin{aligned} Y_1 &= X_1, \quad Y_2 = X_1 + X_2, \quad Y_3 = X_1 + X_2 + X_3, \quad \dots \\ Y_n &= X_1 + X_2 + \dots + X_n. \end{aligned} \quad (4.4)$$

If  $\sigma(Y_1), \sigma(Y_2), \dots, \sigma(Y_n)$  denote the minimal  $\sigma$ -fields generated by  $Y_1, Y_2, \dots, Y_n$  respectively, we can show that

$$\sigma(Y_1) \subset \sigma(Y_2) \subset \dots \subset \sigma(Y_n) \subseteq \mathcal{F}, \quad (4.5)$$

i.e.  $\sigma(Y_1), \dots, \sigma(Y_n)$  form an increasing sequence of  $\sigma$ -fields in  $\mathcal{F}$ . In the above example we can see that if we define a new random variable  $X_2(\cdot) : S \rightarrow \mathbb{R}$  by

$$X_2(\{(HH)\}) = 1, \quad X_2(\{(HT)\}) = X_2(\{(TH)\}) = X_2(\{(TT)\}) = 0,$$

then  $X = X_1 + X_2$  (see Table 4.1) is also a random variable relative to  $\sigma(X)$ ;  $X$  is defined as the number of Hs (see Table 4.1).

Note that  $X_1$  is defined as ‘at least one H’ and  $X_2$  as ‘two Hs’.

The above concept of  $\sigma$ -fields generated by random variables will prove very useful in the discussion of conditional expectation and martingales (see Chapters 7 and 8). The concept of a  $\sigma$ -field generated by a random variable enables us to concentrate on particular aspects of an experiment without having to consider everything associated with the experiment at the same time. Hence, when we choose to define a r.v. and the associated  $\sigma$ -field we make an implicit choice about the features of the random experiment we are interested in.

‘How do we decide that some function  $X(\cdot) : S \rightarrow \mathbb{R}$  is a random variable relative to a given  $\sigma$ -field  $\mathcal{F}$ ?’ From the above discussion of the concept of a

random variable it seems that if we want to decide whether a function  $X$  is a random variable with respect to  $\mathcal{F}$  we have to consider the Borel field  $\mathcal{B}$  on  $\mathbb{R}$  or at least the Borel field  $\mathcal{B}_x$  on  $\mathbb{R}_x$ ; a daunting task. It turns out, however, that this is not necessary. From the discussion of the  $\sigma$ -field  $\sigma(J)$  generated by the set  $J = \{B_x : x \in \mathbb{R}\}$  where  $B_x = (-\infty, x]$  we know that  $\mathcal{B} = \sigma(J)$  and if  $X(\cdot)$  is such that

$$X^{-1}((-\infty, x]) = \{s : X(s) \in (-\infty, x], s \in S\} \in \mathcal{F} \quad \text{for all } (-\infty, x] \in \mathcal{B}. \quad (4.6)$$

then

$$X^{-1}(B) = \{s : X(s) \in B, s \in S\} \in \mathcal{F} \quad \text{for all } B \in \mathcal{B}. \quad (4.7)$$

In other words, when we want to establish that  $X$  is a random variable or define  $P_x(\cdot)$  we have to look no further than the half-closed intervals  $(-\infty, x]$  and the  $\sigma$ -field  $\sigma(J)$  they generate, whatever the range  $\mathbb{R}_x$ . Let us use the shorthand notation  $\{X(s) \leq x\}$  instead of  $\{s : X(s) \in (-\infty, x], s \in S\}$  to consider the above argument in the case of  $X$  – the number of Hs, with respect to  $\mathcal{F}$  in Fig. 4.2,

$$X^{-1}((-\infty, x]) = \{s : X(s) \leq x\} \\ = \begin{cases} \emptyset, & x < 0 \\ \{(TT)\}, & 0 \leq x < 1 \\ \{(TH), (HT)\}, & 1 \leq x < 2 \\ \{(HH)\}, & 2 \leq x, \end{cases} \quad (4.8)$$

we can see that  $X^{-1}((-\infty, x]) \in \mathcal{F}$  for all  $x \in \mathbb{R}$  and thus  $X(\cdot)$  is a random variable with respect to  $\mathcal{F}$ . On the other hand,

$$Y^{-1}((-\infty, y]) = \{s : Y(s) \leq y\} \\ = \begin{cases} \emptyset, & y < 0 \\ \{(TH), (TT)\}, & 0 \leq y < 1 \\ \{(HT), (HH)\}, & 1 \leq y, \end{cases} \quad (4.9)$$

and thus  $Y^{-1}((-\infty, y]) \notin \mathcal{F}$  for  $y=0, y=1$ , i.e.  $Y(\cdot)$  is not a random variable with respect to  $\mathcal{F}$ . With respect to  $\mathcal{F}_y$ , however,  $\{s : Y(s) \leq y\} \in \mathcal{F}_y$  for all  $y \in \mathbb{R}$  and thus it is a random variable.

The term random variable is rather unfortunate because as can be seen from the above definition  $X$  is neither ‘random’ nor a ‘variable’; it is a real valued function and the notion of probability does not enter its definition. Probability enters the picture after the random variable has been defined in an attempt to complete the mathematical model induced by  $X$ .

Table 4.1

$S$	$X_1$	$X_2$	$X$
(HH)	1	1	2
(HT)	1	0	1
(TH)	1	0	1
(TT)	0	0	0

A random variable  $X$  relative to  $\mathcal{F}$  maps  $S$  into a subset of the real line, and the Borel field  $\mathcal{B}$  on  $\mathbb{R}$  plays now the role of  $\mathcal{F}$ . In order to complete the model we need to assign probabilities to the elements  $B$  of  $\mathcal{B}$ . Common sense suggests that the assignment of probabilities to the events  $B \in \mathcal{B}$  must be consistent with the probabilities assigned to the corresponding events in  $\mathcal{F}$ . Formally, we need to define a set function  $P_x(\cdot) : \mathcal{B} \rightarrow [0, 1]$  such that

$$P_x(B) = P(X^{-1}(B)) \equiv P(s : X(s) \in B, s \in S) \quad \text{for all } B \in \mathcal{B}. \quad (4.10)$$

For example, in the case illustrated in Table 4.1

$$P_x(\{0\}) = \frac{1}{4}, \quad P_x(\{1\}) = \frac{1}{2}, \quad P_x(\{2\}) = \frac{1}{4}, \quad P_x(\{0\} \cup \{1\}) = \frac{3}{4}, \quad \text{etc.,}$$

$$P_{x_1}(\{0\}) = \frac{1}{4}, \quad P_{x_1}(\{1\}) = \frac{3}{4}, \quad P_{x_1}(\{0\} \cup \{1\}) = 1, \quad P_{x_1}(\{0\} \cap \{1\}) = 0.$$

The question which arises is whether, in order to define the set function  $P_x(\cdot)$ , we need to consider all the elements of the Borel field  $\mathcal{B}$ . The answer is that we do not need to do that because, as argued above, any such element of  $\mathcal{B}$  can be expressed in terms of the semi-closed intervals  $(-\infty, x]$ . This implies that by choosing such semi-closed intervals ‘intelligently’, we can define  $P_x(\cdot)$  with the minimum of effort. For example,  $P_x(\cdot)$  for  $x$ , as defined in Table 4.1, can be defined as follows:

$$P_x((-\infty, x]) = \begin{cases} 0, & x < 0 \\ \frac{1}{4}, & 0 \leq x < 1 \\ \frac{3}{4}, & 1 \leq x < 2 \\ 1, & 2 \leq x. \end{cases} \quad (4.11)$$

As we can see, the semi-closed intervals were chosen to divide the real line at the points corresponding to the values taken by  $X$ . This way of defining the semi-closed intervals is clearly non-unique but it will prove very convenient in the next section.

The discerning reader will have noted that since we introduced the concept of a random variable  $X(\cdot)$  on  $(S, \mathcal{F}, P(\cdot))$  we have in effect

developed an alternative but equivalent probability space  $(\mathbb{R}, \mathcal{B}, P_x(\cdot))$  induced by  $X$ . The event and probability structure of  $(S, \mathcal{F}, P(\cdot))$  is preserved in the induced probability space  $(\mathbb{R}, \mathcal{B}, P_x(\cdot))$  and the latter has a much ‘easier to handle’ mathematical structure; we traded  $S$ , a set of arbitrary elements, for  $\mathbb{R}$ , the real line,  $\mathcal{F}$  a  $\sigma$ -field of subsets of  $S$  with  $\mathcal{B}$ , the Borel field on the real line, and  $P(\cdot)$  a set function defined on arbitrary sets with  $P_x(\cdot)$ , a set function on semi-closed intervals of the real line. In order to illustrate the transition from the probability space  $(S, \mathcal{F}, P(\cdot))$  to  $(\mathbb{R}_x, \mathcal{B}, P_x(\cdot))$  let us return to Fig. 4.1 and consider the probability space induced by the random variable  $X$ -number of heads, defined above. As can be seen from Fig. 4.3, the random variable  $X(\cdot)$  maps  $S$  into  $\{0, 1, 2\}$ . Choosing the semi-closed intervals  $(-\infty, 0]$ ,  $(-\infty, 1]$ ,  $(-\infty, 2]$  we can generate a Borel field on  $\mathbb{R}$  which forms the domain of  $P_x(\cdot)$ . The concept of a random variable enables us to assign numbers to arbitrary elements of a set ( $S$ ) and we choose to assign semi-closed intervals to events in  $\mathcal{F}$  as induced by  $X$ . By defining  $P_x(\cdot)$  over these semi-closed intervals we complete the procedure of assigning probabilities which is consistent with the one used in Fig. 4.1. The important advantage of the latter procedure is that the mathematical structure of the probability space  $(\mathbb{R}, \mathcal{B}, P_x(\cdot))$  is a lot more flexible as a framework for developing a probability model. The purpose of what follows in this part of the book is to develop such a flexible mathematical framework. It must be stressed, however, that the original probability space  $(S, \mathcal{F}, P(\cdot))$  has a role to play in the new mathematical framework both as a reference point and as the basis of the probability model we propose to build. Any new concept to be introduced has to be related to  $(S, \mathcal{F}, P(\cdot))$  to ensure that it makes sense in its context.

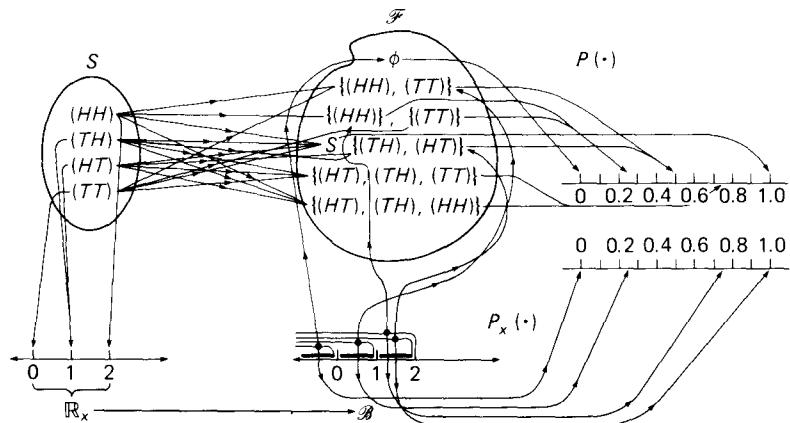


Fig. 4.3. The change from  $(S, \mathcal{F}, P(\cdot))$  to  $(\mathbb{R}_x, \mathcal{B}, P_x(\cdot))$  induced by  $X$ .

## 4.2 The distribution and density functions

In the previous section the introduction of the concept of a random variable (r.v.),  $X$ , enabled us to trade the probability space  $(S, \mathcal{F}, P(\cdot))$  for  $(\mathbb{R}, \mathcal{B}, P_x(\cdot))$  which has a much more convenient mathematical structure. The latter probability space, however, is not as yet simple enough because  $P_x(\cdot)$  is still a set function albeit on real line intervals. In order to simplify it we need to transform it into a point function (a function from a point to a point) with which we are so familiar.

The first step in transforming  $P_x(\cdot)$  into a point function comes in the form of the result discussed in the previous section, that  $P_x(\cdot)$  need only be defined on semi-closed intervals  $(-\infty, x]$ ,  $x \in \mathbb{R}$ , because the Borel field  $\mathcal{B}$  can be viewed as the minimal  $\sigma$ -field generated by such intervals. With this in mind we can proceed to argue that in view of the fact that all such intervals have a common starting ‘point’  $(-\infty)$  we could conceivably define a point function

$$F(\cdot): \mathbb{R} \rightarrow [0, 1], \quad (4.12)$$

which is, seemingly, only a function of  $x$ . In effect, however, this function will do exactly the same job as  $P_x(\cdot)$ . Heuristically, this is achieved by defining  $F(\cdot)$  as a point function by

$$P_x((-\infty, x]) = F(x) - F(-\infty), \quad \text{for all } x \in \mathbb{R}, \quad (4.13)$$

and assigning the value zero to  $F(-\infty)$ . Moreover, given that as  $x$  increases the interval it implicitly represents becomes bigger we need to ensure that  $F(x)$  is a non-decreasing function with one being its maximum value (i.e.  $F(x_1) \leq F(x_2)$  if  $x_1 \leq x_2$  and  $\lim_{x \rightarrow \infty} F(x) = 1$ ). For mathematical reasons we also require  $F(\cdot)$  to be continuous from the right.

*Definition 2*

*Let  $X$  be a r.v. defined on  $(S, \mathcal{F}, P(\cdot))$ . The point function  $F(\cdot): \mathbb{R} \rightarrow [0, 1]$  defined by*

$$F(x) = P_x((-\infty, x]) = Pr(X \leq x), \quad \text{for all } x \in \mathbb{R} \quad (4.14)$$

*is called the **distribution function** (DF) of  $X$  and satisfies the following properties:*

(i)  $F(x)$  is non-decreasing; (4.15)

(ii)  $F(-\infty) = \lim_{x \rightarrow -\infty} F(x) = 0$ ,  $F(\infty) = \lim_{x \rightarrow \infty} F(x) = 1$ , (4.16)

*and*

(iii)  $F(x)$  is continuous from the right

$$(i.e. \lim_{h \downarrow 0} F(x+h) = F(x), \forall x \in \mathbb{R}). \quad (4.17)$$

It can be shown (see Chung (1974)) that this defines a unique point function for every set function  $P_x(\cdot)$ .

The great advantage of  $F(\cdot)$  over  $P(\cdot)$  and  $P_x(\cdot)$  is that the former is a point function and can be represented in the form of an algebraic formula; the kind of functions we are so familiar with in elementary mathematics. This will provide us with a very convenient way of attributing probabilities to events.

Fig. 4.4 represents the graph of the DF of the r.v.  $X$  in the coin-tossing example discussed in the previous section, illustrating its properties in the case of a discrete r.v.  $X$ -number of *Hs*.

### *Definition 3*

A random variable  $X$  is called **discrete** if its range  $\mathbb{R}_x$  is some subset of the set of integers  $Z = \{0 \pm 1, \pm 2, \dots\}$ .

In this book we shall restrict ourselves to only two types of random variables, namely, discrete and (absolutely) continuous.

### *Definition 4*

A random variable  $X$  is called (absolutely) **continuous** if its distribution function  $F(x)$  is continuous for all  $x \in \mathbb{R}$  and there exists a non-negative function  $f(\cdot)$  on the real line such that

$$F(x) = \int_{-\infty}^x f(u) du, \quad \forall x \in \mathbb{R}. \quad (4.18)$$

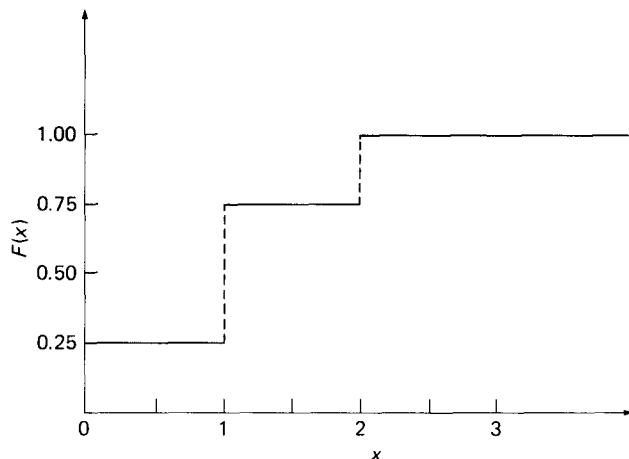


Fig. 4.4. The distribution function of  $X$ -number of 'heads'.

It must be stressed that for  $X$  to be continuous is not enough for the distribution function  $F(x)$  to be continuous. The above definition postulates that  $F(x)$  must also be derivable by integrating some non-negative function  $f(x)$ . So far the examples used to illustrate the various concepts referred to discrete random variables. From now on, however, emphasis will be placed almost exclusively on continuous random variables. The reason for this is that continuous random variables (r.v.'s) are susceptible to a more flexible mathematical treatment than discrete r.v.'s and this helps in the construction of probability models and facilitates the mathematical and statistical analysis.

In defining the concept of a continuous r.v. we introduced the function  $f(x)$  which is directly related to  $F(x)$ .

### *Definition 5*

Let  $F(x)$  be the DF of the r.v.  $X$ . The non-negative function  $f(x)$  defined by

$$F(x) = \int_{-\infty}^x f(u) du, \quad \forall x \in \mathbb{R} - \text{continuous} \quad (4.19)$$

or

$$F(x) = \sum_{u \leq x} f(u), \quad \forall x \in \mathbb{R} - \text{discrete} \quad (4.20)$$

is said to be the **(probability) density function (pdf) of  $X$** .

In the coin-tossing example,  $f(0)=\frac{1}{4}$ ,  $f(1)=\frac{1}{2}$ , and  $f(2)=\frac{1}{4}$  (see Fig. 4.5). In order to compare  $F(x)$  and  $f(x)$  for a discrete with those of a continuous r.v. let us consider the case where  $X$  takes values in the interval  $[a, b]$  and all values of  $X$  are attributed the same probability; we express this by saying that  $X$  is *uniformly distributed* in the interval  $[a, b]$  and we write  $X \sim U(a, b)$ . The DF of  $X$  takes the form

$$F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leq x < b \\ 1, & x \geq b \end{cases} \quad (4.21)$$

(see Fig. 4.6). The corresponding pdf of  $X$  is given by

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{elsewhere.} \end{cases} \quad (4.22)$$

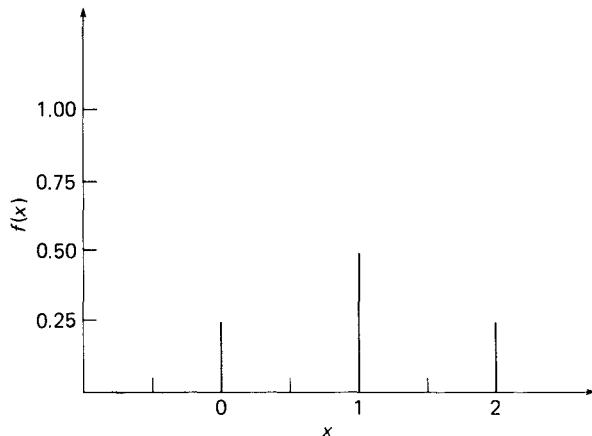
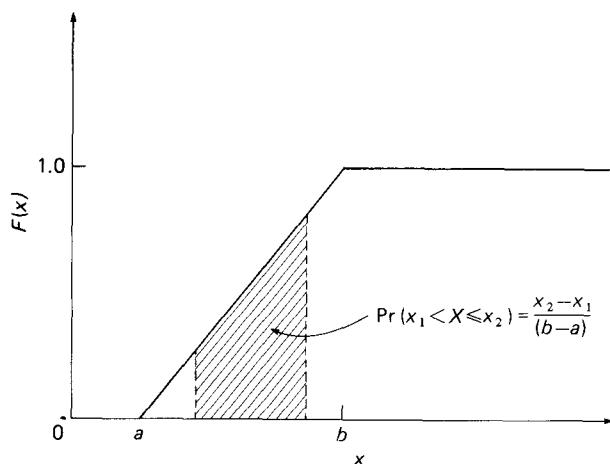
Fig. 4.5. The density function of  $X$ -number of 'heads'.

Fig. 4.6. The distribution function of a uniformly distributed random variable.

Comparing Figs. 4.4 and 4.5 with 4.6 and 4.7 we can see that in the case of a discrete random variable the DF is a step function and the density function attributes probabilities at discrete points. On the other hand, for a continuous r.v. the density function cannot be interpreted as attributing probabilities because, by definition, if  $X$  is a continuous r.v.  $P(X = x) = 0$  for all  $x \in \mathbb{R}$ . This can be seen from the definition of  $f(x)$  at every continuity

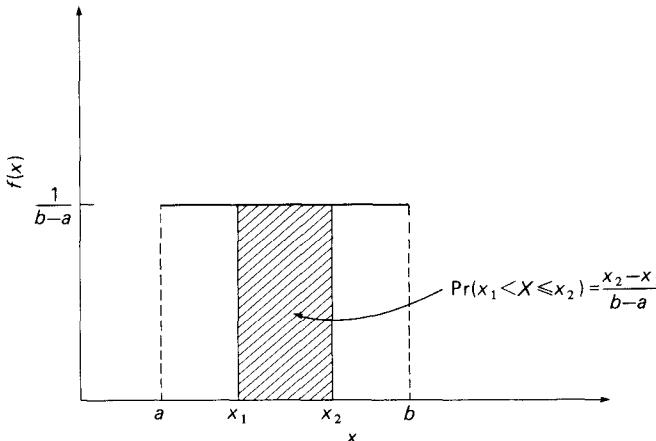


Fig. 4.7. The density function of a uniformly distributed random variable.

point of  $F(x)$ , where

$$f(x) = \frac{d}{dx} F(x) = \lim_{h \rightarrow 0} \frac{F(x+h) - F(x)}{h} = \lim_{h \rightarrow 0} \Pr(x < X \leq x+h), \quad h > 0. \quad (4.23)$$

Hence,

$$f(x) \cdot h \simeq \Pr(x < X \leq x+h), \quad \text{i.e. } f(\cdot) : \mathbb{R} \rightarrow [0, \infty). \quad (4.24)$$

Although we can use the distribution function  $F(x)$  as the fundamental concept of our probability model we prefer to sacrifice some generality and adopt the density function  $f(x)$  instead, because what we lose in generality we gain in simplicity and added intuition. It enhances intuition to view density functions as distributing probability *mass* over the range of  $X$ . The density function satisfies the following properties:

$$(i) \quad f(x) \geq 0, \quad \forall x \in \mathbb{R}; \quad (4.25)$$

$$(ii) \quad \int_{-\infty}^{\infty} f(x) dx = 1; \quad (4.26)$$

$$(iii) \quad \Pr(a < X < b) = \int_a^b f(x) dx; \quad (4.27)$$

$$(iv) \quad f(x) = \frac{d}{dx} F(x), \quad \text{at every point where the DF is continuous.} \quad (4.28)$$

Properties (ii) and (iii) can be translated for discrete r.v.'s by substituting ' $\sum_x$ ' for ' $\int \cdot dx$ '. It must be noted that a continuous r.v. is not one with a continuous DF  $F(\cdot)$ . *Continuity* refers to the condition that also requires the existence of a non-negative function  $f(\cdot)$  such that

$$F(x) = \int_{-\infty}^x f(u) du. \quad (4.29)$$

In cases where the distribution function  $F(x)$  is continuous but no integrating function  $f(x)$  exists, i.e.  $(d/dx)F(x)=0$  for some  $x \in \mathbb{R}$ , then  $F(x)$  is said to be a *singular distribution*. Singular distributions are beyond the scope of this book (see Chung (1974)).

### 4.3      The notion of a probability model

Let us summarise the discussion so far in order to put it in perspective. The axiomatic approach to probability formalising the concept of a random experiment  $\mathcal{E}$  proposed the probability space  $(S, \mathcal{F}, P(\cdot))$ , where  $S$  represents the set of all possible outcomes,  $\mathcal{F}$  is the set of events and  $P(\cdot)$  assigns probabilities to events in  $\mathcal{F}$ . The uncertainty relating to the outcome of a particular performance of  $\mathcal{E}$  is formalised in  $P(\cdot)$ . The concept of a random variable  $X$  enabled us to map  $S$  into the real line  $\mathbb{R}$  and construct an equivalent probability space induced by  $X, (\mathbb{R}, \mathcal{B}, P_x(\cdot))$ , which has a much 'easier to handle' mathematical structure, being defined on the real line. Although  $P_x(\cdot)$  is simpler than  $P(\cdot)$  it is still a set function albeit on the Borel field  $\mathcal{B}$ . Using the idea of  $\sigma$ -fields generated by particular sets of events we defined  $P_x(\cdot)$  on semi-closed intervals of the form  $(-\infty, x]$  and managed to define the point function  $F(\cdot)$ , the three being related by

$$P(s: X(s) \in (-\infty, x], s \in S) = P_x(-\infty, x] = F(x). \quad (4.30)$$

The distribution function  $F(x)$  was simplified even further by introducing the density function  $f(x)$  via  $F(x) = \int_{-\infty}^x f(u) du$ . This introduced further flexibility into the probability model because  $f(x)$  is definable in closed algebraic form. This enables us to transform the original uncertainty related to  $\mathcal{E}$  to uncertainty related to *unknown parameters*  $\theta$  of  $f(\cdot)$ ; in order to emphasise this we write the pdf as  $f(x; \theta)$ . We are now in a position to define our *probability model* in the form of a *parametric family of density functions* which we denote by

$$\Phi = \{f(x; \theta), \theta \in \Theta\}. \quad (4.31)$$

$\Phi$  represents a set of density functions indexed by the *unknown parameter(s)*  $\theta$  which are assumed to belong to a *parameter space*  $\Theta$  (usually a multiple of the real line). In order to illustrate these concepts let us consider an example

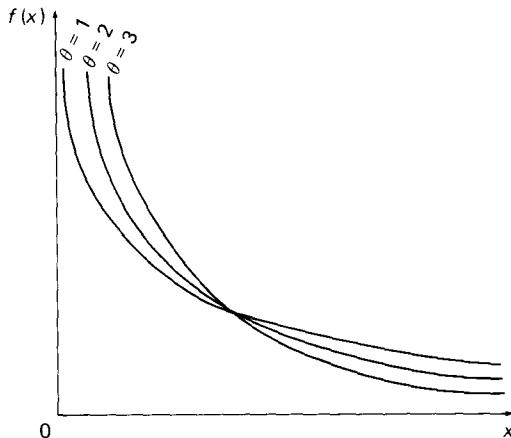


Fig. 4.8. The density function of a Pareto distributed random variable for different values of the parameter.

of a parametric family of density functions, the Pareto distribution:

$$\Phi = \left\{ f(x; \theta) = \frac{\theta}{x_0} \left( \frac{x_0}{x} \right)^{\theta+1}, x > 0, \theta \in \Theta \right\}, \quad (4.32)$$

$x_0$  – a known number  $\Theta = \mathbb{R}_+$  – the positive real line. For each value  $\theta$  in  $\Theta$ ,  $f(x; \theta)$  represents a different density (hence the term parametric family) as can be seen from Fig. 4.8.

When such a probability model is postulated it is intended as a description of the chance mechanism generating the observed data. For example, the model in Fig. 4.8 is commonly postulated in modelling personal incomes exceeding a certain level  $x_0$ . If we compare the above graph with the histogram of personal income data in Chapter 2 for incomes over £4500 we can see that postulating a Pareto probability density seems to be a reasonable model. In practice there are numerous such parametric families of densities we can choose from, some of which will be considered in the next section. The choice of one such family, when modelling a particular real phenomenon, is usually determined by previous experience in modelling similar phenomena or by a preliminary study of the data.

When a particular parametric family of densities  $\Phi$  is chosen, as the appropriate probability model for modelling a real phenomenon, we are in effect assuming that the observed data available were generated by the ‘chance mechanism’ described by one of those densities in  $\Phi$ . The original uncertainty relating to the outcome of a particular trial of the experiment

has now been transformed into the uncertainty relating to the choice of one  $\theta$  in  $\Theta$ , say  $\theta^*$ , which determines uniquely the one density, that is,  $f(x; \theta^*)$ , which gave rise to the observed data. The task of determining  $\theta^*$  or testing some hypothesis about  $\theta^*$  using the observed data lies with statistical inference in Part III. In the meantime, however, we need to formulate a mathematical framework in the context of which the probability model  $\Phi$  can be analysed and extended. This involves not only considering a number of different parametric families of densities, appropriate for modelling different real phenomena but also developing a mathematical apparatus which enables us to describe, compare, analyse and extend such models. The reader should keep this in mind when reading the following chapters to enable him/her not to lose sight of the woods for the trees. The woods comprise the above formulation of the probability model and its various generalisations and extensions, the trees are the various concepts and techniques which enable us to describe and analyse the probability model in its various formulations.

#### **4.4      Some univariate distributions†**

In the previous section we discussed how the concept of a random variable (r.v.)  $X$  defined on the probability space  $(S, \mathcal{F}, P(\cdot))$  enabled us to construct a general probability model in the form of a parametric family of densities (31). This is intended to be an appropriate mathematical model purporting to provide a good approximation of real phenomena in a stochastic (probabilistic) environment. In practice we need a menu of densities to describe different real phenomena and the purpose of this section is to consider a sample of such densities and briefly consider their applicability to such phenomena. For a complete menu and a thorough discussion see Johnson and Kotz (1969), (1970), (1972).

##### **(1)      Discrete distributions**

###### **(i) Bernoulli distribution**

Consider a random experiment  $\mathcal{E}_B$  where there are only two possible outcomes, we call ‘success’ and ‘failure’ for convenience, that is,  $S = \{\text{‘success’}, \text{‘failure’}\}$ . If we define on  $S$  the random variable  $X$  by  $X(\text{success}) = 1$ ,  $X(\text{failure}) = 0$  and postulate the probabilities  $Pr(X = 1) = p$  and  $Pr(X = 0) = 1 - p$  we can deduce that the density function of  $X$  takes the

† The term probability distribution is used to denote a set of probabilities on a complete system (a  $\sigma$ -field) of events.

form

$$\begin{aligned} f(x; p) &= p^x(1-p)^{1-x}, \quad \text{for } x=0, 1 \\ &= 0, \quad \text{otherwise.} \end{aligned} \tag{4.33}$$

In practice  $p$  is unknown and the probability model takes the form

$$\Phi = \{f(x; \theta) = \theta^x(1-\theta)^{1-x}, x=0, 1, \theta \in [0, 1]\}. \tag{4.34}$$

Such a probability model might be appropriate in modelling the sex of a newborn baby, boy or girl, or whether the next president of the USA will be a Democrat or a Republican.

### (ii) The binomial distribution

The binomial distribution is unquestionably the most important *discrete* distribution. It represents a direct extension of the Bernoulli distribution in the sense that the random experiment  $\mathcal{E}_B$  is repeated  $n$  times and we define the random variable  $Y$  to be the number of ‘successes’ in  $n$  trials. If we assume that the probability of ‘success’ is the same in all  $n$  trials the density of  $Y$  takes the form

$$f(y; p) = \begin{cases} \binom{n}{y} p^y(1-p)^{n-y}, & y=0, 1, \dots, n \\ 0, & \text{otherwise.} \end{cases} \tag{4.35}$$

and we denote this by writing  $Y \sim B(n, p)$ : ‘~’ reads ‘distributed as’.

Note that

$$\binom{n}{y} = \frac{n!}{(n-y)! y!}, \quad k! = k \cdot (k-1) \cdot (k-2) \cdots 2 \cdot 1. \tag{4.36}$$

The relationship between the Bernoulli and binomial distributions is of considerable interest. If the Bernoulli r.v. at the  $i$ th trial is denoted by  $X_i, i=1, 2, \dots, n$ , then  $Y$  is the summation of the  $X_i$ s, i.e.  $Y_n = X_1 + X_2 + \cdots + X_n$ ; the subscript in  $Y$  is used to emphasise its dependence on  $n$ . This is because the  $X_i$ s take the value 1 for a ‘success’ and 0 for a ‘failure’, which implies that  $\sum_{i=1}^n X_i$  represents the number of ‘successes’ in  $n$  trials. The interest in this relationship arises because the  $Y_i$ s generate a sequence of increasing  $\sigma$ -fields of the form  $\sigma(Y_1) \subset \sigma(Y_2) \subset \cdots \subset \sigma(Y_n)$ ;  $\sigma(Y_i)$  represents the  $\sigma$ -field generated by the r.f.  $Y_i$ . This is the property, known as *martingale condition* (see Section 8.4), that underlies a remarkable theorem known as the *De Moivre–Laplace*

*central limit theorem.* De Moivre and Laplace, back in the eighteenth century, realised that in order to calculate the probabilities  $f(y; n)$  for a large  $n$  the formula given above was rather impractical. In their attempt to find an easier way to calculate such probabilities they derived a very important approximation to the formula by showing that, for large  $n$ ,

$$\binom{n}{y} p^y (1-p)^{n-y} \simeq \frac{1}{\sqrt{[2\pi np(1-p)]}} e^{-\frac{1}{2}z^2}, \quad (4.37)$$

where

$$z = \frac{y - np}{\sqrt{[np(1-p)]}}, \quad \text{'$\simeq$' reads approximately equal.}$$

Using the formula on the RHS of the equality was much easier. This formula represents the density function of the most celebrated of all distributions which has a bell-shaped symmetric curve, the *normal*. Fig. 4.9 represents the graph of a binomial density for a variety of values for  $n$  and  $p$ . As we can see, as  $n$  increases the density function becomes more and more bell-shape like, especially when the value of  $P$  is around  $\frac{1}{2}$ . This result gave rise to one of the most important and elegant chapters in probability theory, the so-called limit theorems to be considered in Chapter 9.

## (2) Continuous distributions

### (i) The normal distribution

The normal distribution is by far the most important distribution in both probability theory and statistical inference. As seen above, De Moivre and Laplace regarded the distribution only as a convenient approximation to the binomial distribution. By the beginning of the nineteenth century, however, the work of Laplace, Legendre and Gauss on the theory of 'errors' placed the normal distribution at the centre of probability theory. It was found to be the most appropriate distribution for modelling a large number of experimental situations in astronomy, physics and eugenics. Moreover, the work of the Russian School (Chebyshev, Markov, Lyapounov and Kolmogorov) on limit theorems, relating to the behaviour of certain standardised sums of random variables, ensured a central role for the normal distribution.

A random variable  $X$  is normally distributed if its probability density function is given by

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, \quad x \in \mathbb{R}, \quad (4.38)$$

where  $\mu$  and  $\sigma^2$  are constant parameters with  $\mu \in \mathbb{R}$  and  $\sigma^2 \in \mathbb{R}_+$ ; we express

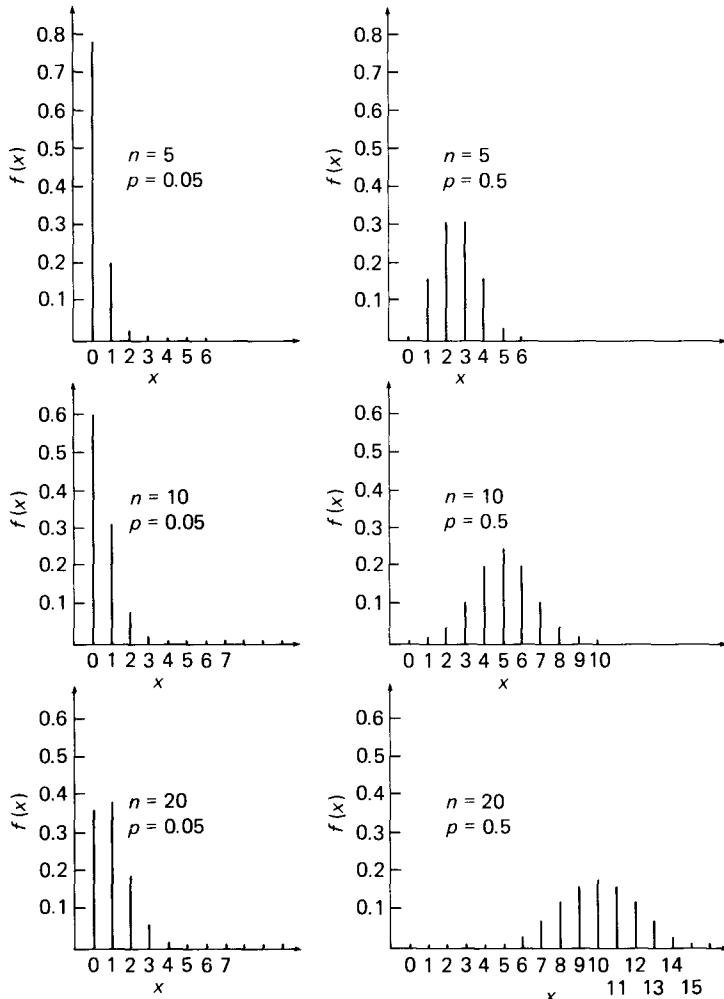


Fig. 4.9. The density function of a binomially distributed random variable for different values of the parameters  $n$  and  $p$ .

this by  $X \sim N(\mu, \sigma^2)$ . The parameters  $\mu$  and  $\sigma^2$  will be studied in more detail when we consider mathematical expectation. At this stage we will treat them as the parameters determining the location and flatness of the density. For a fixed  $\sigma^2$  the normal density for three different values of  $\mu$  is given in Fig. 4.10.

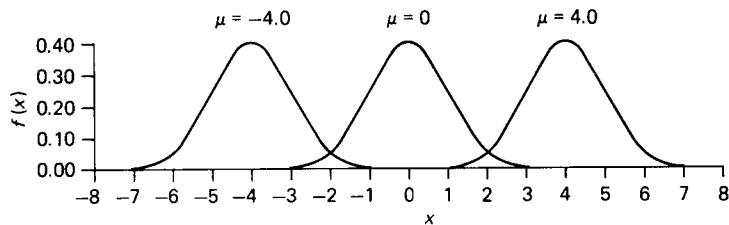


Fig. 4.10. The density function of a normally distributed random variable with  $\sigma^2 = 1$  and different values for the mean  $\mu$ .

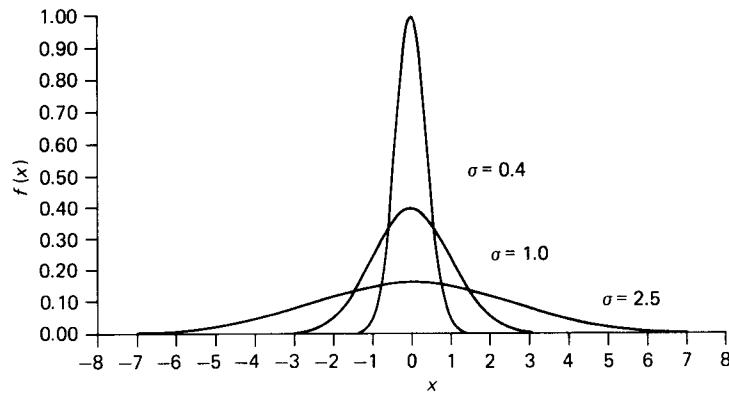


Fig. 4.11. The density function of a normally distributed random variable with mean  $\mu = 0$  and different values for the variance.

Fig. 4.11 represents the graph of the normal density for  $\mu = 0$  and three alternative values of  $\sigma$ ; as can be seen, the greater the value of  $\sigma$  the flatter the graph of the density. As far as the shape of the normal distribution and density functions are concerned we note the following characteristics:

- (i) The normal density is symmetric about  $\mu$ , i.e.

$$f(\mu + k) = \frac{1}{\sigma\sqrt{(2\pi)}} \exp\left\{-\frac{k^2}{2\sigma^2}\right\} = f(\mu - k), \quad (4.39)$$

$$\Rightarrow Pr(\mu \leq X \leq \mu + k) = Pr(\mu - k \leq X \leq \mu), \quad k > 0, \quad (4.40)$$

and for the DF,

$$F(u) = \frac{1}{\sigma\sqrt{(2\pi)}} \int_{-\infty}^u \exp\left\{-\frac{(x-u)^2}{2\sigma^2}\right\} du, \quad F(-x) = 1 - F(x+2\mu). \quad (4.41)$$

- (ii) The density function attains its maximum at  $x = \mu$ ,

$$\frac{df(x)}{dx} = f(x) \left( \frac{-2(x-\mu)}{2\sigma^2} \right) = 0 \Rightarrow x = \mu \quad \text{and} \quad f(\mu) = \frac{1}{\sigma\sqrt{(2\pi)}}. \quad (4.42)$$

- (iii) The density function has two points of inflection at  $x = \mu \pm \sigma$ ,

$$\frac{d^2f(x)}{dx^2} = \frac{\sigma^{-3}}{\sqrt{(2\pi)}} \exp \left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\} \left[ 1 - \frac{(x-\mu)^2}{\sigma^2} \right] = 0 \Rightarrow x = \mu \pm \sigma. \quad (4.43)$$

Thus,  $\sigma$  not only measures the flatness of the graph of the pdf but it determines the distance of the points of inflection from  $\mu$ . Fig. 4.12 represents the graphs of the normal DF and pdf in an attempt to give the reader some idea about the concentration of these functions in terms of the standard deviation parameter  $\sigma$  around the mean  $\mu$ .

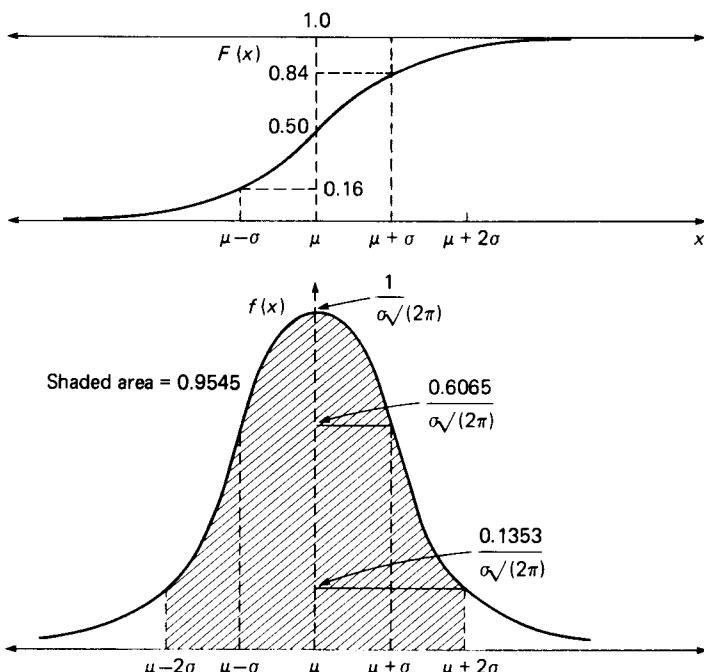


Fig. 4.12. Important features of the normal distribution and density functions.

The density function of the random variable  $Z = (X - \mu)/\sigma$  is

$$f(z) = \frac{1}{\sqrt{(2\pi)}} \exp\left\{-\frac{1}{2}z^2\right\}, \quad (4.44)$$

which does not depend on the unknown parameters  $\mu, \sigma$ . This is called the *standard normal distribution*, which we write in the form  $Z \sim N(0, 1)$ . This shows that any normal r.v. can be transformed to a standard normal r.v. when  $\mu$  and  $\sigma$  are known. The implication of this is that using the tables for  $f(z)$  and  $F(z)$  we can deduce  $f(x)$  and  $F(x)$  using the transformation  $Z = (X - \mu)/\sigma$ . For example, if we want to calculate  $Pr(X \leq 1.5)$  for  $X \sim N(1, 4)$  we form  $z = (x - 1)/2 = 0.25 \Rightarrow F(z) = F(0.25) = 0.5987$ , that is,  $F(x) = F(1.5) = 0.5987$ .

### (ii) Exponential family of distributions

Some of the most important distributions in probability theory, including the Bernoulli, binomial and normal distributions, belong to the exponential family of distributions. The exponential family is of considerable interest in statistical inference because several results in estimation and testing (see Chapters 11–14) depend crucially on the assumption that the underlying probability model is defined in terms of density functions belonging to this family; see Barndorff-Nielsen (1978).

## 4.5 Numerical characteristics of random variables

In modelling real phenomena using probability models of the form  $\Phi = \{f(x; \theta), \theta \in \Theta\}$  we need to be able to postulate such models having only a general quantitative description of the random variable in question at our disposal a priori. Such information comes in the form of certain numerical characteristics of random variables such as the mean, the variance, the skewness and kurtosis coefficients and higher moments. Indeed, sometimes such numerical characteristics actually determine the type of probability density in  $\Phi$ . Moreover, the analysis of density functions is usually undertaken in terms of these numerical characteristics.

### (1) Mathematical expectation

Let  $X$  be a random variable (r.v.) on  $(S, \mathcal{F}, P(\cdot))$  with  $F(x)$  and  $f(x)$  its distribution function (DF) and (probability) density function (pdf) respectively. The mean of  $X$  denoted by  $E(X)$  is defined by

$$E(X) = \int_{-\infty}^{\infty} xf(x) dx \text{ -- for a continuous r.v.} \quad (4.45)$$

and

$$E(X) = \sum_i x_i f(x_i) \quad - \text{for a discrete r.v.,} \quad (4.46)$$

when the integral and sum exist.

Note that the summation is over all possible values of  $X$ .

The integral in the definition of mathematical expectation for a continuous random variable can be interpreted as an improper Riemann integral. If a unifying approach to both discrete and continuous r.v.'s is required the concept of an improper Riemann–Stieltjes integral (see Clarke (1975))

$$E(X) = \int_{-\infty}^{\infty} x dF(x) \quad (4.47)$$

can be used. We sacrifice a certain generality by not going directly to the Lebesgue integral which is tailor-made for probability theory. This is done, however, to moderate the mathematical difficulty of the book.

The mean can be interpreted as *the centre of gravity* of the unit mass as distributed by the density function. If we denote the mass located at a distance  $x_i$ ,  $i = 1, 2, \dots$  from the origin by  $m(x_i)$  then the centre of gravity is located at

$$\frac{\sum_i x_i m(x_i)}{\sum_i m(x_i)}. \quad (4.48)$$

If we identify  $m(x_i)$  with  $p(x_i)$  then  $E(X) = \sum_i x_i p(x_i)$ , given that  $\sum_i p(x_i) = 1$ . In this sense the mean of the r.v.  $X$  provides a measure of location (or central tendency) for the density function of  $X$ .

### Examples

- (i) If  $X$  is a Bernoulli distributed r.v. ( $X \sim b(1, p)$ ) then

$$\begin{array}{ccc} X & 0 & 1 \\ f(x) & (1-p) & p \end{array}$$

$$\Rightarrow E(X) = 0 \cdot (1-p) + 1 \cdot p = p.$$

- (ii) If  $X \sim U(a, b)$ , i.e.  $X$  is a uniformly distributed r.v., then

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_a^b x \left( \frac{1}{b-a} \right) dx = \frac{1}{2} \left( \frac{1}{b-a} \right) x^2 \Big|_a^b = \frac{a+b}{2}.$$

- (iii) If  $X \sim N(\mu, \sigma^2)$ , i.e.  $X$  is a normally distributed r.v., then

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x \left( \frac{1}{\sigma \sqrt{(2\pi)}} \exp \left\{ -\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2 \right\} \right) dx, \quad \text{for } z = \frac{X-\mu}{\sigma} \\ &= \int_{-\infty}^{\infty} \frac{(\sigma z + \mu)}{\sqrt{(2\pi)}} e^{-\frac{1}{2}z^2} dz = \frac{\sigma}{\sqrt{(2\pi)}} \int_{-\infty}^{\infty} z e^{-\frac{1}{2}z^2} dz + \\ &\quad \mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{(2\pi)}} e^{-\frac{1}{2}z^2} dz \\ &= 0 + \mu \cdot 1 = \mu, \end{aligned}$$

since the first term is an odd function, i.e.  $h(-x) = -h(x)$ . Thus, the parameter  $\mu$  for  $X \sim N(\mu, \sigma^2)$  represents its mean.

In the above examples the mean of the r.v.  $X$  existed. The condition which guarantees the existence of  $E(X)$  is that

$$\int_{-\infty}^{\infty} |x| f(x) dx < \infty \quad \text{or} \quad \sum_i |x_i| f(x_i) < \infty. \quad (4.49)$$

One example where the mean does not exist is the case of a Cauchy distributed r.v. with a pdf given by

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

In order to show this let us consider the above condition:

$$\begin{aligned} \int_{-\infty}^{\infty} |x| f(x) dx &= \frac{1}{\pi} \int_{-\infty}^{\infty} |x| \frac{1}{1+x^2} dx = \frac{1}{\pi} 2 \int_0^{\infty} \left( \frac{x}{1+x^2} \right) dx \\ &\quad \text{by symmetry} \\ &= \frac{1}{\pi} \lim_{a \rightarrow \infty} 2 \int_0^a \left( \frac{x}{1+x^2} \right) dx = \frac{1}{\pi} \lim_{a \rightarrow \infty} \log_e(1+a^2) \\ &= \infty. \end{aligned}$$

That is,  $E(X)$  does not exist for the Cauchy distribution.

#### *Some properties of the expectation*

$$(E1) \quad E(c) = c, \text{ if } c \text{ is a constant.}$$

$$(E2) \quad E(aX_1 + bX_2) = aE(X_1) + bE(X_2) \text{ for any two r.v.'s } X_1 \text{ and } X_2.$$

$X_2$  whose means exist and  $a, b$  are real constants.

For example, if  $X_i \sim b(1, p)$ ,  $i = 1, 2, \dots, n$ , i.e.  $X_i$  represents the Bernoulli r.v. of the  $i$ th trial, then for  $Y_n = (\sum_{i=1}^n X_i) \sim B(n, p) \Rightarrow E(Y_n) = \sum_{i=1}^n E(X_i) = \sum_{i=1}^n p = np$ . That is, the mean of a binomially distributed r.v. equals the number of trials multiplied by the probability of ‘success’.

Properties E1 and E2 define  $E(\cdot)$  as a **linear transformation**.

(E3)  $\Pr(X \geq \lambda E(X)) \leq 1/\lambda$  for a positive r.v.  $X$  and  $\lambda > 0$ ; this is the so-called **Markov inequality**.

Although the mean of a r.v.  $X$ ,  $E(X)$  is by far the most widely used measure of location two other measures are sometimes useful. The first is the *mode* defined to be the value of  $X$  for which the density function achieves its maximum. The second is the *median*,  $x_m$  of  $X$  defined to be the value of  $X$  such that

$$F(x_m) = \int_{-\infty}^{x_m} f(x) dx = \frac{1}{2} = \int_{x_m}^{\infty} f(x) dx. \quad (4.50)$$

It is obvious that if the density function of  $X$  is symmetric then

$$E(X) = x_m. \quad (4.51)$$

If it is both *symmetric* and *unimodal* (i.e. it has only one mode) then

$$\text{mean} = \text{median} = \text{mode},$$

assuming that the mean exists. On the other hand, if the pdf is not unimodal this result is not necessarily valid, as Fig. 4.13 exemplifies. In contrast to the mean the median always exists, in particular for the Cauchy distribution  $x_m = 0$ .

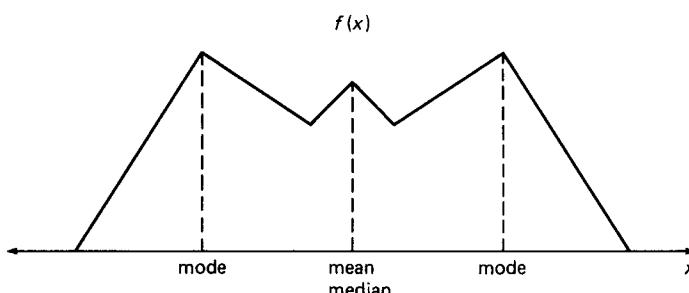


Fig. 4.13. A symmetric density whose mean and median differ from the mode.

## (2) The variance

When a measure of location for a r.v.  $X$  is available, it is often required to get an idea as to how widely the values of  $X$  are spread around the location measure, that is, a measure of dispersion (or spread). Related to the mean as a measure of location is the dispersion measure called the *variance* and defined by

$$\begin{aligned}\text{Var}(X) &= E[X - E(X)]^2 \\ &= \int_{-\infty}^{\infty} (x - E(X))^2 f(x) dx \quad \text{continuous}\end{aligned}\tag{4.52}$$

$$= \sum_i (x_i - E(X))^2 f(x_i) \quad \text{discrete.}\tag{4.53}$$

The variance can be interpreted as the *moment of inertia* of the mass distribution with respect to the perpendicular axis through the mean.

*Note:* the square root of the variance is referred to as *standard deviation*.

*Examples*

- (i) Let  $X \sim b(1, p)$ ; it was shown above that  $E(X) = p$ , thus

$$\text{Var}(X) = (0 - p^2)(1 - p) + (1 - p)^2 p = p(1 - p).$$

- (ii) Let  $X \sim U(a, b)$ ; it was shown above that  $E(X) = (a + b)/2$ , thus

$$\text{Var}(X) = \int_{-\infty}^{\infty} \left[ X - \left( \frac{a+b}{2} \right) \right]^2 \left( \frac{1}{b-a} \right) dx = \frac{(b-a)^2}{12} \quad (\text{verify}).$$

An equality which turns out to be convenient for deriving  $\text{Var}(X)$  is given by

$$\text{Var}(X) = E(X^2) - [E(X)]^2,$$

where

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx.$$

- (iii) Let  $X \sim N(\mu, \sigma^2)$ ; then

$$\begin{aligned}E(X^2) &= \int_{-\infty}^{\infty} x^2 \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} dx \\ &= \int_{-\infty}^{\infty} (\mu^2 + 2\mu\sigma z + \sigma^2 z^2) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \quad \text{for } z = \frac{x-\mu}{\sigma} \\ &= \mu^2 + \sigma^2 \int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = \mu^2 + \sigma^2.\end{aligned}$$

Thus  $\text{Var}(X) = \mu^2 + \sigma^2 - \mu^2 = \sigma^2$ ; see Binmore (1983).

*Some properties of the variance*

- (V1)  $\text{Var}(c) = 0$  for any constant  $c$ .
- (V2)  $\text{Var}(aX) = a^2 \text{Var}(X)$ , for a constant.
- (V3)  $\Pr(|X - E(X)| \geq k) \leq [\text{Var}(X)]/k^2$  – **Chebyshev's inequality** for  $k > 0$ . This inequality gives a relation between the variance and the probability of dispersion as defined by  $|X - E(X)| \geq k$ , providing in effect an upper bound for such probabilities.

### (3) Higher moments

Continuing the analogy with the various concepts from mechanics we define the moments of inertia from  $x=0$  to be the so-called *raw moments*:

$$\mu'_r \equiv E(X^r) = \int_{-\infty}^{\infty} x^r f(x) dx, \quad r = 0, 1, 2, \dots, \quad (4.54)$$

the  $r$ th raw moment, if it exists, with  $\mu'_0 = 1$  and  $\mu'_1 \equiv E(X)$ ; the mean is usually denoted by  $\mu$ . Similarly, the  $r$ th moment around  $x=\mu$ , called the  $r$ th *central moment*, is defined (if it exists) by

$$\mu_r \equiv E(X - \mu)^r = \int_{-\infty}^{\infty} (x - \mu)^r f(x) dx, \quad r = 2, 3, \dots, \quad (4.55)$$

$\mu_2 \equiv E(X - \mu)^2$  is the variance, usually denoted by  $\sigma^2$ . These higher moments are sometimes useful in providing us with further information relating to the distribution and density functions of r.v.'s. In particular, the 3rd and 4th central moments, when standardised in the form:

$$\chi_3 = \frac{\mu_3}{\sigma^3} \quad (4.56)$$

and

$$\chi_4 = \frac{\mu_4}{\sigma^4} \quad (4.57)$$

are referred to as measures of *skewness* and *kurtosis* and provide us with measures of asymmetry and flatness of peak, respectively. Deriving the raw moments first is usually easier and then the central moments can be derived via (see Kendall and Stuart (1969)):

$$\mu_r = \sum_{j=1}^r (-1)^j \binom{r}{j} \mu^j \mu'_{r-j}. \quad (4.58)$$

An important tool in the derivation of the raw moments is the *characteristic*

function defined by

$$\psi_x(t) = E(e^{itx}) = \int_{-\infty}^{\infty} e^{itx} dF(x), \quad i = \sqrt{-1}. \quad (4.59)$$

Using the power series form of  $e^x$  we can express it in the form

$$\psi_x(t) = 1 + \sum_{r=1}^{\infty} \frac{(it)^r}{r!} \mu'_r. \quad (4.60)$$

This implies that we can derive  $\mu'_r$  via

$$\mu'_r = \left. \frac{d^r \psi_x(t)}{dt^r} \right|_{t=0}, \quad r = 1, 2, \dots \quad (4.61)$$

A function related to  $\psi_x(t)$  of particular interest in asymptotic theory (see Chapter 10) is

$$\log_e \psi_x(t) = 1 + \sum_{r=1}^{\infty} \frac{(it)^r}{r!} \kappa_r \quad (4.62)$$

where  $\kappa_r$ ,  $r = 1, 2, \dots$  are called the *cumulants*.

### Example

Let  $X \sim N(\mu, \sigma^2)$ , the characteristic function takes the form

$$\begin{aligned} \psi_x(t) &= \exp\{it\mu - \frac{1}{2}t^2\sigma^2\}, \\ \left. \frac{1}{i} \frac{d\psi_x(t)}{dt} \right|_{t=0} &= \left. \frac{1}{i} \exp\{it\mu - \frac{1}{2}t^2\sigma^2\}(i\mu - t\sigma^2) \right|_{t=0} = \mu, \\ \left. \frac{1}{i^2} \frac{d^2\psi_x(t)}{dt^2} \right|_{t=0} &= \mu^2 + \sigma^2 \Rightarrow \mu_2 = \sigma^2. \end{aligned}$$

Similarly we can show that  $\mu_3 = 0$ ,  $\mu_4 = 3\sigma^4$ ,  $\mu_5 = 0$ ,  $\mu_6 = 15\sigma^6$ , etc.  $\kappa_1 = \mu$ ,  $\kappa_2 = \sigma^2$ ,  $\kappa_r = 0$ ,  $r \geq 3$ ,  $\alpha_3 = 0$ ,  $\alpha_4 = 3$ ; see Kendall and Stuart (1969).

Having considered the various numerical characteristics of r.v.'s as related to their distribution, it is natural to ask whether under certain circumstances knowing the moments  $\mu'_r$ ,  $r = 1, 2, \dots$  we can determine the DF  $F(x)$ . The answer is that  $F(x)$  is uniquely determined by its moments  $\mu'_r$ ,  $r = 1, 2, \dots$  if and only if

$$\sum_{r=1}^{\infty} (\mu'_{2r})^{-\frac{1}{2r}} = \infty. \quad (4.63)$$

This is known as *Carleman's condition*.

***Important concepts***

Random variable, the probability space induced by a r.v., a  $\sigma$ -field generated by a r.v., an increasing sequence of  $\sigma$ -fields, the minimal Borel field generated by half-closed intervals  $(-\infty, x]$   $x \in \mathbb{R}$ , distribution function, density function, discrete and continuous r.v.'s, probability model, parametric family of densities, unknown parameters, normal distribution, expectation and variance of a r.v., skewness and kurtosis, higher raw and central moments, characteristic function, cumulants.

***Questions***

1. Since we can build the whole of probability theory on  $(S, \mathcal{F}, P(\cdot))$  why do we need to introduce the concept of a random variable?
2. Define the concept of a r.v. and explain the role of the Borel field generated by the half-closed intervals  $(-\infty, x]$  in deciding whether a function  $X(\cdot): S \rightarrow \mathbb{R}$  is a r.v.
3. 'Although any function  $X(\cdot): S \rightarrow \mathbb{R}$  can be defined to be a r.v. relative to some  $\sigma$ -field  $\mathcal{F}_x$  we stand to lose valuable information if we do not define the r.v. with care'. Discuss.
4. Explain the relationship between  $P(\cdot)$ ,  $P_x(\cdot)$  and  $F_x(\cdot)$ .
5. Discuss the relationship between  $F_x(\cdot)$  and  $f(\cdot)$  for both discrete and continuous r.v.'s. What properties do density functions satisfy?
6. Explain the idea of a probability model  $\Phi = \{f(x; \theta), \theta \in \Theta\}$  and discuss its relationship with the idea of a random experiment  $\mathcal{E}$ , as well as the real phenomenon to be modelled.
7. Give an example of a real phenomenon for which each of the following distributions might be appropriate:
  - (i) Bernoulli;
  - (ii) binomial;
  - (iii) normal.
 Explain your choice.
8. Explain why we need the concepts of mean, variance and higher moments in the context of modelling real phenomena using the probability model  $\Phi = \{f(x; \theta), \theta \in \Theta\}$ . What features of the density function do the following numerical characteristics purport to measure?
 

mean, median, mode, variance, skewness, kurtosis.

Explain the difference between these and the concepts with the same names in the context of the descriptive study of data.

Explain Markov's and Chebyshev's inequalities.

12. Compare the properties of the mean with those of the variance.  
 13. Under what circumstances do the moments characterise the distribution function of a r.v.?

### *Exercises*

1. Consider a random experiment with  $S = \{a, b, c, d\}$  and  $P(a) = P(b) = \frac{1}{4}$ ,  $P(c) = \frac{1}{3}$ ,  $P(d) = \frac{1}{6}$ .
  - (i) Derive the  $\sigma$ -field of the power set  $\mathcal{F}$ .
  - (ii) Derive the minimal  $\sigma$ -field generated by  $\{a, b\}$ , say  $\mathcal{F}_1$ . Consider the following function defined as  $S$ :
$$X(a) = X(b) = 0, \quad X(c) = X(d) = 7\,405\,926, \dagger$$

$$Y(a) = 0, \quad Y(b) = Y(c) = 1, \quad Y(d) = 2.$$
  - (iii) Show that  $X$  and  $Y$  are both r.v.'s relative to  $\mathcal{F}$ .
  - (iv) Show that  $X$  is a r.v. relative to  $\mathcal{F}_1$  but  $Y$  is not.
  - (v) Find the minimal  $\sigma$ -field generated by  $Y$ ,  $\mathcal{F}_y$ .
  - (vi) Find  $P_y(\{0\})$ ,  $P_y((0, 1])$ ,  $P_y((-\infty, 1])$ ,  $P_y([0, 1] \cup [1, 2))$ .
  - (vii) Derive the distribution and density functions  $F(y)$ ,  $f(y)$  and plot them.
  - (viii) Calculate  $E(Y)$ ,  $\text{Var}(Y)$ ,  $\alpha_3(Y)$  and  $\alpha_4(Y)$ .
2. The distribution function of the *exponential distribution* is

$$F(x) = 1 - \exp\left\{-\frac{x}{\theta}\right\}, \quad x \geq 0.$$

- (i) Derive the density function  $f(x)$  and plot its graph.
- (ii) Derive the characteristic function  $\psi(t) = E(e^{itX})$ .
- (iii) Derive  $E(X)$ ,  $\text{Var}(X)$ ,  $\alpha_3(X)$  and  $\alpha_4(X)$ .

*Note:*

$$\int e^{ax} dx = \frac{1}{a} e^{ax}, \quad \int x e^{ax} dx = \frac{(ax - 1)}{a^2} e^{ax},$$

$$\int x^2 e^{ax} dx = \frac{(a^2 x^2 - 2ax + s)}{a^3} e^{ax},$$

$$\int x^3 e^{ax} dx = \frac{e^{ax}}{a} \left( x^3 - \frac{3}{a} x^2 + \frac{6}{a^2} x - \frac{6}{a^3} \right).$$

† If the reader is wondering about the significance of this number it is the number of demons inhabiting the earth as calculated by German physician Weirus in the sixteenth century (see Jastrow (1962)).

3. Indicate which of the following functions represent proper density functions and explain your answer:
  - (i)  $f(x) = kx^2, \quad 0 < x < 2;$
  - (ii)  $f(x) = 2(1-x)^2, \quad x > 1;$
  - (iii)  $f(x) = 3(1-e^{-x}), \quad x \geq 1;$
  - (iv)  $f(x) = \frac{1}{6}(x^3 + 1), \quad 0 < x < 2;$
  - (v)  $f(x) = \frac{2}{3}x^3, \quad x \in \mathbb{R}.$
4. Prove that  $\text{Var}(X) = E(X^2) - [E(X)]^2$ .
5. If for the r.v.  $X$ ,  $E(X) = 2$ ,  $E(X^2) = 4$ , find  $E(3X + 4)$ ,  $\text{Var}(4X)$ .
6. Let  $X \sim N(0, 1)$ . Use the tables to calculate the probabilities
  - (i)  $F(2.5);$
  - (ii)  $F(0.15);$
  - (iii)  $1 - F(2.0).$

and compare them with the bounds from Chebyshev's inequality. What is the percentage of error in the three cases?

#### Additional references

Bickel and Doksum (1977); Chung (1974); Cramer (1946); Dudewicz (1976); Giri (1974); Mood, Graybill and Boes (1974); Pfeiffer (1978); Rohatgi (1976).

## CHAPTER 5

---

### Random vectors and their distributions

---

The probability model formulated in the previous chapter was in the form of a parametric family of densities associated with a random variable (r.v.)  $X: \Phi = \{f(x; \theta), \theta \in \Theta\}$ . In practice, however, there are many observable phenomena where the outcome comes in the form of several quantitative attributes. For example, data on personal income might be related to number of children, social class, type of occupation, age class, etc. In order to be able to model such real phenomena we need to extend the above framework for a single r.v. to one for multidimensional r.v.'s or *random vectors*, that is,

$$\mathbf{X} = (X_1, X_2, \dots, X_n)',$$

where each  $X_i, i = 1, 2, \dots, n$  measures a particular quantifiable attribute of the random experiment's ( $\mathcal{E}$ ) outcomes.

For expositional purposes we shall restrict attention to the two-dimensional (bivariate) case, which is quite adequate for a proper understanding of the concepts involved, giving only scanty references to the  $n$ -dimensional random vector case (just for notational purposes). In the next section we consider the concept of a random vector and its joint distribution and density functions in direct analogy to the random variable case. In Sections 5.2 and 5.3 we consider two very important forms of the joint density function, the marginal and conditional densities respectively. These forms of the joint density function will play a very important role in Part IV.

#### 5.1 Joint distribution and density functions

Consider the random experiment  $\mathcal{E}$  of tossing a fair coin twice. The sample

space takes the form  $S = \{(HT), (TH), (HH), (TT)\}$ . Define the function  $X_1(\cdot)$  to be the number of ‘heads’ and  $X_2(\cdot)$  to be the number of ‘tails’. Both of these functions map  $S$  into the real line  $\mathbb{R}$  in the form

$$(X_1(\cdot), X_2(\cdot))\{(TH)\} = (1, 1), \quad \text{i.e. } (X_1(TH), X_2(TH)) = (1, 1),$$

$$(X_1(\cdot), X_2(\cdot))\{(HT)\} = (1, 1),$$

$$(X_1(\cdot), X_2(\cdot))\{(HH)\} = (2, 0),$$

$$(X_1(\cdot), X_2(\cdot))\{(TT)\} = (0, 2).$$

This is shown in Fig. 5.1. The function  $(X_1(\cdot), X_2(\cdot)): S \rightarrow \mathbb{R}^2$  is a two-dimensional vector function which assigns to each element  $s$  of  $S$ , the pair of *ordered* numbers  $(x_1, x_2)$  where  $x_1 = X_1(s)$ ,  $x_2 = X_2(s)$ . As in the one-dimensional case, for the vector function to define a random vector it has to satisfy certain conditions which ensure that the probabilistic and event structure of  $(S, \mathcal{F}, P(\cdot))$  is preserved. In direct analogy with the single variable case we say that the mapping

$$\mathbf{X}(\cdot) \equiv (X_1(\cdot), X_2(\cdot)): S \rightarrow \mathbb{R}^2 \quad (5.1)$$

defines a random vector if for each event in the Borel field product  $\mathcal{B} \times \mathcal{B} \equiv \mathcal{B}^2$ , say  $\mathbf{B} \equiv (B_1, B_2)$ , the event defined by

$$\mathbf{X}^{-1}(\mathbf{B}) = \{s: X_1(s) \in B_1, X_2(s) \in B_2, s \in S\} \quad (5.2)$$

belongs to  $\mathcal{F}$ .

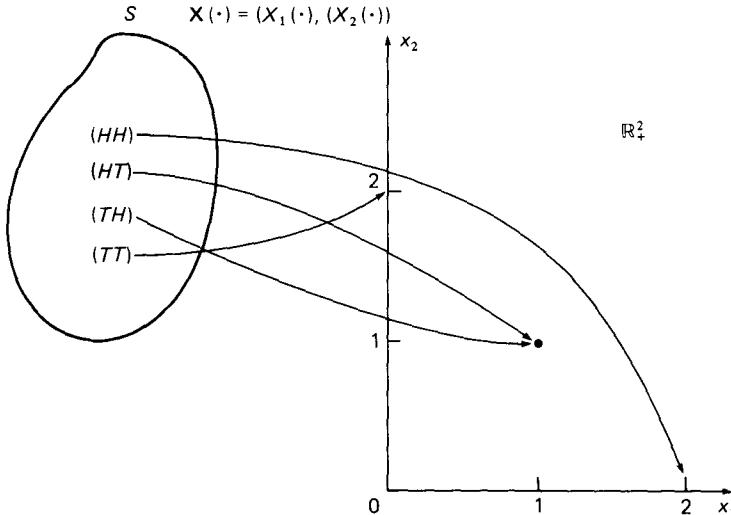


Fig. 5.1. A bivariate random vector.

Extending the result that  $\mathcal{B}$  can be profitably seen as being the  $\sigma$ -field generated by half-closed intervals of the form  $(-\infty, x]$  to the case of the direct product  $\mathcal{B} \times \mathcal{B}$  we can show that the random vector  $\mathbf{X}(\cdot)$  satisfying

$$\mathbf{X}^{-1}((-\infty, \mathbf{x}]) \in \mathcal{F} \text{ for all } \mathbf{x} \in \mathbb{R}^2$$

$$\text{implies } \mathbf{X}^{-1}(\mathbf{B}) \in \mathcal{F} \text{ for all } \mathbf{B} \in \mathcal{B}^2. \quad (5.3)$$

This allows us to define a random vector as follows:

*Definition 1*

A random vector  $X(\cdot)$  is a vector function

$$\mathbf{X}(\cdot): S \rightarrow \mathbb{R}^2, \quad (5.4)$$

such that for any two real numbers  $(x_1, x_2) \equiv \mathbf{x}$ , the event

$$\mathbf{X}^{-1}((-\infty, \mathbf{x}]) = \{s: -\infty < X_1(s) \leq x_1,$$

$$-\infty < X_2(s) \leq x_2, s \in S\} \in \mathcal{F}.$$

Note.  $(-\infty, \mathbf{x}] = ((-\infty, x_1], (-\infty, x_2])$  represents an infinite rectangle (see Fig. 5.2). The random vector (as in the case of a single random variable) induces a probability space  $(\mathbb{R}^2, \mathcal{B}^2, P_x(\cdot))$ , where  $\mathcal{B}^2$  are Borel subsets on the plane and  $P_x(\cdot)$  a probability set function defined over events in  $\mathcal{B}^2$ , in a way which preserves the probability structure of the original probability

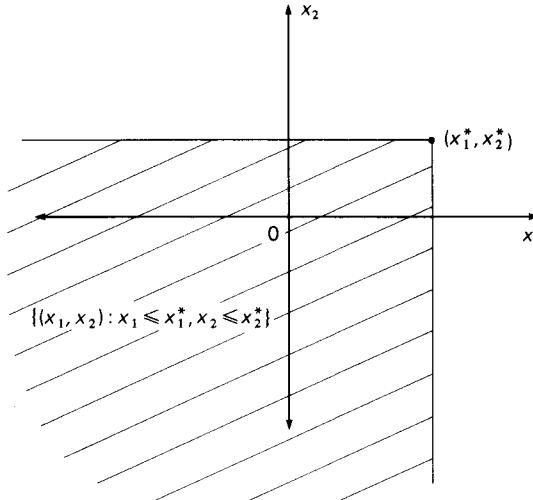


Fig. 5.2. The infinite rectangle  $(-\infty, \mathbf{x}^*]$ ,  $\mathbf{x}^* = (x_1^*, x_2^*)$ .

space  $(S, \mathcal{F}, P(\cdot))$ . This is achieved by attributing to each  $\mathbf{B} \in \mathcal{B}^2$  the probability

$$P_x(\mathbf{B}) = P(\{s: (X_1(s), X_2(s)) \in \mathbf{B}\}) \quad (5.5)$$

or

$$P_x((-\infty, \mathbf{x}]) = Pr(X_1 \leq x_1, X_2 \leq x_2). \quad (5.6)$$

This enables us to reduce  $P_x(\cdot)$  to a point function  $F(x_1, x_2)$ , we call the *joint (cumulative) distribution function*.

*Definition 2*

Let  $\mathbf{X} \equiv (X_1, X_2)$  be a random vector defined on  $(S, \mathcal{F}, P(\cdot))$ . The function defined by

$$F(\cdot, \cdot): \mathbb{R}^2 \rightarrow [0, 1], \quad (5.7)$$

such that

$$\begin{aligned} F(\mathbf{x}) \equiv F(x_1, x_2) &= P_x((-\infty, \mathbf{x}]) = Pr(X_1 \leq x_1, X_2 \leq x_2) \\ &\equiv Pr(\mathbf{X} \leq \mathbf{x}) \end{aligned}$$

is said to be the **joint distribution function** of  $\mathbf{X}$ .

In the coin-tossing example above, the random vector  $\mathbf{X}(\cdot)$  takes the value  $(1, 1), (2, 0), (0, 2)$  with probabilities  $\frac{1}{2}, \frac{1}{4}$  and  $\frac{1}{4}$  respectively. In order to derive the joint distribution function (DF) we have to define all the events of the form  $\{s: X_1(s) \leq x_1, X_2(s) \leq x_2, s \in S\}$  for all  $(x_1, x_2) \in \mathbb{R}^2$

$$\{s: X_1(s) \leq x_1, X_2(s) \leq x_2, s \in S\} = \begin{cases} \emptyset, & x_1 < 0, x_2 < 0 \\ \{(HT), (TH)\}, & 0 \leq x_1 < 2, 0 \leq x_2 < 2 \\ \{HH\}, & x_1 \geq 2, 0 \leq x_2 < 2 \\ \{TT\}, & 0 \leq x_1 < 2, x_2 \geq 2 \\ S, & x_1 \geq 2, x_2 \geq 2. \end{cases}$$

Note a degree of arbitrariness in choosing the infinite rectangles  $(-\infty, \mathbf{x}]$ .

The joint DF of  $X_1$  and  $X_2$  is given by

$$F(x_1, x_2) = \begin{cases} 0, & x_1 < 0, x_2 < 0 \\ \frac{1}{2}, & 0 \leq x_1 < 2, 0 \leq x_2 < 2 \\ \frac{3}{4}, & x_1 \geq 2, 0 \leq x_2 < 2 \\ 1, & x_1 \geq 2, x_2 \geq 2. \end{cases}$$

Table 5.1. Joint density function of  $(X_1, X_2)$ 

$X_2 \backslash X_1$	0	1	2
0	0	0	$\frac{1}{4} \leftarrow Pr(X_1=0, X_2=2)$
1	0	$\frac{1}{2}$	0
2	$\frac{1}{4}$	0	0

From the definition of the joint DF we can deduce that  $F(x_1, x_2)$  is a monotone non-decreasing function in each variable separately, and

$$(i) \quad \lim_{x_1 \rightarrow -\infty} F(x_1, x_2) = \lim_{x_2 \rightarrow -\infty} F(x_1, x_2) = 0; \quad (5.8)$$

$$(ii) \quad \lim_{\substack{x_1 \rightarrow \infty \\ x_2 \rightarrow \infty}} F(x_1, x_2) = 1. \quad (5.9)$$

As in the case of one r.v., we concentrate exclusively on discrete and continuous joint DF only; singular distributions are not considered.

### Definition 3

The joint DF of  $X_1$  and  $X_2$  is called a **discrete distribution** if there exists a density function  $f()$  such that

$$f(x_1, x_2) \geq 0, \quad (x_1, x_2) \in \mathbb{R}^2 \quad (5.10)$$

and it takes the value zero everywhere except at a finite or countably infinite point in the plane with

$$f(x_1, x_2) = Pr(X_1 = x_1, X_2 = x_2). \quad (5.11)$$

In the coin-tossing example the density function in a rectangular array form is represented in Table 5.1. Fig. 5.3 represents the graph of the joint density function of  $\mathbf{X} \equiv (X_1, X_2)$ . The joint DF is obtained from  $f(x_1, x_2)$  via the relation

$$F(x_1, x_2) = \sum_{x_{1i} < x_1} \sum_{x_{2i} < x_2} f(x_{1i}, x_{2i}). \quad (5.12)$$

### Definition 4

The joint DF of  $X_1$  and  $X_2$  is called (absolutely) **continuous** if there exists a **non-negative function**  $f(x_1, x_2)$  such that

$$F(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f(u, v) du dv. \quad (5.13)$$

## 5.2 Some bivariate distributions

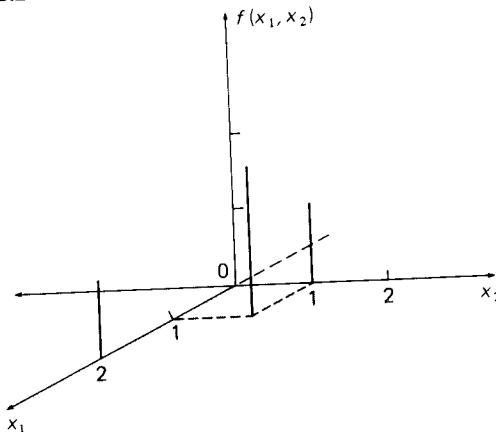


Fig. 5.3. The bivariate density function of Table 5.1.

$f(x_1, x_2)$  is called the joint (probability) density function of  $X_1, X_2$ . This definition implies the following properties for  $f(x_1, x_2)$ :

$$(F1) \quad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 dx_2 = 1. \quad (5.14)$$

$$(F2) \quad Pr(a < X_1 \leq b, c < X_2 \leq d) = \int_a^b \int_c^d f(x_1, x_2) dx_1 dx_2. \quad (5.15)$$

$$(F3) \quad f(x_1, x_2) = \frac{\partial^2}{\partial x_1 \partial x_2} F(x_1, x_2) \quad (5.16)$$

if  $f(\cdot)$  is continuous at  $(x_1, x_2)$ .

## 5.2 Some bivariate distributions

### (1) Bivariate normal distribution

$$\begin{aligned} f(x_1, x_2; \theta) &= \frac{(1 - \rho^2)^{-\frac{1}{2}}}{2\pi\sigma_1\sigma_2} \\ &\exp\left\{-\frac{1}{2(1 - \rho^2)} \left[ \left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 \right.\right. \\ &\quad \left.\left.- 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1}\right) \left(\frac{x_2 - \mu_2}{\sigma_2}\right) + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2 \right]\right\}, \\ &x_1, x_2 \in \mathbb{R}, \quad (5.17) \end{aligned}$$

$$\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho) \in \mathbb{R}^2 \times \mathbb{R}_+^2 \times [0, 1].$$

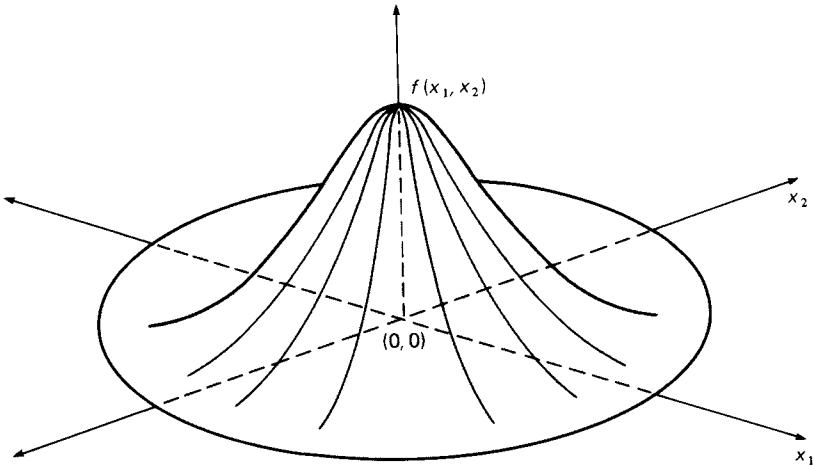


Fig. 5.4. The density function of a standard normal density.

It is interesting to note that the expression inside the square brackets expressed in the form of

$$\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1}\right) \left(\frac{x_2 - \mu_2}{\sigma_2}\right) + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2 = c^2 \quad (5.18)$$

defines a sequence of ellipses of points with equal probability which can be viewed as map-like contours of the graph of  $f(x_1, x_2)$  represented in Fig. 5.4.

### (2) **Bivariate Pareto distribution**

$$f(x_1, x_2) = \lambda(\lambda+1)(a_1 a_2)^{\lambda+1} (a_2 x_1 + a_1 x_2 - a_1 a_2)^{-(\lambda+2)}, \\ (\lambda > 0, x_1 > a_1 > 0, x_2 > a_2 > 0), \quad (5.19)$$

$$\theta = (\lambda, a_1, a_2).$$

### (3) **Bivariate binomial distribution**

$$f(x_1, x_2) = \frac{n!}{x_1! x_2!} p_1^{x_1} p_2^{x_2}, \quad x_1 + x_2 = n, \quad p_1 + p_2 = 1. \quad (5.20)$$

The extension of the concept of a random variable  $X$  to that of a random vector  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  enables us to generalise the probability model

$$\Phi = \{f(x; \theta), \theta \in \Theta\} \quad (5.21)$$

to that of a parametric family of joint density functions

$$\Phi = \{f(x_1, x_2, \dots, x_n; \theta), \theta \in \Theta\}. \quad (5.22)$$

This is a very important generalisation since in most applied disciplines, including econometrics, the real phenomena to be modelled are usually multidimensional in the sense that there is more than one quantifiable feature to be considered.

### 5.3 Marginal distributions

Let  $\mathbf{X} \equiv (X_1, X_2)$  be a bivariate random vector defined on  $(S, \mathcal{F}, P(\cdot))$  with a joint distribution function  $F(x_1, x_2)$ . The question which naturally arises is whether we could separate  $X_1$  and  $X_2$  and consider them as individual random variables. The answer to this question leads us to the concept of a *marginal distribution*. The marginal distribution functions of  $X_1$  and  $X_2$  are defined by

$$F_1(x_1) = \lim_{x_2 \rightarrow \infty} F(x_1, x_2) \quad (5.23)$$

and

$$F_2(x_2) = \lim_{x_1 \rightarrow \infty} F(x_1, x_2). \quad (5.24)$$

Having separated  $X_1$  and  $X_2$  we need to see whether they can be considered as single r.v.'s defined on the same probability space. In defining a random vector we imposed the condition that

$$\{s: X_1(s) \leq x_1, X_2(s) \leq x_2\} \in \mathcal{F}. \quad (5.25)$$

The definition of the marginal distribution function we used the event

$$\{s: X_1(s) \leq x_1, X_2(s) < \infty\}, \quad (5.26)$$

which we know belongs to  $\mathcal{F}$ . This event, however, can be written as the intersection of two sets of the form

$$\{s: X_1(s) \leq x_1\} \cap \{s: X_2(s) < \infty\} \quad (5.27)$$

but the second set is  $S$  i.e.  $\{s: X_2(s) < \infty\} = S$ ,

which implies that

$$\{s: X_1(s) \leq x_1, X_2(s) < \infty\} = \{s: X_1(s) \leq x_1\}, \quad (5.28)$$

which indeed belongs to  $\mathcal{F}$  and it is the condition needed for  $X_1$  to be a r.v. with a probability function  $F_1(x_1)$ ; the same is true for  $X_2$ . In order to see

this, consider the joint distribution function

$$F(x_1, x_2) = 1 - e^{-\theta x_1} - e^{-\theta x_2} + \exp\{-\theta(x_1 + x_2)\},$$

$$(x_1, x_2) \in \mathbb{R}_+^2, \quad (5.29)$$

$$F_1(x_1) = \lim_{x_2 \rightarrow \infty} F(x_1, x_2) = 1 - e^{-\theta x_1}, \quad x_1 \in \mathbb{R}_+,$$

since  $\lim_{n \rightarrow \infty} (e^{-n}) = 0$ . Similarly,

$$F_2(x_2) = 1 - e^{-\theta x_2}, \quad x_2 \in \mathbb{R}_+.$$

Note that  $F_1(x_1)$  and  $F_2(x_2)$  are proper distribution functions.

Given that the probability model has been defined in terms of the joint density functions, it is important to consider the above operation of marginalisation in terms of these density functions. The *marginal density functions* of  $X_1$  and  $X_2$  are defined by

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2$$

and

$$f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1,$$

that is, the marginal density of  $X_i$  ( $i=1, 2$ ) is derived by *integrating out*  $X_j$  ( $i \neq j$ ) from the joint density. In the discrete case this amounts to *summing out* with respect to the other variable:

$$f_1(x_1) = \sum_{i=1}^{\infty} f(x_1, x_{2i}). \quad (5.31)$$

### Example

Consider the working population of the UK classified by income and age as follows:

*Income:* £2000–4000, £4000–8000, £8000–12 000, £12 000–20 000, £20 000–50 000, over £50 000.

*Age:* young, middle-aged, senior.

Define the random variables  $X_1$ —income class, taking values 1–6, and  $X_2$ —age class, taking values 1–3. Let the joint density be (Table 5.2):

Table 5.2. Joint density of  $(X_1, X_2)$ 

		$X_2 = x_{2i}$			$f_1(x_1)$
		1	2	3	
$X_1 = x_{1i}$	1	0.250	0.020	0.005	0.275
	2	0.075	0.250	0.020	0.345
	3	0.040	0.075	0.100	0.215
	4	0.020	0.030	0.035	0.085
	5	0.010	0.015	0.020	0.045
	6	0.005	0.010	0.020	0.035
	$f_2(x_2)$	0.400	0.400	0.200	1.000

The marginal density function of  $X_1$  is shown in the column representing *row totals* and it refers to the probabilities that a randomly selected person will belong to the various income classes. The marginal density of  $X_2$  is the row representing *column totals* and it refers to the probabilities that a randomly selected person will belong to the various age classes. That is, the marginal distribution of  $X_1(X_2)$  incorporates no information relating to  $X_2(X_1)$ . Moreover, it is quite obvious that knowing the joint density function of  $X_1$  and  $X_2$  we can derive their marginal density functions; the reverse, however, is not true in general. Knowledge of  $f_1(x_1)$  and  $f_2(x_2)$  is enough to derive  $f(x_1, x_2)$  only when

$$f(x_1, x_2) = f_1(x_1) \cdot f_2(x_2), \quad (5.32)$$

in which case we say that  $X_1$  and  $X_2$  are *independent r.v.'s*. *Independence* in terms of the distribution functions takes the same form

$$F(x_1, x_2) = F_1(x_1) \cdot F_2(x_2). \quad (5.33)$$

In the case of the income-age example it is clear that

$$f(x_1, x_2) \neq f_1(x_1) \cdot f_2(x_2),$$

e.g.

$$0.250 \neq (0.275)(0.4),$$

and hence,  $X_1$  and  $X_2$  are not independent r.v.'s, i.e. income and age are related in some probabilistic sense; it is more probable to be middle-aged and rich than young and rich!

In the continuous r.v.'s example we can easily verify that

$$F_1(x_1) \cdot F_2(x_2) = (1 - e^{-\theta x_1})(1 - e^{-\theta x_2}) = F(x_1, x_2), \quad (5.34)$$

and thus  $X_1$  and  $X_2$  are indeed independent.

Note that two events,  $A_1$  and  $A_2$ , in the context of the probability space  $(S, \mathcal{F}, P(\cdot))$  are said to be independent (see Section 3.3) if

$$P(A_1 \cap A_2) = P(A_1) \cdot P(A_2). \quad (5.35)$$

It must be stressed that marginal density functions are proper density functions satisfying all the properties of such functions. In the income-age example it can be seen that  $f_1(x_1) \geq 0$ ,  $f_2(x_2) \geq 0$  and  $\sum_i f_1(x_{1i}) = 1$  and  $\sum_i f_2(x_{2i}) = 1$ .

Because of its importance in what follows let us consider the marginal density functions in the case of the bivariate normal density:

$$\begin{aligned} f_1(x_1) &= \int_{-\infty}^{\infty} \frac{(1-\rho^2)^{-\frac{1}{2}}}{2\pi\sigma_1\sigma_2} \\ &\times \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[ \left( \frac{x_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \left( \frac{x_1 - \mu_1}{\sigma_1} \right) \left( \frac{x_2 - \mu_2}{\sigma_2} \right) \right. \right. \\ &\quad \left. \left. + \left( \frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right] \right\} dx_2 \end{aligned} \quad (5.36)$$

$$\begin{aligned} &= \frac{1}{\sqrt{(2\pi)\sigma_1}} \exp \left\{ -\frac{1}{2} \left( \frac{x_1 - \mu_1}{\sigma_1} \right)^2 \right\} \int_{-\infty}^{\infty} \frac{(1-\rho^2)^{-\frac{1}{2}}}{\sqrt{(2\pi)\sigma_2}} \\ &\times \exp \left\{ -\frac{(1-\rho^2)^{-1}}{2\sigma_2^2} \left[ x_2 - \mu_2 - \rho \frac{\sigma_2}{\sigma_1} (x_1 - \mu_1) \right]^2 \right\} dx_2 \end{aligned} \quad (5.37)$$

$$= \frac{1}{\sqrt{(2\pi)\sigma_1}} \exp \left\{ -\frac{1}{2} \left( \frac{x_1 - \mu_1}{\sigma_1} \right)^2 \right\}, \quad (5.38)$$

since the integral equals one, the integrand being a proper conditional density function (see Section 5.4 below).

Similarly, we can show that

$$f_2(x_2) = \frac{1}{\sqrt{(2\pi)\sigma_2}} \exp \left\{ -\frac{1}{2} \left( \frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right\}. \quad (5.39)$$

Hence, the marginal density functions of jointly normal r.v.'s are univariate normal.

In conclusion we observe that *marginalisation* provides us with ways to simplify a probability model when such model is defined in terms of joint density functions by 'taking out' any unwanted random variables. In general, the marginal density of the r.v.'s of interest  $X_1, X_2, \dots, X_k$  can be

derived from the joint density function of  $X_1, X_2, \dots, X_k, X_{k+1}, \dots, X_n$  via

$$\begin{aligned} f_{1,2,\dots,k}(x_1, x_2, \dots, x_k) \\ = \underbrace{\int_{-\infty}^x \int_{-\infty}^x \cdots \int_{-\infty}^x}_{n-k} f(x_1, x_2, \dots, x_k, x_{k+1}, \dots, x_n) dx_{k+1}, \dots, dx_n. \end{aligned} \quad (5.40)$$

In the income-age example if age is not relevant in our investigation we can simplify the probability by marginalising out  $X_2$ .

#### 5.4 Conditional distributions

In the previous section we considered the question of simplifying probability models of the form (22) by marginalising out some subset of the r.v.'s  $X_1, X_2, \dots, X_n$ . This amounts to 'throwing away' the information related to the r.v.'s; integrated out as being irrelevant. In this section we consider the question of simplifying  $\Phi$  by *conditioning* with respect to some subset of the r.v.'s.

In the context of the probability space  $(S, \mathcal{F}, P(\cdot))$  the conditional probability of event  $A_1$  given event  $A_2$  is defined by (see Section 3.3):

$$P(A_1 | A_2) = \frac{P(A_1 \cap A_2)}{P(A_2)}, \quad P(A_2) > 0; \quad A_1, A_2 \in \mathcal{F}. \quad (5.41)$$

By choosing  $A_1 = \{s: X_1(s) \leq x_1\}$  we could use the above formula to derive an analogous definition in terms of distribution functions, that is

$$F_{x_1 | A_2}(x_1 | A_2) = P(X_1 \leq x_1 | A_2), \quad (5.42)$$

where

$$P(X_1 \leq x_1 | A_2) = \frac{P(\{s: X_1(s) \leq x_1\} \cap A_2)}{P(A_2)}. \quad (5.43)$$

As far as event  $A_2$  is concerned there are two related forms we are particularly interested in,  $A_2 = \{X_2 = \tilde{x}_2\}$ , where  $\tilde{x}_2$  is a specific value taken by  $X_2$ , and  $A_2 = \sigma(X_2)$ , i.e. the  $\sigma$ -field generated by  $X_2$ . In the case where  $A_2 = \sigma(X_2)$ , there are no particular problems arising in the definition of the conditional distribution function

$$F_{X_1 | \sigma(X_2)} = \frac{P(\{s: X_1(s) \leq x_1\} \cap \sigma(X_2))}{P(\sigma(X_2))}, \quad (5.44)$$

since  $\sigma(X_2) \in \mathcal{F}$ , although it is not particularly clear what form  $F_{X_1 | \sigma(X_2)}$  will take. In the case where  $A_2 = \{s: X_2(s) = \tilde{x}_2\}$ , however, it is immediately obvious that since  $P(s: A_2(s) = x_2) = 0$  when  $X_2$  is a continuous r.v., there will

be problems. When  $X_2$  is a discrete r.v. there are no problems arising and we can define the *conditional density function* in direct analogy to the conditional probability formula:

$$\begin{aligned} f(x_1/\tilde{x}_2) &= \Pr(X_1 = x_1 / X_2 = \tilde{x}_2) \\ &= \frac{\Pr(X_1 = x_1, X_2 = \tilde{x}_2)}{\Pr(X_2 = \tilde{x}_2)} = \frac{f(x_1, \tilde{x}_2)}{f_2(\tilde{x}_2)}. \end{aligned} \quad (5.45)$$

The upper tilda is used to emphasise the fact that it refers to just one value taken by  $X_2$ .

### *Example*

Let us consider the income–age example of the previous section in order to derive the conditional density for the discrete r.v. case. Assume that we want to derive the conditional density of  $X_2$  given  $X_1 = 6$  (income class of over £50 000). This conditional density takes the form:

$$\begin{aligned} f(x_2/\tilde{x}_1) &= \frac{0.005}{0.035} = 0.143 \quad \text{for } X_2 = 1 \text{ (young)} \\ &= \frac{0.010}{0.035} = 0.286 \quad \text{for } X_2 = 2 \text{ (middle aged)} \\ &= \frac{0.020}{0.035} = 0.571 \quad \text{for } X_2 = 3 \text{ (senior).} \end{aligned}$$

This example shows that conditioning is very different from marginalising because in the latter case all the information related to the r.v.’s integrated out is lost but in the former case some of that information in the form of the value taken by the conditioning variable is included in the conditional density.

In the case of continuous random variables it does not make sense to use the above procedure because

$$\frac{\Pr(X_1 = x_1, X_2 = \tilde{x}_2)}{\Pr(X_2 = \tilde{x}_2)} = \frac{0}{0}.$$

The mathematical apparatus needed to bypass this problem is rather formidable but we can get the gist of defining the conditional distribution in the continuous case by using the following heuristic argument. Let us define the two events to be

$$A_1 = \{s: X_1(s) \leq x_1\} \in \mathcal{F} \quad (5.46)$$

and

$$A_2 = \{s: \tilde{x}_2 - h < X_2(s) \leq \tilde{x}_2 + h\}, \quad (5.47)$$

since  $A_2 \in \mathcal{F}$  we can define the conditional probability to be

$$P(\{X_1 \leq x_1\} / \tilde{x}_2 - h < X_2 \leq \tilde{x}_2 + h) = \frac{P(\{X_1 \leq x_1\} \cap \{\tilde{x}_2 - h < X_2 \leq \tilde{x}_2 + h\})}{P(\{\tilde{x}_2 - h < X_2 \leq \tilde{x}_2 + h\})}. \quad (5.48)$$

This enables us to define the conditional distribution of  $X_1$  given  $X_2 = \tilde{x}_2$  by taking the limit as  $h \rightarrow 0$ . That is,

$$\begin{aligned} F_{X_1/X_2}(x_1 / X_2 = \tilde{x}_2) &= \lim_{0 < h \rightarrow 0} Pr(X_1 \leq x_1 / \tilde{x}_2 - h < X_2 \leq \tilde{x}_2 + h) \\ &= \lim_{0 < h \rightarrow 0} \frac{\int_{u=\tilde{x}_2-h}^{\tilde{x}_2+h} \int_{-\infty}^{x_1} f(v, u) du dv}{\int_{\tilde{x}_2-h}^{\tilde{x}_2+h} f_2(u) du} \\ &= \int_{-\infty}^{x_1} \frac{f(v, \tilde{x}_2)}{f_2(\tilde{x}_2)} dv. \end{aligned} \quad (5.49)$$

Using this heuristic argument we can define the *conditional density* of  $X_1$  given  $X_2 = x_2$  to be

$$f_{X_1/x_2}(x_1 / x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}, \quad x_1 \in \mathbb{R}_{X_1}. \quad (5.50)$$

Similarly,

$$f_{X_2/x_1}(x_2 / x_1) = \frac{f(x_1, x_2)}{f_1(x_1)}, \quad x_2 \in \mathbb{R}_{X_2}, \quad (5.51)$$

provided  $f_1(x_1) > 0$ ,  $f_2(x_2) > 0$ .

### Examples

- (1) Consider the bivariate logistic distribution

$$\begin{aligned} F(x_1, x_2) &= (1 + e^{-x_1} + e^{-x_2})^{-1} \\ \Rightarrow F_{X_1}(x_1) &= (1 + e^{-x_1})^{-1} \quad \text{since } \lim_{x_2 \rightarrow \infty} (e^{-x_2}) = 0. \\ \Rightarrow f_{X_1/X_2}(x_1 / x_2) &= \frac{f(x_1, x_2)}{f_2(x_2)} = \frac{2e^{-x_1-x_2}(1+e^{-x_1}+e^{-x_2})^{-3}}{e^{-x_2}(1+e^{-x_2})^{-2}} \\ &= 2(1+e^{-x_2})^{-1} \left( \frac{1+e^{-x_1}e^{-x_2}}{1+e^{-x_2}} \right)^{-3} e^{-x_1}. \end{aligned}$$

- (2) The distribution function of the bivariate exponential distribution takes the form

$$\begin{aligned} F(x_1, x_2) &= 1 - e^{-x_1} - e^{-x_2} + \exp\{-(x_1 + x_2 + \theta x_1 x_2)\}, \\ \Rightarrow & \quad x_1, x_2 > 0, \quad \theta \in [0, 1] \end{aligned}$$

$$F_{x_2}(x_2) = 1 - e^{-x_2}.$$

Hence

$$\begin{aligned} f_{X_1/X_2}(x_1/x_2) &= \frac{[(1+x_1)(1+\theta x_2) - \theta] \exp\{-x_1 - x_2 - \theta x_1 x_2\}}{e^{-x_2}} \\ &= [(1+\theta x_1)(1+\theta x_2) - \theta] \exp\{-x_1(1+\theta x_2)\}. \end{aligned}$$

- (3) Let  $f(x_1, x_2) = \lambda(\lambda+1)(a_1, a_2)^{\lambda+1}(a_2 x_1 + a_1 x_2 - a_1 a_2)^{-(\lambda+2)}$ ,  
 $x_1 > a_1 > 0, x_2 > a_2 > 0$ .

$$\begin{aligned} f_2(x_2) &= \int_{a_1}^{\infty} f(x_1, x_2) dx_1 = \lambda a_2^\lambda x_2^{-(\lambda+1)} \\ \Rightarrow f_{X_1/X_2}(x_1/x_2) &= \frac{\lambda(\lambda+1)(a_1 a_2)^{\lambda+1}(a_2 x_1 + a_1 x_2 - a_1 a_2)^{-(\lambda+2)}}{\lambda a_2^\lambda x_2^{-(\lambda+1)}} \\ &= a_2(\lambda+1)(a_1 x_2)^{(\lambda+1)}(a_2 x_1 + a_1 x_2 - a_1 a_2)^{-(\lambda+2)}. \end{aligned}$$

There are two things to note in relation to conditional density functions brought out by the above examples:

- (a) the conditional density is a proper density function, i.e.

$$(i) \quad f_{X_1/X_2}(x_1/\tilde{x}_2) \geq 0; \quad (5.52)$$

$$(ii) \quad \int_{-\infty}^{\infty} f_{X_1/X_2}(x_1/\tilde{x}_2) dx_1 = 1; \quad (5.53)$$

both properties can be verified in the above examples.

- (b) If we vary  $X_2$ , that is, allow  $X_2$  to take all its values in  $\mathbb{R}_{X_2}$ , we get a different conditional density for each  $X_2 = x_2$  of the form

$$f_{X_1/X_2}(x_1/x_2), \quad x_1 \in \mathbb{R}_{X_1}, \quad x_2 \in \mathbb{R}_{X_2}, \quad (5.54)$$

A moment's reflection suggests that knowledge of all these conditional densities is equivalent to knowledge of  $f(x_1, x_2)$ , a relationship brought out by the general equality

$$f(x_1, x_2) = f_{X_1/X_2}(x_1/x_2) \cdot f_2(x_2) \quad (5.55)$$

$$= f_{X_2/X_1}(x_2/x_1) \cdot f_1(x_1), \quad (x_1, x_2) \in \mathbb{R}^2. \quad (5.56)$$

This equality will prove of great value in the context of the probability

model  $\Phi = \{f(x_1, x_2, \dots, x_n; \theta), \theta \in \Theta\}$  because it offers us a general way to decompose the joint density function. It can be seen as a generalisation of the equality  $f(x_1, x_2) = f_1(x_1) \cdot f_2(x_2)$ , holding when  $X_1$  and  $X_2$  are *independent*, considered in the previous section; (55)–(56) being valid for any joint density function. Indeed, we can use the condition which makes the two equalities coincide as an alternative definition of independence, i.e.  $X_1$  and  $X_2$  are *independent* if

$$f_{X_1, X_2}(x_1/x_2) = f_1(x_1), \quad x_1 \in \mathbb{R}. \quad (5.57)$$

This definition of independence can be viewed as saying that the information relating to  $X_2$  is irrelevant in attributing probabilities to  $X_1$ . Looking back at the way  $f_1(x_1)$  was derived from the bivariate normal density we can see that the expression inside the integral in (37) was the conditional density of  $X_2$  given  $X_1 = x_1$ . It can be verified directly that in this case

$$\begin{aligned} f(x_1, x_2) &= \left( \frac{1}{\sqrt{(2\pi\sigma_1)}} \exp \left\{ -\frac{1}{2} \left( \frac{x_1 - \mu_1}{\sigma_1} \right)^2 \right\} \frac{(1-\rho^2)^{-\frac{1}{2}}}{\sigma_2 \sqrt{(2\pi)}} \right. \\ &\quad \times \left. \exp \left\{ -\frac{(1-\rho^2)^{-1}}{2\sigma_2^2} \left( x_2 - \mu_2 - \rho \frac{\sigma_2}{\sigma_1} (x_1 - \mu_1) \right)^2 \right\} \right) \\ &\quad (5.58) \end{aligned}$$

$$= (f_1(x_1))(f_{X_2/X_1}(x_2/x_1)). \quad (5.59)$$

The marginal and conditional distributions in this case are denoted by

$$X_1 \sim N(\mu_1, \sigma_1^2), \quad X_2 \sim N(\mu_2, \sigma_2^2), \quad (5.60)$$

$$(X_1/X_2) \sim N\left(\mu_1 + \rho \frac{\sigma_1}{\sigma_2} (x_2 - \mu_2), \sigma_1^2(1-\rho^2)\right), \quad (5.61)$$

$$(X_2/X_1) \sim N\left(\mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x_1 - \mu_1), \sigma_2^2(1-\rho^2)\right). \quad (5.62)$$

We could generalise the above equalities to the general  $n$ -dimensional case:

$$\begin{aligned} f(x_1, x_2, \dots, x_n) \\ = f(x_n/x_{n-1}, \dots, x_n) f(x_{n-1}/x_{n-2}, \dots, x_1) \dots f_1(x_1). \quad (5.63) \end{aligned}$$

When  $X_1, X_2, \dots, X_n$  are *independent*, then the above equality becomes

$$f(x_1, x_2, \dots, x_n) = f_n(x_n) \cdot f_{n-1}(x_{n-1}) \dots f_1(x_1) \equiv \prod_{i=1}^n f_i(x_i). \quad (5.64)$$

A sequence of r.v.'s  $X_1, X_2, \dots, X_n$  is said to be *identically distributed* if, for any  $x \in \mathbb{R}$ ,

$$F_1(x) = F_2(x) = \dots = F_n(x). \quad (5.65)$$

The concept of *conditioning* enables us to simplify the probability model in the form of a parametric family of multivariate density function  $\Phi = \{f(x_1, x_2, \dots, x_n; \theta), \theta \in \Theta\}$  in two related ways:

- (i) to decompose the joint density function into a product of conditional densities which can make the manipulation much easier; and
- (ii) in the case where the stochastic (probabilistic) information in some of the r.v.'s is not of interest we can use the conditional density with respect to some observed values for these r.v.'s. For instance in the case of the income-age example if we were to consider the question of poverty in relation to age we would concentrate on  $f_{x_2/x_1}(x_2/x_1 = 1)$  exclusively.

### ***Important concepts***

random vector, the induced probability space ( $\mathbb{R}^n, \mathcal{B}^n, P_x(\cdot)$ );  
 the joint distribution and density functions;  
 marginal distribution and density functions, marginal normal density;  
 independent r.v.'s; identically distributed r.v.'s;  
 conditional distribution and density functions, conditional normal density.

### ***Questions***

1. Why do we need to extend the concept of a random variable to that of a random vector defined on  $(S, \mathcal{F}, P(\cdot))$ ?
2. Explain how we can extend the definition of a r.v. to that of a random vector stressing the role of half-closed intervals  $(-\infty, \mathbf{x}]$ .
3. Explain the relationships between  $P(\cdot)$ ,  $P_x(\cdot)$  and  $F_{X_1, X_2}(\cdot, \cdot)$ .
4. Compare the concepts of marginalisation and conditioning.
5. Define the concepts of marginal and conditional density functions and discuss their properties.
6. Define the concept of independence among r.v.'s via both marginal as well as conditional density functions.
7. Define the concept of identically distributed r.v.'s.

### ***Exercises***

1. Consider the random experiment of tossing a fair coin three times.

- (i) Derive the sample space  $S$ .
  - (ii) Define the random variables  $X_1$  – the number of ‘heads’,  $X_2 = |\text{number of ‘heads’} - \text{number of ‘tails’}|$ .
  - (iii) Derive the  $\sigma$ -fields generated by  $X_1$ ,  $\sigma(X_1)$ ,  $X_2$ ,  $\sigma(X_2)$  and  $\sigma(X_1, X_2)$ .
  - (iv) Define the joint distribution and density functions  $F(x_1, x_2)$ ,  $f(x_1, x_2)$  and plot  $f(x_1, x_2)$ .
  - (v) Derive the marginal density function  $f_1(x_1)$  and  $f_2(x_2)$ .
  - (vi) Derive the conditional densities
- $f_1(x_1/1)$ ,  $f_1(x_1/2)$ ,  $f_1(x_1/3)$ ,  $f_2(x_2/0)$ ,  $f_2(x_2/1)$ .

2. For the income–age example discussed above

- (i) Plot the graph of  $f(x_1, x_2)$ ,  $f_1(x_1)$ ,  $f_2(x_2)$ .
- (ii) Derive  $f_1(x_1/1)$ ,  $f_1(x_1/2)$ ,  $f_2(x_2/3)$  and compare their graphs with that of  $f_1(x_1)$ .

3. Let the joint distribution of  $X_1$  and  $X_2$  be bivariate normal

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right),$$

$$\mu_1 = 1, \quad \mu_2 = 2, \quad \sigma_1^2 = 2, \quad \rho = 0.5, \quad \sigma_2^2 = 4.$$

- (i) Derive  $f_1(x_1)$  and  $f_2(x_2)$ .
- (ii) Derive  $f_{X_1/X_2}(x_1/x_2)$ , for  $x_2 = 0, 1, 2$ .
- (iii) Under what circumstances are  $X_1$  and  $X_2$  independent?

4. Let the joint density function of  $X_1$  and  $X_2$  be

$$f(x_1, x_2) = x_1 \exp\{-x_1(1+x_2)\}, \quad x_1 > 0, \quad x_2 > 0.$$

- (i) Derive  $f_1(x_1)$  and  $f_2(x_2)$ .
- (ii) Derive  $f_{X_1/X_2}(x_1/x_2)$  for  $x_2 = 1, 2, 10$ , and  $f_{X_2/X_1}(x_2/x_1)$  for  $x_1 = 1, 2, 10$ .

#### Additional references

Bickel and Doksum (1977); Chung (1974); Clarke (1975); Cramer (1946); Dudewicz (1976); Giri (1974); Mood, Graybill and Boes (1974); Pfeiffer (1978); Rohatgi (1976).

## CHAPTER 6

### Functions of random variables

One of the most important problems in probability theory and statistical inference is to derive the distribution of a function  $h(X_1, X_2, \dots, X_n)$  when the distribution of the random vector  $\mathbf{X} = (X_1, \dots, X_n)$  is known. This problem is important for at least two reasons:

- (i) it is often the case that in modelling observable phenomena we are primarily interested in functions of random variables; and
- (ii) in statistical inference the quantities of primary interest are commonly functions of random variables.

It is no exaggeration to say that the whole of statistical inference is based on our ability to derive the distribution of various functions of r.v.'s. In the first subsection we are going to consider the distribution of functions of a single r.v. and then consider the case of functions of random vectors.

#### 6.1 Functions of one random variable

Let  $X$  be a r.v. on the probability space  $(S, \mathcal{F}, P(\cdot))$ . By definition,  $X(\cdot): S \rightarrow \mathbb{R}$ , i.e.  $X$  is a real valued function on  $S$ . Suppose that  $h(\cdot): \mathbb{R} \rightarrow \mathbb{R}$ , where  $h$  is a continuous function with at most a countable number of discontinuities. More formally we need  $h(\cdot)$  to be a *Borel* function.

*Definition 1*

A function  $h(\cdot): \mathbb{R}_x \rightarrow \mathbb{R}$  is said to be a **Borel function** if for any  $a \in \mathbb{R}$  and  $x \in \mathbb{R}_x$  the set  $B_h = \{x: h(x) \leq a\}$  is a Borel set, i.e.  $B_h \in \mathcal{B}$ , where  $\mathcal{B}$  is the Borel field on  $\mathbb{R}$  (see Section 3.2).

Requiring that  $h(\cdot)$  is a Borel function is an obvious condition to impose given that we need  $h(X)$  to be a random variable itself.

We know that  $X$  is a function from  $S$  to  $\mathbb{R}$  and thus,  $h(X(s))$  can be

#### 6.1 Functions of one random variable

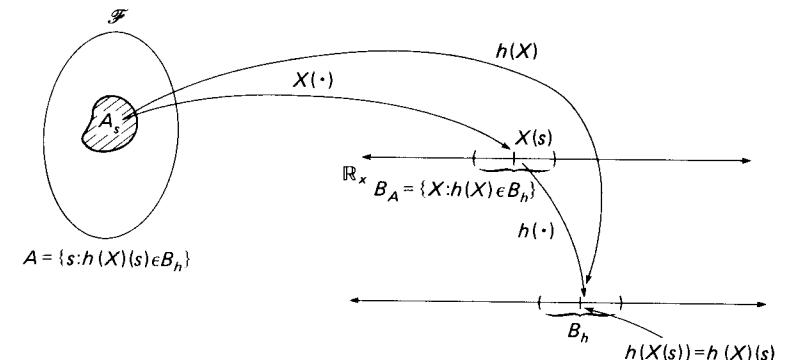


Fig. 6.1. A Borel function of a random variable.

considered a function from  $S$  to  $\mathbb{R}$  and the above ensures that the *composite function*  $h(X): S \rightarrow \mathbb{R}$  is indeed a random variable, i.e. the set  $A = \{s: h(X)(s) \in B_h\} \in \mathcal{F}$  for any  $B_h \in \mathcal{B}$  (see Fig. 6.1). Let us denote the r.v.  $h(X)$  by  $Y$ , then  $Y$  induces a probability set function  $P_y(\cdot)$  such that  $P_y(B_h) = P_x(B_A) = P(A)$ , in order to preserve the probability structure of the original  $(S, \mathcal{F}, P(\cdot))$ . Note that the reason we need  $h(\cdot)$  to be a Borel function is to preserve the event structure of  $\mathcal{B}$ .

Having ensured that the function  $h(\cdot)$  of the r.v.  $X$  is itself a r.v.  $Y = h(X)$  we want to derive the distribution of  $Y$  when the distribution of  $X$  is known. Let us consider the discrete case first. When  $X$  is a discrete r.v. the  $Y = h(X)$  is again a discrete r.v. and all we need to do is to give the set of values of  $Y$  and the corresponding probabilities. Consider the coin-tossing example where  $X$  is the r.v. defined by  $X = (\text{number of } H - \text{number of } T)$ , then since  $S = \{HT, TH, HH, TT\}$ ,  $X(HT) = X(TH) = 0$ ,  $X(HH) = 2$ ,  $X(TT) = -2$  and the probability function is

$$\begin{array}{llll} X = x & -2 & 0 & 2 \\ Pr(X = x) & \frac{1}{4} & \frac{1}{2} & \frac{1}{4}. \end{array}$$

Let  $Y = X^2$ , then  $Y$  takes values  $(-2)^2 = 4$ ,  $(0)^2 = 0$ ,  $2^2 = 4$  with the same probabilities as  $X$  but since 4 occurs twice we add the probabilities, i.e.

$$\begin{array}{lll} Y = y & 0 & 4 \\ Pr(Y = y) & \frac{1}{2} & \frac{1}{2}. \end{array}$$

In general, the distribution function of  $Y$  is defined as

$$F(y) = P(s: Y(s) \leq y) = P(s: X(s) \in h^{-1}((-\infty, y])), \quad (6.1)$$

where the inverse function  $h^{-1}(\cdot)$  need not be unique.

In the case where  $X$  is a continuous r.v., deriving the distribution of  $Y = h(X)$  is not as simple as the discrete case because, firstly,  $Y$  is not always a continuous r.v. as well and, secondly, the solution to the problem depends crucially on the nature of  $h(\cdot)$ . A sufficient condition for  $Y$  to be a continuous r.v. as well is given by the following lemma.

#### Lemma

Let  $X$  be a continuous r.v. and  $Y = h(X)$  where  $h(X)$  is differentiable for all  $x \in \mathbb{R}_x$  and  $[dh(x)]/(dx) > 0$  or  $[dh(x)]/(dx) < 0$  for all  $x$ . Then the density function of  $Y$  is given by

$$\blacksquare f_y(y) = f_x(h^{-1}(y)) \left| \frac{d}{dy} h^{-1}(y) \right| \quad \text{for } a < y < b, \quad (6.2)$$

where  $\| \cdot \|$  stands for the absolute value and  $a$  and  $b$  refer to the smallest and biggest value  $y$  can take, respectively.

#### Example 1

Let  $X \sim N(\mu, \sigma^2)$  and  $Y = (X - \mu)/\sigma$ , which implies that  $[dh(x)]/(dx) = 1/\sigma > 0$  for all  $x \in \mathbb{R}$  since  $\sigma > 0$  by definition;  $h^{-1}(y) = \sigma y + \mu$  and  $[dh^{-1}(y)]/(dy) = \sigma$ . Thus since

$$f_x(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\},$$

$$f_y(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\sigma y + \mu - \mu}{\sigma}\right)^2\right\} \cdot (\sigma) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}y^2\right\},$$

i.e.  $Y \sim N(0, 1)$  the standard normal distribution.

In cases where the conditions of Lemma 1 are not satisfied we need to derive the distribution from the relationship

$$F_y(y) = Pr(h(x) \geq y) = Pr(X \in h^{-1}((-\infty, x])). \quad (6.3)$$

#### Example 2

Let  $X \sim N(\mu, \sigma^2)$  and  $Y = X^2$  (see Fig. 6.2). Since  $[dh(x)]/(dx) = 2x$  we can see that  $h(x)$  is monotonically increasing for  $x > 0$  and monotonically decreasing for  $x < 0$  and Lemma 2 is not satisfied. However, for  $y > 0$

$$F_y(y) = Pr(h(x) \leq y) = Pr(x \in h^{-1}(-\infty, x])$$

$$= Pr(-\sqrt{y} \leq X \leq \sqrt{y}) = F_x(\sqrt{y}) - F_x(-\sqrt{y})$$

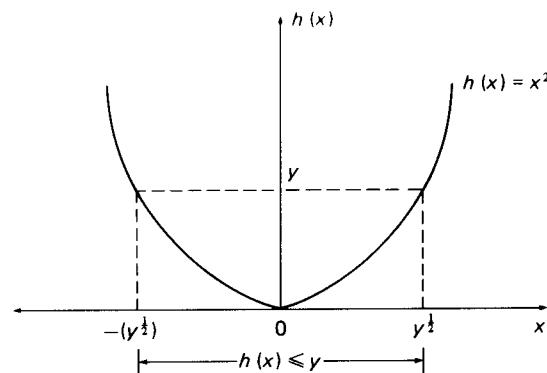


Fig. 6.2. The function  $Y = X^2$  where  $X$  is normally distributed.

(see Fig. 6.2). In this form we can apply the above lemma with  $[dh^{-1}(y)]/(dy) = 1/(2\sqrt{y})$  for  $x > 0$  and  $x < 0$  separately to get

$$f_y(y) = f_x(\sqrt{y}) \left( \frac{1}{2\sqrt{y}} \right) + f_x(-\sqrt{y}) \left( \frac{1}{2\sqrt{y}} \right) \quad \text{for } y > 0$$

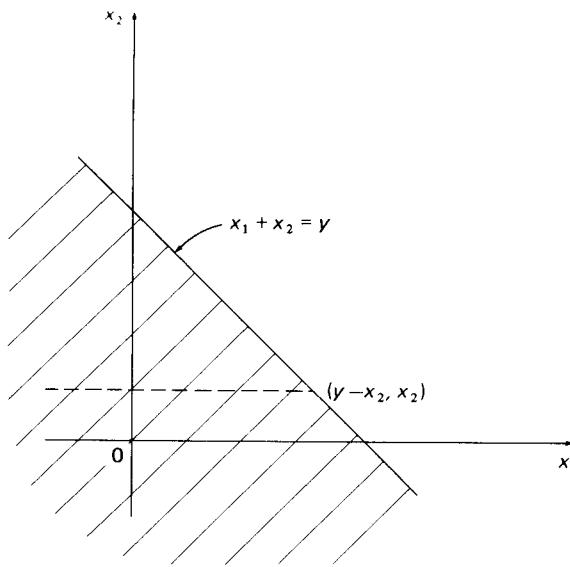
$$= \frac{1}{2} \left( \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}y\right\} + \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}y\right\} \right) y^{-\frac{1}{2}}$$

$$= \frac{\frac{1}{2}}{\Gamma(\frac{1}{2})} (\frac{1}{2}y)^{\frac{1}{2}-1} \exp(-\frac{1}{2}y) \quad y > 0.$$

That is,  $f_y(y)$  is the so-called gamma density, where  $\Gamma(\cdot)$  is the gamma function ( $\Gamma(n) = \int_0^\infty v^n e^{-v} dv$ ). A gamma r.v. denoted by  $Y \sim G(r, p)$  has a density of the form  $f(y) = [p/\Gamma(r)](py)^{r-1} \exp(-py)$ ,  $y > 0$ . The above distribution is  $G(\frac{1}{2}, \frac{1}{2})$  and is known as the *chi-square distribution*; an important distribution in statistical inference; see Appendix 6.1.

#### 6.2\* Functions of several random variables

As in the case of a single r.v. for a Borel function  $h(\cdot): \mathbb{R}^n \rightarrow \mathbb{R}$  and a random vector  $\mathbf{X} = (X_1, X_2, \dots, X_n)$ ,  $h(\mathbf{X})$  is a random variable. Let us consider certain commonly used functions of random variables concentrating on the two variables case for convenience of exposition.

Fig. 6.3. The function  $Y = X_1 + X_2$ .(1) **The distribution of  $X_1 + X_2$** 

By definition the distribution function of  $Y = X_1 + X_2$  (see Fig. 6.3) is

$$F_Y(y) = \Pr(X_1 + X_2 \leq y) = \int_{-\infty}^{\infty} \int_{-\infty}^{y-x_2} f(x_1, x_2) dx_1 dx_2 \quad (6.3)$$

$$\Rightarrow F_Y(y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(u-x_2, x_2) du dx_2 \quad (6.4)$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(y-x_2, x_2) dx_2, \quad y \in \mathbb{R}.$$

In particular, if  $X_1$  and  $X_2$  are independent, then

$$f_Y(y) = \int_{-\infty}^{\infty} f_1(y-x_2) f_2(x_2) dx_2 = \int_{-\infty}^{\infty} f_1(x_1) f_2(y-x_1) dx_1,$$

by symmetry. Using an analogous argument we can show that for  $Y = X_1 - X_2$

$$f_Y(y) = \int_{-\infty}^{\infty} f(y+x_2, x_2) dx_2 = \int_{-\infty}^{\infty} f(x_1, x_1-y) dx_1.$$

For  $X_1$  and  $X_2$  independent

$$f(y+x_2, x_2) = f_1(y+x_2) f_2(x_2) \\ f_1(x_1) f_2(x_1-y).$$

*Example 3*

Let  $X_1 \sim N(\mu_1, \sigma_1^2)$ ,  $X_2 \sim N(\mu_2, \sigma_2^2)$ ,  $X_1$  and  $X_2$  are independent r.v.'s. Define  $Y = X_1 + X_2$ , then

$$f_Y(y) = \int_{-\infty}^{\infty} \left[ \frac{1}{\sigma_1 \sqrt{(2\pi)}} \exp \left\{ -\frac{1}{2} \left( \frac{y-x_2-\mu_1}{\sigma_1} \right)^2 \right\} \right] \\ \times \left[ \frac{1}{\sigma_2 \sqrt{(2\pi)}} \exp \left\{ -\frac{1}{2} \left( \frac{x_2-\mu_2}{\sigma_2} \right)^2 \right\} \right] dx_2 \\ = \frac{(\sigma_1^2 + \sigma_2^2)^{-\frac{1}{2}}}{\sqrt{(2\pi)}} \exp \left\{ -\frac{(y-(\mu_1+\mu_2))^2}{2(\sigma_1^2 + \sigma_2^2)} \right\} \cdot 1.$$

Hence,  $Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ . In general if  $X_1, X_2, \dots, X_n$  are independent r.v.'s with  $X_i \sim N(\mu_i, \sigma_i^2)$ ; then

$$Y = \sum_{i=1}^n X_i \sim N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right).$$

*Example 4*

Let  $X_i \sim U(-1, 1)$ ,  $i = 1, 2$  (uniformly distributed), and define  $Y = X_1 + X_2$ . Using Fig. 6.3 we can show that

$$f_Y(y) = \begin{cases} 0, & |y| \geq 2 \\ \frac{2-|y|}{4}, & |y| \leq 2 \end{cases}$$

(see Fig. 6.4). For  $X_i \sim U(-1, 1)$ ,  $i = 1, 2, 3$  and  $Y = X_1 + X_2 + X_3$  we can show

$$f_Y(y) = \begin{cases} 0, & |y| \geq 3 \\ \frac{(3-|y|)^2}{16}, & 1 \leq |y| \leq 3 \\ \frac{3-y^2}{8}, & 0 \leq |y| \leq 1. \end{cases}$$

This density function is shown below (see Fig. 6.5) and as can be seen it is not

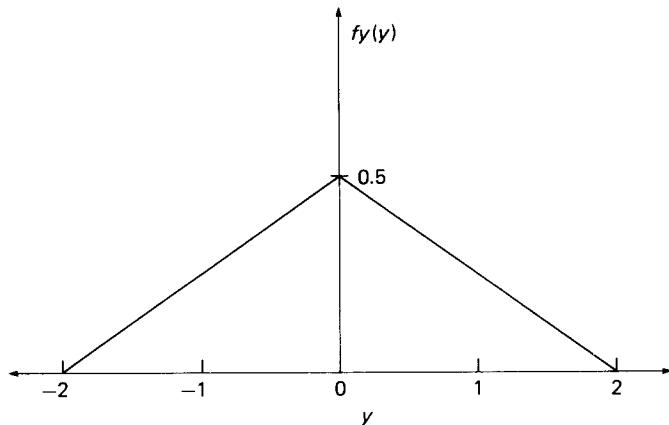


Fig. 6.4. The density function of  $Y = X_1 + X_2$  where  $X_1$  and  $X_2$  are uniformly distributed.

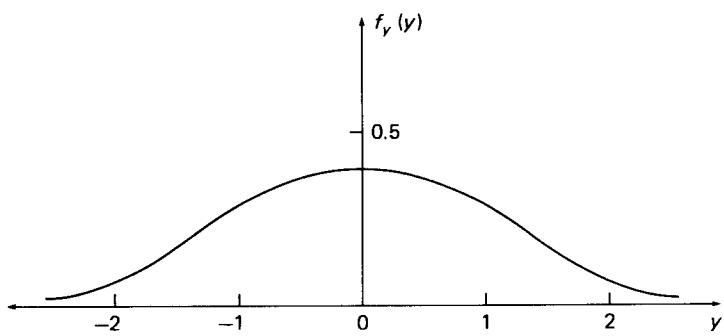


Fig. 6.5. The density function of  $Y = X_1 + X_2 + X_3$  where  $X_i, i = 1, 2, 3$ , are uniformly distributed.

only continuous but also differentiable everywhere. The shape of the curve is very much like the normal density. This is a general result which states that for  $X_i \sim U(-1, 1), i = 1, 2, \dots$  uniformly distributed independent r.v.'s,  $Y_n = \sum_{i=1}^n X_i$  has a distribution which is closer to a normal distribution the greater the value of  $n$ ; a particular case of the central limit theorem (see Chapter 9).

## (2) The distribution of $X_1/X_2$

Consider two r.v.'s  $X_1$  and  $X_2$  and let  $Y = X_1/X_2$ . The distribution of  $Y$

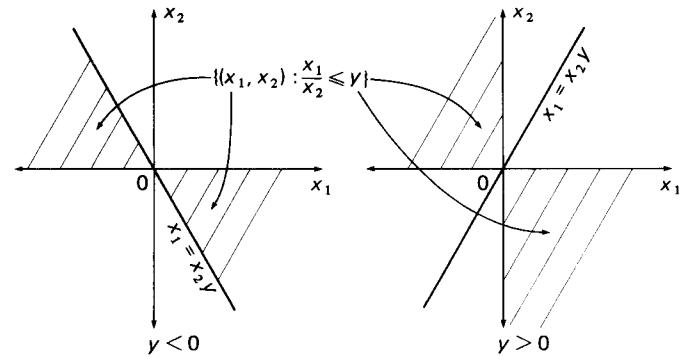


Fig. 6.6. The function  $Y = X_1/X_2$  for  $Y < 0$  and  $Y > 0$ .

takes the form

$$\begin{aligned} F_y(y) &= \int_0^\infty \int_{-\infty}^{yx_2} f(x_1, x_2) dx_1 dx_2 \\ &\quad + \int_{-\infty}^0 \int_{yx_2}^\infty f(x_1, x_2) dx_1 dx_2, \quad y \in \mathbb{R}, \end{aligned} \quad (6.5)$$

as suggested by Fig. 6.6. For  $u = (x_1/x_2)$ ,

$$\begin{aligned} F_y(y) &= \int_0^\infty \int_{-\infty}^y f(ux_2, x_2) x_2 du dx_2 \\ &\quad + \int_{-\infty}^0 \int_y^\infty f(ux_2, x_2) x_2 du dx_2, \end{aligned} \quad (6.6)$$

$$f_y(y) = \int_{-\infty}^\infty |x_2| f(yx_2, x_2) dx_2, \quad y \in \mathbb{R}. \quad (6.7)$$

In the case where  $X_1$  and  $X_2$  are independent this becomes

$$f_y(y) = \int_{-\infty}^\infty |x_2| f_1(yx_2) f_2(x_2) dx_2, \quad y \in \mathbb{R}. \quad (6.8)$$

*Example 5* (the mathematical manipulations are not important!) Let  $X_1 \sim N(0, 1)$  and  $X_2 \sim \chi^2(n)$  – chi-square with  $n$  degrees of freedom,  $X_1$  and  $X_2$  being independent. Define  $Y = X_1/\sqrt{(X_2/n)}$  and let us derive its distribution. The density function of the denominator  $Z = \sqrt{(X_2/n)}$  is given

by

$$f_x(z) = \frac{n^{n/2}}{2^{\lfloor(n/2)-1\rfloor}\Gamma(n/2)} z^{n-1} \exp\left\{-\frac{nz^2}{2}\right\}, \quad z > 0.$$

Since  $f(x_1, z) = f_1(x_1) \cdot f_2(z)$ , it takes values only for  $z > 0$ , which implies that

$$\begin{aligned} f_y(y) &= \int_0^\infty z \left[ \frac{1}{\sqrt{(2\pi)}} \exp\left\{-\frac{yz^2}{2}\right\} \right] \left[ \frac{n^{n/2} z^{n-1}}{2^{\lfloor(n/2)-1\rfloor}\Gamma(n/2)} \exp\left\{-\frac{nz^2}{2}\right\} \right] dz \\ &= \frac{n^{(n/2)}}{\sqrt{(2\pi)2^{\lfloor(n/2)-1}\Gamma(n/2)}} \int_0^\infty \exp\left\{-\frac{1}{2}(n+y^2)z^2\right\} dz \\ &= \frac{\Gamma[(n+1)/2]}{\Gamma(n/2)\sqrt{(n\pi)}} \left(1 + \frac{y^2}{n}\right)^{-1(n+1)/2}, \quad y \in \mathbb{R}. \end{aligned}$$

This is the density of *Student's t-distribution*.

#### Example 6

Let  $X_1 \sim \chi^2(n_1)$  and  $X_2 \sim \chi^2(n_2)$  be two independent r.v.'s and define

$$Y = \frac{(X_1/n_1)}{(X_2/n_2)} = \frac{n_2}{n_1} \frac{X_1}{X_2}.$$

$$\begin{aligned} f_y(y) &= \int_0^\infty \left(\frac{n_1}{n_2}\right) x_2 f_1\left(\frac{n_1}{n_2} x_2 y\right) f_2(x_2) dx_2 \\ &= \frac{(n_1/n_2) 2^{-[(n_1+n_2)/2]}}{\Gamma\left(\frac{n_1}{2}\right) \Gamma\left(\frac{n_2}{2}\right)} \int_0^\infty y^{[(n_1+n_2)/2]-1} \left(\frac{n_1}{n_2} x_2\right)^{(n_1/2)-1} \\ &\quad \exp\left\{\frac{1}{2} x_2 \left(1 + \frac{n_1}{n_2} y\right)\right\} dx_2 \\ f_y(y) &= \frac{\Gamma\left(\frac{n_1+n_2}{2}\right) \left(\frac{n_1}{n_2}\right)^{n_1/2}}{\Gamma\left(\frac{n_1}{2}\right) \Gamma\left(\frac{n_2}{2}\right) \left[1 + \left(\frac{n_1}{n_2}\right) y\right]^{\lfloor(n_1+n_2)/2\rfloor}}, \quad y > 0. \end{aligned}$$

This represents the density of Fisher's *F-distribution* with  $n_1$  and  $n_2$  degrees of freedom.

#### Example 7

Let  $X_1 \sim N(0, \sigma_1^2)$ ,  $X_2 \sim N(0, \sigma_2^2)$ ,  $X_1$  and  $X_2$  independent r.v.'s, and define  $Y = X_1/X_2$ . The density function of  $Y$  takes the form

$$\begin{aligned} f_y(y) &= \frac{1}{2\pi\sigma_1\sigma_2} \left[ \int_0^\infty x_2 \left( \exp\left\{-\frac{1}{2}\left(\frac{y^2 x_2^2}{\sigma_1^2} + \frac{x_2^2}{\sigma_2^2}\right)\right\} \right) dx_2 \right. \\ &\quad \left. - \int_{-\infty}^0 x_2 \left( \exp\left\{-\frac{1}{2}\left(\frac{y^2 x_2^2}{\sigma_1^2} + \frac{x_2^2}{\sigma_2^2}\right)\right\} \right) dx_2 \right] \\ &= \frac{1}{\pi\sigma_1\sigma_2} \int_0^\infty x_2 \left( \exp\left\{-\frac{x_2^2}{2}\left(\frac{y^2}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)\right\} \right) dx_2 \\ &= \frac{\sigma_1\sigma_2}{\pi(\sigma_2^2 y^2 + \sigma_1^2)} \int_0^\infty e^{-u} du, \quad \text{where } u = \frac{x_2^2}{2}\left(\frac{y^2}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right) \\ &= \frac{1}{\pi} \frac{\left(\frac{\sigma_1}{\sigma_2}\right)}{\left[\left(\frac{\sigma_1}{\sigma_2}\right)^2 + y^2\right]}, \quad y \in \mathbb{R}. \end{aligned}$$

The density of  $y$  is known as the *Cauchy density function*.

#### (3) The distribution of $Y = \min(X_1, X_2)$

The distribution function of  $Y = \min(X_1, X_2)$  for two r.v.'s  $X_1, X_2$  takes the general form

$$\begin{aligned} F_Y(y) &= Pr(\min(X_1, X_2) \leq y) = 1 - Pr(\min(X_1, X_2) > y) \\ &= 1 - Pr(X_1 > y, X_2 > y), \end{aligned} \tag{6.9}$$

as illustrated in Fig. 6.7. In the case where  $X_1$  and  $X_2$  are independent

$$F_Y(y) = 1 - (1 - F_1(x_1))(1 - F_2(x_2)).$$

#### Example 8

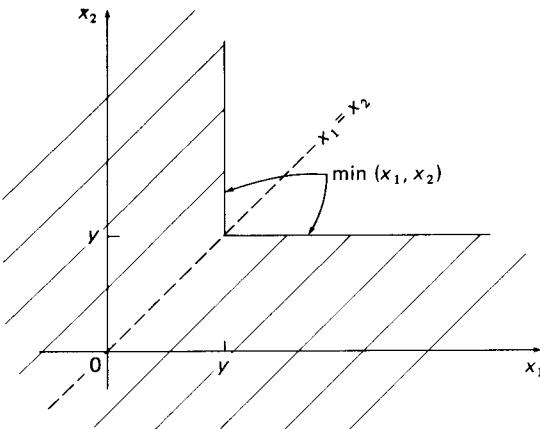
For

$$F_i(x_i) = 1 - e^{-ax_i^b}, \quad x_i \geq 0, \quad i = 1, 2,$$

$$F_3(y) = 1 - \exp\left\{-a(x_1^b + x_2^b)\right\},$$

$F_3(\cdot, \cdot)$ , known as the *Weibull distribution function*.

After considering various simple functions of r.v.'s separately, let us

Fig. 6.7. The function  $Y = \min(X_1, X_2)$ .

consider them together. Let  $(X_1, X_2, \dots, X_n)$  be a random vector with a joint probability density function  $f(x_1, x_2, \dots, x_n)$  and define the *one-to-one* transformation:

$$\begin{aligned} y_1 &= h_1(x_1, x_2, \dots, x_n) \\ y_2 &= h_2(x_1, x_2, \dots, x_n) \\ &\vdots \\ y_n &= h_n(x_1, x_2, \dots, x_n) \end{aligned} \tag{6.10}$$

whose inverse take the form  $h_i^{-1}(\cdot) = g_i(\cdot)$ ,  $i = 1, 2, \dots, n$

$$\begin{aligned} x_1 &= g_1(y_1, y_2, \dots, y_n) \\ &\vdots \\ x_n &= g_n(y_1, y_2, \dots, y_n) \end{aligned} \tag{6.11}$$

Assume:

- (i)  $h_i(\cdot)$  and  $g_i(\cdot)$  are continuous;
- (ii) the partial derivatives  $\partial x_i / \partial y_j$ ,  $i, j = 1, 2, \dots, b$ , exist and are continuous; and
- (iii) the Jacobian of the inverse transformation

$$J = \det \begin{pmatrix} \frac{\partial(x_1, x_2, \dots, x_n)}{\partial(y_1, y_2, \dots, y_n)} \end{pmatrix} \neq 0.$$

These assumptions enable us to deduce that

$$f(y_1, y_2, \dots, y_n) = f(g_1(y_1, y_2, \dots, y_n), \dots, g_n(y_1, y_2, \dots, y_n)) |J|. \tag{6.12}$$

### Example 9

Let  $X_i \sim N(0, 1)$ ,  $i = 1, 2$  be two independent r.v.'s and

$$Y_1 = h_1(X_1, X_2) = X_1 + X_2, \quad Y_2 = h_2(X_1, X_2) = \frac{X_1}{X_2}.$$

Since

$$x_1 = g_1(y_1, y_2) = \frac{y_1 y_2}{1 + y_2}, \quad X_2 = g_2(y_1, y_2) = -\frac{y_1}{(1 + y_2)^2},$$

$$J = \det \begin{pmatrix} \frac{y_2}{1 + y_2} & \frac{y_1}{(1 + y_2)^2} \\ \frac{1}{1 + y_2} & -\frac{y_1}{(1 + y_2)^2} \end{pmatrix} = -\frac{y_1}{(1 + y_2)^2},$$

this implies that

$$\begin{aligned} f(y_1, y_2) &= \frac{1}{2\pi} \exp \left\{ -\frac{1}{2} \left( \frac{(y_1 y_2)^2 + y_1^2}{(1 + y_2)^2} \right) \right\} \frac{|y_1|}{(1 + y_2)^2} \\ &\quad \frac{1}{2\pi} \frac{|y_1|}{(1 + y_2)^2} \exp \left\{ -\frac{1}{2} y_1^2 \frac{(1 - y_2)}{(1 + y_2)} \right\}. \end{aligned}$$

The main drawback of this approach is well demonstrated by the above example. The method provides us with a way to derive the joint density function of the  $Y_i$ 's and not the marginal density functions. These can be derived by integrating out the other variables. For instance,

$$f_1(y_1) = \int_{-\infty}^{\infty} f(y_1, y_2) dy_2, \quad f_2(y_2) = \int_{-\infty}^{\infty} f(y_1, y_2) dy_1, \tag{6.13}$$

and in the above example these take the form

$$f_1(y_1) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} y_1^2 \right\} \text{ normal density,}$$

$$f_2(y_2) = \frac{1}{\pi} \frac{1}{1 + y_2^2} \text{ Cauchy density.}$$

The derivations of these marginal density functions, however, involve some complicated mathematical manipulations.

### 6.3 Functions of normally distributed random variables, a summary

The above examples on functions of random variables show clearly that deriving the distribution of  $h(X_1, \dots, X_n)$  when  $f(x_1, \dots, x_n)$  is known is not an easy exercise. Indeed this is one of the most difficult problems in probability theory as argued below. Some of the above results, although involved (as far as mathematical manipulations are concerned), have been included because they play a very important role in *statistical inference*. Because of their importance generalisations of these results will be summarised below for reference purposes.

#### Lemma 6.1

If  $X_i \sim N(\mu_i, \sigma_i^2)$ ,  $i = 1, 2, \dots, n$  are independent r.v.'s then  $(\sum_{i=1}^n X_i) \sim N(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$  – **normal**.

#### Lemma 6.2

If  $X_i \sim N(0, 1)$ ,  $i = 1, 2, \dots, n$  are independent r.v.'s then  $(\sum_{i=1}^n X_i^2) \sim \chi^2(n)$  – **chi-square** with  $n$  degrees of freedom.

#### Lemma 6.2\*

If  $X_i \sim N(\mu_i, \sigma_i^2)$ ,  $i = 1, 2, \dots, n$  are independent r.v.'s then  $(\sum_{i=1}^n X_i^2/\sigma_i^2) \sim \chi^2(n; \delta)$  – **non-central chi-square** with non-centrality parameter,  $\delta = \sum_{i=1}^n \mu_i^2/\sigma_i^2$ .

#### Lemma 6.3

If  $X_1 \sim N(0, 1)$ ,  $X_2 \sim \chi^2(n)$  are  $X_1, X_2$  independent r.v.'s then  $X_1/[\sqrt{(X_2/n)}] \sim t(n)$  – **Student's t** with  $n$  degrees of freedom.

#### Lemma 6.3\*

If  $X_1 \sim N(\mu, \sigma^2)$ ,  $X_2 \sim \sigma^2 \chi^2(n)$ ,  $X_1, X_2$  independent r.v.'s then  $X_1/[\sqrt{(X_2/n)}] \sim t(n; \delta)$  – **non-central t** with non-centrality parameter  $\delta = \mu/\sigma$ .

#### Lemma 6.4

If  $X_1 \sim \chi^2(n_1)$ ,  $X_2 \sim \chi^2(n_2)$ ,  $X_1, X_2$  independent r.v.'s then  $(X_1/n_1)/(X_2/n_2) \sim F(n_1, n_2)$  – **Fisher's F** with  $n_1$  and  $n_2$  degrees of freedom.

#### Lemma 6.4\*

If  $X_1 \sim \chi^2(n_1; \delta)$ ,  $X_2 \sim \chi^2(n_2)$ ,  $X_1, X_2$  being independent r.v.'s then  $(X_1/n_1)/(X_2/n_2) \sim F(n_1, n_2; \delta)$  – **non-central F**,  $\delta$  being the non-centrality parameter.

### 6.4 Looking ahead

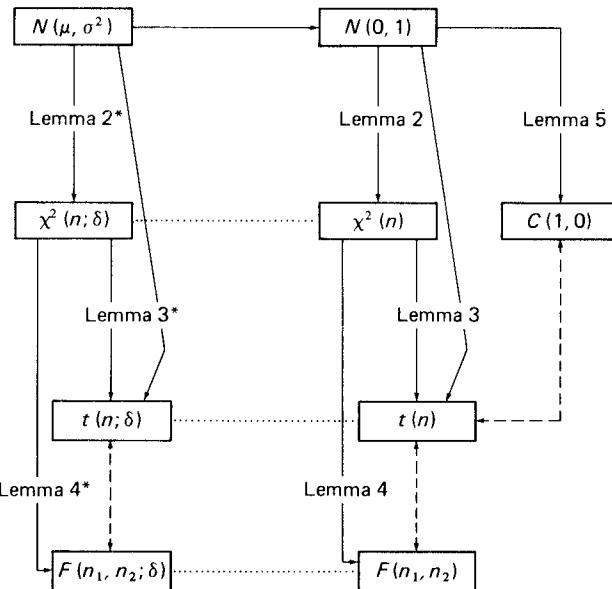


Fig. 6.8. The normal and related distributions.

#### Lemma 6.5

If  $X_i \sim N(0, 1)$ ,  $i = 1, 2$  are two independent r.v.'s then  $(X_1/X_2) \sim C(0, 1)$  – **Cauchy distribution**.

The relationships among the distributions referred to in these lemmas are depicted in Fig. 6.8. For a summary of these distributions see Appendix 6.1 below, for a more extensive discussion see the excellent book by Johnson and Kotz (1970).

Note that if  $X \sim t(n)$ ,  $Y = X^2 \sim F(1, n)$  and for  $n = 1$ ,  $t(1) = C(1, 0)$ .

### 6.4 Looking ahead

In this chapter we considered the distribution of functions of random variables. Although the mathematical manipulations are in general rather involved it is a very important facet of probability theory for two reasons:

- (i) It often occurs in practice that the probability model is not defined in terms of the original r.v.'s but in some functions of these.
- (ii) Statistical inference is crucially dependent on the distribution of functions of random variables. Estimators and test statistics are

functions of r.v.'s of the form  $h(X_1, X_2, \dots, X_n)$  and the distribution of such functions is the basis of any inference related to the unknown parameters  $\theta$ .

From the above discussion it is obvious that determining the distribution of  $h(X_1, X_2, \dots, X_n)$  is by no means a trivial exercise. It turns out that more often than not we cannot determine the distribution exactly. Because of the importance of the problem, however, we are forced to develop approximations; the subject matter of Chapter 10.

It is no exaggeration to say that most of the results derived in the context of the various statistical models in econometrics, discussed in Part IV, depend crucially on the results summarised in Section 6.3 above. Estimation, testing and prediction in the context of these models is based on the results related to functions of normally distributed random variables and the normal, Student's  $t$ , Fisher's  $F$  and chi-square distributions are used extensively in Part IV.

### Appendix 6.1 – The normal and related distributions

(1) *Univariate normal* –  $X \sim N(\mu, \sigma^2)$

$$F(x) = \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2\right\} du;$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \quad x \in \mathbb{R},$$

$$E(X) = \mu, \quad \text{Var}(X) = \sigma^2, \quad \text{skewness} = \alpha_3 = 0, \quad \text{kurtosis} = \alpha_4 = 3$$

*Higher moments:*

$$\mu_r \equiv E(X - \mu)^r = \begin{cases} \frac{\sigma^r r!}{2^{r/2} \left(\frac{r}{2}\right)!}, & r \text{ even} \\ 0, & r \text{ odd,} \end{cases}$$

*Characteristic function*  $\psi_x(t) = \exp(it\mu - \frac{1}{2}\sigma^2 t^2)$ .

*Cumulants*  $\kappa_1 = \mu, \quad \kappa_2 = \sigma^2, \quad \kappa_r = 0, \quad r = 3, 4, \dots$

*Some properties*

(a)  $Z = [(X - \mu)/\sigma] \sim N(0, 1)$  – the standard normal distribution.

*Reproductive property:* If  $X_1, X_2, \dots, X_n$  are independent r.v.'s,  $X_i \sim N(\mu_i, \sigma_i^2)$ ,  $i = 1, 2, \dots, n$ , then  $(\sum_{i=1}^n X_i) \sim N(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$ .

### Appendix 6.1

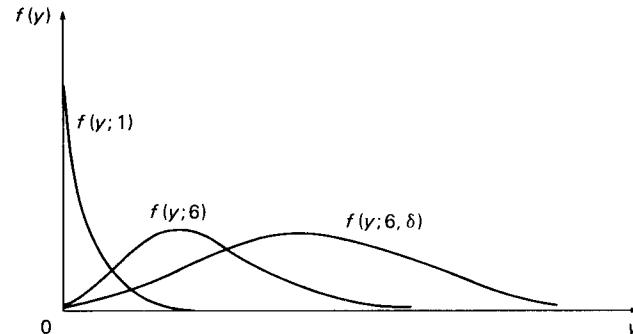


Fig. 6.9. The density functions of a central and non-central chi-square.

(2) *Chi-square distribution* –  $Y \sim \chi^2(n)$

$$f(y, n) = \frac{1}{2^{(n/2)} \Gamma(n/2)} y^{(n/2)-1} e^{-y/2}, \quad y > 0, \quad n = 1, 2, \dots$$

$$E(Y) = n \text{ (the degrees of freedom)}, \quad \text{Var}(Y) = 2n.$$

The density function is illustrated for several values of  $n$  in Fig. 6.9.

*Reproductive property*

If  $Y_1, Y_2, \dots, Y_k$  are independent r.v.'s  $Y_i \sim \chi^2(n_i)$ ,  $i = 1, 2, \dots, k$ , then  $(\sum_{i=1}^k Y_i) \sim \chi^2(n_1 + n_2 + \dots + n_k)$ .

(3) *Non-central chi-square distribution* –  $Y \sim \chi^2(n; \delta)$

$$f(y; \delta, n) = \frac{1}{\sqrt{\pi}} 2^{-(n/2)} \exp\left\{-\frac{1}{2}(y+\delta)\right\} y^{\lfloor(n/2)-1\rfloor} \times \sum_{k=0}^{\infty} \frac{(\delta y)^k \Gamma(k + \frac{1}{2})}{(2k)! \Gamma\left(k + \frac{n}{2}\right)},$$

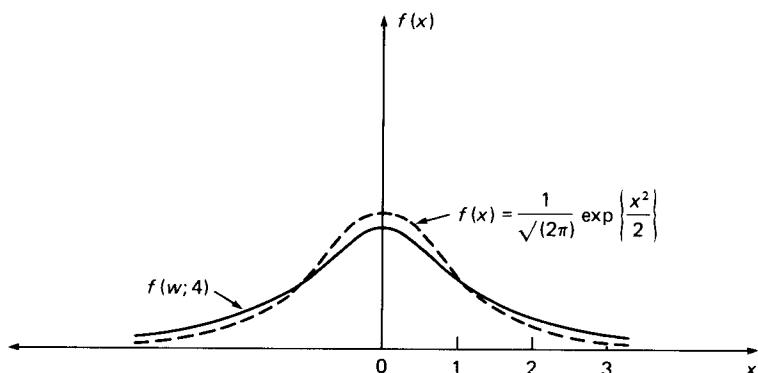
$$y > 0, \quad \delta > 0, \quad n = 1, 2, \dots$$

$$E(Y) = n + \delta, \quad \text{Var}(Y) = 2(n + 2\delta).$$

Hence, the important difference with the central chi-square is that the density function is shifted to the right and the variance increases.

*Reproductive property*

If  $Y_1, Y_2, \dots, Y_k$  are independent r.v.'s,  $Y_i \sim \chi^2(n_i; \delta_i)$ ,  $i = 1, 2, \dots, k$ , then  $(\sum_{i=1}^k Y_i) \sim \chi^2(n_1 + n_2 + \dots + n_k; \delta_1 + \delta_2 + \dots + \delta_k)$ .

Fig. 6.10. Comparison of a  $t$  and standard normal density.(4) Student's  $t$ -distribution -  $W \sim t(n)$ 

$$f(w; n) = \frac{1}{\sqrt{(n\pi)}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \frac{1}{\left(1 + \frac{w^2}{n}\right)^{[(n+1)/2]}}, \quad n > 0, \quad w \in \mathbb{R}.$$

$$E(W) = 0, \quad \text{Var}(W) = \frac{n}{n-2}, \quad n > 2, \quad \alpha_4 = 3 + \frac{6}{n-4}, \quad n \geq 4.$$

These moments show that for large  $n$  the  $t$ -distribution is very close to the normal (see Fig. 6.10).

(5) Non-central  $t$ -distribution -  $W \sim t(n; \delta), \delta > 0$ 

$$f(w; n, \delta) = \frac{n^{n/2}}{\sqrt{\pi} \Gamma(n/2)} \frac{e^{-(\delta^2/2)}}{(n+w^2)^{[(n+1)/2]}} \times \sum_{k=0}^{\infty} \Gamma\left(\frac{n+k+1}{2}\right) \left(\frac{\delta^k}{k!}\right) \left(\frac{2w^2}{n+w^2}\right)^{k/2}, \quad w \in \mathbb{R}.$$

For large  $n$

$$E(W) \approx \delta, \quad \text{Var}(W) \approx 1 + \frac{\delta^2}{2n}$$

## Appendix 6.1

(6) Fisher's  $F$ -distribution -  $U \sim F(n_1, n_2)$ 

$$f(u; n_1, n_2) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right) \left(\frac{n_1}{n_2}\right)^{n_1/2}}{\Gamma\left(\frac{n_1}{2}\right) \cdot \Gamma\left(\frac{n_2}{2}\right)} \frac{u^{\frac{1}{2}(n_1-2)}}{\left(1 + \frac{n_1}{n_2} u\right)^{\frac{1}{2}(n_1+n_2)}}, \quad u > 0.$$

$$E(U) = \frac{n_2}{n_2-2}, \quad n_2 > 2, \quad \text{Var}(U) = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}, \quad n_2 > 4.$$

The central and non-central  $F$ -distribution density functions are shown in Fig. 6.11 for purposes of comparison.

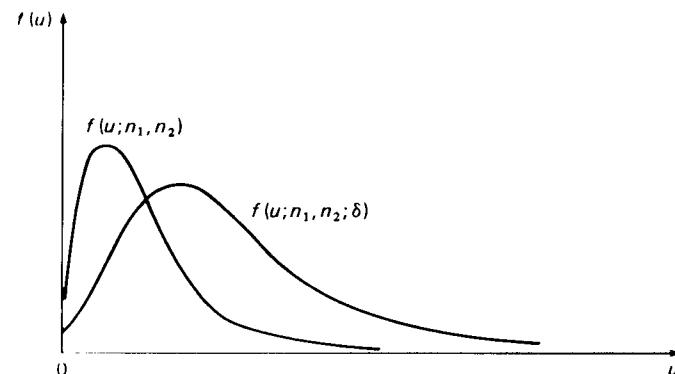
(7) Non-central  $F$ -distribution -  $U \sim F(n_1, n_2; \delta), \delta > 0$ 

$$f(u; n_1, n_2; \delta) =$$

$$e^{-\delta} \sum_{k=0}^{\infty} \frac{\delta^k u^{\frac{1}{2}(n_1+2k)-1} \Gamma\left(\frac{n_1+n_2+2k}{2}\right) \left(\frac{n_1}{n_2}\right)^{\frac{1}{2}(n_1+2k)}}{\left(1 + \frac{n_1}{n_2} u\right)^{\frac{1}{2}(n_1+n_2+2k)} \Gamma\left(\frac{n_2}{2}\right) \Gamma\left(\frac{n_1+2k}{2}\right)}, \quad u > 0,$$

$$E(U) = \frac{n_2(n_1+\delta)}{n_1(n_2-2)}, \quad n_2 > 2,$$

$$\text{Var}(U) = 2 \left(\frac{n_2}{n_1}\right)^2 \frac{(n_1+\delta)^2 + (n_1+2\delta)(n_2-2)}{(n_2-2)^2(n_2-4)}, \quad n_2 > 4.$$

Fig. 6.11. Central and non-central  $F$  density functions.

**Important concepts**

Borel functions, distribution of a Borel function of a r.v., normal and related distributions, Student's  $t$ , chi-square, Fisher's  $F$  and Cauchy distributions.

**Questions**

1. Why should we be interested in Borel functions of r.v.'s and their distributions?
2. 'A Borel function is nothing more than a r.v. relative to the Borel field  $\mathcal{B}$  on the real line.' Discuss.
3. Explain intuitively why a Borel function of a r.v. is a r.v. itself.
4. Explain the relationships between the normal, chi-square, Student's  $t$ , Fisher's  $F$  and Cauchy distributions.
5. What is the difference between central and non-central chi-square and  $F$ -distributions?

**Exercises**

1. Let  $X_1$  be a r.v. with density function

$$\begin{array}{ccccc} X_1 & -1 & 0 & 1 \\ f(x_1) & \frac{1}{4} & \frac{1}{2} & \frac{1}{2} \end{array}$$

Derive the density functions of

- (i)  $X = X_1^2$ ;
- (ii)  $X = e^{X_1}$ ;
- (iii)  $X = 10 + 2X_1^2$ .

2. Let the density function of the r.v.  $X$  be  $f(x) = e^{-x}$ ,  $x > 0$ . Find the distribution of  $Y = \log_e X$ .
3. Let the joint density function of  $X_1$  and  $X_2$  be

$X_1$		2	3	4
$X_2$				
1		$\frac{1}{18}$	$\frac{1}{9}$	$\frac{1}{18}$
2		$\frac{1}{18}$	$\frac{1}{9}$	$\frac{1}{9}$
3		0	$\frac{4}{9}$	$\frac{1}{18}$

Derive the distribution of

- (i)  $Y = X_1^2 + X_2^2$ ;
- (ii)  $Y = \min(X_1, X_2)$ .

4. Let  $X \sim U(0, 1)$ , derive the distribution of  $Y = X^2$ .

Clarke (1975); Cramer (1946); Giri (1974); Mood, Graybill and Boes (1974); Pfeiffer (1978); Rao (1973); Rohatgi (1976).

## CHAPTER 7

---

### The general notion of expectation

---

In Chapter 4 we considered the notion of mathematical expectation in the context of the simple probability model

$$\Phi = \{f(x; \theta), \theta \in \Theta\} \quad (7.1)$$

as a useful characteristic of density functions of a single random variable. Since then we generalised the probability model to

$$\Phi = \{f(x_1, x_2, \dots, x_n; \theta), \theta \in \Theta\} \quad (7.2)$$

and put forward a framework in the context of which joint density functions can be analysed. This included marginalisation, conditioning and functions of random variables (r.v.'s). The purpose of this section is to consider the notion of expectation in the context of this more general framework.

#### 7.1 Expectation of a function of random variables

In the one-dimensional case of a single r.v. we considered many numerical characteristics of density functions, namely,  $E(X)$ ,  $E(X')$ ,  $E(X - E(X))^2$ ,  $E(X - E(X))^r$ ,  $r = 2, 3, \dots$ , which contain summary information concerning the nature of the distribution of  $X$ . It is important to note that each of these characteristics is the expectation of some function of  $X$ , that is,  $E(h(X))$ ;  $h(X) = X$ ,  $h(X) = X'$ , etc. This provides us with the natural way to proceed in the present section, having discussed the idea of a Borel function  $h(\cdot)$ :  $\mathbb{R}^n \rightarrow \mathbb{R}$  which preserves the event structure of the Borel field  $\mathcal{B}$ . For simplicity of exposition we consider the case where  $n = 2$ .

Let  $(X_1, X_2)$  be a bivariate random vector with  $f(x_1, x_2)$  their joint density function and let  $h(\cdot)$ :  $\mathbb{R}^2 \rightarrow \mathbb{R}$  be a Borel function. Define  $Y = h(X_1, X_2)$  and consider its expectation. This can be defined in two

equivalent ways:

$$(i) \quad E(Y) = \int_{-\infty}^{\infty} y f_y(y) dy; \quad (7.3)$$

or

$$(ii) \quad E(h(X_1, X_2)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x_1, x_2) f(x_1, x_2) dx_1 dx_2. \quad (7.4)$$

The choice between (i) and (ii) is a matter of convenience and it is usually determined by the degree of difficulty in deriving the distribution of  $Y$ .

### *Example 1*

Let  $X_i \sim N(0, 1)$ ,  $i = 1, 2$  be two independent r.v.'s and  $h(X_1, X_2) = X_1^2 + X_2^2$ . Using (ii),

$$E(X_1^2 + X_2^2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_1^2 + x_2^2) \left[ \frac{1}{2\pi} \exp\left\{-\frac{1}{2}(x_1^2 + x_2^2)\right\} \right] dx_1 dx_2 = 2.$$

On the other hand,  $Y = (X_1^2 + X_2^2) \sim \chi^2(2)$  – chi-square with two degrees of freedom, that is,

$$f_y(y) = \frac{1}{2\Gamma(1)} \exp\left\{-\frac{1}{2}y\right\}$$

and we know that

$$E(Y) = \int_{-\infty}^{\infty} y f_y(y) dy = 2$$

equals the number of degrees of freedom (see Appendix 6.1).

Before we consider particular forms of  $h(X_1, X_2)$  let us consider some properties of  $E(h(X_1, X_2))$ .

### *Properties of expectation*

- (E1)  $E[ah_1(X_1, X_2) + bh_2(X_1, X_2)] = aE(h_1(X_1, X_2)) + bE(h_2(X_1, X_2))$ ; **linearity**, where  $a$  and  $b$  are constants and  $h_1(\cdot)$ ,  $h_2(\cdot)$  are Borel functions from  $\mathbb{R}^2$  to  $\mathbb{R}$ . In particular

$$E\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i E(X_i). \quad (7.5)$$

- (E2) If  $X_1$  and  $X_2$  are independent r.v.'s, for every Borel function  $h_1(\cdot)$

and  $h_2(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ ,

$$E(h_1(X_1)h_2(X_2)) = E(h_1(X_1)) \cdot E(h_2(X_2)), \quad (7.6)$$

given that the above expectations exist.

This is both a **necessary** as well as **sufficient** condition for independence. One particular case of interest is when  $h_1(X_1) = X_1$  and  $h_2(X_2) = X_2$ ,

$$E(X_1X_2) = E(X_1) \cdot E(X_2). \quad (7.7)$$

This is in some sense *linear independence* which is much weaker than independence. Moreover, given that

$$\text{Cov}(X_1, X_2) = E(X_1X_2) - E(X_1) \cdot E(X_2) \quad (7.8)$$

(see below), linear independence is equivalent to *uncorrelatedness* since it implies that  $\text{Cov}(X_1, X_2) = 0$ . A special case of linear independence of particular interest in what follows is when

$$E(X_1X_2) = 0, \quad (7.9)$$

and we say that  $X_1$  and  $X_2$  are *orthogonal*, writing  $X_1 \perp X_2$ . A case between independence and uncorrelatedness is that in which  $E(X_1/X_2) = E(X_1)$ , that is, the conditional and unconditional expectations of  $X_1$  coincide. The analogous case for orthogonality is

$$E(X_1/X_2) = 0, \quad \text{when } E(X_1) = 0. \quad (7.10)$$

This will prove to be a useful property in the context of limit theorems in Chapter 9, where the condition

$$E(X_n/X_{n-1}, X_{n-2}, \dots, X_1) = 0 \quad (7.11)$$

plays an important role. For reasons that will become apparent in the sequel we call this property *martingale orthogonality*.

### **Forms of $h(X_1, X_2)$ of particular interest**

Let

$$h(X_1, X_2) = X_1^l X_2^k, \quad l, k > 0,$$

then

$$\mu'_{lk} = E(X_1^l X_2^k) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1^l x_2^k f(x_1, x_2) dx_1 dx_2 \quad (7.12)$$

are called *joint raw moments* of order  $l+k$ ; this is a direct generalisation of the concept for one random variable. For

$$h(X_1, X_2) = (X_1 - E(X_1))^l (X_2 - E(X_2))^k,$$

let

$$\mu_1 \equiv E(X_1) \quad \text{and} \quad \mu_2 \equiv E(X_2).$$

$$\mu_{lk} \equiv E(h(X_1, X_2)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (X_1 - \mu_1)^l (X_2 - \mu_2)^k f(x_1, x_2) dx_1 dx_2 \quad (7.13)$$

are called *joint central moments* of order  $l+k$ . Two especially interesting joint central moments are the variance and covariance:

(i) *Covariance:  $l=k=1$*

$$\text{Cov}(X_1, X_2) = E((X_1 - \mu_1)(X_2 - \mu_2)). \quad (7.14)$$

The covariance provides a measure of the *linear relationship between two random variables*. With a direct multiplication the above formula becomes:

$$\text{Cov}(X_1, X_2) = E(X_1 X_2) - E(X_1) \cdot E(X_2). \quad (7.15)$$

If  $X_1$  and  $X_2$  are **independent** then using E2 we deduce that

$$\text{Cov}(X_1, X_2) = 0. \quad (7.16)$$

*It is very important to note that the converse is not true.*

(ii) *Variance:  $l=2, k=0$*

$$\text{Var}(X_1) \equiv E(X_1 - \mu_1)^2. \quad (7.17)$$

For a linear function  $\sum_i a_i X_i$  the variance is of the form

$$\text{Var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \text{Var}(X_i) + \sum_{i \neq j} a_i a_j \text{Cov}(X_i X_j), \quad (7.18)$$

where  $a_i, i=1, 2, \dots, n$  are real constants. In particular if  $X_1, \dots, X_n$  are independent

$$\text{Var}\left(\sum_i a_i X_i\right) = \sum_i a_i^2 \text{Var}(X_i). \quad (7.19)$$

Using the variance we could define the standardised form of a covariance known as the **correlation coefficient** and defined by

$$\text{Corr}(X_1, X_2) = \frac{\text{Cov}(X_1, X_2)}{\sqrt{[\text{Var}(X_1) \cdot \text{Var}(X_2)]}}. \quad (7.20)$$

**Properties of  $\text{Corr}(X_1, X_2)$** 

- (C1)  $\text{Corr}(X_1, X_2)=0$  for  $X_1$  and  $X_2$  independent r.v.'s.  
 (C2)  $-1 \leq \text{Corr}(X_1, X_2) \leq 1$ .  
 (C3)  $\text{Corr}(X_1, X_2)=1$  for  $X_1 = a+bX_2$ ,  $a, b$  being real constants.

**Example 2**

Let

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right),$$

i.e.  $(X_1, X_2)$  is a bivariate normal random vector, that is,

$$\begin{aligned} f(X_1, X_2) &= \frac{(1-\rho^2)^{-\frac{1}{2}}}{2\pi\sigma_1\sigma_2} \\ &\times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2\right.\right. \\ &\quad \left.\left.-2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\}. \end{aligned}$$

It was shown above that  $E(X_1)=\mu_1$ ,  $E(X_2)=\mu_2$ ,  $\text{Var}(X_1)=\sigma_1^2$ ,  $\text{Var}(X_2)=\sigma_2^2$ . Let us derive  $\text{Cov}(X_1, X_2)$ .

$$\begin{aligned} \text{Cov}(X_1, X_2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_1 - \mu_1)(x_2 - \mu_2) f(x_1, x_2) dx_1 dx_2 \\ &= \sigma_1\sigma_2 \int_{-\infty}^{\infty} \frac{(x_1 - \mu_1)}{\sigma_1} \frac{1}{\sqrt{(2\pi)}} \exp\left\{-\frac{1}{2}\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2\right\} dx_1 \\ &\quad \times \left[ \int_{-\infty}^{\infty} \frac{(1-\rho^2)^{-\frac{1}{2}}}{\sqrt{(2\pi)}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2\right.\right.\right. \\ &\quad \left.\left.\left.-2\rho\left(\frac{x_1 - \mu_1}{\sigma_1}\right)\left(\frac{x_2 - \mu_2}{\sigma_2}\right)\right]\right\} dx_2 = \sigma_1\sigma_2\rho. \right] \end{aligned}$$

Hence,  $\text{Corr}(X_1, X_2)=\rho$ . It should be noted that  $\rho=0$  in the case of normality does imply independence, as can be easily verified directly that  $\rho=0 \Rightarrow f(x_1, x_2)=f(x_1) \cdot f_2(x_2)$ .

*Note* that correlation measures the linear dependence between r.v.'s but if we consider only the first two moments that is the only dependence we can analyse.

## 7.2 Conditional expectation

The concept of conditional expectation plays a very important role both in extending the probability model  $\Phi = \{f(x_1, x_2, \dots, x_n; \theta), \theta \in \Theta\}$  to *time dependent* random variables (stochastic process) and in the specification of *linear models* in Part IV.

In Section 5.4 the *conditional distribution function* of  $X_1$  for  $X_2 = x_2$  was defined to be (if the limit exists)

$$F_{X_1/X_2}(x_1/x_2) = \lim_{0 < h \rightarrow 0} Pr(X_1 \leq x_1/x_2 - h \leq X_2 \leq x_2 + h). \quad (7.21)$$

The *conditional density*  $f_{X_1/X_2}(x_1/x_2)$  was defined by

$$F_{X_1/X_2}(x_1/x_2) = \int_{-\infty}^{x_1} f_{X_1/X_2}(u/x_2) du \Rightarrow f_{X_1/X_2}(x_1/x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}. \quad (7.22)$$

Note that

$$f(x_1, x_2) = f_{X_1/X_2}(x_1/x_2) \cdot f_2(x_2) = f_{X_2/X_1}(x_2/x_1) \cdot f_1(x_1), \quad (7.23)$$

$$f_{X_1/X_2}(x_1/x_2) = \frac{f_1(x_1) \cdot f_{X_2/X_1}(x_2/x_1)}{\int_{-\infty}^{\infty} f_1(x_1) \cdot f_{X_2/X_1}(x_2/x_1) dx_1},$$

– Bayes' formula, (7.24)

the denominator being equal to  $f_2(x_2)$ .

The conditional distribution function satisfies the following *properties*:

- (CD1) For a fixed  $x_2$ ,  $F_{X_1/X_2}(x_1/x_2)$  is a distribution function in  $X_1$ , i.e.  $F_{X_1/X_2}(\cdot): \mathbb{R} \rightarrow [0, 1]$ .
- (CD2) For a fixed  $x_1$ ,  $F_{X_1/X_2}(x_1/x_2)$  is a function of  $x_2$ .
- (CD3) For any values of  $x_1, x_2$

$$\int_{-\infty}^{x_2} F_{X_1/X_2}(x_1/u) f_2(u) du = Pr(X_1 \leq x_1, X_2 \leq x_2). \quad (7.25)$$

For a fixed  $x_2$ ,  $f_{X_1/X_2}(x_1/x_2)$  is a proper density function with  $f_{X_1/X_2}(x_1/x_2) \geq 0$ ,  $x_1 \in \mathbb{R}$  and  $\int_{-\infty}^{\infty} f_{X_1/X_2}(x_1/x_2) dx_1 = 1$ .

The *conditional expectation* of  $X_1$  given that  $X_2$  takes a particular value  $x_2$  ( $X_2 = x_2$ ) is defined by

$$E(X_1/X_2 = x_2) = \int_{-\infty}^{\infty} x_1 f_{X_1/X_2}(x_1/x_2) dx_1. \quad (7.26)$$

In general for any Borel function  $h(\cdot)$  whose expectation exists

$$E(h(X_1)/X_2=x_2) = \int_{-\infty}^{\infty} h(x_1) f_{X_1|X_2}(x_1/x_2) dx_1. \quad (7.27)$$

### Properties of the conditional expectation

Let  $X, X_1$  and  $X_2$  be random variables on  $(S, \mathcal{F}, P(\cdot))$ , then:

- (CE1)  $E(a_1 h(X_1) + a_2 h(X_2)/X=x) = a_1 E(h(X_1)/X=x) + a_2 E(h(X_2)/X=x)$ ,  $a_1, a_2$  constants.
- (CE2) If  $X_1 \geq X_2$ ,  $E(X_1/X=x) \geq E(X_2/X=x)$ .
- (CE3)  $E(h(X_1, X_2)/X_2=x_2) = E(h(X_1, x_2)/X_2=x_2)$ .
- (CE4)  $E(h(X_1)/X_2=x_2) = E(h(X_1))$  if  $X_1$  and  $X_2$  are independent.
- (CE5)  $E(h(X_1)) = E[E(h(X_1)/X_2=x_2)]$ , the expectation outside the square brackets being with respect to  $X_2$ .

The conditional expectation  $E(X_1/X_2=x_2)$  is a non-stochastic function of  $x_2$ , i.e.  $E(X_1/\cdot): \mathbb{R}_{X_2} \rightarrow \mathcal{R}$ . The graph  $(x_2, E(X_1/x_2))$  is called the *regression curve*.

### Example 3

In the case of the bivariate normal distribution,

$$\begin{aligned} & f_{X_1|X_2}(x_1/x_2) \\ &= \frac{f(x_1, x_2)}{f_2(x_2)} \\ &= \frac{(1-\rho^2)^{-\frac{1}{2}}}{2\pi\sigma_1\sigma_2} \exp\left\{-\frac{1}{2(1-\rho^2)} \left[ \left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 \right]\right\} \\ &\quad \left\{ \frac{1}{\sigma_2\sqrt{(2\pi)}} \exp\left\{-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right\} \right\} \\ &= \frac{(1-\rho^2)^{-\frac{1}{2}}}{\sigma_1\sqrt{(2\pi)}} \exp\left\{-\frac{1}{2(1-\rho^2)} \left( \frac{x_1-\mu_1 - \rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2)}{\sigma_1} \right)^2 \right\}, \\ & E(X_1/X_2=x_2) = \int_{-\infty}^{\infty} f_{X_1|X_2}(x_1/x_2) dx_1 = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2) \equiv g(x_2). \end{aligned}$$

The graph of this *linear regression function* is shown in Fig. 7.1.

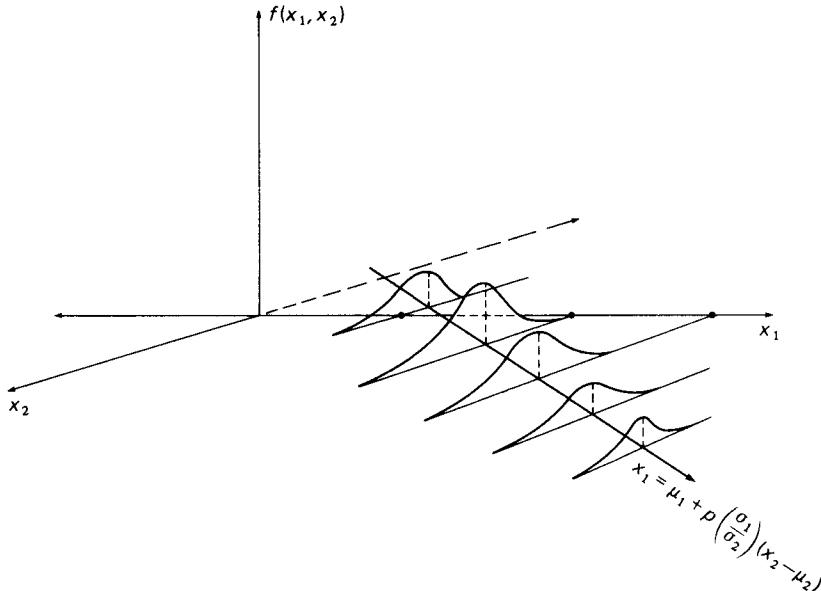


Fig. 7.1. The regression curve of a bivariate normal density.

As in the case of ordinary expectation, we can define higher conditional moments:

(i) *Raw conditional moments:*

$$E(X_1^r/X_2=x_2)=\int_{-\infty}^{\infty} X_1^r f_{X_1|X_2}(x_1/x_2) dx_1, \quad r \geq 1. \quad (7.28)$$

(ii) *Central conditional moments:*

$$E[(X_1 - E(X_1/X_2=x_2))^r/X_2=x_2], \quad r \geq 2. \quad (7.29)$$

Of particular interest is the conditional variance, sometimes called *skedasticity*:

$$\begin{aligned} \text{Var}(X_1/X_2=x_2) &= E[(X_1 - E(X_1/X_2=x_2))^2/X_2=x_2] \\ &= E(X_1^2/X_2=x_2) - [E(X_1/X_2=x_2)]^2. \end{aligned} \quad (7.30)$$

In the above example of the bivariate normal distribution we can show that  $\text{Var}(X_1/X_2=x_2)=\sigma_1^2(1-\rho^2)$ . In the case where the conditional variance is free of the conditioning variables it is said to be *homoskedastic*; otherwise it is called *heteroskedastic*.

*Example 4*

Let

$$F_{X_1/X_2}(x_1, x_2) = 1 - e^{-x_1} - e^{-x_2} + \exp\{-(x_1 + x_2 + \theta x_1 x_2)\},$$

$$x_1 > 0, \quad x_2 > 0, \quad \theta \in [0, 1].$$

This is the distribution function of *Gumbel's bivariate exponential distribution*

$$f(x_1, x_2) = [(1 + \theta x_1)(1 + \theta x_2) - \theta] \exp\{-(x_1 + x_2 + \theta x_1 x_2)\},$$

since

$$f_2(x_2) = e^{-x_2},$$

$$f_{X_1/X_2}(x_1/x_2) = [(1 + \theta x_1)(1 + \theta x_2) - \theta] \exp\{-x_1(1 + \theta x_2)\},$$

$$E(X_1/X_2 = x_2) = \int_0^\infty X_1 f_{X_1/X_2}(x_1/x_2) dx_1 = \frac{(1 + \theta + \theta x_2)}{(1 + \theta x_2)^2}$$

$$E(X_1^r/X_2 = x_2) = \frac{r! (1 + \theta x_2 + r\theta)}{(1 + \theta x_2)^{r+1}} \quad \text{and}$$

$$\text{Var}(X_1/X_2 = x_2) = \frac{(1 + \theta + \theta x_2)^2 - 2\theta^2}{(1 + \theta x_2)^4}$$

For a bivariate exponential distribution the regression curve is non-linear and the conditional variance is heteroskedastic.

*Example 5*

The joint density of a bivariate *Pareto distribution* takes the form

$$f(x_1, x_2) = \theta(\theta + 1)(a_1 a_2)^{\theta+1} (a_2 x_1 + a_1 x_2 - a_1 a_2)^{-(\theta+2)},$$

$$\theta > 0, \quad x_1 > a_1 > 0, \quad x_2 > a_2 > 0.$$

The marginal density function of  $X_2$  is

$$f_2(x_2) = \theta a_2^\theta x_2^{-(\theta+1)},$$

$$f_{X_1/X_2}(x_1/x_2) = a_2(\theta + 1)(a_1 x_2)^{\theta+1} [a_1 x_2 + a_2 x_1 - a_1 a_2]^{-(\theta+2)},$$

$$E(X_1/X_2 = x_2) = a_1 \left[ 1 + \frac{x_2}{a_2 \theta} \right]$$

$$\text{Var}(X_1/X_2 = x_2) = \left( \frac{a_1}{a_2} \right)^2 \frac{(\theta + 1)}{\theta^2(\theta + 1)} x_2^2$$

In the case of the Pareto distribution the regression curve is linear but the conditional variance is heteroskedastic.

*Example 6*

The joint density of a bivariate *logistic distribution* is

$$\begin{aligned}f(x_1, x_2) &= 2[1 + e^{-x_1} + e^{-x_2}]^{-3} \exp\{-(x_1 + x_2)\}, \quad x_1, x_2 > 0, \\E(X_1/X_2 = x_2) &= 1 - \log(1 + e^{-x_2}) \text{ -- non-linear in } x_2 \\\text{Var}(X_1/X_2 = x_2) &= \frac{1}{3}\pi^2 - 1 \text{ -- homoskedastic}\end{aligned}$$

As argued above  $E(X_1/X_2 = x_2)$  is a non-stochastic function of  $x_2$  and for a particular value  $\tilde{x}_2$ ,  $E(X_1/X_2 = \tilde{x}_2)$  is interpreted as the average value of  $X_1$  given  $X_2 = \tilde{x}_2$ . The temptation at this stage is to extend this concept by considering the conditional expectation

$$E(X_1/X_2). \tag{7.31}$$

The problem, however, is that we will be very hard pressed to explain its meaning. What does it mean to say ‘the average value of  $X_1$  given the random variable  $X_2(\cdot): S \rightarrow \mathbb{R}$ ’? The only meaning we can attach to such a conditional expectation is

$$E(X_1/\sigma(X_2)), \tag{7.32}$$

where  $\sigma(X_2)$  represents the  $\sigma$ -field generated by  $X_2$ , since what we condition on must be an event. Now, however,  $E(X_1/\sigma(X_2))$  is no longer a non-stochastic function, being evaluated at the r.v.  $X_2$ . Indeed,  $E(X_1/\sigma(X_2))$  is a random variable with respect to the  $\sigma$ -field  $\sigma(X_2) \subset \mathcal{F}$  which satisfies certain properties. The discerning reader would have noticed that we have already used this generalised concept of conditional expectation in stating CE5, where we took the expected value of the conditional expectation  $E(h(X_1)/X_2 = x_2)$ . What we had in mind there was  $E(h_1(X_1)/\sigma(X_2))$  since otherwise the conditional expectation is a constant. The way we defined  $E(X_1/\sigma(X_2))$  as a direct extension of  $E(X_1/X_2 = x_2)$ , one would hope that the similarity between the two concepts will not end there. It turns out, not surprisingly, that  $E(X_1/\sigma(X_2))$  satisfies certain properties analogous to CE1–CE5:

- (SCE1)  $E(a_1 h(X_1) + a_2 g(X_2)/\sigma(X)) = a_1 E(h(X_1)/\sigma(X)) + a_2 E(g(X_2)/\sigma(X))$ .
- (SCE2) If  $X_1 \geq X_2$ ,  $E(X_1/\sigma(X)) \geq E(X_2/\sigma(X))$ .
- (SCE3)  $E[h(X_1) \cdot g(X_2)] = E[g(X_2)E(h(X_1)/\sigma(X_2))]$ .
- (SCE4)  $E(h(X_1)/\sigma(X_2)) = E(h(X_1))$  if  $X_1$  and  $X_2$  are independent.
- (SCE5)  $E(h(X_1) \cdot g(X_2)/\sigma(X_2)) = g(X_2)E(h(X_1)/\sigma(X_2))$ .

This implies that the two conditional expectation concepts are directly related but one is non-stochastic and the other is a random variable.

*Example 7*

In the case of the bivariate normal distribution considered above,

$$E(X_1/\sigma(X_2)) = \mu_1 + \rho \frac{\sigma_1}{\sigma_2} (X_2 - \mu_2),$$

and it is a random variable. Note that  $\text{Var}(X_1/\sigma(X_2)) = \sigma_1^2(1 - \rho^2)$  is free of  $X_2$  (homoskedastic).

In relation to the conditional variance we can show that if  $E(X_1^2) < \infty$ , i.e. the variance exists,

$$\text{Var}(X_1) = \text{Var}(E(X_1/\sigma(X_2))) + E(\text{Var}(X_1/\sigma(X_2))), \quad (7.33)$$

that is, the variance of  $X_1$  can be decomposed into the variance of the conditional expectation of  $X_1$  plus the expectation of the conditional variance. This implies that

$$\text{Var}(X_1) \geq \text{Var}(E(X_1/\sigma(X_2))). \quad (7.34)$$

Note that some books write  $E(X_1/X_2)$  when they mean either  $E(X_1/X_2=x_2)$  or  $E(X_1/\sigma(X_2))$ , thus causing confusion.

The concept of conditional expectation can be extended further to an arbitrary  $\sigma$ -field  $\mathcal{D}$  where  $\mathcal{D}$  is some sub- $\sigma$ -field of  $\mathcal{F}$ , not necessarily generated by a random variable. This is possible because all elements of  $\mathcal{D}$  are events with respect to which the conditional expectation of a r.v.  $X$  relative to  $\mathcal{F}$  can be defined. If we were to interpret the conditional expectation  $E(X_1/X_2=x_2)$  as ‘smoothing’  $X_1$  to a constant we can think of  $E(X/\mathcal{D})$  as ‘smoothing’  $X$  to a random variable.

Because of the generality of  $E(X/\mathcal{D})$  we are going to summarise its properties (which include CE1–5 and SCE1–5 as special cases) for reference purposes:

Let  $X$  and  $Y$  be two r.v.’s defined on  $(S, \mathcal{F}, P(\cdot))$  and  $\mathcal{D} \subseteq \mathcal{F}$ .

- ( $\sigma$ -CE1) If  $c$  is a constant, then  $E(c/\mathcal{D}) = c$ .
- ( $\sigma$ -CE2) If  $X \leq Y$ , then  $E(X/\mathcal{D}) \leq E(Y/\mathcal{D})$ .
- ( $\sigma$ -CE3) If  $a, b$  are constants,  $E(aX + bY/\mathcal{D}) = aE(X/\mathcal{D}) + bE(Y/\mathcal{D})$ .
- ( $\sigma$ -CE4)  $|E(X/\mathcal{D})| \leq E(|X|/\mathcal{D})$ .
- ( $\sigma$ -CE5) If  $\mathcal{F}_0 = \{\emptyset, S\}$  then  $E(X/\mathcal{F}_0) = E(X)$ .
- ( $\sigma$ -CE6)  $E(X/\mathcal{F}) = X$ .
- ( $\sigma$ -CE7)  $E[E(X/\mathcal{D})] = E(X)$ .
- ( $\sigma$ -CE8) If  $\mathcal{D}_1 \subseteq \mathcal{D}_2$  ( $\mathcal{D}_i \subseteq \mathcal{F}, i=1, 2$ ),  $E[E(X/\mathcal{D}_2)/\mathcal{D}_1] = E[E(X/\mathcal{D}_1)/\mathcal{D}_2] = E(X/\mathcal{D}_1)$ .

- ( $\sigma$ -CE9) Let  $X$  be a r.v. relative to  $\mathcal{D}$  and  $E(|X|) < \infty$ ,  $E(|XY|) < \infty$  then  $E(XY/\mathcal{D}) = XE(Y/\mathcal{D})$ .
- ( $\sigma$ -CE10) If  $X$  is independent of  $\mathcal{D}$  then  $E(X/\mathcal{D}) = E(X)$ .

These properties hold when the various expectations used are defined and are always *relative to some equivalence class*. For this reason all the above statements related to conditional expectations must (formally) be qualified with the statement ‘almost surely’ (a.s.) (see Chapter 10). For example,  $\sigma$ -CE1 formally should read: ‘If  $c$  is a constant, i.e.  $X = c$  a.s. then  $E(X/\mathcal{D}) = c$  a.s.

The concept of conditional expectation will prove invaluable in the study of stochastic processes considered in Chapter 8 because it provides us with a natural way to formulate dependence.

### 7.3 Looking ahead

The concept of conditional expectation provides us with a very useful way to exploit auxiliary information or manipulate information (stochastic or otherwise) related to r.v.’s in the context of the probability model

$$\Phi = \{f(x_1, x_2, \dots, x_n; \theta), \theta \in \Theta\}. \quad (7.36)$$

Moreover, as argued in the next section, the concept of conditional expectation provides us with the most natural and direct link between sequences of independent r.v.’s discussed above and sequences of dependent r.v.’s of *stochastic processes* which enable us to extend  $\Phi$  to include more realistic *dynamic phenomena*. Stochastic processes are the subject of Chapter 8.

#### *Important concepts*

- expectation of a Borel function of a r.v.;
- linearity of the expectation operator;
- independence in terms of expectation, linear independence (uncorrelatedness);
- orthogonality, martingale orthogonality;
- joint raw and central moments, covariance, correlation;
- conditional distribution and density functions, Bayes’ formula;
- conditional expectations  $E(X_1/X_2=x_2)$ ,  $E(X_1/\sigma(X_2))$ ,  $E(X_1/\mathcal{D})$ ;
- regression curve, raw and central conditional moments, skedasticity, homoskedasticity, heteroskedasticity.

**Questions**

1. Explain the concept of expectation for a Borel function of a r.v.
2. Explain the concepts of independence and uncorrelatedness in the context of expectation of Borel functions of r.v.'s.
3. Compare the concepts of independence and martingale orthogonality.
4. What does the regression curve represent?
5. What does the correlation coefficient measure?
6. What does skedasticity measure?
7. Compare the concepts of  $E(X_1/X_2=x_2)$ ,  $E(X_1/\sigma(X_2))$  and  $E(X_1/\mathcal{D})$ , where  $\mathcal{D} \subset \mathcal{F}$ .
8. Explain intuitively the equality

$$\text{Var}(X_1) = \text{Var}(E(X_1/X_2)) + E(\text{Var}(X_1/X_2)).$$

**Exercises**

1. For exercise 3 of Chapter 6 show that  $E(X_1 X_2) = E(X_1)E(X_2)$  but  $X_1$  and  $X_2$  are not independent.
2. For exercise 3 of Chapter 6,
  - (i) derive  $E(X_1/X_2=x_2)$ ,  $\text{Var}(X_1/X_2=x_2)$ , for  $x_2=1, 2$ ,
  - (ii) find  $\text{Cov}(X_1, X_2)$ ,  $\text{Corr}(X_1, X_2)$ .
3. Let  $X_1$  and  $X_2$  be distributed as bivariate normal r.v.'s such that

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right), \quad \mu_1=2, \quad \mu_2=4, \quad \sigma_1^2=2 \\ \sigma_2^2=4, \quad \rho=0.2.$$

Calculate

$$E(X_1/X_2=x_2), \quad x_2=1, 2, 6,$$

$$E(X_2/X_1=x_1), \quad x_1=0, 1, 2,$$

$$\text{Var}(X_1/X_2=x_2), \quad \text{Var}(X_2/X_1=x_1).$$

4. Determine the value of  $c$  for which

$$f(x_1, x_2) = x_1(x_1 + x_2), \quad x_i \in (0, 1], \quad i = 1, 2$$

is a proper joint density function and derive:

- (i)  $E(X_1/X_2=x_2)$ ;
- (ii)  $\text{Var}(X_1/X_2=x_2)$ ;
- (iii)  $E(X_1^2 X_2/X_2=x_2)$ ;
- (iv)  $\Pr(X_1 + X_2 \leq 0.5)$ ;
- (v)  $\text{Corr}(X_1, X_2)$ .

Show that  $E(X_1) = E(E(X_1/X_2))$ .

**Additional references**

Bickel and Doksum (1977); Chung (1974); Clarke (1975); Giri (1974); Rao (1973); Whittle (1970).

## CHAPTER 8\*

---

### Stochastic processes

---

In Chapter 3 it was argued that the main aim of probability theory is to provide us with mathematical models, appropriately called probability models, which can be used as idealised descriptions of observable phenomena. The axiomatic approach to probability was viewed as based on a particular mathematical formulation of the idea of a random experiment in the form of the probability space  $(S, \mathcal{F}, P(\cdot))$ . The concept of a random variable introduced in Chapter 4 enabled us to introduce an isomorphic probability space  $(\mathbb{R}_x, \mathcal{B}, P_x(\cdot))$  which has a much richer (and easier) mathematical structure to help us build and analyse probability models. From the modelling viewpoint the concept of a random variable is particularly useful because most observable phenomena come in the form of quantifiable features amenable to numerical representation.

A particularly important aspect of real observable phenomena, which the random variable concept cannot accommodate, is their *time dimension*; the concept is essentially static. A number of the economic phenomena for which we need to formulate probability models come in the form of dynamic processes for which we have discrete sequence of observations in time. Observed data referring to economic variables such as inflation, national income, money stock, represent examples where the time dependency (dimension) might be very important as argued in Chapters 17 and 23 of Part IV. The problem we have to face is to extend the simple probability model,

$$\Phi = \{f(x; \theta), \theta \in \Theta\}, \quad (8.1)$$

to one which enables us to model *dynamic phenomena*. We have already moved in this direction by proposing the random vector probability model

$$\Phi = \{f(x_1, x_2, \dots, x_n; \theta), \theta \in \Theta\}. \quad (8.2)$$

The way we viewed this model so far has been as representing different characteristics of the phenomenon in question in the form of the jointly distributed r.v.'s  $X_1, \dots, X_n$ . If we reinterpret this model as representing the same characteristic but at successive points in time then this can be viewed as a dynamic probability model. With this as a starting point let us consider the dynamic probability model in the context of  $(S, \mathcal{F}, P(\cdot))$ .

### 8.1 The concept of a stochastic process

The natural way to make the concept of a random variable dynamic is to extend its domain by attaching a date to the elements of the sample space  $S$ .

#### *Definition 1*

*Let  $(S, \mathcal{F}, P(\cdot))$  be a probability space and  $\mathbb{T}$  an index set of real numbers and define the function  $X(\cdot, \cdot)$  by  $X(\cdot, \cdot): S \times \mathbb{T} \rightarrow \mathbb{R}$ . The ordered sequence of random variables  $\{X(\cdot, t), t \in \mathbb{T}\}$  is called a stochastic (random) process.*

This definition suggests that for a stochastic process  $\{X(\cdot, t), t \in \mathbb{T}\}$ , for each  $t \in \mathbb{T}$ ,  $X(\cdot, t)$  represents a random variable on  $S$ . On the other hand, for each  $s \in S$ ,  $X(s, \cdot)$  represents a function of  $t$  which we call a *realisation of the process*.  $X(s, t)$  for given  $s$  and  $t$  is just a real number.

#### *Example 1*

Consider the stochastic process  $\{X(\cdot, t), t \in \mathbb{T}\}$  defined by

$$X(s, t) = Y(s) \cos(Z(s)t + u(s)),$$

where  $Y(\cdot)$  and  $Z(\cdot)$  are two jointly distributed r.v.'s and  $u(\cdot) \sim U(-\pi, \pi)$ , independent of  $Y(\cdot)$  and  $Z(\cdot)$ . For a fixed  $t$ , say  $t = 1$ ,  $X(s) = Y(s) \cos(Z(s) + u(s))$ , being a function of r.v.'s, it is itself a r.v. For a fixed  $s$ ,  $Y(s) = y$ ,  $Z(s) = z$ ,  $u(s) = u$  are just three numbers and there is nothing stochastic about the function  $z(t) = y \cos(zt + u)$  being a simple cosine function of  $t$  (see Fig. 8.1(a)).

This example shows that for each  $t \in \mathbb{T}$  we have a different r.v. and for each  $s \in S$  we have a different realisation of the process. In practice we observe one realisation of the process and we need to postulate a dynamic probability model for which the observed realisation is considered to be one of a family of possible realisations. The original uncertainty of the outcome of an experiment is reduced to the uncertainty of the choice of one of these

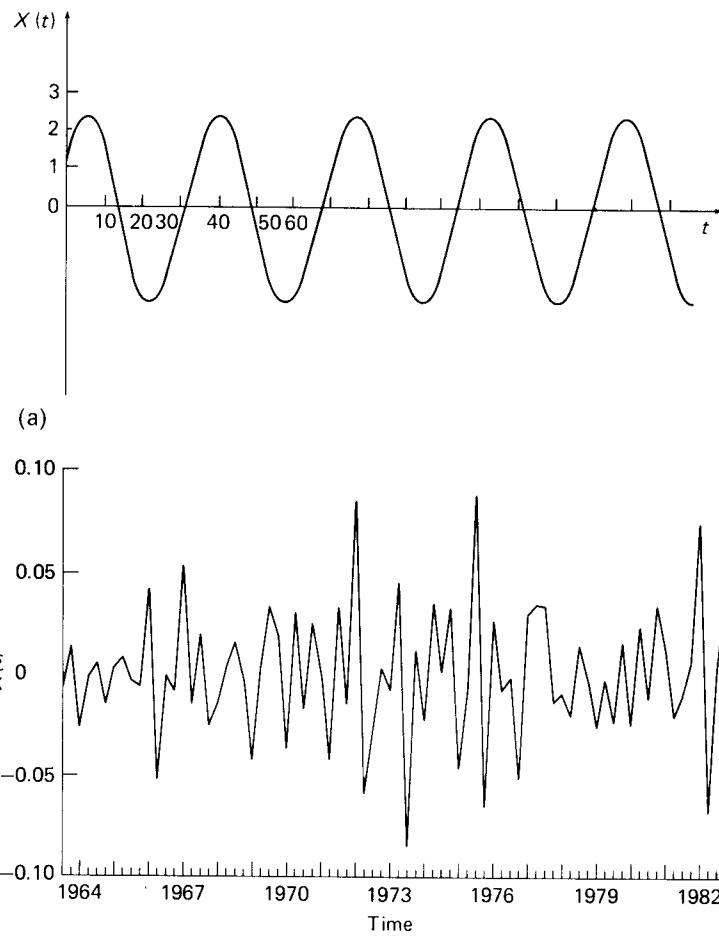


Fig. 8.1(a). The function  $z(t) = y \cos(zt + u)$ . (b) An economic time series.

possible realisations. Thus, there is nothing ‘random’ about a realisation of a process which can be smooth and regular as in the above example (see Fig. 8.1(a)) or have wild fluctuations like most econometric data series (see Fig. 8.1(b)).

- The main elements of a stochastic process  $\{X(\cdot, t), t \in \mathbb{T}\}$  are:
- its range space (sometimes called the state space), usually  $\mathbb{R}$ ;
  - the index set  $\mathbb{T}$ , usually one of  $\mathbb{R}, \mathbb{R}_+ = [0, \infty), \mathbb{Z} = \{\dots, -1, 0, 1, 2, \dots\}, \mathbb{Z}_+ = \{0, 1, 2, \dots\}$ ; and
  - the dependence structure of the r.v.’s  $X(t), t \in \mathbb{T}$ .

In what follows a stochastic process will be denoted by  $\{X(t), t \in \mathbb{T}\}$  ( $s$  is dropped) and its various interpretations as a random variable, a realisation or just a number should be obvious from the context used. The index set  $\mathbb{T}$  used will always be either  $\mathbb{T} = \{0, \pm 1, \pm 2, \dots\}$  or  $\mathbb{T} = \{0, 1, 2, \dots\}$ , thus concentrating exclusively on *discrete stochastic processes* (for continuous stochastic processes see Priestley (1981)).

The dependence structure of  $\{X(t), t \in \mathbb{T}\}$ , in direct analogy with the case of a random vector, should be determined by the joint distribution of the process. The question arises, however, ‘since  $\mathbb{T}$  is commonly an infinite set, do we need an infinite dimensional distribution to define the structure of the process?’ This question was tackled by Kolmogorov (1933) who showed that when the stochastic process satisfies certain regularity conditions the answer is definitely ‘no’. In particular, if we define the ‘tentative’ joint distribution of the process for the subset  $(t_1 < t_2 < t_3, \dots, < t_n)$  of  $\mathbb{T}$  by  $F(X(t_1), \dots, X(t_n)) = \Pr(X(t_1) \leq x_1, \dots, X(t_n) \leq x_n)$  then, if the stochastic process  $\{X(t), t \in \mathbb{T}\}$  satisfies the conditions:

- (i) *symmetry*:  $F(X(t_1), X(t_2), \dots, X(t_n)) = F(X(t_{j_1}), X(t_{j_2}), \dots, X(t_{j_n}))$  where  $j_1, j_2, \dots, j_n$  is any permutation of the indices  $1, 2, \dots, n$  (i.e. reshuffling the ordering of the index does not change the distribution);
- (ii) *compatibility*:  $\lim_{x_n \rightarrow \infty} F(X(t_1), \dots, X(t_n)) = F(X(t_1), \dots, X(t_{n-1}))$  (i.e. the dimensionality of the joint distribution can be reduced by marginalisation);

there exists a probability space  $(S, \mathcal{F}, P(\cdot))$  and a stochastic process  $\{X(t), t \in \mathbb{T}\}$  defined on it whose finite dimensional distribution is the distribution  $F(X(t_1), \dots, X(t_n))$ , as defined above. That is, the probabilistic structure of the stochastic process  $\{X(t), t \in \mathbb{T}\}$  is completely specified by the joint distribution  $F(X(t_1), \dots, X(t_n))$  for all values of  $n$  (a positive integer) and any subset  $(t_1, t_2, \dots, t_n)$  of  $\mathbb{T}$ . This is a remarkable result because it enables us to ‘describe’ the stochastic process without having to define an infinite dimensional distribution. In particular we can concentrate on the joint distribution of a finite collection of elements and thus extend the mathematical apparatus built for random vectors to analyse stochastic processes.

Given that, for a specific  $t$ ,  $X(t)$  is a random variable, we can denote its distribution and density functions by  $F(X(t))$  and  $f(X(t))$  respectively. Moreover, the mean, variance and higher moments of  $X(t)$  (as a r.v.) can be defined as in Section 4.6 by:

$$E(X(t)) = \mu(t), \quad (8.3)$$

$$E[(X(t) - \mu(t))^2] = v(t), \quad (8.4)$$

$$E(X(t)^r) = \mu_r(t), \quad r \geq 1, \quad t \in \mathbb{T}. \quad (8.5)$$

As we can see, these numerical characteristics of  $X(t)$  are in general functions of  $t$ , given that at each  $t \in \mathbb{T}$ ,  $X(\cdot, t)$  has a different distribution  $F(X(t))$ .

The compatibility condition (ii) enables us to extend the distribution function to any number of elements in  $\mathbb{T}$ , say  $t_1, t_2, \dots, t_n$ . That is,  $F(X(t_1), X(t_2), \dots, X(t_n))$  denotes the joint distribution of the same random variables  $X(t)$  at different points in  $\mathbb{T}$ . The question which naturally arises at this stage is ‘how is this joint distribution different from the joint distribution of the random vector  $\mathbf{X} \equiv (X_1, X_2, \dots, X_n)'$  where  $X_1, X_2, \dots, X_n$  are different random variables?’ The answer is not very different. The only real difference stems from the fact that the index set  $\mathbb{T}$  is now a *cardinal set*, the difference between  $t_i$  and  $t_j$  is now crucial, and it is not simply a labelling device as in the case of  $F(X_1, X_2, \dots, X_n)$ . This suggests that the mathematical apparatus developed in Chapters 5–7 for random vectors can be easily extended to the case of a stochastic process. For expositional purposes let us consider the joint distribution of the stochastic process  $\{X(t), t \in \mathbb{T}\}$  for  $t = t_1, t_2$ .

The joint distribution is defined by

$$F(X(t_1), X(t_2)) = \Pr(X(t_1) \leq x_1, X(t_2) \leq x_2), \quad x_1, x_2 \in \mathbb{R}. \quad (8.6)$$

The marginal and conditional distributions for  $X(t_1)$  and  $X(t_2)$  are defined in exactly the same way as in the case of a two-dimensional random vector (see Chapter 5). The various moments related to this joint distribution, however, take on a different meaning due to the importance of the cardinality of the index set  $\mathbb{T}$ . In particular the linear dependence measure

$$v(t_1, t_2) = E[(X(t_1) - \mu(t_1))(X(t_2) - \mu(t_2))], \quad t_1, t_2 \in \mathbb{T}. \quad (8.7)$$

is now called the *autocovariance* function. In standardised form

$$r(t_1, t_2) = \frac{v(t_1, t_2)}{(v(t_1)v(t_2))^{\frac{1}{2}}}, \quad t_1, t_2 \in \mathbb{T} \quad (8.8)$$

is called the *autocorrelation* function. Similarly, the *autoprod moment* is defined by  $m(t_1, t_2) = E(X(t_1)X(t_2))$ . These numerical characteristics of the stochastic process  $\{X(t), t \in \mathbb{T}\}$  play an important role in the analysis of the process and its application to the modelling of real observable phenomena. We say that  $\{X(t), t \in \mathbb{T}\}$  is an uncorrelated process if  $r(t_1, t_2) = 0$  for any  $t_1, t_2 \in \mathbb{T}$ ,  $t_1 \neq t_2$ . When  $m(t_1, t_2) = 0$  for any  $t_1, t_2 \in \mathbb{T}$ ,  $t_1 \neq t_2$  the process is said to be *orthogonal*.

### *Example 2*

One of the most important examples of a stochastic process is the *normal* (or

Gaussian) process. The stochastic process  $\{X(t), t \in \mathbb{T}\}$  is said to be normal if for any finite subset of  $\mathbb{T}$ , say  $t_1, t_2, \dots, t_n, (X(t_1), X(t_2), \dots, X(t_n)) \equiv \mathbf{X}_n(\mathbf{t})'$  has a multivariate normal distribution, i.e.

$$\begin{aligned} f(X(t_1), \dots, X(t_n)) \\ = \frac{(\det \mathbf{V}_n)^{-\frac{1}{2}}}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}(\mathbf{X}_n(\mathbf{t}) - \boldsymbol{\mu}(\mathbf{t}))' \mathbf{V}_n^{-1} (\mathbf{X}_n(\mathbf{t}) - \boldsymbol{\mu}(\mathbf{t}))\right\}, \end{aligned} \quad (8.9)$$

where  $\mathbf{V}_n \equiv [v(t_i, t_j)]$ ,  $i, j = 1, 2, \dots, n$  is an  $n \times n$  autocovariance matrix and  $\boldsymbol{\mu}(\mathbf{t}) \equiv (\mu(t_1), \mu(t_2), \dots, \mu(t_n))'$  is a  $n \times 1$  vector of means. In view of the compatibility condition we can deduce the marginal distribution of each  $X(t_i)$ , which is also normal,

$$f(X(t_i)) = \frac{v(t_i)^{-\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2v(t_i)} (X(t_i) - \mu(t_i))^2\right\}, \quad i = 1, 2, \dots, n. \quad (8.10)$$

As in the case of a normal random variable, the distribution of a normal stochastic process is characterised by the first two moments  $\mu(t)$  and  $v(t)$  but now they are both functions of  $t$ .

The concepts introduced above for the stochastic process  $\{X(t), t \in \mathbb{T}\}$  can be extended directly to a  $k \times 1$  vector stochastic process  $\{\mathbf{X}(t), t \in \mathbb{T}\}$  where  $\mathbf{X}(t) \equiv (X_1(t), X_2(t), \dots, X_k(t))'$ . Each component of  $\mathbf{X}(t)$  defines a stochastic process  $\{X_i(t), t \in \mathbb{T}\}$ ,  $i = 1, 2, \dots, k$ . This introduces a new dimension to the concept of a random vector because at each  $t$ , say  $t_1$ ,  $\mathbf{X}(t_1)$  is a  $k \times 1$  random vector and for  $t = t_1, t_2, \dots, t_n$ ,  $\mathcal{X} \equiv (\mathbf{X}(t_1)', \dots, \mathbf{X}(t_n)')$  defines a random  $n \times k$  matrix. The joint distribution of  $\mathcal{X}$  is defined by

$$F(\mathbf{X}(t_1), \mathbf{X}(t_2), \dots, \mathbf{X}(t_n)) = Pr(\mathbf{X}(t_1) \leq \mathbf{x}_1, \dots, \mathbf{X}(t_n) \leq \mathbf{x}_n) \quad (8.11)$$

with the marginal distributions  $F(\mathbf{X}(t_i)) = Pr(\mathbf{X}(t_i) \leq \mathbf{x}_i)$  being  $k$ -dimensional distribution functions. Most of the numerical characteristics introduced above can be extended to the vector stochastic process  $\{\mathbf{X}(t), t \in \mathbb{T}\}$  by a simple change in notation, say  $E(\mathbf{X}(t)) = \boldsymbol{\mu}(t)$ ,  $E[(\mathbf{X}(t) - \boldsymbol{\mu}(t))(\mathbf{X}(t) - \boldsymbol{\mu}(t))'] = \mathbf{V}(t)$ ,  $t \in \mathbb{T}$ , but we also need to introduce new concepts to describe the relationship between  $X_i(t)$  and  $X_j(\tau)$  where  $i \neq j$  and  $t, \tau \in \mathbb{T}$ . Hence, we define the cross-covariance and cross-correlation functions by

$$c_{ij}(t, \tau) = E[(X_i(t) - \mu_i(t))(X_j(\tau) - \mu_j(\tau))] \quad (8.12)$$

and

$$r_{ij}(t, \tau) = \frac{c_{ij}(t, \tau)}{(v_i(t) \cdot v_j(\tau))^{\frac{1}{2}}}, \quad i, j = 1, 2, \dots, n, \quad t, \tau \in \mathbb{T}. \quad (8.13)$$

Note that  $c_{ij}(t, \tau) = v(t, \tau)$  and  $r_{ij}(t, \tau) = r(t, \tau)$  for  $i=j$ . These concepts measure the linear dependence between the stochastic processes  $\{X_i(t), t \in \mathbb{T}\}$  and  $\{X_j(\tau), \tau \in \mathbb{T}\}$ . Similarly, we define the *cross-product moment* function by  $m_{ij}(t, \tau) = E(X_i(t), X_j(\tau)) = m(t, \tau)$  when  $i=j$ . Note that  $v(t, \tau) = m(t, \tau) - \mu(t)\mu(\tau) = r(t, \tau)[v(t)v(\tau)]^{\frac{1}{2}}$ . Using the notation introduced in Chapter 6 (see also Chapter 15) we can denote the distribution of a normal random matrix  $\mathcal{X}$  by

$$\begin{pmatrix} \mathbf{X}(t_1) \\ \mathbf{X}(t_2) \\ \vdots \\ \mathbf{X}(t_n) \end{pmatrix} \sim N \left( \begin{pmatrix} \boldsymbol{\mu}(t_1) \\ \boldsymbol{\mu}(t_2) \\ \vdots \\ \boldsymbol{\mu}(t_n) \end{pmatrix}, \begin{pmatrix} \mathbf{V}(t_1)\mathbf{C}(t_1, t_2), & \dots, & \mathbf{C}(t_1, t_n) \\ \mathbf{C}(t_2, t_1)\mathbf{V}(t_2) & \ddots & \vdots \\ \vdots & \dots & \mathbf{V}(t_n) \end{pmatrix} \right), \quad (8.14)$$

where  $\mathbf{V}(t_i)$  and  $\mathbf{C}(t_i, t_j)$  are  $k \times k$  matrices of autocorrelations and cross-correlations, respectively. The formula of the distribution of  $\mathcal{X}$  needs special notation which is rather complicated to introduce at this stage.

In defining the above concepts we (implicitly) assumed that the various moments used are well defined (bounded) for all  $t \in \mathbb{T}$ , which is not generally true. When the moments of  $\{X(t), t \in \mathbb{T}\}$  are bounded for all  $t \in \mathbb{T}$  up to order  $l$ , i.e.

$$E(|X(t)|^l) = \int_{-\infty}^{\infty} |X|^l f(X(t)) dx < \infty, \quad \text{for all } t \in \mathbb{T}, \quad (8.15)$$

we say that the *process is of order l*. In defining the above concepts we assumed implicitly that the stochastic processes involved are at least of order 2.

The definition of a stochastic process given above is much too general to enable us to obtain a manageable (operational) probability model for modelling dynamic phenomena. In order to see this let us consider the question of constructing a probability model using the normal process. The natural way to proceed is to define the parametric family of densities  $f(X(t); \boldsymbol{\theta}_t)$  which is now indexed not by  $\boldsymbol{\theta}$  alone but  $t$  as well, i.e.

$$\Phi = \{f(X(t); \boldsymbol{\theta}_t), \boldsymbol{\theta}_t \in \Theta_t, t \in \mathbb{T}\}, \quad (8.16)$$

If  $f(X(t); \boldsymbol{\theta}_t)$  is the normal density  $\boldsymbol{\theta}_t \equiv (\mu(t), V(t,t))$  and  $\Theta_t \equiv \mathbb{R} \times \mathbb{R}_+$ . The fact that the unknown parameters of the stochastic process  $\{X(t), t \in \mathbb{T}\}$  change with  $t$  (such parameters are sometimes called *incidental*) presents us with a difficult problem. The problem is that in the case where we only have a single realisation of the process (the usual case in econometrics) we will

have to deduce the values of  $\mu(t)$  and  $V(t, t)$  with the help of a single observation! This arises because, as argued above for each  $t$ ,  $X(s, t)$  is a random variable with its own distribution.

The main purpose of the next three sections is to consider various special forms of stochastic processes where we can construct probability models which are manageable in the context of statistical inference. Such manageability is achieved by imposing certain restrictions which enable us to reduce the number of unknown parameters involved in order to be able to deduce their values from a single realisation. These restrictions come in two forms:

- (i)      restrictions on the *time-heterogeneity* of the process; and
- (ii)     restrictions on the *memory* of the process.

In Section 8.2 the concept of stationarity inducing considerable time-homogeneity to a stochastic process is considered. Section 8.3 considers various concepts which restrict the memory of a stochastic process in different ways. These restrictions will play an important role in Chapters 22 and 23. The purpose of Section 8.4 is to consider briefly a number of important stochastic processes which are used extensively in Part IV. These include martingales, martingale differences, innovation processes, Markov processes, Brownian motion process, white-noise, autoregressive (AR) and moving average (MA) processes as well as ARMA and ARIMA processes.

## 8.2 Restricting the time-heterogeneity of a stochastic process

For an arbitrary stochastic process  $\{X(t), t \in \mathbb{T}\}$  the distribution function  $F(X(t); \theta_t)$  depends on  $t$  with the parameters  $\theta_t$  characterising it being functions of  $t$  as well. That is, a stochastic process is time-heterogeneous in general. This, however, raises very difficult issues in modelling real phenomena because usually we only have one observation for each  $t$ . Hence, in practice we will have to 'estimate'  $\theta_t$  on the basis of a single observation, which is impossible. For this reason we are going to consider an important class of stationary processes which exhibit considerable time-homogeneity and can be used to model phenomena approaching their *equilibrium steady-state*, but continuously undergoing 'random' fluctuations. This is the class of stationary stochastic processes.

### *Definition 2*

A stochastic process  $\{X(t), t \in \mathbb{T}\}$  is said to be **(strictly) stationary** if for any subset  $(t_1, t_2, \dots, t_n)$  of  $\mathbb{T}$  and some  $\tau$ ,

$$F(X(t_1), \dots, X(t_n)) = F(X(t_1 + \tau), \dots, X(t_n + \tau)). \quad (8.17)$$

That is, the distribution function of the process remains unchanged when shifted in time by an arbitrary value  $\tau$ . In terms of the marginal distributions  $F(X(t))$ ,  $t \in \mathbb{T}$  stationarity implies that

$$F(X(t)) = F(X(t + \tau)), \quad (8.18)$$

and hence  $F(X(t_1)) = F(X(t_2)) = \dots = F(X(t_n))$ . That is, stationarity implies that  $X(t_1), \dots, X(t_n)$  are (individually) identically distributed (ID); a perfect time-homogeneity. As far as the joint distribution is concerned, stationarity implies that it does not depend on the date of the first time index  $t_1$ .

This concept of stationarity, although very useful in the context of probability theory, is very difficult to verify in practice because it is defined in terms of the distribution function. For this reason the concept of  $l$ th-order stationarity, defined in terms of the first  $l$  moments, is commonly preferred.

### *Definition 3*

A stochastic process  $\{X(t), t \in \mathbb{T}\}$  is said to be  **$l$ th-order stationary** if for any subset  $(t_1, t_2, \dots, t_n)$  of  $\mathbb{T}$  and any  $\tau$ ,  $F(X(t_1, \dots, X(t_n)))$  is of **order  $l$**  and its joint moments are equal to the corresponding moments of  $F(X(t_1 + \tau), \dots, X(t_n + \tau))$ , i.e.

$$\begin{aligned} E[\{X(t_1)\}^{l_1} \{X(t_2)\}^{l_2}, \dots, \{X(t_n)\}^{l_n}] \\ = E[\{X(t_1 + \tau)\}^{l_1}, \dots, \{X(t_n + \tau)\}^{l_n}], \end{aligned} \quad (8.19)$$

where  $l_1 + l_2 + \dots + l_n \leq l$ ; see Priestley (1981).

In order to understand this definition let us take  $l=1$  and  $l=2$ .

#### (1)      *First-order stationarity*

$\{X(t), t \in \mathbb{T}\}$  is said to be first order stationary if  $E(|X(t)|) < \infty$  for all  $t \in \mathbb{T}$  and for  $l_1 = 1$ ,  $E(X(t)) = E(X(t + \tau)) = \mu$ , constant free of  $t$ .

#### (2)      *Second-order stationarity*

$\{X(t), t \in \mathbb{T}\}$  is said to be second order stationary if  $E(|X(t)|^2) < \infty$  for all  $t \in \mathbb{T}$  and

$$(l_1 = 1, l_2 = 0) \quad (i) \quad E[X(t)] = E[X(t + \tau)] = \mu_1, \quad \text{constant free of } t,$$

$$(l_1 = 2, l_2 = 0) \quad (ii) \quad E[\{X(t)\}^2] = E[\{X(t + \tau)\}^2] = \mu'_2, \quad \text{constant free of } t$$

and

$$(l_1 = 1, l_2 = 1) \quad (\text{iii}) \quad E[\{X(t_1)\}\{X(t_2)\}] \\ = E[\{X(t_1 + \tau)\}\{X(t_2 + \tau)\}].$$

Taking  $\tau = -t_1$  we can deduce that

$$E[\{X(0)\}\{X(t_2 - t_1)\}] = h(t_2 - t_1), \quad \text{a function of } |t_2 - t_1|.$$

These suggest that second-order stationarity for  $X(t)$  implies that its mean and variance ( $\text{Var}(X(t)) = \mu'_2 - \mu_1^2$ ) are constant and free of  $t$  and its autocovariance ( $v(t_1, t_2) = E[\{X(0)\}\{X(t_2 - t_1)\}] - \mu_1^2$ ) depends on the interval  $|t_2 - t_1|$ ; not  $t_1$  or/and  $t_2$ . Second-order stationarity, which is also called *weak or wide-sense stationarity*, is by far the most important form of stationarity in modelling real phenomena. This is partly due to the fact that in the case of a normal stationary process second-order stationarity is equivalent to strict stationarity given that the first two moments characterise the normal distribution.

In order to see how stationarity can help us define operational probability models for modelling dynamic phenomena let us consider the implications of assuming stationarity for the normal stochastic process  $\{X(t), t \in \mathbb{T}\}$  and its parameters  $\theta_t$ . Given that  $E(X(t)) = \mu$  and  $\text{Var}(X(t)) = \sigma^2$  for all  $t \in \mathbb{T}$  and  $v(t_1, t_2) = v(|t_1 - t_2|)$  for any  $t_1, t_2 \in \mathbb{T}$  we can deduce that for the subset  $(t_1, t_2, \dots, t_n)$  of  $\mathbb{T}$  the joint distribution of the process is characterised by the parameters

$$\theta^* \equiv (\mu, \sigma^2, v(|t_i - t_j|), i, j = 1, 2, \dots, n, i \neq j), \\ \text{a } (n+1) \times 1 \text{ vector.} \quad (8.20)$$

This is to be contrasted with the non-stationary case where the parameter vector is  $\theta = (\mu(t_i), v(t_i, t_j), i, j = 1, 2, \dots, n)$ , a  $(n+n^2) \times 1$  vector. A sizeable reduction in the number of the unknown parameters. It is important, however, to note that even in the case of stationarity the number of parameters increases with the size of the subset  $(t_1, t_2, \dots, t_n)$  although the parameters do not depend on  $t \in \mathbb{T}$ . This is because time-homogeneity does not restrict the ‘memory’ of the process. The dependence between  $X(t_1)$  and  $X(t_2)$  is restricted only to be a function of the distance  $|t_1 - t_2|$  but the function itself is not restricted in any way. For example  $h(\cdot)$  can take forms such as:

$$(a) \quad h(|t_1 - t_2|) = (t_1 - t_2)^2. \quad (8.21)$$

$$(b) \quad h(|t_1 - t_2|) = \exp\{-|t_1 - t_2|\}. \quad (8.22)$$

In case (a) the dependence between  $X(t_1)$  and  $X(t_2)$  increases as the gap between  $t_1$  and  $t_2$  increases and in case (b) the dependence decreases as

$|t_1 - t_2|$  increases. In terms of the ‘memory’ of the process these two cases are very different indeed but from the stationarity viewpoint they are identical (second-order stationary process autocovariance functions). In the next section we are going to consider ‘memory’ restrictions in an obvious attempt to ‘solve’ the problem of the parameters increasing with the size of the subset  $(t_1, t_2, \dots, t_n)$  of  $\mathbb{T}$ .

Before we consider memory restrictions, however, it is important to comment on the notion of a non-stationary stochastic process as the absence of time-homogeneity. Stationarity, in time-series analysis, plays a similar role to linearity in mathematics; every function which is not linear is said to be non-linear. A non-stationary stochastic process in the present context is said to be a process which exhibits time-heterogeneity. In terms of actual observed realisations, the assumption of stationarity is considered appropriate for the underlying stochastic process, when a  $\tau$ -period ( $\tau > 1$ ) window, wide enough to include the width of the realisation, placed directly over the time graph of the realisation and sliced over it along the time axis, shows ‘the same picture’ in its frame; no systematic variation in the picture (see Fig. 8.1(b)). Non-stationarity will be an appropriate assumption for the underlying stochastic process when the picture shown by the window as sliced along the time axis changes ‘systematically’, such as the presence of a trend or a monotonic change in the variance. An important form of non-stationarity is the so-called homogeneous non-stationarity which is described as local time dependence of the mean of the process only (see ARIMA( $p, q$ ) formulation below).

### 8.3      Restricting the memory of a stochastic process

In the case of a typical economic time series, viewed as a particular realisation of a stochastic process  $\{X(t), t \in \mathbb{T}\}$  one would expect that the dependence between  $X(t_1)$  and  $X(t_2)$  would tend to weaken as the distance  $(t_2 - t_1)$  increases. For example, if  $X(t)$  refers to the GNP in the UK at time  $t$  one would expect that dependence between  $X(t_1)$  and  $X(t_2)$  to be much greater when  $t_1 = 1984$  and  $t_2 = 1985$  than when  $t_1 = 1952$  and  $t_2 = 1985$ . Formally, this dependence can be described in terms of the joint distribution  $F(X(t_1), X(t_2), \dots, X(t_n))$  as follows:

*Definition 4*

A stochastic process  $\{X(t), t \in \mathbb{T}\}$  defined on the probability space  $(S, \mathcal{F}, P(\cdot))$  is said to be **asymptotically independent** if for any subset

$(t_1, t_2, \dots, t_n)$  of  $\mathbb{T}$  and any  $\tau, \beta(\tau)$  defined by

$$\begin{aligned} & |F(X(t_1), \dots, X(t_n), X(t_1 + \tau), \dots, X(t_n + \tau)) - F(X(t_1), \\ & \dots, X(t_n))F(X(t_1 + \tau), \dots, X(t_n + \tau))| \leq \beta(\tau) \end{aligned} \quad (8.23)$$

goes to zero as  $\tau \rightarrow \infty$ .

Let us consider the intuition underlying this definition of asymptotic independence. Any two events  $A$  and  $B$  in  $\mathcal{F}$  are said to be independent when  $P(A \cap B) = P(A) \cdot P(B)$  or equivalently  $P(A \cap B) - P(A)P(B) = 0$ . Using this notion of independence in terms of the distribution function of two random variables (r.v.'s)  $X_1$  and  $X_2$  we can view  $|F(X_1, X_2) - F(X_1)F(X_2)|$  as a measure of dependence between the two r.v.'s. In the above definition of asymptotic independence  $\beta(\tau)$  provides an upper bound for such a measure of dependence in the case of a stochastic process. If  $\beta(\tau) \rightarrow 0$  as  $\tau \rightarrow \infty$  the two subsets  $(X(t_1), \dots, X(t_n))$  and  $(X(t_1 + \tau), \dots, X(t_n + \tau))$  become independent.

A particular case of asymptotic independence is that of *m-dependence* which restricts  $\beta(\tau)$  to be zero for all  $\tau > m$ . That is,  $X(t_1)$  and  $X(t_2)$  are independent for  $|t_1 - t_2| > m$ . In practice we would expect to be able to find a 'large enough'  $m$  so as to be able to approximate any asymptotically independent process by an *m*-dependent process. This is equivalent to assuming that  $\beta(\tau)$  for  $\tau > m$  are so small as to be able to equate them to zero.

An alternative way to express the weakening of the dependence between  $X(t_1)$  and  $X(t_2)$  as  $|t_1 - t_2|$  increases is in terms of the autocorrelation function which is a measure of linear dependence (see Chapter 7).

#### Definition 5

A stochastic process  $\{X(t), t \in \mathbb{T}\}$  is said to be **asymptotically uncorrelated** if there exists a sequence of constants  $\{\rho(\tau), \tau \geq 1\}$  defined by

$$\left| \frac{v(t, t + \tau)}{(v(t)v(t + \tau))^2} \right| \leq \rho(\tau), \quad \text{for all } t \in \mathbb{T},$$

such that

$$0 \leq \rho(\tau) \leq 1 \quad \text{and} \quad \sum_{\tau=0}^{\infty} \rho(\tau) < \infty. \quad (8.24)$$

As we can see, the sequence of constants  $\{\rho(\tau), \tau \geq 1\}$  defines an upper bound for the sequence of autocorrelation coefficients  $r(t, t + \tau)$ . Moreover, given that  $\rho(\tau) \rightarrow 0$  as  $\tau \rightarrow \infty$  is a necessary and  $\rho(\tau) < \tau^{-(1+\delta)}$  for  $\delta > 0$ , a sufficient condition for  $\sum_{\tau=1}^{\infty} \rho(\tau) < \infty$  (see White (1984)), the intuition underlying the above definition is obvious.

In the case of a normal stochastic process the notions of asymptotic independence and uncorrelatedness coincide because the dependence between  $X(t_1)$  and  $X(t_2)$  for any  $t_1, t_2 \in \mathbb{T}$  is completely determined by the autocorrelation function  $r(t_1, t_2)$ . This will play a very important role in Part IV (see Chapters 22 and 23) where the notion of a stationary, asymptotically independent normal process is used extensively. At this stage it is important to note that the above concepts of asymptotic independence and uncorrelatedness which restrict the memory of a stochastic process are not defined in terms of a stationary stochastic process but a general time-heterogeneous process. This is the reason why  $\beta(\tau)$  and  $\rho(\tau)$  for  $\tau \geq 1$  define only upper bounds for the two measures of dependence given that when equality is used in their definition they will depend on  $(t_1, t_2, \dots, t_n)$  as well as  $\tau$ .

A more general formulation of asymptotic independence can be achieved using the concept of a  $\sigma$ -field generated by a random vector (see Chapters 4 and 7). Let  $\mathcal{B}_1^t$  denote the  $\sigma$ -field generated by  $X(1), \dots, X(t)$  where  $\{X(t), t \in \mathbb{T}\}$  is a stochastic process. A measure of the dependence among the elements of the stochastic process can be defined in terms of the events  $B \in \mathcal{B}_{-\infty}^t$  and  $A \in \mathcal{B}_{t+\tau}^\infty$  by

$$\alpha(\tau) = \sup_{\tau} |P(A \cap B) - P(A)P(B)|. \quad (8.25)$$

#### *Definition 6*

A stochastic process  $\{X(t), t \in \mathbb{T}\}$  is said to be **(strongly) mixing** if  $\alpha(\tau) \rightarrow 0$  as  $\tau \rightarrow \infty$ .

As we can see, this is a direct generalisation of the asymptotic independence concept which is defined in terms of particular events  $A$  and  $B$  related to the definition of the joint distribution function. In the case where  $\{X(t), t \in \mathbb{T}\}$  is an *independent* process  $\alpha(\tau) = 0$  for  $\tau \geq 1$ . Another interesting special case defined above of a mixing process is the *m-dependent process* where  $\alpha(\tau) = 0$  for  $\tau > m$ . In this sense an independent process is a zero-dependent process. The usefulness of the concept of an *m-dependent process* stems from the fact that commonly in practice any asymptotically independent (or mixing) process can be approximated by such a process for ‘large enough’  $m$ .

A stronger form of mixing, sometimes called *uniform mixing*, can be defined in terms of the following measure of dependence:

$$\varphi(\tau) = \sup_{\tau} |P(A|B) - P(A)|, \quad P(B) > 0. \quad (8.26)$$

*Definition 7*

A stochastic process  $\{X(t), t \in \mathbb{T}\}$  is said to be **uniformly mixing** if  $\varphi(\tau) \rightarrow 0$  as  $\tau \rightarrow \infty$ .

Looking at the two definitions of mixing we can see that  $\alpha(\tau)$  and  $\varphi(\tau)$  define absolute and relative measures of temporal dependence, respectively. The former is based on the definition of dependence between two events  $A$  and  $B$  separated by  $\tau$  periods using the absolute measure

$$[P(A \cap B) - P(A) \cdot P(B)] \geq 0$$

and the latter the relative measure

$$[P(A|B) - P(A)] \geq 0.$$

In the context of second-order stationary stochastic processes asymptotic uncorrelatedness can be defined more intuitively in terms of the temporal covariance as follows:

$$\text{Cov}(X(t), X(t + \tau)) = v(\tau) \rightarrow 0 \quad \text{as } \tau \rightarrow \infty. \quad (8.27)$$

A weaker form of such memory restriction is the so-called ergodicity property. Ergodicity can be viewed as a condition which ensures that the memory of the process as measured by  $v(\tau)$  ‘weakens by averaging over time’.

*Definition 8*

A second-order stationary process  $\{X(t), t \in \mathbb{T}\}$  is said to be **ergodic** if

$$\lim_{T \rightarrow \infty} \left( \frac{1}{T} \sum_{\tau=1}^T v(\tau) \right) = 0. \quad (8.28)$$

If we compare (28) with (25) we can deduce that in the case of a second-order stationary process strong mixing implies ergodicity. The weaker form of temporal dependence in (28), however, is achieved at the expense of a very restrictive form of time-homogeneity (stationarity). In modelling we need both type of restrictions and there is often a trade off between them (see Domowitz and White (1982)).

Memory restrictions enable us to model the temporal dependence of a stochastic process using a finite set of parameters in the form of temporal moments or some parametric process (see Section 4). This is necessary in order to enable us to construct operational probability models for modelling dynamic phenomena. The same time-heterogeneity and memory restrictions enable us to derive asymptotic results which are crucial for

statistical inference purposes. For example one of the most attractive features of mixing processes is that any Borel function of them is also mixing. This implies that the limit theorems for mixing processes (see Section 9.4) can be used to derive asymptotic results for estimators and test statistics which are functions of the process. The intuition underlying these results is that because of stationarity the restriction on the memory enables us to argue that the observed realisation of the process is typical (in a certain sense) of the underlying stochastic process and thus the time averages constitute reliable estimates of the corresponding probability expectations.

## 8.4      Some special stochastic processes

The purpose of this section is to consider briefly several special stochastic processes which play an important role in econometric modelling (see Part IV). These stochastic processes will be divided into parametric and non-parametric processes. The *non-parametric processes* are defined in terms of their joint distribution functions or the first few joint moments. On the other hand, *parametric processes* are defined in terms of a generating mechanism which is commonly a functional form based on a non-parametric process.

### (I)      Non-parametric processes

The concept of conditional expectation discussed in Chapter 7 provides us with an ideal link between the theory of random variables discussed in Chapters 4–7 and that of stochastic processes, the subject matter of the present chapter. This is because the notion of conditional expectation enables us to formalise the temporal dependence in a stochastic process  $\{X(t), t \in \mathbb{T}\}$  in terms of the conditional expectation of the process at time  $t$ ,  $X(t)$  ('the present') given  $(X(t-1), X(t-2), \dots)$  ('the past'). One important application of conditional expectation in such a context is in connection with a stochastic process which forms a martingale.

#### (1)      *Martingales*

##### *Definition 9*

Let  $\{X(t), t \in \mathbb{T}\}$  be a stochastic process defined on  $(S, \mathcal{F}, P(\cdot))$  and  $\{\mathcal{G}_t, t \in \mathbb{T}\}$  an increasing sequence of  $\sigma$ -fields  $\{\mathcal{G}_t \subset \mathcal{F}, t \in \mathbb{T}\}$  satisfying the following conditions:

- (i)  $X(t)$  is a random variable (r.v.) relative to  $\mathcal{D}_t$  for all  $t \in \mathbb{T}$ .
- (ii)  $E(|X(t)|) < \infty$  (i.e. its mean is bounded) for all  $t \in \mathbb{T}$ ; and
- (iii)  $E(X(t)/\mathcal{D}_{t-1}) = X(t-1)$ , for all  $t \in \mathbb{T}$ .

Then  $\{X(t), t \in \mathbb{T}\}$  is said to be a **martingale** with respect to  $\{\mathcal{D}_t, t \in \mathbb{T}\}$  and we write  $\{X(t), \mathcal{D}_t, t \in \mathbb{T}\}$ .

Several aspects of this definition need commenting on. Firstly, a martingale is a relative concept; a stochastic process relative to an increasing sequence of  $\sigma$ -fields. That is,  $\sigma$ -fields such that  $\mathcal{D}_1 \subset \mathcal{D}_2 \subset \mathcal{D}_3, \dots, \subset \mathcal{D}_t \subset \dots$  and each  $X(t)$  is a r.v. relative to  $\mathcal{D}_t$ ,  $t \in \mathbb{T}$ . A natural choice for such  $\sigma$ -fields will be  $\mathcal{D}_t = \sigma(X(t), X(t-1), \dots, X(1))$ ,  $t \in \mathbb{T}$ . Secondly, the expected value of  $X(t)$  must be bounded for all  $t \in \mathbb{T}$ . This, however, implies that the stochastic process has *constant mean* because  $E(X(t)) = E[E(X(t))/\mathcal{D}_{t-1}] = E(X(t-1))$  for all  $t \in \mathbb{T}$  by property  $\sigma$ -CE7 of conditional expectations (see Section 7.2). Thirdly, (iii) implies that

$$E(X(t+\tau)/\mathcal{D}_{t-1}) = X(t-1) \quad \text{for all } t \in \mathbb{T} \text{ and } \tau \geq 0. \quad (8.29)$$

That is, the best predictor of  $X(t+\tau)$ , given the information  $\mathcal{D}_{t-1}$ , is  $X(t-1)$  for any  $\tau \geq 0$ .

Intuitively, a martingale can be viewed as a ‘fair game’. Defining  $X(t)$  to be the money held by a gambler after the  $t$ th trial in a casino game (say, black jack) and  $\mathcal{D}_t$  to be the ‘history’ of the game up to time  $t$ , then the condition (iii) above suggests that the game is ‘fair’ because the gambler before trial  $t$  expects to have the same amount of money at the end of the trial as the amount held before the bet was placed. It will take a very foolish gambler to play a game for which

$$(iii)' \quad E(X(t)/\mathcal{D}_{t-1}) \leq X(t-1). \quad (8.30)$$

This last condition defines what is called a *supermartingale* (‘super’ for the casino?).

The importance of martingales stems from the fact that they are general enough to include most forms of stochastic processes of interest in econometric modelling as special cases, and restrictive enough so as to allow the various ‘limit theorems’ (see Chapter 9) needed for their statistical analysis to go through, thus making probability models based on martingales ‘largely’ operational. In order to appreciate their generality let us consider two extreme examples of martingales.

### *Example 3*

Let  $\{Z(t), t \in \mathbb{T}\}$  be a sequence of *independent r.v.’s* such that  $E(Z(t)) = 0$  for

all  $t \in \mathbb{T}$ . If we define  $X(t)$  by

$$X(t) = \sum_{k=1}^t Z(k), \quad (8.31)$$

then  $\{X(t), \mathcal{D}_t, t \in \mathbb{T}\}$  is a martingale, with  $\mathcal{D}_t = \sigma(Z(t), Z(t-1), \dots, Z(1)) = \sigma(X(t), X(t-1), \dots, X(1))$ . This is because conditions (i) and (ii) are automatically satisfied and we can verify that

$$E(X(t)/\mathcal{D}_{t-1}) = E[(X(t-1) + Z(t))/\mathcal{D}_{t-1}] = X(t-1), \quad t \in \mathbb{T}, \quad (8.32)$$

using the properties  $\sigma$ -CE9 and 10 in Section 7.2.

#### *Example 4*

Let  $\{Z(t), t \in \mathbb{T}\}$  be an arbitrary stochastic process whose only restriction is that  $E(|Z(t)|) < \infty$  for all  $t \in \mathbb{T}$ . If we define  $X(t)$  by

$$X(t) = \sum_{k=1}^t [Z(k) - E(Z(k)/\mathcal{D}_{k-1})], \quad (8.33)$$

where  $\mathcal{D}_k = \sigma(Z(k), Z(k-1), \dots, Z(1)) = \sigma(X(k), X(k-1), \dots, X(1))$ , then  $\{X(t), \mathcal{D}_t, t \in \mathbb{T}\}$  is a martingale. Note that condition (iii) can be verified using the property  $\sigma$ -CE8 (see Section 7.2).

The above two extreme examples illustrate the flexibility of martingales very well. As we can see, the main difference between them is that in example 3,  $X(t)$  was defined as a linear function of independent r.v.'s and in example 4 as a linear function of dependent r.v.'s centred at their conditional means, i.e.

$$Y(t) = X(t) - E(Z(t)/\mathcal{D}_{t-1}), \quad t \in \mathbb{T}. \quad (8.34)$$

It can be easily verified that  $\{Y(t), t \in \mathbb{T}\}$  defines what is known as a *martingale difference process* relative to  $\mathcal{D}_t$  because

$$E(Y(t)/\mathcal{D}_{t-1}) = 0, \quad \text{for all } t \in \mathbb{T}. \quad (8.35)$$

In the case where  $E(|Z(t)|^2) < \infty$  for all  $t \in \mathbb{T}$  we can deduce that for  $t > k$

$$\begin{aligned} E(Y(t)Y(k)) &= E[E(Y(t)Y(k))/\mathcal{D}_{t-1}] \\ &= E[Y(k)E(Y(t)/\mathcal{D}_{t-1})] = 0 \end{aligned} \quad (8.36)$$

That is,  $\{Y(t), t \in \mathbb{T}\}$  is an *orthogonal sequence* as well (see Chapter 7).

*Definition 10*

A stochastic process  $\{Y(t), t \in \mathbb{T}\}$  is said to be a **martingale difference** process relative to the increasing sequence of  $\sigma$ -fields  $\mathcal{D}_1 \subset \mathcal{D}_2 \subset \dots \subset \mathcal{D}_t \subset \dots$  if

- (i)  $Y(t)$  is a r.v. relative to  $\mathcal{D}_t$ ;
- (ii)  $E(|Y(t)|) < \infty$ ; and
- (iii)  $E(Y(t)/\mathcal{D}_{t-1}) = 0, t \in \mathbb{T}$ .

*Definition 11*

A stochastic process  $\{Y(t), t \in \mathbb{T}\}$  is said to be an **innovation process** if it is a martingale difference with respect to  $\mathcal{D}_t = \sigma(X(t), X(t-1), \dots, X(0))$ , where  $X(t) = \sum_{j=0}^t Y(j)$ , and

- (i)  $E(|Y(t)|^2) < \infty$ ;
- (ii)  $E(Y(t)Y(\tau)) = 0, t > \tau, t, \tau \in \mathbb{T}$ .

These special processes related to martingales will play a very important role in the statistical models of interest in econometrics in Part IV. Returning to the main difference between the two examples above we can see the independence assumption in a sequence is ‘largely’ equivalent to that of orthogonality in the context of martingales. It will be shown in Chapter 9 that as far as the various limit theorem results are concerned this ‘crude equivalence’ carries over in that context as well. The ‘law of large numbers’ and the ‘central limit theorem’ results for sequences of independent r.v.’s can be extended directly to orthogonal martingale difference processes.

Martingales are particularly important in specifying statistical models of interest in econometrics (see Part IV) because they enable us to decompose any stochastic process  $\{X(t), t \in \mathbb{T}\}$  whose mean is bounded for all  $t \in \mathbb{T}$  into two orthogonal components,  $\mu(t)$  and  $u(t)$ , called the systematic and non-systematic components, respectively, such that

$$X(t) = \mu(t) + u(t), \quad (8.37)$$

where  $\mu(t) = E(X(t)/\mathcal{D}_{t-1})$  and  $u(t) = X(t) - E(X(t)/\mathcal{D}_{t-1})$ , for some  $\sigma$ -field  $\mathcal{D}_t$  defining our sample information set. The non-systematic component  $u(t)$  defines a martingale difference and thus all the limit theorem results needed for the statistical analysis of such a specification are readily available.

In view of the discussion in the last two sections on time-homogeneity and memory restrictions, the question which arises naturally is to what extent martingales assume any of these restrictions. As shown above, martingales are first-order stationary because

$$E(X(t)) = E(X(t-1)) = \mu, \quad \text{for all } t \in \mathbb{T}. \quad (8.38)$$

Moreover, their conditional memory is restricted enough to allow us to define a martingale difference sequence with any martingale. That is, if  $\{X(t), t \in \mathbb{T}\}$  is a martingale, then we can define the process

$$Y(t) = X(t) - X(t-1), \quad Y(0) = X(0), \quad t \in \mathbb{T}, \quad (8.39)$$

which is a martingale difference and  $X(t) = \sum_{j=0}^t Y(j)$ . In the case where the martingale is also a second-order process then  $\{Y(t), t \in \mathbb{T}\}$  is also an innovation process. In Chapter 9 it will be shown that an innovation process behaves asymptotically like an independent sequence of random variables; the most extreme form of memory restriction.

## (2)      *Markov processes*

Another important class of stochastic processes is that of Markov processes. These processes are based on the so-called Markov property that ‘the future’ of the process, given the ‘present’, is independent of the ‘past’.

*Definition 12*

A stochastic process  $\{X(t), t \in \mathbb{T}\}$  is said to be a **Markov process** if for every Borel function  $h(X(t)) \in \mathcal{B}_t^\infty$  ('the future') such that

$$E|h(X(t))| < \infty,$$

$$E(h(X(t))/\mathcal{B}_{-\infty}^t) = E(h(X(t))/\sigma(X(t-1))), \quad (8.40)$$

where  $\mathcal{B}_z^b = \sigma(X(t): \alpha < t < b)$ . (Note:  $\mathcal{B}_{-\infty}^t$  is ‘past’ plus ‘present’.)

In particular, in the case where  $h(X(t)) = X(t+\tau)$ , the Markov property suggest that

$$E(X(t+\tau)/\mathcal{B}_{-\infty}^t) = E(X(t+\tau)/\sigma(X(t))), \quad \tau > 0. \quad (8.41)$$

An alternative but equivalent way to express the Markov property is to define the events  $B \in \mathcal{B}_{-\infty}^{t-1}$ ,  $A \in \mathcal{B}_{t+1}^\infty$  and state that

$$P(A \cap B/\mathcal{B}_t^t) = P(A/\mathcal{B}_t^t) \cdot P(B/\mathcal{B}_t^t). \quad (8.42)$$

It is important to note that the Markov property is not a direct restriction on the memory of the process. It is a restriction on the conditional memory of the process. For ‘predicting’ the future of the process the present provides all the relevant information.

A natural extension of the Markov process is to allow the relevant information to be  $m$  periods into the past.

*Definition 13*

A stochastic process  $\{X(t), t \in \mathbb{T}\}$  is said to be ***mth-order Markov*** if  $E(|X(t)|) < \infty$  and

$$E(X(t)/\mathcal{B}_{-\infty}^{t-1}) = E(X(t)/\sigma(X(t-1), \dots, X(t-m))). \quad (8.43)$$

In terms of this definition a Markov process is a first-order Markov. The *mth-order Markov* property suggests that for predicting  $X(t)$  only the ‘recent past’ of the process is relevant.

In practice a Markov process is commonly supplemented with direct restrictions on its memory and time-heterogeneity. In particular, if an *mth-order Markov* process is also assumed to be normal, stationary and asymptotically independent we can deduce that

$$E(X(t)/\mathcal{B}_{-\infty}^{t-1}) = \alpha_1 X(t-1) + \alpha_2 X(t-2) + \dots + \alpha_m X(t-m), \quad (8.44)$$

and the roots of the polynomial  $(\lambda^m - \alpha_1 \lambda^{m-1} - \dots - \alpha_m) = 0$  lie inside the unit circle (see AR(*m*) below). This special stochastic process will play a very important role in Part IV.

### (3) Brownian motion process

A particular form of a Markov process with a long history in physics is the so-called Brownian motion (or Wiener) process.

*Definition 14*

A stochastic process  $\{X(t), t \in \mathbb{T}\}$  is called a **Brownian motion process**, defined on  $(S, \mathcal{F}, P(\cdot))$  if

- (a)  $X(t) = 0$ , for  $t = 0$  (the process starts at 0; a convention);
- (b)  $X(t)$  is a process with stationary independent increments, i.e. for  $0 \leq t_1 \leq t_2 \leq \dots \leq t_n$ , the increments  $(X(t_i) - X(t_{i-1}))$ ,  $i = 1, 2, \dots, n$ , are independent r.v.’s, such that

$$E(X(t_i) - X(t_{i-1})) = 0,$$

$$V(t_i, t_{i-1}) = \sigma^2(t_i - t_{i-1});$$

- (c) the increments  $X(t_i) - X(t_{i-1})$ ,  $i = 1, 2, \dots, n$ , are normally distributed. This implies that the density function is

$$\begin{aligned} f(x_1, \dots, x_n; t_1, \dots, t_n) \\ = \left( \frac{1}{\sigma \sqrt{(2\pi)t_1}} \exp\left(-\frac{x_1^2}{2\sigma^2 t_1}\right) \right) \prod_{i=2}^n \frac{(t_i - t_{i-1})^{-\frac{1}{2}}}{\sigma \sqrt{(2\pi)}} \end{aligned}$$

$$\times \exp \left\{ -\frac{(x_i - x_{i-1})^2}{2\sigma^2(t_i - t_{i-1})} \right\}.$$

In the case where  $\sigma^2 = 1$  the process is called standard Brownian motion. It is not very difficult to see that the standard Brownian motion process is both a martingale as well as a Markov process. That is,  $\{X(t), t \in \mathbb{T}\}$  is a martingale with respect to  $\mathcal{B}_{-\infty}^t$  since  $E(X(t)/\mathcal{B}_{-\infty}^t) = X(\tau), \tau \leq t$ . Moreover, since  $E(X(t)/\mathcal{B}_{-\infty}^t) = E(X(t)/\sigma(X(\tau)))$  it is also a Markov process. Note also that  $E[(X(t) - X(\tau))^2/\mathcal{B}_{-\infty}^t] = (t - \tau), t \leq \tau$ .

### *Definition 15*

A stochastic process  $\{u(t), t \in \mathbb{T}\}$  is said to be a **white-noise process** if

- (i)       $E(u(t)) = 0;$
- (ii)      $E(u(t)u(\tau)) = \begin{cases} \sigma^2 & \text{if } t = \tau \\ 0 & \text{if } t \neq \tau \text{ (uncorrelated r.v.'s).} \end{cases}$

Hence, a white-noise process is both time-homogeneous, in view of the fact that it is a second-order stationary process, and has no memory. In the case where  $\{u(t), t \in \mathbb{T}\}$  is also assumed to be normal the process is also strictly stationary.

Despite its simplicity (or because of it) the concept of a white-noise process plays a very important role in the context of parametric time-series models to be considered next, as a basic building block.

## (II)    Parametric stochastic processes

The main difference between the type of stochastic processes considered so far and the ones to be considered in this sub-section is that the latter are defined in terms of a generating mechanism; they are in some sense ‘derived’ stochastic processes.

### (4)    *Autoregressive, first order (AR(1))*

The AR(1) process is by far the most widely used stochastic process in econometric modelling. An adequate understanding of this process will provide the necessary groundwork for more general parametric models such as AR( $m$ ), MA( $m$ ) or ARMA( $p,q$ ) considered next.

### *Definition 16*

A stochastic process  $\{x(t), t \in \mathbb{T}\}$  is said to be **autoregressive of**

**order one (AR(1))** if it satisfies the stochastic difference equation,

$$X(t) = \alpha X(t-1) + u(t), \quad (8.45)$$

where  $\alpha$  is a constant and  $u(t)$  is a **white-noise process**.

The main difference between this definition and the non-parametric definitions given above is that the processes  $\{X(t), t \in \mathbb{T}\}$  are now defined indirectly via the generating mechanism GM (45). This suggests that the properties of this process depend crucially on the properties of  $u(t)$  and the structure of the difference equation. The properties of  $u(t)$  have already been discussed and thus we need to consider how the structure of the GM as a non-homogeneous difference equation (see Miller (1968)) determines the properties of  $\{X(t), t \in \mathbb{T}\}$ ,  $\mathbb{T} = \{0, 1, 2, \dots\}$ .

Viewing (45) as a stochastic non-homogeneous first-order difference equation we can express it (by repeated substitution) in the form

$$X(t) = \alpha^t X(0) + \sum_{i=0}^{t-1} \alpha^i u(t-i). \quad (8.46)$$

This expresses  $X(t)$  as a linear function of the white-noise process  $\{u(t), t \in \mathbb{T}\}$  and  $X(0)$ . Using this form we can deduce certain properties of the AR(1) such as time-homogeneity or/and memory. In particular, from (46) we can deduce that

$$E(X(t)) = \alpha^t E(X(0)) \quad (8.47)$$

and

$$\begin{aligned} & E(X(t)X(t+\tau)) \\ &= E \left\{ \left( \alpha^t X(0) + \sum_{i=0}^{t-1} \alpha^i u(t-i) \right) \left( \alpha^{t+\tau} X(0) + \sum_{i=0}^{t+\tau-1} \alpha^i u(t+\tau-i) \right) \right\} \\ &= E\{\alpha^{2t+\tau} X(0)^2\} + E \left\{ \left( \sum_{i=0}^{t-1} \alpha^i u(t-i) \right) \left( \sum_{i=0}^{t+\tau-1} \alpha^i u(t+\tau-i) \right) \right\} \\ &= \alpha^{2t+\tau} E(X(0)^2) + \sigma^2 (\alpha^\tau + \alpha^{\tau+2} + \dots + \alpha^{\tau+2(t-1)}), \quad \tau \geq 0. \end{aligned} \quad (8.48)$$

This shows clearly that if no restrictions are placed on  $X(0)$  or/and  $\alpha$  the AR(1) process represented by (45) is neither stationary nor asymptotically independent. Let us consider restricting  $X(0)$  and  $\alpha$ .

$X(0)$  in (46) represents the initial condition for the solution of the stochastic difference equation. This seems to provide a ‘unique’ solution for the difference equation (see Luenberger (1969)) and plays an important role in the determination of the dependence structure of stochastic difference

equations. If we assume that  $X(0)=0$  for simplicity, (47) and (48) take the form

$$\begin{aligned} E(X(t)) &= 0, \\ E(X(t)X(t+\tau)) &= \begin{cases} \sigma^2 t, & |\alpha|=1 \\ \sigma^2 \alpha^2 \left( \frac{1-\alpha^{2t}}{1-\alpha^2} \right), & |\alpha| \neq 1, \quad \tau \geq 0. \end{cases} \end{aligned} \quad (8.50)$$

As we can see, assuming  $X(0)=0$  does not make the stochastic process  $\{X(t), t \in \mathbb{T}\}$  as generated by (45), stationary or asymptotically independent. The latter is achieved by supplementing the above initial condition by the coefficient restriction  $|\alpha|<1$ . Assuming that  $X(0)=0$  and  $|\alpha|<1$  we can deduce that

$$E(X(t)X(t+\tau)) = \sigma^2 \alpha^2 \left( \frac{1-\alpha^{2\tau}}{1-\alpha^2} \right) \rightarrow 0 \quad \text{as } \tau \rightarrow \infty, \quad (8.51)$$

and thus  $\{X(t), t \in \mathbb{T}\}$  is indeed asymptotically independent (but not stationary). For  $|\alpha| \geq 1$  the process is neither asymptotically independent nor stationary.

For the stochastic process  $\{X(t), t \in \mathbb{T}\}$ , as generated by (45), to be stationary we need to change the index set  $\mathbb{T}$  to a double infinite set  $\mathbb{T}^*=\{0, \pm 1, \pm 2, \dots\}$ . That is, assume that the process  $X(t)$  stretches back to the infinite remote past (a convenient fiction!). The stochastic process  $\{X(t), t \in \mathbb{T}^*\}$  with GM (45) can be expressed in the form:

$$X(t) = \sum_{i=0}^{\infty} \alpha^i u(t-i) = \sum_{j=-\infty}^t \alpha^{t-j} u(j), \quad t \in \mathbb{T}^*, \quad (8.52)$$

assuming  $X(-T) \rightarrow 0$  as  $T \rightarrow \infty$ . From (52) we can deduce that

$$\begin{aligned} E(X(0)) &= 0, \\ E(X(t)X(t+\tau)) &= E \left\{ \left( \sum_{i=0}^{\infty} \alpha^i u(t-i) \right) \left( \sum_{j=0}^{\infty} \alpha^j u(t+\tau-j) \right) \right\} \\ &= \sigma^2 \left( \sum_{i=0}^{\infty} \alpha^i \alpha^{i+\tau} \right) = \sigma^2 \alpha^\tau \left( \sum_{i=0}^{\infty} \alpha^{2i} \right), \quad \tau \geq 0. \end{aligned} \quad (8.54)$$

Hence, for  $|\alpha|<1$  the stochastic process  $\{X(t), t \in \mathbb{T}^*\}$  is both second-order stationary and asymptotically independent with autocovariance function

$$v(\tau) = \frac{\sigma^2}{(1-\alpha^2)} \alpha^\tau, \quad \tau \geq 0. \quad (8.55)$$

On the other hand, for  $|\alpha| \geq 1$ ,  $(\sum_{i=0}^{\infty} \alpha^{2i}) \rightarrow \infty$  and the process is not even second order since  $E(|X(t)|^2)$  is not bounded.

The main purpose of the above discussion has been to bring out the importance of seemingly innocuous assumptions such as the initial conditions ( $X(0)=0$ ,  $X(-T)\rightarrow 0$  as  $T\rightarrow\infty$ ), the choice of the index set ( $\mathbb{T}$  or  $\mathbb{T}^*$ ) and the parameter restrictions ( $|\alpha|\leq 1$ ), as well as the role of these assumptions in determining the properties of the stochastic process. As seen above, the choice of the index set and the initial condition play a very important role in determining the stationarity of the process. In particular, for the stochastic process as defined by the GM (45) to be stationary it is necessary to use  $\mathbb{T}^*$  as the index set. This, however, although theoretically convenient, is an unrealistic presupposition which we will prefer to avoid if possible. Moreover, the initial condition  $X(0)=0$  or  $X(-T)\rightarrow 0$  as  $T\rightarrow\infty$  is not as innocuous as it seems at first sight.  $X(0)=0$  determines the mean of the process in no uncertain terms by attributing to the origin a very special status. The condition  $X(-T)\rightarrow 0$  as  $T\rightarrow\infty$  ensures that  $X(t)$  can be expressed in the form (52) even without the condition  $|\alpha|<1$ .

For the purposes of econometric modelling (see Part IV) it is interesting to consider the case where the index set is  $\mathbb{T}=\{0, 1, 2, \dots\}$  in relation to stationarity and asymptotic independence. As seen above, in this case  $\{X(t), t \in \mathbb{T}\}$  is not second-order stationary even under the restriction  $X(0)=0$ ,  $|\alpha|<1$ . Under the same conditions, however, the process is asymptotically independent. The non-stationarity stems from the fact the autocorrelation function  $v(t, t+\tau)$  depends on  $t$  because for  $\tau \geq 0$ ,

$$v(t, t+\tau) = \sigma^2 \alpha^\tau \left( \frac{1-\alpha^{2t}}{1-\alpha^2} \right), \quad |\alpha| < 1, \quad t \in \mathbb{T}. \quad (8.56)$$

This dependence on  $t$ , however, decreases as  $t$  increases. This led Priestley (1981) to introduce the concept of *asymptotic stationarity* which enables us to approximate the autocorrelation function by

$$v(t, t+\tau) \approx \sigma^2 \alpha^\tau \left( \frac{1}{1-\alpha^2} \right), \quad |\alpha| < 1, \quad t \in \mathbb{T}. \quad (8.57)$$

At this stage, it is interesting to consider the question of whether, instead of postulating (45) as a generating mechanism, we can actually 'derive' it by imposing certain restrictions on the structure of an arbitrary stochastic process. If we assume that the stochastic process  $\{X(t), t \in \mathbb{T}\}$  is: (i) *normal*, (ii) *Markov* and (iii) *stationary* then

$$E(X(t)/\mathcal{D}_{t-1}) = E(X(t)/\sigma(X(t-1))) \quad (8.58)$$

$$= \alpha X(t-1), \quad (8.59)$$

where  $\mathcal{D}_{t-1} = \sigma(X(t-1), X(t-2), \dots, X(0))$ . The first equality stems from the Markov property, the linearity of the conditional mean from the

normality and the time invariance of  $\alpha$  from stationarity. In order to ensure that  $|\alpha| < 1$  we need to assume that the process is also (iv) *asymptotically independent*. Defining the process  $u(t)$  by

$$u(t) = X(t) - E(X(t)/\mathcal{D}_{t-1}), \quad u(0) = X(0), \quad (8.60)$$

we can construct the GM

$$X(t) = \alpha X(t-1) + u(t), \quad t \in \mathbb{T}, \quad (8.61)$$

where  $u(t)$  is now a martingale difference orthogonal process relative to  $\mathcal{D}_t$ . This is because by construction  $E(u(t)/\mathcal{D}_{t-1}) = 0$  and for  $t > k$ ,  $E(u(t)u(k)) = E\{E[u(t)u(k)/\mathcal{D}_{t-1}]\} = 0$  by σ-CE8 of Section 7.2. Moreover, the process  $\{u(t), t \in \mathbb{T}\}$  can be viewed as a *white-noise process* or an *innovation process* relative to the information set  $\mathcal{D}_t$  because

$$(i) \quad E(u(t)) = E\{E(u(t)/\mathcal{D}_{t-1})\} = 0; \quad (8.62)$$

$$\begin{aligned} (ii) \quad E(u(t)^2) &= E\{E(u(t)^2/\mathcal{D}_{t-1})\} \\ &= E(X_t^2) - \alpha^2 E(X_{t-1}^2) = \sigma_x^2(1 - \alpha^2) = \sigma^2, \end{aligned} \quad (8.63)$$

say. This way of defining a white-noise process is very illuminating because it brings out the role of the information set relative to which the white-noise process is not predictable. In this case the process  $\{u(t), t \in \mathbb{T}\}$  as defined in (60) is white-noise (or non-predictable) relative to  $\mathcal{D}_t$ . That is,  $u(t)$  contains no ‘systematic’ information relative to  $\mathcal{D}_t$ . This, however, does not preclude the possibility of being able to predict  $u(t)$  using some other information set  $\mathcal{D}_t^* \supseteq \mathcal{D}_t$  with respect to which  $u(t)$  is a random variable. To summarise the above argument, if  $\{X(t), t \in \mathbb{T}\}$  is assumed to be a normal, Markov, stationary and asymptotically independent process, then the GM (61) follows by ‘design’.

The question which naturally arises at this point is ‘given that (61) seems identical to (45) what happens to the difficulties related to the time dependence of the autocorrelation function as shown in (50)?’ The answer is that the difficulty never arises because given the stationarity of  $X(t)$  we can derive its autocorrelation function by

$$\begin{aligned} v(\tau) &= E(X(t)X(t-\tau)) \\ &= E\{(\alpha X(t-1) + u(t))X(t-\tau)\} = \alpha v(\tau-1), \end{aligned} \quad (8.64)$$

since  $E(u(t)X(t-\tau)) = 0$ . This implies that

$$v(\tau) = \alpha^\tau v(0). \quad (8.65)$$

Moreover,

$$\begin{aligned} v(0) &= E(X(t)X(t)) = E(\alpha X(t-1)X(t)) + E(u(t)X(t)) \\ &= \alpha v(1) + \sigma^2 = \alpha^2 v(0) + \sigma^2. \end{aligned} \quad (8.66)$$

Hence,

$$v(0) = \frac{\sigma^2}{(1-\alpha^2)} \quad \text{and} \quad v(\tau) = \frac{\sigma^2}{(1-\alpha^2)} \alpha^\tau. \quad (8.67)$$

This shows clearly that in the case where the GM is ‘designed’ no need to change to asymptotic stationarity arises as in the case where (45) is postulated as a GM. What is more, the role of the various assumptions becomes much easier to understand when the probabilistic assumptions are made directly in terms of the stochastic process  $\{X(t), t \in \mathbb{T}\}$  and not the white-noise process.

### (5) Autoregressive, *mth*-order (AR(*m*))

The above discussion of the AR(1) process generalises directly to the AR(*m*) process where  $m \geq 1$ .

*Definition 17*

A stochastic process  $\{X(t), t \in \mathbb{T}\}$  is said to be **autoregressive of order *m*** (AR(*m*)) if it satisfies the stochastic difference equation

$$X(t) = \alpha_1 X(t-1) + \alpha_2 X(t-2) + \cdots + \alpha_m X(t-m) + u(t), \quad (8.68)$$

where  $\alpha_1, \alpha_2, \dots, \alpha_m$  are constant and  $u(t)$  is a white-noise process.

For the discussion of (68) viewed as a generating mechanism (GM) it is convenient to express it in the lag operator notation

$$\alpha(L)X(t) = u(t), \quad (8.69)$$

where  $\alpha(L) = (1 - \alpha_1 L - \alpha_2 L^2 - \cdots - \alpha_m L^m)$  with  $L^k X(t) = X(t-k), k \geq 1$ . The solution of the difference equation (24) for  $\mathbb{T}^* = \{0, \pm 1, \pm 2, \dots\}$  can be written as

$$X(t) = g(t) + \alpha^{-1}(L)u(t). \quad (8.70)$$

$g(t) = c_1 \lambda_1^t + c_2 \lambda_2^t + \cdots + c_m \lambda_m^t$  is the so-called general solution which is expressed as a linear combination of the roots  $(\lambda_1, \lambda_2, \dots, \lambda_m)$  of the polynomial

$$(x^m - \alpha_1 x^{m-1} - \cdots - \alpha_m) = 0 \quad (8.71)$$

(assumed to be distinct for convenience) with the constants  $c_1, c_2, \dots, c_m$

being functions of the initial conditions  $X(0), X(1), \dots, X(m-1)$ . The particular solution is the second component of (70) and takes the form:

$$\alpha^{-1}(L)u(t) = \sum_{j=0}^{\infty} \gamma_j u(t-j), \quad (8.72)$$

where

$$\left. \begin{array}{l} \gamma_0 = 1 \\ \gamma_1 + \alpha_1 \gamma_0 = 0 \\ \vdots \\ \gamma_m + \alpha_1 \gamma_{m-1} + \cdots + \alpha_m \gamma_0 = 0 \\ \gamma_{m+\tau} + \alpha_1 \gamma_{m+\tau-1} + \cdots + \alpha_m \gamma_\tau = 0 \end{array} \right\}. \quad (8.73)$$

In the simple case  $m=1$  considered above  $\alpha(L)=(1-\alpha L)$ ,  $g(t)=X(0)\alpha^t$  ( $\lambda_1=\alpha$ ),  $\gamma_j=\alpha^j, j=0, 1, 2, \dots$ . The restriction  $|\alpha|<1$  in the AR(1) case is now extended to all the roots of (71), i.e.

$$|\lambda_i| < 1, \quad i = 1, 2, \dots, m. \quad (8.74)$$

That is, the roots of the polynomial (71) are said to lie *within the unit circle*.

Under the restrictions (74) the general solution goes to zero, i.e.

$$g(t) \rightarrow 0 \quad \text{as } t \rightarrow \infty, \quad (8.75)$$

and the solution of the difference equation (68) can be written as

$$X(t) = \sum_{j=0}^{\infty} \gamma_j u(t-j). \quad (8.76)$$

This form can be used to determine the first two moments of the stochastic process  $\{X(t), t \in \mathbb{T}^*\}$ . In particular

$$\begin{aligned} E(X(t)) &= 0, \\ E(X(t)X(t+\tau)) &= E \left\{ \left( \sum_{j=0}^{\infty} \gamma_j u(t-j) \right) \left( \sum_{i=0}^{\infty} \gamma_i u(t+\tau-i) \right) \right\} \\ &= \left( \sum_{j=0}^{\infty} \gamma_j \gamma_{j+\tau} \right) \sigma^2. \end{aligned} \quad (8.77)$$

This is bounded when  $(\sum_{j=0}^{\infty} \gamma_j^2) < \infty$ , a condition which holds when the roots of the polynomial (71) lie within the unit circle (see Miller (1968)). In a sense the conditions (74) ensure that the stochastic process  $\{X(t), t \in \mathbb{T}^*\}$  as generated by (68) is both second-order stationary and asymptotically independent because the condition  $(\sum_{j=0}^{\infty} \gamma_j^2) < \infty$  implies that  $v(\tau) \rightarrow 0$  as  $\tau \rightarrow \infty$ .

As in the case of an AR(1) process when the index set  $\mathbb{T} = \{0, 1, 2, \dots\}$  is

used instead of  $\mathbb{T}^* = \{0, \pm 1, \pm 2, \dots\}$  we run into problems with the second-order stationarity of  $\{X(t), t \in \mathbb{T}\}$ , which we can partly overcome using the concept of asymptotic stationarity. This will not be pursued any further because, as argued in the AR(1) case, when the parametric model is not postulated as a GM but ‘designed’ by making the necessary assumptions directly in terms of  $\{X(t), t \in \mathbb{T}\}$  no such problems arise. In particular, if we assume that the process is (i) normal, (ii)  $m$ th-order Markov, (iii) stationary and (iv) asymptotically independent then

$$\begin{aligned} E(X(t)/\sigma(X(t-1), \dots, X(0))) &= E(X(t)/\sigma(X(t-1), \dots, X(t-m))) \\ &= \sum_{i=1}^m \alpha_i X(t-i). \end{aligned} \quad (8.78)$$

The first equality stems from the  $m$ th-order Markov property. The linearity of the conditional mean is due to the normality, the time invariance of the  $\alpha_i$ s is due to the stationarity and asymptotic independence implies that the roots of the polynomial (71) lie inside the unit circle. If we define the increasing sequence of  $\sigma$ -fields  $\mathcal{D}_t = \sigma(X(t), X(t-1), \dots, X(0))$ ,  $t \in \mathbb{T}$ , and the process  $\{u(t), t \in \mathbb{T}\}$  by

$$u(t) = X(t) - E(X(t)/\mathcal{D}_{t-1}), \quad t \in \mathbb{T}, \quad (8.79)$$

we can deduce that  $\{u(t), \mathcal{D}_t, t \in \mathbb{T}\}$  is a martingale difference, orthogonal (an innovation) process. This enables us to ‘design’ the AR( $m$ ) GM as in (68) from first principles with  $u(t)$  being a white-noise process relative to the information set  $\mathcal{D}_t$ .

The autocovariance function can be defined directly, as in the ‘designed’ AR(1) case by multiplying (68) with  $X(t-\tau)$  and taking expectations to yield

$$E(X(t)X(t-\tau)) = \alpha_1 E(X(t-1)X(t-\tau)) + \dots + E(u(t)X(t-\tau)),$$

$$\begin{aligned} \Rightarrow v(0) &= \alpha_1 v(1) + \alpha_2 v(2) + \dots + \alpha_m v(m) + \sigma^2, \\ v(\tau) &= \alpha_1 v(\tau-1) + \alpha_2 v(\tau-2) + \dots + \alpha_m v(\tau-m), \tau > 0. \end{aligned} \quad (8.80)$$

(8.81)

Hence, we can see that the autocovariances satisfy the same difference equation as the process itself. Similarly, the autocorrelation function takes the form

$$r(\tau) = \alpha_1 r(\tau-1) + \dots + \alpha_m r(\tau-m), \quad \tau > 0. \quad (8.82)$$

The system of equations for  $\tau = 1, 2, \dots, m$  are known as *Yule–Walker equations* which play an important role in the estimation of the coefficients  $\alpha_1, \alpha_2, \dots, \alpha_m$  (see Priestley (1981)). The relationship between the

autocorrelations and the asymptotic independence of  $\{X(t), t \in \mathbb{T}\}$  is shown most clearly by the relationship

$$r(\tau) = \gamma_1 \lambda_1^\tau + \gamma_2 \lambda_2^\tau + \cdots + \gamma_m \lambda_m^\tau, \quad \tau = 0, 1, 2, \dots, \quad (8.83)$$

viewed as a general solution of the difference equation (82). Under the restrictions  $|\lambda_i| < 1, i = 1, 2, \dots, m$  (implied by asymptotic independence) we can deduce that

$$r(\tau) \rightarrow 0 \quad \text{as } \tau \rightarrow \infty. \quad (8.84)$$

#### (6) Moving average (MA) processes

*Definition 18*

The stochastic process  $\{X(t), t \in \mathbb{T}\}$  is said to be a **moving average process of order  $k$**  ( $MA(k)$ ) if it can be expressed in the form

$$X(t) = u(t) + b_1 u(t-1) + \cdots + b_k u(t-k), \quad (8.85)$$

where  $b_1, b_2, \dots, b_k$  are constants and  $\{u(t), t \in \mathbb{T}\}$  is a **white-noise process**. That is, the white-noise process is used to build the process  $\{X(t), t \in \mathbb{T}\}$ , being a linear combination of the last  $k$   $u(t-i)$ s.

Given that  $\{X(t), t \in \mathbb{T}\}$  is a linear combination of uncorrelated random variables we can deduce that

$$E(X(t)) = 0, \quad (8.86)$$

$$E(X(t)X(t+\tau)) = \begin{cases} \sigma^2(b_\tau + b_{\tau+1}b_1 + \cdots + b_kb_{k-\tau}), & 0 \leq \tau \leq k \\ 0, & \tau > k \end{cases} \quad (8.87)$$

$$r(\tau) = r(-\tau) = \begin{cases} \left\{ \sum_{j=\tau}^k b_j b_{j-\tau} \right\} / \left\{ \sum_{j=0}^k b_j^2 \right\}, & 0 \leq \tau \leq k \\ 0, & \tau > k \end{cases} \quad (8.88)$$

( $b_0 = 1$ ). These results show that, firstly, a  $MA(k)$  process is second-order stationary irrespective of the values taken by  $b_1, b_2, \dots, b_k$ , and, secondly, its autocovariance and autocorrelation functions have a ‘cut-off’ after  $k$  periods. That is, a  $MA(k)$  process is both *second-order stationary* and  *$k$ -correlated* ( $r(\tau) = 0, \tau > k$ ).

In the simple case of a  $MA(1)$ , (85) takes the form

$$X(t) = u(t) + b_1 u(t-1), \quad (8.89)$$

with

$$v(0) = (1 + b_1^2)\sigma^2, \quad v(1) = b_1\sigma^2, \quad r(\tau) = 0, \quad \tau > 1 \quad (8.90)$$

$$r(1) = \frac{b_1}{(1 + b_1^2)}, \quad r(\tau) = 0, \quad \tau > 1. \quad (8.91)$$

As we can see, a MA( $k$ ) process is severely restrictive in relation to time-heterogeneity and memory. It turns out, however, that any second-order stationary, asymptotically independent process  $\{X(t), t \in \mathbb{T}\}$  can be 'expressed' as a MA( $\infty$ ), i.e.

$$X(t) = \sum_{j=0}^{\infty} b_j u(t-j), \quad t \in \mathbb{T}, \quad (8.92)$$

where  $(\sum_{j=0}^{\infty} b_j^2) < \infty$  and  $\{u(t), t \in \mathbb{T}\}$  is an innovation process. This result constitutes a form of the celebrated *Wold decomposition theorem* (see Priestley (1981)) which provided the theoretical foundation for MA( $k$ ) and ARMA( $p, q$ ) processes to be considered next. The MA( $\infty$ ) in (92) can be constructed from first principles by restricting the time-heterogeneity and the memory of the process. If we assume that  $\{X(t), t \in \mathbb{T}\}$  is (i) second-order stationary, and (ii) asymptotically uncorrelated, then we can define the *innovation process*  $\{u(t), t \in \mathbb{T}\}$  by

$$u(t) = X(t) - E(X(t)/\mathcal{D}_{t-1}), \quad t \in \mathbb{T}, \quad (8.93)$$

where  $\mathcal{D}_t = \sigma(X(t), X(t-1), \dots, X(0))$ . Asymptotic independence enables us to deduce that  $\mathcal{D}_t = \sigma(u(t), u(t-1), \dots, u(0))$  and thus by  $\sigma$ -CE6 (see Section 7.2)

$$X(t) = E(X(t)/\mathcal{D}_t). \quad (8.94)$$

Given that  $\{u(t), t \in \mathbb{T}\}$  is an innovation process (martingale difference, orthogonal process), it can be viewed as an orthogonal basis for  $\mathcal{D}_t$ . This enables us to deduce that  $E(X(t)/\mathcal{D}_t)$  can be expressed as a linear combination of the  $u(t-j)$ s, i.e.

$$E(X(t)/\mathcal{D}_t) = \sum_{j=0}^{\infty} b_j u(t-j), \quad (8.95)$$

from which (92) follows directly. In a sense the process  $\{u(t), t \in \mathbb{T}\}$  provides the 'building blocks' for any second-order stationary process. This can be seen as a direct extension of the result that any element of a linear space can be expressed in terms of an orthogonal basis uniquely, to the case of an infinite dimensional linear space, a Hilbert space (see Kreyszig (1978)).

The MA( $k$ ) process can be viewed as a special case of (92) where the assumption of asymptotic uncorrelatedness is restricted to  $k$ -correlatedness. In such a case  $\{X(t), t \in \mathbb{T}\}$  can be expressed as a linear function of the last  $k$  orthogonal elements  $u(t), u(t-1), \dots, u(t-k)$ .

## (7) Autoregressive moving average processes

As shown above, any second-order stationary, asymptotically uncorrelated

process can be expressed in a MA( $\infty$ ) form

$$X(t) = \sum_{j=0}^{\infty} b_j u(t-j), \quad (8.96)$$

where  $(\sum_{j=0}^{\infty} b_j^2) < \infty$  and  $\{u(t), t \in \mathbb{T}\}$  is an innovation process. Such a representation, however, is of very little value in practice in view of its non-operational nature. The ARMA( $p, q$ ) formulation provides a parsimonious, operational form for (96).

### *Definition 19*

*A stochastic process  $\{X(t), t \in \mathbb{T}\}$  is said to be an **autoregressive moving average process of order  $p, q$**  (ARMA( $p, q$ )) if it can be expressed in the form:*

$$\begin{aligned} X(t) + \alpha_1 X(t-1) + \cdots + \alpha_p X(t-p) &= u(t) + b_1 u(t-1) + \cdots \\ &\quad + b_q u(t-q) \end{aligned} \quad (8.97)$$

where  $\alpha_1, \alpha_2, \dots, \alpha_p, b_1, b_2, \dots, b_q$ , are constants and  $\{u(t), t \in \mathbb{T}\}$  is a white-noise process.

In order to motivate the ARMA( $p, q$ ) formulation as an operational form of the MA( $\infty$ ) representation (96) let us express the latter in terms of the infinite polynomial  $b^*(L) = 1 + b_1 L + b_2 L^2 + \cdots$

$$X(t) = b^*(L)u_t. \quad (8.98)$$

Under certain mild regularity conditions  $b^*(L)$  can be approximated by the ratio of two finite polynomials (see Dhrymes (1971)),

$$b^*(L) = \frac{b_q(L)}{\alpha_p(L)} = \frac{(1 + b_1 L + b_2 L^2 + \cdots + b_q L^q)}{(1 + \alpha_1 L + \alpha_2 L^2 + \cdots + \alpha_p L^p)}, \quad p \geq q. \quad (8.99)$$

For large enough  $p$  and  $q$ ,  $b^*(L)$  can be approximated to any degree of accuracy. Substituting (99) back into (98) we get

$$\alpha_p(L)X(t) = b_q(L)u(t), \quad (8.100)$$

which is an ARMA( $p, q$ ) model. This is an operational form which is widely used in time-series modelling to provide a parsimonious approximation to second-order stationary processes. Time-series modelling based on ARMA( $p, q$ ) formulation was popularised by Box and Jenkins (1976).

The ARMA( $p, q$ ) formulation (97) can be viewed as an extension of the AR( $m$ ) representation in so far as the non-homogeneous part of the difference equation includes additional terms. This, however, makes no difference to the mathematical properties of (97) as a stochastic difference

equation. In particular, the asymptotic independence of the process depends only on the restriction that the roots of the polynomial

$$\alpha_p(\lambda) = (\lambda^p + \alpha_1 \lambda^{p-1} + \cdots + \alpha_p) = 0 \quad (8.101)$$

lie inside the unit circle. No restrictions are needed on the coefficients or the roots of  $b_q(L)$ . Such restrictions are needed in the case where an AR( $\infty$ ) formulation of (97) is required. Assuming that the roots of  $b_p(\lambda) = 0$  lie inside the unit circle enables us to express the ARMA( $p, q$ ) in the form

$$X(t) = \sum_{i=1}^{\infty} \alpha_i X(t-i) + u(t). \quad (8.102)$$

This form, however, can be operational only when it can be approximated by an AR( $m$ ) representation for 'large enough'  $m$ . The conditions on  $\alpha_p(\lambda)$  are commonly known as *stability conditions* and those on  $b_q(\lambda)$  as *invertibility conditions* (see Box and Jenkins (1976)).

The popularity of ARMA( $p, q$ ) formulations in time-series modelling stems partly from the fact that the formulation can be extended to a particular type of non-stationary stochastic processes; the so-called *homogeneous non-stationarity*. This is the case where only the mean is time dependent (the variance and covariance are time invariant) and the time change is local. In such a case the stochastic process  $\{Z(t), t \in \mathbb{T}\}$  exhibiting such behaviour can be transformed into a stationary process by differencing, i.e. define

$$X(t) = (1 - L)^d Z(t), \quad (8.103)$$

where  $d$  is some integer. For  $d = 0$ ,  $X(t) = Z(t)$ ; for  $d = 1$ ,  $X(t) = Z(t) - Z(t-1)$ ; first difference and, for  $d = 2$ ,

$$X(t) = Z(t) - 2Z(t-1) + Z(t-2). \quad (8.104)$$

Once the process is transformed into a stationary one the ARMA( $p, q$ ) formulation is used to model  $X(t)$ . In terms of the original model, however, the formulation is

$$\alpha_p(L)(1 - L)^d Z(t) = b_q(L)u(t), \quad (8.105)$$

which is called an ARIMA ( $p, d, q$ ); *autoregressive, integrated moving average*, of order  $p, d, q$  (see Box and Jenkins (1976)).

In the context of econometric modelling the ARIMA formulation is of limited value because it is commonly preferable to model non-stationarity as part of the statistical model specification rather than transform the data at the outset.

### 8.5      Summary

The purpose of this chapter has been to extend the concept of a random variable (r.v.) in order to enable us to model dynamic processes. The extension came in the form of a stochastic process  $\{X(t), t \in \mathbb{T}\}$  where  $X(t)$  is defined on  $S \times \mathbb{T}$  not just on  $S$  as in the case of a r.v.; the index set  $\mathbb{T}$  provides the time dimension needed. The concept of a stochastic process enables us to extend the notion of the probability model  $\Phi = \{f(x_1, \dots, x_n; \theta), \theta \in \Theta\}$  discussed so far to one with a distinct time dimension

$$\Phi = \{f(x(t); \theta_t), \theta_t \in \Theta_t, t \in \mathbb{T}\}. \quad (8.106)$$

This, however, presents us with an obvious problem. The fact that the unknown parameter vector  $\theta_t$ , indexing the parametric family of densities, depends on  $t$  will make our task of ‘estimating’ their values from (commonly) a single sample realisation impossible.

In order to make the theory build upon the concept of a stochastic process manageable we need to impose certain restrictions on the process itself. The notions of asymptotic independence and stationarity are employed with this purpose in mind. Asymptotic independence, by restricting the memory of the stochastic process, enables us to approximate such processes with parametric ones which reduces the number of unknown parameters to a finite set  $\theta$ . Similarly, stationarity by imposing time-homogeneity on the stochastic process enables us to use time-independent parameters to model a dynamic process in a ‘statistical equilibrium’. The effect of both sets of restrictions is to reduce the probability model (106) to

$$\Phi = \{f(x(t); \theta), \theta \in \Theta, t \in \mathbb{T}\}. \quad (8.107)$$

This form of a probability model is extensively used in Part IV as an important building block of statistical models of particular interest in econometrics.

#### *Important concepts*

Stochastic process, realisation of a process discrete stochastic processes, distribution of a stochastic process, symmetry and compatibility restrictions, autocovariance, autocorrelation, autoprodut, cross-covariance and cross-correlation functions, normal stochastic process, vector stochastic process,  $l$ th-order process, time-homogeneity and memory of a process, strict stationarity, second-order stationarity, non-stationarity, homogeneous non-stationarity, asymptotically independent and uncorrelated processes,  $m$ -dependent process, strong mixing,

asymptotic independence,  $m$ th-order Markov process, ergodicity, martingale, martingale difference, innovation process, Markov property, Brownian motion process, white-noise, parametric and non-parametric processes, AR(1), AR( $m$ ), initial conditions, stability and invertibility conditions, MA( $m$ ), ARMA( $p, q$ ), ARIMA( $p, d, q$ ).

### *Questions*

1. What is the reason for extending the concept of a random variable to that of a stochastic process?
2. Define the concept of a stochastic process and explain its main components.
3. ‘ $X(s, t)$  can be interpreted as a random variable, a non-stochastic function (realisation) as well as a single number.’ Discuss.
4. ‘Wild fluctuations of a realisation of a process have nothing to do with its randomness.’ Discuss.
5. How do we specify the structure of a stochastic process?
6. Compare the joint distribution of a set of  $n$  normally distributed independent r.v.’s with that of a stochastic process  $\{X(t), t \in \mathbb{T}\}$  for  $(t_1, t_2, \dots, t_n)$  in terms of the unknown parameters involved.
7. Let  $\{X(t), t \in \mathbb{T}\}$  be a stationary normal process. Define its joint distribution for  $t < t_1, \dots, < t_n$  and explain the effect on the unknown parameters involved by assuming (i)  $m$ -dependence or (ii)  $m$ th-order Markovness.
8. If  $\{X(t), t \in \mathbb{T}\}$  is a normal stationary process then:
  - (i) asymptotic independence and uncorrelatedness; as well as
  - (ii) strict and second-order stationarity, coincide.’Explain.
9. Discuss and compare the notions of an  $m$ -dependent and an  $m$ th-order Markov process.
10. Explain how restrictions on the time-heterogeneity and memory of a stochastic process can help us construct operational probability models for dynamic phenomena.
11. Compare the memory restriction notions of asymptotic independence, asymptotic uncorrelatedness,  $m$ th-order Markovness, mixing and ergodicity.
12. Explain the notion of homogeneous non-stationarity and its relation to ARIMA( $p, d, q$ ) formulations.
13. Explain the difference between a parametric AR(1) stochastic process and a ‘designed’ non-parametric AR(1) model.
14. Define the notion of a martingale and explain its attractiveness for modelling dynamic phenomena.

15. Compare and contrast the concepts of a white-noise and an innovation process.
16. 'The AR(1) process is a Markov process but not a martingale unless we sacrifice asymptotic independence.' Discuss.
17. 'The AR(1) process defined over  $\mathbb{T} = \{0, 1, 2, \dots\}$  is not a second-order stationary process even if  $|\alpha| < 1$ .' Discuss.
18. 'Any second-order stationary and asymptotically uncorrelated stochastic process can be expressed in  $\text{MA}(\infty)$  form.' Explain.
19. Explain the role of the initial conditions in the context of an AR(1) process.
20. Explain the role of the stability conditions in the context of an AR( $m$ ) process.
21. 'The  $\text{ARMA}(p, q)$  formulation provides a parsimonious representation for second-order stationary stochastic processes.' Explain.
22. Discuss the likely usefulness of  $\text{ARIMA}(p, q)$  formulations in econometric modelling.

**Additional references**

Anderson (1971); Chung (1974); Doob (1953); Feller (1970); Fuller (1976); Gnedenko (1969); Granger and Newbold (1977); Granger and Watson (1984); Hannan (1970); Lamperti (1977); Nerlove *et al.* (1979); Rosenblatt (1974); Whittle (1970); Yaglom (1962).

## CHAPTER 9

---

### Limit theorems

---

#### 9.1 The early limit theorems

The term ‘limit theorems’ refers to several theorems in probability theory under the generic names, ‘law of large numbers’ (LLN) and ‘central limit theorem’ (CLT). These limit theorems constitute one of the most important and elegant chapters of probability theory and play a crucial role in statistical inference. The origins of these theorems go back to the seventeenth-century result proved by James Bernoulli.

*Bernoulli’s theorem*

*Let  $S_n$  be the number of occurrences of an event  $A$  in  $n$  independent trials of a random experiment  $\mathcal{E}$  and  $p = P(A)$  is the probability of occurrence of  $A$  in each of the trials. Then for any  $\varepsilon > 0$*

$$\lim_{n \rightarrow \infty} Pr\left(\left|\frac{S_n}{n} - p\right| < \varepsilon\right) = 1, \quad (9.1)$$

*i.e. the limit of the probability of the event  $|[(S_n/n) - p]| < \varepsilon$  approaches one as the number of trials goes to infinity.*

Shortly after the publication of Bernoulli’s result *De Moivre* and *Laplace* in their attempt to provide an easier way to calculate binomial probabilities proved that when  $[(S_n/n) - p]$  is multiplied by a factor equal to the inverse of its standard error the resulting quantity has a distribution which approaches the normal as  $n \rightarrow \infty$ , i.e.

$$\lim_{n \rightarrow \infty} Pr\left\{\frac{\left(\frac{S_n}{n} - p\right)}{\sqrt{\frac{p(1-p)}{n}}} \leq z\right\} = \int_{-\infty}^z \frac{1}{\sqrt{(2\pi)}} \exp\{-\frac{1}{2}u^2\} du. \quad (9.2)$$

These two results gave rise to a voluminous literature related to the various ramifications and extensions of the Bernoulli and De Moivre–Laplace theorems known today as ‘the’ LLN and ‘the’ CLT respectively. The purpose of this chapter is to consider some of the extensions of the Bernoulli and De Moivre–Laplace results. In the discussion which follows emphasis is placed on the intuitive understanding of the conclusions as well as the crucial assumptions underlying the various limit theorems. The discussion is semi-historical in a conscious attempt to motivate the various extensions and the weakening of the underlying assumptions giving rise to the results.

The main conditions underlying the Bernoulli and De Moivre–Laplace results are the following:

- (LT1)  $S_n = \sum_{i=1}^n X_i$ , that is,  $S_n$  defined as the **sum** of  $n$  random variables (r.v.’s).
- (LT2)  $X_i = 1$ , if  $A$  occurs, and  $X_i = 0$ , otherwise,  $i = 1, 2, \dots, n$ , i.e. the  $X_i$ s are **Bernoulli** r.v.’s and hence  $S_n$  is a **binomially distributed** r.v.
- (LT3)  $X_1, X_2, \dots, X_n$  are **independent** r.v.’s.
- (LT4)  $f(x_1) = f(x_2) = \dots = f(x_n)$ , i.e.  $X_1, X_2, \dots, X_n$  are **identically distributed** with  $\Pr(X_i = 1) = p$ ,  $\Pr(X_i = 0) = 1 - p$  for  $i = 1, 2, \dots, n$ .
- (LT5)  $E(S_n/n) = p$ , i.e. we consider the event of the difference between a r.v. and its **expected value**.

The main difference between the Bernoulli and De Moivre–Laplace theorems lies in their notion of *convergence*, the former referring to the convergence of the probability associated with the sequence of events  $|[(S_n/n) - p]| < \varepsilon$  and the latter to the convergence of the probability associated with a very specific sequence of events, that is, events of the form ( $Z \leq z$ ) which define the distribution function  $F(z)$ . In order to discriminate between them we call the former ‘convergence in probability’ and the latter ‘convergence in distribution’.

### *Definition 1*

*A sequence of r.v.’s  $\{Y_n, n \geq 1\}$  is said to **converge in probability** to a r.v. (or constant)  $Y$  if for every  $\varepsilon > 0$*

$$\lim_{n \rightarrow \infty} \Pr(|Y_n - Y| < \varepsilon) = 1. \quad (9.3)$$

*We denote this with  $Y_n \xrightarrow{P} Y$ .*

### *Definition 2*

*A sequence of r.v.’s  $\{Y_n, n \geq 1\}$  with distribution functions  $\{F_n(y), n \geq 1\}$  is said to **converge in distribution** to a r.v.  $Y$  with distribution*

function  $F(y)$  if

$$\lim_{n \rightarrow \infty} F_n(y) = F(y) \quad (9.4)$$

at all points of continuity of  $F(y)$ ; denoted by  $Y_n \xrightarrow{D} Y$ .

It should be emphasised that neither of the above types of convergence tells us anything about any convergence of the sequence  $\{Y_n\}$  to  $Y$  in the sense used in mathematical analysis, such as for each  $\varepsilon > 0$  and  $s \in S$ , there exists an  $N = N(\varepsilon, s)$  such that

$$|Y_n(s) - Y(s)| < \varepsilon \quad \text{for } n > N. \quad (9.5)$$

Both convergence types refer only to convergence of probabilities or functions associated with probabilities. On the other hand, the definition of a r.v. has nothing to do with probabilities and the above convergence of  $Y_n$  to  $Y$  on  $S$  is convergence of real valued functions defined on  $S$ . The type of stochastic convergence which comes closer to the above mathematical convergence is known as ‘almost sure’ convergence.

### *Definition 3*

A sequence of r.v.’s  $\{Y_n, n \geq 1\}$  converges to  $Y$  (a r.v. or a constant) **almost surely** (or with probability one) if

$$Pr\left(\lim_{n \rightarrow \infty} Y_n = Y\right) = 1; \quad \text{denoted by } Y_n \xrightarrow{a.s.} Y, \quad (9.6)$$

or, equivalently, if for any  $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} Pr\left(\bigcup_{m \geq n} |Y_m - Y| < \varepsilon\right) = 1. \quad (9.7)$$

This is a much stronger mode of convergence than either convergence in probability or convergence in distribution. For a more extensive discussion of these modes of convergence and their interrelationships see Chapter 10. The limit theorems associated with convergence almost surely are appropriately called ‘strong law of large numbers’ (SLLN). The term is used to emphasise the distinction with the ‘weak law of large numbers’ (WLLN) associated with convergence in probability.

In the next section the law of large numbers is used as an example of the developments the various limit theorems have undergone since Bernoulli. For this reason the discussion is intentionally rather long in an attempt to motivate a deeper understanding of the crucial assumptions giving rise to all the limit theorems considered in the sequel.

## 9.2 The law of large numbers

### (1) The weak law of large numbers (WLLN)

Early in the nineteenth century Poisson realised that the condition LT4 asserting identical distributions for  $X_1, \dots, X_n$  was not necessary for the result to go through.

*Poisson's theorem*

Let  $\{X_n, n \geq 1\}$  be a sequence of independent Bernoulli r.v.'s with  $Pr(X_i = 1) = p_i$  and  $Pr(X_i = 0) = 1 - p_i$ ,  $i = 1, 2, \dots, n$ , then, for any  $\varepsilon > 0$ ,

$$\lim_{n \rightarrow \infty} Pr\left(\left|\frac{S_n}{n} - \frac{1}{n} \sum_{i=1}^n p_i\right| < \varepsilon\right) = 1. \quad (9.8)$$

The important breakthrough in relation to the WLLN was made by Chebyshev who realised that not only LT4 but LT2 was unnecessary for the result to follow. That is, the fact that  $X_1, \dots, X_n$  were Bernoulli r.v.'s was not contributing to the result in any essential way. What was crucially important was the fact that we considered the summation of  $n$  r.v.'s to form  $S_n = \sum_{i=1}^n X_i$  and comparing it with its mean.

*Chebyshev's theorem*

Let  $\{X_n, n \geq 1\}$  be a sequence of independent r.v.'s such that  $E(X_i) = \mu_i$ ,  $\text{Var}(X_i) = \sigma_i^2 < c < \infty$ ,  $i = 1, 2, \dots, n$ , then for any  $\varepsilon > 0$ ,

$$\lim_{n \rightarrow \infty} Pr\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n \mu_i\right| < \varepsilon\right) = 1. \quad (9.9)$$

In order to see how these conditions ensure the result let us prove Chebyshev's theorem.

*Proof:* Since the  $X_i$ 's are independent

$$\text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \sigma_i^2 \leq \frac{c}{n}.$$

Using Chebyshev's inequality for  $(1/n) \sum_i X_i$  we get

$$Pr\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n \mu_i\right| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2} \sum_{i=1}^n \frac{\sigma_i^2}{n^2} \leq \frac{c}{n\varepsilon^2},$$

$$\text{since } \lim_{n \rightarrow \infty} \left(\frac{c}{n\varepsilon^2}\right) = 0, \quad \lim_{n \rightarrow \infty} Pr\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n \mu_i\right| > \varepsilon\right) = 0,$$

$$\text{or } \lim_{n \rightarrow \infty} Pr\left(\left|\frac{1}{n} \sum_i X_i - \frac{1}{n} \sum_i \mu_i\right| < \varepsilon\right) = 1.$$

Markov, a student of Chebyshev's, noticed in the proof of Chebyshev's theorem the fact that the  $X_1, X_2, \dots, X_n$  are independent played only a minor role in enabling us to deduce that  $\text{Var}(S_n) = (1/n^2) \sum_{i=1}^n \sigma_i^2$ . The above proof goes through provided that  $(1/n^2) \text{Var}(S_n) \rightarrow 0$  as  $n \rightarrow \infty$ . Since

$$\text{Var}(S_n) = \sum_{i=1}^n \text{Var}(X_i) + \sum_{i \neq j} \text{Cov}(X_i X_j) \dagger \quad (9.10)$$

we need to assume that  $\text{Var}(\sum_i X_i)$  is of smaller *order of magnitude* (see Chapter 10) than  $n^2$  for the result to follow. Hence LT3 is not a crucial condition.

#### *Markov's theorem*

Let  $\{X_n, n \geq 1\}$  be a sequence of r.v.'s such that

$$\frac{1}{n^2} \text{Var}(S_n) \rightarrow 0, \quad \text{as } n \rightarrow \infty, \quad (9.11)$$

then

$$\lim_{n \rightarrow \infty} Pr\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n E(X_i)\right| < \varepsilon\right) = 1. \quad (9.12)$$

Khinchin, a student of Markov's, realised that, in the case of independent and identically distributed (IID) r.v.'s, Markov's condition was not a necessary condition. In fact in the IID case no restriction on the nature of the variances is needed.

#### *Khinchin's theorem*

Let  $\{X_n, n \geq 1\}$  be a sequence of IID r.v.'s, then the existence of  $E(X_i) = \mu$  for all  $i$  is sufficient to imply that for any  $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} Pr\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \mu\right| < \varepsilon\right) = 1. \quad (9.13)$$

Kolmogorov (1926) settled the issue by providing both necessary as well as sufficient conditions for the WLLN.

#### *Kolmogorov's theorem 1*

The sequence of r.v.'s  $\{X_n, n \geq 1\}$  obey the WLLN if and only if

$$E\left(\frac{\left[S_n - \sum_i E(X_i)\right]^2}{n^2 + \left[S_n - \sum_i E(X_i)\right]^2}\right) \rightarrow 0, \quad \text{as } n \rightarrow \infty. \quad (9.14)$$

† There are  $n^2$  terms and thus if all of them are bounded the  $\text{Var}(\sum_{i=1}^n X_i)$  is at least of the same order as  $n^2$ .

(2)      **The strong law of large numbers (SLLN)**

The first result relating to the almost sure convergence of  $S_n$  for the Bernoulli distributed r.v.'s case was proved by Borel in 1909.

*Borel's theorem*

Let  $\{X_n\}$  be a sequence of IID Bernoulli r.v.'s with  $Pr(X_i = 1) = p$  and  $Pr(X_i = 0) = 1 - p$  for all  $i$ , then

$$Pr\left(\lim_{n \rightarrow \infty} \frac{S_n}{n} = p\right) = 1. \quad (9.15)$$

In other words, the event defined by  $\{s: \lim_{n \rightarrow \infty} [S_n(s)]/n = p, s \in S\}$ , has probability one;  $S$  being the sample space. An equivalent way to express this is

$$\lim_{n \rightarrow \infty} Pr\left(\max_{m \geq n} \left( \left| \frac{S_m}{m} - p \right| \geq \varepsilon \right)\right) = 0. \quad (9.16)$$

This brings out the relationship between the SLLN and the WLLN since the former refers to the simultaneous realisation of the inequalities and

$$\left| \frac{S_n}{n} - p \right| \leq \max_{m \geq n} \left| \frac{S_m}{m} - p \right|. \quad (9.17)$$

This implies that ' $\xrightarrow{\text{as}}$ '  $\Rightarrow$  ' $\xrightarrow{P}$ '.

Kolmogorov, by replacing the Markov condition

$$\lim_{n \rightarrow \infty} \frac{1}{n^2} \sum_{k=1}^n \text{Var}(X_k) = 0 \quad (9.18)$$

for the WLLN in the case of independent r.v.'s, with the stronger condition

$$\sum_{k=1}^{\infty} \frac{1}{k^2} \text{Var}(X_k) < \infty, \quad (9.19)$$

proved the first SLLN for a general sequence of independent r.v.'s.

*Kolmogorov's theorem 2*

Let  $\{X_n, n \geq 1\}$  be a sequence of independent r.v.'s such that  $E(X_i)$  and  $\text{Var}(X_i)$  exist for all  $i = 1, 2, \dots$ , then if they satisfy the condition (19) we can deduce that

$$Pr\left(\lim_{n \rightarrow \infty} \frac{1}{n} \left( \sum_{i=1}^n [X_i - E(X_i)] \right) = 0\right) = 1.$$

This SLLN is analogous to Chebyshev's WLLN and in the same way we can prove it using an inequality. The inequality used in this context is *Kolmogorov's inequality*: If  $X_1, X_2, \dots, X_n$  are independent r.v.'s, such that  $\text{Var}(X_i) = \sigma_i^2 < \infty$ ,  $i = 1, 2, \dots, n$ , then for any  $\varepsilon > 0$

$$Pr\left(\max_{1 \leq k \leq n} |S_k - E(S_k)| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2} \sum_{i=1}^n \sigma_i^2. \quad (9.20)$$

Kolmogorov went on to prove that in the case where  $\{X_n, n \geq 1\}$  is a sequence of IID r.v.'s such that  $E(X_i) < \infty$  then

$$\sum_{k=1}^{\infty} \frac{\text{Var}(X_k)}{k^2} = \sum_{k=1}^{\infty} \frac{1}{k^2} \int_{-k}^k x^2 f(x) dx < \infty, \quad (9.21)$$

which implies that for such a sequence the existence of expectation is a necessary as well as sufficient condition for the SLLN.

Having argued that some of the conditions of the Bernoulli theorem did not contribute (in any essential way) to the result, the question that arises naturally is, 'what are the important elements giving rise to the "law of large numbers" (SLLN, WLLN)?' The Markov condition (18) for the WLLN, and Kolmogorov's condition (19) for the SLLN, hold the key to the answer of this question. It is clear from these two conditions that the most important ingredient is the restriction on the variance of the partial sums  $S_n$ , that is, we need the  $\text{Var}(S_n)$  to increase at most as quickly as  $n$ . More formally we need  $\text{Var}(S_n)$  to be at most of order  $n$  and we write  $\text{Var}(S_n) = O(n)$ . In order to see this let us consider some of the cases discussed above. In the IID case if  $\text{Var}(X_i) = \sigma^2$  for all  $i$ , then  $\text{Var}(S_n) = n\sigma^2 = O(n)$ .

In the case of independent r.v.'s with

$$\text{Var}(X_i) = \sigma_i^2 < \infty, \quad i = 1, 2, \dots, \quad \text{then } \text{Var}(S_n) = \sum_{i=1}^n \sigma_i^2 = O(n). \quad (9.22)$$

Moreover, the Markov condition can be written as  $\text{Var}(S_n) = o(n^2)$  where small 'o' reads 'of smaller order than' achieves the same effect since  $\text{Var}(S_n) = O(n) \Rightarrow \text{Var}(S_n) = o(n^2)$  (see Chapter 10). The Kolmogorov condition is a more restrictive form of the Markov condition, requiring the variance of the partial sums to be uniformly of at most of order  $n$ . This being the case, it becomes obvious that the conditions LT3 and LT4, assuming independence and identically distributed r.v.'s, are not fundamental ingredients. Indeed, if we drop the identically distributed condition altogether and weaken independence to *martingale orthogonality* the above limit theorems go through with minor modifications. We say that a sequence of r.v.'s  $\{X_n, n \geq 1\}$  is martingale orthogonal if  $E(X_n | \sigma(X_{n-1}, \dots,$

$X_1) = 0$ ,  $n \geq 1$ . It should come as no surprise to learn that both important tools in proving the WLLN and SLLN, the Chebyshev and Kolmogorov inequalities hold true for orthogonal r.v.'s. This enables us to prove the WLLN and SLLN under much weaker conditions than the ones discussed above. The most useful of these results are the ones related to martingales because they can be seen as direct extensions of the 'independent' case and the results are general enough to cover most types of dependencies we are interested in.

### (3)      *The law of large numbers for martingales*

Let  $\{S_n, \mathcal{D}_n, n \in N\}$  be a martingale such that  $E(S_n) = 0$ , for all  $n$ , and define  $Y_n = S_n - S_{n-1}$ ,  $n \geq 1$  ( $S_0 = 0$ ). As discussed in Section 8.4, if  $S_n$  defines a martingale with respect to  $\mathcal{D}_n$  then by construction  $Y_n$  defines an orthogonal process and thus, assuming a bounded variance for  $Y_n$ , the above limit theorems can go through with minor modifications.

#### *WLLN for martingales*

Let  $\{X_n, n \geq 1\}$  be a sequence of r.v.'s with respect to the increasing sequence of  $\sigma$ -fields  $\{\mathcal{D}_n, n \geq 1\}$  such that  $E(|X|) < \infty$  and  $P(|X_n| > x) \leq cP(|X| > x)$  for  $x \geq 0$  and  $n \geq 1$ ,  $c$ -constant (i.e. all  $X_i$ 's are bounded by some r.v.  $X$ ). Then

$$\lim_{n \rightarrow \infty} Pr\left(\left|\frac{S_n}{n}\right| < \varepsilon\right) = 1, \quad (9.23)$$

where  $S_n = \sum_{i=1}^n Y_i$  is a martingale with respect to  $\mathcal{D}_n$ ,  $n \geq 1$ , and

$$Y_i \equiv X_i - E(X_i / \mathcal{D}_{i-1}), \quad i = 1, 2, \dots \quad (9.24)$$

An equivalent way to state the WLLN is

$$\frac{1}{n} \sum_{i=1}^n [X_i - E(X_i / \mathcal{D}_{i-1})] \xrightarrow{P} 0. \quad (9.25)$$

#### *SLLN for martingales*

For the martingale  $\{X_n, \mathcal{D}_n, n \geq 1\}$  satisfying the assumptions of the WLLN if the sequences  $\{X_n, n \geq 1\}$  and  $\{E(X_n / \mathcal{D}_{n-1}), n \geq 1\}$  are stationary, then

$$\frac{1}{n} \sum_{i=1}^n [X_i - E(X_i / \mathcal{D}_{i-1})] \xrightarrow{a.s.} 0. \quad (9.26)$$

This result shows clearly how the assumption of stationarity of  $\{X_n, n \geq 1\}$

and  $\{E(X_n/\mathcal{Q}_{n-1}), n \geq 1\}$  (see Chapter 8) can strengthen the WLLN result to that of the SLLN.

The above discussion suggests that the most important ingredients of the Bernoulli theorem are that:

- (i) we consider the probabilistic behaviour of centred r.v.'s of the form  $Z_n = S_n - np = \sum_{i=1}^n (X_i - E(X_i))$ ;
- (ii)  $\text{Var}(S_n) = O(n)$ ; and
- (iii) for  $Y_n = X_n - E(X_n)$ , the sequence  $\{Y_n, n \geq 1\}$  is a martingale difference, i.e.  $E(Y_n/\sigma(Y_{n-1}, \dots, Y_1)) = 0, n \geq 1$ .

This suggests that martingales provide a very convenient framework for these limit theorems because by definition they are r.v.'s with respect to an increasing sequence of  $\sigma$ -fields and under some general conditions they converge to some r.v. as  $n \rightarrow \infty$ . The latter being of great importance when convergence to a non-degenerate r.v. is needed. Moreover, for any martingale sequence  $\{X_n, \mathcal{Q}_n, n \geq 1\}$  the martingale differences sequence  $\{Y_n, n \geq 1\}$  defines a martingale orthogonal sequence of r.v.'s which can help us ensure (ii) above.

*Remark:* The SLLN is sometimes credited as providing a mathematical foundation for the frequency approach to probability. This is, however, erroneous because the definition is rendered circular given that we need a notion of probability to define the SLLN in the first place.

### 9.3 The central limit theorem

As with the WLLN and SLLN, it was realised that LT2 was not contributing in any essential way to the De Moivre–Laplace theorem and the literature considered sequences of r.v.'s with restrictions on the first few moments. Let  $\{X_n, n \geq 1\}$  be a sequence of r.v.'s and  $S_n = \sum_{i=1}^n X_i$ , the CLT considers the limiting behaviour of

$$Y_n = \frac{S_n - E(S_n)}{\sqrt{[\text{Var}(S_n)]}}, \quad (9.27)$$

which is a normalised version of  $S_n - E(S_n)$ , the subject matter of the WLLN and SLLN.

#### Lindeberg–Levy theorem

Let  $\{X_n, n \geq 1\}$  be a sequence of IID r.v.'s such that  $E(X_i) = \mu$ ,  $\text{Var}(X_i) = \sigma^2 < \infty$  for all  $i$ . Then for  $F_n(y)$  the DF of  $Y_n$ ,

$$\lim_{n \rightarrow \infty} F_n(y) = \lim_{n \rightarrow \infty} P(Y_n \leq y) = \int_{-\infty}^y \frac{1}{\sqrt{(2\pi)}} \exp\{-\frac{1}{2}u^2\} du. \quad (9.28)$$

*Liapunov's theorem*

Let  $\{X_n, n \geq 1\}$  be a sequence of independent r.v.'s with

$$E(X_i) = \mu_i, \quad \text{Var}(X_i) = \sigma_i^2 < \infty, \quad E(|X_i|^{2+\delta}) < \infty, \quad \delta > 0.$$

Define

$$c_n = \left( \sum_{i=1}^n \sigma_i^2 \right)^{\frac{1}{2}},$$

then if

$$\lim_{n \rightarrow \infty} \left( \frac{1}{c_n^{2+\delta}} \sum_{i=1}^n E(|X_i - \mu_i|^{2+\delta}) \right) = 0, \quad (9.29)$$

$$\lim_{n \rightarrow \infty} F_n(y) = \int_{-\infty}^y \frac{1}{\sqrt{(2\pi)}} \exp\{-\frac{1}{2}u^2\} du. \quad (9.30)$$

Liapunov's theorem is rather restrictive because it requires the existence of moments higher than the second. A more satisfactory result providing both necessary and sufficient conditions is the next theorem; Lindeberg in 1923 established the 'if' part and Feller in 1935 the 'only if' part.

*Lindeberg–Feller theorem*

Let  $\{X_n, n \geq 1\}$  be a sequence of independent r.v.'s with distribution functions  $\{F_n(x), n \geq 1\}$  such that

$$\begin{aligned} (i) \quad & E(X_i) = \mu_i \\ (ii) \quad & \text{Var}(X_i) = \sigma_i^2 < \infty, \quad i = 1, 2, \dots \end{aligned} \Bigg\}. \quad (9.31)$$

Then the relations

$$(a) \quad \lim_{n \rightarrow \infty} \max_{1 \leq i \leq n} \frac{\sigma_i}{c_n} = 0, \quad \text{where } c_n = \left( \sum_{i=1}^n \sigma_i^2 \right)^{\frac{1}{2}}; \quad (9.32)$$

$$(b) \quad \lim_{n \rightarrow \infty} F_n(y) = \int_{-\infty}^y \frac{1}{\sqrt{(2\pi)}} \exp\{-\frac{1}{2}u^2\} du, \quad (9.33)$$

hold true, if and only if,

$$\lim_{n \rightarrow \infty} \left( \frac{1}{c_n^2} \sum_{i=1}^n \int_{|x - \mu_i| > \varepsilon c_i} (x - \mu_i)^2 dF_i(x) \right) = 0, \quad (9.34)$$

i.e.

$$\sum_{i=1}^n \int_{|x - \mu_i| > \varepsilon c_i} (x - \mu_i)^2 dF_i(x) = O(c_n^2) \quad \text{for all } \varepsilon > 0. \quad (9.35)$$

The necessary and sufficient condition is known as the *Lindeberg condition* and provides an intuitive insight into 'what really gives rise to the result'.

Given that

$$\begin{aligned} & \frac{1}{c_n^2} \sum_{i=1}^n \int_{|x-\mu_i|>\varepsilon c_i} (x-\mu_i)^2 dF_i(x) \\ & \geq \varepsilon^2 \sum_{i=1}^n Pr(|X_i - \mu_i| \geq \varepsilon c_i) \\ & \geq \varepsilon^2 \max_{1 \leq i \leq n} Pr(|X_i - \mu_i| \geq \varepsilon c_i), \end{aligned} \quad (9.36)$$

this shows that the heart of the CLT is the condition that *no one r.v. dominates the sequence of sums*, that is, each  $(X_i - \mu_i)/\sigma_i$  is small relative to the sum  $[S_n - E(S_n)]/c_n$  as  $n$  increases. The Liapunov condition can be deduced from the Lindeberg condition and thus it achieves the same effect. Hence the CLT refers to the distributional behaviour of the summation of an increasing number of r.v.'s which individually do not exert any significant effect on the behaviour of the sum. An analogy can be drawn from economic theory where under the assumptions of perfect competition (no individual agent dominates the aggregate) we can prove the existence of a general equilibrium. A more pertinent analogy can be drawn between the CLT and the theory of gas in physics. A particular viewpoint in physics considers a gas as consisting of an enormous number of individual particles in continuous but chaotic motion. One can say nothing about the behaviour of individual particles in relation to their position or velocity but we can determine (at least probabilistically) the behaviour of a large group of them.

Fig. 9.1 illustrates the CLT in the case where  $X_1, \dots, X_n$  are IID uniformly distributed r.v.'s, i.e.  $X_i \sim U(-1, 1)$ ,  $i = 1, 2, \dots, n$ , and  $f_n(y)$  represents the density function of  $Y_n = X_1 + X_2 + \dots + X_n$ .

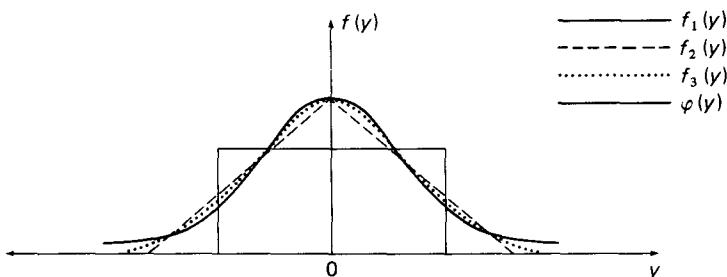


Fig. 9.1. Illustrating the CLT using the density function of  $Y = \sum_{i=1}^n X_i$  where the  $X_i$ 's are uniformly distributed for  $n = 1, 2, 3$ .

Returning to the Bernoulli and De Moivre–Laplace theorems, we can see that the important ingredients were the same in both cases and both families of limit theorems refer to the behaviour of the sum of a sequence of r.v.'s in a different probabilistic sense. The WLLN referred to  $[(S_n/n) \xrightarrow{P} p]$ , the SLLN to  $[(S_n/n) \xrightarrow{\text{a.s.}} p]$  and the CLT to

$$\left[ \frac{p(1-p)}{n} \right]^{-\frac{1}{2}} \left[ \frac{S_n}{n} - p \right] \rightarrow Z \sim N(0, 1). \quad (9.37)$$

Let us consider the relationship between these limit theorems in the particular case of the binomial distribution. From the CLT we know that

$$Pr\left(a < \frac{S_n - np}{\sqrt{[np(1-p)]}} \leq b\right) \simeq \int_a^b \frac{1}{\sqrt{(2\pi)}} \exp\left\{-\frac{u^2}{2}\right\} du \quad (9.38)$$

(' $\simeq$ ' reads 'approximately equal'). In order to see how good the approximation is let us take  $Pr(6 < S_n \leq 8)$ ,  $n = 10$ ,  $p = \frac{1}{2}$ . From the binomial tables we get  $\sum_{k=6}^8 {}^{10}_k (0.5)^k (0.5)^{10-k} = 0.3662$ . The normal approximation to this probability takes the form

$$Pr(6 < S_n \leq 8) \simeq \Phi\left[\frac{8 + \frac{1}{2} - np}{\sqrt{(np)(1-p)}}\right] - \Phi\left[\frac{6 - \frac{1}{2} - np}{\sqrt{(np)(1-p)}}\right] = \Phi(2.21) - \Phi(0.316),$$

where  $\Phi(\cdot)$  refers to the normal cumulative distribution function. It must be noted that  $Pr(S_n \leq b)$  is approximated by  $F\{(b + \frac{1}{2} - np)/\sqrt{[np(1-p)]}\}$  rather than  $F\{(b - np)/\sqrt{[np(1-p)]}\}$  in order to improve the approximation by bridging the discontinuity between integers. From the normal tables we get  $\Phi(2.21) - \Phi(0.316) = 0.9866 - 0.6239 = 0.3627$  which is a very good approximation of the exact binomial probability for  $n$  as small as 10.

Using the above results we can deduce that

$$Pr\left(\left|\frac{S_n}{n} - p\right| < \varepsilon\right) = \int_{-\varepsilon^*}^{\varepsilon^*} \frac{1}{\sqrt{(2\pi)}} \exp\left\{-\frac{u^2}{2}\right\} du, \quad \varepsilon^* = \left[ \frac{p(1-p)}{n} \right]^{\frac{1}{2}}.$$

The WLLN tells us that the RHS goes to one as  $n \rightarrow \infty$ . To see this let us take  $p = \frac{1}{2}$ ,  $\varepsilon = 2$  and  $n = 100, 200, 500$ :

$$Pr\left(\left|\frac{S_{100}}{100} - p\right| < 2\right) = 0.920, \quad Pr\left(\left|\frac{S_{200}}{200} - p\right| < 2\right) = 0.944,$$

$$Pr\left(\left|\frac{S_{500}}{500} - p\right| < 2\right) = 0.965.$$

From the above example of the normal distribution providing an approximation to the binomial distribution we saw that for  $n$  as small as 10 it was a very good approximation. This is, however, by no means the rule. In this case it arose because  $p = \frac{1}{2}$ . For values of  $p$  near zero or one the approximation is much worse for the same  $n$ . In general the accuracy of the approximation depends on  $n$  as well as the unknown parameter  $\theta$ . This presents us with the problem of assessing how good the approximation is for a particular value of  $n$  and a range for  $\theta$ . This problem will be considered further in Chapter 10 where the additional question of improving the approximation will also be considered.

Although the CLT refers to the convergence in distribution of the standardised sum  $S_n^* = (S_n - m_n)/c_n$ , where  $m_n = E(S_n)$  and  $c_n = \sqrt{[Var(S_n)]}$  it is common place in practice to refer to  $S_n$  being asymptotically normally distributed with mean  $m_n$  and variance  $c_n^2$  and to denote this by

$$\underset{x}{\sim} N(m_n, c_n^2). \quad (9.39)$$

Strictly speaking, such a statement is incorrect, but it can be justified in the following sense:

for large  $n$  probabilities of the form  $Pr(S_n \leq a)$  can be approximated by  $\Phi[(a - m_n)/c_n]$  since the approximation error  $Pr(S_n \leq a) - \Phi[(a - m_n)/c_n]$  goes to zero uniformly on  $\mathbb{R}$  as  $n \rightarrow \infty$ .

(See Ash (1972).)

The CLT can be extended to a sequence of random vectors  $\{\mathbf{X}_n, n \geq 1\}$  where  $\mathbf{X}_n$  is a  $k \times 1$  vector.

#### Lindeberg–Feller CLT

Let  $\{\mathbf{X}_n, n \geq 1\}$  be a sequence of  $k \times 1$  independent random vectors with  $E(\mathbf{X}_i) = \boldsymbol{\mu}_i$  and  $\text{Cov}(\mathbf{X}_i) = \Sigma_i$ ,  $i = 1, 2, \dots$ , and distribution functions  $\{F_n, n \geq 1\}$  such that:

$$(i) \lim_{n \rightarrow \infty} (\bar{\Sigma}_n) = \Sigma \neq 0, \quad \text{where } \bar{\Sigma}_n = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i; \quad (9.40)$$

$$(ii) \frac{1}{n} \sum_{i=1}^n \int_{\|\mathbf{x} - \boldsymbol{\mu}_i\| > \varepsilon \sqrt{n}} \|\mathbf{x} - \boldsymbol{\mu}\|^2 dF_i(\mathbf{x}) \rightarrow 0$$

as  $n \rightarrow \infty$  and each  $\varepsilon > 0$ ;  $(9.41)$

then

$$n^{-\frac{1}{2}} \sum_{i=1}^n (\mathbf{X}_i - \boldsymbol{\mu}_i) \underset{x}{\sim} N(\mathbf{0}, \Sigma). \quad (9.42)$$

In practice this result is demonstrated by showing that for any fixed

vector  $\mathbf{c} \equiv (c_1, c_2, \dots, c_k)'$  where  $\mathbf{c} \neq \mathbf{0}$

$$(nc'\Sigma_n\mathbf{c})^{-\frac{1}{2}} \left[ \mathbf{c}' \left( \sum_{i=1}^n (\mathbf{X}_i - \boldsymbol{\mu}_i) \right) \right] \xrightarrow{z} N(0, 1). \quad (9.43)$$

Since

$$\mathbf{c}'\Sigma_n\mathbf{c} \rightarrow \mathbf{c}'\Sigma\mathbf{c} \neq 0, \quad \mathbf{c}' \left[ n^{-\frac{1}{2}} \sum_{i=1}^n (\mathbf{X}_i - \boldsymbol{\mu}_i) \right] \xrightarrow{z} N(0, \mathbf{c}'\Sigma\mathbf{c})$$

for all  $\mathbf{c} \neq 0$ . (9.44)

Then, using the Cramer–Wold theorem (see Chapter 10) we can deduce the CLT result stated above.

As in the case of the other limit theorems (WLLN, SLLN) the CLT can be easily extended to the martingale case.

#### *CLT for martingales*

Let  $\{S_n, \mathcal{D}_n, n \geq 1\}$  be a martingale and define the martingale differences  $X_n = S_n - S_{n-1}$ , that is,  $\{X_n, n \geq 1\}$  is a sequence of r.v.'s with  $E(X_n/\mathcal{D}_{n-1}) = 0$ ,  $n = 1, \dots$ , such that:

$$(i) \quad \lim_{n \rightarrow \infty} \frac{1}{c_n^2} \sum_{i=1}^n E(|E(X_i^2/\mathcal{D}_{i-1}) - \sigma_i^2|) = 0; \quad (9.45)$$

$$(ii) \quad \lim_{n \rightarrow \infty} \frac{1}{c_n^2} \sum_{i=1}^n \int_{(|x| > \epsilon c_i)} x^2 dF_i(x) = 0; \quad (9.46)$$

$$c_n^2 = \sum_{i=1}^n \sigma_i^2, \quad \sigma_i = E(X_i^2), \quad \text{then } \left( \frac{1}{c_n} \sum_{i=1}^n X_i \right) \xrightarrow{z} N(0, 1). \quad (9.47)$$

This theorem is a direct extension of the Lindeberg–Feller theorem. It is important to note that (i) and (ii) ensure that the summations involved are of smaller order of magnitude than  $c_n^2$ .

#### 9.4\* Limit theorems for stochastic processes

The purpose of this section is to consider briefly various extensions of the limit theorems discussed above to some interesting cases where  $\{X_n, n \geq 1\}$  is a stochastic process satisfying certain restrictions (see Chapter 8). The first step towards generalising the limit theorems to dependent r.v.'s has already been considered above for the case where  $\{X_n, n \geq 1\}$  is a martingale relative to the increasing sequence of  $\sigma$ -fields  $\{\mathcal{D}_n, n \geq 1\}$ .

Another interesting form of a stochastic process is the *m-dependent* process (see Chapter 8). For an *m-dependent* zero mean stochastic process  $\{X_n, n \geq 1\}$  with finite third moments,  $E(|X_n|^3) < K$  for all  $n \geq 1$  and some

constant  $K$ , it can be shown that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \xrightarrow{D} Z \sim N(0, \sigma^2), \quad \text{if } \sigma^2 = \lim_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{k=1}^n \sigma_{k+i}^2 \right) < \infty, \quad (9.48)$$

$$\sigma_i^2 = 2 \sum_{j=0}^{m-1} \text{Cov}(X_{i+j}, X_{i+m}) + \text{Var}(X_{i+m}). \quad (9.49)$$

The importance of martingales and  $m$ -dependent processes lies with the fact that stationary-ergodic and mixing stochastic processes behave asymptotically like martingale differences and  $m$ -dependent processes, respectively. Hence, the limit theorems for martingales and  $m$ -dependent processes can be extended to stationary and mixing processes when certain restrictions related to their homogeneity and memory are imposed.

#### *SLLN for mixing processes*

Let  $\{X_n, n \geq 1\}$  be a mixing stochastic process which is either:

$$(a) \phi(m) = O(m^{-\tau}), \quad \tau > (r/(2r-1)); \quad (9.50)$$

or

$$(b) \alpha(m) = O(m^{-\tau}), \quad \tau > (r/(r-1)), \quad (9.51)$$

such that  $|X_n| \leq Z_n$  and  $E|Z_n|^{r+\delta} \leq K < \infty$ ,  $n \geq 1$  for  $r \geq 1$  and any  $\delta > 0$  (i.e.  $X_n$  is dominated by  $Z_n$ ,  $n \geq 1$ ) then

$$\left[ \frac{1}{n} \sum_{i=1}^n [X_i - E(X_i)] \right] \xrightarrow{\text{a.s.}} 0 \quad (9.52)$$

(see White and Domowitz (1984)).

The value of  $r$  is the above SLLN determines the highest moment assumed to exist and the same value restricts the memory of  $\{X_n, n \geq 1\}$ . For  $\{X_n, n \geq 1\}$  an independent process  $r = 1$  and  $\phi(m)$  dies out exponentially.

#### *CLT for mixing processes*

Let  $\{X_n, n \geq 1\}$  be a mixing process satisfying the restrictions imposed for the SLLN to hold. In addition let us assume that:

$$(i) \quad E(X_n) = 0, \quad n \geq 1; \quad (9.53)$$

$$(ii) \quad E(|X_n|^{2r}) \leq K < \infty, \quad n \geq 1, \quad r > 1; \quad (9.54)$$

$$(iii) \quad \text{for } S_n(\tau) = n^{-\frac{1}{2}} \sum_{i=\tau+1}^{n+\tau} X_i, \text{ there exists } V \text{ finite and non-zero such that } [E(S_n(\tau))^2 - V] \rightarrow 0 \text{ for all } \tau \text{ as } n \rightarrow \infty, \text{ then}$$

$$(nV)^{-\frac{1}{2}} S_n(0) \xrightarrow{\text{a.s.}} N(0, 1) \quad (9.55)$$

(see White and Domowitz (1984)).

The importance of the above limit theorems for mixing processes becomes more apparent in view of the following result:

If  $\{X_n, n \geq 1\}$  is a mixing process ( $\phi(m)$  or  $\alpha(m)$ ) then any Borel function  $Y_n = g_n(X_n, \dots, X_{n-k})$  is also mixing. Moreover, if  $X_n$  is  $O(m^{-\tau})$  then  $Y_n$  is also  $O(m^{-\tau})$ ,  $\tau > 0$ .

This result is of considerable interest in statistical inference where asymptotic results for functions of stochastic processes are at a premium.

For stationary stochastic processes which are also ergodic several limit theorems can be proved.

*SLLN for stationary, ergodic processes*

Let  $\{X_n, n \geq 1\}$  be a stationary and ergodic process such that  $E|X_n| < \infty$ ,  $n \geq 1$ , then

$$\left( \frac{1}{n} \sum_{i=1}^m X_i \right) \xrightarrow{\text{a.s.}} \mu \equiv E(X_n), \quad n \geq 1 \quad (9.56)$$

(see Stout (1974)).

*CLT for a stationary, mixing process*

Let  $\{X_n, n \geq 1\}$  be a stationary, strong mixing process such that:

$$(i) \quad E(X_n) = 0, \quad n \geq 1; \quad (9.57)$$

$$(ii) \quad E|X_n|^{2+\delta} < \infty \quad \text{for } \delta > 0, \quad n \geq 1; \quad (9.58)$$

$$(iii) \quad \lim_{n \rightarrow \infty} \left( \frac{1}{n} E(S_n^2) \right) = \sigma^2 > 0 \quad \text{where } S_n = \sum_{i=1}^n X_i. \quad (9.59)$$

If

$$\sum_{m=1}^{\infty} [\alpha(m)]^{[(1+\delta)/(2+\delta)]} < \infty \quad (9.60)$$

then

$$n^{-\frac{1}{2}} \left( \frac{S_n}{\sigma} \right) \xrightarrow{\text{a.s.}} N(0, 1) \quad (9.61)$$

(see Hall and Heyde (1980)).

Note that mixing implies ergodicity (see Chapter 8).



## 9.5 Summary

The limit theorems discussed above provide us with useful information relating to the probabilistic behaviour of a particular aggregate function, the sum, of an increasing sequence of r.v.'s, as the number of r.v.'s goes to

infinity, when certain conditions are imposed on the individual r.v.'s to ensure that 'no one dominates the sum'. The WLLN refers to the convergence in probability of  $S_n/n$ , i.e.

$$\frac{S_n}{n} \xrightarrow{P} \frac{E(S_n)}{n}.$$

The SLLN strengthens the convergence to 'almost surely', i.e.

$$\frac{S_n}{n} \xrightarrow{\text{a.s.}} \frac{E(S_n)}{n}.$$

The CLT, on the other hand, provides us with information relating to the rate of convergence. This information comes in the form of the 'appropriate' factor by which to premultiply  $S_n - E(S_n)$  so that it converges to a non-degenerate r.v. (the convergence in WLLN and SLLN is to a degenerate r.v.; a constant). This factor comes in the form of the standard deviation of  $S_n$ , i.e.

$$\left\{ \frac{1}{\sqrt{[\text{Var}(S_n)]}} \right\} (S_n - E(S_n)) \xrightarrow{D} Z \sim N(0, 1). \quad (9.62)$$

### ***Important concepts***

Convergence in probability, convergence almost surely, convergence in distribution, weak law of large numbers, strong law of large numbers, central limit theorem, Lindeberg condition.

### ***Questions***

1. Explain the statement  $\lim_{n \rightarrow \infty} Pr\{[(S_n/n) - P] < \varepsilon\} = 1$  and contrast it with  $Pr\{\lim_{n \rightarrow \infty} [(S_n/n) - p]\} = 1$ .
2. Discuss the underlying assumptions of the Bernoulli WLLN in relation to their contribution to the result.
3. Explain the difference between Chebyshev's and Markov's WLLN.
4. Whose behaviour do the WLLN and SLLN refer to?
5. Explain intuitively why a sequence of martingale differences with finite variances obeys the WLLN and SLLN.
6. Explain the Lindeberg–Feller CLT in relation to the assumptions and conclusions.
- 7\*. In the CLT why is the limit distribution a normal and not some other distribution?
8. 'All limit theorems impose conditions on the individual r.v.'s of a sequence in order to ensure that no one dominates the behaviour of the aggregate and this leads to their conclusions.' Discuss.

9. Compare the Lindeberg–Feller CLT with the one for martingales.
10. Compare Markov's condition for the WLLN with Kolmogorov's condition for the SLLN.
11. Compare the conclusions of the WLLN, SLLN and CLT.

**Exercises**

1. Which of the following sequences of independent r.v.'s satisfy the WLLN or/and the SLLN?

$$\begin{aligned}
 \text{(i)} \quad & Pr(X_n = \pm 2^n) = \frac{1}{2} \quad \left( \text{Note: } \sum_{i=1}^n i^2 = \frac{r(r^n - 1)}{(r - 1)} \right); \\
 \text{(ii)} \quad & Pr(X_n = \pm \sqrt{n}) = \frac{1}{2}; \\
 \text{(iii)} \quad & Pr(X_n = \pm 2^n) = 2^{-(2n+1)}, \quad Pr(X_n = 0) = 1 - 2^{-2n}; \\
 \text{(iv)} \quad & Pr(X_n = \pm n) = \frac{1}{2\sqrt{n}}, \quad Pr(X_n = 0) = 1 - \frac{1}{\sqrt{n}} \\
 & \left( \text{Note: } \sum_{i=1}^n i = \frac{n(n+1)}{2} \right).
 \end{aligned}$$

2. Determine the value of  $\alpha$  for which the sequence of r.v.'s  $\{X_n\}$

$$Pr(X_n = \pm n^\alpha) = \frac{1}{2}$$

satisfies the SLLN.

- 3\*. Show that for the sequence of independent r.v.'s  $\{X_n, n \geq 1\}$  with  $Pr(X_n = \pm n^\alpha) = 1/[\alpha n^{2(\alpha-1)}]$ , the Lindeberg conditions holds iff  $\alpha < \frac{3}{2}$ .

**Additional references**

Chung (1974); Cramer (1946); Feller (1968); Giri (1974); Gnedenko (1969); Loeve (1963); Pfeiffer (1978); Rao (1973); Renyi (1970); Rohatgi (1976).

## CHAPTER 10\*

---

### Introduction to asymptotic theory

---

#### 10.1 Introduction

At the heart of statistical inference lies the problem of deriving the distribution of some Borel function  $h(\cdot)$  of the random vector  $\mathbf{X} = (X_1, \dots, X_n)'$ , i.e. determine

$$F_n(y) = Pr(h(X_1, \dots, X_n) \leq y), \quad y \in \mathbb{R} \quad (10.1)$$

from the distribution of  $\mathbf{X}$ . In Chapter 6 we saw that this is by no means a trivial problem even for the simplest functions  $h(\cdot)$ . Indeed the results in this area are almost exclusively related to simple functions of normally distributed r.v.'s, most of which have been derived in Chapter 6. For more complicated functions even in the case of normality very few results are available. Given, however, that statistical inference depends crucially on being able to determine the distribution of such functions  $h(\mathbf{X})$  we need to tackle the problem somehow. Intuition suggests that the limit theorems discussed in Chapter 9, when extended, might enable us to derive *approximate solutions* to the distribution problem.

The limit theorems considered in Chapter 9 tell us that under certain conditions, which ensure that no one r.v. in a sequence  $\{X_n, n \geq 1\}$  dominates the behaviour of the sum ( $\sum_{i=1}^n X_i$ ), we can deduce that:

$$(i) \quad \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{\text{P}} \frac{1}{n} \sum_{i=1}^n E(X_i) \quad (\text{WLLN}); \quad (10.2)$$

$$(ii) \quad \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{\text{a.s.}} \frac{1}{n} \sum_{i=1}^n E(X_i) \quad (\text{SLLN}); \quad (10.3)$$

$$(iii) \quad \frac{\frac{1}{n} \sum (X_i - E(X_i))}{\left( \text{Var}\left(\frac{1}{n} \sum X_i\right) \right)^{\frac{1}{2}}} \xrightarrow{\text{D}} Z \sim N(0, 1) \quad (\text{CLT}). \quad (10.4)$$

In order to be able to extend these results to arbitrary Borel functions  $h(\mathbf{X})$ , not just  $\sum_{i=1}^n X_i$ , we firstly need to extend the various modes of convergence (convergence in probability, almost sure convergence, convergence in distribution) to apply to any sequence of r.v.'s  $\{X_n, n \geq 1\}$ .

The various modes of convergence related to the above limit theorems are considered in Section 10.2. The main purpose of this section is to relate the various mathematical notions of convergence to the probabilistic convergence modes needed in asymptotic theory. One important mode of convergence not encountered in the context of the limit theorems is 'convergence in the  $r$ th mean', which refers to convergence of moments. Section 10.3 discusses various concepts related to the convergence of moments such as asymptotic moments, limits of moments and probability limits in an attempt to distinguish between these concepts often confused in asymptotic theory. In Chapter 9 it was stressed that an important ingredient which underlies the conditions giving rise to the various limit theorems is the notion of the 'order of magnitude'. For example, the Markov condition needed for the WLLN,

$$\lim_{n \rightarrow \infty} \frac{1}{n^2} \sum_{k=1}^n \text{Var}(X_k) = 0, \quad (10.5)$$

is a restriction on the order of magnitude of  $\text{Var}(S_n)$  of the form

$$\text{Var}(S_n) = o(n^2). \quad (10.6)$$

In Section 10.4 the 'big O', 'little o' notation is considered in some detail as a prelude to Sections 10.5 and 10.6. The purpose of Section 10.5 is to consider the question of extending the limit theorems of Chapter 9 from  $(\sum_{i=1}^n X_i)$  to more general functions of  $(X_1, X_2, \dots, X_n)$  such as  $\sum_{i=1}^n X_i^r$ ,  $r \geq 1$ . This is indeed the main aim of asymptotic theory.

Asymptotic theory results are resorted to by necessity, when finite sample results are not available in a usable form. This is because asymptotic results provide only approximations. 'How good the approximations are' is commonly unknown because we could answer such a question only when the finite result is available. But in such a case the asymptotic result is not needed! There are, however, various 'rough' error bounds which can shed some light on the magnitude of the approximation error. Moreover, it is often possible to 'improve' upon the asymptotic results using what we call

asymptotic expansions such as the Edgeworth expansion. The purpose of Section 10.6 is to introduce the reader to this important literature on error bounds and asymptotic expansions. The discussion is only introductory and much more intuitive than formal in an attempt to demystify this literature which plays an important role in econometrics. For a more complete and formal discussion see Phillips (1980), Rothenberg (1984), *inter alia*.

## 10.2 Modes of convergence

The notions of ‘limit’ and ‘convergence’ play a very important role in probability theory, not only because of the limit theorems discussed in Chapter 9 but also because they underlie some of the most fundamental concepts such as probability and distribution functions, density functions, mean, variance, as well as higher moments. This was not made explicit in Chapters 3–7 because of the mathematical subtleties involved.

In order to understand the various modes of convergence in probability theory let us begin by reminding ourselves of the notion of convergence in mathematical analysis. A sequence  $\{a_n, n \in \mathbb{N}\}$  is defined to be a function from the natural numbers  $\mathbb{N} = \{1, 2, 3, \dots\}$  to the real line  $\mathbb{R}$ .

### *Definition 1*

A sequence  $\{a_n, n \in \mathbb{N}\}$  is said to converge to a limit  $a$  if for every arbitrary small number  $\varepsilon > 0$  there corresponds a number  $N(\varepsilon)$  such that the inequality  $|a_n - a| < \varepsilon$  holds for all terms  $a_n$  of the sequence with  $n \geq N(\varepsilon)$ ; we denote this by  $\lim_{n \rightarrow \infty} a_n = a$ .

### *Example 1*

$\lim_{n \rightarrow \infty} \sqrt[n]{n} = 1$ , for any  $n > 0$ ;  $\lim_{n \rightarrow \infty} (\log_b n/n) = 0$ ,  $b > 0$ ,  $b \neq 1$ ;  $\lim_{n \rightarrow \infty} [1 + (1/n)^n] = e = 2.71828$ ;  $\lim_{n \rightarrow \infty} [(n^2 + n + 6)/(3n^2 - 2n + 2)] = \frac{1}{3}$ ;  $\lim_{n \rightarrow \infty} [\sqrt{(n+1)} - \sqrt{n}] = 0$ , for  $n \in \mathbb{N}$ .

This notion of convergence can be extended directly to any function whose domain is not necessarily  $\mathbb{N}$  but any subset of  $\mathbb{R}$ , i.e.  $h(x): \mathbb{R} \rightarrow \mathbb{R}$ . The way this is done is to allow the variable  $x$  to define a sequence of numbers  $\{x_n, n \in \mathbb{N}\}$  converging to some limit  $x_0$  and consider the sequence  $\{h(x_n), n \in \mathbb{N}\}$  as  $x_n \rightarrow x_0$  and denote it by  $\lim_{x \rightarrow x_0} h(x) = l$ .

### *Definition 2*

A function  $h(x)$  is said to converge to a limit  $l$  as  $x \rightarrow x_0$ , if for every

**Introduction to asymptotic theory**

$\varepsilon > 0$ , however small, there exists a number  $\delta(\varepsilon) > 0$  such that

$$|h(x) - l| < \varepsilon$$

holds for every  $x$  satisfying the condition  $0 < |x - x_0| < \delta(\varepsilon)$ .

*Example 2*

For  $h(x) = e^x$ ,  $\lim_{x \rightarrow -\infty} h(x) = 0$  and for the polynomial function

$$h(x) = a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n, \quad \lim_{x \rightarrow 0} h(x) = a_n.$$

Note that the condition  $0 < |x - x_0| < \delta(\varepsilon)$  excludes the point  $x = x_0$  in the above definition and thus for  $h(x) = (x^2 - 9)/(x - 3)$ ,  $\lim_{x \rightarrow 3} h(x) = 6$ , even though  $h(x)$  is not defined at  $x = 3$ .

For

$$h(x) = a^x, \quad a > 0, \quad \lim_{x \rightarrow 0} h(x) = 1,$$

$$h(x) = (1+x)^{1/x}, \quad \lim_{x \rightarrow 0} h(x) = e,$$

$$h(x) = \left(1 + \frac{a}{x}\right)^x, \quad \lim_{x \rightarrow 0} h(x) = e^a,$$

$$h(x) = [\log_e(1+x)/x], \quad \lim_{x \rightarrow 0} h(x) = 1.$$

Using the notion of convergence of a function to a limit we can define the *continuity of a function at a point*.

*Definition 3*

A function  $h(x)$ , defined over some interval  $D(h) \subseteq \mathbb{R}$ ,  $x_0 \in D(h)$  is said to be **continuous at the point  $x_0$**  if for each  $\varepsilon > 0$ , there exists a  $\delta(\varepsilon) > 0$  such that

$$|h(x) - h(x_0)| < \varepsilon$$

for every  $x$  satisfying the restriction  $|x - x_0| < \delta(\varepsilon)$ . We denote this by  $\lim_{x \rightarrow x_0} h(x) = h(x_0)$ . A function  $h(x)$  is said to be **continuous** if it is continuous at every point of its domain,  $D$ .

*Example 3*

The functions  $h(x) = ax + b$  and  $h(x) = e^x$  are continuous for all  $x \in \mathbb{R}$  (verify!).

In the case of a general function we can define the sequence  $\{h_n(x), n \in \mathbb{N}\}$  and consider its behaviour for each  $x \in A$  where  $A$  is a subset of  $D(h)$ .

*Definition 4*

A sequence of functions  $\{h_n(x), n \in \mathbb{N}\}$  with common domain  $D$  is said to **converge pointwise** to  $h(x)$  on  $A \subseteq D$  if for each  $\varepsilon > 0$ , there exists a  $N(\varepsilon, x)$  such that if  $n > N(\varepsilon, x)$ , then

$$|h_n(x) - h(x)| < \varepsilon \quad \text{holds for all } x \in A.$$

*Example 4*

For

$$h_n(x) = \sum_{k=0}^n \frac{x^k}{k!}, \quad \lim_{n \rightarrow \infty} h_n(x) = e^x \quad \text{for all } x \in \mathbb{R}.$$

In the case where  $N(\varepsilon, x)$  does not depend on  $x$  (only on  $\varepsilon$ ) then  $\{h_n(x), n \in \mathbb{N}\}$  is said to *converge uniformly on A*. The importance of uniform convergence stems from the fact that if each  $h_n(x)$  in the sequence is continuous and  $h_n(x)$  converges uniformly to  $h(x)$  on  $D$  then the limit  $h(x)$  is also continuous. That is, if

$$\lim_{x \rightarrow x_0} h_n(x) = h_n(x_0) \quad \text{for } x_0 \in D \tag{10.7}$$

and

$$\lim_{n \rightarrow \infty} h_n(x) = h(x) \quad \text{for all } x \in D, \text{ uniformly,} \tag{10.8}$$

then

$$\lim_{x \rightarrow x_0} h(x) = h(x_0). \tag{10.9}$$

With the above notions of continuity and limit in mathematical analysis in mind let us consider the question of convergence in the context of the probability spaces  $(S, \mathcal{F}, P(\cdot))$  and  $(\mathbb{R}, \mathcal{B}, P_x(\cdot))$ . Given that a random variable  $X(\cdot)$  is a function from  $S$  to  $\mathbb{R}$  we can define *pointwise* and *uniform convergence* on  $S$  for the sequence  $\{X_n(s), n \in \mathbb{N}\}$  by

$$|X_n(s) - X(s)| < \varepsilon \quad \text{for } n > N(\varepsilon, s), \quad s \in S \tag{10.10}$$

and

$$|X_n(s) - X(s)| < \varepsilon \quad \text{for } n > N(\varepsilon), \quad s \in S, \tag{10.11}$$

respectively. These notions of convergence are of little interest because the probabilistic structure of  $\{X_n(s), n \in \mathbb{N}\}$  is ignored. Although the probability set functions  $P(\cdot)$  and  $P_x(\cdot)$  do not come into the definition of a

random variable they play a crucial role in its behaviour. If we take its probabilistic structure into consideration both of the above forms of convergence are much too strong because they imply that for  $n > N$

$$|X_n(s) - X(s)| < \varepsilon \quad \text{whatever the outcome } s \in S. \quad (10.12)$$

The form of probabilistic convergence closer to this is the *almost sure convergence* which allows for convergence of  $X_n(s)$  to  $X(s)$  for all  $s$  except of some  $s$ -set  $A \subseteq S$  for which  $P(A) = 0$ ;  $A$  is said to be a set of probability zero. The term almost sure is used to emphasise the convergence on  $S - A$  not the whole of  $S$ .

#### *Definition 5*

*A sequence of r.v.'s  $\{X_n(s), n \in \mathbb{N}\}$  is said to converge **almost surely** (a.s.) to a r.v.  $X(s)$ , denoted by  $X_n \xrightarrow{\text{a.s.}} X$ , if*

$$\blacktriangleright \quad \Pr\left(s: \lim_{n \rightarrow \infty} X_n(s) = X(s)\right) = 1. \quad (10.13)$$

*An equivalent way of defining almost sure convergence is by*

$$\blacktriangleright \quad \lim_{n \rightarrow \infty} \Pr(s: |X_m(s) - X(s)| < \varepsilon, \text{ all } m \geq n) = 1 \quad (10.14)$$

*(see Chung (1974)). The almost sure convergence is the mode of convergence associated with the strong law of large numbers (SLLN).*

Another mode of convergence not considered in relation to the limit theorems (see Chapter 9) is that of *convergence in rth mean*.

#### *Definition 6*

*Let  $\{X_n(s), n \in \mathbb{N}\}$  be a sequence of r.v.'s such that  $E(|X_n|^r) < \infty$  for all  $n \in \mathbb{N}$  and  $E(|X|^r) < \infty$  for  $r > 0$ , then the sequence converges to  $X$  in **rth mean**, denoted by  $X_n \xrightarrow{r} X$ , if*

$$\blacktriangleright \quad \lim_{n \rightarrow \infty} E(|X_n - X|^r) = 0. \quad (10.15)$$

*Of particular interest in what follows is the convergence in mean ( $r = 1$ ) and mean square ( $r = 2$ ).*

A weaker mode of convergence related to the weak law of large numbers (WLLN) is that of convergence in probability.

*Definition 7*

A sequence of r.v.'s  $\{X_n(s), n \in \mathcal{N}\}$  is said to **converge in probability** to a r.v.  $X$ , denoted by  $X_n \xrightarrow{P} X$ , if

$$\blacktriangleright \quad \lim_{n \rightarrow \infty} Pr(s: |X_n(s) - X(s)| < \varepsilon) = 1. \quad (10.16)$$

The relationship between convergence almost surely and in probability can be deduced by comparing (16) with (14).

The mode of convergence related to the central limit theorem (CLT) is that of convergence in distribution.

*Definition 8*

A sequence of r.v.'s  $\{X_n(s), n \in \mathcal{N}\}$  with distribution functions  $\{F_n(x), n \in \mathcal{N}\}$  is said to **converge in distribution** to  $X(s)$ , denoted by  $X_n \xrightarrow{D} X$ , if

$$\blacktriangleright \quad \lim_{n \rightarrow \infty} F_n(x) = F(x) \quad (10.17)$$

at every continuity point  $x$  of  $F(x)$ .

This is nothing more than the pointwise convergence of a sequence of functions considered above. In the case where the *convergence* is also *uniform* then  $F(x)$  is continuous and vice versa. It is important, however, to note that  $F(x)$  in (17) might not be a proper distribution function (see Chapter 4).

Without any further restrictions on the sequence of r.v.'s  $\{X_n(s), n \in \mathcal{N}\}$  the above four modes of convergence are related as shown in Fig. 10.1. As we can see, convergence in distribution is the weakest mode of convergence being implied by all three other modes. Moreover, almost sure and  $r$ th mean convergence are not directly related but they both imply convergence in probability. In order to be able to relate almost sure and  $r$ th mean convergence we need to impose some more restrictions on the sequence  $\{X_n(s), n \in \mathcal{N}\}$ , such as the existence of moments up to order  $r$ .

The implication  $\xrightarrow{\text{a.s.}} \Rightarrow \xrightarrow{P} \xrightarrow{r}$  stems from the fact that (14) is a stronger form of

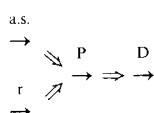


Fig. 10.1

convergence than (16), which holds for all  $n \geq m$ . The implication  $\xrightarrow{r} \Rightarrow \xrightarrow{P}$  is based on the inequality

$$E(|X_n - X|^r) \geq \varepsilon^r Pr(|X_n - X| > \varepsilon), \quad \varepsilon > 0, \quad (10.18)$$

which implies that  $\underset{P}{Pr}(|X_n - X| > \varepsilon) \leq \varepsilon^{-r} E(|X_n - X|^r) \rightarrow 0$  as  $n \rightarrow \infty$ . The implication  $\xrightarrow{P} \Rightarrow \xrightarrow{r}$  is rather obvious in the case where  $F(x)$  is a proper distribution function because for every  $\varepsilon > 0, \delta > 0$ , there exists  $N$  so that for all  $n \geq N$ ,  $Pr(|X_n - X| > \varepsilon) < \delta$ , and thus

$$F_n(x - \varepsilon) - \delta \leq F(x) \leq F_n(x + \varepsilon) + \delta,$$

implying that as  $\varepsilon, \delta \rightarrow 0$ ,

$$X_n \xrightarrow{D} X. \quad (10.19)$$

The reverse implication  $X_n \xrightarrow{D} X \Rightarrow X_n \xrightarrow{P} X$  holds only in the case where  $X$  is a *constant*.

In order to be able to establish the result  $\xrightarrow{P} \Rightarrow \xrightarrow{r}$  we need to ensure that the convergence in probability is 'sufficiently fast'. If

$$\sum_{n=1}^{\infty} Pr(|X_n - X| > \varepsilon) < \infty \quad \text{for every } \varepsilon > 0, \quad \xrightarrow{P} \Rightarrow \xrightarrow{a.s.}. \quad (10.20)$$

Moreover, a similar condition on the convergence in  $r$ th mean implies convergence almost surely. In particular, if

$$\sum_{n=1}^{\infty} E(|X_n - X|^r) < \infty, \quad \text{then } X_n \xrightarrow{a.s.} X. \quad (10.21)$$

In order to go from convergence in probability or almost sure convergence to  $r$ th mean convergence we need to ensure that the sequence of r.v.'s  $\{X_n, n \in \mathbb{N}\}$  is bounded and the moments up to order  $r$  exist. In particular if  $X_n \xrightarrow{P} X$  and conditions (i)–(iii) above hold, then  $X_n \xrightarrow{r} X$  (see Serfling (1980)).

An important property of  $r$ th mean convergence is that if  $X_n \xrightarrow{r} X$  for some  $r \geq 1$  then  $X_n \xrightarrow{l} X$  for  $0 < l < r$ . For example, mean square convergence implies mean convergence. This is related to the result that if

$$E(|X_n|^r) < \infty \quad \text{then } E(|X_n|^l) < \infty \quad \text{for } 0 < l < r. \quad (10.22)$$

That is, if the  $r$ th moment exists (is bounded) then all the moments of order less than  $r$  also exist. This is the reason why when we assume that

$\text{Var}(X_n) < \infty$  we do not need to add that  $E(X_n) < \infty$ , given that it is always implied.

In applying asymptotic theory we often need to extend the above convergence results to transformed sequences of random vectors  $\{g(\mathbf{X}_n), n \in \mathbb{V}\}$ . The above convergence results are said to hold for a random vector sequence  $\{\mathbf{X}_n, n \in \mathbb{V}\}$  if they hold for each component  $X_{in}, i = 1, 2, \dots, k$  of  $\mathbf{X}_n$ .

*Lemma 10.1*

Let  $\{X_n, n \in \mathbb{V}\}$  be a random vector sequence and  $g(\cdot): \mathbb{R}^k \rightarrow \mathbb{R}$  a continuous function at  $\mathbf{X}$ , then:

$$(i) \quad \mathbf{X}_n \xrightarrow{\text{a.s.}} \mathbf{X} \Rightarrow g(\mathbf{X}_n) \xrightarrow{\text{a.s.}} g(\mathbf{X}); \quad (10.23)$$

$$(ii) \quad \mathbf{X}_n \xrightarrow{\text{P}} \mathbf{X} \Rightarrow g(\mathbf{X}_n) \xrightarrow{\text{P}} g(\mathbf{X}); \quad (10.24)$$

$$(iii) \quad \mathbf{X}_n \xrightarrow{\text{D}} \mathbf{X} \Rightarrow g(\mathbf{X}_n) \xrightarrow{\text{D}} g(\mathbf{X}). \quad (10.25)$$

These results also hold in the case where  $g(\cdot)$  is a *Borel function* (see Chapter 6) when certain conditions are placed on the set of discontinuities of  $g(\cdot)$ : see Mann and Wald (1943). Borel functions have a distinct advantage over continuous functions in the present context because the limit of such functions are commonly Borel functions themselves without requiring uniform convergence. Continuous functions are Borel functions but not vice versa. In order to get some idea about the generality of Borel functions note that if  $h$  and  $g$  are Borel functions then the following are also Borel functions: (i)  $ah + bg$ ,  $a, b \in \mathbb{R}$ , (ii)  $|h|$ , (iii)  $\max(h, g)$ , (iv)  $\min(h, g)$ , (v)  $h \cdot g$ .

Of particular interest in asymptotic theory are the following results.

*Lemma 10.2 (Cramer–Wold)*

The sequence of random vectors  $\{\mathbf{X}_n, n \in \mathbb{V}\}$  where  $\mathbf{X}_n \equiv (X_{1n}, X_{2n}, \dots, X_{kn})'$ , converges in distribution to the random vector  $\mathbf{X}$  with distribution function  $F(\mathbf{x})$  if for any real constants  $c_1, \dots, c_k$ ,

$$\begin{aligned} & (c_1 X_{1n} + c_2 X_{2n} + \dots + c_k X_{kn}) \\ & \xrightarrow{\text{D}} c_1 X_1 + c_2 X_2 + \dots + c_k X_k. \end{aligned} \quad (10.26)$$

*Lemma 10.3*

Let  $\{\mathbf{X}_n, \mathbf{Y}_n, n \in \mathbb{V}\}$  be a sequence of pair of random  $k \times 1$  vectors.

Then:

$$(1) \text{ If } (\mathbf{X}_n - \mathbf{Y}_n) \xrightarrow{P} \mathbf{0} \text{ and } \mathbf{X}_n \xrightarrow{D} \mathbf{X} \Rightarrow \mathbf{Y}_n \xrightarrow{D} \mathbf{Y};$$

$$(2) \text{ If } \mathbf{X}_n \xrightarrow{D} \mathbf{X} \text{ and } \mathbf{Y}_n \xrightarrow{P} \mathbf{0} \Rightarrow \mathbf{X}_n \mathbf{Y}_n \xrightarrow{P} \mathbf{0};$$

$$(3) \text{ If } \mathbf{X}_n \xrightarrow{D} \mathbf{X} \text{ and } \mathbf{Y}_n \xrightarrow{P} \mathbf{C} \text{ (constant)} \Rightarrow (\mathbf{X}_n + \mathbf{Y}_n) \xrightarrow{D} \mathbf{X} + \mathbf{C}$$

$$\Rightarrow \mathbf{Y}_n \mathbf{X}_n \xrightarrow{P} \mathbf{C} \mathbf{X};$$

$$(\text{for } \mathbf{Y}_n \text{ and } \mathbf{C} \text{ } k \times k \text{ non-singular}) \Rightarrow \mathbf{Y}_n^{-1} \mathbf{X}_n \xrightarrow{D} \mathbf{C}^{-1} \mathbf{X}.$$

### 10.3 Convergence of moments

Consider the sequence of r.v.'s  $\{X_n, n \geq 1\}$  such that

$$X_n \xrightarrow{D} X \quad (\text{i.e. } \lim_{n \rightarrow \infty} F_n(x) = F(x)), \quad (10.27)$$

where  $F_n(x)$  and  $F(x)$  refer to the cumulative distribution functions of  $X_n$  and  $X$  respectively. We define the *moments* of  $X_n$  (when they exist) by

$$E(X_n^r) = \int_{-\infty}^{\infty} x^r dF_n(x), \quad r \geq 1. \quad (10.28)$$

This definition of raw moments involves the Riemann–Stieltjer integral which is a direct generalisation of the ordinary Riemann integral. In the case where  $F_n(x)$  is a monotonically non-decreasing function of  $x$  and it has a continuous derivative (as in the case of a continuous r.v.) then  $dF_n(x)$  is equivalent to the differential  $f_n(x) dx$ ;  $f_n(x) = [dF_n(x)]/dx$  being the corresponding density function.

The *limit* of the *r*th moment ( $E(X_n^r)$ ) is defined by

$$\lim_{n \rightarrow \infty} E(X_n^r), \quad r \geq 1, \quad (10.29)$$

and it refers to the ordinary mathematical limit of the sequence  $\{E(X_n^r), n \geq 1\}$ . This limit is by no means equivalent to the *asymptotic moments* of  $X_n$  defined by

$$E(X_n^r) \equiv E(X^r) = \int_{-\infty}^{\infty} x^r dF(x), \quad r \geq 1. \quad (10.30)$$

As we can see from (30) the asymptotic moments of  $X_n$  are defined in terms

of its asymptotic distribution  $F(x)$  and not its finite sample distribution  $F_n(x)$ . In view of the fact that  $F_n(x)$  might have moments up to order  $m$  and  $F(x)$  might not (or vice versa), there is no reason why  $E(X'_n)$ ,  $\lim_{n \rightarrow \infty} E(X'_n)$  and  $E_r(X'_n)$  will be equal for all  $r \leq q$  and all  $n$ . Indeed, we can show that the limit inferior of  $E(|X_n|^r)$  for some  $r \geq 1$  provide upper bounds for the corresponding asymptotic moments. In particular:

*Lemma 10.4*

If  $X_n \xrightarrow{D} X$  then;

$$(i) \quad \liminf_{n \rightarrow \infty} E(|X_n|) \geq E(|X|);$$

$$(ii) \quad \liminf_{n \rightarrow \infty} \text{Var}(X_n) \geq \text{Var}(X)$$

(see Chung (1974)).

Under certain conditions, however, these concepts are equal.

*Lemma 10.5*

If  $X_n \xrightarrow{D} X$  and the sequence  $\{X'_n, n \geq 1\}$  is **uniformly integrable**

$$\left( \text{i.e. } \lim_{\epsilon \rightarrow \infty} \sup_n \int_{\{|X'_n| > \epsilon\}} |X'_n| dP = 0 \right), E(|X|^r) < \infty \text{ and } \lim_{n \rightarrow \infty} E(X'_n) = E(X^r).$$

*Lemma 10.6*

If  $X_n \xrightarrow{r} X$  and  $E(|X|^r) < \infty$ , then  $\lim_{n \rightarrow \infty} E(X'_n) = E(X^r)$ .

*Lemma 10.7*

If  $X_n \xrightarrow{P} X$  and  $E(|X|^r) < \infty$ ,  $\{X'_n, n \geq 1\}$  is **uniformly integrable**, then  $\lim_{n \rightarrow \infty} E(X'_n) = E(X^r)$ .

*Lemma 10.8*

If  $X_n \xrightarrow{\text{a.s.}} X$  and  $\lim_{n \rightarrow \infty} \inf E(|X_n|^r) \leq E(|X|^r)$  then

$$\lim_{n \rightarrow \infty} E(X'_n) = E(X^r).$$

(For these lemmas see Serfling (1980).) Looking at these results we can see that the important condition for the equality of the limit of the  $r$ th moment and the  $r$ th asymptotic moment is the uniform integrability of  $\{X'_n, n \geq 1\}$  which allows us to interchange limits with expectations.

Beyond the distinction between moments, limits of moments and asymptotic moments we sometimes encounter the concept of *approximate moments*.

Consider the Taylor series expansion of  $g(m_r)$ ,  $m_r = (1/n) \sum_{i=1}^n X_i^r$

$$g(m_r) = g(\mu'_r) + g^{(1)}(\mu'_r)(m_r - \mu'_r) + \frac{1}{2}g^{(2)}(\mu'_r)(m_r - \mu'_r)^2 + \dots \quad (10.31)$$

This expansion is often used to derive *approximate moments* for  $g(m_r)$ . Under certain regularity conditions (see Sargan (1974)),

$$E(g(m_r)) \simeq g(\mu'_r) + \frac{1}{2}g^{(2)}(\mu'_r)E(m_r - \mu'_r)^2, \quad (10.32)$$

$$\text{Var}(g(m_r)) \simeq [g^{(1)}(\mu'_r)]^2 \text{Var}(m_r), \quad (10.33)$$

$$\begin{aligned} E(g(m_r) - E(g(m_r)))^3 &\simeq (g^{(1)}(\mu'_r))^3 E(g(m_r) - g(\mu'_r))^3 \\ &\quad + \frac{3}{2}(g^{(1)}(\mu'_r))^2 g^{(2)}(\mu'_r) E[g(m_r) \\ &\quad - g(\mu'_r)]^4, \end{aligned} \quad (10.34)$$

where ‘ $\simeq$ ’ reads approximately equal. These moments are viewed as moments of a statistic purporting to approximate  $g(m_r)$  and under certain conditions can be treated as approximations to the moments of  $g(m_r)$  (see Sargan (1974)). Such approximations must be distinguished from  $E(X_n^r)$  as well as  $E(X^r)$ . The approximate moments derived above can be very useful in choosing the functions  $g(\cdot)$  so as to make the asymptotic results more accurate in the context of variance stabilising transformations and asymptotic expansions (see Rothenberg (1984)). In deriving the asymptotic distributions of  $g(m_r)$  only the first two moments are utilised and one can improve upon the normal approximation by utilising the above approximate higher moments in the context of asymptotic expansions. A brief introduction to asymptotic expansions is given in Section 10.6.

#### 10.4    The ‘big O’ and ‘little o’ notation

As argued above, the essence of asymptotic theory is approximation; approximation of Borel functions, random variables, distribution functions, mean, variances and higher moments (see Section 10.5). A particularly useful notion in the context of any approximation theory is that of the *accuracy* or *order of magnitude* of the approximations. In mathematical analysis the order of magnitude of the various quantities involved in an approximation is ‘kept track of’ by the use of the ‘big O, little o’ notation. It turns out that this notation can be extended to probabilistic approximations with minor modifications. The purpose of this section is to review the O, o notation and consider its extension to asymptotic theory.

Let  $\{a_n, b_n, n \in \mathcal{N}\}$  be a double sequence of real numbers.

*Definition 9*

The sequence  $\{a_n, n \in \mathcal{N}\}$  is said to be ‘at most of order  $b_n$ ’ and denoted by

$$\blacktriangleright \quad a_n = O(b_n) \quad \text{as } n \rightarrow \infty \quad \text{if } \lim_{n \rightarrow \infty} \left( \frac{|a_n|}{b_n} \right) < K, \\ \text{for some constant } K > 0. \quad (10.35)$$

*Definition 10*

The sequence  $\{a_n, n \in \mathcal{N}\}$  is said to be ‘of smaller order than  $b_n$ ’ and denoted by

$$\blacktriangleright \quad a_n = o(b_n) \quad \text{as } n \rightarrow \infty \quad \text{if } \lim_{n \rightarrow \infty} \left( \frac{a_n}{b_n} \right) = 0. \quad (10.36)$$

*Example 5*

$$\left( \frac{1}{2n^2 - 3} \right) = O\left( \frac{1}{n^2} \right); \quad (n+1) = O(n) = o(n^2); \quad \exp\{-n\} = o(n^{-\delta}), \quad \delta > 0; \\ \left( \frac{2n+n^2}{5n^2+n^3} \right) = O(n^{-1}); \quad \log_e n = o(n^\alpha), \quad \alpha > 0; \quad (6n^2 + 3n) = o(n^3) = O(n^2).$$

A very important implication stemming from these examples is that if

$$a_n = O(n^\alpha) \quad \text{then } a_n = o(n^{\alpha+\delta}), \quad \alpha, \delta > 0.$$

The O, o notation satisfies the following properties:

$$(P1) \quad \text{If } a_n = O(e_n) \quad \text{and} \quad b_n = O(c_n), \quad \text{then}$$

$$a_n b_n = O(e_n c_n);$$

$$|a_n|^r = O(e_n^r);$$

$$a_n + b_n = O(\min\{c_n, e_n\}).$$

The same results hold for ‘small o’ in place of ‘big O’ above.

$$(P2) \quad \text{If } a_n = O(e_n) \quad \text{and} \quad b_n = o(c_n), \quad \text{then}$$

$$a_n + b_n = O(e_n);$$

$$a_n b_n = o(e_n c_n).$$

The O, o notation can be extended to general real valued functions  $h(\cdot)$  and

$g( )$  with common domain  $D \neq \emptyset$ . We say  $h(x) = O(g(x))$  as  $x \rightarrow x_0$  if for a constant  $K > 0$ ,

$$\lim_{x \rightarrow x_0} \left| \frac{h(x)}{g(x)} \right| \leq K, \quad x \in (D - x_0).$$

Moreover, we say that  $h(x) = o(g(x))$  if

$$\lim_{x \rightarrow x_0} \left( \frac{h(x)}{g(x)} \right) = 0, \quad x \in (D - x_0).$$

*Example 6*

$$h(x) = e^x - 1, \quad |h(x)| \leq e|x| \quad \text{for } x \in [-1, 1], \quad h(x) = O(x),$$

and

$$h(x) = \cos x, \quad h(x) = 1 + o(x) \quad \text{as } x \rightarrow 0.$$

In the case where

$$h(x) - g(x) = O(l(x)) \quad \text{we write } h(x) = g(x) + O(l(x))$$

and for

$$h(x) - g(x) = o(l(x)) \quad \text{we write } h(x) = g(x) + o(l(x)).$$

This notation is particularly useful in the case of the Taylor expansion, where we can show that if  $h(x)$  is differentiable of order  $n$  (i.e. the derivatives  $(d^j h)/(dx^j) \equiv h^{(j)}$ ,  $j = 1, 2, \dots, n$ , exist for some positive integer  $n$ ) at  $x = x_0$ , then

$$\begin{aligned} h(x_0 + \delta) &= h(x_0) + h^{(1)}(x_0)\delta + \frac{h^{(2)}(x_0)}{2!} \delta^2 + \dots \\ &\quad + \frac{h^{(n)}(x_0)}{n!} \delta^n + o(\delta^n) \quad \text{as } \delta \rightarrow 0. \end{aligned} \tag{10.37}$$

The  $O$ ,  $o$  notation considered above can be extended to the case of stochastic convergence, convergence almost surely and in probability.

*Definition 11*

Let  $\{X_n, n \in \mathbb{N}\}$  be a sequence of r.v.'s and  $[c_n, n \in \mathbb{N}]$  a sequence of positive real numbers. We say that

- (i)  $X_n$  is **at most of order**  $c_n$  in probability if there exists non-

stochastic sequence  $\{a_n, n \in \mathbb{N}\}$  such that

$$a_n = O(1) \quad \text{and} \quad \left( \frac{X_n}{c_n} - a_n \right) \xrightarrow{P} 0$$

and denoted by  $X_n = O_p(c_n)$ .

(ii)  $X_n$  is of order smaller than  $c_n$  in probability if

$$\left( \frac{X_n}{c_n} \right) \xrightarrow{P} 0; \quad \text{denoted by } X_n = o_p(c_n).$$

In the same way we can define  $X_n = O_{a.s.}(c_n)$  and  $X_n = o_{a.s.}(c_n)$ .

It turns out that the properties P1 and P2 for O, o can be extended directly to  $O_p$ ,  $o_p$  and  $O_{a.s.}$ ,  $o_{a.s.}$  with minor modifications (see White (1984)). Moreover, order of magnitude results related to non-stochastic sequences can be transformed into stochastic order of magnitude using the following theorem due to Mann and Wald (1943).

*Theorem (Mann–Wald)*

Let  $\{X_n, n \in \mathbb{N}\}$  be a sequence of  $k$ -dimensional random vectors where  $\mathbf{X}_n \equiv (X_{jn}; j = 1, 2, \dots, k)$  such that

$$X_{jn} = O_p(c_{jn}), \quad j = 1, 2, \dots, m,$$

$$X_{jn} = o_p(c_{jn}), \quad j = m + 1, \dots, k.$$

If  $\{g_n(\mathbf{X}), n \in \mathbb{N}\}$  is a sequence of Borel functions  $g_n(\cdot): \mathbb{R}^k \rightarrow \mathbb{R}$  and

$$g_n(\mathbf{a}_n) = O(b_n),$$

for some non-stochastic sequence of  $k$ -dimensional vectors  $\{\mathbf{a}_n, n \in \mathbb{N}\}$  such that

$$a_{jn} = O(c_{jn}), \quad j = 1, 2, \dots, m,$$

$$a_{jn} = o(c_{jn}), \quad j = m + 1, \dots, k,$$

then we can deduce the order of  $g_n(\mathbf{X}_n)$  being  $O_p(b_n)$ , i.e.

$$g_n(\mathbf{X}_n) = O_p(b_n)$$

(see Fuller (1976)).

This theorem can be used to translate non-stochastic results related to the Taylor expansion to stochastic ones.

***Useful results relating to order***

- (O1) If  $\text{Var}(X_n) = \sigma_n^2 < \infty$  then  $X_n = O_p(\sigma_n)$  – as big as the standard deviation.
- (O2) If  $X_n = O_p(1/\sqrt{n}) \Rightarrow X_n = o_p(1)$ .
- (O3) If  $X_n \xrightarrow{D} X \Rightarrow X_n = O_p(1)$ .
- (O4) If  $X_n \xrightarrow{D} X \Rightarrow X_n + o_p(1) \xrightarrow{D} X$ .

**10.5 Extending the limit theorems**

As mentioned above, in statistical inference most of the Borel functions  $h(X_1, X_2, \dots, X_n)$  whose asymptotic behaviour is of interest are functions of  $\sum_{i=1}^n X_i^r$ ,  $r \geq 1$ . The limit theorems discussed in Chapter 9 refer exclusively to the asymptotic behaviour of the case  $r = 1$ . What we need to do now is to extend this to any  $r > 1$  as a prelude to a discussion of the asymptotic behaviour of a function of such quantities and then extend results to any Borel function  $h(X_1, X_2, \dots, X_n)$ , when possible.

For expositional purposes let us consider the case where  $\{X_n, n \geq 1\}$  is a sequence of IID r.v.'s on  $(S, \mathcal{F}, P(\cdot))$ . The quantities  $\sum_{i=1}^n X_i^r$ ,  $r \geq 1$ , are directly related to the raw moments  $\mu'_r \equiv E(X_i^r)$ ,  $r \geq 1$ , in the sense that if we were to select an ‘estimator’ of  $\mu'_r$  we would select

$$m_r = \frac{1}{n} \sum_{i=1}^n X_i^r, \quad r \geq 1 \tag{10.38}$$

as a natural choice. Given that the raw moments  $\{\mu'_r, r \geq 1\}$  play an important role in the manipulation and description of the distribution function  $F(x)$  their estimators sequence  $\{m_r, r \geq 1\}$  is expected to play an important role in statistical inference. For reasons which will become apparent in the next chapter the  $m_r$ s are called *sample raw moments*. Let us consider  $m_r = (1/n) \sum_{i=1}^n X_i^r$  as a Borel function of the IID r.v.'s  $X_1, \dots, X_n$  and thus a r.v. itself. Taking its expectation,

$$\begin{aligned} E(m_r) &= \frac{1}{n} \sum_{i=1}^n E(X_i^r) - \text{because of the linearity of } E(\cdot) \\ &= \frac{1}{n} \sum_{i=1}^n \mu'_r \quad - \text{because of the IID assumption,} \\ &= \mu'_r. \end{aligned} \tag{10.39}$$

Let

$$Y_i = X_i^r, \quad i = 1, 2, \dots, \quad \text{then } m_r = \frac{S_n}{n} \quad \text{where } S_n = \sum_{i=1}^n Y_i,$$

that is,  $S_n$  is the sum of IID r.v.'s  $Y_1, Y_2, \dots, Y_n$ . From the WLLN and SLLN we can deduce that

$$(i) \quad m_r \xrightarrow{P} \mu'_r, \quad r \geq 1; \quad (10.40)$$

and

$$(ii) \quad m_r \xrightarrow{\text{a.s.}} \mu'_r, \quad r \geq 1. \quad (10.41)$$

Using the Lindeberg–Levy CLT we can deduce that if  $\text{Var}(Y_i) = \sigma_y^2 < \infty$  then

$$(iii) \quad \sqrt{n} \frac{(m_r - \mu'_r)}{\sigma_y} \xrightarrow{D} Z \sim N(0, 1), \quad r \geq 1. \quad (10.42)$$

If

$$\text{Var}(X_i) = \sigma^2 < \infty \Rightarrow \text{Var}(Y_i) = E(X_i^r - \mu'_r)^2 = \mu'_{2r} - \mu'^2_r = \sigma_y^2,$$

assuming that  $\mu'_{2r} < \infty$ ,  $r \geq 1$ . Moreover, given that

$$\begin{aligned} E(m_r m_k) &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n E(X_i^r X_j^k) = \frac{1}{n^2} \sum_{i=1}^n E(X_i^{r+k}) + \frac{1}{n^2} \sum_{i \neq j} \sum E(X_i^r X_j^k) \\ &= \frac{1}{n} \mu'_{r+k} + \left( \frac{n-1}{n} \right) \mu'_r \mu'_k, \end{aligned} \quad (10.43)$$

we can deduce that

$$\text{Cov}(m_r m_k) = \frac{1}{n} (\mu'_{r+k} - \mu'_r \mu'_k). \quad (10.44)$$

Since  $\text{Cov}(m_r m_k) = E(m_r m_k) - \mu'_r \mu'_k$ , we can deduce that

$$(iv) \quad \sqrt{n}(\mathbf{m}_r - \boldsymbol{\mu}'_r) \xrightarrow{D} \mathbf{Z} \sim N(\mathbf{0}, \boldsymbol{\Sigma}),$$

$$\mathbf{m}_r = (m_1, m_2, \dots, m_r), \quad \boldsymbol{\mu}'_r = (\mu'_1, \mu'_2, \dots, \mu'_r) \quad (10.45)$$

with  $\boldsymbol{\Sigma} = [\sigma_{ij}]$ ,  $\sigma_{ij} = \mu'_{i+j} - \mu'_i \mu'_j$ ,  $i, j = 1, 2, \dots, r$ . The results (i)–(iv) can be seen as extensions of the limit theorems in Chapter 9 for the sum  $\sum_{i=1}^n X_i$  to the general sum  $\sum_{i=1}^n X_i^r$ ,  $r \geq 1$ .

Having derived the asymptotic results (i)–(iv) for the sample raw moments we can now extend these results to any *continuous function* of them using Lemma 10.1 above.

*Example 7*

If  $g(\cdot) = \log_e(\cdot)$  then for  $m_r > 0, \mu'_r > 0, Z > 0$ :

$$(i) \quad \log_e(m_r) \xrightarrow{P} \log_e(\mu'_r);$$

$$(ii) \quad \log_e(m_r) \xrightarrow{\text{a.s.}} \log_e(\mu'_r);$$

$$(iii) \quad \log_e\left(\frac{\sqrt{[n(m_r - \mu'_r)]}}{\sqrt{(\mu'_{2r} - \mu'^2_r)}}\right) \xrightarrow{D} \log_e Z$$

The last result, however, is not particularly useful because we usually need to derive the distribution of  $g(m_r)$  and not that of  $g\{\sqrt{[n(m_r - \mu'_r)]}/\sqrt{(\mu'_{2r} - \mu'^2_r)}\}$ . Let us derive the asymptotic distribution of  $g(m_r)$ , taking the opportunity to use some of the concepts and results propounded above.

From O1 above we know that  $m_r = \mu'_r + O_p(1/\sqrt{n})$  and hence from the Mann–Wald theorem we can deduce that if  $g(\cdot)$  has continuous derivatives of order  $k$  then

$$\begin{aligned} g(m_r) &= g(\mu'_r) + g^{(1)}(\mu'_r)(m_r - \mu'_r) + \dots \\ &\quad + \frac{1}{(k-1)!} g^{(k-1)}(\mu'_r)(m_r - \mu'_r)^{k-1} + O_p(n^{k/2}). \end{aligned} \quad (10.46)$$

Assuming that  $k=2$  we can deduce that

$$g(m_r) - g(\mu'_r) - g^{(1)}(\mu'_r)(m_r - \mu'_r) = O_p(n^{-1}) \quad (10.47)$$

and

$$\sqrt{n}(g(m_r) - g(\mu'_r)) - \sqrt{n}(m_r - \mu'_r)g^{(1)}(\mu'_r) = O_p(n^{-\frac{1}{2}})$$

and by O2,

$$\sqrt{n}(g(m_r) - g(\mu'_r)) = \sqrt{n}(m_r - \mu'_r)g^{(1)}(\mu'_r) + o_p(1).$$

Let

$$V_n = \sqrt{n}(g(m_r) - g(\mu'_r)), \quad U_n = \sqrt{n}(m_r - \mu'_r) \quad \text{and} \quad c = g^{(1)}(\mu'_r) \neq 0$$

then

$$V_n = cU_n + o_p(1). \quad (10.48)$$

From Lemma 3 we may conclude that if  $U_n \xrightarrow{D} U$  then  $V_n \xrightarrow{D} cU$ , and thus

$$(v) \quad \sqrt{n}(g(m_r) - g(\mu'_r)) \xrightarrow{a} N(0, (\mu'_{2r} - \mu'^2_r)[g^{(1)}(\mu'_r)]^2). \quad (10.49)$$

For example, if  $g(m_r) = \log_e m_r$  then

$$\sqrt{n}(\log_e m_r - \log_e \mu'_r) \xrightarrow{a} N\left(0, \frac{\mu'_{2r} - \mu'^2_r}{\mu'^2_r}\right). \quad (10.50)$$

This result can be generalised to the vector case (iv) directly. If we define  $\mathbf{G}(\mathbf{m}_r) \equiv (g_1(\mathbf{m}_r), g_2(\mathbf{m}_r), \dots, g_k(\mathbf{m}_r))$

$$\sqrt{n}(\mathbf{G}(\mathbf{m}_r) - \mathbf{G}(\boldsymbol{\mu}_r)) \underset{x}{\sim} N(\mathbf{0}, \mathbf{D}\Sigma\mathbf{D}'), \quad (10.51)$$

where

$$\mathbf{D} = \left[ \frac{\partial g_i(\mathbf{m}_r)}{\partial m_j} \Big|_{\mathbf{m}_r = \boldsymbol{\mu}_r} \right], \quad i = 1, 2, \dots, k, \quad j = 1, 2, \dots, r.$$

The above procedure is known as the  *$\delta$ -method* (see Rothenberg (1984)). The above results can also be directly generalised to an IID sequence of random vectors  $\{\mathbf{X}_n, n \geq 1\}$ ,  $\mathbf{X}_n : k \times 1$  without any crucial changes. In order to see how useful these results are we can easily show that one of the most important Borel functions of interest in Part IV, the estimator

$$\hat{b} = \frac{\frac{1}{n} \sum_{i=1}^n X_i Y_i - \left( \frac{1}{n} \sum_{i=1}^n X_i \right) \left( \frac{1}{n} \sum_{i=1}^n Y_i \right)}{\frac{1}{n} \sum_{i=1}^n X_i^2 - \left[ \frac{1}{n} \sum_{i=1}^n X_i \right]^2} \quad (10.52)$$

can be easily accommodated into the above framework. To see this let

$$\begin{aligned} z_1 &= \frac{1}{n} \sum_{i=1}^n X_i, & z_2 &= \frac{1}{n} \sum_{i=1}^n Y_i, & z_3 &= \frac{1}{n} \sum_{i=1}^n X_i^2, \\ z_4 &= \frac{1}{n} \sum_{i=1}^n Y_i^2, & z_5 &= \frac{1}{n} \sum_{i=1}^n X_i Y_i, \end{aligned}$$

then

$$\hat{b} \equiv g(z_1, z_2, z_3, z_4, z_5) = \frac{z_5 - z_1 z_2}{z_3 - z_1^2}. \quad (10.53)$$

Let us summarise the argument so far. In our attempt to derive asymptotic results for arbitrary Borel functions  $h(X_1, X_2, \dots, X_n)$  we extended the limit theorems for  $\sum_{i=1}^n X_i$  to more general sums  $\sum_{i=1}^n X_i^r$ ,  $r \geq 1$ , and then these results were extended to continuous functions of these general sums.

The results so far suggest the following way to extend them even further. Consider the IID sequence of  $k$ -dimensional random vectors  $\{\mathbf{X}_n, n \geq 1\}$  and define the Borel functions  $l_i(\cdot) : \mathbb{R}^k \rightarrow \mathbb{R}$ ,  $i = 1, 2, \dots, r$  (analogous to  $\sum_{i=1}^n X_i^r$ ,  $r \geq 1$ ) and  $\mathbf{Z}_n \equiv (l_1(\mathbf{X}_n), l_2(\mathbf{X}_n), \dots, l_r(\mathbf{X}_n))$  (the  $z_i$ s above) with  $E(\mathbf{Z}_n) = \boldsymbol{\mu}$ . Let  $h(\cdot) : \mathbb{R}^r \rightarrow \mathbb{R}$  be a Borel function ( $g(\cdot)$  above), and define  $h(\bar{\mathbf{Z}}_n)$  where  $\bar{\mathbf{Z}}_n = (1/n) \sum_{i=1}^n \mathbf{Z}_i$ . The above result in terms of  $h(\bar{\mathbf{Z}}_n)$  being continuously differentiable at  $\bar{\mathbf{Z}}_n = \boldsymbol{\mu}$  takes the form

$$\sqrt{n}(h(\bar{\mathbf{Z}}_n) - h(\boldsymbol{\mu})) \underset{x}{\sim} N(0, \mathbf{D}\Sigma\mathbf{D}'), \quad (10.54)$$

where

$$\mathbf{V} = \text{Cov}(\mathbf{Z}_1), \quad \mathbf{D} = \left[ \frac{\partial h(\mathbf{Z})}{\partial \bar{\mathbf{Z}}_i} \Big|_{\mathbf{Z}=\boldsymbol{\mu}} \right] \quad i=1, 2, \dots, r$$

(see Bhattacharya (1977) and references therein).

## 10.6 Error bounds and asymptotic expansions

The CLT tells us that under certain conditions the distribution of the standardised sum  $Y_n = \{[S_n - E(S_n)]/\sqrt{\text{Var}(S_n)}\}$ , tends to a standard normal r.v., i.e.

$$\lim_{n \rightarrow \infty} \bar{F}_n(y) = \Phi(y) = \int_{-\infty}^y \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}u^2\} du. \quad (10.55)$$

It provides us with no information as to how large the approximation error  $|\bar{F}_n(y) - \Phi(y)|$  is for a particular value  $n$ . The first formal bound of the approximation error was provided by Liapunov in 1901 proving that: For  $X_1, X_2, \dots$  an IID sequence of r.v.'s with  $E(X_i) = \mu$ ,  $\text{Var}(X_i) = \sigma^2$  and  $E(|X_i - \mu|^3) \equiv v_3$ , all finite for  $i = 1, 2, \dots$ ,

$$\blacktriangleright \quad \sup_{t \in \mathbb{R}} |\bar{F}_n(y) - \Phi(y)| \leq C \left( \frac{v_3}{\sigma^3} \right) \frac{\log n}{\sqrt{n}}, \quad (10.56)$$

where  $C$  is a positive constant. This result was later sharpened by Cramer but the best result was provided by Berry (1941) and Esseen (1945).

### Berry–Esseen theorem

*Under the same conditions as the Liapunov result we can deduce that*

$$\blacktriangleright \quad \sup_y |\bar{F}_n(y) - \Phi(y)| \leq \frac{C}{\sqrt{n}} \left( \frac{v_3}{\sigma^3} \right), \quad C = \frac{33}{4}. \quad (10.57)$$

That is, the factor  $\log n$  was unnecessary. As we can see, this is a sharpening of the Lindeberg–Levy CLT (see Chapter 9) in so far as the latter states that  $\sup_y |\bar{F}_n(y) - \Phi(y)| \rightarrow 0$  as  $n \rightarrow \infty$  without specifying the rate of convergence. The Berry–Esseen theorem using higher moments provides us with the additional information that the rate of convergence is  $O(n^{-\frac{1}{2}})$ . Various authors improved the upper bound of the error by reducing the constant  $C$ . The best bound for  $C$  so far was provided by Beeck (1972),  $0.4097 \leq C < 0.7975$ .

For an independent but not identically distributed sequence of r.v.'s  $X_1, \dots, X_n$  with  $E(X_i) = \mu_i$ ,  $\text{Var}(X_i) = \sigma_i^2$  and  $E(|X_i - \mu_i|^3) = v_{3i} < \infty$ ,  $i = 1, 2, \dots$ ,

the Berry–Esseen result takes the form

$$\blacktriangleright \quad \sup_y |\bar{F}_n(y) - \Phi(y)| \leq C \frac{\sum_{i=1}^n v_{3i}}{\left( \sum_{i=1}^n \sigma_i^2 \right)^{\frac{3}{2}}} \quad (10.58)$$

Although these results provide us with a simple and usually accurate absolute magnitude of the approximation error, they become trivially true (i.e. uninformative) when  $y$  is allowed to vary with  $n$ . For example if  $y \rightarrow -\infty$  as  $n \rightarrow \infty$  then  $\bar{F}_n(y)$  and  $\Phi(y)$  tend to zero (see Phillips (1980)). In order to reflect the dependence of the upper bound on  $y$  as well, the Berry–Esseen result can be extended to

$$\blacktriangleright \quad |\bar{F}_n(y) - \Phi(y)| \leq C \left( \frac{v_3}{\sigma^3} \right) \left( \frac{1}{1+y^2} \right) \quad \text{for all } y \in \mathbb{R} \quad (10.59)$$

(see Ibragimov and Linnik (1971)). An important disadvantage of these results is that they provide us with no information as to how we *can improve* the asymptotic approximations provided by the CLT or how we can choose between asymptotic approximations provided by the CLT or how we can choose between asymptotically equivalent results. This line of reasoning leads us naturally to the idea of *asymptotic expansions* related to the approximation error.

The idea underlying some asymptotic expansions of the approximation error is that we can think of the asymptotic distribution as the first term in an expansion of the form

$$\blacktriangleright \quad \bar{F}_n(y) = \Phi(y) + \sum_{i=1}^r \frac{A_i(y)}{(n^{\frac{1}{2}})^i} + R_{rn}(y), \quad (10.60)$$

where the terms of the expansion are polynomials in  $y$  in powers of  $(n^{-\frac{1}{2}})$  and the remainder  $R_{rn} = O(n^{-r/2})$ . This is the so-called *Edgeworth expansion* widely used in econometrics. Let us consider how such asymptotic expansions can be derived in the present context.

Let  $f_n(x)$  be the density function of an appropriately normalised Borel function of the IID sequence of r.v.'s  $X_1, X_2, \dots, X_n$ .  $f_n(x)$  is assumed to belong to a particular family of functions defined over the interval  $(-\infty, \infty)$  and satisfy the condition

$$\int_{-\infty}^{\infty} |f(x)|^2 dx < \infty, \quad (10.61)$$

known as square integrability. The space of all such functions is denoted by  $L_2(-\infty, \infty)$  and constitutes a Hilbert space (see Kreyszig (1978)). One of

the most important properties of  $L_2(-\infty, \infty)$  is that any element in this space can be represented or sufficiently accurately approximated by the use of the orthogonal polynomials:

$$H_k(x) = \frac{(-1)^k}{\phi(x)} \frac{d^k \phi(x)}{dx^k}, \quad k=0, 1, 2, \dots, \quad (10.62)$$

known as *Hermite polynomials* of degree  $k$ ;  $\phi(x) = [1/\sqrt{(2\pi)}] \exp\{-\frac{1}{2}x^2\}$ , the density function of a standard normal r.v. This is a natural extension of the concept of an orthogonal basis spanning a linear space. The orthogonality of these polynomials comes in the form of

$$\int_{-\infty}^{\infty} H_k(x) H_m(x) \phi(x) dx = k! \quad \text{for } m=k \\ = 0 \quad \text{otherwise.} \quad (10.63)$$

The first five of these polynomials are

$$H_0 = 1, \quad H_1 = x, \quad H_2 = x^2 - 1, \quad H_3 = x^3 - 3x, \quad H_4 = x^4 - 6x^2 + 3. \quad (10.64)$$

The density function  $f_n(x)$  is assumed to be an element of  $L_2(-\infty, \infty)$  and thus it can be represented in the form

$$f_n(x) = \sum_{k=0}^{\infty} b_k H_k(x). \quad (10.65)$$

Although in principle we can approximate  $f_n(x)$  sufficiently accurately by choosing a finite value for the summation, in practice we prefer to approximate the ratio  $[f_n(x)/\phi(x)]$  in view of the fact that it is usually smoother than  $f_n(x)$  and thus easier to approximate sufficiently accurately by a low degree polynomial. This ratio can be approximated by

$$\frac{f_n(x)}{\phi(x)} \approx \sum_{k=0}^r b_k H_k(x), \quad (10.66)$$

where

$$b_k = \frac{1}{k!} \int_{-\infty}^{\infty} f_n(x) H_k(x) dx. \quad (10.67)$$

These coefficients are chosen so as to minimise the error

$$\int_{-\infty}^{\infty} \phi(x) \left[ \frac{f_n(x)}{\phi(x)} - \sum_{k=0}^r b_k H_k(x) \right]^2 dx. \quad (10.68)$$

Hence, the density function  $f_n(x)$  can be approximated by  $f_n^*(x)$  where

$$f_n^*(x) = \phi(x) \sum_{k=0}^r b_k H_k(x). \quad (10.69)$$

As we can see, the coefficients  $b_k, k = 0, 1, 2, \dots$  as defined above are directly related to the moments of  $f_n(x)$ . For instance

$$\begin{aligned} b_0 &= \int_{-\infty}^{\infty} f_n(x) dx = 1, \\ b_1 &= \int_{-\infty}^{\infty} xf_n(x) dx = \mu'_1 \\ b_2 &= \frac{1}{2} \int_{-\infty}^{\infty} (x^2 - 1)f_n(x) dx = \frac{1}{2}(\mu'_2 - 1), \\ b_3 &= \frac{1}{6}(\mu'_3 - 3\mu'_1), \quad \text{etc.} \end{aligned}$$

This implies that if we know the moments of  $f_n(x)$  up to some order  $r$  we can use  $f_n^*(x)$  to approximate it.

In view of the relationship between the derivatives of  $\phi(x)$  and Hermite polynomials  $f_n^*(x)$  can be expressed in the form

$$f_n^*(x) = \sum_{k=0}^r \frac{c_k}{k!} \phi^{(k)}(x), \quad (10.70)$$

where

$$\begin{aligned} c_k &= (-1)^k \int_{-\infty}^{\infty} f_n(x) H_k(x) dx, \\ \phi^{(k)}(x) &= \frac{d^k \phi(x)}{dx^k}, \quad k = 1, 2, \dots \end{aligned} \quad (10.71)$$

That is, the approximation is a linear combination of  $\phi(x)$  and its derivatives up to order  $r$ ; see Cramer (1946). This form of the approximation is known as the *Gram–Charlier series A* approximation. In terms of the cumulative distribution function  $F_n(x) = \int_{-\infty}^x f_n(u) du$  the approximation is

$$\blacktriangleright F_n^*(x) = \sum_{k=0}^r \frac{c_k}{k!} \Phi^{(k)}(x). \quad (10.72)$$

The question which is of primary interest about these approximations is not so much whether the above series converges to  $F_n(x)$  or  $f_n(x)$  as  $r \rightarrow \infty$  but whether for a small  $r < \infty$  the above series will provide a good approximation to these functions or not. A particular way to proceed which enables us to choose the order of the approximation error is to break down the terms  $b_k H_k(x)$  into components and then reassemble them according to their *order of magnitude*, say in powers of  $n^{-\frac{1}{2}}$ . That is, express  $f_n^*(x)$  in the form

$$\blacktriangleright f_n^*(x) = \phi(x) \left[ 1 + \sum_{k=1}^r \frac{A_k(x)}{(\sqrt{n})^k} \right], \quad (10.73)$$

where  $A_k(x)$  is a polynomial in  $x$ . In order to be able to choose the order of the approximation error (remainder)  $R_{nr}(x)$  defined by  $R_{nr}(x)=f_n(x)-f_n^*(x)$  we need to ensure that the above series constitutes a proper *asymptotic expansion*. An asymptotic expansion is defined to be a series which has the property that when truncated at some finite number  $r$  the remainder has the same order of magnitude as the first neglected term. In the present case the remainder must be of order

$$(n^{-\frac{1}{2}})^{r+1}, \quad \text{i.e. } R_{nr}(x)=O(n^{-(r+1)/2}). \quad (10.74)$$

This is ensured in the case of an expansion

$$f(x)=\sum_{k=0}^r \alpha_k \psi_k(x) + R_r(x), \quad (10.75)$$

when the sequence  $\{\psi_k(x), k \geq 0\}$  is an *asymptotic sequence* as  $x \rightarrow x_0$

$$\psi_{k+1}(x)=O(\psi_k) \quad \text{as } x \rightarrow x_0 \quad (10.76)$$

because then  $R_r(x)=O(\psi_k)$ .

Under certain restrictions (see Feller (1970)) the expansion of the form

$$\blacktriangleright \quad f_n(x)=\phi(x)\left[1+\sum_{k=1}^r \frac{A_k(x)}{(\sqrt{n})^k}\right]+R_m \quad (10.77)$$

constitutes a proper asymptotic expansion and  $R_m(x)=O(n^{-(l(r+1)/2)})$ . This is known as the *Edgeworth expansion* and has been widely applied in the econometric literature (see Sargan (1976), Phillips (1977), Rothenberg (1984), *inter alia*).

In order to illustrate some of the concepts introduced above let us consider the standardised Borel function

$$Z_n=\frac{\sqrt{n}}{\sigma}\left(\frac{1}{n} \sum_{i=1}^n X_i - \mu\right) \quad (10.78)$$

of the IID sequence  $X_1, \dots, X_n$  with  $E(X_i)=\mu$ ,  $\text{Var}(X_i)=\sigma^2$ ,  $E(X_i-\mu)^k=\mu_i$ ,  $i=3, 4, \dots$ . The first problem we have to face is to derive the necessary moments of  $Z_n$  in order to be able to evaluate the coefficients  $c_k$ ,  $k=0, 1, 2, \dots, r$ . Apart from the first two moments  $E(Z_n)=0$ ,  $E(Z_n^2)=1$ , the higher moments involve some messy manipulations (see Cramer (1946)). Using these moments given by Cramer we can evaluate the first few coefficients:

$$c_0=1, \quad c_1=0, \quad c_2=0, \quad c_3=-\frac{1}{\sqrt{n}}\kappa_3,$$

$$c_4=\frac{1}{n}\kappa_4, \quad c_5=-\frac{1}{n^{\frac{3}{2}}}\kappa_5, \quad c_6=\left(\frac{\kappa_6}{n^2}+\frac{10\kappa_3^2}{n}\right)$$

where  $\kappa_i$ ,  $i = 3, 4, 5, 6$  refer to the cumulants of  $(X_i - \mu)/\sigma$  (see Chapter 4);  $\kappa_3 = \mu_3/\sigma^3$ ,  $\kappa_4 = (\mu_4/\sigma^4) - 3$  (see Cramer (1946)). The Gram–Charlier approximation of  $f_n(x)$  for  $r=6$  takes the form

$$\begin{aligned} f_n^*(x) &= \phi(x) - \frac{1}{3!} \frac{\kappa_3}{n^{\frac{3}{2}}} \phi^{(3)}(x) + \frac{1}{4!} \frac{\kappa_4}{n} \phi^{(4)}(x) - \frac{1}{5!} \frac{\kappa_5}{n^{\frac{5}{2}}} \phi^{(5)}(x) \\ &\quad + \frac{1}{6!} \left( \frac{\kappa_6}{n^{\frac{3}{2}}} + \frac{10\kappa_3^2}{n} \right) \phi^{(6)}(x). \end{aligned} \quad (10.79)$$

Collecting the terms with the same order of magnitude we can construct the Edgeworth expansion

$$\blacktriangleright \quad f_n(x) = \phi(x) - \frac{1}{3!} \frac{\kappa_3}{n^{\frac{3}{2}}} \phi^{(3)}(x) + \frac{1}{n!} \left[ \frac{1}{4!} \kappa_4 \phi^{(4)}(x) + \frac{10}{6!} \kappa_3^2 \phi^{(6)}(x) \right] + R_{2n}, \quad (10.80)$$

where  $R_{2n} = O(n^{-\frac{1}{2}})$ . In terms of the Hermite polynomials and the central moments this takes the form

$$\blacktriangleright \quad f_n(x) = \phi(x) \left[ 1 + \frac{1}{n^{\frac{1}{2}}} \left( \frac{\mu_3}{\sigma^3} \right) \frac{H_3(x)}{3!} + \frac{1}{n} \left\{ \left( \frac{\mu_4}{\sigma^4} - 3 \right) \frac{H_4(x)}{4!} \right. \right. \\ \left. \left. + 10 \left( \frac{\mu_3}{\sigma^3} \right)^2 \frac{H_6(x)}{6!} \right\} \right] + R_{2n}. \quad (10.81)$$

It is important to note that the presence of cumulants in (79) and (80) is no coincidence. It is due to the fact that for the sum of standardised IID r.v.'s

$$\kappa_r = O(n^{1-(r/2)}), \quad r \geq 2$$

From these expansions we can see that in the case where the distribution of the  $X_i$ 's is symmetric [ $(\mu_3/\sigma^3)=0$ ] the CLT approximation is of order  $1/n$  and when the kurtosis  $\alpha_4 = (\mu_4/\sigma^4) = 3$  the approximation is even better, of order  $1/n^{\frac{1}{2}}$  and not of order  $1/\sqrt{n}$  as the CLT suggests. In a sense we can interpret the CLT as providing us with the first term in the above Edgeworth expansion and if we want to improve it we should include higher-order terms.

The Gram–Charlier and Edgeworth expansions for arbitrary Borel functions of  $X_1, \dots, X_n$  can be derived similarly when the moments needed to determine the required coefficients are available (see Bhattacharya (1977)). When these moments are not easily available the approximate moments considered in Section 10.5 can be used instead. It must be emphasised that Edgeworth expansions are not restricted to Borel functions  $h(\mathbf{X})$  with asymptotic normal distributions. Certain Borel

functions of considerable interest in econometrics (see Chapter 16) have asymptotic chi-square distributions for which Edgeworth approximations can be developed (see Rothenberg (1984)). The only difference from the derivation of the normal Edgeworth expansion is that the appropriate space is  $L_2(0, \infty)$  and the corresponding orthogonal polynomials are the so-called *Laguerre polynomials*:

$$L_n(x) = \frac{e^x}{k!} \frac{d^k}{dx^k} (x^k e^{-x}), \quad k = 1, 2, \dots \quad (10.82)$$

Error bounds and asymptotic expansions of the type discussed above are of considerable interest in econometrics where we often have to resort to asymptotic theory (see Part IV).

### **Important concepts**

Convergence almost surely, convergence in  $r$ th mean, convergence in probability, convergence in distribution, convergence on  $S$ , convergence uniformly on  $S$ , big  $O_p$ , small  $o_p$ , order of magnitude, Mann–Wald theorem, Taylor series expansions, sample raw moments,  $\delta$ -method, limits of moments, asymptotic moments, approximate moments, uniform integrability, Berry–Esseen theorem, square integrable functions, orthogonal polynomials, polynomial approximation, Hermite polynomials, Gram–Charlier series A, asymptotic sequence, asymptotic expansion, Edgeworth expansion.

### **Questions**

1. Why do we need to bother with asymptotic theory and run the risk to use inaccurate approximations?
2. Compare and contrast the two modes of convergence, convergence in probability and almost sure convergence. Explain intuitively why  
 $\xrightarrow{\text{a.s.}} \xrightarrow{P} \xrightarrow{D}$ .
3. Compare the concepts of convergence in probability and convergence in distribution and give an intuitive explanation to the result  $\xrightarrow{P} \xrightarrow{D} \xrightarrow{D}$ .
4. Explain the Cramer–Wald device of proving convergence in distribution for a vector sequence  $\{\mathbf{X}_n, n \geq 1\}$ .
5. Explain the  $O_p$  and  $o_p$  notation and compare it with the  $O$ ,  $o$  notation of mathematical analysis.
6. Discuss the Mann–Wald theorem on how to derive stochastic order of magnitude results from non-stochastic ones.
7. Explain the Taylor series expansion for a Borel function  $h(\mathbf{X})$  of a sequence of r.v.'s  $\mathbf{X} \equiv (X_1, X_2, \dots, X_n)$ .

8. Explain how the results of the limit theorems for  $\sum_{i=1}^n X_i$  can be extended to the sample raw moments  $\sum_{i=1}^n X_i^r$ ,  $r \geq 1$ . How can the latter be extended to arbitrary continuous functions of them?
9. Discuss the  $\delta$ -method for deriving the asymptotic distribution of  $g(m_r)$  ( $g(\cdot)$  continuous) from that of  $m_r$ .
10. Compare and contrast the following concepts:
  - (i)  $r$ th-order moments;
  - (ii) limits of  $r$ th-order moments;
  - (iii) asymptotic  $r$ th-order moments; and
  - (iv) approximate moments.
11. Explain intuitively the concept of uniform integrability.
12. Discuss the Berry–Esseen theorem and its role in deriving upper bounds for the approximation error in the CLT.
13. Explain the role of Edgeworth expansions in asymptotic theory.
14. Discuss the derivation of an orthogonal polynomial approximation to an appropriately standardised Borel function of a sequence of r.v.'s  $\{X_n, n \geq 1\}$  in the context of the space  $L_2(-\infty, \infty)$ .
15. Explain the difference between Gram–Charlier and Edgeworth expansions.
16. What order of approximation does the CLT provide in the case where the skewness and kurtosis coefficients of  $\sum_{i=1}^n X_i$  are zero and three respectively? Explain.
17. Discuss the question of how Edgeworth approximations can help us discriminate between asymptotically equivalent Borel functions of a sequence of r.v.'s  $\{X_n, n \geq 1\}$ .

### *Exercises*

1. Determine the order of magnitude (big O and small o) of the following sequences:
  - (i)  $\frac{n^3 + 6n^2 + 2}{6n^3 + 1}$ ,  $n = 1, 2, \dots$ ;
  - (ii)  $\log_e n$ ,  $n = 1, 2, \dots$ .
2. Determine the order of magnitude of the following series:
  - (i)  $\sum_{i=1}^n i$ ;
  - (ii)  $\sum_{i=1}^n i^{-2}$ ;
  - (iii)  $\sum_{i=1}^n i^{\frac{1}{2}}$ .

210      **Introduction to asymptotic theory**

3. For the IID sequence  $\{X_n, n \geq 1\}$  where  $X_n \sim N(0, 1)$  we know  $v_n = (\sum_{i=1}^n X_i^2) \sim \chi^2(n)$ . From the CLT we know that  $h_n(\mathbf{X}) = (v_n - n)/[\sqrt{(2n)}] \sim N(0, 1)$ . Derive the Edgeworth approximation of  $h(\mathbf{X})$  of order  $1/n$ . (Hint:  $\alpha_{3n} = (2\sqrt{2})/\sqrt{n}$ ,  $\alpha_{4n} = 12/n$ .)

**Additional references**

Bhattacharya and Rao (1976); Bishop *et al.* (1975); Billingsley (1968); Rao (1984); Wallace (1958).

## **PART III**

---

### **Statistical inference**

---

## CHAPTER 11

---

### The nature of statistical inference

---

#### 11.1 Introduction

In the discussion of descriptive statistics in Part I it was argued that in order to be able to go beyond the mere summarisation and description of the observed data under consideration it was important to develop a mathematical model purporting to provide a generalised description of the data generating process (DGP). Motivated by the various results on frequency curves, a probability model in the form of the parametric family of density functions  $\Phi = \{f(x; \theta), \theta \in \Theta\}$  and its various ramifications was formulated in Part II, providing such a mathematical model. Along with the formulation of the probability model  $\Phi$  various concepts and results were discussed in order to enable us to extend and analyse the model, preparing the way for statistical inference to be considered in the sequel. Before we go on to consider that, however, it is important to understand the difference between the descriptive study of data and statistical inference. As suggested above, the concept of a density function in terms of which the probability model is defined was motivated by the concept of a frequency curve. It is obvious that any density function  $f(x; \theta)$  can be used as a frequency curve by reinterpreting it as a non-stochastic function of the observed data. This precludes any suggestions that the main difference between the descriptive study of data and statistical inference proper lies with the use of density functions in describing the observed data. ‘What is the main difference then?’

In descriptive statistics the aim is to summarise and describe the data under consideration and frequency curves provide us with a convenient way to do that. The choice of a frequency curve is entirely based on the data in hand. On the other hand, in statistical inference a probability model  $\Phi$  is

postulated a priori as a generalised description of the underlying DGP giving rise to the observed data (not the observed data themselves). Indeed, there is nothing stochastic about a set of members making up the data. The stochastic element is introduced into the framework in the form of uncertainty relating to the underlying DGP and the observed data are viewed as one of the many possible realisations. In descriptive statistics we start with the observed data and seek a frequency curve which describes these data as closely as possible. In statistical inference we postulate a probability model  $\Phi$  a priori, which purports to describe either the DGP giving rise to the data or the population which the observed data came from. These constitute fundamental departures from descriptive statistics allowing us to make generalisations beyond the numbers in hand. This being the case the analysis of observed data in statistical inference proper will take a very different form as compared with descriptive statistics briefly considered in Part I. In order to see this let us return to the income data discussed in Chapter 2. There we considered the summarisation and description of personal income data on 23 000 households using descriptors like the mean, median, mode, variance, the histogram and the frequency curve. These enabled us to get some idea about the distribution of incomes among these households. The discussion ended with us speculating about the possibility of finding an appropriate frequency curve which depends on few parameters enabling us to describe the data and analyse them in a much more convenient way. In Section 4.3 we suggested that the parametric family of density functions of the Pareto distribution

$$\Phi = \left\{ f(x; \theta) = \frac{\theta}{x_0} \left( \frac{x_0}{x} \right)^{\theta+1}, x > 0, \theta \in \mathbb{R}_+ \right\} \quad (11.1)$$

could provide a reasonable probability model for incomes over £4500. As can be seen, there is only one unknown parameter  $\theta$  which once specified  $f(x; \theta)$  is completely determined. In the context of statistical inference we postulate  $\Phi$  a priori as a stochastic model not of the data in hand but of the distribution of income of the population from which the observed data constitute one realisation, i.e. the UK households. Clearly, there is nothing wrong with using  $f(x; \theta)$  as a *frequency curve* in the context of descriptive statistics by returning to the histogram of these data and after plotting  $f(x; \theta)$  for various values of  $\theta$ , say  $\theta = 1, 1.5, 2$ , choose the one which comes closer to the frequency polygon. For the sake of the argument let us assume that the curve chosen is  $\theta = 1.5$ , i.e.

$$f(x) = \frac{1.5}{4500} \left( \frac{4500}{x} \right)^{2.5} = 452.804 x^{-2.5}. \quad (11.2)$$

This provides us with a very convenient descriptor of these data as can be easily seen when compared with the cumbersome histogram function

$$f^*(x) = \sum_{i=1}^8 \frac{\phi_i}{8(x_{i+1} - x_i)} I([x_i, x_{i+1})) \quad (11.3)$$

(see Chapter 2). But it is no more than a convenient descriptor of the data in hand. For example, we cannot make any statements about the distribution of personal income in the UK on the basis of the frequency curve  $f^*(x)$ . In order to do that we need to consider the problem in the context of statistical inference proper. By postulating  $\Phi$  above as a probability model for the distribution of income in the UK and interpreting the observed data as a sample from the population under study we could go on to consider questions about the unknown parameter  $\theta$  as well as further observations from the probability model, see Section 11.4 below.

In Section 11.2 the important concept of a sampling model is introduced as a way to link the probability model postulated, say  $\Phi = \{f(x; \theta), \theta \in \Theta\}$ , to the observed data  $x \equiv (x_1, \dots, x_n)'$  available. The sampling model provides the second important ingredient needed to define a statistical model; the starting point of any ‘parametric’ statistical inference.

In Section 11.3, armed with the concept of a statistical model, we go on to discuss a particular approach to statistical inference, known as the frequency approach. The frequency approach is briefly contrasted with another important approach to statistical inference, the Bayesian.

A brief overview of statistical inference is considered in Section 11.4 as a prelude to the discussion of the next three chapters. The most important concept in statistical inference is the concept of a statistic which is discussed in Section 11.5. This concept and its distribution provide the cornerstone for estimation, testing and prediction.

## 11.2 The sampling model

As argued above, the probability model  $\Phi = \{f(x; \theta), \theta \in \Theta\}$  constitutes a very important component of statistical inference. Another important element in the same context is what we call a *sampling model*, which provides the link between the probability model and the observed data. It is designed to model the relationship between them and refers to the way the observed data can be viewed in relation to  $\Phi$ . In order to be able to formulate sampling models we need to define formally the concept of a sample in statistical inference.

*Definition 1*

**A sample** is defined to be a set of random variables (r.v.'s)  $(X_1, X_2, \dots, X_n)$  whose density functions coincide with the 'true' density function  $f(x; \theta_0)$  as postulated by the probability model.

Note that the term sample has a very precise meaning in this context and it is not the meaning attributed in everyday language. In particular the term does not refer to any observed data as the everyday use of the term might suggest.

The significance of the concept becomes apparent when we learn that the observed data in this context are considered to be one of the many possible realisations of the sample. In this interpretation lies the inductive argument of statistical inference which enables us to extend the results based on the observed data in hand to the underlying mechanism giving rise to them. Hence the observed data in this context are no longer just a set of numbers we want to make some sense of, they represent a particular outcome of an experiment; the experiment as defined by the sampling model postulated to complement the probability model  $\Phi = \{f(x; \theta), \theta \in \Theta\}$ .

Given that a sample is a set of r.v.'s related to  $\Phi$  it must have a distribution which we call the distribution of the sample.

*Definition 2*

**The distribution of the sample  $\mathbf{X} \equiv (X_1, \dots, X_n)'$**  is defined to be the joint distribution of the r.v.'s  $X_1, \dots, X_n$  denoted by

$$f(x_1, \dots, x_n; \theta) \equiv f(\mathbf{x}; \theta).$$

The distribution of the sample incorporates both forms of relevant information, the probability as well as sample information. It must come as no surprise to learn that  $f(\mathbf{x}; \theta)$  plays a very important role in statistical inference. The form of  $f(\mathbf{x}; \theta)$  depends crucially on the nature of the sampling model as well as  $\Phi$ . The simplest but most widely used form of a sampling model is the one based on the idea of a random experiment  $\mathcal{E}$  (see Chapter 3) and is called a random sample.

*Definition 3*

A set of random variables  $(X_1, X_2, \dots, X_n)$  is called a **random sample** from  $f(x; \theta)$  if the r.v.'s  $X_1, X_2, \dots, X_n$  are independent and identically distributed (IID). In this case the distribution of the

sample takes the form

$$f(x_1, x_2, \dots, x_n; \boldsymbol{\theta}) = \prod_{i=1}^n f(x_i; \boldsymbol{\theta}) = [f(x; \boldsymbol{\theta})]^n,$$

the first equality due to **independence** and the second to the fact that the r.v.'s are **identically distributed**.

One of the important ingredients of a random experiment  $\mathcal{E}$  is that the experiment can be repeated under identical conditions. This enables us to construct a random sample by repeating the experiment  $n$  times. Such a procedure of constructing a random sample might suggest that this is feasible only when experimentation is possible. Although there is some truth in this presupposition, the concept of a random sample is also used in cases where the experiment can be repeated under identical conditions, if only conceptually. In order to see this let us consider the personal income example where  $\Phi$  represents a Pareto family of density functions. 'What is a random sample in this case?' If we can ensure that every household in the UK has the same chance of being selected in one performance of a conceptual experiment then we can interpret the  $n$  households selected as representing a random sample  $(X_1, X_2, \dots, X_n)$  and their incomes (the observed data) as being a realisation of the sample. In general we denote the sample by  $\mathbf{X} \equiv (X_1, \dots, X_n)'$  and its realisation by  $\mathbf{x} \equiv (x_1, \dots, x_n)'$ , where  $\mathbf{x}$  is assumed to take values in the *observation space*  $\mathcal{X}$ , i.e.  $\mathbf{x} \in \mathcal{X}$ ; usually  $\mathcal{X} = \mathbb{R}^n$ .

A less restrictive form of a sampling model is what we call an independent sample, where the identically distributed condition in the random sample is relaxed.

#### *Definition 4*

A set of r.v.'s  $(X_1, \dots, X_n)$  is said to be an **independent sample** from  $f(x_i; \boldsymbol{\theta})$ ,  $i = 1, 2, \dots, n$ , respectively, if the r.v.'s  $X_1, \dots, X_n$  are independent. In this case the distribution of the sample takes the form

$$f(x_1, x_2, \dots, x_n; \boldsymbol{\theta}) = \prod_{i=1}^n f(x_i; \boldsymbol{\theta}_i). \quad (11.4)$$

Usually the density functions  $f(x_i; \boldsymbol{\theta}_i)$ ,  $i = 1, 2, \dots, n$  belong to the same family but their numerical characteristics (moments, etc.) may differ.

If we relax the independence assumption as well we have what we can call a non-random sample.

*Definition 5*

*A set of r.v.'s  $(X_1, \dots, X_n)$  is said to be a **non-random sample** from  $f(x_1, \dots, x_n; \boldsymbol{\theta})$  if the r.v.'s  $X_1, \dots, X_n$  are non-IID. In this case the only decomposition of the distribution of the sample possible is*

$$f(x_1, x_2, \dots, x_n; \boldsymbol{\theta}) = \prod_{i=1}^n f(x_i/x_1, \dots, x_{i-1}; \boldsymbol{\theta}_i), \quad (11.5)$$

*given  $x_0$ , where  $f(x_i/x_1, \dots, x_{i-1}; \boldsymbol{\theta}_i)$ ,  $i = 1, 2, \dots, n$ , represent the conditional distribution of  $X_i$  given  $X_1, X_2, \dots, X_{i-1}$ .*

A non-random sample is clearly the most general of the sampling models considered above and includes the independent and random samples as special cases given that

$$f(x_i/x_1, \dots, x_{i-1}; \boldsymbol{\theta}_i) = f(x_i; \boldsymbol{\theta}_i), \quad i = 1, 2, \dots, n, \quad (11.6)$$

when  $X_1, \dots, X_n$  are independent r.v.'s. Its generality, however, renders the concept non-operational unless certain restrictions are imposed on the heterogeneity and dependence among the  $X_i$ 's. Such restrictions have been extensively discussed in Sections 8.2–3. In Part IV the restrictions often used are stationarity and asymptotic independence.

In the context of statistical inference we need to postulate both a probability as well as a sampling model and thus we define a statistical model as comprising both.

*Definition 6*

*A **statistical model** is defined as comprising*

- (i)      a probability model  $\Phi = \{f(x; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ ; and
- (ii)     a sampling model  $\mathbf{X} \equiv (X_1, X_2, \dots, X_n)'$ .

The concept of a statistical model provides the starting point of all forms of statistical inference to be considered in the sequel. To be more precise, the concept of a statistical model forms the basis of what is known as parametric inference. There is also a branch of statistical inference known as non-parametric inference where no  $\Phi$  is assumed a priori (see Gibbons (1971)). Non-parametric statistical inference is beyond the scope of this book.

It must be emphasised at the outset that the two important components of a statistical model, the probability and sampling models, are clearly interrelated. For example, we cannot postulate the probability model  $\Phi =$

$\{f(x; \theta), \theta \in \Theta\}$  if the sample  $\mathbf{X}$  is non-random. This is because if the r.v.'s  $X_1, \dots, X_n$  are not independent the probability model must be defined in terms of their joint distribution, i.e.  $\Phi = \{f(x_1, \dots, x_n; \theta), \theta \in \Theta\}$ . Moreover, in the case of an independent but not identically distributed sample we need to specify the individual density functions for each r.v. in the sample, i.e.  $\Phi = \{f_k(x_k; \theta), \theta \in \Theta, k = 1, 2, \dots, n\}$ . The most important implication of this relationship is that when the sampling model postulated is found to be inappropriate it means that the probability model has to be respecified as well. Several examples of this are encountered in Chapters 21 to 23.

### 11.3 The frequency approach

In developing the concept of a probability model in Part II it was argued that no interpretation of probability was needed. The whole structure was built upon the axiomatic approach which defined probability as a set function  $P(\cdot): \mathcal{F} \rightarrow [0, 1]$  satisfying various axioms and devoid of any interpretations (see Section 3.2). In statistical inference, however, the interpretation of the notion of probability is indispensable. The discerning reader would have noted that in the above introductory discussion we have already adopted a particular attitude towards the meaning of probability. In interpreting the observed data as one of many possible realisations of the DGP as represented by the probability model we have committed ourselves towards the frequency interpretation of probability. This is because we implicitly assumed that if we were to repeat the experiment under identical conditions indefinitely (i.e. with the number of observations going to infinity) we would be able to reconstruct the probability model  $\Phi$ . In the case of the income example discussed above, this amounts to assuming that if we were to observe everybody's income and plot the relative frequency curve for incomes over £4500 we would get a Pareto density function. This suggests that the frequency approach to statistical inference can be viewed as a natural extension of the descriptive study of data with the introduction of the concept of a probability model. In practice we never have an infinity of observations in order to recover the probability model completely and hence caution should be exercised in interpreting the results of the frequency-approach-based statistical methods which we consider in the sequel. These results depend crucially on the probability model which we interpret as referring to a situation where we keep on repeating the experiment to infinity. This suggests that the results should be interpreted as holding under the same circumstances, i.e. 'in the long run' or 'on average'. Adopting such an interpretation implies that we should propose statistical procedures which give rise to 'optimum results' according to

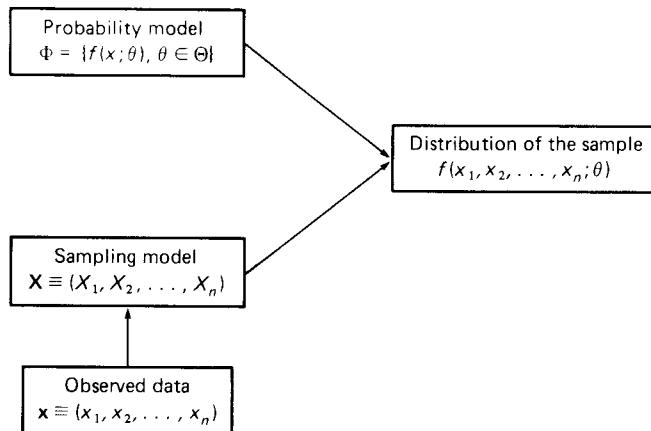


Fig. 11.1. The frequentist approach to statistical inference.

criteria related to this ‘long-run’ interpretation. Hence, it is important to keep this in mind when reading the following chapters on criteria for optimal estimators, tests and predictors.

The various approaches to statistical inference based on alternative interpretations of the notion of probability differ mainly in relation to *what constitutes relevant information* for statistical inference and *how it should be processed*. In the case of the *frequency approach* (sometimes called the classical approach) the relevant information comes in the form of a probability model  $\Phi = \{f(x; \theta), \theta \in \Theta\}$  and a sampling model  $X \equiv (X_1, X_2, \dots, X_n)'$ , providing the link between  $\Phi$  and the observed data  $x \equiv (x_1, x_2, \dots, x_n)'$ . The observed data are in effect interpreted as a realisation of the sampling model, i.e.  $X = x$ . This relevant information is then processed via the distribution of the sample  $f(x_1, x_2, \dots, x_n; \theta)$  (see Fig. 11.1).

The ‘*subjective*’ interpretation of probability, on the other hand, leads to a different approach to statistical inference. This is commonly known as the *Bayesian approach* because the discussion is based on revising prior beliefs about the unknown parameters  $\theta$  in the light of the observed data using Bayes’ formula. The prior information about  $\theta$  comes in the form of a probability distribution  $f(\theta)$ ; that is,  $\theta$  is assumed to be a random variable. The revision to the *prior*  $f(\theta)$  comes in the form of the *posterior distribution*  $f(\theta|x)$  via Bayes’ formula:

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)} \propto f(x|\theta)f(\theta). \quad (11.7)$$

$f(x|\theta)$  being the distribution of the sample and  $f(x)$  being constant for

$\mathbf{X} = \mathbf{x}$ . For more details and an excellent discussion of the frequency and Bayesian approaches to statistical inference see Barnett (1973). In what follows we concentrate exclusively on the frequency approach.

### 11.4 An overview of statistical inference

As defined above the simplest form of a statistical model comprises:

- (i) a probability model  $\Phi = \{f(x; \theta), \theta \in \Theta\}$ ; and
- (ii) a sampling model  $\mathbf{X} = (X_1, X_2, \dots, X_n)'$  – a random sample.

Using this simple statistical model, let us attempt a brief overview of statistical inference before we consider the various topics individually in order to keep the discussion which follows in perspective. The statistical model in conjunction with the observed data enable us to consider the following questions:

- (1) Are the observed data consistent with the postulated statistical model? (*misspecification*)
- (2) Assuming that the statistical model postulated is consistent with the observed data, what can we infer about the unknown parameters  $\theta \in \Theta$ ?
  - (a) Can we decrease the uncertainty about  $\theta$  by reducing the parameter space from  $\Theta$  to  $\Theta_0$  where  $\Theta_0$  is a subset of  $\Theta$ ? (*confidence estimation*)
  - (b) Can we decrease the uncertainty about  $\theta$  by choosing a particular value in  $\Theta$ , say  $\hat{\theta}$ , as providing the most representative value of  $\theta$ ? (*point estimation*)
  - (c) Can we consider the question that  $\theta$  belongs to some subset  $\Theta_0$  of  $\Theta$ ? (*hypothesis testing*)
- (3) Assuming that a particular representative value  $\hat{\theta}$  of  $\theta$  has been chosen what can we infer about further observations from the DGP as described by the postulated statistical model? (*prediction*)

The above questions describe the main areas of statistical inference. Comparing these questions with the ones we could ask in the context of descriptive statistics we can easily appreciate the role of probability theory in statistical inference.

The second question posed above (the first question is considered in the appendix below) assumes that the statistical model postulated is ‘valid’ and considers various forms of inference relating to the unknown parameters  $\theta$ .

*Point estimation (or just estimation)*: refers to our attempt to give a numerical value to  $\theta$ . This entails constructing a mapping  $\mathbf{h}(\cdot): \mathcal{X} \rightarrow \Theta$  (see Fig. 11.2). We call function  $\mathbf{h}(\mathbf{X})$  an *estimator* of  $\theta$  and its value  $\mathbf{h}(\mathbf{x})$  an

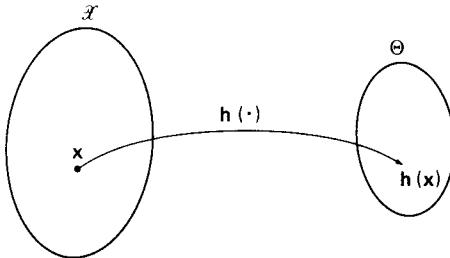


Fig. 11.2. Point estimation.

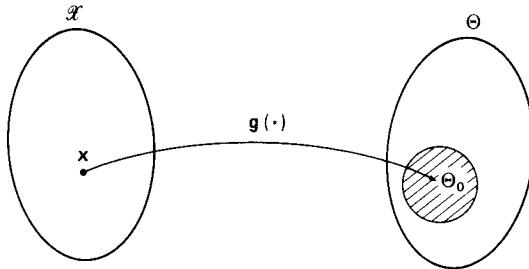


Fig. 11.3. Interval estimation.

*estimate* of  $\theta$ . Chapters 12 and 13 on point estimation deal with the issues of defining and constructing ‘optimal’ estimators, respectively.

*Confidence estimation:* refers to the construction of a numerical region for  $\theta$ , in the form of a subset  $\Theta_0$  of  $\Theta$  (see Fig. 11.3). Again, confidence estimation comes in the form of a multivalued function (one-to-many)  $g(\cdot): \mathcal{X} \rightarrow \Theta$ .

*Hypothesis testing*, on the other hand, relates to some a priori statement about  $\theta$  of the form  $H_0: \theta \in \Theta_0$ ,  $\Theta_0 \subset \Theta$ , against some opposite statement  $H_1: \theta \notin \Theta_0$  or, equivalently,  $\theta \in \Theta_1$ ,  $\Theta_1 \cap \Theta_0 = \emptyset$  and  $\Theta_1 \cup \Theta_0 = \Theta$ . In a situation like this we need to devise a rule which tells us when to accept  $H_0$  as ‘valid’ or reject  $H_1$  as ‘invalid’ in view of the observed data. Using the postulated partition of  $\Theta$  into  $\Theta_0$  and  $\Theta_1$  we need, in some sense, to construct a mapping  $q(\cdot): \mathcal{X} \rightarrow \Theta$  whose inverse image induces the partition

$$q^{-1}(\Theta_0) = C_0 - \text{acceptance region},$$

$$q^{-1}(\Theta_1) = C_1 - \text{rejection region},$$

where  $C_0 \cup C_1 = \mathcal{X}$  (see Fig. 11.4).

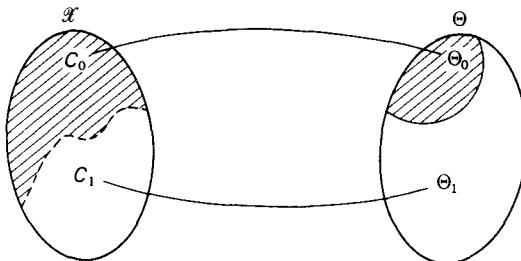


Fig. 11.4. Hypothesis testing.

The decision to *accept*  $H_0$  as a valid hypothesis about  $\theta$  or *reject*  $H_0$  as an invalid hypothesis about  $\theta$ , in view of the observed data, will be based on whether the observed data  $X$  belongs to the acceptance or rejection regions respectively, i.e.  $X \in C_0$  or  $X \in C_1$  (see Chapter 14).

Hypothesis testing can also be used to consider the question of the appropriateness of the probability model postulated. Apart from the direct test based on the empirical cumulative distribution function (see Appendix 11.1) we can use indirect tests based on characterisation theorems. For example, if a particular parametric family is characterised by the form of its first three moments, then we can construct a test based on these. For several characterisation results related to the normal distribution see Mathai and Pederzoli (1977). Similarly, hypothesis testing can be used to assess the appropriateness of the sampling model as well (see Chapter 22).

As far as question 3 is concerned we need to construct a mapping  $l(\cdot)$ :  $\Theta \rightarrow \mathcal{X}$  which will provide us with further values of  $X$  not belonging to the sample  $\mathbf{X}$ , for a given value of  $\theta$ .

## 11.5 Statistics and their distributions

As can be seen from the bird's-eye view of statistical inference considered in the previous section, the problem is essentially one of *constructing some mapping* of the form:

$$q(\cdot): \mathcal{X} \rightarrow \Theta \quad (11.8)$$

or its inverse, which satisfies certain criteria (restrictions) depending on the nature of the problem. Because of their importance in what follows such mappings will be given a very special name, we call them (sample) 'statistics'.

*Definition 7*

A **statistic** is said to be any Borel function (see Chapter 6)

$$q(\cdot): \mathcal{X} \rightarrow \mathbb{R},$$

Note that  $q(\cdot)$  does not depend on any unknown parameters.

Estimators, confidence intervals, rejection regions and predictors are all statistics which are directly related to the distribution of the sample. ‘Statistics’ are themselves random variables (r.v.’s) being Borel functions of r.v.’s or random vectors and they have their own distributions. The discussion of criteria for optimum ‘statistics’ is largely in terms of their distributions.

Two important examples of statistics which we will encounter on numerous occasions in what follows are:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i; \quad \text{called the } \textit{sample mean}, \quad (11.9)$$

and

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2; \quad \text{called the } \textit{sample variance}. \quad (11.10)$$

On the other hand, the functions

$$l_1(X) = \frac{1}{n} \sum_{i=1}^n \frac{X_i}{\sigma^2} \quad (11.11)$$

and

$$l_2(X) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \mu)^2 \quad (11.12)$$

are not statistics unless  $\sigma^2$  and  $\mu$  are known, respectively.

The concept of a statistic can be generalised to a vector-valued function of the form

$$\mathbf{q}(\cdot): \mathcal{X} \rightarrow \Theta = \mathbb{R}^m, \quad m \geq 1. \quad (11.13)$$

As with any random variable, any discussion relating to the nature of  $q(\mathbf{X})$  must be in terms of its distribution. Hence, it must come as no surprise to learn that statistical inference to a considerable extent depends critically on our ability to determine the distribution of a statistic  $q(\mathbf{X})$  from that of  $\mathbf{X} \equiv (X_1, X_2, \dots, X_n)$ , and determining such distributions is one of the most difficult problems in probability theory as Chapter 6 clearly exemplified. In that chapter we discussed various ways to derive the distribution function of  $Y = q(\mathbf{X})$

$$F(y) = Pr(q(\mathbf{X}) \leq y), \quad (11.14)$$

when the distribution of  $\mathbf{X}$  is known and several results have been derived. The reader can now appreciate the reason he/she had to put up with some rather involved examples. All the results derived in that chapter will form the backbone of the discussion that follows. The discerning reader must have noted that most of these results are related to *simple* functions  $q(\mathbf{X})$  of normally distributed r.v.'s,  $X_1, X_2, \dots, X_n$ . It turns out that most of the results in this area are related to this simple case. Because of the importance of the *normal distribution*, however, these results can take us a long way down 'statistical inference avenue'. Let us restate some of these results in terms of the statistics  $\bar{X}_n$  and  $s_n^2$  for reference purposes.

*Example 1*

Consider the following statistical model:

$$(i) \quad \Phi = \left\{ f(x; \boldsymbol{\theta}) = \frac{1}{\sigma\sqrt{(2\pi)}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \boldsymbol{\theta} \equiv (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+ \right\};$$

(ii)  $\mathbf{X} \equiv (X_1, X_2, \dots, X_n)'$  is a random sample from  $f(x; \boldsymbol{\theta})$ .

For the statistics

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

and

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

the following distributional results hold:

$$(i) \quad \bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right); \quad (11.15)$$

$$(ii) \quad \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim N(0, 1); \quad (11.16)$$

$$(iii) \quad n\left(\frac{\bar{X}_n - \mu}{\sigma}\right)^2 \sim \chi^2(1); \quad (11.17)$$

$$(iv) \quad (n-1)\frac{s_n^2}{\sigma^2} \sim \chi^2(n-1). \quad (11.18)$$

Note that

$$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2 = \left[ n\left(\frac{\bar{X}_n - \mu}{\sigma}\right)^2 + (n-1)\frac{s_n^2}{\sigma^2} \right] \sim \chi^2(n) \quad (11.19)$$

and

$$\text{Cov}(\bar{X}_n, s_n^2) = 0.$$

$$(v) \quad \frac{\sqrt{n}(\bar{X}_n - \mu)}{s_n} \sim t(n-1), \quad (11.20)$$

$$(vi) \quad \frac{(s_n^2/\sigma^2)}{(\tau_m^2/\tau^2)} \sim F(n-1, m-1), \quad (11.21)$$

where  $\tau_m^2$  is the corresponding sample variance of a random sample  $(Z_1, Z_2, \dots, Z_m)$  from  $N(\mu_2, \tau^2)$  and  $s_n^2, \tau_m^2$  are independent.

All these results follow from Lemmas 6.1–6.4 of Section 6.3 where the normal, chi-square, Student's  $t$  and  $F$  distributions are related.

Using the distribution of  $g(\mathbf{X})$ , when known, as in the above cases, we can consider questions relating to the nature of this statistic such as whether it provides a 'good' (to be defined in the sequel) or a 'bad' estimator or test statistic. Once this is decided we can go on to make probabilistic statements about  $\theta$ , the 'true' parameter of  $\Phi$ , which is what statistical inference is largely about. The question which naturally arises at this point is: 'What happens if we cannot determine the distribution of the statistic  $g(\mathbf{X})$ ?' Obviously, without a distribution for  $g(\mathbf{X})$  no statistical inference is possible and thus it is imperative to 'solve' the problem of the distribution somehow. In such cases *asymptotic theory* developed in Chapters 9 and 10 comes to our rescue by offering us 'second best' solutions in the form of approximations to the distribution of  $g(\mathbf{X})$ . In Chapter 6 we discussed various results related to the asymptotic behaviour of the statistic  $\bar{X}_n$  such as:

$$(i) \quad \bar{X}_n \xrightarrow{\text{a.s.}} \mu;$$

$$(ii) \quad \bar{X}_n \xrightarrow{\text{P}} \mu; \quad \text{and}$$

$$(iii) \quad \sqrt{n} \frac{(\bar{X}_n - \mu)}{\sigma} \xrightarrow{\text{D}} Z \sim N(0, 1); \quad (11.22)$$

irrespective of the original distribution of the  $X_i$ 's. Given only that  $E(X_i) = \mu$ ,  $\text{Var}(X_i) = \sigma^2 < \infty$ ; note that  $E(\bar{X}_n) = \mu$ . In Chapter 10 these results were extended to more general functions  $h(\mathbf{X})$ . In particular to continuous functions of the sample raw moments

$$m_r = \frac{1}{n} \sum_{i=1}^n X_i^r, \quad r \geq 1. \quad (11.23)$$

In relation to  $m_r$  it was shown that in the case of a random sample:

- (i)  $m_r \xrightarrow{\text{a.s.}} \mu'_r;$
- (ii)  $m_r \xrightarrow{\text{P}} \mu'_r; \text{ and}$
- (iii)  $\frac{\sqrt{n}(m_r - \mu'_r)}{\sigma_r} \xrightarrow{\text{D}} Z \sim N(0, 1);$  (11.24)

for

$$\mu'_r = \int_{-\infty}^{\infty} x^r f(x) dx \quad \text{with } E(m_r) = \mu'_r, \quad \sigma_r^2 = \mu'_{2r} - (\mu'_r)^2, \quad r \geq 1, \quad (11.25)$$

assuming that  $\mu'_{2r} < \infty.$

It turns out that in practice the statistics  $q(\mathbf{X})$  of interest are often functions of these sample moments. Examples of such continuous functions of the sample raw moments are the *sample central moments* being defined by

$$\hat{\mu}_r = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^r, \quad r \geq 1. \quad (11.26)$$

These provide us with a direct extension of the sample variance and they represent the sample equivalents to the central moments

$$\mu_r = \int_{-\infty}^{\infty} (x - \mu)^r f(x) dx. \quad (11.27)$$

With the help of asymptotic theory we could generalise the above asymptotic results related to  $m_r, r \geq 1,$  to those of  $Y_n = q(\mathbf{X})$  where  $q(\cdot)$  is a Borel function. For example, we could show that under the same conditions

- (i)  $\hat{\mu}_r \xrightarrow{\text{a.s.}} \mu_r;$
- (ii)  $\hat{\mu}_r \xrightarrow{\text{P}} \mu_r;$
- (iii)  $\frac{\sqrt{n}(\hat{\mu}_r - \mu_r)}{\sigma_r^*} \xrightarrow{\text{D}} Z \sim N(0, 1);$  (11.28)

where

$$\sigma_r^{*2} = \mu_{2r+2} - \mu_{r+1}^2 - 2(r+1)\mu_r\mu_{r+2} + (r+1)^2\mu_r^2\mu_2, \quad (11.29)$$

assuming that  $\mu_{2r} < \infty;$  see exercise 1.

Asymptotic results related to  $Y_n = q(\mathbf{X})$  can be used when the distribution of  $Y_n$  is not available (or very difficult to use). Although there are many ways to obtain asymptotic results in particular cases it is often natural to proceed by following the pattern suggested by the limit theorems in Chapter 9:

*Step 1*

Under certain conditions  $Y_n = q(\mathbf{X})$  can be shown to converge in probability to some function of  $h(\theta)$  of  $\theta$ , i.e.

$$Y_n \xrightarrow{\text{P}} h(\theta), \quad \text{or} \quad Y_n \xrightarrow{\text{a.s.}} h(\theta). \quad (11.30)$$

*Step 2*

Construct two sequences  $\{h_n(\theta), c_n(\theta), n \geq 1\}$  such that

$$Y_n^* = \left( \frac{Y_n - h_n(\theta)}{c_n(\theta)} \right) \xrightarrow{\text{D}} Z \sim N(0, 1). \quad (11.31)$$

Let  $F_\infty(y^*)$  denote the asymptotic distribution of  $Y_n^*$ , then *for large n*

$$F_n(y) \approx F_\infty(y^*), \quad (11.32)$$

and  $F_\infty(y^*)$  can be used as the basis of any inference relating to  $Y_n = q(\mathbf{X})$ . A question which naturally comes to mind is how large  $n$  should be to justify the use of these results. Commonly no answer is available because the answer would involve the derivation of  $F_n(y)$  whose unavailability was the very reason we had to resort to asymptotic theory. In certain cases higher-order approximations based on asymptotic expansions can throw some light on this question (see Chapter 10). In general, caution should be exercised when asymptotic results are used for relatively small values of  $n$ , say  $n < 100$ ?

### Appendix 11.1 – The empirical distribution function

The first question posed in Section 11.4 relates to the validity of the probability and sampling models postulated. One way to consider the validity of the probability model postulated is via the *empirical distribution function*  $F_n^*(x)$  defined by

$$F_n^*(x) = \frac{1}{n} (\text{number of } x_i \leq x), \quad x \in \mathbb{R}.$$

Alternatively, if we define the random variable (r.v.)  $Z_i$  to be

$$Z_i = \begin{cases} 1 & \text{if } x_i \in (-\infty, x], \quad x \in \mathbb{R}, \\ 0 & \text{otherwise,} \end{cases}$$

then  $F_n^*(x) = (1/n) \sum_{i=1}^n Z_i$ . If the original distribution postulated in  $\Phi$  is  $F(x)$ , a reasonable thing to do is to compare it with  $F_n^*(x)$ . For example, consider the distance

$$D_n = \max_{x \in \mathbb{R}} |F_n^*(x) - F(x)|,$$

$D_n$  as defined is a mapping of the form  $D_n(\cdot) : \mathcal{X} \rightarrow [0, 1]$  where  $\mathcal{X}$  is the observation space. Given that  $Z_i$  has a Bernoulli distribution  $F_n^*(x)$  being the sum of  $Z_1, Z_2, \dots, Z_n$  is binomially distributed, i.e.

$$\Pr\left(F_n^*(x) = \frac{k}{n}\right) = \binom{n}{k} [F(x)^k][1 - F(x)]^{n-k}, \quad k = 0, 1, \dots, n,$$

where  $E(F_n^*(x)) = F(x)$  and  $\text{Var}(F_n^*(x)) = (1/n)F(x)[1 - F(x)]$ . Using the central limit theorem (see Section 9.3) we can show that

$$\frac{\sqrt{n}(F_n^*(x) - F(x))}{\sqrt{\{F(x)[1 - F(x)]\}}} \xrightarrow{D} Z \sim N(0, 1).$$

Using this asymptotic result it can be shown that  $\sqrt{n} D_n \xrightarrow{D} y$  where

$$F(y) = \left[ 1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} \exp\{-2k^2y^2\} \right], \quad y \in \mathbb{R}_+.$$

This asymptotic distribution of  $\sqrt{n} D_n$  can be used to test the validity of  $\Phi$ ; see Section 21.2.

### Important concepts

Sample, the distribution of the sample, sampling model, random sample, independent sample, non-random sample, observation space, statistical model, empirical distribution function, point estimation, confidence estimation, hypothesis testing, a statistic, sample mean, sample variance, sample raw moments, sample central moments, the distribution of a statistic, the asymptotic distribution of a statistic.

### Questions

1. Discuss the difference between descriptive statistics and statistical inference.

2. Contrast  $f(x; \theta)$  as a descriptor of observed data with  $f(x; \theta)$  as a member of a parametric family of density functions.
3. Explain the concept of a sampling model and discuss its relationship to the probability model and the observed data.
4. Compare the sampling models:
  - (i) random sample;
  - (ii) independent sample;
  - (iii) non-random sample;
 and explain the form of the distribution of the sample in each case.
5. Explain the concept of the empirical distribution function.
6. ‘Estimation and hypothesis testing is largely a matter of constructing mappings of the form  $g(\cdot): \mathcal{X} \rightarrow \Theta$ .’ Discuss.
7. Explain why a statistic is a random variable.
8. Ensure that you understand the results (11.15)–(11.21) (see Appendix 6.1).
9. ‘Being able to derive the distribution of statistics of interest is largely what statistical inference is all about.’ Discuss.
10. Discuss the concept of a statistical model.

### **Exercises**

- 1.\* Using the results (22)–(29) show that for a random sample  $\mathbf{X}$  from a distribution whose first four moments exist,

$$\begin{pmatrix} \sqrt{n}(\bar{X}_n - \mu) \\ \sqrt{n}(s_n^2 - \sigma^2) \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \mu_3 \\ \mu^3 & \mu_4 - \sigma^4 \end{pmatrix}\right)$$

### **Additional references**

Barnett (1973); Bickel and Doksum (1977); Cramer (1946); Dudewicz (1976).

## CHAPTER 12

---

### Estimation I – properties of estimators

---

Estimation in what follows refers to point estimation unless indicated otherwise. Let  $(S, \mathcal{F}, P(\cdot))$  be the probability space of reference with  $X$  a r.v. defined on this space. The following statistical model is postulated:

- (i)  $\Phi = \{f(x; \theta), \theta \in \Theta\}, \quad \Theta \subseteq \mathbb{R};$
- (ii)  $\mathbf{X} \equiv (X_1, X_2, \dots, X_n)'$  is a random sample from  $f(x; \theta).$

Estimation in the context of this statistical model takes the form of constructing a mapping  $h(\cdot): \mathcal{X} \rightarrow \Theta$ , where  $\mathcal{X}$  is the observation space and  $h(\cdot)$  is a Borel function. The composite function (a statistic)  $\hat{\theta} \equiv h(\mathbf{X}): S \rightarrow \Theta$  is called an *estimator* and its value  $h(\mathbf{x})$ ,  $\mathbf{x} \in \mathcal{X}$  an *estimate*. It is important to distinguish between the two because the former is a random variable (r.v.) and the latter is a real number.

#### Example 1

Let  $f(x; \theta) = [1/\sqrt{(2n)}] \exp\{-\frac{1}{2}(x - \theta)^2\}$ ,  $\theta \in \mathbb{R}$ , and  $\mathbf{X}$  be a random sample from  $f(x; \theta)$ . Then  $\mathcal{X} = \mathbb{R}^n$  and the following functions define estimators of  $\theta$ :

- (i)  $\hat{\theta}_1 = \frac{1}{n} \sum_{i=1}^n X_i;$
- (ii)  $\hat{\theta}_2 = \frac{1}{k} \sum_{i=1}^k X_i, \quad k = 1, 2, \dots, n-1;$
- (iii)  $\hat{\theta}_3 = X_i, \quad i = 1, 2, \dots, n;$

$$(iv) \quad \hat{\theta}_4 = \frac{1}{n}(X_1 + X_n);$$

$$(v) \quad \hat{\theta}_5 = \frac{1}{n} \sum_{i=1}^n X_i^2;$$

$$(vi) \quad \hat{\theta}_6 = \frac{1}{n} \sum_{i=1}^n iX_i;$$

$$(vii) \quad \hat{\theta}_7 = \frac{1}{n+1} \sum_{i=1}^n X_i.$$

It is obvious that we can construct infinitely many such estimators. However, constructing ‘good’ estimators is not so obvious. From the above examples it is clear that we need some criteria to choose between these estimators. In other words, we need to formalise what we mean by a ‘good’ estimator. Moreover, it will be of considerable help if we could devise general methods of constructing such good estimators; a question considered in the next chapter.

## 12.1 Finite sample properties

In order to be able to set up criteria for choosing between estimators we need to understand the role of an estimator first. An estimator is constructed with the sole aim of providing us with the ‘most representative value’ of  $\theta$  in the parameter space  $\Theta$ , based on the available information in the form of the statistical model. Given that the estimator  $\hat{\theta} = h(\mathbf{X})$  is a r.v. (being a Borel function of a random vector  $\mathbf{X}$ ) any formalisation of what we mean by a ‘most representative value’ must be in terms of the distribution of  $\hat{\theta}$ , say  $f(\hat{\theta})$ . This is because any statement about ‘how near  $\hat{\theta}$  is to the true  $\theta$ ’ can only be a probabilistic one.

The obvious property to require a ‘good’ estimator  $\hat{\theta}$  of  $\theta$  to satisfy is that  $f(\hat{\theta})$  is centred around  $\theta$ .

*Definition 1*

*An estimator  $\hat{\theta}$  of  $\theta$  is said to be an **unbiased estimator** of  $\theta$  if*

$$E(\hat{\theta}) = \int_{-\infty}^{\infty} \hat{\theta} f(\hat{\theta}) d\hat{\theta}. \quad (12.1)$$

*That is, the distribution of  $\hat{\theta}$  has mean equal to the unknown parameter to be estimated.*

Note that an alternative, but equivalent, way to define  $E(\hat{\theta})$  is

$$E(\hat{\theta}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\mathbf{x}) f(\mathbf{x}; \theta) d\mathbf{x}, \quad (12.2)$$

where  $f(\mathbf{x}; \theta) = f(x_1, x_2, \dots, x_n; \theta)$  is the *distribution of the sample,  $\mathbf{X}$* .

Sometimes we can derive  $E(\hat{\theta})$  without having to derive either of the above distributions by just using the properties of  $E(\cdot)$  (see Chapter 4). For example, in the case of the estimators suggested in Section 12.1, using independence and the properties of the normal distribution we can deduce that  $\hat{\theta}_1 \sim N(\theta, (1/n))$ , this is because  $\hat{\theta}_1$  is a linear function of normally distributed r.v.'s (see Chapter 6.3), and

$$E(\hat{\theta}_1) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \sum_{i=1}^n \theta = \frac{n\theta}{n} = \theta \quad (12.3)$$

(see Fig. 12.1). The second equality due to independence and the property  $E(c) = c$  if  $c$  is a constant and the third equality because of the identically distributed assumption. Similarly for the variance of  $\hat{\theta}_1$

$$E(\hat{\theta}_1 - \theta)^2 = E\left(\frac{1}{n^2} \sum_{i=1}^n (X_i - \theta)^2\right) = \frac{1}{n^2} \sum_{i=1}^n E(X_i - \theta)^2 = \frac{1}{n^2} \sum_{i=1}^n 1 = \frac{1}{n}. \quad (12.4)$$

Using similar arguments we can deduce that

$$\hat{\theta}_2 \sim N\left(\theta, \frac{1}{k}\right), \quad k = 1, 2, \dots, n-1,$$

$$\hat{\theta}_3 \sim N(\theta, 1), \quad \hat{\theta}_4 \sim N\left(\frac{2\theta}{n}, \frac{2}{n^2}\right), \quad \hat{\theta}_5 \sim n\chi^2(n; n\theta^2),$$

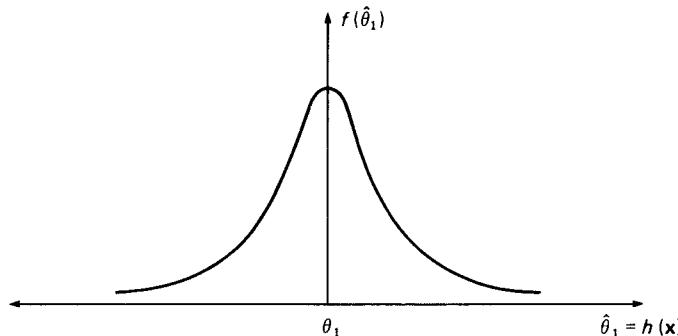


Fig. 12.1. The sampling distribution of  $\hat{\theta}_1$ .

$$\hat{\theta}_6 \sim N\left(\left(\frac{n+1}{2}\right)\theta, \frac{(n+1)(2n+1)}{6n}\right), \quad \hat{\theta}_7 \sim N\left(\frac{n}{n+1}\theta, \frac{n}{(n+1)^2}\right).$$

Hence, the estimators  $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$  are indeed unbiased but  $\hat{\theta}_4, \hat{\theta}_5, \hat{\theta}_6$  and  $\hat{\theta}_7$  are biased. We define *bias* to be  $B(\theta) = E(\hat{\theta}) - \theta$  and thus  $B(\hat{\theta}_4) = [(2-n)/n]\theta, B(\hat{\theta}_5) = n^2(1+\theta^2) - \theta, B(\hat{\theta}_6) = [(n-1)/2]\theta, B(\hat{\theta}_7) = -\theta/(n+1)$ . As can be seen from the above discussion, it is often possible to derive the mean of an estimator  $\hat{\theta}$  without having to derive its distribution. It must be remembered, however, that unbiasedness is a property based on the distribution of  $\hat{\theta}$ . This distribution is often called *sampling distribution* of  $\hat{\theta}$  in order to distinguish it from any other distribution of functions of r.v.'s.

Although unbiasedness seems at first sight to be a highly desirable property it turns out to be a rather severe restriction in some cases and in most situations there are too many unbiased estimators for this property to be used as the sole criterion for judging estimators. The question which naturally arises is, 'how can we choose among unbiased estimators?'. Returning to the above example, we can see that the unbiased estimators  $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$  have the same mean but they do not have the same variances. Given that the variance is a measure of dispersion, intuition suggests that the estimator with the smallest variance is in a sense better because its distribution is more 'concentrated' around  $\theta$ . This argument leads to the second property, that of relative efficiency.

### *Definition 2*

*An unbiased estimator  $\hat{\theta}_1$  of  $\theta$  is said to be relatively more efficient than some other unbiased estimator  $\hat{\theta}_2$  if*

$$\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2) \quad \text{or} \quad \text{eff}(\hat{\theta}_1 | \hat{\theta}_2) = \frac{\text{Var}(\hat{\theta}_1)}{\text{Var}(\hat{\theta}_2)} < 1.$$

In the above definition  $\hat{\theta}_1$  is relatively more efficient than either  $\hat{\theta}_2$  or  $\hat{\theta}_3$  since

$$\text{Var}(\hat{\theta}_1) = \frac{1}{n} < \frac{1}{k} = \text{Var}(\hat{\theta}_2), \quad k = 1, 2, \dots, n-1,$$

and

$$\text{Var}(\hat{\theta}_2) = \frac{1}{k} < 1 = \text{Var}(\hat{\theta}_3), \quad \text{for } k > 1.$$

i.e.  $\hat{\theta}_2$  is relatively more efficient than  $\hat{\theta}_3$  (see Fig. 12.2).

In the case of *biased* estimators relative efficiency can be defined in terms of the *mean square error* (MSE) which takes the form

$$E(\hat{\theta} - \theta)^2 = \text{Var}(\hat{\theta}) + [B(\hat{\theta})]^2, \tag{12.6}$$

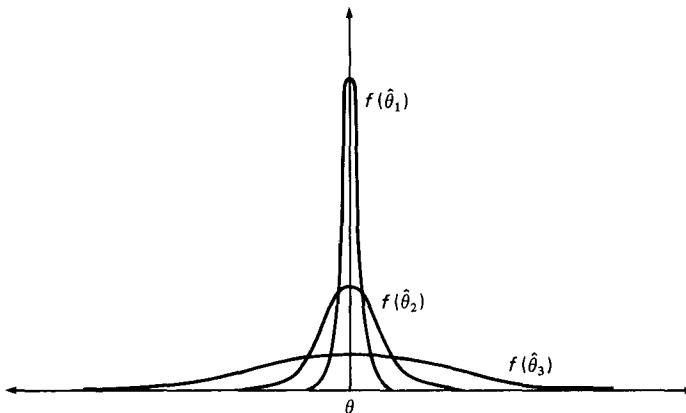


Fig. 12.2. The sampling distribution of  $\hat{\theta}_1$ ,  $\hat{\theta}_2$  and  $\hat{\theta}_3$ .

that is, an estimator  $\hat{\theta}^*$  is relatively more efficient than  $\hat{\theta}$  if

$$E(\hat{\theta}^* - \theta)^2 < E(\hat{\theta} - \theta)^2$$

or

$$\text{MSE}(\hat{\theta}^*) < \text{MSE}(\hat{\theta}).$$

As can be seen, this definition includes the definition in the case of unbiased estimators as a special case. Moreover, the definition in terms of the MSE enables us to compare an unbiased with a biased estimator in terms of efficiency. For example,

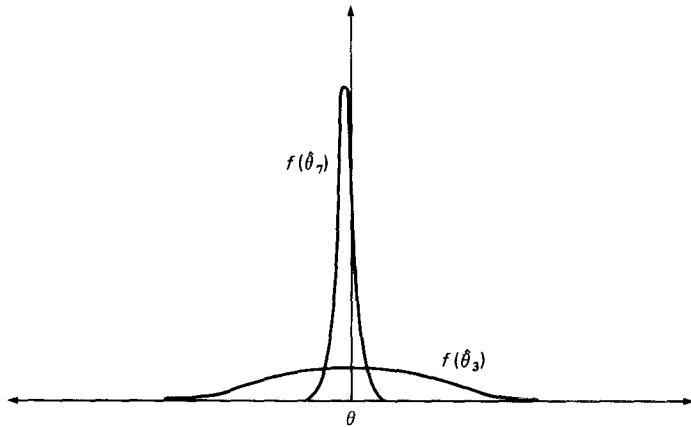
$$\text{MSE}(\hat{\theta}_7) < \text{MSE}(\hat{\theta}_3),$$

and intuition suggests that  $\hat{\theta}_7$  is a ‘better’ estimator than  $\hat{\theta}_3$  despite the fact that  $\hat{\theta}_3$  is unbiased and  $\hat{\theta}_7$  is not; Fig. 12.3 illustrates the case. In circumstances like this it seems a bit unreasonable to insist on unbiasedness. Caution should be exercised, however, when different distributions are involved as in the case of  $\hat{\theta}_5$ .

Let us consider the concept of MSE in some detail. As defined above, the MSE of  $\hat{\theta}$  depends not only on  $\hat{\theta}$  but the value of  $\theta$  in  $\Theta$  chosen as well. That is, for some  $\theta_0 \in \Theta$

$$\text{MSE}(\hat{\theta}, \theta_0) = E(\hat{\theta} - \theta_0)^2 = E[(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta_0)]^2 = \text{Var}(\hat{\theta}) + [B(\hat{\theta}, \theta_0)]^2, \quad (12.7)$$

the cross-product term being zero.  $B(\hat{\theta}, \theta_0) = E(\hat{\theta}) - \theta_0$  is the bias of  $\hat{\theta}$  relative to the value  $\theta_0$ . Using the MSE as a criterion for optimal estimators we would like to have an estimator  $\hat{\theta}$  with the smallest MSE for all  $\theta$  in  $\Theta$ .

Fig. 12.3. Comparing the sampling distribution of  $\hat{\theta}_7$  and  $\hat{\theta}_3$ .

(uniformly in  $\Theta$ ). That is,

$$\text{MSE}(\hat{\theta}, \theta) \leq \text{MSE}(\tilde{\theta}, \theta) \quad \text{for all } \theta \in \Theta,$$

where  $\tilde{\theta}$  denotes any other estimator of  $\theta$ . For any two estimators  $\hat{\theta}$  and  $\tilde{\theta}$  of  $\theta$  if  $\text{MSE}(\hat{\theta}, \theta) \leq \text{MSE}(\tilde{\theta}, \theta)$ ,  $\theta \in \Theta$  with strict inequality holding for some  $\theta \in \Theta$ ,  $\tilde{\theta}$  is said to be *inadmissible*. For example,  $\hat{\theta}_3$  above is inadmissible because

$$\text{MSE}(\hat{\theta}_1, \theta) < \text{MSE}(\hat{\theta}_3, \theta), \quad \text{for all } \theta \in \Theta \text{ if } n > 1.$$

In view of this we can see that  $\hat{\theta}_4$  and  $\hat{\theta}_5$  are inadmissible because in MSE terms are dominated by  $\hat{\theta}_7$ . The question which naturally arises is: 'Can we find an estimator which dominates every other in MSE terms?' A moment's reflection suggests that this is impossible because the MSE criterion depends on the value of  $\theta$  chosen. In order to see this let us choose a particular value of  $\theta$  in  $\Theta$ , say  $\theta_0$ , and define the estimator

$$\theta^* = \theta_0 \quad \text{for all } x \in \mathcal{X}.$$

Then  $\text{MSE}(\theta^*, \theta_0) = 0$  and any uniformly best estimator would have to satisfy  $\text{MSE}(\theta^*, \theta) = 0$  for all  $\theta \in \Theta$ , since  $\theta_0$  was arbitrarily chosen. That is, estimate  $\theta$  perfectly whatever its 'true' value! Who needs criteria in such a case? Hence, the fact that there are no uniformly best estimators in MSE terms is due to the nature of the problem itself.

Using the concept of relative efficiency we can compare different

estimators we happen to consider. This, however, is not very satisfactory since there might be much better estimators in terms of MSE for which we know nothing about. In order to be able to avoid choosing the better of two inefficient estimators we need some *absolute measure of efficiency*. Such a measure is provided by the *Cramer–Rao lower bound* which takes the form

$$CR(\theta) = \frac{\left[ 1 + \frac{dB(\theta)}{d\theta} \right]^2}{E\left[ \left( \frac{\partial \log f(\mathbf{x}; \theta)}{\partial \theta} \right)^2 \right]}, \quad (12.8)$$

$f(\mathbf{x}; \theta)$  is the distribution of the sample and  $B(\theta)$  the bias. It can be shown that for any estimator  $\theta^*$  of  $\theta$

$$\text{MSE}(\theta^*, \theta) \geq CR(\theta)$$

under the following regularity conditions on  $\Phi$ :

- (CR1) The set  $A = \{\mathbf{x}: f(\mathbf{x}; \theta) > 0\}$  does not depend on  $\theta$ .
- (CR2) For each  $\theta \in \Theta$  the derivatives  $[\partial^i \log f(\mathbf{x}; \theta)]/(\partial \theta^i)$ ,  $i = 1, 2, 3$ , exist for all  $\mathbf{x} \in A$ .
- (CR3)  $0 < E[(\partial/\partial \theta) \log f(\mathbf{x}; \theta)]^2 < \infty$  for all  $\theta \in \Theta$ .

In the case of *unbiased estimators* the inequality takes the form

$$\text{Var}(\theta^*) \geq E\left[ \left( \frac{\partial \log f(\mathbf{x}; \theta)}{\partial \theta} \right)^2 \right]^{-1};$$

the inverse of this lower bound is called *Fisher's information* and is denoted by  $I_n(\theta)$ .

### Definition 3

An **unbiased estimator**  $\hat{\theta}$  of  $\theta$  is said to be **(fully) efficient** if

$$\text{Var}(\hat{\theta}) = \left[ E\left( \frac{\partial \log f(\mathbf{x}; \theta)}{\partial \theta} \right)^2 \right]^{-1} \equiv I_n(\theta)^{-1}.$$

That is, an unbiased estimator is efficient when its variance equals the Cramer–Rao lower bound.

In the example considered above the distribution of the sample is

$$\begin{aligned} f(\mathbf{x}; \theta) &= \prod_{i=1}^n f(x_i; \theta) = \prod_{i=1}^n \left( \frac{1}{\sqrt{(2\pi)}} \exp\left(-\frac{1}{2}(x_i - \theta)^2\right) \right) \\ &= (2\pi)^{-n/2} \exp\left\{-\frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2\right\}, \end{aligned}$$

$$\begin{aligned}\log f(\mathbf{x}; \theta) &= -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2 \\ \Rightarrow \quad \frac{d \log f(\mathbf{x}; \theta)}{d\theta} &= \sum_{i=1}^n (x_i - \theta)\end{aligned}$$

and

$$E\left[\left(\frac{d \log f(\mathbf{x}; \theta)}{d\theta}\right)^2\right] = E\left[\left(\sum_{i=1}^n (x_i - \theta)\right)^2\right] = n \text{ by independence.}$$

An alternative way to derive the Cramer–Rao lower bound is to use the equality

$$E\left[\left(\frac{d \log f(\mathbf{x}; \theta)}{d\theta}\right)^2\right] = -E\left[\frac{d^2 \log f(\mathbf{x}; \theta)}{d\theta^2}\right], \quad (12.9)$$

which holds true under CR1–CR3 and  $f(\mathbf{x}; \theta)$  is the ‘true’ density function. In the above example  $[d^2 \log f(\mathbf{x}; \theta)]/(d\theta^2) = -n$  and hence the equality holds.

So, for this example,  $CR(\theta) = 1/n$  and, as can be seen from above, the only estimator which achieves the bound is  $\hat{\theta}_1$ , that is,  $\text{Var}(\hat{\theta}_1) = CR(\theta)$ ; hence  $\hat{\theta}_1$  is a fully efficient estimator. The properties of unbiasedness, relative efficiency and full efficiency enabled us to reduce the number of the originally suggested estimators considerably. Moreover, by narrowing the class of estimators considered, to the class of unbiased estimators, we succeeded in ‘solving’ the problem of ‘no uniformly best estimator’, discussed above in relation to the MSE criterion. This, however, is not very surprising given that by assuming unbiasedness we exclude the bias term which is largely responsible for the problem.

Sometimes the class of estimators considered is narrowed even further by requiring unbiased as well as *linear estimators*. That is, estimators which are linear functions of the r.v.’s of the sample. For example, in the case of example 1 above,  $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4, \hat{\theta}_6$  and  $\hat{\theta}_7$  are linear estimators. Within the class of linear and unbiased estimators we can show that  $\hat{\theta}_1$  has minimum variance. In order to show that let us take a general linear estimator

$$\tilde{\theta} = c + \sum_{i=1}^n a_i X_i, \quad (12.10)$$

which includes the above linear estimators as special cases, and determine the values for  $c, a_i, i = 1, 2, \dots, n$ , which ensure that  $\tilde{\theta}$  is *best linear unbiased estimator* (BLUE) of  $\mu$ . Firstly, for  $\tilde{\theta}$  to be unbiased we must have  $E(\tilde{\theta}) = 0$  which implies that  $c = 0$  and  $\sum_{i=1}^n a_i = 1$ . Secondly, since  $\text{Var}(\tilde{\theta}) = \sum_{i=1}^n a_i^2 \sigma^2 = \sigma^2 \sum_{i=1}^n a_i^2$  we must choose the  $a_i$ s so as to minimise  $\sum_{i=1}^n a_i^2$  as well as satisfy  $\sum_{i=1}^n a_i = 1$  (for unbiasedness). Setting up the Lagrangian for

this problem we have

$$\min_{a_i} l(\mathbf{a}, \lambda) = \sum_{i=1}^n a_i^2 - \lambda \left( \sum_{i=1}^n a_i - 1 \right), \quad (12.11)$$

$$\frac{\partial l}{\partial a_i} = 2a_i - \lambda = 0, \quad i = 1, 2, \dots, n, \quad \text{i.e. } a_i = \frac{\lambda}{2}.$$

Summing over  $i$ ,

$$\sum_{i=1}^n a_i = \sum_{i=1}^n \frac{\lambda}{2} = 1 \Rightarrow \lambda = \frac{2}{n}, \quad \text{i.e. } a_i = \frac{1}{n}, \quad i = 1, 2, \dots, n,$$

for  $c=0$  and  $a_i = 1/n$ ,  $i = 1, 2, \dots, n$ ,  $\hat{\theta} = (1/n) \sum_{i=1}^n X_i$ , which is identical to  $\hat{\theta}_1$ . Hence  $\hat{\theta}_1$  is BLUE (minimum variance among the class of linear and unbiased estimators of  $\mu$ ). This result will be of considerable interest in Chapter 21, in relation to the so-called Gauss–Markov theorem.

The properties of unbiasedness and efficiency can be generalised directly to the multiparameter case where  $\boldsymbol{\theta} \equiv (\theta_1, \dots, \theta_m)$ .  $\hat{\boldsymbol{\theta}}$  is said to be an unbiased estimator of  $\boldsymbol{\theta}$  if

$$E(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}, \quad \text{i.e. } E(\hat{\theta}_i) = \theta_i, \quad i = 1, 2, \dots, m.$$

In the case of full efficiency we can show that the Cramer–Rao inequality for an unbiased estimator  $\hat{\boldsymbol{\theta}}$  of  $\boldsymbol{\theta}$  takes the form

$$\text{Cov}(\hat{\boldsymbol{\theta}}) - \left[ E\left( \frac{\partial \log f(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \left( \frac{\partial \log f(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)' \right] \geq \mathbf{0} \quad (12.12)$$

or

$$\text{Var}(\hat{\theta}_i) \geq I_n(\boldsymbol{\theta})_{ii}^{-1}, \quad i = 1, 2, \dots, m$$

( $m$  being the number of parameters), where  $I_n(\boldsymbol{\theta})_{ii}^{-1}$  represents the  $i$ th diagonal element of the inverse of the matrix

$$\begin{aligned} I_n(\boldsymbol{\theta}) &= E\left[ \left( \frac{\partial \log f(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \left( \frac{\partial \log f(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)' \right] \\ &= E\left[ -\frac{\partial^2 \log f(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right], \end{aligned} \quad (12.13)$$

called the *sample information matrix*; the second equality holding under the restrictions CR1–CR3. In order to illustrate these, consider the following example:

*Example 2*

- (i)  $\Phi = \left\{ f(x; \theta) = \frac{1}{\sigma\sqrt{(2\pi)}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \theta \equiv (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+ \right\};$
- (ii)  $\mathbf{X} \equiv (X_1, X_2, \dots, X_n)'$  is a random sample from  $f(x; \theta)$ .

In example 1 discussed above we deduced that

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i \quad (12.14)$$

is a ‘good’ estimator of  $\mu$ , and intuition suggests that since  $\hat{\mu}$  is in effect the sample moment corresponding to  $\mu$  the sample variance

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu})^2 \quad (12.15)$$

should be a ‘good’ estimator of  $\sigma^2$ . In order to check our intuition let us examine whether  $\hat{\sigma}^2$  satisfies any of the above discussed properties.

$$\begin{aligned} E\left(\sum_{i=1}^n (X_i - \hat{\mu})^2\right) &= \left(\sum_{i=1}^n [(X_i - \mu) - (\hat{\mu} - \mu)]^2\right) \\ &= E\left(\sum_{i=1}^n [(X_i - \mu)^2 + (\hat{\mu} - \mu)^2 - 2(X_i - \mu)(\hat{\mu} - \mu)]\right). \end{aligned} \quad (12.16)$$

Since

$$E(X_i - \mu)^2 = \sigma^2, \quad E(\hat{\mu} - \mu)^2 = \frac{\sigma^2}{n} \quad \text{and} \quad E[(X_i - \mu)(\hat{\mu} - \mu)] = \frac{\sigma^2}{n}$$

from independence, we can deduce that

$$E\left[\sum_{i=1}^n (X_i - \hat{\mu})^2\right] = \sum_{i=1}^n \left(\sigma^2 + \frac{\sigma^2}{n} - 2\frac{\sigma^2}{n}\right) = (n-1)\sigma^2. \quad (12.17)$$

This, however, implies that  $E(\hat{\sigma}^2) = [(n-1)/n]\sigma^2 \neq \sigma^2$ , that is,  $\hat{\sigma}^2$  is a biased estimator of  $\sigma^2$ . Moreover, it is clear that the estimator

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \hat{\mu})^2 \quad (12.18)$$

is unbiased. From Chapter 6.3 we also know that

$$(n-1)\frac{s^2}{2} \sim \chi^2(n-1) \quad (12.19)$$

and thus

$$\text{Var}(s^2) = \frac{\sigma^4}{(n-1)^2} 2(n-1) = \frac{2\sigma^4}{n-1}, \quad (12.20)$$

since the variance of a chi-square r.v. equals twice its degrees of freedom. Let us consider the question whether  $\hat{\mu} = \bar{X}_n$  and  $s^2$  are efficient estimators:

$$\begin{aligned} f(\mathbf{x}; \theta) &= \prod_{i=1}^n f(x_i; \theta) \\ &= \prod_{i=1}^n \left[ \frac{1}{\sigma\sqrt{(2\pi)}} \exp \left\{ -\frac{1}{2} \left( \frac{x_i - \mu}{\sigma} \right)^2 \right\} \right] \\ &= \frac{(\sigma^2)^{-n/2}}{(2\pi)^{n/2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\}, \end{aligned} \quad (12.21)$$

$\Rightarrow$

$$\log f(\mathbf{x}; \theta) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2, \quad (12.22)$$

$$\frac{\partial \log f(\mathbf{x}; \theta)}{\partial \theta} = \begin{pmatrix} \frac{\partial \log f(\mathbf{x}; \theta)}{\partial \mu} \\ \frac{\partial \log f(\mathbf{x}; \theta)}{\partial \sigma^2} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu) \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (X_i - \mu)^2 \end{pmatrix}, \quad (12.23)$$

$$\begin{aligned} \frac{\partial^2 \log f(\mathbf{x}; \theta)}{\partial \theta \partial \theta} &= \begin{pmatrix} \frac{\partial^2 \log f(\mathbf{x}; \theta)}{\partial \mu^2} & \frac{\partial^2 \log f(\mathbf{x}; \theta)}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 \log f(\mathbf{x}; \theta)}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 \log f(\mathbf{x}; \theta)}{\partial \sigma^4} \end{pmatrix} \\ &= \begin{pmatrix} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4} \sum_i (X_i - \mu) \\ -\frac{1}{\sigma^4} \sum_i (X_i - \mu) & \frac{n}{2\sigma^4} + \frac{1}{\sigma^6} \sum_i (X_i - \mu)^2 \end{pmatrix}, \end{aligned} \quad (12.24)$$

$\Rightarrow$

$$\mathbf{I}_n(\theta) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} \text{ and } [\mathbf{I}_n(\theta)]^{-1} = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}. \quad (12.25)$$

This clearly shows that although  $\bar{X}_n$  achieves the Cramer–Rao lower bound  $s^2$  does not. It turns out, however, that no other unbiased estimator exists which is relatively more efficient than  $s^2$ ; although there are more efficient biased estimators such as

$$\hat{\sigma}_1^2 = \frac{1}{n+1} \sum_{i=1}^n (X_i - \bar{X}_n)^2. \quad (12.26)$$

Efficiency can be seen as a property indicating that the estimator ‘utilises’ all the information contained in the statistical model. An important concept related to the information of a statistical model is the concept of a *sufficient statistic*. This concept was introduced by Fisher (1922) as a way to reduce the sampling information by discarding only the information of no relevance to any inference about  $\theta$ . In other words, a statistic  $\tau(\mathbf{X})$  is said to be sufficient for  $\theta$  if it makes no difference whether we use  $\mathbf{X}$  or  $\tau(\mathbf{X})$  in inference concerning  $\theta$ . Obviously in such a case we would prefer to work with  $\tau(\mathbf{X})$  instead of  $\mathbf{X}$ , the former being of lower dimensionality.

#### *Definition 4*

A statistic  $\tau(\cdot): \mathcal{X} \rightarrow \mathbb{R}^m$ ,  $n > m$ , is called **sufficient for  $\theta$**  if the conditional distribution  $f(\mathbf{x}/\tau(\mathbf{x})=\tau)$  is independent of  $\theta$ , i.e.  $\theta$  does not appear in  $f(\mathbf{x}/\tau(\mathbf{x})=\tau)$  and the domain of  $f(\cdot)$  does not involve  $\theta$ .

In example 1 above intuition suggests that  $\tau(\mathbf{X}) = \sum_{i=1}^n X_i$  must be a sufficient statistic for  $\theta$  since in constructing a ‘very good’ estimator of  $\theta$ ,  $\hat{\theta}_1$ , we only needed to know the sum of the sample and not the sample itself. That is, as far as inference about  $\theta$  is concerned knowing all the numbers  $(X_1, X_2, \dots, X_n)$  or just  $\sum_{i=1}^n X_i$  makes no difference. Verifying this directly by deriving  $f(\mathbf{x}/\tau(\mathbf{x})=\tau)$  and showing that it is independent of  $\theta$  can be a very difficult exercise. One indirect way of verifying sufficiency is provided by the following lemma.

#### *Fisher–Neyman factorisation lemma*

The statistic  $\tau(\mathbf{X})$  is sufficient for  $\theta$  if and only if there exists a factorisation of the form

$$f(\mathbf{x}; \theta) = f(\tau(\mathbf{x}); \theta) \cdot h(\mathbf{x}), \quad (12.27)$$

where  $f(\tau(\mathbf{x}); \theta)$  is the density function of  $\tau(\mathbf{X})$  and depends on  $\theta$  and  $h(\mathbf{X})$ , some function of  $\mathbf{X}$  independent of  $\theta$ .

Even this result, however, is of no great help because we have to have the statistic  $\tau(\mathbf{X})$  as well as its distribution to begin with. The following method suggested by Lehmann and Scheffe (1950) provides us with a very convenient way to derive *minimal sufficient statistics*. A sufficient statistic  $\tau(\mathbf{X})$  is said to

be minimal if the sample  $\mathbf{X}$  cannot be reduced beyond  $\tau(\mathbf{X})$  without losing sufficiency. They suggested choosing an arbitrary value  $\mathbf{x}_0$  in  $\mathcal{X}$  and form the ratio

$$\frac{f(\mathbf{x}; \boldsymbol{\theta})}{f(\mathbf{x}_0; \boldsymbol{\theta})} = g(\mathbf{x}, \mathbf{x}_0; \boldsymbol{\theta}), \quad \mathbf{x} \in \mathcal{X}, \quad \boldsymbol{\theta} \in \Theta, \quad (12.28)$$

and the values of  $\mathbf{x}_0$  which make  $g(\mathbf{x}, \mathbf{x}_0; \boldsymbol{\theta})$  independent of  $\boldsymbol{\theta}$  are the required minimal sufficient statistics.

In example 2 above

$$g(\mathbf{x}, \mathbf{x}_0; \boldsymbol{\theta}) = \exp \left\{ -\frac{1}{2\sigma^2} \left[ \sum_{i=1}^n X_i^2 - \sum_{i=1}^n X_{i0}^2 \right] + \frac{\mu}{\sigma^2} \left[ \sum_{i=1}^n X_i - \sum_{i=1}^n X_{i0} \right] \right\}. \quad (12.29)$$

This clearly shows that  $\tau(\mathbf{X}) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$  is a minimal sufficient statistic since for these values of  $\mathbf{x}_0$ ,  $g(\mathbf{x}, \mathbf{x}_0; \boldsymbol{\theta}) = 1$ . Hence, we can conclude that  $(\bar{X}_n, s^2)$  being simple functions of  $\tau(\mathbf{X})$  are sufficient statistics. It is important to note that we cannot take  $\sum_{i=1}^n X_i$  or  $\sum_{i=1}^n X_i^2$  separately as minimal sufficient statistics; they are jointly sufficient for  $\boldsymbol{\theta} = (\mu, \sigma^2)$ .

In contrast to unbiasedness and efficiency, sufficiency is a property of statistics in general, not just estimators, and it is inextricably bound up with the nature of  $\Phi$ . For some parametric family of density functions such as the exponential family of distributions sufficient statistics exist, for other families they might not. Intuition suggests that, since efficiency is related to full utilisation of the information in the statistical model, and sufficiency can be seen as a maximal reduction of such information without losing any relevant information as far as inference about  $\boldsymbol{\theta}$  is concerned, there must be a direct relationship between the two properties. A relationship along the lines that when an efficient estimator is needed we should look no further than the sufficient statistics, is provided by the following lemma.

#### Rao and Blackwell lemma

Let  $\tau(\mathbf{X})$  be a sufficient statistic for  $\boldsymbol{\theta}$  and  $\mathbf{t}(\mathbf{X})$  be an estimator of  $\boldsymbol{\theta}$ , then

$$E(\mathbf{h}(\mathbf{X}) - \boldsymbol{\theta})^2 \leq E(\mathbf{t}(\mathbf{X}) - \boldsymbol{\theta})^2, \quad \boldsymbol{\theta} \in \Theta, \quad (12.30)$$

where  $\mathbf{h}(\mathbf{X}) = E(\mathbf{t}(\mathbf{X}) / \tau(\mathbf{X}) = \tau)$ , i.e. the conditional expectation of  $\mathbf{t}(\mathbf{X})$  given  $\tau(\mathbf{X}) = \tau$ .

From the above discussion of the properties of unbiasedness, relative and full efficiency and sufficiency we can see that these properties are directly

related to the distribution of the estimator  $\hat{\theta}$  of  $\theta$ . As argued repeatedly, deriving the distribution of Borel functions of r.v.'s such as  $\hat{\theta} = h(\mathbf{X})$  is a very difficult exercise and very few results are available in the literature. These results are mainly related to simple functions of normally distributed r.v.'s (see Section 6.3). For the cases where no such results are available (which is the rule rather than the exception) we have to resort to asymptotic results. This implies that we need to extend the above list of criteria for 'good' estimators to include *asymptotic properties* of estimators. These asymptotic properties will refer to the behaviour of  $\hat{\theta}$  as  $n \rightarrow \infty$ . In order to emphasise the distinction between these asymptotic properties and the properties considered so far we call the latter *finite sample* (or *small sample*) *properties*. The finite sample properties are related directly to the distribution of  $\hat{\theta}_n$ , say  $f(\hat{\theta}_n)$ . On the other hand, the asymptotic properties are related to the asymptotic distribution of  $\hat{\theta}_n$ .

## 12.2 Asymptotic properties

A natural property to require estimators to have is that as  $n \rightarrow \infty$  (i.e. as the sample size increases) the probability of  $\hat{\theta}$  being close to the true value  $\theta$  should increase as well. We formalise this idea using the concept of convergence in probability associated with the weak law of large numbers (WLLN) (see Section 9.2).

*Definition 5*

An estimator  $\hat{\theta}_n = h(\mathbf{X})$  is said to be **consistent** for  $\theta$  if

$$\lim_{n \rightarrow \infty} Pr(|\hat{\theta}_n - \theta| < \varepsilon) = 1, \quad (12.31)$$

and we write  $\hat{\theta}_n \xrightarrow{P} \theta$ .

This is in effect an extension of the WLLN for the sample mean  $\bar{X}_n$  to some Borel function  $h(\mathbf{X})$ . It is important to note that consistency does not refer to  $\hat{\theta}_n$  approaching  $\theta$  in the sense of mathematical convergence. The convergence refers to the probability associated with the event  $|\hat{\theta}_n - \theta| < \varepsilon$  as derived from the distribution of  $\hat{\theta}_n$  as  $n \rightarrow \infty$ . Moreover, consistency is a very minimal property (although a very important one) since if  $\hat{\theta}_n$  is a consistent estimator of  $\theta$  then so is  $\hat{\theta}_n^* = \hat{\theta}_n + 7405926/n$  if  $Pr(|\hat{\theta}_n - \theta| \geq 7405926/n) = 1/n, n > 1$ , which implies that for a small  $n$  the difference  $|\hat{\theta}_n - \theta|$  might be enormous, but the probability of this occurring decreasing to zero as  $n \rightarrow \infty$ .

Fig. 12.4 illustrates the concept in the case where  $\hat{\theta}_n$  has a well-behaved

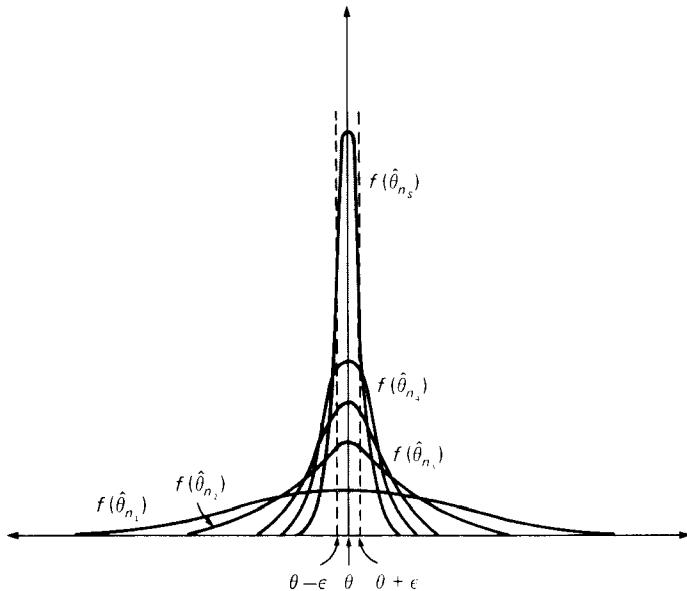


Fig. 12.4. Consistency in the case of a symmetric uniformly converging distribution.

symmetric distribution for  $n_1 < n_2 < n_3 < n_4 < n_5$ . This diagram seems to suggest that if the sampling distribution  $f(\hat{\theta}_n)$  becomes less and less dispersed as  $n \rightarrow \infty$  and eventually collapses at the point  $\theta$  (i.e. becomes degenerate), then  $\hat{\theta}_n$  is a consistent estimator of  $\theta$ . The following lemma formalises this argument.

*Lemma 12.1*

If  $\hat{\theta}_n$  is an estimator of  $\theta$  which satisfies the following properties

- (i)  $\lim_{n \rightarrow \infty} E(\hat{\theta}_n) = \theta$ ;
- (ii)  $\lim_{n \rightarrow \infty} \text{Var}(\hat{\theta}_n) = 0$ , then  $\hat{\theta}_n \xrightarrow{P} \theta$ .

It is important, however, to note that these are only *sufficient conditions* for consistency (*not necessary*); that is, consistency is not equivalent to the above conditions, since for consistency  $\text{Var}(\hat{\theta}_n)$  need not even exist. The above lemma, however, enables us to prove consistency in many cases of interest in practice. If we return to example 1 above we can see that

$$\hat{\theta}_1 \xrightarrow{P} \theta \quad \text{since } \Pr(|\hat{\theta}_1 - \theta| < \varepsilon) \geq 1 - \frac{1}{n\varepsilon^2}, \quad \varepsilon > 0,$$

by Chebyshev's inequality and  $\lim_{n \rightarrow \infty} [1 - (1/n\epsilon^2)] = 1$ . Alternatively, using Lemma 12.1 we can see that both conditions are satisfied. Similarly, we can show that  $\hat{\theta}_2 \xrightarrow{P} \theta$ ,  $\hat{\theta}_3 \not\xrightarrow{P} \theta$  (' $\not\xrightarrow{P}$ ' reads 'does not converge in probability to'),  $\hat{\theta}_4 \not\xrightarrow{P} \theta$ ,  $\hat{\theta}_5 \not\xrightarrow{P} \theta$ ,  $\hat{\theta}_6 \not\xrightarrow{P} \theta$ ,  $\hat{\theta}_7 \not\xrightarrow{P} \theta$ . Moreover, for  $\hat{\sigma}^2$  and  $s^2$  of example 2 we can show that  $\hat{\sigma}^2 \rightarrow \sigma^2$  and  $s^2 \rightarrow \sigma^2$ .

A stronger form of consistency associated with almost sure convergence is a very desirable asymptotic property for estimators.

#### *Definition 6*

An estimator  $\hat{\theta}_n$  is said to be a **strongly consistent estimator** of  $\theta$  if

$$\Pr\left(\lim_{n \rightarrow \infty} \hat{\theta}_n = \theta\right) = 1$$

and is denoted by  $\hat{\theta}_n \xrightarrow{a.s.} \theta$ .

The strong consistency of  $\hat{\theta}_n$  in example 1 is verified directly by the SLLN and that of  $s^2$  from the fact that it is a continuous function of the sample moments  $X_n$  and  $m_2 = (1/n) \sum_{i=1}^n X_i^2$  (see Chapter 10). Consistency and strong consistency can be seen as extensions of the weak law and strong law of large numbers for  $\sum_{i=1}^n X_i$  to the general statistic  $\hat{\theta}_n$ , respectively.

Extending the central limit theorem to  $\hat{\theta}_n$  leads to the property of asymptotic normality.

#### *Definition 7*

An estimator  $\hat{\theta}_n$  is said to be **asymptotically normal** if two sequences  $\{V_n(\theta), n \geq 1\}$ ,  $\{\theta_n, n \geq 1\}$  exist such that

$$(V_n(\theta))^{-\frac{1}{2}}(\hat{\theta}_n - \theta_n) \xrightarrow{D} Z \sim N(0, 1). \quad (12.32)$$

This way of defining asymptotic normality presents several logical problems deriving from the non-uniqueness of the sequences  $\{V_n(\theta)\}$  and  $\{\theta_n\}, n \geq 1$ . A more useful definition can be used in the case where the order of magnitude (see Section 10.4) of  $V_n(\theta)$  is known. In most cases of interest in practice such as the case of a random sample,  $V_n(\theta)$  is of order  $1/n$ , denoted by  $V_n(\theta) = O(1/n)$ . In such a case asymptotic normality can be written in the following form:

$$\sqrt{n}(\hat{\theta}_n - \theta_n) \underset{x}{\sim} N(0, V(\theta)), \quad (12.33)$$

where ' $\sim$ ' reads 'asymptotically distributed as' and  $V(\theta) > 0$  represents the

asymptotic variance. In relation to this form of asymptotic normality we consider two further asymptotic properties.

*Definition 8*

An estimator  $\hat{\theta}_n$  with  $\text{Var}(\hat{\theta}_n) = O(1/n)$  is said to be **asymptotically unbiased** if

$$\sqrt{n}(\hat{\theta}_n - \theta) \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (12.34)$$

This is automatically satisfied in the case of an asymptotically normal estimator  $\hat{\theta}_n$  for  $\text{Var}(\hat{\theta}_n) = V_n(\theta)$  and  $E(\hat{\theta}_n) = \theta$ . Thus, asymptotic normality can be written in the form

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, V(\theta)). \quad (12.35)$$

It must be emphasised that asymptotic unbiasedness is a stronger condition than  $\lim_{n \rightarrow \infty} E(\hat{\theta}_n) = \theta$ ; the former specifying the rate of convergence.

In relation to the variance of the asymptotic normal distribution we can define the concept of asymptotic efficiency.

*Definition 9*

An asymptotically normal estimator  $\hat{\theta}_n$  is said to be **asymptotically efficient** if  $V(\theta) = I_\infty(\theta)^{-1}$ , where

$$I_\infty(\theta) = \lim_{n \rightarrow \infty} \left( \frac{1}{n} I_n(\theta) \right), \quad (12.36)$$

i.e. the asymptotic variance achieves the limit of the Cramer–Rao lower bound (see Rothenberg (1973)).

At this stage it is important to distinguish between three different forms of the information matrix. The sample information matrix  $I_n(\theta)$  (see (13)), the single observation one  $I(\theta)$  with  $f(x_i; \theta)$  in (13), i.e.

$$E\left(\left(\frac{d \log f(x_i; \theta)}{d\theta}\right)^2\right)$$

and the asymptotic information matrix  $I_\infty(\theta)$  in (36).

## 12.3 Predictors and their properties

Consider the simple statistical model:

- (i) *Probability model:*  $\Phi = \{f(x; \theta) = 1/\sqrt{(2\pi)} \exp\{-\frac{1}{2}(x-\theta)^2\}, \theta \in \mathbb{R}\}$   
i.e.  $X \sim N(\theta, 1)$ .
- (ii) *Sampling model:*  $\mathbf{X} \equiv (X_1, X_2, \dots, X_n)'$  is a random sample.

Hence the distribution of the sample is

$$f(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta). \quad (12.37)$$

Prediction of the value of  $X$  beyond the sample observations, say  $X_{n+1}$ , refers to the construction of a Borel function  $l(\cdot)$  from the parameter space  $\Theta$  to the observation space  $\mathcal{X}$

$$l(\cdot): \Theta \rightarrow \mathcal{X}. \quad (12.38)$$

If  $\theta$  is known we can use the assumption that  $X_{n+1} \sim N(\theta, 1)$  to make probabilistic statements about  $X_{n+1}$ . Otherwise we need to estimate  $\theta$  first and then use it to construct  $l(\cdot)$ . In the present example we know from Sections 12.1 and 12.2 above that

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad (12.39)$$

is a ‘good’ estimator of  $\theta$ . Intuition suggests that a ‘good’ predictor of  $X_{n+1}$  might be to use  $l(\hat{\theta}_n) = \hat{\theta}_n$ , that is,

$$\hat{X}_{n+1} = \hat{\theta}_n. \quad (12.40)$$

The random variable  $\hat{X}_{n+1} = l(\hat{\theta}_n)$  is called *the predictor* of  $X_{n+1}$  and its value the prediction value. Note that the main difference between estimation and prediction is that in the latter case what we are ‘estimating’ ( $X_{n+1}$ ) is a random variable itself not a constant parameter  $\theta$ .

In order to consider the optimal properties of a predictor  $\hat{X}_{n+1} = l(\hat{\theta}_n)$  we define the prediction error to be

$$e_{n+1} = X_{n+1} - \hat{X}_{n+1}. \quad (12.41)$$

Given that both  $X_{n+1}$  and  $\hat{X}_{n+1}$  are random variables  $e_{n+1}$  is also a random variable and has its own distribution. Using the expectation operator with respect to the distribution of  $e_{n+1}$  we can define the following properties:

- (1) *Unbiasedness.* The predictor  $\hat{X}_{n+1}$  of  $X_{n+1}$  is said to be unbiased if

$$E(e_{n+1}) = 0. \quad (12.42)$$

- (2) *Minimum MSE.* The predictor  $\hat{X}_{n+1}$  of  $X_{n+1}$  is said to be minimum mean square error if

$$E(e_{n+1}^2) \equiv E(X_{n+1} - \hat{X}_{n+1})^2 \leq E(X_{n+1} - \tilde{X}_{n+1})^2 \quad (12.43)$$

for any other predictor  $\tilde{X}_{n+1}$  of  $X_{n+1}$ .

Another property of predictors commonly used in practice is linearity.

- (3) *Linear.* The predictor  $\hat{X}_{n+1}$  of  $X_{n+1}$  is said to be linear if  $l(\cdot)$  is a linear function of the sample.

In the case of the example considered above we can deduce that

$$e_{n+1} \sim N\left(0, 1 + \frac{1}{n}\right), \quad (12.44)$$

given that  $e_{n+1}$  is a linear function of normally distributed r.v.'s,  $e_{n+1} = X_{n+1} - (1/n) \sum_{i=1}^n X_i$ . Hence,  $\hat{X}_{n+1}$  is both linear and unbiased. Moreover, using the same procedure as in Section 13.1 for linear least-squares estimators, we can show that  $\hat{X}_{n+1}$  is also *minimum MSE* among the class of *linear unbiased predictors*.

The above properties of predictors are directly related to the same properties for estimators discussed in Section 12.1. This is not surprising, however, given that a predictor can be viewed as an 'estimator' of a random variable which does not belong to the sample.

### **Important concepts**

Estimator, estimate, unbiased estimator, bias, relative efficiency, mean square error, full efficiency, Cramer–Rao lower bound, information matrix, sufficient statistic, finite sample properties, asymptotic properties, consistency, strong consistency, asymptotic normality, asymptotic unbiasedness, asymptotic efficiency, BLUE.

### **Questions**

1. Define the concept of an estimator as a mapping and contrast it with the concept of an estimate.
2. Define the finite sample properties of unbiasedness, relative and full efficiency, sufficiency and explain their meaning.
3. 'Underlying every expectation operator  $E(\cdot)$  there is an implicit distribution.' Explain.
4. Explain the Cramer–Rao lower bound and the concept of the information matrix.
5. Explain the Lehmann–Scheffe method of constructing minimal sufficient statistics.
6. Contrast unbiasedness and efficiency with sufficiency.
7. Explain the difference between small sample and asymptotic properties.
8. Define and compare consistency and strong consistency.
9. Discuss the concept of asymptotic normality and its relationship to the order of magnitude of  $\text{Var}(\hat{\theta}_n)$ .

10. Explain the concept of asymptotic efficiency in relation to asymptotically normal estimators. 'What happens when the asymptotic distribution is not normal?'
11. Explain intuitively why  $\sqrt{n}(\hat{\theta}_n - \theta) \rightarrow 0$  as  $n \rightarrow \infty$  is a stronger condition than  $\lim_{n \rightarrow \infty} E(\hat{\theta}_n) = \theta$ .
12. Explain the relationships between  $I_n(\theta)$ ,  $I(\theta)$  and  $I_\infty(\theta)$ .

### *Exercises*

1. Let  $\mathbf{X} \equiv (X_1, X_2, \dots, X_n)'$  be a random sample from  $N(\theta, 1)$  and consider the following estimators of  $\theta$ :

$$\begin{aligned}\hat{\theta}_1 &= X_1, \quad \hat{\theta}_2 = \frac{1}{i} \sum_{j=1}^i jX_j, \quad i = 1, 2, \dots, n-1, \\ \hat{\theta}_3 &= \frac{1}{3}X_1 + \frac{2}{3}X_n, \quad \hat{\theta}_4 = \frac{1}{n^2} \sum_{i=1}^n iX_i, \quad \hat{\theta}_5 = \hat{\theta}_1 + \hat{\theta}_2.\end{aligned}$$

- (i) Derive the distribution of these estimators.
- (ii) Using these distributions consider the question whether these estimators satisfy the properties of unbiasedness, full efficiency and consistency.
- (iii) Choose the relatively most efficient estimator.
2. Consider the following estimator defined by

$$\begin{aligned}\hat{\theta}_n &= \frac{1}{n} \quad \text{and} \quad Pr\left(\hat{\theta}_n = \frac{1}{n}\right) = \frac{n}{n+1}, \\ \hat{\theta}_n &= n \quad \text{and} \quad Pr(\hat{\theta}_n = n) = \frac{1}{n+1},\end{aligned}$$

and show that:

- (i)  $\hat{\theta}_n$  as defined above has a proper sampling distribution;
- (ii)  $\hat{\theta}_n$  is a biased estimator of zero;
- (iii)  $\lim_{n \rightarrow \infty} \text{Var}(\hat{\theta}_n)$  does not exist; and
- (iv)  $\hat{\theta}_n$  is a consistent estimator of zero.
3. Let  $\mathbf{X} \equiv (X_1, X_2, \dots, X_n)'$  be a random sample from  $N(0, \sigma^2)$  and consider

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2$$

as an estimator of  $\sigma^2$ .

- (i) Derive the sampling distribution of  $\hat{\sigma}^2$  and show that it is an unbiased, consistent and fully efficient estimator of  $\sigma^2$ .

- (ii) Compare it with  $\hat{\sigma}^2$  of example 2 above and explain intuitively why the differences occur.
- (iii) Derive the asymptotic distribution of  $\hat{\sigma}^2$ .
4. Let  $\mathbf{X} = (X_1, \dots, X_n)'$  be a random sample from the exponential distribution with density

$$f(x; \theta) = \frac{1}{\theta} e^{-x/\theta}, \quad x > 0.$$

Construct a minimal sufficient statistic for  $\theta$  using the Lehmann–Scheffe method.

#### Additional references

Bickel and Doksum (1977); Cox and Hinkley (1974); Kendall and Stuart (1973); Lloyd (1984); Rao (1973); Rohatgi (1976); Silvey (1975); Zacks (1971).

## CHAPTER 13

---

### Estimation II – methods

---

The purpose of this chapter is to consider various methods for constructing ‘good’ estimators for the unknown parameters  $\theta$ . The methods to be discussed are the *least-squares method*, the *method of moments* and the *maximum likelihood method*. These three methods played an important role in the development of statistical inference from the early nineteenth century to the present day. The historical background is central to the discussion of these methods because they were developed in response to the particular demands of the day and in the context of different statistical frameworks. If we consider these methods in the context of the present-day framework of a statistical model as developed above we lose most of the early pioneers’ insight and the resulting anachronism can lead to misunderstanding. The method developed in relation to the contemporary statistical model framework is the maximum likelihood method attributed to Fisher (1922). The other two methods will be considered briefly in relation to their historical context in an attempt to delineate their role in contemporary statistical inference and in particular their relation to the method of maximum likelihood.

The method of maximum likelihood will play a very important role in the discussion and analysis of the statistical models considered in Part IV; a sound understanding of this method will be of paramount importance. After the discussion of the concepts of the likelihood function, maximum likelihood estimator (MLE) and score function we go on to discuss the properties of MLE’s. The properties of MLE’s are divided into finite-sample and asymptotic properties and discussed in the case of a random as well as a non-random sample. The latter case will be used extensively in Part IV. The actual derivation of MLE’s and their asymptotic distributions

is emphasised throughout as a prelude to the discussion of estimation in Part IV.

### 13.1 The method of least-squares

The method of least-squares was first introduced by Legendre in 1805 and Gauss in 1809 in the context of astronomical measurements. The problem as posed at the time was one of approximating a set of noisy observations  $y_i$ ,  $i = 1, 2, \dots, n$ , with some known functions  $g_i(\theta_1, \theta_2, \dots, \theta_m)$ ,  $i = 1, \dots, n$ , which depended on the unknown parameters  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)'$ ,  $m < n$ . Legendre argued that in the case of  $g_i(\boldsymbol{\theta}) = \theta_1$ ,  $i = 1, 2, \dots, n$ , minimising

$$\sum_{i=1}^n (y_i - \theta_1)^2, \quad \text{with respect to } \theta_1 \quad (13.1)$$

gives rise to  $\hat{\theta}_1 = (1/n) \sum y_i = \bar{y}_n$ , the sample mean, which was generally considered to be the most representative value of  $(y_1, y_2, \dots, y_n)$ . On the basis of this result he went on to suggest minimising the squared errors

$$l(\boldsymbol{\theta}) = \sum_{i=1}^n (y_i - g_i(\boldsymbol{\theta}))^2; \quad \text{the least-squares,} \quad (13.2)$$

in the general case. Assuming differentiability of  $g_i(\boldsymbol{\theta})$ ,  $i = 1, 2, \dots, n$ ,  $[\partial l(\boldsymbol{\theta})]/\partial \theta = 0$  gives rise to the so-called *normal equations* of the form

$$(-2) \sum_{i=1}^n [y_i - g_i(\boldsymbol{\theta})] \frac{\partial}{\partial \theta_k} g_i(\boldsymbol{\theta}) = 0, \quad k = 1, 2, \dots, m. \quad (13.3)$$

In this form the least-squares method has nothing to do with the statistical model framework developed above, it is merely an interpolation method in approximation theory.

Gauss, on the other hand, proposed a probabilistic set-up by reversing the Legendre argument about the mean. Crudely, his argument was that if  $\mathbf{X} \equiv (X_1, X_2, \dots, X_n)'$  is a random sample from some density function  $f(x)$  and the mean  $\bar{x}_n$  is the most representative value for all such  $X_i$ s, then the density function must be normal, i.e.

$$f(x) = \frac{1}{\sigma \sqrt{(2\pi)}} \exp \left\{ -\frac{1}{2\sigma^2} x^2 \right\}, \quad x \in \mathbb{R}. \quad (13.4)$$

Using this argument he went on to pose the problem in the form

$$\begin{aligned} y_i &= g_i(\boldsymbol{\theta}) + \varepsilon_i, \quad i = 1, 2, \dots, n, \\ \varepsilon_i &\sim NI(0, \sigma^2), \quad i = 1, 2, \dots, n. \end{aligned} \quad (13.5)$$

(Note: NI( $\cdot$ ) ‘reads’ normal independent, justifying the normality assumption on the grounds of being made up of a large number of independent factors cancelling each other out.) In this form the problem can be viewed as one of estimation in the context of the statistical model:

$$(i) \quad \Phi = \left\{ f(y_i; \theta) = \frac{1}{\sigma\sqrt{(2\pi)}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - g_i(\theta))^2\right\}, \theta \in \Theta \right\} \quad (13.6)$$

by transferring the probabilistic assumption from  $\varepsilon_i$  to  $y_i$ , the observable r.v., and

(ii) consider

$$\mathbf{y} = (y_1, y_2, \dots, y_n)'$$

to be an independent sample from  $f(y_i; \theta), i = 1, 2, \dots, n$ . Gauss went on to derive what we have called the distribution of the sample

$$f(\mathbf{y}; \theta) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - g_i(\theta))^2\right\}, \quad (13.7)$$

interpreting it as a function of  $\theta$ , and suggested that maximisation of  $f(\mathbf{y}; \theta)$  with respect to  $\theta$  gives rise to the same estimator of  $\theta$  as minimising the squared errors

$$\sum_{i=1}^n [y_i - g_i(\theta)]^2. \quad (13.8)$$

As we will see below, the above maximisation can be seen as a forerunner of the maximum likelihood method.

Since these early contributions the method of least-squares has been extended in various directions, both as an interpolation method as well as a statistical model with the probabilistic structure attached to the error term. The model mostly discussed in this literature is the linear model where  $g_i(\theta)$  is linear, i.e.

$$y_i = \sum_{k=1}^m \theta_k x_{ki} + \varepsilon_i, \quad i = 1, 2, \dots, n, \quad (13.9)$$

and the normality assumption replaced by the assumptions

$$E(\varepsilon_i) = 0, \quad \text{Var}(\varepsilon_i) = \sigma^2, \quad E(\varepsilon_i \varepsilon_j) = 0, \quad i \neq j, \quad i, j = 1, 2, \dots, n. \quad (13.10)$$

In some of the present-day literature this model is considered as an extension of the Gauss formulation by weakening the normality assumption. For further discussion of the Gauss linear model see Chapter 18.

For simplicity of exposition let us consider the case where  $m = 1$  and the model becomes

$$y_i = \theta_1 x_{1i} + \varepsilon_i, \quad i = 1, 2, \dots, n. \quad (13.11)$$

The least-squares method suggests minimising

$$l(\theta_1) = \sum_{i=1}^n (y_i - \theta_1 x_{1i})^2, \quad (13.12)$$

with respect to  $\theta_1$ . The normal equations in this case take the form

$$\begin{aligned} \frac{dl(\theta_1)}{d\theta_1} &= (-2) \sum_{i=1}^n x_{1i}(y_i - \theta_1 x_{1i}) = 0, \\ \Rightarrow \theta_1 &= \frac{\sum_{i=1}^n x_{1i} y_i}{\sum_{i=1}^n x_{1i}^2} \end{aligned} \quad (13.13) \quad (13.14)$$

is the least-squares estimator of  $\theta_1$ .

Note that in the case where  $x_{1i} = 1, i = 1, 2, \dots, n$ ,  $\hat{\theta}_1 = (1/n) \sum_{i=1}^n y_i$ , i.e. the sample mean.

Given that the  $x_{1i}$ 's are in general assumed to be known constants  $\hat{\theta}_1$  is a linear function of the r.v.'s  $y_1, \dots, y_n$  of the form

$$\hat{\theta}_1 = \sum_{i=1}^n c_i y_i \quad (13.15)$$

where

$$c_i = \frac{x_{1i}}{\sum_{i=1}^n x_{1i}^2}.$$

Hence,

$$E(\hat{\theta}_1) = \sum_{i=1}^n c_i E(y_i) = \sum_{i=1}^n c_i x_{1i} \theta_1 = \theta_1, \quad (13.16)$$

i.e.  $\hat{\theta}_1$  is an unbiased estimator of  $\theta_1$ . Moreover, since  $(\hat{\theta}_1 - \theta_1) = \sum_{i=1}^n x_{1i} \varepsilon_i$ ,

$$\text{Var}(\hat{\theta}_1) = \frac{\sigma^2}{\sum_{i=1}^n x_{1i}^2}. \quad (13.17)$$

It can be shown that, under the above assumptions relating to  $\varepsilon_i$  if  $(\sum_{i=1}^n x_{1i}^2) \neq 0$ , the least-squares estimator  $\hat{\theta}_1$  of  $\theta_1$  has the smallest variance

within the class of linear and unbiased estimators (Gauss–Markov theorem, see Section 21.2).

### 13.2      The method of moments

From the discussion in the previous section it is clear that the least-squares method is not a general method of estimation because it presupposes the existence of approximating functions  $g_i(\theta), i=1, 2, \dots, n$ , which play the role of the mean in the context of a probability model. In the context of a probability model  $\Phi$ , however, unknown parameters of interest are not only associated with the mean but also with the higher moments. This prompted Pearson in 1894 to suggest the *method of moments* as a general estimation method. The idea underlying the method can be summarised as follows:

Let us assume that  $\mathbf{X} = (X_1, X_2, \dots, X_n)'$  is a random sample from  $f(x; \theta)$ ,  $\theta \in \mathbb{R}^k$ . The *raw moments* of  $f(x; \theta)$ ,  $\mu'_r \equiv E(x^r)$ ,  $r \geq 1$ , are by definition functions of the unknown parameters, since

$$\mu'_r(\theta) = \int_{-\infty}^{\infty} x^r f(x; \theta) dx, \quad r \geq 1. \quad (13.18)$$

In order to apply the method we need to express the unknown parameters  $\theta$  in the form

$$\theta_i = g_i(\mu'_1, \mu'_2, \dots, \mu'_k), \quad i = 1, 2, \dots, k, \quad (13.19)$$

where the  $g_i$ s are continuous functions. The method of moments based on the substitution idea, proposes estimating  $\theta_i$  using

$$\hat{\theta}_i = g_i(m_1, m_2, \dots, m_k), \quad i = 1, 2, \dots, k, \quad (13.20)$$

where  $m_r = (1/n) \sum_{i=1}^n X_i^r$ ,  $r \geq 1$ , represent the *sample raw moments*, as the estimators of  $\hat{\theta}_i$ ,  $i = 1, 2, \dots, k$ . The justification of the method is based on the fact that if  $\mu'_1, \dots, \mu'_k$  are one-to-one functions of  $\theta$  then since

$$m_r \xrightarrow{\text{a.s.}} \mu'_r, \quad r \geq 1, \quad (13.21)$$

it follows that

$$\hat{\theta}_i \xrightarrow{\text{a.s.}} \theta_i, \quad i = 1, 2, \dots, k \quad (13.22)$$

(see Chapter 10).

#### *Example 1*

Let  $X_i \sim N(\mu, \sigma^2)$ ,  $i = 1, 2, \dots, n$ , then  $\mu'_1 = \mu$ ,  $\mu'_2 = \sigma^2 + \mu^2$  and  $m_1 =$

$(1/n) \sum_{i=1}^n X_i$ ,  $m_2 = (1/n) \sum_{i=1}^n X_i^2$ . The method suggests

$$\hat{\mu} = m_1 = \bar{X}_n,$$

$$\hat{\sigma}^2 = m_2 - (m_1)^2 = \frac{1}{n} \sum_i X_i^2 - (\bar{X}_n^2) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n^2).$$

Formally, the above result can be stated as follows: Let the functions  $\mu'_i(\theta)$ ,  $i = 1, 2, \dots, k$ , have continuous partial derivatives up to order  $k$  on  $\Theta$  and the Jacobian of the transformation

$$\det \left| \frac{\partial(\mu_1, \dots, \mu_k)}{\partial(\theta_1, \dots, \theta_k)} \right| \neq 0 \quad \text{for } \theta \in \Theta. \quad (13.23)$$

If the equations  $\mu'_i(\theta) = m_i$ ,  $i = 1, 2, \dots, k$ , have a unique solution  $\hat{\theta}_n \equiv (\hat{\theta}_1, \dots, \hat{\theta}_k)'^P$  with probability approaching one, as  $n \rightarrow \infty$ , then  $\hat{\theta} \xrightarrow{P} \theta$  (i.e.  $\hat{\theta}_n$  is a consistent estimator of  $\theta$ ).

Although the method of moments usually yields (strongly) consistent estimators they are in general inefficient. This was taken up by Fisher in several papers in the 1920s and 30s arguing in favour of the maximum likelihood method for producing efficient estimators (at least asymptotically). The controversy between Pearson and Fisher about the relative merits of their respective methods of estimation ended in the mid-1930s with Fisher the winner and the absolute dominance since then of the maximum likelihood method.

The basic reason for the inefficiency of the estimators based on the method of moments is not hard to find. It is due to the fact that the method does not use any information relating to the probability model  $\Phi$  apart from the assumption that raw moments of order  $k$  exist. It is important, however, to remember that this method was proposed by Pearson in the late nineteenth century when no such probability model was postulated a priori. The problem of statistical inference at the time was seen as one starting from a sample  $\mathbf{X} = (X_1, \dots, X_n)'$  and estimating  $f(\mathbf{x}; \theta)$  without assuming a priori some form for  $f(\cdot)$ . This point is commonly missed when comparisons between the various methods are made; it was unfortunately missed even by Pearson himself in his exchanges with Fisher. It is no surprise then to discover that a method developed in the context of an alternative framework when applied to present-day set-up is found wanting.

### 13.3 The maximum likelihood method

The maximum likelihood method of estimation was formulated by Fisher

in a series of papers in the 1920s and 30s and extended by various authors such as Cramer, Rao and Wald. In the current statistical literature the method of maximum likelihood is by far the most widely used method of estimation and plays a very important role in hypothesis testing.

### (1)      *The likelihood function*

Consider the statistical model:

- (i)       $\Phi = \{f(x; \theta), \theta \in \Theta\};$
- (ii)      $\mathbf{X} = (X_1, X_2, \dots, X_n)' \text{ a sample from } f(x; \theta),$

where  $\mathbf{X}$  takes values in  $\mathcal{X} = \mathbb{R}^n$ , the observation space. The distribution of the sample  $D(x_1, x_2, \dots, x_n; \theta)$  describes how the density changes as  $\mathbf{X}$  takes different values in  $\mathcal{X}$  for a given  $\theta \in \Theta$ . In deriving the likelihood function we reason as follows:

since  $D(\mathbf{x}; \theta)$  incorporates all the information in the statistical model it makes a lot of intuitive sense to reverse the argument in deriving  $D(\mathbf{x}; \theta)$  and consider the question which value of  $\theta \in \Theta$  is mostly supported by a given sample realisation  $\mathbf{X} = \mathbf{x}$ ?

#### *Example 1*

Consider the case where  $\mathbf{X}$  is a random sample from a Bernoulli distribution with density function

$$f(x; \theta) = \theta^x(1 - \theta)^{1-x}, \quad x = 0, 1, \quad (13.24)$$

and for simplicity assume that the ‘true’  $\theta$  can only be either  $\theta = 0.2$  or  $\theta = 0.8$ . Suppose the sample realisation was

$$\mathbf{x} = (1, 0, 1, 0, 1, 1, 1, 0, 1, 1),$$

what can we say about the two possible values of  $\theta$ ? Intuition suggests that since the average of the observed realisation is 0.7 it is more reasonable to assume that  $\mathbf{x}$  is a sample realisation from  $f(x; 0.8)$  rather than  $f(x; 0.2)$ . That is, the value of  $\theta$  under which  $\mathbf{x}$  would have had the highest ‘likelihood’ of arising must be intuitively our best choice of  $\theta$ . Using this intuitive argument the likelihood function is defined by

$$L(\theta; \mathbf{x}) = k(\mathbf{x})D(\mathbf{x}; \theta), \quad \theta \in \Theta, \quad (13.25)$$

where  $k(\mathbf{x}) > 0$  is a function of  $\mathbf{x}$  only (not  $\theta$ ). In particular

$$L(\cdot; \mathbf{x}): \Theta \rightarrow [0, \infty). \quad (13.26)$$

Even though the probability remains attached to  $\mathbf{X}$  and not  $\theta$  in defining  $L(\theta; \mathbf{x})$  it is interpreted as if it is reflected inferentially on  $\theta$ ; reflecting the 'likelihood' of a given  $\mathbf{X} = \mathbf{x}$  arising for different values of  $\theta$  in  $\Theta$ . In order to see this, consider the following example.

*Example 2*

Let  $\mathbf{X} \equiv (X_1, \dots, X_n)'$  be a random sample from  $N(0, \theta)$ ;  $\theta = \sigma^2$ . Because of the randomness of the sample its distribution takes the form

$$\begin{aligned} D(x_1, \dots, x_n; \theta) &= \prod_{i=1}^n \left[ \frac{1}{\sqrt{(2\pi\theta)}} \exp\left\{-\frac{x_i^2}{2\theta}\right\} \right] \\ &= (2\pi\theta)^{-n/2} \prod_{i=1}^n \exp\left\{-\frac{x_i^2}{2\theta}\right\}, \end{aligned}$$

and the likelihood function is

$$L(\theta; \mathbf{x}) = k(\mathbf{x})(2\pi\theta)^{-n/2} \prod_{i=1}^n \exp\left\{-\frac{x_i^2}{2\theta}\right\}.$$

If we were to reduce the dimensionality of  $\mathbf{x}$  to one (to enable us to draw pictures) the derivation of  $L(\theta; \mathbf{x})$  is shown in Fig. 13.1 for two sample realisations  $\mathbf{X} = \mathbf{x}_1$  and  $\mathbf{X} = \mathbf{x}_2$ . Fig. 13.1(a) shows a family of  $D(\mathbf{x}; \theta)$  for  $\theta = 0.5, 1, 2, 3, 4$  and for two given values of  $\mathbf{X}$ ,  $\mathbf{x}_1$  and  $\mathbf{x}_2$  the likelihood functions  $L(\theta; \mathbf{x}_1)$  and  $L(\theta; \mathbf{x}_2)$  are reflected in Fig. 13.1(b). As can be seen from these diagrams, different sample realisations provides different 'likelihoods' for  $\theta \in \Theta$ . In the derivation of these likelihoods the constant  $k(\mathbf{x})$  was chosen arbitrarily to be equal to one. The presence of  $k(\mathbf{x})$  in the definition of  $L(\theta; \mathbf{x})$

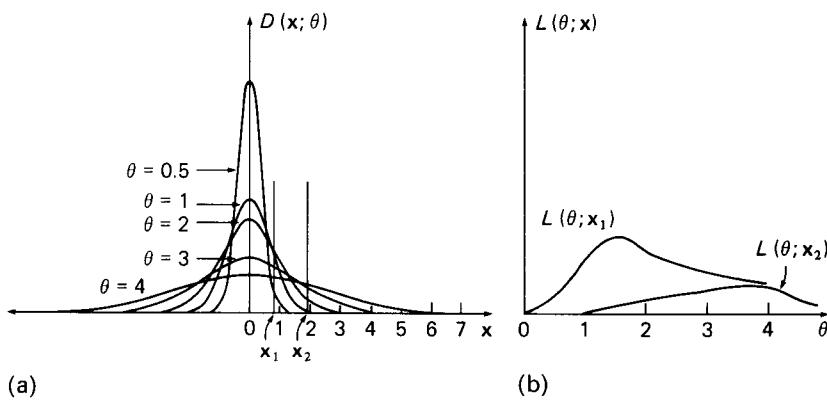


Fig. 13.1. Deriving the likelihood function from the distribution of the sample.

implies that the likelihood function is non-unique; any monotonic transformation of it represents the same information. In particular:

$$(i) \quad \log L(\theta; \mathbf{x}), \text{ the } \log \text{likelihood function; and} \quad (13.27)$$

$$(ii) \quad \frac{\partial \log L(\theta; \mathbf{x})}{\partial \theta} \equiv \mathbf{s}(\theta; \mathbf{x}), \text{ the score function,} \quad (13.28)$$

incorporate the same information as  $L(\theta; \mathbf{x})$  itself; if we have any one of the functions  $L(\theta; \mathbf{x})$ ,  $\log L(\theta; \mathbf{x})$ ,  $\mathbf{s}(\theta; \mathbf{x})$  we can derive the others.

In example 2 above

$$\begin{aligned} \log L(\theta; \mathbf{x}) &= \log k(\mathbf{x}) - \frac{n}{2} \log 2\pi - \frac{n}{2} \log \theta - \frac{1}{2\theta} \sum_{i=1}^n X_i^2 \\ &= c - \frac{n}{2} \log \theta - \frac{1}{2\theta} \sum_{i=1}^n X_i^2 \end{aligned}$$

and

$$\frac{\partial \log L(\theta; \mathbf{x})}{\partial \theta} = -\frac{n}{2\theta} + \frac{1}{2\theta^2} \sum_{i=1}^n X_i^2.$$

Hence,

$$\int_0^\infty \frac{d \log L(\theta; \mathbf{x})}{d \theta} d\theta = -\frac{n}{2} \log \theta - \frac{1}{2\theta} \sum_{i=1}^n X_i^2 + c^*$$

and

$$\exp\{\log L(\theta; \mathbf{x})\} = c^* \theta^{-n/2} \prod_{i=1}^n \left( -\frac{X_i^2}{2\theta} \right).$$

As will be seen in the sequel, although the proportionality factor is indispensable in the actual definition of the likelihood function, it plays no role in the derivation of estimators. In what follows working with  $\log L(\theta; \mathbf{x})$  and  $[\partial \log L(\theta; \mathbf{x})]/\partial \theta$  will prove more convenient than using  $L(\theta; \mathbf{x})$  itself; see Figs. 13.2–13.4.

## (2)      ***The maximum likelihood estimator (MLE)***

Given that the likelihood function represents the support given to the various  $\theta \in \Theta$  given  $\mathbf{X} = \mathbf{x}$ , it is natural to define the *maximum likelihood estimator* of  $\theta$  to be a Borel function  $\hat{\theta}: \mathcal{X} \rightarrow \Theta$  such that

$$L(\hat{\theta}; \mathbf{x}) = \max_{\theta \in \Theta} L(\theta; \mathbf{x}), \quad (13.29)$$

and there may be one, none or many such MLE's.

### 13.3 The maximum likelihood method

261

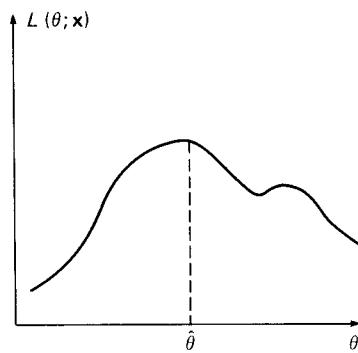


Fig. 13.2. A likelihood function.

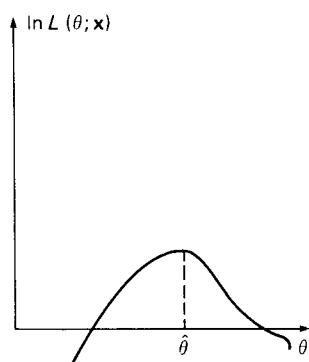


Fig. 13.3. The log-likelihood function.

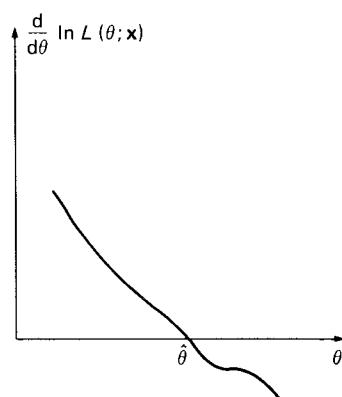


Fig. 13.4. The score function.

Note that

$$\log L(\hat{\theta}; \mathbf{x}) \geq \log L(\theta^*; \mathbf{x}), \quad \text{for all } \theta^* \in \Theta. \quad (13.30)$$

In the case where  $L(\theta; \mathbf{x})$  is *differentiable* the MLE can be derived as a solution of the equations

$$\frac{\partial \log L(\theta; \mathbf{x})}{\partial \theta} \equiv \mathbf{s}(\theta; \mathbf{x}) = \mathbf{0}, \quad (13.31)$$

referred to as *the likelihood equations*.

In example 2,

$$\begin{aligned} \frac{\partial \log L(\theta; \mathbf{x})}{\partial \theta} &= -\frac{n}{\theta} + \frac{n}{2\theta^2} \sum_{i=1}^n X_i^2 = 0, \\ \Rightarrow \hat{\theta} &= \frac{1}{n} \sum_{i=1}^n X_i^2 \end{aligned}$$

is the MLE of  $\theta$ , since

$$\left. \frac{d^2 \log L(\theta; \mathbf{x})}{d\theta^2} \right|_{\theta=\hat{\theta}} = \left. \left( \frac{n}{2\theta^2} - \frac{1}{\theta^3} \sum_{i=1}^n X_i^2 \right) \right|_{\theta=\hat{\theta}} = -\frac{1}{2} - \frac{n}{\hat{\theta}^2} < 0$$

(for a maximum).

### *Example 3*

Let  $\mathbf{X} \equiv (X_1, \dots, X_n)'$  be a random sample from a Pareto distribution with  $f(x; \theta) = \theta x^{-\theta-1}$ ,  $1 \leq x \leq \infty$ ,  $\theta \in \mathbb{R}_+$ ,

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

and

$$\begin{aligned} \log L(\theta; \mathbf{x}) &= c + n \log \theta - (\theta + 1) \sum_{i=1}^n \log x_i, \\ \Rightarrow \frac{d \log L(\theta; \mathbf{x})}{d\theta} &= \frac{n}{\theta} - \sum_{i=1}^n \log x_i = 0 \\ \Rightarrow \hat{\theta} &= n \left( \sum_{i=1}^n \log x_i \right)^{-1} \end{aligned}$$

is the MLE of  $\theta$  since

$$\left. \frac{d^2 \log L(\theta; \mathbf{x})}{d\theta^2} \right|_{\theta=\hat{\theta}} = -\frac{n}{\hat{\theta}^2} < 0.$$

Before the reader jumps to the erroneous conclusion that deriving the MLE is a matter of a simple differentiation let us consider some examples where the derivation is not as straightforward.

*Example 4*

Let  $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_n)$  where  $\mathbf{Z}_i = (X_i, Y_i)$  be a random sample from

$$N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right],$$

i.e.

$$f(x, y; \rho) = \frac{(1-\rho^2)^{-\frac{1}{2}}}{2\pi} \exp\left\{-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right\},$$

$$\log L(\rho; \mathbf{x}, \mathbf{y}) = c - n \log 2\pi - \frac{n}{2} \log(1-\rho^2)$$

$$- \frac{1}{2(1-\rho^2)} \sum_{i=1}^n (x_i^2 - 2\rho x_i y_i + y_i^2),$$

$$\begin{aligned} \frac{d \log L}{d \rho} &= +\frac{n(-2)}{2(1-\rho^2)} \rho - \rho \frac{\sum_{i=1}^n (x_i^2 + y_i^2)}{(1-\rho^2)^2} + \frac{(1+\rho^2)}{(1-\rho^2)^2} \sum_{i=1}^n x_i y_i = 0 \\ &= n\hat{\rho}(1-\hat{\rho}^2) + (1+\hat{\rho}^2) \sum_{i=1}^n x_i y_i - \hat{\rho} \left( \sum_{i=1}^n x_i^2 + \sum_{i=1}^n y_i^2 \right) = 0. \end{aligned}$$

This is a cubic equation in  $\rho$  and hence there are three possible values for the MLE and additional search is needed to locate the maximum value using numerical methods. The use of numerical methods in deriving MLE's was a major drawback for the method when it was first suggested by Fisher. Nowadays, however, this presents no difficulties. For a discussion of several numerical methods as used in econometrics see Harvey (1981).

*Example 5*

Let  $\mathbf{X} \equiv (X_1, X_2, \dots, X_n)'$  be a random sample from  $f(x; \theta) = 1/\theta$  where  $0 \leq x \leq \theta$ . The likelihood function is

$$L(\theta; \mathbf{x}) = \theta^{-n} \quad \text{if } 0 \leq x_i \leq \theta, \quad i = 1, 2, \dots, n.$$

Using  $[dL(\theta; \mathbf{x})]/d\theta = 0$  to derive the MLE is out of the question since  $L(\theta; \mathbf{x})$  is not continuous at the maximum (see Fig. 13.5). A moment's reflection suggests that the MLE of  $\theta$  is  $\hat{\theta} = \max(X_1, X_2, \dots, X_n)$ .

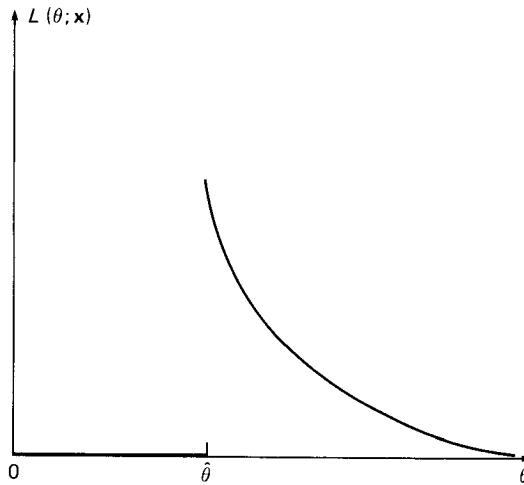


Fig. 13.5. The likelihood function of example 5.

*Example 6*

Let  $\mathbf{X} = (X_1, X_2, \dots, X_n)'$  be a random sample from  $f(x; \theta) = e^{-(x-\theta)}$  where  $x \geq \theta$ . The likelihood function is

$$L(\theta; \mathbf{x}) = \exp \left\{ - \sum_{i=1}^n (x_i - \theta) \right\} \quad \text{if } x_i \geq \theta, \quad i = 1, 2, \dots, n.$$

Again  $[dL(\theta; \mathbf{x})]/d\theta = 0$  is inappropriate for deriving the MLE. Common sense suggests  $L(\theta; \mathbf{x})$  is maximised by choosing  $\theta$  as large as possible such that  $L(\theta; \mathbf{x}) > 0$ . Since  $\theta$  is bounded below by the  $x_i$ s,  $\hat{\theta} = \min(X_1, X_2, \dots, X_n)$  represents the MLE of  $\theta$ .

Looking at examples 5 and 6 we can see that the problem of the derivation of the MLE arose because *the range of the  $X_i$ s depended on the unknown parameter  $\theta$* . It turns out that in such cases there are not only problems with deriving the MLE but also the estimators derived do not in general satisfy all the properties MLE's enjoy (see below). For example,  $\hat{\theta} = \max(X_1, \dots, X_n)$  is not asymptotically normal. Such cases are excluded by the assumption CR1 of Chapter 12.

So far the examples considered refer to the case where  $\theta$  is a scalar. In econometrics, however,  $\theta$  is commonly a  $k \times 1$  vector, a case which presents certain additional difficulties. For differentiable likelihood functions the MLE of  $\theta \equiv (\theta_1, \theta_2, \dots, \theta_k)'$  is derived by solving the system of equations

$$\frac{\partial \log L}{\partial \theta} = \mathbf{0} \tag{13.32}$$

or

$$\frac{\partial \log L}{\partial \theta_i} = 0, \quad i = 1, 2, \dots, k,$$

ensuring that the Hessian matrix

$$\mathbf{H}(\hat{\theta}) = \left. \frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j} \right|_{\theta=\hat{\theta}} = \left( \left. \frac{\partial \log L}{\partial \theta_i \partial \theta_j} \right|_{\theta=\hat{\theta}} \right)_{i,j}, \quad i, j \leq 1, 2, \dots, k \quad (13.33)$$

is negative definite, i.e. for any  $\mathbf{z} \in \mathbb{R}^k$ ,  $\mathbf{z}' \mathbf{H}(\hat{\theta}) \mathbf{z} < 0$ ,  $\mathbf{z} \neq \mathbf{0}$ .

### Example 7

Let  $\mathbf{X} \equiv (X_1, X_2, \dots, X_n)'$  be a random sample from  $N(\mu, \sigma^2)$ , i.e.

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma \sqrt{(2\pi)}} \exp \left\{ -\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2 \right\},$$

$$x \in \mathbb{R}, \quad \theta \equiv (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+.$$

The likelihood function takes the form

$$\begin{aligned} L(\theta; \mathbf{x}) &= k(\mathbf{x}) \prod_{i=1}^n \left( \frac{1}{\sigma \sqrt{(2\pi)}} \exp \left\{ -\frac{1}{2} \left( \frac{x_i-\mu}{\sigma} \right)^2 \right\} \right) \\ &= k(\mathbf{x}) \left( \frac{1}{\sigma \sqrt{(2\pi)}} \right)^n \left\{ \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right) \right\} \\ L(\theta; \mathbf{x}) &= k(\mathbf{x}) (\sigma^2 2\pi)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\} \end{aligned}$$

and

$$\log L = c - \frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2,$$

$$\frac{\partial \log L}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0$$

$$\Rightarrow \text{the MLE is } \hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n,$$

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 = 0$$

$$\Rightarrow \text{the MLE is } \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Since

$$\frac{\partial^2 \log L}{\partial \mu^2} \Big|_{\theta=\hat{\theta}} = -\frac{n}{\hat{\sigma}^2} < 0,$$

$$\frac{\partial^2 \log L}{\partial \sigma^4} \Big|_{\theta=\hat{\theta}} = \frac{n}{2\hat{\sigma}^4} - \frac{1}{\hat{\sigma}^6} \sum_{i=1}^n (x_i - \bar{X}_n)^2 = \frac{-n}{2\hat{\sigma}^4} < 0$$

and

$$\frac{\partial^2 \log L}{\partial \mu \partial \sigma^2} \Big|_{\theta=\hat{\theta}} = -\frac{1}{\hat{\sigma}^4} \sum_{i=1}^n (x_i - \bar{X}_n) = 0,$$

we can deduce that

$$\mathbf{H}(\hat{\theta}) = \begin{pmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2\hat{\sigma}^4} \end{pmatrix} \quad \text{and} \quad \mathbf{z}' \mathbf{H}(\hat{\theta}) \mathbf{z} < 0 \quad \text{for any } \mathbf{z} \in \mathbb{R}^2.$$

This example will be of considerable interest in Part IV where most estimation problems will be a direct extension of this case.

Despite the intuitive appeal of the method of maximum likelihood its ultimate justification as a general estimation method must be based on the optimum properties of the resulting MLE's.

### (3)      *Finite sample properties*

Let us discuss the finite sample properties of MLE's in the context of the simple statistical model:

- (i)      probability model,  $\Phi = \{f(x; \theta), \theta \in \Theta\}$ ;
- (ii)     sampling model,  $\mathbf{X} \equiv (X_1, \dots, X_n)'$ , is a random sample from  $f(x; \theta)$ .

One of the most attractive properties of MLE's is invariance.

#### *Invariance*

Let  $\hat{\theta}$  be a MLE of  $\theta$ . If  $g(\cdot): \Theta \rightarrow \mathbb{R}$  is a Borel function of  $\theta$  then a MLE of  $g(\theta)$  exists and is given by  $g(\hat{\theta})$ . This means that if the MLE of  $\theta$  is available then for functions such as  $\theta^k, e^\theta, \log \theta$ , its MLE is derived by substituting  $\hat{\theta}$  in place of  $\theta$ , i.e.  $\hat{\theta}^k, e^{\hat{\theta}}, \log \hat{\theta}$  are the MLE's of these functions. In the case of example 2 above the MLE of  $\theta$  was

$$\hat{\theta} = n \left( \sum_{i=1}^n \log X_i \right)^{-1}.$$

The invariance property of MLE's enables us to deduce that the MLE of

$$\phi = \frac{1}{\theta} \quad \text{is} \quad \hat{\phi} = \frac{1}{n} \sum_{i=1}^n \log X_i.$$

In relation to invariance it is important to note that in general

$$E(g(\hat{\theta})) \neq g(E(\hat{\theta})). \quad (13.34)$$

For example, if  $g(\theta) = \theta^2$  it is well known that  $E(\hat{\theta}^2) \neq (E(\hat{\theta}))^2$  in general. This contributes to the fact that the MLE's are not in general unbiased estimators. For instance, in example 7 above the MLE of  $\sigma^2$ ,  $\hat{\sigma}^2 = (1/n) \sum_{i=1}^n (X_i - \bar{X})^2$  is a biased estimator since  $(n\hat{\sigma}^2)/\sigma^2 \sim \chi^2(n-1)$  (see Section 11.5) and hence  $E(\hat{\sigma}^2) = [(n-1)/n]\sigma^2 \neq \sigma^2$ . Thus, in general, unbiased and MLE's do not coincide. In one particular case, when unbiasedness is accompanied by full efficiency, however, the two coincide.

#### *Unbiasedness, full-efficiency*

In the case where  $\Phi$  satisfies the regularity conditions CR1–CR3 and  $\hat{\theta}$  is an unbiased estimator of  $\theta$  whose variance achieves the Cramer–Rao lower bound, then the likelihood equation has a unique solution equal to  $\hat{\theta}$ . This suggests that any unbiased fully efficient estimator  $\hat{\theta}$  can be derived as a solution of the likelihood equation (a comforting thought!). In example 7 above the MLE of  $\mu$  was  $\hat{\mu}_n = \bar{X}_n$  which implies that  $\hat{\mu}_n \sim N(\mu, \sigma^2/n)$  since  $\hat{\mu}_n$  is a linear function of independent r.v.'s. Hence,  $E(\hat{\mu}_n) = \mu$  and  $\hat{\mu}_n$  is an unbiased estimator. Moreover, given that

$$\mathbf{I}_n(\theta) \equiv E \left[ \frac{-\partial^2 \log L(\theta; \mathbf{x})}{\partial \theta \partial \theta'} \right] = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$$

$$\text{and } [\mathbf{I}_n(\theta)]^{-1} = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix},$$

we can see that  $\text{Var}(\hat{\mu}_n)$  achieves the Cramer–Rao lower bound. On the other hand, the MLE of  $\sigma^2$ ,  $\hat{\sigma}_n^2 = (1/n) \sum_{i=1}^n (X_i - \bar{X}_n)^2$  as discussed above, is not an unbiased estimator.

The property mostly emphasised by Fisher in support of the method of maximum likelihood was the property of sufficiency.

**Sufficiency**

If  $\tau(\mathbf{X})$  is a sufficient statistic for  $\theta$  and a unique MLE  $\hat{\theta}$  of  $\theta$  exists then  $\hat{\theta}$  is a function of  $\tau(\mathbf{X})$ . In the case of a non-unique MLE, a MLE  $\hat{\theta}$  can be found which is a function of  $\tau(\mathbf{X})$ . It is important to note that this does not say that any MLE  $\hat{\theta}$  is a function of  $\tau(\mathbf{X})$ ; in the case of non-uniqueness some MLE's are not functions of  $\tau(\mathbf{X})$ . It was shown in Chapter 12 that  $\tau(\mathbf{X}) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$  are jointly minimal sufficient statistics for  $\theta \equiv (\mu, \sigma^2)$  in the case where  $\mathbf{X} \equiv (X_1, \dots, X_n)'$  is a random sample from  $N(\mu, \sigma^2)$ . In example 7 above the MLE's of  $\mu$  and  $\sigma^2$  were

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n X_i, \quad \hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

which are clearly functions of  $\tau(\mathbf{X})$ .

An important implication for ML estimation when a sufficient statistic exists is that the asymptotic covariance of  $\hat{\theta}_n$  (see below) can be consistently estimated by the Hessian evaluated at  $\theta = \hat{\theta}_n$ . That is,

$$\frac{1}{n} \frac{\partial^2 \log L(\hat{\theta}_n)}{\partial \theta \partial \theta'} \xrightarrow{P} \mathbf{I}_{\infty}(\theta) \quad (13.35)$$

(see Dhrymes (1970)).

**(4) Asymptotic properties (IID case)**

Although MLE's enjoy several optimum finite sample properties, as seen above, their asymptotic properties provide the main justification for the almost universal appeal of the method of maximum likelihood. As argued below, under certain regularity conditions, MLE's can be shown to be consistent, asymptotically normal and asymptotically efficient.

Let us begin the discussion of asymptotic properties enjoyed by MLE's by considering the simplest possible case where the statistical model is as follows:

- (i) probability model,  $\Phi = \{f(x; \theta), \theta \in \Theta\}$ ,  $\theta$  being a scalar;
- (ii) sampling model,  $\mathbf{X} \equiv (X_1, \dots, X_n)'$  is a *random sample* from  $f(x; \theta)$ .

Although this case is of little interest in Part IV, a brief discussion of it will help us understand the *non-random* sample case considered in the sequel.

The regularity conditions needed to prove the above-mentioned asymptotic properties for MLE's can take various forms (see Cramer (1946), Wald (1949), Norden (1972–73), Weiss and Wolfowitz (1974), Serfling (1980), *inter alia*). For our purposes it suffices to supplement the regularity conditions of Chapter 12, CR1–CR3, with the following condition:

(CR4) For every  $\theta \in \Theta$ ,

$$\left| \frac{\partial^i \log f(x; \theta)}{\partial \theta^i} \right| \leq h_i(x), \quad i = 1, 2, 3,$$

where the functions  $h_1(x)$  and  $h_2(x)$  are integrable over  $(-\infty, \infty)$ , i.e.

$$\int_{-\infty}^{\infty} h_i(x) dx < \infty, \quad i = 1, 2,$$

and

$$\int_{-\infty}^{\infty} h_3(x) f(x; \theta) dx < K,$$

where  $K$  does not depend on  $\theta$ .

The conditions CR1–CR4 are only sufficient conditions for consistency and asymptotic normality. In order to get some idea about how restrictive these conditions are it is instructive to consider various examples which do not satisfy them.

The examples 4 and 5 considered above are excluded by the condition CR1 because the range of the random variables  $X_1, \dots, X_n$  depends on the parameter  $\theta$ . Moreover, in the case of example 6, if  $\mu = 0$  then

$$\frac{\partial^3 \log f}{\partial \sigma^6} = -\frac{1}{\sigma^6} + \frac{3x^2}{\sigma^8} \rightarrow \infty \quad \text{as } \sigma^2 \rightarrow 0,$$

i.e. the third derivative is not bounded in the open interval  $0 < \sigma^2 < \infty$ ; condition CR4 is not satisfied (see Norden (1973)).

Under the regularity conditions CR1–CR4 we can prove (see Serfling (1980)) that the likelihood equation  $[\partial \log L(\theta; \mathbf{x})] / \partial \theta = 0$  admits a sequence of solutions  $\{\hat{\theta}_n, n \geq 1\}$  such that:

(1) *Consistency:*

$$\hat{\theta}_n \xrightarrow{\text{a.s.}} \theta_0 \quad (\text{strong consistency});$$

which implies that

$$\hat{\theta}_n \xrightarrow{P} \theta_0 \quad (\text{weak consistency})$$

(see Chapter 10).

(2) *Asymptotic normality:*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{a} N(0, I(\theta)^{-1}).$$

(3)      *Asymptotic efficiency:*

$$I(\theta) = \lim_{n \rightarrow \infty} \left[ \frac{1}{n} I_n(\theta) \right],$$

i.e. the asymptotic variance of  $\hat{\theta}_n$  equals the limit of the Cramer–Rao lower bound.

Although a formal proof of these results is beyond the scope of this book it is important to consider an informal heuristic argument on how such results come about.

If we were to take a Taylor expansion of  $[\partial \log L(\hat{\theta}_n; \mathbf{x})]/\partial \theta$  at  $\hat{\theta} = \theta_0$  we would get

$$\frac{\partial \log L(\hat{\theta}_n)}{\partial \theta} = \frac{\partial \log L(\theta_0)}{\partial \theta} + \frac{\partial^2 \log L(\theta_0)}{\partial \theta^2} (\hat{\theta}_n - \theta_0) + O_p(n), \quad (13.36)$$

where  $O_p(n)$  refers to all terms of order  $n$  (see Chapter 10). The above expansion is based on CR2 and CR4. In view of the fact the  $\log L(\theta) = \sum_{i=1}^n \log f(x_i; \theta)$  and the  $f(x_i; \theta)$ ,  $i = 1, 2, \dots, n$ , can be interpreted as functions of IID (independent and identically distributed r.v.'s) we can express the above Taylor expansion in the form

$$-\frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \log f(x_i; \theta_0)}{\partial \theta^2} (\hat{\theta}_n - \theta_0) = \frac{1}{n} \sum_{i=1}^n \frac{\partial \log f(x_i; \theta_0)}{\partial \theta} + O_p(1). \quad (13.37)$$

Using the strong law of large numbers (SLLN) for IID r.v.'s (see Section 9.2)

$$A_n = \frac{1}{n} \sum_{i=1}^n \frac{\partial \log f(x_i; \theta_0)}{\partial \theta} \xrightarrow{\text{a.s.}} 0 = E\left(\frac{\partial \log f(x; \theta_0)}{\partial \theta}\right), \quad (13.38)$$

$$B_n = \left( -\frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \log f(x_i; \theta_0)}{\partial \theta^2} \right) \xrightarrow{\text{a.s.}} I(\theta) = E\left(-\frac{\partial^2 \log f(x; \theta_0)}{\partial \theta^2}\right). \quad (13.39)$$

These in turn imply that

$$(\hat{\theta}_n - \theta_0) \xrightarrow{\text{a.s.}} 0 \quad \text{or} \quad \hat{\theta}_n \xrightarrow{\text{a.s.}} \theta_0. \quad (13.40)$$

To show asymptotic normality we divide all the terms of the Taylor expansion by  $1/\sqrt{n}$  (not  $1/n$ ) to get

$$B_n \sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial \log f(x_i; \theta_0)}{\partial \theta} + O_p(\sqrt{n}). \quad (13.41)$$

Using the central limit theorem (CLT) for IID r.v.'s (see Section 9.3) we can show that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial \log f(x_i; \theta_0)}{\partial \theta} \xrightarrow{x} N(0, I(\theta)). \quad (13.42)$$

Given that  $B_n \xrightarrow{\text{a.s.}} I(\theta)$  we can deduce that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{x} N(0, I(\theta)^{-1}). \quad (13.43)$$

As far as asymptotic efficiency is concerned we can see that the asymptotic variance of  $\hat{\theta}_n$ , a consistent and asymptotically normal (CAN) estimator, is equal to  $I(\theta)^{-1}$  where

$$I(\theta) = E\left(\frac{\partial \log f(x; \theta_0)}{\partial \theta}\right)^2 = E\left(-\frac{\partial^2 \log f(x; \theta_0)}{\partial \theta^2}\right), \quad (13.44)$$

the information for a single observation. We know, however, that in the IID case the sample information matrix  $I_n(\theta) = E[\partial \log L(\theta_0)/\partial \theta]^2$  equals  $n$  times  $I(\theta)$ , given that each observation contributes equally to the sample, i.e.  $nI(\theta) = I_n(\theta)$ . This implies that the limit of the Cramer–Rao lower bound is  $I(\theta)$  because

$$\lim_{n \rightarrow \infty} \left[ \frac{1}{n} I_n(\theta) \right] = \frac{1}{n} I_n(\theta) = I(\theta). \quad (13.45)$$

It must be stressed that this is only true for the IID case. In more general cases care should be exercised in assessing the order of magnitude of  $I_n(\theta)$  (see Chapter 19).

Returning to the example 2 above we can see that

$$I_n(\sigma_0^2) = \frac{n}{2\sigma_0^4} \quad \text{and} \quad I(\sigma_0^2) = \frac{1}{2\sigma_0^4}$$

which implies that

$$\sqrt{n}(\hat{\sigma}_n^2 - \sigma_0^2) \xrightarrow{x} N(0, 2\sigma_0^4).$$

In example 3,  $I_n(\theta_0) = n/\theta_0^2$  and  $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{x} N(0, \theta_0^2)$ . In example 4, although  $\hat{\rho}_n$  cannot be derived explicitly, this does not stop us from deriving its asymptotic distribution. This takes the form

$$\sqrt{n}(\hat{\rho}_n - \rho) \xrightarrow{x} N\left(0, \frac{(1-\rho^2)^2}{1+\rho^2}\right).$$

The above asymptotic properties can be extended directly to the case where  $\boldsymbol{\theta} \equiv (\theta_1, \theta_2, \dots, \theta_k)'$ ,  $k \geq 1$ , to read:

$$(i) \quad \hat{\boldsymbol{\theta}}_n \xrightarrow{\text{a.s.}} \boldsymbol{\theta} \quad \text{and} \quad \hat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}$$

(the zero subscript dropped for notational convenience);

$$(ii) \quad \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \underset{x}{\sim} N(\mathbf{0}, \mathbf{I}(\boldsymbol{\theta})^{-1}), \quad (13.46)$$

where

$$\mathbf{I}(\boldsymbol{\theta}_0) = E\left(\left(\frac{\partial \log f(x; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \log f(x; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right)'\right) = E\left(-\frac{\partial^2 \log f(x; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right). \quad (13.47)$$

That is,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{in} - \theta_i) \underset{x}{\sim} N(0, [\mathbf{I}(\boldsymbol{\theta})]_{ii}^{-1}), \quad i = 1, 2, \dots, k, \quad (13.48)$$

$\mathbf{I}(\boldsymbol{\theta})_{ii}^{-1}$  being the  $i$ th diagonal element of  $\mathbf{I}(\boldsymbol{\theta})^{-1}$ .

Asymptotic efficiency in the multiparameter case is in terms of the asymptotic covariance of  $\hat{\boldsymbol{\theta}}_n$  which takes the form  $\text{Cov}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_n) = \mathbf{I}(\boldsymbol{\theta})^{-1}$ . For any other CAN estimator  $\tilde{\boldsymbol{\theta}}_n$ , the matrix difference

$$\text{Cov}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{\theta}}_n) - \mathbf{I}(\boldsymbol{\theta})^{-1} \geq 0 \quad (13.49)$$

is non-negative definite. In the case of example 7 above the asymptotic distribution of  $\hat{\boldsymbol{\theta}}_n \equiv (\hat{\mu}_n, \hat{\sigma}_n^2)'$  takes the form

$$\begin{pmatrix} \sqrt{n}(\hat{\mu}_n - \mu) \\ \sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) \end{pmatrix} \underset{x}{\sim} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}\right].$$

This shows that asymptotically  $\hat{\sigma}_n^2 = (1/n) \sum_{i=1}^n (X_i - \bar{X}_n)^2$  achieves the lower bound even though for a fixed  $n$  it does not (see Chapter 12).

#### (5)\*    Asymptotic properties (non-IID case)

The cases of particular interest in Part IV are the ones of an independent and a non-random sample, that is, (i)  $\mathbf{X}_n \equiv (X_1, \dots, X_n)'$  is a set of independent (ID) (but not identically distributed) r.v.'s, and (ii)  $\mathbf{X}_n$  is a set of non-IID r.v.'s. The asymptotic properties of MLE's for the IID case considered above can be extended to the cases of interest without much difficulty. The way this is achieved is to reduce the more complicated cases to something for which the limit theorems considered in Chapter 9 can be applied.

The extension of the above asymptotic results to the independent (ID)

case presents no particular difficulties apart from the fact that the result  $nI(\theta) = I_n(\theta)$  no longer holds. This in turn implies that the asymptotic variance of  $\hat{\theta}_n$  can no longer be  $I(\theta)$ . This, however, is not such a difficult problem because it can always be replaced by the assumption that  $I_n(\theta)$  is of order  $n$ . That is,

$$\lim_{n \rightarrow \infty} \left( \frac{1}{n} I_n(\theta) \right) < \infty, \quad (13.50)$$

denoted by  $I_n(\theta) = O(n)$  (see Chapter 10). Given that

$$I_n(\theta) = E \left( - \sum_{i=1}^n \frac{\partial^2 \log f(x_i; \theta)}{\partial \theta^2} \right), \quad (13.51)$$

we can see that the assumption  $I_n(\theta) = O(n)$  ensures that as  $n \rightarrow \infty$ ,  $I_n(\theta) \rightarrow \infty$  at the same speed, i.e. information continues to accrue as  $n$  increases.

The non-random sample case is more difficult to tackle but again all we need to do is to reduce it to a case where the various limit theorems can be applied. In our discussion of these limit theorems it was argued that independence is not needed for the various results and that martingale orthogonality suffices for most purposes. This suggests that if we were to reduce  $A_n$  and  $B_n$  above to martingale orthogonal processes then the results will follow.

The natural way to orthogonalise the non-random sample  $X_n$  is to start from  $X_1$ , define  $\tilde{X}_1 = X_1 - E(X_1)$ , and then construct

$$\tilde{X}_2 = X_2 - E(X_2 / \sigma(X_1)),$$

$$\tilde{X}_3 = X_3 - E(X_3 / \sigma(X_1, X_2)),$$

⋮

$$\tilde{X}_i = X_i - E(X_i / \sigma(X_1, X_2, \dots, X_{i-1})),$$

⋮

$$\tilde{X}_n = X_n - E(X_n / \sigma(X_1, \dots, X_{n-1})).$$

Let  $\mathcal{D}_{i-1} = \sigma(X_1, X_2, \dots, X_{i-1})$ ,  $i = 2, 3, \dots, n$ , denote the  $\sigma$ -field generated by the r.v.'s  $X_1, \dots, X_{i-1}$ . By construction the  $\tilde{X}_i$  includes only the new information in  $X_i$  and satisfies the following properties:

$$(i) \quad E(\tilde{X}_i / \mathcal{D}_{i-1}) = 0, \quad i = 2, 3, \dots, n; \quad (13.52)$$

$$(ii) \quad E(\tilde{X}_i \tilde{X}_j / \mathcal{D}_{i-1}) = 0, \quad j < i, \quad i, j = 2, 3, \dots, n. \quad (13.53)$$

That is, the  $\tilde{X}_i$ 's define a martingale difference (see Section 8.4). Hence by conditioning on past information at each  $i$  we can reduce the non-random

sample  $\mathbf{X}_n \equiv (X_1, \dots, X_n)'$  to a *martingale orthogonal sample*  $\tilde{\mathbf{X}}_n \equiv (\tilde{X}_1, \dots, \tilde{X}_n)$  which asymptotically can be treated in a similar way as a random sample. For this we need to impose certain time-homogeneity and memory restrictions on  $\{X_n, n \geq 1\}$  so as to ensure for example that

$$n^{-1} E(S_n) \rightarrow \sigma^2 \quad \text{as } n \rightarrow \infty \quad (13.54)$$

where  $S_n = \sum_{i=1}^n \tilde{X}_i$ . In such a case  $\{S_n, n \geq 1\}$  behaves asymptotically as the sum of  $n$  martingale differences each with variance  $\sigma^2$ . This in turn enables us to use the various limit theorems (see Chapter 9) to derive a number of asymptotic results needed. Heuristically, this enables us to treat the parameters  $\theta_i$  in the decomposition

$$D(x_1, \dots, x_n; \boldsymbol{\theta}) = \prod_{i=1}^n f(x_i/x_1, \dots, x_1; \boldsymbol{\theta}_i) \quad (13.55)$$

as being equal, i.e. deduce that

$$\boldsymbol{\theta}_n \rightarrow \boldsymbol{\theta} \quad \text{as } n \rightarrow \infty. \quad (13.56)$$

This in turn allows us to define the likelihood function to be

$$L_n(\boldsymbol{\theta}; \mathbf{x}) = k(\mathbf{x}) \prod_{i=1}^n f(x_i/x_1, \dots, x_{i-1}; \boldsymbol{\theta}), \quad \text{given } x_0. \quad (13.57)$$

For simplicity let us consider the case where  $\theta$  is scalar.

$$\frac{d}{d\theta} \log L_n(\theta) = \sum_{i=1}^n \frac{d}{d\theta} \log f(x_i/x_1, \dots, x_{i-1}; \theta) \quad (13.58)$$

is the random quantity we are particularly interested in. Observe that the terms in the summation can be written as

$$\frac{d}{d\theta} \log f(x_i/x_1, \dots, x_{i-1}; \theta) = \frac{d}{d\theta} [\log L_i(\theta) - \log L_{i-1}(\theta)] \equiv z_i(\theta), \quad (13.59)$$

which implies that

$$\frac{d}{d\theta} \log L_n(\theta) = \sum_{i=1}^n \frac{d}{d\theta} [\log L_i(\theta) - \log L_{i-1}(\theta)] \equiv \sum_{i=1}^n z_i(\theta) \quad (13.60)$$

and assuming the expected values exist,

$$\begin{aligned} & E\left(\frac{d}{d\theta} \log f(x_i/x_1, \dots, x_{i-1}; \theta)/\mathcal{Q}_{i-1}\right) \\ &= E\left(\frac{d}{d\theta} \log L_i(\theta)/\mathcal{Q}_{i-1}\right) - E\left(\frac{d}{d\theta} \log L_{i-1}(\theta)/\mathcal{Q}_{i-1}\right) = 0. \end{aligned} \quad (13.61)$$

These imply that  $E(z_i(\theta)/\mathcal{D}_{i-1})=0$ ,  $i=1, 2, \dots$ , and hence  $\{(d/d\theta) \log L_n(\theta), \mathcal{D}_n, n \geq 1\}$  is a zero mean martingale and the  $z_i(\theta)$ s are *martingale differences* (see Section 8.4). Defining the random variable (r.v.)

$$I_n(\theta) = \sum_{i=1}^n E(z_i^2(\theta)/\mathcal{D}_{i-1}), \quad (13.62)$$

we observe that it corresponds to the Fisher information matrix defined above. Moreover, under conditions similar to CR1–CR3 above,

$$E(z_i^2(\theta)/\mathcal{D}_{i-1}) = E\left(-\frac{dz_i(\theta)}{d\theta} \middle| \mathcal{D}_{i-1}\right), \quad i=1, 2, \dots, \quad (13.63)$$

and  $I_n(\theta)$  can be defined alternatively as

$$I_n(\theta) = \sum_{i=1}^n E\left(-\frac{dz_i(\theta)}{d\theta} \middle| \mathcal{D}_{i-1}\right). \quad (13.64)$$

Under certain regularity conditions relating the first three derivatives of  $\log L_n(\theta)$ ,  $n=1, 2, \dots$ , it can be shown that

$$[I_n(\theta)]^{-1} \sum_{i=1}^n z_i(\theta) \xrightarrow{P} 0, \quad (13.65)$$

provided  $I_n(\theta) \rightarrow \infty$  as  $n \rightarrow \infty$ . This enables us to deduce the following property for MLE's.

### Consistency

The likelihood equation  $\sum_{i=1}^n z_i(\hat{\theta})=0$  has a root  $\hat{\theta}_n$  which is consistent for  $\theta$ , i.e.

$$\lim_{n \rightarrow \infty} Pr(|\hat{\theta}_n - \theta| < \varepsilon) = 0. \quad (13.66)$$

Moreover, under slightly stronger restrictions

$$[I_n(\theta)]^{-1} \sum_{i=1}^n z_i(\theta) \xrightarrow{\text{a.s.}} 0, \quad (13.67)$$

provided  $I_n(\theta) \rightarrow \infty$  as  $n \rightarrow \infty$  and hence a MLE  $\hat{\theta}_n$  is also *strongly consistent*, i.e.  $\hat{\theta}_n \xrightarrow{\text{a.s.}} \theta$ . The most interesting of the sufficient conditions for

the consistency of MLE's is the condition that

$$I_n(\theta) = \sum_{i=1}^n E(z_i^2(\theta)/\mathcal{D}_{i-1}) \xrightarrow{P} \infty \quad (13.68)$$

(or a.s.). This can be interpreted intuitively as saying that *information about  $\theta$  increases with the sample size*.

As argued above consistency is only a minimal property. A more useful property is that of asymptotic normality. In order to get asymptotic normality we need to find a normalisation sequence  $\{c_n, n \geq 1\}$  such that convergence in probability  $(\hat{\theta}_n - \theta) \xrightarrow{P} 0$  can be transformed into convergence in distribution of the form

$$c_n(\hat{\theta}_n - \theta) \xrightarrow{D} Z \sim N(0, 1). \quad (13.69)$$

The natural choice for  $c_n$  in this case is  $c_n = [I_n(\theta)]^{1/2}$  which ensures that

$$\underset{x}{[I_n(\theta)]^{1/2}}(\hat{\theta}_n - \theta) \sim N(0, 1). \quad (13.70)$$

*In the IID case*

$$\sum_{i=1}^n z_i(\theta) = \sum_{i=1}^n \frac{d}{d\theta} \log f(x_i; \theta) \quad (13.71)$$

and

$$I_n(\theta) = \sum_{i=1}^n E(z_i^2(\theta)/\mathcal{D}_{i-1}) = \sum_{i=1}^n E\left(\frac{d \log f(x_i; \theta)}{d\theta}\right)^2. \quad (13.72)$$

If we denote the *information matrix for one observation* by  $I(\theta)$ , i.e.

$$I(\theta) = E\left(\frac{d \log f(x_i; \theta)}{d\theta}\right)^2. \quad (13.73)$$

and assume CR1–CR3 then, in the random sample case

$$I_n(\theta) = \sum_{i=1}^n I(\theta) = nI(\theta), \quad (13.74)$$

and hence the asymptotic normality takes the form

$$(nI(\theta))^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{x} N(0, 1) \quad (13.75)$$

or

$$\sqrt{n}(\hat{\theta}_n - \theta) \underset{x}{\sim} N(0, I(\theta)^{-1}), \quad (13.76)$$

since  $0 < I(\theta) < \infty$  from CR3 and independent of  $n$ . It is obvious that in this case the condition

$$I_n(\theta) = nI(\theta) \rightarrow \infty \quad \text{as } n \rightarrow \infty \quad (13.77)$$

is automatically satisfied.

In the general non-random sample case we would like an analogous asymptotic normality result

$$c_n(\hat{\theta}_n - \theta) \underset{x}{\sim} N(0, V(\theta)), \quad (13.78)$$

where  $c_n$  refers to the order of magnitude of  $[I_n(\theta)]^{\frac{1}{2}}$ . For example, in the IID case,

$$I_n(\theta) = nI(\theta) = O(n) \quad \text{and} \quad c_n = \sqrt{n},$$

i.e.  $\lim_{n \rightarrow \infty} [I_n(\theta)/n] < k < \infty$  since  $0 < I(\theta) < \infty$ . To illustrate this consider  $\hat{\theta}_n \equiv (\hat{\mu}_n, \hat{\sigma}_n^2)'$  in example 7 above.

$$\lim_{n \rightarrow \infty} \left( \frac{\mathbf{I}_n(\theta)}{n} \right) = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix} = I(\theta), \quad \text{i.e. } I_n(\theta) \text{ is of order } n.$$

In cases where the sampling model is either an independent or a non-random sample the order of magnitude of  $I_n(\theta)$  can be any power of  $n^{\frac{1}{2}}$ . For most cases of interest, however, it suffices to concentrate on cases where

$$\mathbf{I}_n(\theta) = O_p(n) \quad (13.79)$$

and define the asymptotic information matrix  $\mathbf{I}_{\infty}(\theta)$  to be

$$\frac{\mathbf{I}_n(\theta)}{n} \xrightarrow{P} \mathbf{I}_{\infty}(\theta), \quad (13.80)$$

with  $\mathbf{V}(\theta) = \mathbf{I}_{\infty}(\theta)^{-1}$ . The notation above is used to emphasise the fact that  $\mathbf{I}_n(\theta)$  might be stochastic given that the conditional expectation is relative to a  $\sigma$ -field (see Section 7.2).

### *Asymptotic normality*

Under certain regularity conditions any consistent solution of the likelihood equation  $\hat{\theta}_n$  is asymptotically normal, i.e.

$$(I_n(\theta))^{\frac{1}{2}}(\hat{\theta}_n - \theta) \underset{x}{\sim} N(0, 1). \quad (13.81)$$

The case of particular interest in econometrics is when  $\theta$  is a  $k \times 1$  vector of unknown parameters. In such a case we need to normalise each individual parameter separately in general. With this in mind let us define the following normalising matrix:

$$\mathbf{D}_n(\theta) = \text{diag}\left(\left(\frac{\partial \log L}{\partial \theta_1}\right)^{\frac{1}{2}}, \dots, \left(\frac{\partial \log L}{\partial \theta_k}\right)^{\frac{1}{2}}\right),$$

where  $\left(\frac{\partial \log L}{\partial \theta_i}\right) \xrightarrow{P} \infty \quad \text{as } n \rightarrow \infty.$  (13.82)

Under certain regularity conditions we can show that

$$\mathbf{D}_n(\theta)(\hat{\theta}_n - \theta) \xrightarrow{P} N(\mathbf{0}, \mathbf{C}(\theta)^{-1}), \quad (13.83)$$

where

$$\mathbf{D}_n^{-1}(\theta) \mathbf{A}_n(\hat{\theta}) \mathbf{D}_n^{-1}(\theta) \xrightarrow{P} \mathbf{C}(\theta)$$

and

$$\mathbf{A}_n(\theta) = \left( \frac{-\partial^2 \log L}{\partial \theta_i \partial \theta_j} \right), \quad i, j = 1, \dots, k.$$

#### *Asymptotic efficiency*

As shown above, the asymptotic distribution of the CAN estimator  $\hat{\theta}_n$  is  $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{P} N(\mathbf{0}, V(\theta))$  where  $V(\theta)$  achieves the asymptotic Cramer–Rao lower bound  $I_{\infty}(\theta)^{-1}$ , that is,

$$V(\theta) = I_{\infty}(\theta)^{-1}. \quad (13.84)$$

This last equality defines the asymptotic efficiency of the MLE  $\hat{\theta}_n$ .

#### **(6) Summary of asymptotic properties**

For reference purposes let us summarise the asymptotic properties of  $\hat{\theta}_n$ , a MLE of  $\theta_0$ , in the case where  $\mathbf{I}_n(\theta) = O_p(n)$ :

(1) *Consistency.* For some root  $\hat{\theta}_n$  of the likelihood equation

$$(i) \quad \hat{\theta}_n \xrightarrow{\text{a.s.}} \theta_0 \quad \left( \Pr\left(\lim_{n \rightarrow \infty} \hat{\theta}_n = \theta_0\right) = 1 \right);$$

$$(ii) \quad \hat{\theta}_n \xrightarrow{P} \theta_0 \quad \left( \lim_{n \rightarrow \infty} \Pr(|\hat{\theta}_n - \theta_0| < \varepsilon) = 1 \right).$$

- (2) *Asymptotic normality.* For a consistent MLE  $\hat{\theta}_n$  of  $\theta_0$

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \mathbf{V}(\theta)).$$

- (3) *Asymptotic efficiency.* For a CAN estimator  $\hat{\theta}_n$  of  $\theta_0$

$$\mathbf{V}(\theta) = \mathbf{I}_{\infty}(\theta)^{-1}.$$

Note that asymptotic normality also implies *asymptotic unbiasedness*, i.e.  
 $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{P} 0$ .

### Example 8

Let  $\mathbf{X}_n \equiv (X_1, X_2, \dots, X_n)'$  be a non-random sample generated by the AR(1) time-series model:

$$X_i = \alpha X_{i-1} + U_i, \quad |\alpha| < 1,$$

where  $U_i \sim N(0, \sigma^2)$ ; a normal white-noise process (see Chapter 8). The distribution of the sample  $D(\mathbf{X}_n; \theta)$  is multivariate normal of the form:

$$\mathbf{X}_n \sim N(\mathbf{0}, \sigma^2 \mathbf{V}_n(\alpha)),$$

where

$$\mathbf{V}_n(\alpha) = [v_{ij}]_{i,j}, \quad v_{ij} = \frac{\alpha^{|i-j|}}{(1-\alpha^2)}, \quad i, j = 1, 2, \dots, n.$$

This is because

$$E(X_i) = 0,$$

$$\begin{aligned} E(X_i X_{i-k}) &= E((\alpha X_{i-1} + U_i) X_{i-k}) \\ &= \frac{\alpha^k \sigma^2}{(1-\alpha^2)} = E(X_i X_{i+k}), \quad k = 0, 1, 2, \dots \end{aligned}$$

in view of the restriction  $|\alpha| < 1$  which ensures stationarity (see Chapter 8). As argued above the only decomposition possible in the case of a non-random sample is

$$D(x_n; \theta) = f(x_0; \theta) \prod_{i=1}^n f(x_i/x_1, \dots, x_{i-1}; \theta)$$

where  $f(x_0; \theta)$  refers to the marginal distribution of the initial conditions. In practice there are several ways the initial conditions can be treated:

- (i) assume  $X_0$  is a known constant, say  $X_0 = 0$  (i.e. a degenerate r.v.);

- (ii) assume that  $X_0 \sim N(0, \sigma^2/(1-\alpha)^2)$ ; this ensures stationarity of  $\{X_n, n \geq 1\}$ ;
- (iii) assume that  $X_0 = X_n$  which defines  $\{X_n, n \geq 1\}$  as a circular stochastic process (see Anderson (1971)).

For expositional purposes let us adopt (i) to define the likelihood function as

$$\begin{aligned} L(\boldsymbol{\theta}; \mathbf{X}_n) &= k(\mathbf{X}_n) \prod_{i=1}^n \left( \frac{1}{\sigma \sqrt{(2\pi)}} \exp \left\{ -\frac{1}{2\sigma^2} (X_i - \alpha X_{i-1})^2 \right\} \right), \\ \log L(\boldsymbol{\theta}; \mathbf{X}_n) &= \text{const} - \left( \frac{n}{2} \right) \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \alpha X_{i-1})^2 \\ \frac{\partial \log L}{\partial \alpha} &= \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \alpha X_{i-1}) X_{i-1} = 0, \\ \frac{\partial \log L}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (X_i - \alpha X_{i-1})^2 = 0. \end{aligned}$$

Hence, the MLE's of  $\alpha$  and  $\sigma^2$  are

$$\begin{aligned} \hat{\alpha}_n &= \frac{\sum_{i=1}^n X_i X_{i-1}}{\sum_{i=1}^n X_i^2}, \quad \hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\alpha} X_{i-1})^2, \\ \frac{\partial^2 \log L}{\partial \alpha^2} &= \frac{1}{\sigma^2} \left( - \sum_{i=1}^n X_{i-1}^2 \right), \\ \frac{\partial^2 \log L}{\partial \alpha \partial \sigma^2} &= -\frac{1}{\sigma^4} \left( \sum_{i=1}^n X_i X_{i-1} - \alpha \sum_{i=1}^n X_{i-1}^2 \right), \\ \frac{\partial^2 \log L}{\partial \sigma^4} &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^n (X_i - \alpha X_{i-1})^2. \end{aligned}$$

Taking expectations with respect to  $D(\mathbf{X}_n; \boldsymbol{\theta})$  it follows that

$$\begin{aligned} E\left( -\frac{\partial^2 \log L}{\partial \alpha^2} \right) &= \frac{1}{\sigma^2} \sum_{i=1}^n E(X_{i-1}^2) = \frac{1}{\sigma^2} \sum_{i=2}^n \left( \frac{\sigma^2}{1-\alpha^2} \right) = \left( \frac{n-1}{1-\alpha^2} \right), \\ E\left( -\frac{\partial^2 \log L}{\partial \alpha \partial \sigma^2} \right) &= \frac{1}{\sigma^2} \left( \frac{(n-1)\alpha}{(1-\alpha^2)} - \frac{(n-1)\alpha}{(1-\alpha^2)} \right) = 0; \end{aligned}$$

note the role of the initial conditions.

$$E - \frac{\hat{c}^2 \log L}{\hat{c}\sigma^4} = -\frac{n}{2\sigma^4} + \frac{n\sigma^2}{\sigma^6} = \frac{n}{2\sigma^4}; \quad I_n(\theta) = \begin{pmatrix} \left(\frac{n-1}{1-\alpha^2}\right) & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}.$$

In order to establish consistency of  $\hat{\alpha}_n$  we note that the score function:

$$S_n = \sum_{i=1}^n (X_i - \alpha X_{i-1}) X_{i-1} = 0$$

defines a zero-mean *martingale*  $\{S_n, n \geq 1\}$ . Using the WLLN for martingales and Markov processes it follows that

$$\frac{S_n}{n} \xrightarrow{P} 0 \quad \text{and} \quad \left( \frac{1}{n} \sum_{i=1}^n X_{i-1}^2 \right) \xrightarrow{P} \left( \frac{\sigma^2}{1-\alpha^2} \right),$$

respectively. Hence

$$\hat{\alpha}_n \xrightarrow{P} \alpha.$$

Using a similar argument we can show that  $\hat{\sigma}_n^2 \xrightarrow{P} \sigma^2$ . For asymptotic normality we need to express  $\hat{\alpha}_n$  in the form

$$\sqrt{n}(\hat{\alpha}_n - \alpha) = \left( \frac{1}{n} \sum_{i=1}^n X_{i-1}^2 \right)^{-1} \left( \frac{1}{\sqrt{n}} S_n \right).$$

Using the CLT for martingales (see Section 9.3) it follows that

$$\frac{1}{\sqrt{n}} S_n \underset{\alpha}{\sim} N\left(0, \frac{\sigma^4}{1-\alpha^2}\right)$$

deducing that  $\sqrt{n}(\hat{\alpha}_n - \alpha) \underset{\alpha}{\sim} N(0, (1-\alpha^2))$  and  $\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) \underset{\alpha}{\sim} N(0, 2\sigma^4)$

(see Anderson (1971)). It is interesting to note that in the case where  $|\alpha| > 1$  the order of magnitude of  $I_n(\alpha)$  is no longer  $n$ , in which case we need a different normalising sequence for asymptotic normality. In particular for the sequence  $\{c_n, n \geq 1\}$  where  $c_n = (\sum_{i=1}^n X_{i-1}^2)^{\frac{1}{2}}$  it follows that

$$c_n(\hat{\alpha}_n - \alpha) \underset{\alpha}{\sim} N(0, \sigma^2)$$

(see Anderson (1959)). It is also important to stress that the asymptotic distribution of  $\hat{\alpha}_n$  depends crucially on the distribution of the white-noise error process  $\{U_n, n \geq 1\}$ . In the case where  $|\alpha| > 1$  if  $U_n$  is not normally distributed the above asymptotic normality result does not follow.

***Important concepts***

Least-squares, normal equations, method of moments, the likelihood function, the log-likelihood function, the score function, maximum likelihood estimator, likelihood equations, invariance, sample information matrix, single observation information matrix, asymptotic information matrix.

***Questions***

1. Explain the least-squares estimation method.
2. Explain the logic underlying the method of moments.
3. Why is it that the method of moments usually leads to inefficient estimators?
4. Define the likelihood function and explain its relationship with the distribution of the sample.
5. Discuss the relationship between the likelihood, log likelihood and score functions.
6. Define the concept of a MLE and explain the common-sense logic underlying the definition.
7. State and explain the small sample properties of MLE's.
8. Explain how a non-random sample  $\mathbf{X}$  can be transformed into a martingale orthogonal sample  $\mathbf{X}$ .
9. Explain why  $\{(d/d\theta) \log L_n(\theta), \mathcal{D}_n, n \geq 1\}$  defines a zero mean martingale and the  $z_i(\theta), i = 1, 2, \dots, n$ , define a martingale difference.
10. State and explain the asymptotic properties of MLE's.
11. Discuss the relationship between the order of magnitude and asymptotic normality of MLE's.
12. If we were to interpret the likelihood function as a density function what does the MLE correspond to?

***Exercises***

1. Consider the linear model

$$y_i = \theta_1 x_{1i} + \varepsilon_i, \quad \varepsilon_i \sim NI(0, \sigma^2), \quad i = 1, \dots, n.$$

$\theta = (\theta_1, \sigma^2)$ . Define the MLE's of  $\theta$  and compare its properties with those of the least-squares estimators.

2. Show that

$$\text{Var}(\tilde{\theta}_1) = \frac{\sigma^2}{\sum_{i=1}^n x_{1i}^2},$$

where  $\tilde{\theta}_1$  is the least-squares estimator of exercise 1, and compare it with the Cramer–Rao lower bound.

3. Let  $\mathbf{X} \equiv (X_1, X_2, \dots, X_n)'$  be a random sample from the Poisson distribution with density function

$$f(x; \theta) = \frac{\theta^x e^{-\theta}}{x!}.$$

- (i) Derive the likelihood, the log-likelihood and the score functions.  
(ii) Derive the MLE of  $\theta$  and its asymptotic distribution.

4. Let  $\mathbf{X} \equiv (X_1, \dots, X_n)'$  be a random sample from the Bernoulli distribution with a density function

$$f(x; \theta) = \theta^x (1 - \theta)^{1-x}, \quad x = 0, 1.$$

Derive the MLE and state its properties.

5. Let  $\mathbf{X} \equiv (X_1, \dots, X_n)'$  be a random sample from  $N(\mu, 1)$ . Show that

$$E\left(\frac{d \log L(\mu; \mathbf{x})}{d\mu}\right)^2 = -E\left(\frac{d^2 \log L(\mu; \mathbf{x})}{d\mu^2}\right).$$

6. Let  $\mathbf{X} \equiv (X_1, \dots, X_n)'$  be a random sample from the log normal distribution with density function

$$f(x; m, \sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}\left(\log\left(\frac{x}{m}\right)\right)^2\right\}, \quad x > 0.$$

- (i) Derive the MLE's of  $m$  and  $\sigma^2$ . (Hint: use invariance.)  
(ii) Derive the method of moments estimators for  $m$  and  $\sigma^2$ .

7. Let  $\mathbf{X} \equiv (X_1, \dots, X_n)'$  be a random sample from the exponential distribution with density function

$$f(x; \theta) = \theta e^{-\theta x}, \quad x \geq 0.$$

Derive the MLE of  $\theta$  and show that it is both consistent and asymptotically normal. What is the MLE of  $1/\theta$ ?

8. Let  $\mathbf{X} \equiv (X_1, X_2, \dots, X_n)'$  be an *independent sample* from  $N(\mu_i, \sigma_i^2)$ .
- (i) For  $\sigma_i^2 = \sigma^2$ ,  $i = 1, 2, \dots, n$ , derive the MLE's of  $(\sigma^2, \mu_1, \dots, \mu_n)$  and their asymptotic distribution. (Hint: check  $I_\infty(\boldsymbol{\theta})$ .)
  - (ii) For  $\mu_i = \mu$ ,  $i = 1, 2, \dots, n$ , derive the MLE's of  $(\mu, \sigma_1^2, \dots, \sigma_n^2)$  and their asymptotic distribution.
9. Let  $\mathbf{X} \equiv (X_1, \dots, X_n)'$  be a random sample from the Weibull distribution with a density function

$$f(x; \theta) = \theta c x^{c-1} \exp\{-\theta x^c\}, \quad x \geq 0,$$

where  $c$  is a known constant. Derive the MLE of  $\theta$  and its asymptotic distributions.

- 10\*. Let  $\mathbf{X} \equiv (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n)'$  where

$$\mathbf{X}_i \equiv \begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix}$$

be a random sample from the bivariate normal distribution

$$\begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right), \quad i = 1, 2, \dots, n.$$

- (i) Derive the MLE  $\hat{\theta}$  of  $\theta \equiv (\sigma_1^2, \sigma_2^2, \rho)$  and its asymptotic distribution.
- (ii) Assuming that  $\sigma_1^2 = \sigma_2^2 = 1$  derive the MLE  $\hat{\rho}$  of  $\rho$  and its asymptotic distribution.
- (iii) Compare the asymptotic variance of  $\hat{\rho}$  and  $\tilde{\rho}$  and explain intuitively why they differ.
- (iv) For  $\rho=0$  and  $\rho=1$  derive the MLE's of  $\sigma_1^2$  and  $\sigma_2^2$  and compare their asymptotic variances.

#### **Additional references**

Bickel and Doksum (1977); Cox and Hinkley (1974); Kendall and Stuart (1973); Lloyd (1984); Rao (1973); Rohatgi (1976); Silvey (1975); Zacks (1971).

## CHAPTER 14

---

### Hypothesis testing and confidence regions

---

The current framework of hypothesis testing is largely due to the work of Neyman and Pearson in the late 1920s, early 30s, complementing Fisher's work on estimation. As in estimation, we begin by postulating a statistical model but instead of seeking an estimator of  $\theta$  in  $\Theta$  we consider the question whether  $\theta \in \Theta_0 \subset \Theta$  or  $\theta \in \Theta_1 = \Theta - \Theta_0$  is mostly supported by the observed data. The discussion which follows will proceed in a similar way, though less systematically and formally, to the discussion of estimation. This is due to the complexity of the topic which arises mainly because one is asked to assimilate too many concepts too quickly just to be able to define the problem properly. This difficulty, however, is inherent in testing, if any proper understanding of the topic is to be attempted, and thus unavoidable. Every effort is made to ensure that the formal definitions are supplemented with intuitive explanations and examples. In Sections 14.1 and 14.2 the concepts needed to define a test and some criteria for 'good' tests are discussed using a simple example. In Section 14.3 the question of constructing 'good' tests is considered. Section 14.4 relates hypothesis testing to confidence estimation, bringing out the duality between the two areas. In Section 14.5 the related topic of prediction is considered.

#### 14.1 Testing, definitions and concepts

Let  $X$  be a random variable (r.v.) defined on the probability space  $(S, \mathcal{F}, P(\cdot))$  and consider the statistical model associated with  $X$ :

- (i)  $\Phi = \{f(x; \theta), \theta \in \Theta\}$ ;
- (ii)  $\mathbf{X} = (X_1, X_2, \dots, X_n)'$  is a random sample, from  $f(x; \theta)$ .

The problem of hypothesis testing is one of deciding whether or not some

conjecture about  $\theta$  of the form  $\theta$  belongs to some subset  $\Theta_0$  of  $\Theta$  is supported by the data  $\mathbf{x} = (x_1, x_2, \dots, x_n)'$ . We call such a conjecture the *null hypothesis* and denote it by  $H_0: \theta \in \Theta_0$ , where if the sample realisation  $\mathbf{x} \in C_0$  we accept  $H_0$ , if  $\mathbf{x} \in C_1$  we *reject* it. The mapping which enables us to define  $C_0$  and  $C_1$  we call a *test statistic*  $\tau(\mathbf{X}): \mathcal{X} \rightarrow \mathbb{R}$  (see Fig. 11.4).

In order to illustrate the concepts introduced so far let us consider the following example. Let  $X$  be the random variable representing the marks achieved by students in an econometric theory paper and let the statistical model be:

$$(i) \quad \Phi = \left\{ f(X; \theta) = \frac{1}{8\sqrt{(2\pi)}} \exp \left\{ -\frac{1}{2} \left( \frac{X - \theta}{8} \right)^2 \right\}, \quad \theta \in \Theta \equiv [0, 100]; \right.$$

$$(ii) \quad \mathbf{X} = (X_1, X_2, \dots, X_n)',$$

$n=40$  is a random sample from  $f(x; \theta)$ . The hypothesis to be tested is

$$H_0: \theta = 60 \quad (\text{i.e. } X \sim N(60, 64)), \quad \Theta_0 = \{60\}$$

against

$$H_1: \theta \neq 60 \quad (\text{i.e. } X \sim N(\mu, 64)), \quad \mu \neq 60, \quad \Theta_1 = [0, 100] - \{60\}.$$

Common sense suggests that if some ‘good’ estimator of  $\theta$ , say  $\bar{X}_n = (1/n) \sum_{i=1}^n X_i$ , for the sample realisation  $\mathbf{x}$  takes a value ‘around’ 60 then we will be inclined to accept  $H_0$ . Let us formalise this argument:

The *acceptance region* takes the form  $60 - \varepsilon \leq \bar{X}_n < 60 + \varepsilon$ ,  $\varepsilon > 0$ , or

$$C_0 = \{\mathbf{x}: |\bar{X}_n - 60| \leq \varepsilon\},$$

and

$$C_1 = \{\mathbf{x}: |\bar{X}_n - 60| \geq \varepsilon\} \quad \text{is the rejection region.}$$

The next question is, ‘how do we choose  $\varepsilon$ ?’ If  $\varepsilon$  is too small we run the risk of *rejecting  $H_0$  when it is true*; we call this *type I error*. On the other hand, if  $\varepsilon$  is too large we run the risk of *accepting  $H_0$  when it is false*; we call this *type II error*. Formally, if  $\mathbf{x} \in C_1$  (reject  $H_0$ ) and  $\theta \in \Theta_0$  ( $H_0$  is true) – type I error; if  $\mathbf{x} \in C_0$  (accept  $H_0$ ) and  $\theta \in \Theta_1$  ( $H_0$  is false) – type II error (see Table 14.1). The

Table 14.1

	$H_0$ accepted	$H_0$ rejected
$H_0$ true	correct	type I error
$H_0$ false	type II error	correct

hypothesis to be tested is formally stated as follows:

$$H_0: \theta \in \Theta_0, \quad \Theta_0 \subseteq \Theta. \quad (14.1)$$

Against the null hypothesis  $H_0$  we postulate the *alternative*  $H_1$  which takes the form:

$$H_1: \theta \notin \Theta_0 \quad (14.2)$$

or, equivalently,

$$H_1: \theta \in \Theta_1 \equiv \Theta - \Theta_0. \quad (14.3)$$

It is important to note at the outset that  $H_0$  and  $H_1$  are in effect hypotheses about the distribution of the sample  $f(\mathbf{x}; \theta)$ , i.e.

$$H_0: f(\mathbf{x}; \theta), \quad \theta \in \Theta_0, \quad H_1: f(\mathbf{x}; \theta), \quad \theta \in \Theta_1. \quad (14.4)$$

A hypothesis  $H_0$  or  $H_1$  is called *simple* if knowing  $\theta \in \Theta_0$  or  $\theta \in \Theta_1$  specifies  $f(\mathbf{x}; \theta)$  completely, otherwise it is called a *composite* hypothesis. That is, if  $f(\mathbf{x}; \theta)$ ,  $\theta \in \Theta_0$  or  $f(\mathbf{x}; \theta)$ ,  $\theta \in \Theta_1$  contain only one density function we say that  $H_0$  or  $H_1$  are simple hypotheses, respectively; otherwise they are said to be composite.

In testing a null hypothesis  $H_0$  against an alternative  $H_1$  the issue is to decide whether the sample realisation  $\mathbf{x}$  ‘supports’  $H_0$  or  $H_1$ . In the former case we say that  $H_0$  is *accepted*, in the latter  $H_0$  is *rejected*. In order to be able to make such a decision we need to formulate a mapping which relates  $\Theta_0$  to some subset of the observation space  $\mathcal{X}$ , say  $C_0$ , we call an *acceptance region*, and its complement  $C_1$  ( $C_0 \cup C_1 = \mathcal{X}$ ,  $C_0 \cap C_1 = \emptyset$ ) we call the *rejection region* (see Fig. 11.4). Obviously, in any particular situation we cannot say for certain in which of the four boxes in Table 14.1 we are in; at best we can only make a probabilistic statement relating to this. Moreover, if we were to choose  $\alpha$  ‘too small’ we run a higher risk of committing a type I error than of committing a type II error and vice versa. That is, there is a *trade off* between the probability of type I error, i.e.

$$Pr(\mathbf{x} \in C_1; \theta \in \Theta_0) = \alpha, \quad (14.5)$$

and the probability  $\beta$  of type II error, i.e.

$$Pr(\mathbf{x} \in C_0; \theta \in \Theta_1) = \beta. \quad (14.6)$$

Ideally we would like  $\alpha = \beta = 0$  for all  $\theta \in \Theta$  which is not possible for a fixed  $n$ . Moreover, we cannot control both simultaneously because of the trade-off between them. ‘How do we proceed, then?’ In order to help us decide let us consider the close analogy between this problem and the dilemma facing the jury in a trial of a criminal offence.

The jury in a criminal offence trial are instructed to choose between:

- $H_0$ : the accused is not guilty; and
- $H_1$ : the accused is guilty;

with their decision based on the evidence presented in the court. This evidence in hypothesis testing comes in the form of  $\Phi$  and  $X$ . The jury are instructed to accept  $H_0$  unless they have been convinced beyond any reasonable doubt otherwise. This requirement is designed to protect an innocent person from being convicted and it corresponds to choosing a small value for  $\alpha$ , the probability of convicting the accused when innocent. By adopting such a strategy, however, they are running the risk of letting a number of 'crooks off the hook'. This corresponds to being prepared to accept a relatively high value of  $\beta$ , the probability of not convicting the accused when guilty, in order to protect an innocent person from conviction. This is based on the moral argument that it is preferable to let off a number of guilty people rather than to sentence an innocent person. However, we can never be sure that an innocent person has not been sent to prison and the strategy is designed to keep the probability of this happening very low. A similar strategy is also adopted in hypothesis testing where a small value of  $\alpha$  is chosen and for a given  $\alpha$ ,  $\beta$  is minimised. Formally, this amounts to choosing  $\alpha^*$  such that

$$Pr(x \in C_1; \theta \in \Theta_0) = \alpha(\theta) \leq \alpha^* \quad \text{for } \theta \in \Theta_0 \quad (14.7)$$

and

$$Pr(x \in C_0; \theta \in \Theta_1) = \beta(\theta) \quad \text{is minimised for } \theta \in \Theta_1 \quad (14.8)$$

by choosing  $C_1$  or  $C_0$  appropriately.

In the case of the above example if we were to choose  $\alpha$ , say  $\alpha^* = 0.05$ , then

$$Pr(|\bar{X}_n - 60| > \varepsilon; \theta = 60) = 0.05. \quad (14.9)$$

This represents a probabilistic statement with  $\varepsilon$  being the only unknown. 'How do we determine  $\varepsilon$ , then?' Being a probabilistic statement it must be based on some distribution. The only random variable involved in the statement is  $\bar{X}_n$  and hence it has to be its sampling distribution. For the above probabilistic statement to have any operational meaning to enable us to determine  $\varepsilon$ , the distribution of  $\bar{X}_n$  must be known. In the present case we know that

$$\bar{X}_n \sim N\left(\theta, \frac{\sigma^2}{n}\right) \quad \text{where } \frac{\sigma^2}{n} = \frac{64}{40} = 1.6, \quad (14.10)$$

which implies that for  $\theta = 60$  (i.e. when  $H_0$  is true)

$$\tau(X) \equiv \left( \frac{\bar{X}_n - 60}{1.265} \right) \sim N(0, 1), \quad (14.11)$$

and thus the distribution of  $\tau(\cdot)$  is known completely (no unknown parameters). When this is the case this distribution can be used in conjunction with the above probabilistic statement to determine  $\varepsilon$ . In order to do this we need to relate  $|\bar{X}_n - 60|$  to  $\tau(\mathbf{X})$  (a statistic) for which the distribution is known. The obvious way to do this is to standardise the former, i.e. consider  $|\bar{X}_n - 60|/1.265$  which is equal to  $|\tau(\mathbf{X})|$ . This suggests changing the above probabilistic statement to the equivalent statement

$$Pr\left(\frac{|\bar{X}_n - 60|}{1.265} \geq c_z; \theta = 60\right) = 0.05 \quad \text{where } c_z = \frac{\varepsilon}{1.265}. \quad (14.12)$$

Given that the distribution of the statistic  $\tau(\mathbf{X})$  is symmetric and we want to determine  $c_z$  such that  $Pr(|\tau(\mathbf{X})| \geq c_z) = 0.05$  we should choose the value of  $c_z$  from the tables of  $N(0, 1)$  which leaves  $\alpha^*/2 = 0.025$  probability on either side of the distribution as shown in Fig. 14.1. The value of  $c_z$  given from the  $N(0, 1)$  tables is  $c_z = 1.96$ . This in turn implies that the rejection region for the test is

$$C_1 = \left\{ \mathbf{x}: \frac{|\bar{X}_n - 60|}{1.265} \geq 1.96 \right\} = \{ \mathbf{x}: |\tau(\mathbf{X})| \geq 1.96 \} \quad (14.13)$$

or

$$C_1 = \{ \mathbf{x}: |\bar{X}_n - 60| \geq 2.48 \}. \quad (14.14)$$

That is, for sample realisations  $\mathbf{x}$  which give rise to  $\bar{X}_n$  falling outside the interval  $(57.52, 62.48)$  we reject  $H_0$ .

Let us summarise the argument so far in order to keep the discussion in perspective. We set out to construct a test for  $H_0: \theta = 60$  against  $H_1: \theta \neq 60$  and intuition suggested the rejection region ( $|\bar{X}_n - 60| \geq \varepsilon$ ). In order to determine  $\varepsilon$  we had to

- (i) choose an  $\alpha$ ; and then
- (ii) define the rejection region in terms of some statistic  $\tau(\mathbf{X})$ .

The latter is necessary to enable us to determine  $\varepsilon$  via some known distribution. This is the distribution of the *test statistic*  $\tau(\mathbf{X})$  under  $H_0$  (i.e. when  $H_0$  is true).

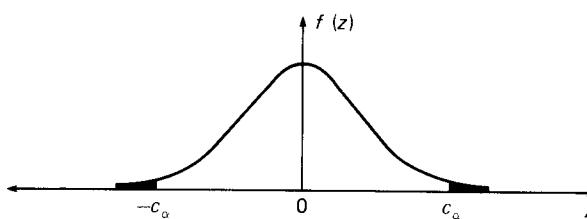


Fig. 14.1. The rejection region (14.13).

Given that  $C_1 = \{\mathbf{x}: |\tau(\mathbf{X})| \geq 1.96\}$  defines a test with  $\alpha = 0.05$ , the question which naturally arises is: ‘What do we need the probability of type II error  $\beta$  for?’ The answer is that we need  $\beta$  to decide whether the test defined in terms of  $C_1$  is a ‘good’ or a ‘bad’ test. As we mentioned at the outset, the way we decided to ‘solve’ the problem of the trade-off between  $\alpha$  and  $\beta$  was to choose a small value for  $\alpha$  and define  $C_1$  so as to minimise  $\beta$ . At this stage we do not know whether the test defined above is a ‘good’ test or not. Let us consider setting up the apparatus to enable us to consider the question of optimality.

## 14.2    Optimal tests

Since the acceptance and rejection regions constitute a partition of the observation space  $\mathcal{X}$ , i.e.  $C_0 \cup C_1 = \mathcal{X}$  and  $C_0 \cap C_1 = \emptyset$ , it implies that  $Pr(\mathbf{x} \in C_0) = 1 - Pr(\mathbf{x} \in C_1)$  for all  $\theta \in \Theta_1$ . Hence, minimisation of  $Pr(\mathbf{x} \in C_0)$  for all  $\theta \in \Theta_1$  is equivalent to maximising  $Pr(\mathbf{x} \in C_1)$  for all  $\theta \in \Theta_1$ .

*Definition 1*

*The probability of rejecting  $H_0$  when false at some point  $\theta_1 \in \Theta_1$ , i.e.  $Pr(\mathbf{x} \in C_1; \theta = \theta_1)$  is called the power of the test at  $\theta = \theta_1$ .*

Note that

$$Pr(\mathbf{x} \in C_1; \theta = \theta_1) = 1 - Pr(\mathbf{x} \in C_0; \theta = \theta_1) = 1 - \beta(\theta_1). \quad (14.15)$$

In the case of the example above we can define the power of the test at some  $\theta_1 \in \Theta_1$ , say  $\theta_1 = 54$ , to be  $Pr[|(\bar{X}_n - 60)|/1.265 \geq 1.96; \theta = 54]$ . ‘How do we calculate this probability?’ The temptation is to suggest using the same distribution as above, i.e.  $\tau(\mathbf{X}) \equiv (\bar{X}_n - 60)/1.265 \sim N(0, 1)$ . This is, however, wrong because  $\theta$  is no longer equal to 60; we assumed that  $\theta = 54$  and thus  $(\bar{X}_n - 54)/1.265 \sim N(0, 1)$ . This implies that

$$\tau(\mathbf{X}) \sim N\left(\frac{(54-60)}{1.265}, 1\right) \text{ for } \theta = 54.$$

Using this we can define the power of the test at  $\theta = 54$  to be

$$\begin{aligned} Pr\left(\left|\frac{\bar{X}_n - 60}{1.265}\right| \geq 1.96; \theta = 54\right) &= Pr\left(\frac{(\bar{X}_n - 54)}{1.265} \leq -1.96 - \frac{(54-60)}{1.265}\right) \\ &+ Pr\left(\frac{(\bar{X}_n - 54)}{1.265} \geq 1.96 - \frac{(54-60)}{1.265}\right) = 0.9973. \end{aligned}$$

Hence, the power of the test defined by  $C_1$  above is indeed very high for  $\theta = 54$ . In order to be able to decide on how good such a test is, however, we

need to calculate the power for all  $\theta \in \Theta_1$ . Following the same procedure the power of the test defined by  $C_1$  for  $\theta = 56, 58, 60, 62, 64, 66$  is as follows:

$$Pr(|\tau(\mathbf{X})| \geq 1.96; \theta = 56) = 0.8849,$$

$$Pr(|\tau(\mathbf{X})| \geq 1.96; \theta = 58) = 0.3520,$$

$$Pr(|\tau(\mathbf{X})| \geq 1.96; \theta = 60) = 0.05,$$

$$Pr(|\tau(\mathbf{X})| \geq 1.96; \theta = 62) = 0.3520,$$

$$Pr(|\tau(\mathbf{X})| \geq 1.96; \theta = 64) = 0.8849,$$

$$Pr(|\tau(\mathbf{X})| \geq 1.96; \theta = 66) = 0.9973.$$

As we can see, the power of the test increases as we go further away from  $\theta = 60$  ( $H_0$ ) and the power at  $\theta = 60$  equals the probability of type I error. This prompts us to define the power function as follows:

*Definition 2*

$\mathcal{P}(\theta) = Pr(\mathbf{x} \in C_1, \theta \in \Theta)$  is called the **power function** of the test defined by the rejection region  $C_1$ .

*Definition 3*

$\alpha = \max_{\theta \in \Theta_0} \mathcal{P}(\theta)$  is defined to be the **size** (or the significance level) of the test.

In the case where  $H_0$  is simple, say  $\theta = \theta_0$ , then  $\alpha = \mathcal{P}(\theta_0)$ . These definitions enable us to define a criterion for ‘a best’ test of a given size  $\alpha$  to be the one (if it exists) whose power function  $\mathcal{P}(\theta)$ ,  $\theta \in \Theta_1$  is maximum at every  $\theta$ .

*Definition 4*

A test of  $H_0: \theta \in \Theta_0$  against  $H_1: \theta \in \Theta_1$  as defined by some rejection region  $C_1$  is said to be **uniformly most powerful (UMP)** test of size  $\alpha$  if

$$(i) \quad \max_{\theta \in \Theta_0} \mathcal{P}(\theta) = \alpha;$$

$$(ii) \quad \mathcal{P}(\theta) \geq \mathcal{P}^*(\theta) \quad \text{for all } \theta \in \Theta_1;$$

where  $\mathcal{P}^*(\theta)$  is the power function of any other test of size  $\alpha$ .

As we saw above, in order to be able to determine the power function we need to know the distribution of the test statistic  $\tau(\mathbf{X})$  (in terms of which  $C_1$  is defined) under  $H_1$  (i.e. when  $H_0$  is false). The concept of a UMP test provides us with the criterion needed to choose between tests for the same  $H_0$ .

Let us consider the question of optimality for the size 0.05 test derived

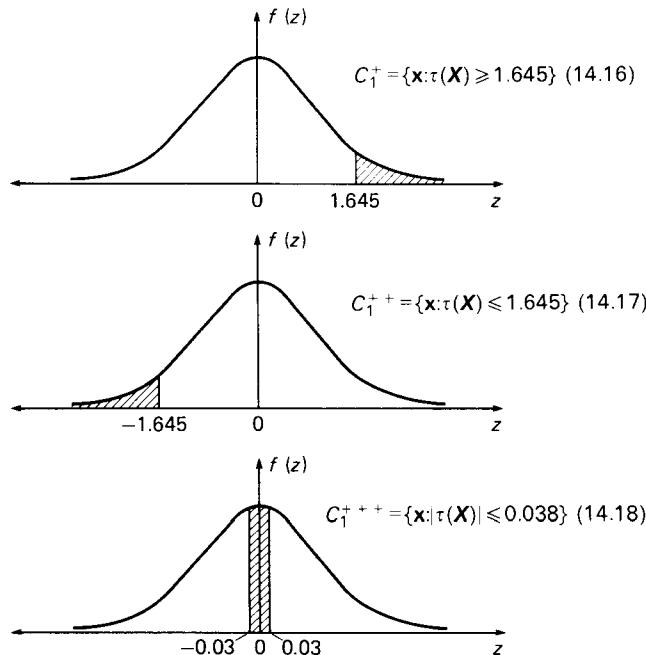


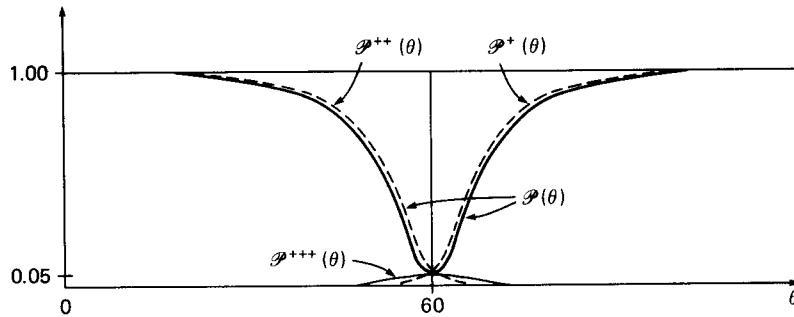
Fig. 14.2. The rejection regions (14.16), (14.17) and (14.18).

above with rejection region

$$C_1 = \{\mathbf{x}: |\tau(\mathbf{X})| \geq 1.96\}. \quad (14.19)$$

To that end we shall compare the power of this test with the power of the size 0.05 tests (Fig. 14.2), defined by the rejection regions. All the rejection regions define size 0.05 tests for  $H_0: \theta = 60$  against  $H_1: \theta \neq 60$ . In order to discriminate between ‘bad’, ‘good’ and ‘better’ tests we have to calculate their power functions and compare them. The diagram of the power functions  $\mathcal{P}(\theta)$ ,  $\mathcal{P}^+(\theta)$ ,  $\mathcal{P}^{++}(\theta)$ ,  $\mathcal{P}^{++*}(\theta)$  is illustrated in Fig. 14.3.

Looking at the diagram we can see that only one thing is clear ‘cut’;  $C_1^{++*}$  defines a very bad test, its power function being *dominated* by the other tests. Comparing the other three tests we can see that  $C_1^+$  is more powerful than the other two for  $\theta > 60$  but  $\mathcal{P}^+(\theta) < \alpha$  for  $\theta < 60$ .  $C_1^{++}$  is more powerful than the other two for  $\theta < 60$  but  $\mathcal{P}^{++}(\theta) < \alpha$  for  $\theta > 60$ , but none of the tests is more powerful over the whole range. That there is no UMP test of size 0.05 for  $H_0: \theta = 60$  against  $H_1: \theta \neq 60$ . As will be seen in the sequel, no UMP tests exist in most situations of interest in practice. The procedure adopted in such cases is to reduce the class of all tests to some subclass by imposing some more criteria and consider the question of UMP tests within

Fig. 14.3. The power functions  $\mathcal{P}(\theta)$ ,  $\mathcal{P}^+(\theta)$ ,  $\mathcal{P}^{++}(\theta)$ ,  $\mathcal{P}^{++\pm}(\theta)$ .

the subclass. One of the most important restrictions used in this context is the criterion of unbiasedness.

#### *Definition 5*

A test of  $H_0: \theta \in \Theta_0$  against  $\theta \in \Theta_1$  is said to be **unbiased** if

$$\max_{\theta \in \Theta_0} \mathcal{P}(\theta) \leq \max_{\theta \in \Theta_1} \mathcal{P}(\theta). \quad (14.20)$$

In other words, a test is unbiased if it rejects  $H_0$  more often when it is false than when it is true; a minimal but sensible requirement. Another form these added restrictions can take which reduces the problem to one where UMP do exist is related to the probability model  $\Phi$ . These include restrictions such as that  $\Phi$  belongs to the *one-parameter exponential family*.

In the case of the above example we can see that the test defined by  $C_1^{++\pm}$  is biased and  $C_1^+$  is now UMP within the class of *unbiased tests*. This is because  $C_1^+$  and  $C_1^{++}$  are biased for  $\theta < 60$  and  $\theta > 60$  respectively. It is obvious, however, that for

$$H_0: \theta = 60$$

against

$$H_1: \theta > 60$$

or

$$H_1^*: \theta < 60,$$

the tests defined by  $C_1^+$  and  $C_1^{++}$  are UMP, respectively. That is, for the *one-sided* alternatives there exist UMP tests given by  $C_1^+$  and  $C_1^{++}$ . It is important to note that in the case of  $H_1$  and  $H_1^*$  above the parameter space implicitly assumed is different. In the case of  $H_1$  the parameter space implicitly assumed is  $\Theta = [60, 100]$  and in the case of  $H_1^*$ ,  $\Theta = [0, 60]$ . This is needed in order to ensure that  $\Theta_0$  and  $\Theta_1$  constitute a partition of  $\Theta$ .

Collecting all the above concepts together we say that a *test* has been *defined* when the following components have been specified:

- (T1) a test statistic  $\tau(\mathbf{X})$ .
- (T2) the size of the test  $\alpha$ .
- (T3) the distribution of  $\tau(\mathbf{X})$  under  $H_0$  and  $H_1$ .
- (T4) the rejection region  $C_1$  (or, equivalently,  $C_0$ ).

Let us illustrate this using the marks example above. The test statistic is

$$\tau(\mathbf{X}) = \frac{n(\bar{X}_n - \theta)}{\sigma} = \frac{(\bar{X}_n - 60)}{1.27}; \quad (14.21)$$

we call it a statistic because  $\sigma$  is known and  $\theta$  is known under  $H_0$  and  $H_1$ . If we choose the size  $\alpha = 0.05$  the fact that  $\tau(\mathbf{X}) \sim N(0, 1)$  under  $H_0$  enables us to define the rejection region  $C_1 = \{\mathbf{x}: |\tau(\mathbf{X})| \geq c_\alpha\}$  where  $c_\alpha$  is determined from  $Pr(|\tau(\mathbf{X})| \geq c_\alpha; \theta = 60) = 0.05$  to be 1.96, from the standard normal tables, i.e. if  $\phi(z)$  denotes the density function of  $N(0, 1)$  then

$$\int_{-c_\alpha}^{c_\alpha} \phi(z) dz = 1 - \alpha. \quad (14.22)$$

In order to derive the power function we need the distribution of  $\tau(\mathbf{X})$  under  $H_1$ . Under  $H_1$  we know that

$$\tau^*(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \theta_1)}{\sigma} \sim N(0, 1), \quad (14.23)$$

for any  $\theta_1 \in \Theta_1$  and hence we can relate  $\tau(\mathbf{X})$  with  $\tau^*(\mathbf{X})$  by

$$\tau(\mathbf{X}) = \tau^*(\mathbf{X}) + \sqrt{n} \frac{(\theta_1 - \theta_0)}{\sigma} \quad (14.24)$$

to deduce that

$$\tau(\mathbf{X}) \sim N\left(\sqrt{n} \frac{(\theta_1 - \theta_0)}{\sigma}, 1\right) \quad (14.25)$$

under  $H_1$ . This enables us to define the power function as

$$\begin{aligned} \mathcal{P}(\theta_1) &= Pr(\mathbf{x}: |\tau(\mathbf{X})| \geq c_\alpha) \\ &= Pr\left(\tau^*(\mathbf{X}) \leq -c_\alpha - \sqrt{n} \frac{(\theta_1 - \theta_0)}{\sigma}\right) \\ &\quad + Pr\left(\tau^*(\mathbf{X}) \geq c_\alpha - \sqrt{n} \frac{(\theta_1 - \theta_0)}{\sigma}\right), \quad \theta_1 \in \Theta_1. \end{aligned} \quad (14.26)$$

Using the power function this test was shown to be UMP unbiased.

The most important component in defining a test is the test statistic for

which we need to know its distribution under both  $H_0$  and  $H_1$ . Hence, constructing an optimal test is largely a matter of being able to find a statistic  $\tau(\mathbf{X})$  which should have the following properties:

- (i)  $\tau(\mathbf{X})$  depends on  $\mathbf{X}$  via a 'good' estimator of  $\theta$ ; and
- (ii) the distribution of  $\tau(\mathbf{X})$  under both  $H_0$  and  $H_1$  does not depend on any unknown parameters. We call such a statistic a *pivot*.

It is no exaggeration to say that hypothesis testing is based on our ability to construct such pivots. When  $\mathbf{X}$  is a random sample from  $N(\mu, \sigma^2)$  pivots are readily available in the form of

$$\sqrt{n} \left( \frac{\bar{X}_n - \mu}{\sigma} \right) \sim N(0, 1), \quad \sqrt{n} \left( \frac{\bar{X}_n - \mu}{s} \right) \sim t(n-1), \quad (n-1) \frac{s^2}{\sigma^2} \sim \chi^2(n-1), \quad (14.27)$$

but in general these pivots are very hard to come by.

The first pivot was used above to construct tests for  $\mu$  when  $\sigma^2$  is known (both one-sided and two-sided tests). The second pivot can be used to set up similar tests for  $\mu$  when  $\sigma^2$  is unknown. For example, testing  $H_0: \mu = \mu_0$  against  $H_1: \mu \neq \mu_0$  the rejection region can be defined by

$$C_1 = \{ \mathbf{x}: |\tau_1(\mathbf{X})| \geq c_\alpha \} \quad \text{where } \tau_1(\mathbf{X}) = \sqrt{n} \frac{(\bar{X}_n - \mu)}{s}, \quad (14.28)$$

and  $c_\alpha$  can be determined by:  $\int_{-c_\alpha}^{c_\alpha} f(t) dt = 1 - \alpha$ ;  $f(t)$  being the density of the Student's  $t$ -distribution with  $n-1$  degrees of freedom. For  $H_0: \mu = \mu_0$  against  $H_1: \mu < \mu_0$  the rejection region takes the form

$$C_1 = \{ \mathbf{x}: \tau_1(\mathbf{X}) \geq c_\alpha \} \quad \text{with } \alpha = \int_{c_\alpha}^{\infty} f(t) dt, \quad (14.29)$$

determining  $c_\alpha$ .

The pivot

$$\tau_2(\mathbf{X}) = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1) \quad (14.30)$$

can be used to test hypotheses about  $\sigma^2$ . For example, in the case of a random sample from  $N(\mu, \sigma^2)$  testing  $H_0: \sigma^2 \geq \sigma_0^2$  against  $H_1: \sigma^2 < \sigma_0^2$  the rejection for an optimal test takes the form

$$C_1 = \{ \mathbf{x}: \tau_2(\mathbf{X}) \leq c_\alpha \}, \quad (14.31)$$

where  $c_\alpha$  is determined via

$$\int_0^{c_\alpha} d\chi^2(n-1) = \alpha.$$

### 14.3    Constructing optimal tests

In constructing the tests considered so far we used ad hoc intuitive arguments which led us to a pivot. As with estimation, it would be helpful if there were general methods for constructing optimal tests. It turns out that the availability of a method for constructing optimal tests depends crucially on the nature of the hypotheses ( $H_0$  and  $H_1$ ) or/and the probability model postulated. As far as the nature of  $H_0$  and  $H_1$  is concerned existence and optimality depend crucially on whether these hypotheses are simple or composite. As mentioned in Section 14.2 a hypothesis  $H_0$  or  $H_1$  is called *simple* if  $\Theta_0$  or  $\Theta_1$  contain just one point respectively. In the case of the ‘marks’ example above,  $\Theta_0 = \{60\}$  and  $\Theta_1 = \{[0, 60) \cup (60, 100]\}$ , i.e.  $H_0$  is simple and  $H_1$  is *composite* since it contains more than one point. Care should be exercised when  $\theta$  is a vector of unknown parameters because in such a case  $\Theta_0$  or  $\Theta_1$  must contain single vectors as well in order to be simple. For example, in the case of sampling from  $N(\mu, \sigma^2)$  and  $\sigma^2$  is not known,  $H_0: \mu = \mu_0$  is not a simple hypothesis since  $\Theta_0 = \{(\mu_0, \sigma^2), \sigma^2 \in \mathbb{R}_+\}$ .

#### (1)    Simple null and simple alternative

The theory concerning two simple hypotheses was fully developed in the 1920s by Neyman and Pearson. Let

$$\Phi = \{f(x; \theta), \theta \in \Theta\}$$

be the probability model and  $\mathbf{X} = (X_1, X_2, \dots, X_n)'$  be the sampling model and consider the simple null and simple alternative  $H_0: \theta = \theta_0$  and  $H_1: \theta = \theta_1, \Theta = \{\theta_0, \theta_1\}$ , i.e. there are only two possible distributions for  $\Phi$ , that is,  $f(x; \theta_0)$  and  $f(x; \theta_1)$ . Given the available data  $\mathbf{x}$  we want to choose between the two distributions. The following theorem provides us with sufficient conditions for the existence of a UMP test for this, the simplest of the cases in testing.

*Neyman–Pearson theorem*

Let  $\mathbf{X} = (X_1, X_2, \dots, X_n)'$  be a sample from a continuous distribution  $f(x; \theta), \theta \in \Theta = \{\theta_0, \theta_1\}$ . If there exists a test with rejection region

$$C_1 = \left\{ \mathbf{x}: \frac{f(\mathbf{x}; \theta_0)}{f(\mathbf{x}; \theta_1)} \leq c_x \right\} \quad (14.32)$$

for some positive constant  $c_x$ , such that

$$Pr(\mathbf{x} \in C_1; \theta = \theta_0) = \alpha, \quad (14.33)$$

then  $C_1$  defines a UMP test for  $H_0: \theta = \theta_0$  against  $H_1: \theta = \theta_1$  of size  $\alpha$ .

In this simple case

$$\mathcal{P}(\theta) = \begin{cases} \alpha & \text{for } \theta = \theta_0 \\ 1 - \beta & \text{for } \theta = \theta_1. \end{cases} \quad (14.34)$$

The Neyman–Pearson theorem suggests that it is intuitively sensible to base the acceptance or rejection of  $H_0$  on the relative values of the distributions of the sample evaluated at  $\theta = \theta_0$  and  $\theta = \theta_1$ , i.e. reject  $H_0$  if the ratio  $f(\mathbf{x}; \theta_0)/f(\mathbf{x}; \theta_1)$  is relatively small. This amounts to rejecting  $H_0$  when the evidence in the form of  $\mathbf{x}$  favour  $H_1$ , giving it a higher ‘support’. It is very important to note that the Neyman–Pearson theorem does not solve the problem completely because the problem of relating the ratio  $f(\mathbf{x}; \theta_0)/f(\mathbf{x}; \theta_1)$  to a pivotal quantity (test statistic) remains. Consider the case where  $X \sim N(\theta, \sigma^2)$ ,  $\sigma^2$  known, and we want to test  $H_0: \theta = \theta_0$  against  $H_1: \theta = \theta_1$  ( $\theta_0 < \theta_1$ ). From the Neyman–Pearson theorem we know that the rejection region defined in terms of the ratio

$$\begin{aligned} l(\mathbf{x}; \theta_0, \theta_1) &= \frac{f(\mathbf{x}; \theta_0)}{f(\mathbf{x}; \theta_1)} \\ &= \exp \left\{ \frac{n}{2\sigma^2} [(\theta_1^2 - \theta_0^2) - 2\bar{X}_n(\theta_1 - \theta_0)] \right\} \end{aligned} \quad (14.35)$$

can provide us with a UMP test if  $Pr(\mathbf{x} \in C_1; \theta = \theta_0) = \alpha$  exists for some  $\alpha$ . The ratio as it stands is not a proper pivot as we defined it. We know, however, that any monotonic transformation of the ratio generates the same family of rejection regions. Thus we can define

$$\begin{aligned} \tau(\mathbf{X}) &= \sqrt{n} \frac{(\bar{X}_n - \theta)}{\sigma} = \frac{-\sigma}{\sqrt{[n(\theta_1 - \theta_0)]}} \left[ \log l(\mathbf{x}, \theta_0, \theta_1) \right. \\ &\quad \left. + \frac{\sqrt{n}}{\sigma} \left( \frac{\theta_1 + \theta_0}{2} - \theta \right) \right], \end{aligned} \quad (14.36)$$

in terms of which we can define the rejection region as

$$C_1 = \{\mathbf{x}: \tau(\mathbf{X}) \geq c_x^*\}. \quad (14.37)$$

$C_1$  defines a UMP test of size  $\alpha$  if

$$Pr(\mathbf{x} \in C_1; \theta = \theta_0) = \alpha \quad \text{exists.} \quad (14.38)$$

*Remark:* in the case of a discrete random variable

$$Pr(\mathbf{x} \in C_1; \theta = \theta_0) = \alpha,$$

$\alpha$  might not exist since the random variable takes discrete values.

For example if  $\alpha=0.05$ ,  $c_x^*=1.645$  and the power of the test is

$$\mathcal{P}(\theta_1) = \left\{ \tau_1(\mathbf{X}) \geq c_x^* - \frac{\sqrt{n}(\theta_1 - \theta_0)}{\sigma} \right\} = 1 - \beta, \quad (14.39)$$

where

$$\tau_1(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \theta_0)}{\sigma} \sim N(0, 1) \quad \text{under } H_1. \quad (14.40)$$

In this case we can control  $\beta$  if we *can increase the sample size* since

$$1 - \beta = Pr(\tau_1(\mathbf{X}) \leq c_x^{**}) \Rightarrow c_x^* + c_x^{**} = \frac{\sqrt{n}}{\sigma}(\theta_1 - \theta_0). \quad (14.41)$$

For the hypothesis  $H_0: \theta = \theta_0$  against  $H_1: \theta = \theta_1$  when  $\theta_1 < \theta_0$  the test statistic takes the form

$$\begin{aligned} \tau(\mathbf{X}) &= \frac{\sqrt{n}(\bar{X}_n - \theta)}{\sigma} \\ &= \frac{\sigma}{\sqrt{[n(\theta_1 - \theta_0)]}} \left[ \log l(\mathbf{x}; \theta_0, \theta_1) \right. \\ &\quad \left. + \frac{\sqrt{n}}{\sigma} \left( \frac{\theta_1 + \theta_0}{2} - \theta \right) \right] ((\theta_1 - \theta_0) < 0), \end{aligned}$$

which gives rise to the rejection region

$$C_1 = \{ \mathbf{x}: \tau(\mathbf{X}) \leq c_x^* \}. \quad (14.42)$$

## (2)      **Composite null and composite alternative (one parameter case)**

For the hypothesis

$$H_0: \theta \geq \theta_0$$

against

$$H_1: \theta < \theta_0$$

being the other extreme of two simple hypotheses, no such results as the Neyman–Pearson theorem exist and it comes as no surprise that no UMP tests exist in general. The only result of some interest in this case is that if we restrict the probability model to require the density functions to have *monotone likelihood ratio* in the test statistic  $\tau(\mathbf{X})$  then UMP tests do exist. This result is of limited value, however, since it does not provide us with a *method* to derive  $\tau(\mathbf{X})$ .

(3) **Simple  $H_0$  against composite  $H_1$** 

In the case where we want to test  $H_0: \theta = \theta_0$  against  $H_1: \theta > \theta_0$  (or  $\theta < \theta_0$ ) uniformly most powerful (UMP) tests do not exist in general. In some particular cases, however, such UMP tests do exist and the Neyman–Pearson theorem can help us derive them. If the UMP test for the simple  $H_0: \theta = \theta_0$  against the simple  $H_1: \theta = \theta_1$  does not depend on  $\theta_1$  then the same test is UMP for the one-sided alternative  $\theta > \theta_0$  (or  $\theta < \theta_0$ ). In the example discussed above the tests defined by

$$C_1 = \{\mathbf{x}: \tau(\mathbf{X}) \geq c_\alpha^*\} \quad (14.43)$$

and

$$C_2 = \{\mathbf{x}: \tau(\mathbf{X}) \leq c_\alpha^*\} \quad (14.44)$$

are also UMP for the hypotheses  $H_0: \theta = \theta_0$  against  $H_1: \theta > \theta_0$  and  $H_0: \theta = \theta_0$  against  $H_1: \theta < \theta_0$ , respectively. This is indeed confirmed by the diagram of the power function derived for the ‘marks’ example above. Another result in the simple class of hypotheses is available in the case where sampling is from a *one-parameter exponential family of densities* (normal, binomial, Poisson, etc.). In such cases UMP tests do exist for one-sided alternatives.

*Two-sided alternatives*

For testing  $H_0: \theta = \theta_0$  against  $H_1: \theta \neq \theta_0$  no UMP tests exist in general. This is rather unfortunate since most tests in practice are of this type. One interesting result in this case is that if we restrict the probability model to the *one-parameter exponential family* and narrow down the class of tests by imposing *unbiasedness*, then we know that UMP tests do exist. The test defined by the rejection region

$$C_1 = \{\mathbf{x}: |\tau(\mathbf{X})| \geq c_\alpha\} \quad (14.45)$$

(see ‘marks’ example) is indeed UMP unbiased; the one-sided alternative tests being biased over the whole of  $\Theta$ .

**14.4 The likelihood ratio test procedure**

The discussion so far suggests that no UMP tests exist for a wide variety of cases which are important in practice. However, the likelihood ratio test procedure yields very satisfactory tests for a great number of cases where none of the above methods is applicable. It is particularly valuable in the case where both hypotheses are composite and  $\theta$  is a vector of parameters. This procedure not only has a lot of intuitive appeal but also frequently leads to UMP tests or UMP unbiased tests (when such exist).

Consider

$$H_0: \theta \in \Theta_0$$

against

$$H_1: \theta \in \Theta_1.$$

Let the likelihood function be  $L(\theta; \mathbf{x})$ , then the likelihood ratio is defined by

$$\lambda(\mathbf{x}) = \frac{\max_{\theta \in \Theta_0} L(\theta; \mathbf{x})}{\max_{\theta \in \Theta} L(\theta; \mathbf{x})} = \frac{L(\hat{\theta}; \mathbf{x})}{L(\tilde{\theta}; \mathbf{x})}. \quad (14.46)$$

The numerator measures the highest 'support'  $\mathbf{x}$  renders to  $\theta \in \Theta_0$  and the denominator measures the maximum value of the likelihood function (see Fig. 14.4). By definition  $\lambda(\mathbf{x})$  can never exceed unity and the smaller it is the less  $H_0$  is 'supported' by the data. This suggests that the rejection region based on  $\lambda(\mathbf{x})$  must be of the form

$$C_1 = \{\mathbf{x}: \lambda(\mathbf{x}) \leq k\}, \quad 0 \leq k \leq 1, \quad (14.47)$$

and the size being defined by

$$\max_{\theta \in \Theta_0} \mathcal{P}(\theta) = \alpha.$$

$\alpha$  and  $k$  as well as the power function can only be defined when the

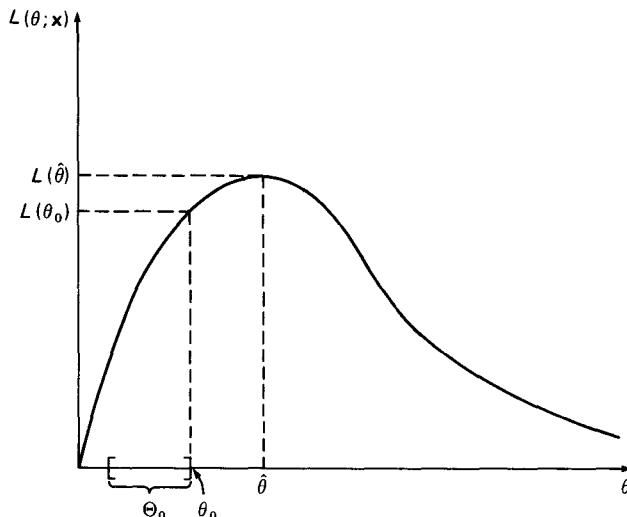


Fig. 14.4. The likelihood ratio test.

distribution of  $\lambda(\mathbf{x})$  under both  $H_0$  and  $H_1$  is known. This is usually the exception rather than the rule. The exceptions arise when  $\Phi$  is a normal family of densities and  $\mathbf{X}$  is a random sample in which case  $\lambda(\mathbf{x})$  is often a monotone function of some of the pivots we encountered above. Let us illustrate the procedure and the difficulties arising by considering several examples.

*Example 1*

Let

$$\Phi = \left\{ f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{(2\pi)}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+ \right\},$$

$$\mathbf{X} = (X_1, X_2, \dots, X_n)'$$

be the probability model and  $\mathbf{X} = (X_1, X_2, \dots, X_n)'$  be a random sample from  $f(x; \mu, \sigma^2)$ ,  $H_0: \mu = \mu_0$  against  $H_1: \mu \neq \mu_0$ .

$$L(\theta; \mathbf{x}) = (2\pi\sigma^2)^{-n/2} \exp\left\{\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right\}$$

$$\Rightarrow \lambda(\mathbf{x}) = \left[ \frac{\sum_{i=1}^n (x_i - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{X}_n)^2} \right]^{-n/2}.$$

At first sight it might seem an impossible task to determine the distribution of  $\lambda(\mathbf{x})$ . Note, however, that

$$\sum_{i=1}^n (x_i - \mu_0)^2 = \sum_{i=1}^n (x_i - \bar{X}_n)^2 + n(\bar{X}_n - \mu_0)^2,$$

which implies that

$$\lambda(\mathbf{x}) = 1 + \left( \frac{n(\bar{X}_n - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{X}_n)^2} \right)^{-n/2} = \left( 1 + \frac{W^2}{n-1} \right)^{-n/2},$$

where  $W = \sqrt{n}[(\bar{X}_n - \mu_0)/s] \sim t(n-1)$  under  $H_0$ ,

$$W \sim t(n-1; \delta) \text{ under } H_1, \quad \delta = \frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma}, \quad \mu_1 \in \Theta_1.$$

Since  $\lambda(\mathbf{x})$  is a monotone decreasing function of  $W$  the rejection region takes the form

$$C_1\{\mathbf{x}: |W| \geq c_z\},$$

and  $\alpha$ ,  $c_\alpha$  and  $\mathcal{P}(\theta)$  can be derived from the distribution of  $W$ .

*Example 2*

In the context of the statistical model of example 1 consider

$$H_0: \sigma^2 = \sigma_0^2$$

against

$$H_1: \sigma^2 \neq \sigma_0^2, \quad \Theta \equiv \mathbb{R} \times \mathbb{R}_+$$

and

$$\Rightarrow \begin{aligned} \Theta_0 &= \{\boldsymbol{\theta} = (\mu, \sigma_0^2), \mu \in \mathbb{R}\} \\ \lambda(\mathbf{x}) &= \frac{1}{n} \left[ \sum_{i=1}^n \frac{(X_i - \bar{X}_n)^2}{\sigma_0^2} \right]^{n/2} \exp \left\{ \frac{1}{2} \left( \sum_{i=1}^n \frac{(X_i - \bar{X}_n)^2}{\sigma_0^2} - n \right) \right\}. \end{aligned}$$

The inequality  $\lambda(\mathbf{x}) \leq k$  is equivalent to  $v \leq k_1$  or  $v \geq k_2$  where

$$v = \frac{1}{\sigma_0^2} \sum_{i=1}^n (x_i - \bar{X}_n)^2 \sim \chi^2(n-1) \quad \text{under } H_0$$

and

$$v \sim \chi^2(n-1; \delta) \quad \text{under } H_1, \quad \delta = \frac{n\sigma_1^2}{\sigma_0^2}, \quad \sigma_1^2 \in \Theta_1,$$

with  $k_1$  and  $k_2$  defined by

$$\int_{k_1}^{k_2} d\chi^2(n-1) = 1 - \alpha,$$

e.g. if  $\alpha = 0.1$ ,  $n-1 = 30$ ,  $k_1 = 18.5$ ,  $k_2 = 29.3$ .

Hence, the rejection region is  $C_1 = \{\mathbf{x}: v \leq k_1 \text{ or } v \geq k_2\}$ . Using the analogy between this and the various tests of  $\mu$  we encountered so far we can postulate that in the case of the one-sided hypotheses:

- (i)  $H_0: \sigma^2 \geq \sigma_0^2$ ,  $H_1: \sigma^2 < \sigma_0^2$  the rejection region is  $C_1 = \{\mathbf{x}: v \leq k_1\}$ ;
- (ii)  $H_0: \sigma^2 \leq \sigma_0^2$ ,  $H_1: \sigma^2 > \sigma_0^2$ ,  $C_1 = \{\mathbf{x}: v \geq k_2\}$ .

The question arising at this stage is: ‘What use is the likelihood ratio test procedure if the distribution of  $\lambda(\mathbf{X})$  is only known when a well-known pivot exists already?’ The answer is that it is reassuring to know that the procedure in these cases leads to certain well-known pivots because the likelihood ratio test procedure is of considerable importance when no such pivots exist. Under certain conditions we can derive the asymptotic distribution of  $\lambda(\mathbf{X})$ . We can show that under certain conditions

$$-2 \log \lambda(\mathbf{X}) \stackrel{\alpha}{\sim} \chi^2(r) \tag{14.48}$$

$\stackrel{H_0}{\sim}$  reads ‘asymptotically distributed under  $H_0$ ’),  $r$  being the number of parameters tested. This will be pursued further in Section 16.2.

### 14.5 Confidence estimation

In point estimation when an estimator  $\hat{\theta}$  of  $\theta$  is constructed we usually think of it not just as a point but as a point surrounded by some region of possible error, i.e.  $\hat{\theta} \pm e$ , where  $e$  is related to the standard error of  $\hat{\theta}$ . This can be viewed as a crude form of a confidence interval for  $\theta$

$$(\hat{\theta} - e \leq \theta \leq \hat{\theta} + e); \quad (14.49)$$

crude because there is no guarantee that such an interval will include  $\theta$ . Indeed, we can show that the probability the  $\theta$  does not belong to this interval is actually non-negative. In order to formalise this argument we need to attach probabilities to such intervals. In general, interval estimation refers to constructing random intervals of the form

$$(\underline{\tau}(X) \leq \theta \leq \bar{\tau}(X)), \quad (14.50)$$

together with an associated probability for such a statement being valid.  $\underline{\tau}(X)$  and  $\bar{\tau}(X)$  are two statistics referred to as the lower and upper ‘bound’ respectively; they are in effect stochastic bounds on  $\theta$ . The associated probability will take the form

$$Pr(\underline{\tau}(X) \leq \theta \leq \bar{\tau}(X)) = 1 - \alpha, \quad (14.51)$$

where the probabilistic statement is based on the distribution of  $\underline{\tau}(X)$  and  $\bar{\tau}(X)$ . The main problem is to construct such statistics for which the distribution does not depend on the unknown parameter(s)  $\theta$ . This, however, is the same problem as in hypothesis testing. In that context we ‘solved’ the problem by seeking what we called *pivots* and intuition suggests that the same quantities might be of use in the present context. It turns out that not only this is indeed the case but the similarity between interval estimation and hypothesis testing does not end here. Any size  $\alpha$  test about  $\theta$  can be transformed directly to an interval estimator of  $\theta$  with  $1 - \alpha$  confidence level.

#### Definition 6

The interval  $(\underline{\tau}(X), \bar{\tau}(X))$  is called a  $(1 - \alpha)$  **confidence interval** for  $\theta$  if for all  $\theta \in \Theta$

$$Pr(\underline{\tau}(X) \leq \theta \leq \bar{\tau}(X)) \geq 1 - \alpha. \quad (14.52)$$

$(1 - \alpha)$  is called the probability of coverage of the interval and the statement

suggests that in the long-run (in repeated experiments) the random interval  $(\underline{\tau}(\mathbf{X}), \bar{\tau}(\mathbf{X}))$  will include the ‘true’ but unknown  $\theta$ . For any particular realisation  $\mathbf{x}$ , however, we do not know ‘for sure’ whether  $(\underline{\tau}(\mathbf{X}), \bar{\tau}(\mathbf{X}))$  includes or not the ‘true’  $\theta$ ; we are only  $(1 - \alpha)$  confident that it does. The duality between hypothesis testing and confidence intervals can be seen in the ‘marks’ example discussed above. For the null hypothesis

$$H_0: \theta = \theta_0, \quad \theta_0 \in \Theta$$

against

$$H_1: \theta \neq \theta_0,$$

we constructed a size  $\alpha$  test based on the acceptance region

$$C_0(\theta_0) = \left\{ \mathbf{x}: \theta_0 - c_\alpha \frac{\sigma}{\sqrt{n}} \leq \bar{X}_n \leq \theta_0 + c_\alpha \frac{\sigma}{\sqrt{n}} \right\}, \quad (14.53)$$

with  $c_\alpha$  defined by

$$\int_{-\alpha}^{c_\alpha} \phi(z) dz = 1 - \alpha, \quad Z \sim N(0, 1). \quad (14.54)$$

This implies that  $Pr(\mathbf{x} \in C_0, \theta = \theta_0) = 1 - \alpha$  and hence by a simple manipulation of  $C_0$  we can define the  $(1 - \alpha)$  confidence interval

$$C(\mathbf{X}) = \left\{ \theta_0 : \bar{X}_n - c_\alpha \frac{\sigma}{\sqrt{n}} \leq \theta_0 \leq \bar{X}_n + c_\alpha \frac{\sigma}{\sqrt{n}} \right\}, \quad (14.55)$$

$$Pr(\theta_0 \in C) = 1 - \alpha. \quad (14.56)$$

In general, any acceptance region for a size  $\alpha$  test can be transformed into a  $(1 - \alpha)$  confidence interval for  $\theta$  by changing  $C_0$ , a function of  $\mathbf{x} \in \mathcal{X}$ , to  $C$ , a function of  $\theta_0 \in \Theta$ .

One-sided tests correspond to one-sided confidence intervals of the form

$$Pr(\underline{\tau}(\mathbf{X}) \leq \theta) \geq 1 - \alpha \quad (14.57)$$

or

$$Pr(\theta \leq \bar{\tau}(\mathbf{X})) \geq 1 - \alpha. \quad (14.58)$$

In general when  $\Theta = \mathbb{R}^m$ ,  $m \geq 1$ , the family of subsets  $C(\mathbf{X})$  of  $\Theta$  where  $C(\mathbf{X})$  depends on  $\mathbf{X}$  but not  $\theta$  is called a *random region*. For example,

$$C(\mathbf{X}) = \{ \theta : \underline{\tau}(\mathbf{X}) \leq \theta \leq \bar{\tau}(\mathbf{X}) \} \quad \text{or} \quad C(\mathbf{X}) = \{ \theta : \underline{\tau}(\mathbf{X}) \leq \theta \}. \quad (14.59)$$

The problem of confidence estimation is one of constructing a random region  $C(\mathbf{X})$  such that, for a given  $\alpha \in (0, 1)$ ,

$$Pr(\mathbf{x} : \theta \in C(\mathbf{X}) / \theta) \geq 1 - \alpha, \quad \text{for all } \theta \in \Theta. \quad (14.60)$$

It is interesting to note that  $C(\mathbf{X})$  could be interpreted as

$$C(\mathbf{X}) = \{\boldsymbol{\theta}: \underline{\tau}_i(\mathbf{X}) \leq \theta_i \leq \bar{\tau}_i(\mathbf{X}), i = 1, 2, \dots, m\}, \quad (14.61)$$

in which case if  $(\underline{\tau}_i(\mathbf{X}), \bar{\tau}_i(\mathbf{X}))$  represent independent  $(1-\alpha_i)$  confidence intervals

$$i = 1, 2, \dots, m \quad \text{and} \quad (1-\alpha) = \prod_{i=1}^m (1-\alpha_i). \quad (14.62)$$

The duality between hypothesis testing and confidence estimation does not end at the construction stage. The various properties of tests have corresponding counterparts in confidence estimation.

#### *Definition 7*

A family of  $(1-\alpha)$  level confidence regions  $C(\mathbf{X})$  is said to be **uniformly most accurate (UMA)** among  $(1-\alpha)$  level confidence regions  $C^*(\mathbf{X})$  if

$$\Pr(\mathbf{x}: \boldsymbol{\theta} \in C(\mathbf{X})/\boldsymbol{\theta}) \leq \Pr(\mathbf{x}: \boldsymbol{\theta} \in C^*(\mathbf{X})/\boldsymbol{\theta}) \quad \text{for all } \boldsymbol{\theta} \in \Theta. \quad (14.63)$$

This clearly shows that when power is reinterpreted as accuracy it provides us with the basic optimality criterion in confidence estimation. It turns out (not surprisingly) that *UMP* tests lead to *UMA* confidence regions. This is because

$$\boldsymbol{\theta} \in C(\mathbf{X}) \subseteq \Theta \quad \text{if and only if } \mathbf{x} \in C_0(\boldsymbol{\theta}) \subseteq \mathcal{X}, \quad (14.64)$$

where  $C_0(\boldsymbol{\theta})$  represents the acceptance region of  $H_0: \boldsymbol{\theta} = \boldsymbol{\theta}_0$ . In effect the confidence region  $C(\mathbf{X})$  can be formed by

$$C(\mathbf{x}) = \{\boldsymbol{\theta}_0: \mathbf{x} \in C_0(\boldsymbol{\theta}_0)\}, \quad (14.65)$$

and the acceptance region  $C_0(\boldsymbol{\theta})$  by

$$C(\boldsymbol{\theta}_0) = \{\mathbf{x}: \boldsymbol{\theta}_0 \in C(\mathbf{x})\}, \quad (14.66)$$

hence

$$\Pr(\mathbf{x}: \mathbf{x} \in C_0(\boldsymbol{\theta}_0)/\boldsymbol{\theta} = \boldsymbol{\theta}_0) = \Pr(\mathbf{x}: \boldsymbol{\theta}_0 \in C(\mathbf{x})/\boldsymbol{\theta} = \boldsymbol{\theta}_0) \geq 1 - \alpha. \quad (14.67)$$

This duality between  $C_0(\boldsymbol{\theta}_0)$  and  $C(\mathbf{X})$  is illustrated below for the above example assuming that  $n = 1$  to enable us to draw the graph given in Fig. 14.5.

Continuing with this duality it comes as no surprise to learn that unbiased tests give rise to unbiased confidence regions and vice versa.

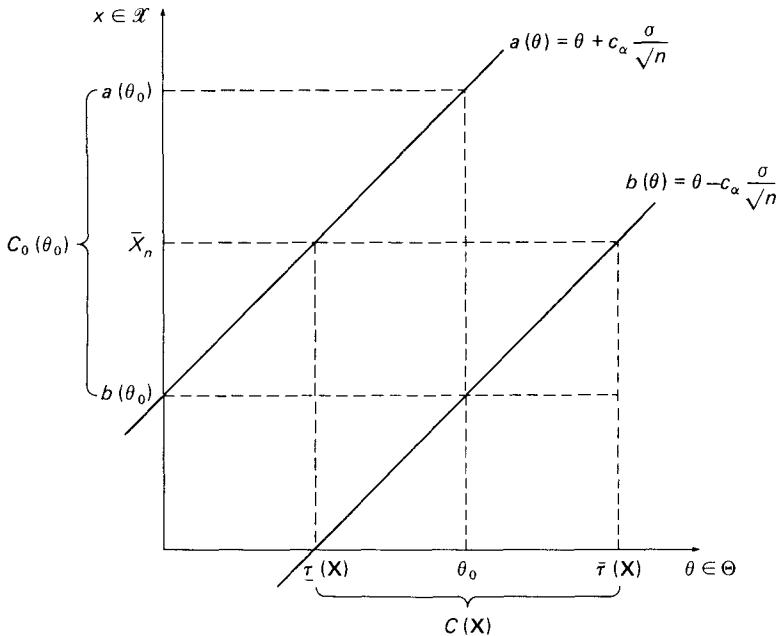


Fig. 14.5. The duality between hypothesis testing and interval estimation.

#### Definition 8

A confidence region  $C(\mathbf{X})$  for  $\theta_1$  is said to be **unbiased** at confidence level  $(1-\alpha)$  if

$$\Pr(\mathbf{x}: \theta_1 \in C(\mathbf{x}) / \theta_2) \leq 1 - \alpha \quad \text{for } \theta_1, \theta_2 \in \Theta. \quad (14.68)$$

In general, a ‘good’ test will give rise to a good confidence region and vice versa (see Lehmann (1959)).

#### 14.6 Prediction

In the context of a statistical model as defined by the probability and sampling model components, prediction refers to the construction of an ‘optimal’ Borel function  $l(\cdot)$  (see Chapter 6) of the form:

$$l(\cdot): \Theta \rightarrow \mathcal{X}, \quad (14.69)$$

which purports to provide a ‘good guess’ for the value of a random variable  $X_{n+1}$  which does not belong to the postulated sample. If we denote the sample with  $\mathbf{X}_n \equiv (X_1, \dots, X_n)'$  and its distribution with  $D(\mathbf{X}_n; \theta)$  then we need to construct  $l(\theta)$  which for a good estimator  $\hat{\theta}_n$  of  $\theta$ ,  $\hat{X}_{n+1} = l(\hat{\theta}_n)$  is a

'good' predictor of  $X_{n+1}$ . Given that  $\hat{\theta}_n = h(\mathbf{X}_n)$  we can define  $l(\cdot)$  as a function of the sample directly, i.e.  $l(\hat{\theta}_n) = l(\mathbf{X}_n)$ . Properties of optimal predictors were discussed in Section 12.3, but no methods of constructing such predictors were mentioned. The purpose of this section is to consider this problem briefly.

The problem of constructing optimal predictors refers to the question of 'how do we choose the function  $l(\cdot)$  so as the resulting predictor to satisfy certain desirable properties?' To be able to answer this question we need to specify what the desirable properties are. The single most widely used criterion for a good predictor is *minimum mean square error* (MSE). This criterion suggests choosing  $l(\cdot)$  in such a way so as to minimise

$$E(X_{n+1} - l(\mathbf{X}_n))^2, \quad (14.70)$$

where  $E(\cdot)$  is defined in terms of the joint distribution of  $X_{n-1}$  and  $\mathbf{X}_n$ , say,  $D(X_{n+1}, \mathbf{X}_n; \psi)$ . It turns out that the solution of this minimisation problem is theoretically extremely simple. This is because (70) can be expressed in the form

$$\begin{aligned} & E(X_{n+1} - l(\mathbf{X}_n))^2 \\ &= E(\{X_{n+1} - E(X_{n+1}/\sigma(\mathbf{X}_n))\} + \{E(X_{n+1}/\sigma(\mathbf{X}_n)) - l(\mathbf{X}_n)\})^2 \\ &= E(X_{n+1} - E(X_{n+1}/\sigma(\mathbf{X}_n)))^2 + E(E(X_{n+1}/\sigma(\mathbf{X}_n)) - l(\mathbf{X}_n))^2 \\ &\quad + 2E(\{X_{n+1} - E(X_{n+1}/\sigma(\mathbf{X}_n))\}\{E(X_{n+1}/\sigma(\mathbf{X}_n)) - l(\mathbf{X}_n)\}). \end{aligned} \quad (14.71)$$

Using the properties CE5 and SCE5 of Section 7.2 we can show that the last term is equal to zero. Hence, (71) is minimised when

$$l(\mathbf{X}_n) = E(X_{n+1}/\sigma(\mathbf{X}_n)). \quad (14.72)$$

When this is the case the second term is also equal to zero. That is, the form of the *predictor*  $\hat{X}_{n+1} = l(\mathbf{X}_n)$  which minimises (70) is

$$\hat{X}_{n+1} = E(X_{n+1}/\sigma(\mathbf{X}_n)), \quad (14.73)$$

where  $\sigma(\mathbf{X}_n)$  is the  $\sigma$ -field generated by  $\mathbf{X}_n$  (see Chapter 4).

As argued in Sections 5.4 and 7.2, the functional form of  $E(X_{n+1}/\sigma(\mathbf{X}_n))$  depends entirely on the form of the joint distribution  $D(X_{n+1}, \mathbf{X}_n; \psi)$ . For example, in the case where  $D(X_{n+1}, \mathbf{X}_n; \psi)$  is *multivariate normal* then the conditional expectation is linear, i.e.

$$E(X_{n+1}/\sigma(\mathbf{X}_n)) = \boldsymbol{\beta}' \mathbf{X}_n \quad (14.74)$$

(see Chapter 15). In practice, when  $D(X_{n+1}, \mathbf{X}_n; \psi)$  is not known *linear predictors* are commonly used as approximations to the particular

functional form of

$$E(X_{n+1}/\sigma(\mathbf{X}_n)) = g(\mathbf{X}_n). \quad (14.75)$$

In such cases the joint distribution is implicitly assumed to be closely approximated by a normal distribution.

The prediction value of  $X_{n+1}$  will take the form

$$\hat{x}_{n+1} = E(X_{n+1}/\mathbf{X}_n = \mathbf{x}_n), \quad (14.76)$$

where  $\mathbf{x}_n$  refers to the observed realisation of the sample  $\mathbf{X}_n$ . The intuition underlying (76) is that the best ‘guess’ for the value  $X_{n+1}$  must be the *average* of all its possible values, in view of the past realisations of  $X(\mathbf{X}_n = \mathbf{x}_n)$  (see Fig. 14.6).

It is important to note that in the case where  $X_{n+1}$  and  $\mathbf{X}_n$  are *independent* then

$$E(X_{n+1}/\sigma(\mathbf{X}_n)) = E(X_{n+1}). \quad (14.77)$$

That is, the conditional expectation coincides with the marginal expectation of  $X_{n+1}$  (see Chapters 6 and 7). This is the reason why in the case of the random sample  $\mathbf{X}_n$  where  $X_i \sim N(0, 1)$ ,  $i = 1, 2, \dots, n$ , if  $X_{n+1}$  is also assumed to have the same distribution, its best predictor (in MSE sense) is its mean, i.e.  $\hat{X}_{n+1} = (1/n) \sum_{i=1}^n X_i$  (see Section 12.3). It goes without saying that for prediction purposes we prefer to have non-random sampling models because the past history of the stochastic process  $\{X_n, n \geq 1\}$  will be of considerable value in such a case.

*Prediction regions* for  $X_{n+1}$  take the same form as confidence regions for

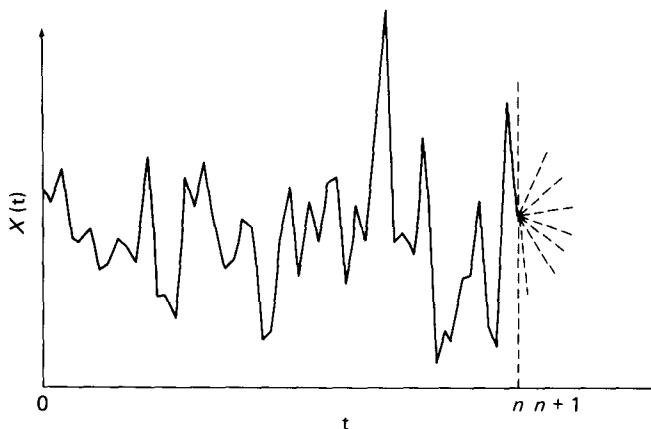


Fig. 14.6. Predicting  $X_{n+1}$  when  $X_1, X_2, \dots, X_n$  is known.

$\theta$  and the same analysis as in Section 14.5 goes through with minor interpretation changes.

### *Important concepts*

Null and alternative hypotheses, acceptance region, rejection region, test statistic, type I error, type II error, power of a test, the power function, size of a test, uniformly most powerful test, unbiased test, simple hypothesis, composite hypothesis, Neyman–Pearson lemma, likelihood ratio test, pivots, confidence region, confidence level, uniformly most accurate confidence regions, unbiased confidence region, optimal predictor, minimum mean square error.

### *Questions*

1. Explain the relationship between  $H_0$  and  $H_1$  and the distribution of the sample.
2. Describe the relationship between the acceptance and rejection regions and  $\Theta_0$  and  $\Theta_1$ .
3. Define the concepts of a test statistic, type I and type II errors and probabilities of type I and II errors.
4. Explain intuitively why we cannot control both probabilities of type I and type II errors. How do we ‘solve’ this problem in hypothesis testing?
5. Define and explain the concepts of the power of a test and the power function of a test.
6. Explain the concept of the size of a test.
7. Define and explain the concept of a UMP test.
8. State the components needed to define a test.
9. Explain why we need to know the distribution of the test statistic under both the null and the alternative hypotheses.
10. Define the concept of a pivot and explain its role in hypothesis testing.
11. Explain the concepts of one-sided and two-sided tests.
12. Explain the circumstances under which UMP tests exist.
13. Explain the Neyman–Pearson theorem and the likelihood ratio test procedure as ways of constructing optimal tests.
14. Explain intuitively the meaning of the statement

$$Pr(\underline{\tau}(\mathbf{X}) \leq \theta \leq \bar{\tau}(\mathbf{X})) = 1 - \alpha.$$

15. Define the concept of a  $(1 - \alpha)$  confidence region for  $\theta$ .
16. Explain the relationship between  $C_0(\theta_0)$ , the acceptance region for  $H_0: \theta = \theta_0$  against  $H_1: \theta \neq \theta_0$  and the confidence interval  $C(\mathbf{X})$  for  $\theta_0$ .

17. Define and explain the concept of a  $(1 - \alpha)$  uniformly most accurate confidence region.

### *Exercises*

1. For the ‘marks’ example of Section 14.2 construct a size 0.05 test for  $H_0: \theta = 60$  against  $H_1: \theta < 60$ . Is it unbiased? Using this, construct a 0.95 significance level confidence interval for  $\theta$ .
2. Let  $X \sim N(\mu, \sigma^2)$  and consider the following hypotheses:
  - (i)  $H_0: \mu \leq \mu_0, H_1: \mu > \mu_0, \sigma^2 > 0, \mu_0$  – known;
  - (ii)  $H_0: \sigma^2 \geq \sigma_0^2, H_1: \sigma^2 < \sigma_0^2, \mu \in \mathbb{R}, \sigma_0^2$  – known;
  - (iii)  $H_0: \mu = \mu_0, \sigma^2 = \sigma_0^2, H_1: \mu \neq \mu_0, \sigma^2 \neq \sigma_0^2;$
  - (iv)  $H_0: \mu = \mu_0, \sigma^2 \geq \sigma_0^2, H_1: \mu \neq \mu_0, \sigma^2 < \sigma_0^2.$

State whether the above null and alternative hypotheses are simple or composite and explain your answer.

3. Let  $\mathbf{X} \equiv (X_1, \dots, X_n)'$  be a random sample from  $N(\theta, 1)$  where  $\theta \in \Theta = \{\theta_1, \theta_2\}$ . Construct a size  $\alpha$  test for

$$H_0: \theta = \theta_1 \text{ against } H_1: \theta = \theta_2.$$

Using this, construct a  $(1 - \alpha)$  significance level confidence interval for  $\theta$ .

4. Let  $\mathbf{X} \equiv (X_1, \dots, X_n)'$  be a random sample from a Bernoulli distribution with a density function

$$f(x; \theta) = \theta^x(1 - \theta)^{1-x}, \quad x = 0, 1.$$

Construct a size  $\alpha$  test for  $H_0: \theta \leq \theta_0$  against  $H_1: \theta > \theta_0$ . (Hint:  $(\sum_{i=1}^n X_i)$  is binomially distributed.)

5. Let  $\mathbf{X} \equiv (X_1, \dots, X_n)'$  be a random sample from  $N(\mu, \sigma^2)$ 
  - (i) Show that the test defined by the rejection region

$$C_1 = \left\{ \mathbf{x}: \sqrt{n} \frac{(\bar{X}_n - \mu_0)}{s} \geq k \right\}, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

defines a UMP unbiased test for  $H_0: \mu \leq \mu_0$  against  $H_1: \mu > \mu_0$ .

- (ii) Derive a UMP unbiased test for  $H_0: \mu \geq \mu_0$  against  $H_1: \mu < \mu_0$ .
6. Let  $\mathbf{X} \equiv (X_1, \dots, X_n)'$  and  $\mathbf{Y} = (Y_1, \dots, Y_m)'$  be two random samples from  $N(\mu, \sigma_1^2)$  and  $N(\mu, \sigma_2^2)$  respectively. Show that for the hypotheses:

- (i)  $H_0: \sigma_1^2 \leq \sigma_2^2, H_1: \sigma_1^2 > \sigma_2^2;$
- (ii)  $H_0: \sigma_1^2 \geq \sigma_2^2, H_1: \sigma_1^2 < \sigma_2^2;$

- (iii)  $H_0: \sigma_1^2 = \sigma_2^2, H_1: \sigma_1^2 \neq \sigma_2^2,$

the rejection regions are:

$$C_1 = \{\mathbf{x}, \mathbf{y}: \tau(\mathbf{x}, \mathbf{y}) \geq k_1\},$$

$$C_2 = \{\mathbf{x}, \mathbf{y}: \tau(\mathbf{x}, \mathbf{y}) \leq k_2\},$$

$$C_3 = \{\mathbf{x}, \mathbf{y}: k_3 \leq \tau(\mathbf{x}, \mathbf{y}) \leq k_4\},$$

respectively, where

$$\tau(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2 / n - 1}{\sum_{i=1}^m (Y_i - \bar{Y}_m)^2 / m - 1},$$

define UMP unbiased tests (see Lehmann (1959), pp. 169–70). What is the distribution of  $\tau(\mathbf{x}, \mathbf{y})$ ?

- (iv) Construct a size  $\alpha$  test for  $H_0: \sigma_1^2 = \sigma_2^2 = \sigma^2, H_1: \sigma_1^2 \neq \sigma_2^2$  using the likelihood ratio test procedure.

#### Additional references

Bickel and Doksum (1977); Cox and Hinkley (1974); Kendall and Stuart (1973); Lehmann (1959); Rao (1973); Rohatgi (1976); Silvey (1975).

## CHAPTER 15\*

---

### The multivariate normal distribution

---

#### 15.1 Multivariate distributions

The multivariate normal distribution is by far the most important distribution in statistical inference for a variety of reasons including the fact that some of the statistics based on sampling from such a distribution have tractable distributions themselves. It forms the backbone of Part IV on statistical models in econometrics and thus a closer study of this distribution will greatly simplify the discussion that follows. Before we consider the multivariate normal distribution, however, let us introduce some notation and various simple results related to random vectors and their distributions in general.

Let  $\mathbf{X} \equiv (X_1, X_2, \dots, X_n)'$  be an  $n \times 1$  random vector defined on the probability space  $(S, \mathcal{F}, P(\cdot))$ . The *mean* vector  $E(\mathbf{X})$  is defined by

$$E(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{bmatrix} \equiv \boldsymbol{\mu}, \text{ an } n \times 1 \text{ vector} \quad (15.1)$$

and the covariance matrix  $\text{Cov}(\mathbf{X})$  by

$$\begin{aligned} \text{Cov}(\mathbf{X}) &\equiv E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})' \\ &= \begin{bmatrix} \text{Var}(X_1), \text{Cov}(X_1 X_2) & \cdots & \text{Cov}(X_1 X_n) \\ \text{Cov}(X_2 X_1), \text{Var}(X_2) & \cdots & \text{Cov}(X_2 X_n) \\ \vdots & \ddots & \vdots \\ \text{Cov}(X_n X_1), \text{Cov}(X_n X_2) & \cdots & \text{Var}(X_n) \end{bmatrix} \equiv \boldsymbol{\Sigma}, \quad (15.2) \end{aligned}$$

where  $\Sigma$  is an  $n \times n$  symmetric non-negative definite matrix, i.e.  $\Sigma' = \Sigma$  and  $\alpha'\Sigma\alpha \geq 0$  for any  $\alpha \in \mathbb{R}^n$ . The  $ij$ th element of  $\Sigma$  is

$$\sigma_{ij} = E(X_i - \mu_i)(X_j - \mu_j), \quad i, j = 1, 2, \dots, n. \quad (15.3)$$

In relation to  $\alpha'\Sigma\alpha \geq 0$  we can show that if there exists an  $\alpha \in \mathbb{R}^n$ ,  $\alpha \neq \mathbf{0}$  such that  $\text{Var}(\alpha'X) = \alpha'\Sigma\alpha = 0$  then  $\Pr(\alpha'X = c) = 1$  where  $c$  is a constant (only constants have zero variance), i.e. there is a linear relationship holding among the r.v.'s  $X_1, \dots, X_n$  with probability one.

*Lemma 15.1*

If the random vector  $\mathbf{X}$  has a continuous density function then  $\Sigma > 0$ . This is because  $\Pr(\alpha'X = c) = 0$  for all  $\alpha$  and  $c$  in this case.

*Lemma 15.2*

If  $\mathbf{X}$  has mean  $\mu$  and covariance  $\Sigma$  for  $\mathbf{Z} = \mathbf{AX} + \mathbf{b}$

- (i)  $E(\mathbf{Z}) = \mathbf{A}E(\mathbf{X}) + \mathbf{b} = \mathbf{A}\mu + \mathbf{b};$
- (ii)  $\text{Cov}(\mathbf{Z}) = E[(\mathbf{AX} + \mathbf{b} - (\mathbf{A}\mu + \mathbf{b}))(\mathbf{AX} + \mathbf{b} - (\mathbf{A}\mu + \mathbf{b}))']$   
 $= \mathbf{A}E(\mathbf{X} - \mu)(\mathbf{X} - \mu)' \mathbf{A}' = \mathbf{A}\Sigma\mathbf{A}'.$

Let  $\mathbf{X}$  and  $\mathbf{Y}$  be  $n \times 1$  and  $m \times 1$  random vectors with  $E(\mathbf{X}) = \mu_x$ ,  $E(\mathbf{Y}) = \mu_y$ , then

$$\text{Cov}(\mathbf{X}, \mathbf{Y}) = [(\text{Cov}(X_i Y_j)_{ij}] = E[(\mathbf{X} - \mu_x)(\mathbf{Y} - \mu_y)']. \quad (15.4)$$

### Correlation

So far correlation has been defined for random variables (r.v.'s) only and the question arises whether it can be generalised to random vectors. Let  $E(\mathbf{X}) = \mathbf{0}$  (without any loss of generality),  $\text{Cov}(\mathbf{X}) = \Sigma$ ,  $\mathbf{X} n \times 1$ , and partition  $\mathbf{X}$  into

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \quad \mathbf{X}_2: (n-1) \times 1 \quad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \Sigma_{22} \end{pmatrix}. \quad (15.5)$$

Define  $Z = \alpha' \mathbf{X}_2$  and let us consider the correlation between  $X_1$  and  $Z$

$$\text{Corr}(Z, X_1) = \frac{\alpha' \sigma_{12}}{\sigma_{11}^{\frac{1}{2}} (\alpha' \Sigma_{22} \alpha)^{\frac{1}{2}}}. \quad (15.6)$$

This is maximised for the value of  $\alpha$  which minimises

$$E(X_1 - \alpha' \mathbf{X}_2)^2 \quad (15.7)$$

(see Chapter 7), which is  $\alpha = \Sigma_{22}^{-1} \sigma_{21}$  and we define the *multiple correlation*

coefficient to be

$$R = \frac{(\boldsymbol{\sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21})^{\frac{1}{2}}}{\sigma_{11}^{\frac{1}{2}}} \geq \text{Corr}(X_1, \boldsymbol{\alpha}'\mathbf{X}_2), \quad 0 \leq R \leq 1. \quad (15.8)$$

In the case where

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}, \quad \mathbf{X}_1: k \times 1, \quad \mathbf{X}: (n-k) \times 1, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \quad k \geq 1,$$

we could define the r.v.'s  $Z_1 = \boldsymbol{\alpha}'_1 \mathbf{X}_1$  and  $Z_2 = \boldsymbol{\alpha}'_2 \mathbf{X}_2$  whose correlation coefficient is

$$\text{Corr}(Z_1, Z_2) = \frac{\boldsymbol{\alpha}'_1 \boldsymbol{\Sigma}_{12} \boldsymbol{\alpha}_2}{(\boldsymbol{\alpha}'_1 \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha}_1)^{\frac{1}{2}} (\boldsymbol{\alpha}'_2 \boldsymbol{\Sigma}_{22} \boldsymbol{\alpha}_2)^{\frac{1}{2}}}. \quad (15.9)$$

From the above inequality it follows that for  $\boldsymbol{\alpha}_2 = \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} \boldsymbol{\alpha}_1$

$$C_{12} = \frac{(\boldsymbol{\alpha}'_1 \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} \boldsymbol{\alpha}_1)^{\frac{1}{2}}}{(\boldsymbol{\alpha}'_1 \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha}_1)^{\frac{1}{2}}} \geq \text{Corr}(Z_1, Z_2). \quad (15.10)$$

$\boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}$  has at most  $k$  non-zero eigenvalues which measure the association between  $\mathbf{X}_1$  and  $\mathbf{X}_2$  and are called *canonical correlations*.

Let

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_3 \end{pmatrix} \quad \text{where } \mathbf{X}_3: (n-2) \times 1$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \boldsymbol{\Sigma}_{33} \end{pmatrix}.$$

Another form of correlation of interest in this context is the correlation between  $X_1$  and  $X_2$  given that the effect of  $\mathbf{X}_3$  is taken away. For this we form the r.v.'s

$$Y_1 = X_1 - \mathbf{b}'_1 \mathbf{X}_3 \quad \text{and} \quad Y_2 = X_2 - \mathbf{b}'_2 \mathbf{X}_3 \quad \text{and} \quad \text{Corr}(Y_1, Y_2)$$

is maximised by  $\mathbf{b}_1 = \boldsymbol{\Sigma}_{33}^{-1} \boldsymbol{\sigma}_{31}$  and  $\mathbf{b}_2 = \boldsymbol{\Sigma}_{33}^{-1} \boldsymbol{\sigma}_{32}$  as seen above. Hence we define the *partial correlation coefficient* between  $X_1$  and  $X_2$  given  $\mathbf{X}_3$  to be

$$\rho_{12.3} = \frac{\sigma_{12} - \boldsymbol{\sigma}_{13} \boldsymbol{\Sigma}_{33}^{-1} \boldsymbol{\sigma}_{32}}{[\sigma_{11} - \boldsymbol{\sigma}_{13} \boldsymbol{\Sigma}_{33}^{-1} \boldsymbol{\sigma}_{31}]^{\frac{1}{2}} [\sigma_{22} - \boldsymbol{\sigma}_{23} \boldsymbol{\Sigma}_{33}^{-1} \boldsymbol{\sigma}_{32}]^{\frac{1}{2}}} \geq \text{Corr}(Y_1, Y_2). \quad (15.11)$$

## 15.2 The multivariate normal distribution

The univariate normal density function discussed above was of the form

$$f(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}. \quad (15.12)$$

The density function of  $\mathbf{X} \equiv (X_1, X_2, \dots, X_n)'$  when the  $X_i$ 's are IID normally distributed r.v.'s was shown to be of the form

$$\begin{aligned} f(\mathbf{x}; \mu, \sigma^2) &= \prod_{i=1}^n f(x_i; \mu, \sigma^2) \\ &= (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right\}. \end{aligned} \quad (15.13)$$

Similarly, the density function of  $\mathbf{X}$  when the  $X_i$ 's are only independent, i.e.  $X_i \sim N(\mu_i, \sigma_i^2)$ ,  $i = 1, 2, \dots, n$ , takes the form

$$\begin{aligned} f(\mathbf{x}; \mu_1, \dots, \mu_n, \sigma_1^2, \dots, \sigma_n^2) &= \prod_{i=1}^n f(x_i; \mu_i, \sigma_i^2) \\ &= (2\pi)^{-n/2} (\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2} \sum_{i=1}^n \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2\right\}. \end{aligned} \quad (15.14)$$

Comparing the above three density functions we can discern a pattern developing which is very suggestive for the density function of an arbitrary normal vector  $\mathbf{X}$  with  $E(\mathbf{X}) = \boldsymbol{\mu}$  and  $\text{Cov}(\mathbf{X}) = \boldsymbol{\Sigma}$ , which takes the form

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-n/2} (\det \boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right\}, \quad (15.15)$$

and we write  $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . If the  $X_i$ 's are IID r.v.'s  $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_n$  and  $(\det \boldsymbol{\Sigma}) = (\sigma^2)^n$ . On the other hand, if the  $X_i$ 's are independent but not identically distributed

$$\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \dots, \sigma_n^2) \quad \text{and} \quad (\det \boldsymbol{\Sigma}) = \prod_{i=1}^n (\sigma_i^2) = (\sigma_1^2, \dots, \sigma_n^2).$$

In the case of  $n = 2$

$$\begin{aligned} \boldsymbol{\mu} &= \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \quad \text{where } \rho = \frac{\sigma_{12}}{\sigma_1\sigma_2}, \\ \Rightarrow (\det \boldsymbol{\Sigma}) &= \sigma_1^2\sigma_2^2(1 - \rho^2) > 0 \quad \text{for } -1 < \rho < 1 \end{aligned}$$

and

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{(1-\rho^2)} \begin{pmatrix} \frac{1}{\sigma_1^2} & \frac{-\rho}{\sigma_1 \sigma_2} \\ \frac{-\rho}{\sigma_1 \sigma_2} & \frac{1}{\sigma_2^2} \end{pmatrix}. \quad (15.16)$$

Thus the bivariate normal density function is

$$\begin{aligned} f(x_1, x_2; \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{[\sigma_1^2 \sigma_2^2 (1 - \rho^2)]^{-\frac{1}{2}}}{2\pi} \times \exp \left\{ -\frac{(1 - \rho^2)^{-1}}{2} \right. \\ &\quad \times \left. \left[ \left( \frac{x_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \left( \frac{x_1 - \mu_1}{\sigma_1} \right) \left( \frac{x_2 - \mu_2}{\sigma_2} \right) + \left( \frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right] \right\} \end{aligned} \quad (15.17)$$

(see Chapter 6). The standard bivariate density function can be obtained by defining the new r.v.'s

$$Z_i = \left( \frac{X_i - \mu_i}{\sigma_i} \right), \quad i = 1, 2,$$

whose density function is

$$f(z_1, z_2; \rho) = \frac{(1 - \rho^2)^{-\frac{1}{2}}}{2\pi} \exp \left\{ -\frac{(1 - \rho^2)^{-1}}{2} (z_1^2 - 2\rho z_1 z_2 + z_2^2) \right\}. \quad (15.18)$$

### (1) Properties

(N1) Let  $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  then  $\mathbf{Y} = (\mathbf{A}\mathbf{X} + \mathbf{b}) \sim N(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'')$  for  $\mathbf{A}: m \times n$  and  $\mathbf{b}: m \times 1$  constant matrices, e.g. if  $\mathbf{Y} = c\mathbf{X}$ ,  $c \neq 0$ ,  $\mathbf{Y} \sim N(c\boldsymbol{\mu}, c^2\boldsymbol{\Sigma})$ .

This property shows that if  $\mathbf{X}$  is normally distributed then any linear function of  $\mathbf{X}$  is also normally distributed.

(N2) Let  $\mathbf{X}_t \sim N(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ ,  $t = 1, 2, \dots, T$ , be independently distributed random vectors, then for any arbitrary constant matrices  $\mathbf{A}_t$ ,  $t = 1, 2, \dots, T$ ,

$$\left( \sum_{t=1}^T \mathbf{A}_t \mathbf{X}_t \right) \sim N \left( \sum_{t=1}^T \mathbf{A}_t \boldsymbol{\mu}_t, \sum_{t=1}^T (\mathbf{A}_t \boldsymbol{\Sigma}_t \mathbf{A}_t') \right).$$

The converse also holds. If the  $\mathbf{X}_t$ s are IID then  $\boldsymbol{\mu}_t = \boldsymbol{\mu}$ ,  $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}$ ,  $t = 1, 2, \dots, T$ , and

$$\left( \frac{1}{T} \sum_{t=1}^T \mathbf{X}_t \right) \sim N \left( \boldsymbol{\mu}, \frac{1}{T} \boldsymbol{\Sigma} \right).$$

- (N3) Let  $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  then the  $X_i$ s are **independent** if and only if  $\sigma_{ij}=0$ ,  $i \neq j$ ,  $i, j = 1, 2, \dots, n$ , i.e.  $\boldsymbol{\Sigma} = \text{diag}(\sigma_{11}, \dots, \sigma_{nn})$ . In general, zero covariance does not imply independence but in the case of normality the two are equivalent.
- (N4) If  $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  then the **marginal distribution** of any  $k \times 1$  subset  $\mathbf{X}_1$  where

$$\mathbf{X} \equiv \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

$\mathbf{X}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ . This follows from property N1 for  $\mathbf{A} = (\mathbf{I}_k; \mathbf{0})$ ,  $(k \times n)$ ,  $\mathbf{b} = \mathbf{0}$ . Similarly,  $\mathbf{X}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$ .

These can be verified directly using

$$f(\mathbf{x}_1; \boldsymbol{\theta}_1) = \int f(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x}_2 \quad \text{and} \quad f(\mathbf{x}_2; \boldsymbol{\theta}_2) = \int f(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x}_1,$$

although the manipulations involved are rather too cumbersome. Taking  $k=1$ , this property implies that each component of  $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  is also normally distributed; the converse, however, is not true.

- (N5) For the same partition of  $\mathbf{X}$  considered in N4 the **conditional distribution** of  $\mathbf{X}_1$  given  $\mathbf{X}_2$  takes the form

$$(\mathbf{X}_1/\mathbf{X}_2), \quad N(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2), \quad \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

This follows from property N1 for

$$\mathbf{A} = \begin{pmatrix} \mathbf{I}_k & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \mathbf{0} & \mathbf{I}_{n-k} \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \mathbf{0} \quad (15.19)$$

since

$$\mathbf{AX} = \begin{pmatrix} \mathbf{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2 \end{pmatrix}, \quad \text{Cov}(\mathbf{AX}) = \begin{pmatrix} \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & 0 \\ 0 & \boldsymbol{\Sigma}_{22} \end{pmatrix}. \quad (15.20)$$

From this we can deduce that if  $\boldsymbol{\Sigma}_{12} = \mathbf{0}$  then  $\mathbf{X}_1$  and  $\mathbf{X}_2$  are independent since  $(\mathbf{X}_1/\mathbf{X}_2) \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ . Moreover, for any  $\boldsymbol{\Sigma}_{12}$ ,  $(\mathbf{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{X}_2)$  and  $\mathbf{X}_2$  are independent given that their covariance is zero. Similarly,  $(\mathbf{X}_2/\mathbf{X}_1) \sim N(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{X}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})$ . In the case  $n=2$

$$(X_1/X_2) \sim N\left(\mu_1 + \rho \frac{\sigma_1}{\sigma_2} (X_2 - \mu_2), \sigma_1^2(1 - \rho^2)\right). \quad (15.21)$$

These results can be verified using the formula

$$f(\mathbf{x}_1/\mathbf{x}_2; \boldsymbol{\phi}) = \frac{f(\mathbf{x}; \boldsymbol{\theta})}{f(\mathbf{x}_2; \boldsymbol{\theta}_2)}. \quad (15.22)$$

N5 suggests that the *regression function*

$$E(\mathbf{X}_1/\mathbf{X}_2 = \mathbf{x}_2) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \quad \text{is linear in } \mathbf{x}_2 \quad (15.23)$$

and the *skedasticity function*

$$\text{Cov}(\mathbf{X}_1/\mathbf{X}_2 = \mathbf{x}_2) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \quad \text{is free of } \mathbf{x}_2. \quad (15.24)$$

These are very important properties of the multivariate normal distribution and will play a crucial role in Part IV.

## (2)    *Multiple correlation*

Without any loss of generality let  $\mathbf{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ ,  $\mathbf{X}: n \times 1$  and define the partition

$$\mathbf{X} \equiv \begin{pmatrix} X_1 \\ \mathbf{X}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \boldsymbol{\sigma}_{12} \\ \boldsymbol{\sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

where  $\mathbf{X}_2: (n-1) \times 1$  and  $X_1: 1 \times 1$ . The squared multiple correlation coefficient takes the form

$$R^2 = 1 - \frac{\text{Var}(X_1/\mathbf{X}_2)}{\text{Var}(X_1)} = \frac{\boldsymbol{\sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}}{\sigma_{11}}. \quad (15.25)$$

## (3)    *Partial correlation*

Let  $\mathbf{X}_2$  be partitioned further into

$$\mathbf{X}_2 \equiv \begin{pmatrix} X_2 \\ \mathbf{X}_3 \end{pmatrix}, \quad X_2: 1 \times 1, \quad \mathbf{X}_3: (n-2) \times 1$$

with

$$\boldsymbol{\Sigma}_{22} = \begin{pmatrix} \sigma_{22} & \boldsymbol{\sigma}_{23} \\ \boldsymbol{\sigma}_{32} & \boldsymbol{\Sigma}_{33} \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \boldsymbol{\sigma}_{12} & \boldsymbol{\sigma}_{13} \\ \boldsymbol{\sigma}_{21} & \sigma_{22} & \boldsymbol{\sigma}_{23} \\ \boldsymbol{\sigma}_{31} & \boldsymbol{\sigma}_{32} & \boldsymbol{\Sigma}_{33} \end{pmatrix}.$$

The partial correlation between  $X_1$  and  $X_2$  given  $\mathbf{X}_3$  takes the form

$$\begin{aligned} \rho_{12.3} &= \frac{\text{Cov}(X_1, X_2/\mathbf{X}_3)}{[\text{Var}(X_1/\mathbf{X}_3)]^{1/2}[\text{Var}(X_2/\mathbf{X}_3)]^{1/2}} \\ &= \frac{\boldsymbol{\sigma}_{12} - \boldsymbol{\sigma}_{13}\boldsymbol{\Sigma}_{33}^{-1}\boldsymbol{\sigma}_{32}}{[\sigma_{11} - \boldsymbol{\sigma}_{13}\boldsymbol{\Sigma}_{33}^{-1}\boldsymbol{\sigma}_{31}]^{1/2}[\sigma_{22} - \boldsymbol{\sigma}_{23}\boldsymbol{\Sigma}_{33}^{-1}\boldsymbol{\sigma}_{32}]^{1/2}}, \end{aligned} \quad (15.26)$$

with

$$\begin{pmatrix} X_1/X_3 \\ X_2/X_3 \end{pmatrix} \sim N \left( \begin{pmatrix} \sigma_{13}\Sigma_{33}^{-1}X_3 \\ \sigma_{23}\Sigma_{33}^{-1}X_3 \end{pmatrix}, \begin{pmatrix} \sigma_{11}-\sigma_{13}\Sigma_{33}^{-1}\sigma_{31} & \sigma_{12}-\sigma_{13}\Sigma_{33}^{-1}\sigma_{32} \\ \sigma_{21}-\sigma_{23}\Sigma_{33}^{-1}\sigma_{32} & \sigma_{22}-\sigma_{23}\Sigma_{33}^{-1}\sigma_{32} \end{pmatrix} \right). \quad (15.27)$$

### 15.3 Quadratic forms related to the normal distribution

(Q1) Let  $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , where  $\boldsymbol{\Sigma} > \mathbf{0}$ ,  $\mathbf{X}: n \times 1$ , then

- (i)  $(\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2(n) - \text{chi-square};$
- (ii)  $\mathbf{X} \boldsymbol{\Sigma}^{-1} \mathbf{X} \sim \chi^2(n; \delta) - \text{non-central chi-square};$

where  $\delta = \boldsymbol{\mu}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$ . These results depend crucially on  $\boldsymbol{\Sigma}$  being a positive definite matrix because for  $\boldsymbol{\Sigma} > \mathbf{0}$  there exists a non-singular matrix  $\mathbf{H}$ ,  $\boldsymbol{\Sigma} = \mathbf{H}\mathbf{H}' \Rightarrow \mathbf{Z} = \mathbf{H}^{-1}(\mathbf{X} - \boldsymbol{\mu}) \sim N(\mathbf{0}, \mathbf{I}_n)$ , i.e. the  $Z_i$ s are independent and  $(\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) = \sum_{i=1}^n Z_i^2$ . Similarly for (ii). For the MLE of  $\boldsymbol{\mu}$ ,

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= \left( \frac{1}{T} \sum_{t=1}^T \mathbf{X}_t \right) \sim N \left( \boldsymbol{\mu}, \frac{1}{T} \boldsymbol{\Sigma} \right) \quad \text{from } N2 \\ &\Rightarrow T(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \sim \chi^2(n). \end{aligned}$$

(Q2) Let  $\mathbf{X} \sim N(\boldsymbol{\mu}, \mathbf{I}_n)$ , then for  $\mathbf{A}$  a symmetric ( $\mathbf{A}' = \mathbf{A}$ ) matrix

- (i)  $(\mathbf{X} - \boldsymbol{\mu})' \mathbf{A} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2(\text{tr } \mathbf{A});$
- (ii)  $\mathbf{X}' \mathbf{A} \mathbf{X} \sim \chi^2(\text{tr } \mathbf{A}; \delta), \quad \delta = \boldsymbol{\mu}' \mathbf{A} \boldsymbol{\mu}.$

**if and only if**  $\mathbf{A}$  is idempotent (i.e.  $\mathbf{A}^2 = \mathbf{A}$ ). Note  $\text{tr } \mathbf{A}$  refers to the trace of  $\mathbf{A}$  ( $\text{tr } \mathbf{A} = \sum_{i=1}^n a_{ii}$ ).

(Q3) Let  $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ,  $\boldsymbol{\Sigma} > \mathbf{0}$  and  $\mathbf{A}$  is a symmetric matrix, then

- (i)  $(\mathbf{X} - \boldsymbol{\mu})' \mathbf{A} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2(\text{tr } \mathbf{A});$
- (ii)  $\mathbf{X}' \mathbf{A} \mathbf{X} \sim \chi^2(\text{tr } \mathbf{A}; \delta), \quad \delta = \boldsymbol{\mu}' \mathbf{A} \boldsymbol{\mu};$

**if and only if**  $\mathbf{A}\boldsymbol{\Sigma}$  is idempotent (i.e.  $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A} = \mathbf{A}$ ).

(Q4) Let

$$\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} > \mathbf{0}, \quad \text{and} \quad \mathbf{X} \equiv \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}, \quad \boldsymbol{\mu} \equiv \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

for  $\mathbf{X}_1: k \times 1$  the difference

$$[(\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) - (\mathbf{X}_1 - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{X}_1 - \boldsymbol{\mu}_1)] \sim \chi^2(n-k).$$

- (Q5) Let  $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , then for  $\mathbf{A}$  and  $\mathbf{B}$  symmetric and idempotent matrices,  $q_1 = \mathbf{X}'\mathbf{AX}$  and  $q_2 = \mathbf{X}'\mathbf{BX}$  are **independent** if and only if  $\mathbf{A}\boldsymbol{\Sigma}\mathbf{B} = \mathbf{0}$ .
- (Q6) Let  $\mathbf{X} \sim N(\boldsymbol{\mu}, \mathbf{I}_n)$ , then for  $\mathbf{A}$  a symmetric and idempotent matrix and  $\mathbf{B}$  a  $k \times n$  matrix, then  $\mathbf{X}'\mathbf{AX}$  and  $\mathbf{BX}$  are independent if  $\mathbf{BA} = \mathbf{0}$ .
- (Q7) Let  $\mathbf{X} \sim N(\boldsymbol{\mu}, \mathbf{I}_n)$  and  $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I}_m)$  then for  $\mathbf{A}$  and  $\mathbf{B}$  symmetric idempotent matrices

$$\begin{pmatrix} \frac{\mathbf{X}'\mathbf{AX}}{\text{tr } \mathbf{A}} \\ \frac{\mathbf{Z}'\mathbf{BZ}}{\text{tr } \mathbf{B}} \end{pmatrix} \sim F(\text{tr } \mathbf{A}, \text{tr } \mathbf{B}; \delta), \quad \delta = \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}.$$

## 15.4 Estimation

Let  $\mathbf{X} \equiv (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_T)'$  be a random sample from  $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , i.e.  $\mathbf{X}_t \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ,  $t = 1, 2, \dots, T$ .  $\mathbf{X}$  being a  $T \times n$  matrix. The likelihood function takes the form

$$\begin{aligned} L(\boldsymbol{\theta}; \mathbf{X}) &= k(\mathbf{X}) \prod_{t=1}^T [(2\pi)^{-n/2} (\det \boldsymbol{\Sigma})^{-1/2} \exp\{-\frac{1}{2}(\mathbf{X}_t - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}_t - \boldsymbol{\mu})\}] \\ &= k(\mathbf{X})(2\pi)^{-nT/2} (\det \boldsymbol{\Sigma})^{-T/2} \exp\left\{-\frac{1}{2} \sum_{t=1}^T (\mathbf{X}_t - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}_t - \boldsymbol{\mu})\right\}, \end{aligned} \quad (15.28)$$

$$\log L(\boldsymbol{\theta}; \mathbf{X}) = c - \frac{nT}{2} \log 2\pi - \frac{T}{2} \log(\det \boldsymbol{\Sigma}) - \frac{1}{2} \sum_{t=1}^T (\mathbf{X}_t - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}_t - \boldsymbol{\mu}). \quad (15.29)$$

Since

$$\sum_t (\mathbf{X}_t - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}_t - \boldsymbol{\mu}) = \sum_t (\mathbf{X}_t - \bar{\mathbf{X}}_T)\boldsymbol{\Sigma}^{-1}(\mathbf{X}_t - \bar{\mathbf{X}}_T) +$$

$$T(\bar{\mathbf{X}}_T - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}}_T - \boldsymbol{\mu}) = \text{tr } \boldsymbol{\Sigma}^{-1}\mathbf{\Lambda} + T(\bar{\mathbf{X}}_T - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}}_T - \boldsymbol{\mu})$$

for

$$\bar{\mathbf{X}}_T = \frac{1}{T} \sum_t \mathbf{X}_t, \quad \mathbf{\Lambda} = \sum_t (\mathbf{X}_t - \bar{\mathbf{X}}_T)(\mathbf{X}_t - \bar{\mathbf{X}}_T)', \quad (15.30)$$

$\Rightarrow$

$$\begin{aligned} \log L(\boldsymbol{\theta}; \mathbf{X}) &= c^* - \frac{T}{2} \log(\det \boldsymbol{\Sigma}) - \frac{1}{2} \text{tr } \boldsymbol{\Sigma}^{-1}\mathbf{\Lambda} \\ &\quad - \frac{T}{2} (\bar{\mathbf{X}}_T - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}}_T - \boldsymbol{\mu}), \end{aligned} \quad (15.31)$$

$$\frac{\partial \log L(\boldsymbol{\theta}; \mathbf{X})}{\partial \boldsymbol{\mu}} = T \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{X}}_T - \boldsymbol{\mu}) = \mathbf{0} \Rightarrow \hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}_T, \quad (15.32)$$

$$\frac{\partial \log L(\boldsymbol{\theta}; \mathbf{X})}{\partial \boldsymbol{\Sigma}^{-1}} = \frac{T}{2} \boldsymbol{\Sigma} - \frac{1}{2} \boldsymbol{\Lambda} = \mathbf{0} \Rightarrow \hat{\boldsymbol{\Sigma}} = \frac{1}{T} \sum_t (\mathbf{X}_t - \bar{\mathbf{X}}_T)(\mathbf{X}_t - \bar{\mathbf{X}}_T)', \quad (15.33)$$

$$\frac{\partial^2 \log L(\boldsymbol{\theta}; \mathbf{X})}{\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}'} = -T \boldsymbol{\Sigma}^{-1}, \quad \partial \left( \frac{\partial \log L(\boldsymbol{\theta}; \mathbf{X})}{\partial \boldsymbol{\Sigma}^{-1}} \right) = -\frac{T}{2} \boldsymbol{\Sigma} (\partial \boldsymbol{\Sigma}^{-1}) \boldsymbol{\Sigma}. \quad (15.34)$$

Hence,  $\bar{\mathbf{X}}_T$  and  $\hat{\boldsymbol{\Sigma}}$  are the MLE's of  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  respectively.

### (1) Properties

Looking at  $\bar{\mathbf{X}}_T$  and  $\hat{\boldsymbol{\Sigma}}$  we can see that they correspond directly to the MLE's in the univariate case. It turns out that the analogy between the univariate and multivariate cases extends to the properties of  $\bar{\mathbf{X}}_T$  and  $\hat{\boldsymbol{\Sigma}}$ .

In order to discuss the small sample properties of  $\bar{\mathbf{X}}_T$  and  $\hat{\boldsymbol{\Sigma}}$  we need their distributions. Since  $\bar{\mathbf{X}}_T$  is a linear function of normally distributed random vectors it is itself normally distributed

$$\bar{\mathbf{X}}_T \sim N\left(\boldsymbol{\mu}, \frac{1}{T} \boldsymbol{\Sigma}\right). \quad (15.35)$$

The distribution of  $\hat{\boldsymbol{\Sigma}}$  is a direct generalisation of the chi-square distribution, the so-called *Wishart distribution* with  $T-1$  degrees of freedom (see Appendix 24.1), i.e.

$$T \hat{\boldsymbol{\Sigma}} \sim W(\boldsymbol{\Sigma}, T-1) \quad (15.36)$$

From these we can deduce that  $E(\bar{\mathbf{X}}_T) = \boldsymbol{\mu}$  – unbiased estimator of  $\boldsymbol{\mu}$  and  $E(\hat{\boldsymbol{\Sigma}}) = [(T-1)/T] \boldsymbol{\Sigma}$  – biased estimator of  $\boldsymbol{\Sigma}$ .  $\mathbf{S} = [1/(T-1)] \sum_t (\mathbf{X}_t - \bar{\mathbf{X}}_T)(\mathbf{X}_t - \bar{\mathbf{X}}_T)'$  is an unbiased estimator of  $\boldsymbol{\Sigma}$ .

$\bar{\mathbf{X}}_T$  and  $\hat{\boldsymbol{\Sigma}}$  are independent and jointly sufficient for  $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ .

### (2) Useful distributions

Using the distribution  $(T-1)\mathbf{S} \sim W(\boldsymbol{\Sigma}, T-1)$  the following results relating to the sample correlations can be derived (see Muirhead (1982)):

#### (i) Simple correlation

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}, \quad \mathbf{S} = [s_{ij}]_{i,j}, \quad i, j = 1, 2, \dots, n. \quad (15.37)$$

If

$$\sigma_{ij} = 0, \quad (T-2)^{\frac{1}{2}} \frac{r_{ij}}{(1-r_{ij}^2)^{\frac{1}{2}}} \sim t(T-2).$$

For  $\mathbf{M} = [r_{ij}]_{ij}$  when

$$\Sigma = \text{diag}(\sigma_{11}, \dots, \sigma_{nn}), \quad -2 \log(\det(\mathbf{M})) \underset{\chi^2}{\sim} \chi^2(\frac{1}{2}n(n-1)). \quad (15.38)$$

(ii) *Multiple correlation*

$$\hat{R}_T = \left( \frac{\mathbf{s}_{12} \mathbf{S}_{22}^{-1} \mathbf{s}_{21}}{s_{11}} \right)^{\frac{1}{2}}. \quad (15.39)$$

Under

$$R = 0, \quad \left( \frac{T-n}{n-1} \right) \left( \frac{\hat{R}_T^2}{1-\hat{R}_T^2} \right) \sim F(n-1, T-n). \quad (15.40)$$

In particular,

$$E(\hat{R}_T^2) = \frac{n-1}{r-1}, \quad \text{Var}(\hat{R}_T^2) = \frac{2(T-n)(n-1)}{(T^2-1)(T-1)}. \quad (15.41)$$

The distribution of  $\hat{R}^2$  when  $R \neq 0$  is rather complicated and instead we commonly use its asymptotic distribution:

$$\sqrt{(T-1)(\hat{R}_T^2 - R^2)} \underset{\chi^2}{\sim} N(0, 4R^2(1-R^2)^2), \quad 0 < R^2 < 1. \quad (15.42)$$

On the other hand, under

$$R = 0, \quad (T-1)\hat{R}_T^2 \underset{\chi^2}{\sim} \chi^2(n-1). \quad (15.43)$$

A closely related sample equivalent to  $R^2$  is the quantity

$$\tilde{R}_T^2 = \left( \frac{\mathbf{x}'_1 \mathbf{X}_2 (\mathbf{X}'_2 \mathbf{X}_2)^{-1} \mathbf{X}'_2 \mathbf{x}_1}{(\mathbf{x}'_1 \mathbf{x}_1)} \right)^{\frac{1}{2}}, \quad (15.44)$$

$\mathbf{x}_1: T \times 1, \mathbf{X}_2: T \times k$ .

The sampling distribution of  $\tilde{R}^2$  was derived by Fisher (1928) but it is far too complicated to be of direct interest. Its mean, however, and variance are of some interest

$$E(\tilde{R}^2) = R^2 + \frac{k}{T-1} (1-R^2) + \left( \frac{2}{T+1} \right) R^2 (1-R^2) + O(T^{-2}) \quad (15.45)$$

$$\text{Var}(\tilde{R}^2) = \frac{4R^2(1-R^2)^2}{T-1} + O(T^{-2}) \quad (15.46)$$

(see Muirhead (1982)). On the  $O(\cdot)$  notation (see Chapter 10). Hence, the mean of  $\tilde{R}^2$  increases as  $k$  increases and for  $R^2=0$

$$E(\tilde{R}^2) = \frac{k}{T-1} + O(T^{-2}). \quad (15.47)$$

(iii) *Partial correlation*

$$\rho_{12.3} = \frac{s_{12} - s_{13}S_{33}^{-1}s_{32}}{(s_{11} - s_{13}S_{33}^{-1}s_{31})^{1/2}(s_{22} - s_{23}S_{33}^{-1}s_{32})^{1/2}}. \quad (15.48)$$

Under

$$\rho_{12.3} = 0, \quad (T-n-4)^{1/2} \frac{\hat{\rho}_{12.3}}{(1-\hat{\rho}_{12.3}^2)^{1/2}} \sim t(T-n-4). \quad (15.49)$$

## 15.5 Hypothesis testing and confidence regions

Hypothesis testing in the context of the multivariate normal distribution will form the backbone of testing in Part IV where the normal distribution plays a very important role.

For expositional purposes let us consider an example of testing and confidence estimation in the context of the statistical model of Section 15.4.

Consider the null hypothesis  $H_0: \mu = \mathbf{0}$  against  $H_1: \mu \neq \mathbf{0}$  when  $\Sigma$  is unknown. Using the likelihood ratio test procedure with

$$\max_{\theta \in \Theta} L(\theta; \mathbf{x}) = c^*(\det \hat{\Sigma})^{-T/2} (\det T\hat{\Sigma})^{-\frac{1}{2}(T-n-1)} \exp\left\{-\frac{1}{2}Tn\right\}, \quad (15.50)$$

$$\max_{\theta \in \Theta_0} L(\theta; \mathbf{x}) = c^*(\det(\hat{\Sigma} + \bar{\mathbf{X}}_T \bar{\mathbf{X}}_T')^{-T/2} (\det T\hat{\Sigma})^{\frac{1}{2}(T-n-1)} \exp\left\{-\frac{1}{2}Tn\right\}), \quad (15.51)$$

we get

$$\lambda(\mathbf{x}) = \left( \frac{\det(T\Sigma)}{\det(\hat{\Sigma} + \bar{\mathbf{X}}_T \bar{\mathbf{X}}_T')} \right)^{T/2} = \left( \frac{1}{1 + \bar{\mathbf{X}}_T' \Sigma^{-1} \bar{\mathbf{X}}_T} \right)^{T/2} = \left( \frac{1}{1 + H^2/T - 1} \right), \quad (15.52)$$

where

$$H^2 = T\bar{\mathbf{X}}_T' \mathbf{S}^{-1} \bar{\mathbf{X}}_T \quad \left( \mathbf{S} = \frac{T}{T-1} \hat{\Sigma} \right) \quad (15.53)$$

is the so-called Hotelling's statistic which can form the basis of the test, being a monotone function of  $\lambda(\mathbf{X})$ . Indeed,

$$\left( \frac{T-n}{n(T-1)} \right) H^2 \sim F(n, T-n; \delta), \quad \delta = T\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \quad (15.54)$$

(see Muirhead (1982)). Using the Hotelling statistic we can define a test based on the *rejection region*

$$C_1 = \left\{ \mathbf{x}: \left( \frac{T-n}{n(T-1)} \right) H^2 \geq c_\alpha \right\}, \quad (15.55)$$

where  $\alpha = \int_{c_\alpha}^\infty dF(n, T-n)$ . For  $H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0$  against  $H_1: \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$  the test statistic takes the form

$$H^2 = T(\bar{\mathbf{X}}_T - \boldsymbol{\mu}_0)' \mathbf{S}^{-1} (\bar{\mathbf{X}}_T - \boldsymbol{\mu}_0). \quad (15.56)$$

Using the acceptance region for the latter

$$C_0 = \left\{ \mathbf{x}: \left( \frac{T-n}{n(T-1)} \right) H^2 \leq c_\alpha \right\}, \quad (15.57)$$

we can define a  $(1-\alpha)$  confidence region for  $\boldsymbol{\mu}$  of the form

$$C(\mathbf{X}) = \left\{ \boldsymbol{\mu}: T(\bar{\mathbf{X}}_T - \boldsymbol{\mu})' \mathbf{S}^{-1} (\bar{\mathbf{X}}_T - \boldsymbol{\mu}) \leq \frac{(T-1)n}{(T-n)} c_\alpha \right\}. \quad (15.58)$$

### **Important concepts**

Mean and covariance of a random vector, multiple correlation, canonical correlation, partial correlation, the multivariate normal density function, marginal and conditional normal density functions, Wishart distribution.

### **Questions**

1. Explain the various correlation measures in the context of random vectors and compare the general formulae with the ones associated with the multivariate normal distribution. Comment on the similarities.
2. Discuss the relationship between normality and linearity.
3. Discuss the marginal and conditional distributions of a subvector  $\mathbf{X}_1$  of a normally distributed random vector  $\mathbf{X} = (\mathbf{X}'_1, \mathbf{X}'_2)'$ .
4. Under what circumstances is the quadratic form  $(\mathbf{X} - \boldsymbol{\mu})' \mathbf{A} (\mathbf{X} - \boldsymbol{\mu})$  chi-square distributed?

5. State the conditions under which the quadratic forms  $\mathbf{X}'\mathbf{A}\mathbf{X}$  and  $\mathbf{X}'\mathbf{B}\mathbf{X}$  will be independent.
6. State the conditions under which the ratio of the quadratic forms  $\mathbf{X}'\mathbf{A}\mathbf{X}$  and  $\mathbf{Z}'\mathbf{B}\mathbf{Z}$  will be  $F$ -distributed.
7. Under which circumstances will the quadratic form  $\mathbf{X}'\mathbf{A}\mathbf{X}$  and  $\mathbf{B}\mathbf{X}$  be independent?
8. Discuss the properties of the MLE's  $\bar{\mathbf{X}}_n$  and  $\hat{\Sigma}$  of  $\mu$  and  $\Sigma$  respectively in the context of the statistical model defined by the random sample  $\mathbf{X} \equiv (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n)'$  from  $N(\mu, \Sigma)$ .

**Additional references**

Anderson (1984); Mardia *et al.* (1979); Morrison (1976); Seber (1984).

## CHAPTER 16\*

---

### Asymptotic test procedures

---

As discussed in Chapter 14, the main problem in hypothesis testing is to construct a test statistic  $\tau(\mathbf{X})$  whose distribution we know under both the null hypothesis  $H_0$  and the alternative  $H_1$  and it does not depend on the unknown parameter(s)  $\theta$ . The first part of the problem, that of constructing  $\tau(\mathbf{X})$ , can be handled relatively easily using the various methods discussed above (Neyman–Pearson lemma, likelihood ratio) when certain conditions are satisfied. The second part of the problem, that of determining the distribution of  $\tau(\mathbf{X})$  under both  $H_0$  and  $H_1$ , is much more difficult to ‘solve’ and often we have to resort to asymptotic theory. This amounts to deriving the asymptotic distribution of  $\tau(\mathbf{X})$  and using that to determine the rejection region  $C_1$  (or  $C_0$ ) and the associated probabilities. For a given sample size  $n$ , however, these will be as accurate as the asymptotic distribution of  $\tau_n(\mathbf{X})$  is an accurate approximation of its finite sample distribution. Moreover, since the distribution of  $\tau_n(\mathbf{X})$  for a given  $n$  is not known (otherwise we use that) we do not know how ‘good’ the approximation is. This suggests that when asymptotic results are used we should be aware of their limitations and the inaccuracies they can lead to (see Chapter 10).

#### 16.1 Asymptotic properties

Consider the test defined by the rejection region

$$C_1^n = \{\mathbf{x}: |\tau_n(\mathbf{X})| \geq c_n\}, \quad (16.1)$$

and whose power function is

$$\mathcal{P}_n(\theta) = Pr(\mathbf{x} \in C_1^n), \quad \theta \in \Theta. \quad (16.2)$$

Since the distribution of  $\tau_n(\mathbf{X})$  is not known we *cannot* determine  $c_n$  or

$\mathcal{P}(\theta)$ . If the asymptotic distribution of  $\tau_n(\mathbf{X})$  is available, however, we can use that instead to define  $c_n$  from some fixed  $\alpha$  and the *asymptotic power function*

$$\pi(\theta) = Pr(\mathbf{x} \in C_1^\infty), \quad \theta \in \Theta. \quad (16.3)$$

In this sense we can think of  $\{\tau_n(\mathbf{X}), n \geq 1\}$  as a sequence of test statistics defining a sequence of rejection regions  $\{C_1^n, n \geq 1\}$  with power functions  $\{\mathcal{P}_n(\theta), n \geq 1, \theta \in \Theta\}$  and we can choose  $c_n$  accordingly to ensure that the sequence of tests have the same size  $\alpha$  if

$$\max_{\theta \in \Theta_0} \mathcal{P}_n(\theta) = \alpha \quad \text{for all } n \geq 1. \quad (16.4)$$

Note that  $\lim_{n \rightarrow \infty} \mathcal{P}_n(\theta) = \pi(\theta)$ . In this context the various criteria for tests discussed above must be reformulated in terms of the asymptotic power function  $\pi(\theta)$ ; see Bickel and Doksum (1977).

### Definition 1

The sequence of tests for  $H_0: \theta \in \Theta_0$  against  $H_1: \theta \in \Theta_1$  defined by  $\{C_1^n, n \geq 1\}$  is said to be **consistent** of size  $\alpha$  if

$$\max_{\theta \in \Theta_0} \pi(\theta) = \alpha \quad (16.5)$$

and

$$\pi(\theta) = 1, \quad \theta \in \Theta_1. \quad (16.6)$$

As in the case of estimation, consistency is a reasonable property but only a minimal property. In order to be able to make comparisons between various tests we need better approximations to the power than 1. With this in mind we define asymptotic unbiasedness.

### Definition 2

A sequence of tests as defined above is said to be **asymptotically unbiased** of size  $\alpha$  if

$$\max_{\theta \in \Theta_0} \pi(\theta) = \alpha \quad (16.7)$$

and

$$\alpha < \pi(\theta) < 1, \quad \theta \in \Theta_1. \quad (16.8)$$

### Definition 3

A sequence of tests as defined above is said to be **uniformly most**

**power (UMP) of size  $\alpha$**  if

$$\max_{\theta \in \Theta_0} \pi(\theta) = \alpha \quad (16.9)$$

and

$$\pi(\theta) \geq \pi^*(\theta), \quad \theta \in \Theta_1, \quad (16.10)$$

for any size  $\alpha$  test with asymptotic power function  $\pi^*(\theta)$ .

In asymptotic tests we are often interested in *local alternatives* of the form

$$H_1: \theta_1 = \theta_0 + \frac{\mathbf{b}}{\sqrt{n}}, \quad \mathbf{b} \neq 0 \quad (16.11)$$

in order to assess the power of the test around the null. When

$$\mathbf{I}_n(\theta) = \mathbf{O}_p(n) \quad \text{then } \sqrt{n}(\hat{\theta} - \theta_0) \sim N(\mathbf{b}, \mathbf{I}_{\infty}(\theta)^{-1}) \quad (16.12)$$

for  $\hat{\theta}$  the MLE of  $\theta$ . In this case we consider only local power and a test with greatest local power is called *locally uniformly most powerful*.

## 16.2 The likelihood ratio and related test procedures

In this section three general test procedures which give rise to asymptotically optimal tests will be considered; the *likelihood ratio*, *Wald* and *Lagrange multiplier* test procedures. All three test procedures can be interpreted as utilising the information incorporated in the log likelihood function in different but asymptotically equivalent ways.

For expositional purposes the test procedures will be considered in the context of the simplest statistical model where

- (i)  $\Phi = \{f(x; \theta), \theta \in \Theta\}$  is the probability model; and
- (ii)  $\mathbf{X} \equiv (X_1, X_2, \dots, X_n)'$  is a random sample.

The results can be easily generalised to the non-random sample case where  $\mathbf{I}_n(\theta) = \mathbf{O}_p(n)$  as explained in Chapter 13 above in the context of maximum likelihood estimation. For most results the generalisation amounts to substituting  $\mathbf{I}(\theta)$  for  $\mathbf{I}_{\infty}(\theta)$  and reinterpreting the results.

### (1) Simple null hypothesis

Let the null hypothesis be  $H_0: \theta = \theta_0, \theta \in \Theta \equiv \mathbb{R}^m$  against  $H_1: \theta \neq \theta_0$ .

- (i) *The likelihood ratio test*

The likelihood ratio test statistic discussed in Section 14.4 takes the form

$$\lambda(\mathbf{x}) = \frac{L(\theta_0; \mathbf{x})}{\max_{\theta \in \Theta} L(\theta; \mathbf{x})} = \frac{L(\theta_0; \mathbf{x})}{L(\hat{\theta}; \mathbf{x})} \quad (16.13)$$

where  $\hat{\theta}$  is the MLE of  $\theta$ . In cases where  $\lambda(\mathbf{x})$  or some monotonic function of it have a tractable distribution there is no need for asymptotic theory. Commonly, however, this is not the case and asymptotic theory is called for.

(ii) *The Wald test*

Under certain regularity conditions which include CR1–CR3 (see Chapter 13)  $\log L(\theta; \mathbf{x})$  can be expanded in a Taylor series at  $\theta = \hat{\theta}$

$$\begin{aligned} \log L(\theta; \mathbf{x}) &= \log L(\hat{\theta}; \mathbf{x}) + (\hat{\theta} - \theta) \left[ \frac{\partial}{\partial \theta} \log L(\theta; \mathbf{x}) \right]_{\theta=\hat{\theta}} \\ &\quad + \frac{1}{2} (\hat{\theta} - \theta)' \left[ \frac{\partial^2 \log L(\theta^*; \mathbf{x})}{\partial \theta \partial \theta'} \right]_{\theta=\hat{\theta}} (\hat{\theta} - \theta) + o_p(1), \end{aligned} \quad (16.14)$$

where  $|\theta^* - \theta| < |\hat{\theta} - \theta|$  and  $o_p(1)$  refers to asymptotically negligible terms (see Chapter 10). Since

$$\left. \frac{\partial}{\partial \theta} \log L(\theta; \mathbf{x}) \right|_{\theta=\hat{\theta}} = 0, \quad (16.15)$$

being the first order conditions for the MLE, and

$$\left( \frac{1}{n} \frac{\partial^2}{\partial \theta \partial \theta} \log L(\hat{\theta}; \mathbf{x}) \right) \xrightarrow{P} \mathbf{I}(\theta), \quad (16.16)$$

the above expansion can be simplified (see Serfling (1980)) to:

$$\log L(\theta; \mathbf{x}) = \log L(\hat{\theta}; \mathbf{x}) + \frac{1}{2} n(\hat{\theta} - \theta)' \mathbf{I}(\theta)(\hat{\theta} - \theta) + o_p(1). \quad (16.17)$$

This implies that, since

$$-2 \log \lambda(\mathbf{x}) = 2[\log L(\hat{\theta}; \mathbf{x}) - \log L(\theta_0; \mathbf{x})], \quad (16.18)$$

$$-2 \log \lambda(\mathbf{x}) = n(\hat{\theta} - \theta_0)' \mathbf{I}(\theta)(\hat{\theta} - \theta_0) + o_p(1). \quad (16.19)$$

For the asymptotic properties of MLE's it is known that under certain regularity conditions

$$\sqrt{n}(\hat{\theta} - \theta_0) \underset{\mathbf{x}}{\sim} N(0, \mathbf{I}(\theta)^{-1}). \quad (16.20)$$

Using this we can deduce (see property Q1, Chapter 15) that

$$LR = -2 \log \lambda(\mathbf{x}) \underset{\mathbf{x}}{\sim} n(\hat{\theta} - \theta_0)' \mathbf{I}(\theta)(\hat{\theta} - \theta_0) \underset{\mathbf{x}}{\sim} \chi^2(m), \quad (16.21)$$

being a quadratic form in asymptotically normal random variables (r.v.'s).

Wald (1943), using the above approximation of  $-2 \log \lambda(\mathbf{x})$ , proposed an alternative test statistic by replacing  $\mathbf{I}(\boldsymbol{\theta})$  with  $\mathbf{I}(\hat{\boldsymbol{\theta}})$ :

$$W = n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathbf{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{H_0}{\underset{\alpha}{\sim}} \chi^2(m), \quad (16.22)$$

given that  $\mathbf{I}(\hat{\boldsymbol{\theta}}) \xrightarrow{P} \mathbf{I}(\boldsymbol{\theta})$ . This is the so-called *Wald statistic*.

(iii) *The Lagrange multiplier test*

Rao (1947) using the asymptotic distribution of the score function (instead of that of  $\hat{\boldsymbol{\theta}}$ ), i.e.

$$\mathbf{q}(\boldsymbol{\theta}) = \frac{1}{\sqrt{n}} \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}; \mathbf{x}) \underset{\alpha}{\sim} N(\mathbf{0}, \mathbf{I}(\boldsymbol{\theta})) \quad (16.23)$$

proposed the *efficient score* (or Lagrange multiplier) test statistic

$$LM = \frac{1}{n} \mathbf{q}(\boldsymbol{\theta}_0)' \mathbf{I}(\boldsymbol{\theta}_0)^{-1} \mathbf{q}(\boldsymbol{\theta}_0) \stackrel{H_0}{\underset{\alpha}{\sim}} \chi^2(m), \quad (16.24)$$

which is again a quadratic form in asymptotically normally distributed r.v.'s.

For all three test statistics ( $LR, W, LM$ ) the rejection region takes the form

$$C_1 = \{ \mathbf{x}: l(\mathbf{x}) \geq c_\alpha \}, \quad (16.25)$$

where  $l(\mathbf{x})$  stands for all three test statistics and the critical value  $c_\alpha$  is defined by  $\int_{c_\alpha}^\infty d\chi^2(m) = \alpha$ ,  $\alpha$  being the size of the test. Under local alternatives with a Pittman type drift of the form:

$$H_1: \boldsymbol{\theta}_n = \boldsymbol{\theta}_0 + \frac{\mathbf{b}}{\sqrt{n}}, \quad (16.26)$$

all three test statistics are asymptotically distributed as:

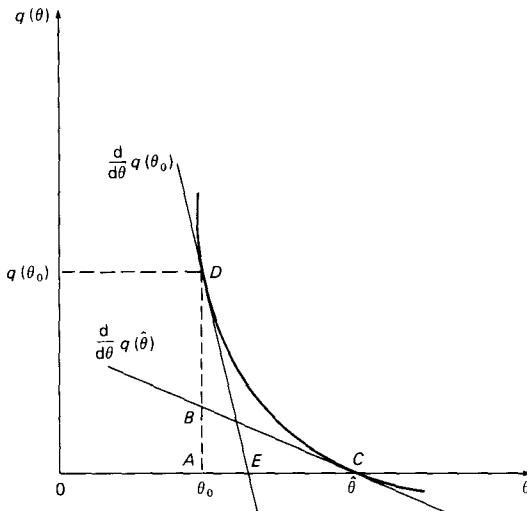
$$l(\mathbf{x}) \stackrel{H_1}{\underset{\alpha}{\sim}} \chi^2(m; \delta), \quad \delta = \mathbf{b}' \mathbf{I}(\boldsymbol{\theta}_0) \mathbf{b}, \quad (16.27)$$

since

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_n) + \mathbf{b} \underset{\alpha}{\sim} N(\mathbf{b}, \mathbf{I}(\boldsymbol{\theta}_0)^{-1}) \quad (16.28)$$

and

$$\sqrt{n} \mathbf{q}(\boldsymbol{\theta}_0) = \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathbf{I}(\boldsymbol{\theta}_0) + op(1) \underset{\alpha}{\sim} N(\mathbf{b}' \mathbf{I}(\boldsymbol{\theta}_0), \mathbf{I}(\boldsymbol{\theta}_0)). \quad (16.29)$$

Fig. 16.1. The  $LR$ ,  $W$  and  $LM$  tests compared.

Hence, the power function for all three test statistics takes the form

$$\mathcal{P}(\theta) = \int_{c_\alpha}^{\infty} d\chi^2(m; \delta), \quad (16.30)$$

and thus,  $LR$ ,  $W$  and  $LM$  are asymptotically equivalent in the sense that they have the same asymptotic properties.

Fig. 16.1, due to Pagan (1982), shows the relationship between  $LR$ ,  $W$  and  $LM$  in the case of a scalar  $\theta$ .

$$LM = 2 \left( \text{area } \underset{A}{\overset{D}{\triangle}} \underset{E}{\triangle} \right) = q(\theta_0)^2 \frac{\partial}{\partial \theta} q(\theta_0), \quad (16.31)$$

$$W = 2 \left( \text{area } \underset{A}{\overset{B}{\triangle}} \underset{C}{\triangle} \right) = (\hat{\theta} - \theta_0)^2 \frac{\partial}{\partial \theta} q(\hat{\theta}), \quad (16.32)$$

$$LR = 2 \left( \text{area } \underset{A}{\overset{D}{\triangle}} \underset{C}{\triangle} \right) = 2 \int_{\theta_0}^{\hat{\theta}} q(\theta) d\theta. \quad (16.33)$$

Note that all three test statistics can be interpreted as functions of the score function.

## (2) Composite null hypothesis

Consider the case where the  $H_0$  is composite, i.e.

$$H_0: \theta \in \Theta_0 \text{ against } H_1: \theta \in \Theta_1, \quad \Theta_0 \subseteq \mathbb{R}^r, \quad \Theta \subseteq \mathbb{R}^m.$$

It is both convenient as well as practical to parametrise  $\Theta_0$  in the form

$$\Theta_0 = \{\theta: \mathbf{R}(\theta) = \mathbf{0}, \theta \in \Theta\} \quad (16.34)$$

where  $\mathbf{R}(\theta) = \mathbf{0}$  represents  $r$  non-linear equations, i.e.  $\mathbf{R}(\theta) = (R_1(\theta), R_2(\theta), \dots, R_r(\theta))'$ . In most situations in practice the parametrised form arises naturally in the form of restrictions such as  $R_1(\theta) = \theta_1\theta_3 + \theta_2$ ,  $R_2(\theta) = \log \theta_1 - \theta_2$ ,  $R_3(\theta) = \theta_1^2 + \theta_2 - 1$ ,  $R_4(\theta) = \theta_1 - 2\theta_2$ , etc. If we define  $\hat{\theta}$  to be the maximum likelihood estimator (MLE) of  $\theta$ , i.e.  $\hat{\theta}$  is the solution of  $[\partial \log L(\theta; \mathbf{x})]/\partial \theta = \mathbf{0}$ , then from

$$\sqrt{n}(\hat{\theta} - \theta) \underset{\mathbf{x}}{\sim} N(0, \mathbf{I}(\theta)^{-1}), \quad (16.35)$$

and

$$\frac{1}{\sqrt{n}} \frac{\partial}{\partial \theta} \log L(\theta; \mathbf{x}) \underset{\mathbf{x}}{\sim} N(\mathbf{0}, \mathbf{I}(\theta)), \quad (16.36)$$

we can deduce that

$$\sqrt{n}(\mathbf{R}(\hat{\theta}) - \mathbf{R}(\theta)) \underset{\mathbf{x}}{\stackrel{H_0}{\sim}} N(\mathbf{0}, \mathbf{R}'_\theta \mathbf{I}(\theta)^{-1} \mathbf{R}_\theta), \quad (16.37)$$

since  $\mathbf{R}(\theta)$  can be approximated at  $\theta = \hat{\theta}$  by

$$\mathbf{R}(\theta) = \mathbf{R}(\hat{\theta}) + \mathbf{R}_\theta(\theta - \hat{\theta}) + o_p(1), \quad (16.38)$$

where

$$\mathbf{R}_\theta = \frac{\partial \mathbf{R}(\theta)}{\partial \theta}.$$

## (i) The Wald test procedure

If the null hypothesis  $H_0$  is true we expect the MLE  $\hat{\theta}$ , without imposing the restrictions, to be close to satisfying the restrictions, i.e. if  $H_0$  is true,  $\mathbf{R}(\hat{\theta}) \approx \mathbf{0}$ . This implies that a natural measure for any departure from  $H_0$  should be

$$\|\mathbf{R}(\hat{\theta}) - \mathbf{0}\|. \quad (16.39)$$

If this is ‘significantly’ different from zero it will be a good indication that  $H_0$  is false. The problem is to formalise the concept ‘significantly different from zero’. The obvious way to proceed is to construct a pivot based on  $\|\mathbf{R}(\hat{\theta})\|$  in order to enable us to turn this statement into a precise probabilistic statement.

In constructing such a pivot there are two basic problems to overcome. The first is that  $\|\mathbf{R}(\hat{\theta})\|$  depends on the units of measurement and the second is that absolute values are not easy to manipulate. A quantity which 'solves' both problems is the quadratic form

$$\mathbf{R}(\hat{\theta})'[\mathbf{V}(\mathbf{R}(\hat{\theta}))]^{-1}\mathbf{R}(\hat{\theta}), \quad (16.40)$$

where  $\mathbf{V}(\mathbf{R}(\hat{\theta}))$  represents the covariance of  $\mathbf{R}(\hat{\theta})$ . Determining  $\mathbf{V}(\mathbf{R}(\hat{\theta}))$  can be a very difficult task since we often know very little about the distribution of  $\hat{\theta}$ . Asymptotically, however, we know the distribution of  $\mathbf{R}(\hat{\theta})$  and

$$\mathbf{V}(\mathbf{R}(\hat{\theta})) = \mathbf{R}'_{\theta} \mathbf{I}(\theta)^{-1} \mathbf{R}_{\theta}, \quad (16.41)$$

hence we can deduce that

$$n \mathbf{R}(\hat{\theta})' [\mathbf{R}'_{\theta} \mathbf{I}(\theta)^{-1} \mathbf{R}_{\theta}]^{-1} \mathbf{R}(\hat{\theta}) \xrightarrow[n]{H_0} \chi^2(r). \quad (16.42)$$

Wald's suggestion amounts to replacing  $\mathbf{V}(\mathbf{R}(\hat{\theta}))$  with a consistent estimator, i.e.

$$W = n \mathbf{R}(\hat{\theta})' [\mathbf{R}'_{\theta} \mathbf{I}(\hat{\theta})^{-1} \mathbf{R}_{\hat{\theta}}]^{-1} \mathbf{R}(\hat{\theta}) \xrightarrow[n]{H_0} \chi^2(r). \quad (16.43)$$

Note that the Wald procedure can be used in conjunction with any asymptotically normal estimator  $\theta^*$  (not just MLE's) since if

$$\sqrt{n}(\theta^* - \theta) \sim N(\mathbf{0}, \Sigma_{\theta}), \quad (16.44)$$

$$W = n \mathbf{R}(\theta^*)' [\mathbf{R}'_{\theta} \hat{\Sigma}_{\theta} \mathbf{R}_{\hat{\theta}}]^{-1} \mathbf{R}(\theta^*) \xrightarrow[n]{H_0} \chi^2(r). \quad (16.45)$$

## (ii) Lagrange multiplier test procedure

In contrast to the Wald test procedure the Lagrange multiplier procedure is based solely on the restricted MLE of  $\theta$ , say  $\tilde{\theta}$ . Although the Lagrange multiplier test statistic can take various equivalent forms we consider only two such forms in what follows. Estimation of  $\theta$  subject to the restrictions  $\mathbf{R}(\theta) = \mathbf{0}$  is based on the optimisation of the Lagrangian function

$$\mathcal{L} = \log L(\theta; \mathbf{x}) - \mathbf{R}(\theta)\boldsymbol{\mu}, \quad (16.46)$$

where  $\boldsymbol{\mu}: r \times 1$  vector of multipliers. The restricted MLE of  $\theta$  is defined to be the solution of the system of equations:

$$\frac{\partial}{\partial \theta} \log L(\tilde{\theta}; \mathbf{x}) - \mathbf{R}'_{\tilde{\theta}} \boldsymbol{\mu} = \mathbf{0}, \quad (16.47)$$

$$\mathbf{R}(\tilde{\theta}) = \mathbf{0}. \quad (16.48)$$

In the case of the Wald procedure we began our search for an asymptotic pivot using  $\mathbf{R}(\hat{\theta})$  which should be close to zero when  $H_0$  is true. In the present case, however,  $\mathbf{R}(\tilde{\theta}) = \mathbf{0}$  by definition and thus it cannot be used. But, although in the Wald procedure the score function evaluated at  $\theta = \hat{\theta}$  is zero, i.e.

$$\frac{\partial \log L(\hat{\theta}; \mathbf{x})}{\partial \theta} = \mathbf{0}, \quad (16.49)$$

this is not the case for  $[\partial \log L(\tilde{\theta}; \mathbf{x})]/\partial \theta$  and we can use it to construct an asymptotic pivot. Equivalently, the Lagrange multipliers  $\mu(\tilde{\theta})$  can be used instead. The intuition underlying the use of  $\mu(\tilde{\theta})$  is that these multipliers can be interpreted as shadow prices for the constraints and should register all departures from  $H_0$ ; if  $\tilde{\theta}$  is closed to  $\theta$   $\mu(\tilde{\theta})$  is small and vice versa. Hence, a reasonable thing to do is to consider the quantity  $|\mu(\tilde{\theta}) - \mathbf{0}|$ . Using the same argument as in the Wald procedure for  $|\mathbf{R}(\hat{\theta}) - \mathbf{0}|$  we set up the quadratic form

$$\mu(\tilde{\theta})' [\mathbf{V}(\mu(\tilde{\theta}))]^{-1} \mu(\tilde{\theta}). \quad (16.50)$$

Using the fact that

$$\frac{1}{n} \frac{\partial \log L(\tilde{\theta}; \mathbf{x})}{\partial \theta} \underset{\alpha}{\sim} N(0, \tilde{\mathbf{I}}(\theta)), \quad (16.51)$$

we can deduce that

$$\frac{1}{\sqrt{n}} (\mu(\tilde{\theta}) - \mu(\theta)) \underset{\alpha}{\sim} N(\mathbf{0}, [\mathbf{R}'_{\theta} \tilde{\mathbf{I}}(\theta)^{-1} \mathbf{R}_{\theta}]^{-1}). \quad (16.52)$$

Hence,

$$\frac{1}{n} \mu(\tilde{\theta})' [\mathbf{R}'_{\theta} \tilde{\mathbf{I}}(\theta)^{-1} \mathbf{R}_{\theta}] \mu(\tilde{\theta}) \underset{\alpha}{\sim} \chi^2(r). \quad (16.53)$$

The *Lagrange multiplier* test statistic takes the form

$$LM = \frac{1}{n} \mu(\tilde{\theta})' [\mathbf{R}'_{\theta} \tilde{\mathbf{I}}(\theta)^{-1} \mathbf{R}_{\theta}] \mu(\tilde{\theta}) \underset{\alpha}{\sim} \chi^2(r), \quad (16.54)$$

or, equivalently,

$$LM = \frac{1}{n} \left( \frac{\partial}{\partial \theta} \log L(\tilde{\theta}; \mathbf{x}) \right)' \tilde{\mathbf{I}}(\theta)^{-1} \left( \frac{\partial}{\partial \theta} \log L(\tilde{\theta}; \mathbf{x}) \right), \quad (16.55)$$

which is the *efficient score* form.

The likelihood ratio test statistic takes the form

$$LR = 2(\log L(\hat{\theta}; \mathbf{x}) - \log L(\tilde{\theta}; \mathbf{x})) \stackrel{H_0}{\underset{\alpha}{\sim}} \chi^2(r). \quad (16.56)$$

Using the Taylor series expansions we can show that

$$LR \simeq W \simeq LM \simeq n(\hat{\theta} - \tilde{\theta})' \mathbf{I}(\theta)(\hat{\theta} - \tilde{\theta}). \quad (16.57)$$

Thus, although all three test statistics are based on three different asymptotic pivots, as  $n \rightarrow \infty$  the test statistics become equivalent. All three tests share the same asymptotic properties; they are all *consistent* as well as asymptotically *locally UMP* against local alternatives of the form considered above. In the absence of any information relating to higher-order approximations of the distribution of these test statistics under both  $H_0$  and  $H_1$  the choice between them is based on computational convenience. The Wald test statistic is constructed in terms of  $\hat{\theta}$  the unrestricted MLE of  $\theta$ , the Lagrange multiplier in terms of  $\tilde{\theta}$  the restricted MLE of  $\theta$  and the likelihood ratio in terms of both.

In order to be able to discriminate between the above three tests we need to derive higher-order approximations such as Edgeworth approximations (see Chapter 10). Rothenberg (1984) gives an excellent discussion of various ways to derive such higher-order approximations.

Of particular interest in practice is the case where  $\theta \equiv (\theta_1, \theta_2)$  and  $H_0: \theta_1 = \theta_1^0$  against  $H_1: \theta_1 \neq \theta_1^0, \theta_1: r \times 1$  with  $\theta_2: (m-r) \times 1$  left unrestricted. In this case the three test statistics take the form

$$LR = -2(\log L(\tilde{\theta}; \mathbf{x}) - \log L(\hat{\theta}; \mathbf{x})), \quad (16.58)$$

$$W = n(\hat{\theta}_1 - \theta_1^0)' [\mathbf{I}_{11}(\hat{\theta}) - \mathbf{I}_{12}(\hat{\theta}) \mathbf{I}_{22}^{-1}(\hat{\theta}) \mathbf{I}_{21}(\hat{\theta})](\hat{\theta}_1 - \theta_1^0), \quad (16.59)$$

$$LM = \frac{1}{n} \mu(\tilde{\theta})' [\mathbf{I}_{11}(\tilde{\theta}) - \mathbf{I}_{12}(\tilde{\theta}) \mathbf{I}_{22}^{-1}(\tilde{\theta}) \mathbf{I}_{21}(\tilde{\theta})]^{-1} \mu(\tilde{\theta}), \quad (16.60)$$

where

$$\hat{\theta} \equiv (\hat{\theta}_1, \hat{\theta}_2), \quad \tilde{\theta} = (\theta_1^0, \tilde{\theta}_2), \quad \mu(\tilde{\theta}) = \left. \frac{\partial \log L(\theta; \mathbf{x})}{\partial \theta_1} \right|_{\theta=\tilde{\theta}}.$$

This is because for  $\mathbf{R}(\theta) = \theta_1 - \theta_1^0$

$$\mathbf{I}(\theta) = \begin{pmatrix} \mathbf{I}_{11}(\theta) & \mathbf{I}_{12}(\theta) \\ \mathbf{I}_{21}(\theta) & \mathbf{I}_{22}(\theta) \end{pmatrix}, \quad \mathbf{R}'_\theta = (\mathbf{I}_r, 0)$$

and hence

$$\mathbf{R}'_\theta \mathbf{I}(\theta)^{-1} \mathbf{R}_\theta = [\mathbf{I}_{11}(\theta) - \mathbf{I}_{12}(\theta) \mathbf{I}_{22}^{-1}(\theta) \mathbf{I}_{21}(\theta)]^{-1}. \quad (16.61)$$

For further discussion of the above asymptotic test procedures see the survey by Engle (1984).

***Important concepts***

Asymptotic power function, consistent test, asymptotically unbiased test, asymptotically uniformly most powerful test, locally uniformly most powerful test, Wald test statistic, Lagrange multiplier (efficient score) test statistic.

***Questions***

1. Why do we need asymptotic theory in hypothesis testing?
2. Explain the concept of an asymptotic power function and use it to define consistency, asymptotic unbiasedness and UMP in testing.
3. What do we mean by a size  $\alpha$  test in this context?
4. Compare the  $LR$ ,  $W$  and  $LM$  tests in the case of a simple null hypothesis (draw diagrams if it helps).
5. Explain the common-sense logic underlying the  $LR$ ,  $W$  and  $LM$  test procedures in the case of a composite null hypothesis.
6. Discuss the similarities and differences between the  $LR$ ,  $W$  and  $LM$  test procedures.
7. Explain the circumstances under which you would use these asymptotic test procedures in preference to the test procedures discussed in Chapter 14.
8. Explain the derivation of the  $W$  and  $LM$  test statistics in the case of  $H_0: \theta_1 = \theta_1^0$  against  $H_1: \theta_1 \neq \theta_1^0$ ,  $\theta_1$  being a subset of parameter vector  $\theta = (\theta_1, \theta_2)$ , considered above.
9. Verify the form of the Wald and Lagrange multiplier test statistics for  $H_0: \theta_1 = \theta_1^0$  against  $H_1: \theta_1 \neq \theta_1^0, \theta \equiv (\theta_1, \theta_2)$  using the partitioned matrix inversion rule

$$\begin{aligned} & \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1} & (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1} \end{pmatrix}. \end{aligned}$$

***Additional references***

Aitchison and Silvey (1958); Buse (1982); Engle (1984); Moran (1970); Rao (1973); Silvey (1959).

## **PART IV**

---

### **The linear regression and related statistical models**

---

## CHAPTER 17

---

### Statistical models in econometrics

---

#### 17.1 Simple statistical models

The main purpose of Parts II and III has been to formulate and discuss the concept of a statistical model which will form the backbone of the discussion in Part IV. A statistical model has been defined as made up of two related components:

- (i) a probability model,  $\Phi = \{D(y; \theta), \theta \in \Theta\}$ † specifying a parametric family of densities indexed by  $\theta$ ; and
- (ii) a sampling model,  $\mathbf{y} = (y_1, y_2, \dots, y_T)'$  defining a sample from  $D(y; \theta_0)$ , for some ‘true’  $\theta_0$  in  $\Theta$ .

The probability model provides the framework in the context of which the stochastic environment of the real phenomenon being studied can be defined and the sampling model describes the relationship between the probability model and the observable data. By postulating a statistical model we transform the uncertainty relating to the mechanism giving rise to the observed data to uncertainty relating to some unknown parameter(s)  $\theta$  whose estimation determines the stochastic mechanism  $D(y; \theta)$ .

An example of such a statistical model in econometrics is provided by the modelling of the distribution of personal income. In studying the distribution of personal income higher than a lower limit  $y_0$  the following statistical model is often postulated:

- (i) 
$$\Phi = \left\{ D(y/y_0; \theta) = \left( \frac{\theta}{y_0} \right) \left( \frac{y_0}{y} \right)^{\theta+1}, \quad \theta \in \mathbb{R}_+, \quad y \geq y_0 \right\};$$
- (ii)  $\mathbf{y} \equiv (y_1, y_2, \dots, y_T)'$  is a random sample from  $D(y/y_0; \theta)$ .

† The notation in Part IV will be somewhat different from the one used in Parts II and III. This change in notation has been made to conform with the established econometric notation.

Note:

$$E(y) = y_0 \left( \frac{\theta}{\theta-1} \right), \quad \text{if } \theta > 1,$$

$$\text{Var}(y) = y_0^2 \left( \frac{\theta}{(\theta-1)^2(\theta-2)} \right), \quad \text{if } \theta > 2.$$

For  $\mathbf{y}$  a random sample the likelihood function is

$$L(\theta; \mathbf{y}) = \prod_{t=1}^T \left( \frac{\theta}{y_0} \right) \left( \frac{y_0}{y_t} \right)^{\theta+1} = \theta^T y_0^{T\theta} (y_1, y_2, \dots, y_T)^{-(\theta+1)},$$

$$\log L(\theta; \mathbf{y}) = T \log \theta + T \theta \log y_0 - (\theta+1) \sum_{t=1}^T \log y_t,$$

$$\frac{d \log L}{d \theta} = \frac{T}{\theta} + T \log y_0 - \sum_t \log y_t = 0,$$

$$\Rightarrow \hat{\theta} = T \left[ \sum_t \log \left( \frac{y_t}{y_0} \right) \right]^{-1}$$

is the maximum likelihood estimator (MLE) of the parameter  $\theta$ . Since  $(d^2 \log L)/d\theta^2 = -T/\theta^2$ , the asymptotic distribution of  $\hat{\theta}$  takes the form (see Chapter 13):

$$\sqrt{T}(\hat{\theta} - \theta) \sim N(0, \theta^2).$$

Although in general the finite sample distribution is not frequently available, in this particular case we can derive  $D(\hat{\theta})$  analytically. It takes the form

$$D(\hat{\theta}) = \frac{\theta^{T-1} T^{T-1}}{\Gamma(T-1) y^T} \exp \left( -\frac{T\theta}{y} \right), \quad y \geq 0, \quad \text{i.e. } \left( \frac{2T\theta}{\hat{\theta}} \right) \sim \chi^2(2T)$$

(see Appendix 6.1). This distribution of  $\hat{\theta}$  can be used to consider the finite sample properties of  $\hat{\theta}$  as well as test hypotheses or set up confidence intervals for the unknown parameter  $\theta$ . For instance, in view of the fact that

$$E(\hat{\theta}) = \left( \frac{T}{T-2} \right) \theta$$

we can deduce that  $\hat{\theta}$  is a biased estimator of  $\theta$ .

It is of interest in this particular case to assess the ‘accuracy’ of the asymptotic distribution of  $\hat{\theta}$  for a small  $T$ , ( $T=8$ ), by noting that

$$\text{Var}(\hat{\theta}) = \frac{T^2 \theta^2}{(T-2)^2(T-3)}$$

(see Johnson and Kotz (1970)). Using the data on income distribution (see Chapter 2), for  $y \geq 5000$  (reproduced below) to estimate  $\theta$ ,

Income lower limit	5000	6000	7000	8000	10 000	12 000	15 000	20 000
No. of incomes	2600	1890	1150	990	410	220	100	50

we get

$$\hat{\theta} = T \left[ \sum_t \log \left( \frac{y_t}{y_0} \right) \right]^{-1} = 1.6$$

as the ML estimate.

Using the invariance property of MLE's (see Section 13.3) we can deduce that

$$\hat{E}(\hat{\theta}) = 2.13, \quad \hat{\text{Var}}(\hat{\theta}) = 0.91.$$

As we can see, for a small sample ( $T=8$ ) the estimate of the mean and the variance are considerably larger than the ones given by the asymptotic distribution:

$$\hat{E}(\hat{\theta}) = 1.6, \quad \hat{\text{Var}}(\hat{\theta}) = \frac{\hat{\theta}^2}{T} = 0.32.$$

On the other hand, for a much larger sample, say  $T=100$ ,

$$\hat{E}(\hat{\theta}) = 1.63, \quad \hat{\text{Var}}(\hat{\theta}) = 0.028,$$

as compared with

$$\hat{E}(\hat{\theta}) = 1.6, \quad \hat{\text{Var}}(\hat{\theta}) = 0.026.$$

These results exemplify the danger of using asymptotic results for small samples and should be viewed as a warning against uncritical use of asymptotic theory. For a more general discussion of asymptotic theory and how to improve upon the asymptotic results see Chapter 10.

The statistical inference results derived above in relation to the income distribution example depend crucially on the appropriateness of the statistical model postulated. That is, the statistical model should represent a good approximation of the real phenomenon to be explained in a way which takes account the nature of the available data. For example, if the data were collected using stratified sampling then the random sample assumption is inappropriate (see Section 17.2 below). When any of the

assumptions underlying the statistical model are invalid the above estimation results are unwarranted.

In the next three sections it is argued that for the purposes of econometric modelling we need to extend the simple statistical model based on a random sample, illustrated above, in certain specific directions as required by the particular features of econometric modelling. In Section 17.2 we consider the nature of economic data commonly available and discuss its implications for the form of the sampling model. It is argued that for most forms of economic data the random sample assumption is inappropriate. Section 17.3 considers the question of constructing probability models if the identically distributed assumption does not hold. The concept of a statistical generating mechanism (GM) is introduced in Section 17.4 in order to supplement the probability and sampling models. This additional component enables us to accommodate certain specific features of econometric modelling. In Section 17.5 the main statistical models of interest in econometrics are summarised as a prelude to the discussion which follows.

## **17.2    Economic data and the sampling model**

Economic data are usually non-experimental in nature and come in one of three forms:

- (i)      *time series*, measuring a particular variable at successive points in time (annual, quarterly, monthly or weekly);
- (ii)     *cross-section*, measuring a particular variable at a given point in time over different units (persons, households, firms, industries, countries, etc.);
- (iii)    *panel data*, which refer to cross-section data over time.

Economic data such as M1 money stock ( $M$ ), real consumers' expenditure ( $Y$ ) and its implicit deflator ( $P$ ), interest rate on 7 days' deposit account ( $I$ ), over time, are examples of time-series data (see Appendix, Table 17.2). The income data used in Chapter 2 are cross-section data on 23 000 households in the UK for 1979–80. Using the same 23 000 households of the cross-section observed over time we could generate panel data on income. In practice, panel data are rather rare in econometrics because of the difficulties involved in gathering such data. For a thorough discussion of econometric modelling using panel data see Chamberlain (1984).

The econometric modeller is rarely involved directly with the data collection and refinement and often has to use published data knowing very little about their origins. This lack of knowledge can have serious repercussions on the modelling process and lead to misleading conclusions. Ignorance related to how the data were collected can lead to an erroneous

choice of an appropriate sampling model. Moreover, if the choice of the data is based only on the name they carry and not on intimate knowledge about what exactly they are measuring, it can lead to an inappropriate choice of the statistical GM (see Section 17.4, below) and some misleading conclusions about the relationship between the estimated econometric model and the theoretical model as suggested by economic theory (see Chapter 1). Let us consider the relationship between the nature of the data and the sampling model in some more detail.

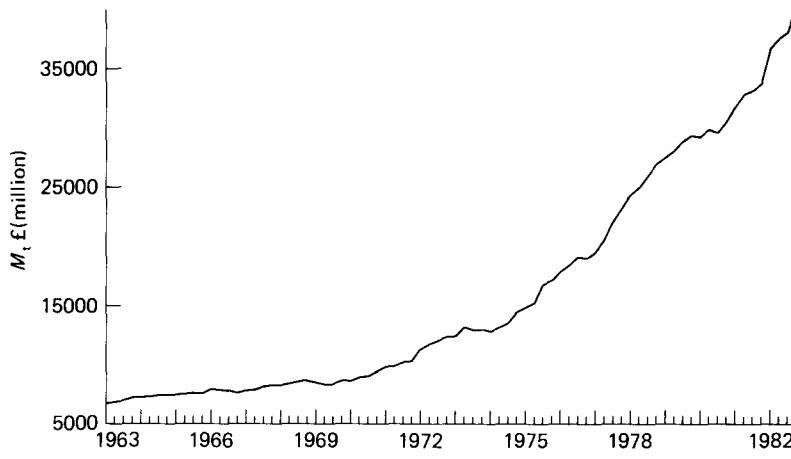
In Chapter 11 we discussed three basic forms of a sampling model:

- (i) *random sample* – a set of independent and identically distributed (IID) random variables (r.v.'s);
- (ii) *independent sample* – a set of independent but not identically distributed r.v.'s; and
- (iii) *non-random sample* – a set of non-IID r.v.'s.

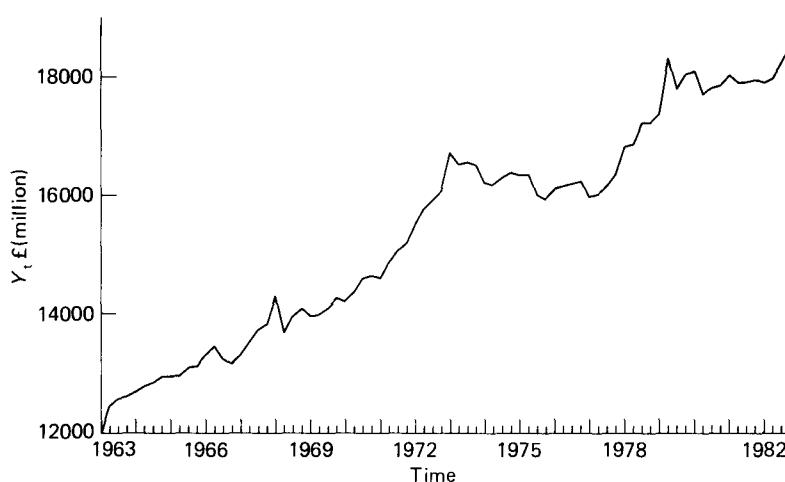
For cross-section data selected by the *simple random sampling* method (where every unit in the target population has the same probability of being selected), the sampling model of a random sample seems the most appropriate choice. On the other hand, for cross-section data selected by the stratified sampling method (the target population divided into a number of groups (strata) with every unit in each group having the same probability of being selected), the identically distributed assumption seems rather inappropriate. The fact that the groups are chosen *a priori* in some systematic way renders the identically distributed assumption inappropriate. For such cross-section data the sampling model of an independent sample seems more appropriate. The independence assumption can be justified if sampling within and between groups is random.

For time-series data the sampling models of a random or an independent sample seem rather unrealistic on a priori grounds, leaving the non-random sample as the most likely sampling model to postulate at the outset. For the time-series data plotted against time in Fig. 17.1(a)–(d) the assumption that they represent realisations of stochastic processes (see Chapter 8) seems more realistic than their being realisations of IID r.v.'s. The plotted series exhibit considerable time dependence. This is confirmed in Chapter 23 where these series are used to estimate a money adjustment equation. In Chapters 19–22 the sampling model of an independent sample is intentionally maintained for the example which involves these data series and several misleading conclusions are noted throughout.

In order to be able to take explicitly into consideration the nature of the observed data chosen in the context of econometric modelling, the statistical models of particular interest in econometrics will be specified in terms of the observable r.v.'s giving rise to the data rather than the error term, the usual



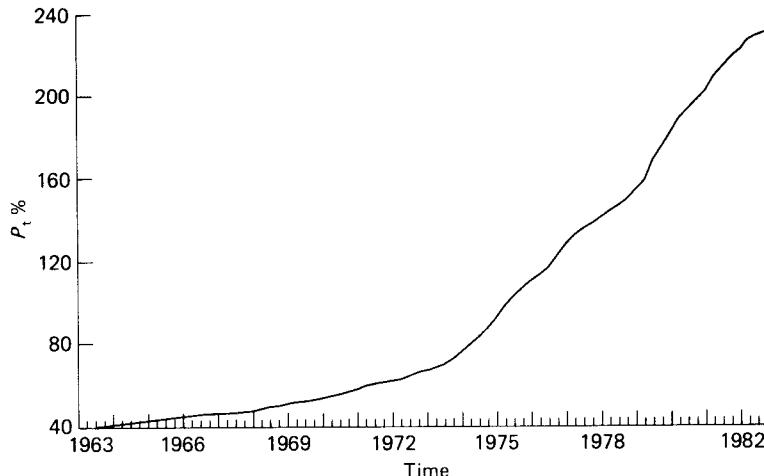
(a)



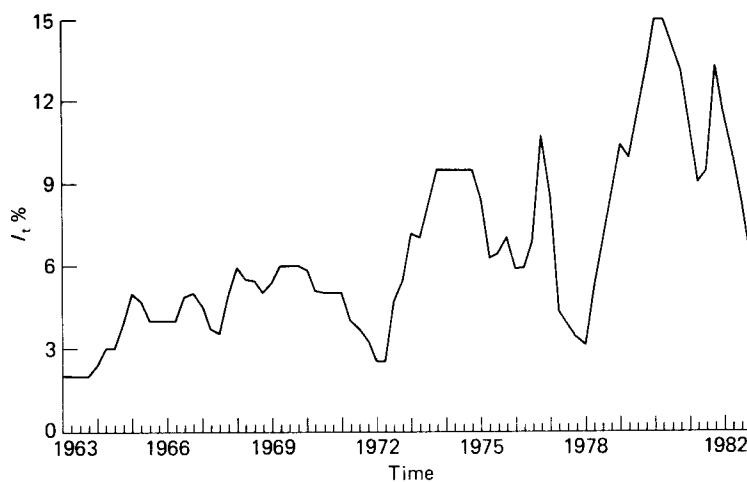
(b)

Fig. 17.1(a). Money stock £(million). (b) Real consumers' expenditure.

approach in econometrics textbooks (see Theil (1971), Maddala (1977), Judge *et al.* (1982) *inter alia*). The approach adopted in the present book is to extend the statistical models considered so far in Part III in order to accommodate certain specific features of econometric modelling. In particular a third component, called a *statistical generating mechanism* (*GM*) will be added to the probability and sampling models in order to enable us to summarise the information involved in a way which provides



(c)



(d)

Fig. 17.1(c). Implicit price deflator. (d) Interest rate on 7 days' deposit account.

'an adequate' approximation to the actual DGP giving rise to the observed data (see Chapter 1). This additional component will be considered extensively in Section 17.4 below. In the next section the nature of the probability models required in econometric modelling will be discussed in view of the above discussion of the sampling model.

### 17.3 Economic data and the probability model

In Chapter 1 it was argued that the specification of statistical models should take account not only of the theoretical a priori information available but the nature of the observed data chosen as well. This is because the specification of statistical models proposed in the present book is based on the observable random variable giving rise to the observed data and not by attaching a white-noise error term to the theoretical model. This strategy implies that the modeller should consider assumptions such as independence, stationarity, mixing (see Chapter 8) in relation to the observed data at the outset.

As argued in Section 17.2, the sampling model of a random sample seems rather unrealistic for most situations in econometric modelling in view of the economic data usually available. Because of the interrelationship between the sampling and the probability model we need to extend the simple probability model  $\Phi = \{D(y; \theta), \theta \in \Theta\}$  associated with a random sample to ones related to independent and non-random samples.

An *independent* (but non-identically distributed) sample  $y \equiv (y_1, \dots, y_T)'$  raises questions of *time-heterogeneity* in the context of the corresponding probability model. This is because in general every element  $y_t$  of  $y$  has its own distribution with different parameters  $D(y_t; \theta_t)$ . The parameters  $\theta_t$  which depend on  $t$  are called *incidental parameters*. A probability model related to  $y$  takes the general form

$$\Phi = \{D(y_t; \theta_t), \theta_t \in \Theta, t \in \mathbb{T}\}, \quad (17.1)$$

where  $\mathbb{T} = \{1, 2, \dots\}$  is an index set.

A *non-random sample*  $y$  raises questions not only of *time-heterogeneity* but of *time-dependence* as well. In this case we need the joint distribution of  $y$  in order to define an appropriate probability model of the general form

$$\Phi = \{D(y_1, y_2, \dots, y_T; \theta_T), \theta_T \in \Theta, \mathbb{T}_1 = \{1, 2, \dots, T\} \subseteq \mathbb{T}\}. \quad (17.2)$$

In both of the above cases the observed data can be viewed as realisations of the stochastic process  $\{y_t, t \in \mathbb{T}\}$  and for modelling purposes we need to restrict its generality using assumptions such as normality, stationarity and asymptotic independence or/and supplement the sample and theoretical information available. In order to illustrate these let us consider the simplest case of an independent sample and one incidental parameter:

$$(i) \quad \Phi = \left\{ D(y_t; \theta_t) = \frac{1}{\sigma\sqrt{(2\pi)}} \exp \left\{ -\frac{1}{2} \left( \frac{y_t - \mu_t}{\sigma} \right)^2 \right\}, \right. \\ \left. \theta_t \equiv (\mu_t, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+, t \in \mathbb{T} \right\};$$

- (ii)  $\mathbf{y} \equiv (y_1, y_2, \dots, y_T)'$  is an independent sample from  $D(y_t; \boldsymbol{\theta}_t)$ ,  $t = 1, 2, \dots, T$ , respectively.

The probability model postulates a normal density with mean  $\mu_t$  (an *incidental parameter*) and variance  $\sigma^2$ . The sampling model allows each  $y_t$  to have a different mean but the same variance and to be independent of the other  $y_s$ s. The distribution of the sample for the above statistical model  $D(\mathbf{y}; \boldsymbol{\theta})$  where  $\mathbf{y} \equiv (y_1, y_2, \dots, y_T)'$  and  $\boldsymbol{\theta} \equiv (\mu_1, \mu_2, \dots, \mu_T, \sigma^2)$  is

$$\begin{aligned} D(\mathbf{y}, \boldsymbol{\theta}) &= \prod_{t=1}^T D(y_t; \mu_t, \sigma^2) \\ &= (\sigma^2)^{-T/2} (2\pi)^{-T/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{t=1}^T (y_t - \mu_t)^2 \right\}. \end{aligned} \quad (17.3)$$

As we can see, there are  $T + 1$  unknown parameters,  $\boldsymbol{\theta} = (\sigma^2, \mu_1, \mu_2, \dots, \mu_T)$ , to be estimated and only  $T$  observations which provide us with sufficient warning that there will be problems. This is indeed confirmed by the maximum likelihood (ML) method. The log likelihood is

$$\log L(\boldsymbol{\theta}; \mathbf{y}) = \text{const} - \frac{T}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=1}^T (y_t - \mu_t)^2, \quad (17.4)$$

$$\frac{\partial \log L}{\partial \mu_t} = -\frac{1}{2\sigma^2} (-2)(y_t - \mu_t) = 0, \quad t = 1, 2, \dots, T, \quad (17.5)$$

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_t (y_t - \mu_t)^2 = 0. \quad (17.6)$$

These first-order conditions imply that  $\hat{\mu}_t = y_t$ ,  $t = 1, 2, \dots, T$ , and  $\hat{\sigma}^2 = 0$ . Before we rush into pronouncing these as MLE's it is important to look at the second-order conditions for a maximum.

$$\frac{\partial^2 \log L}{\partial \mu_t^2} \Bigg|_{\substack{\hat{\mu}_t \\ \hat{\sigma}^2}} = -\frac{1}{\sigma^2} \Bigg|_{\sigma^2 = \hat{\sigma}^2}, \quad \frac{\partial^2 \log L}{\partial \sigma^4} \Bigg|_{\substack{\hat{\mu}_t \\ \hat{\sigma}^2}} = \frac{T}{2\sigma^4} - \frac{1}{\sigma^6} \sum_t (y_t - \mu_t)^2 \Bigg|_{\substack{\mu_t = \hat{\mu}_t \\ \sigma^2 = \hat{\sigma}^2}},$$

which are unbounded and hence  $\hat{\mu}_t$  and  $\hat{\sigma}^2$  are not MLE's; see Section 13.3. This suggests that there is not enough information in the statistical model (i)-(ii) above to estimate the statistical parameters  $\boldsymbol{\theta} = (\mu_1, \mu_2, \dots, \mu_T, \sigma^2)$ .

An obvious way to supplement this information is in the form of *panel data* for  $y_t$ , say  $y_{it}$ ,  $i = 1, 2, \dots, N$ ,  $t = 1, 2, \dots, T$ . In the case where  $N$  realisations of  $y_t$  are available at each  $t$ ,  $\boldsymbol{\theta}$  could be estimated by

$$\tilde{\mu}_t = \frac{1}{N} \sum_{i=1}^N y_{it}, \quad t = 1, 2, \dots, T \quad (17.7)$$

and

$$\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N (y_{it} - \hat{\mu}_t)^2. \quad (17.8)$$

It can be verified that these are indeed the MLE's of  $\theta$ .

An alternative way to supplement the information of the statistical model (i)–(ii) is to reduce the dimensionality of the statistical parameter space  $\Theta$ . This can be achieved by imposing restrictions on  $\theta$  or modelling  $\theta$  by relating it to other observable random variables (r.v.'s) via conditioning (see Chapter 7). Note that non-stochastic variables are viewed as degenerate r.v.'s. The latter procedure enables us to accommodate theoretical information within a probability model by relating such information to the statistical parameters  $\theta$ . In particular, such information is related to the mean (marginal or conditional) of r.v.'s involved and sometimes to the variance. Theoretical information is rarely related to higher-order moments (see Chapter 4).

The modelling of statistical parameters via conditioning leads naturally to an additional component to supplement the probability and sampling models. This additional component we call a *statistical generating mechanism* (GM) for reasons which will become apparent in the discussion which follows. At this stage it suffices to say that the statistical GM is postulated as a crude approximation to the actual DGP which gave rise to the observed data in question, taking account of the nature of such data as well as theoretical a priori information.

In the case of the statistical model (i)–(ii) above we could 'solve' the inadequate information problem by relating  $\mu_t$  to a vector of observable variables  $x_{1t}, x_{2t}, \dots, x_{kt}$ ,  $t = 1, 2, \dots, T$ , say, linearly, to postulate

$$\mu_t = \mathbf{b}' \mathbf{x}_t, \quad (17.9)$$

where  $\mathbf{b} \equiv (b_1, b_2, \dots, b_k)', k < T$ , is a vector of unknown parameters. By postulating this relationship we reduce the parameter space from  $\Theta \equiv \mathbb{R}^T \times \mathbb{R}_+$  and increasing with  $T$  to  $\Theta_0 \equiv \mathbb{R}^k \times \mathbb{R}_+$  and independent of  $T$ .

The statistical GM in this case takes the general form

$$y_t = \mathbf{b}' \mathbf{x}_t + u_t, \quad t \in \mathbb{T}, \quad (17.10)$$

where  $\mu_t = \mathbf{b}' \mathbf{x}_t$  and  $u_t = y_t - \mathbf{b}' \mathbf{x}_t$  are called systematic and non-systematic components of  $y_t$ , respectively. By construction

$$E(\mu_t u_t) = 0 \quad \text{and} \quad E(u_t) = 0, \quad E(u_t^2) = \sigma^2, \quad E(u_t u_s) = 0, \\ t \neq s, \quad t, s \in \mathbb{T},$$

where  $E(\cdot)$  is defined relative to  $D(y_t; \theta)$ , the marginal distribution of  $y_t$ . Equation (10) represents a situation where the choice of the values  $x_{1t}, x_{2t}, \dots$

$\dots, x_{kt}$  determines the systematic part of  $y_t$  with the unmodelled part  $u_t$  being a white-noise process (see Chapter 8). This is the statistical GM of the Gauss linear model (see Chapter 18). The above statistical GM will be extended in the next section in order to define some of the most widely used statistical models in econometrics.

## 17.4 The statistical generating mechanism

The concept of a statistical GM is postulated to supplement the probability and sampling models and represents a crude approximation to the actual DGP which gave rise to the available data. It represents a summarisation of the sample information in a way which enables us to accommodate any a priori information related to the actual DGP as suggested by economic theory (see Chapter 1).

Let  $\{y_t, t \in \mathbb{T}\}$  be a stochastic process defined on  $(S, \mathcal{F}, P(\cdot))$  (see Chapter 8). The statistical GM is defined by

$$y_t = \mu_t + u_t, \quad t \in \mathbb{T}, \tag{17.11}$$

where

$$\mu_t = E(y_t | \mathcal{D}), \quad \mathcal{D} \subseteq \mathcal{F}, \tag{17.12}$$

$\mathcal{D}$  being some  $\sigma$ -field. This defines the statistical process generating  $y_t$  with  $\mu_t$  being the postulated *systematic mechanism* giving rise to the observed data on  $y_t$  and  $u_t$  the *non-systematic part* of  $y_t$  defined by  $u_t = y_t - \mu_t$ . Defining  $u_t$  this way ensures that it is orthogonal to the systematic component  $\mu_t$ ; denoted by  $\mu_t \perp u_t$  (see Chapter 7). The orthogonality condition is needed for the logical consistency of the statistical GM in view of the fact that  $u_t$  represents the part of  $y_t$  left unexplained by the choice of  $\mu_t$ . The terms systematic, non-systematic and orthogonality are formalised in terms of the underlying probability and sampling models defining the statistical model.

It must be emphasised at the outset that the terms systematic and non-systematic are relative to the information set as defined by the underlying probability and sampling models as well as to any a priori information related to the statistical parameters of interest  $\theta$ . This information is incorporated in the definition of the systematic component and the remaining part of  $y_t$  we call non-systematic or error. Hence, the nature of  $u_t$  depends crucially on how  $\mu_t$  is defined and incorporates the *unmodelled* part of  $y_t$ . This definition of the error term differs significantly from the usual use of the term in econometrics as either errors-in-equation or errors of measurement. The use of the concept in the present book comes much closer to the term ‘noise’ used in engineering and control literatures (see

Kálman (1982)). Our aim in postulating a statistical GM is to minimise the non-systematic component  $u_t$  by making the most of the systematic information in defining the systematic component  $\mu_t$ . For more discussion on the error term see Hendry (1983).

Let  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  be a  $k \times 1$  vector stochastic process defined on  $(S, \mathcal{F}, P(\cdot))$  which represents the observable random variables involved. Let  $y_t$  be the random variable whose behaviour is of interest, where

$$\mathbf{Z}_t \equiv \begin{pmatrix} y_t \\ \mathbf{X}_t \end{pmatrix}. \text{ For a conditioning information set } \mathcal{D}_t$$

the systematic component of  $y_t$  can be defined by

$$\mu_t = E(y_t / \mathcal{D}_t), \quad t \in \mathbb{T}, \quad (17.13)$$

where  $\mathcal{D}_t$  is some sub- $\sigma$ -field of  $\mathcal{F}$ . The non-systematic component  $u_t$  represents the unmodelled part of  $y_t$  given  $\mu_t$ , i.e.

$$u_t = y_t - E(y_t / \mathcal{D}_t), \quad t \in \mathbb{T}. \quad (17.14)$$

These two components give rise to the general statistical GM

$$y_t = E(y_t / \mathcal{D}_t) + u_t, \quad t \in \mathbb{T}, \quad (17.15)$$

where by construction,

$$(i) \quad E(u_t / \mathcal{D}_t) = E[(y_t - E(y_t / \mathcal{D}_t)) / \mathcal{D}_t] = 0; \quad (17.16)$$

$$(ii) \quad E(\mu_t u_t / \mathcal{D}_t) = \mu_t E(u_t / \mathcal{D}_t) = 0; \quad (17.17)$$

using the properties of conditional expectation (see Chapter 7). It is important to note at this stage that the expectation operator  $E(\cdot)$  in (16) and (17) is defined relative to the probability distribution of the underlying probability model. By changing  $\mathcal{D}_t$  (and the related probability model) we can define some of the most important statistical models of interest in econometrics. Let us consider some of these special cases.

- (a) Assuming that  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is a normal IID stochastic process and choosing  $\mathcal{D}_t = \{\mathbf{X}_t = \mathbf{x}_t\}$ , a degenerate  $\sigma$ -field, (15) takes the special form

$$y_t = \boldsymbol{\beta}' \mathbf{x}_t + u_t, \quad t \in \mathbb{T}, \quad (17.18)$$

where the underlying probability model is based on  $D(y_t / \mathbf{X}_t; \boldsymbol{\theta})$ . This defines the *linear regression model* (see Chapter 19).

- (b) Assuming that  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is a normal IID stochastic process and choosing  $\mathcal{D}_t = \sigma(\mathbf{X}_t)$ , (15) becomes

$$y_t = \boldsymbol{\beta}' \mathbf{X}_t + u_t, \quad t \in \mathbb{T}, \quad (17.19)$$

with  $D(\mathbf{Z}_t; \psi)$  being the distribution defining the probability model. (19) represents the statistical GM of the *stochastic linear regression model* (see Chapter 20).

- (c) Assuming that  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is a normal stationary  $l$ th-order Markov process and choosing the appropriate  $\sigma$ -field to be  $\mathcal{D}_t = \sigma(\mathbf{y}_{t-1}^0, \mathbf{X}_t^0 = \mathbf{x}_t^0), \mathbf{y}_{t-1}^0 = (y_{t-i}, i = 1, 2, \dots), \mathbf{X}_t^0 = (\mathbf{X}_{t-i}, i = 0, 1, 2, \dots)$ , (15) takes the form

$$y_t = \beta'_0 \mathbf{x}_t + \sum_{i=1}^l (\alpha_i y_{t-i} + \beta'_i \mathbf{x}_{t-i}) + u_t, \quad (17.20)$$

where the underlying probability model is based on  $D(y_t / \mathbf{y}_{t-1}^0, \mathbf{X}_t^0; \theta_0)$ . This defines the statistical GM of the *dynamic linear regression model* (see Chapter 23).

- (d) Assuming that  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is a normal IID stochastic process and  $\mathbf{y}_t$  is an  $m \times 1$  subvector of  $\mathbf{Z}_t$ , the  $\sigma$ -field  $\mathcal{D}_t = \sigma(\mathbf{X}_t = \mathbf{x}_t)$  reduces (15) to

$$\mathbf{y}_t = \mathbf{B}' \mathbf{x}_t + \mathbf{u}_t, \quad t \in \mathbb{T}, \quad (17.21)$$

with  $D(\mathbf{y}_t / \mathbf{X}_t; \theta^*)$  the distribution defining the underlying probability model. This is the statistical GM of the *multivariate linear regression model* (see Chapter 24).

An important feature of any statistical GM is the set of parameters defining it. These parameters are called the *statistical parameters of interest*. For instance, in the case of (18) and (19) the statistical parameters of interest are  $\theta \equiv (\beta, \sigma^2), \beta = \Sigma_{22}^{-1} \sigma_2, \sigma^2 = \sigma_{11} - \sigma_{12} \Sigma_{22}^{-1} \sigma_{21}$ . These are functions of the parameters of  $D(\mathbf{Z}_t; \psi)$  assumed to be

$$\mathbf{Z}_t \equiv \begin{pmatrix} y_t \\ \mathbf{X}_t \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \Sigma_{22} \end{pmatrix}\right) \quad (17.22)$$

(see Chapter 15).

In practice the statistical parameters of interest  $\theta$  might not coincide with the *theoretical parameters of interest*, say  $\xi$ . In such a case we need to relate the two sets of parameters in such a way that the latter are uniquely determined by the former. That is, there exists a mapping

$$\xi = \mathbf{H}(\theta), \quad (17.23)$$

which define  $\xi$  uniquely. This situation for example arises in the case of the *simultaneous equations model* where the statistical parameters of interest are the parameters defining (21) but the theoretical parameters are different (see Chapter 25). In such a case the statistical GM is reparametrised/restricted in an attempt to define it in terms of the theoretical parameters of interest. The reparametrised/restricted statistical GM is said to be an econometric model.

It must be stressed that the statistical GM postulated depends crucially on the information set chosen at the outset and it is well defined within such a context. When the information set is changed the statistical GM should be respecified to take account of the change. This implies that in econometric modelling we have to decide on the information set within which the specification of the statistical model will take place. This is one of the reasons why the statistical model is defined directly in terms of the random variables giving rise to the available observed data chosen and *not* in terms of the error term. The relevant information underlying the specification of the statistical GM comes in three forms:

- (i)      theoretical information;
- (ii)     sample information; and
- (iii)    measurement information.

In terms of Fig. 1.2 the theoretical information relates to the choice of the observed data series (and hence of  $\mathbf{Z}_t$ ) and the form of the estimable model. The sample information relates to the probabilistic structure of  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  and the measurement information to the measurement system of  $\mathbf{Z}_t$  and any exact relationships among the observed data chosen (see Chapter 26 for further discussion). Any theoretical information which can be tested as restrictions on  $\theta$  is not imposed *a priori* in order to be able to test it. An important implication of this is that the statistical GM is not restricted *a priori* to coincide with either the theoretical or estimable model apart from a white-noise term at the outset. Moreover, before any theoretical meaning is attached to the statistical GM we need to ensure that the latter is first *well-defined statistically*; the underlying assumptions defining the statistical model are indeed valid for the data chosen. Testing the underlying assumptions is the task of *misspecification testing* (see Chapters 20–22). When these assumptions are tested and their validity established we can proceed with the *reparametrisation/restriction* in order to derive a theoretically meaningful GM, the empirical econometric model (see Fig. 1.2).

## 17.5   Looking ahead

As a prelude to the extensive discussion of the linear regression model and related statistical models of interest in econometrics let us summarise these in Table 17.1.

In the chapters which follow the statistical analysis (specification, misspecification, estimation and testing) of the above statistical models will be considered in some detail. In Chapter 18 the linear model is briefly considered in its simplest form ( $k = 2$ ) in an attempt to motivate the linear regression model considered extensively in Chapters 18–22. The main

Table 17.1. *Linear regression and related statistical models*

	Statistical GM	Probability model	Sampling model
Gauss linear model	$y_t = \mathbf{b}'\mathbf{x}_t + u_t, \quad t \in \mathbb{T}$ $\theta_1 \equiv (\mathbf{b}, \sigma_{11})$	$\Phi = \{D(y_t; \theta_1), \theta_1 \in \mathbb{R}^k \times \mathbb{R}_+, t \in \mathbb{T}\}$ $D(y_t; \theta_1)$ is normal	$\mathbf{y} \equiv (y_1, \dots, y_T)'$ is an independent sample from $D(\mathbf{y}_t; \theta_1), t = 1, 2, \dots, T$
Linear regression model	$y_t = \boldsymbol{\beta}'\mathbf{x}_t + u_t, \quad t \in \mathbb{T}$ $\theta_2 = (\boldsymbol{\beta}, \sigma^2)$	$\Phi = \{D(y_t/\mathbf{X}_t; \theta_2), \theta_2 \in \mathbb{R}^k \times \mathbb{R}_+, t \in \mathbb{T}\}$ $D(y_t/\mathbf{X}_t; \theta_2)$ is normal	$\mathbf{y} \equiv (y_1, \dots, y_T)'$ is an independent sample sequentially drawn from $D(y_t/\mathbf{X}_t; \theta_2), t = 1, 2, \dots, T$
Stochastic linear regression model	$y_2 = \boldsymbol{\beta}'\mathbf{x}_2 + u_t, \quad t \in \mathbb{T}$ $\theta_2 = (\boldsymbol{\beta}, \sigma^2)$	$\Phi = \{D(y_t, \mathbf{X}_t; \psi), \psi \equiv (\theta_2, \psi_2) \in \Theta, t \in \mathbb{T}\}$ $D(y_t/\mathbf{X}_t; \theta_2)$ is normal	$\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_T$ , is a random sample from $D(Z_t; \psi), t = 1, 2, \dots, T$
Dynamic linear regression model	$y_t = \boldsymbol{\beta}_0'\mathbf{x}_t + \sum_{i=1}^t y_i z_{t-i} + u_t, \quad t \in \mathbb{T}$ $\theta_3 \equiv \boldsymbol{\beta}_0, \mathbf{y}_1, \dots, \mathbf{y}_t, \sigma_0^2$	$\Phi = \{D(y_t/\mathbf{Y}_{t-1}^0, \mathbf{X}_t^0; \theta_3), \theta_3 \in \Theta, t \in \mathbb{T}\}$ $D(y_t/\mathbf{Y}_{t-1}^0, \mathbf{X}_t^0; \theta_3)$ is normal	$\mathbf{y}$ is a non-random sample sequentially drawn from $D(y_t/\mathbf{Y}_{t-1}^0, \mathbf{X}_t^0; \theta_3), t = 1, 2, \dots, T$
Multivariate linear regression model	$\mathbf{y}_t = \mathbf{B}'\mathbf{x}_t + \mathbf{u}_t, \quad t \in \mathbb{T}$ $\theta_4 \equiv (\mathbf{B}, \Omega)$	$\Phi = \{D(\mathbf{y}_t/\mathbf{X}_t; \theta_4), \theta_4 \in \Theta, t \in \mathbb{T}\}$ $D(\mathbf{y}_t/\mathbf{X}_t; \theta_4)$ is normal	$\mathbf{Y} \equiv (y_1, y_2, \dots, y_T)'$ is an independent sample sequentially drawn from $D(\mathbf{y}_t/\mathbf{X}_t; \theta_4), t = 1, 2, \dots, T$
Simultaneous equations model	$\xi \equiv \mathbf{H}(\theta_4)$ are the theoretical parameters of interest uniquely defined in terms of $\theta_4$		

reason for the extensive discussion of the linear regression model is that this statistical model forms the backbone of Part IV. In Chapter 19 the estimation, specification testing and prediction in the context of the linear regression model are discussed. Departures from the assumptions (misspecification) underlying the linear regression model are discussed in Chapters 20–22. Chapter 23 considers the dynamic linear regression model which is by far the most widely used statistical model in econometric modelling. This statistical model is viewed as a natural extension of the linear regression model in the case where the non-random sample is the appropriate sampling model. In Chapter 24 the multivariate linear regression model is discussed as a direct extension of the linear regression model. The simultaneous equation model viewed as a reparametrisation of the multivariate linear regression model is discussed in Chapter 25. In Chapter 26 the methodological discussion sketched in Chapter 1 is considered more extensively.

### ***Important concepts***

Time-series, cross-section and panel data, simple random sampling, stratified sampling, incidental parameters, statistical generating mechanism, systematic and non-systematic components, statistical parameters of interest, theoretical parameters of interest, reparametrisation/restriction.

### ***Questions***

1. Explain why for most forms of economic data the notion of a random sample is inappropriate.
2. Explain the concept of a statistical GM and its role in the statistical model specification.
3. Explain the concepts of the systematic and non-systematic components.
4. Discuss the type of information relevant for the specification of a statistical GM.

**Appendix 17.1**

Table 17.2. *Quarterly seasonally adjusted data on money stock M1 (M), real consumers' expenditure (Y), its implicit price deflator (P) and interest rate on 7 days' deposit account (I) for the period 1963i–1982iv. (Source: Economic Trends, Annual Supplement, 1983, CSO)*

	<i>M</i>	<i>Y</i>	<i>P</i>	<i>I</i>
1	6740.0	12 086.0	0.402 53	0.202 00E-01
2	6870.0	12 446.0	0.403 26	0.200 00E-01
3	6990.0	12 575.0	0.405 81	0.200 00E-01
4	7210.3	12 618.0	0.408 15	0.200 00E-01
5	7280.0	12 691.0	0.412 58	0.237 00E-01
6	7330.0	12 787.0	0.416 36	0.300 00E-01
7	7440.0	12 847.0	0.421 27	0.300 00E-01
8	7450.0	12 949.0	0.426 98	0.390 00E-01
9	7490.0	12 959.0	0.432 98	0.500 00E-01
10	7570.0	12 960.0	0.437 81	0.470 00E-01
11	7620.0	13 095.0	0.442 38	0.400 00E-01
12	7610.0	13 117.0	0.446 37	0.400 00E-01
13	7910.0	13 304.0	0.449 94	0.400 00E-01
14	7830.0	13 458.0	0.454 97	0.400 00E-01
15	7740.0	13 258.0	0.459 72	0.486 00E-01
16	7600.0	13 164.0	0.465 36	0.500 00E-01
17	7780.0	13 311.0	0.465 33	0.455 00E-01
18	7880.0	13 527.0	0.467 36	0.368 00E-01
19	8160.0	13 726.0	0.470 42	0.350 00E-01
20	8250.0	13 821.0	0.474 35	0.489 00E-01
21	8210.0	14 290.0	0.477 82	0.594 00E-01
22	8340.0	13 691.0	0.489 52	0.550 00E-01
23	8530.0	13 962.0	0.497 14	0.544 00E-01
24	8640.0	14 083.0	0.501 10	0.500 00E-01
25	8490.0	13 960.0	0.511 03	0.535 00E-01
26	8310.0	13 988.0	0.516 94	0.600 00E-01
27	8380.0	14 089.0	0.520 83	0.600 00E-01
28	8660.0	14 276.0	0.527 46	0.600 00E-01
29	8640.0	14 217.0	0.534 99	0.585 00E-01
30	8920.0	14 359.0	0.544 82	0.508 00E-01
31	9020.0	14 597.0	0.552 99	0.500 00E-01
32	9420.0	14 641.0	0.565 33	0.500 00E-01
33	9820.0	14 603.0	0.576 53	0.500 00E-01
34	9900.0	14 867.0	0.592 99	0.400 00E-01
35	10 210.0	15 071.0	0.603 48	0.367 00E-01
36	10 310.0	15 183.0	0.610 49	0.325 00E-01
37	11 300.0	15 503.0	0.615 75	0.250 00E-01
38	11 740.0	15 766.0	0.624 45	0.250 00E-01
39	12 050.0	15 930.0	0.641 81	0.470 00E-01
40	12 370.0	16 071.0	0.657 58	0.544 00E-01
41	12 440.0	16 724.0	0.665 15	0.718 00E-01
42	13 200.0	16 525.0	0.677 76	0.703 00E-01
43	12 960.0	16 566.0	0.695 34	0.827 00E-01

*continued*

Table 17.2. *continued*

44	13 020.0	16 517.0	0.721 44	0.950 00E-01
45	12 850.0	16 211.0	0.752 76	0.950 00E-01
46	13 230.0	16 169.0	0.790 53	0.950 00E-01
47	13 550.0	16 288.0	0.824 84	0.950 00E-01
48	14 460.0	16 381.0	0.867 22	0.950 00E-01
49	14 850.0	16 342.0	0.919 47	0.846 00E-01
50	15 250.0	16 358.0	0.982 88	0.625 00E-01
51	16 770.0	16 015.0	1.0297	0.642 00E-01
52	17 150.0	15 937.0	1.0703	0.700 00E-01
53	17 880.0	16 105.0	1.1027	0.588 00E-01
54	18 430.0	16 163.0	1.1322	0.592 00E-01
55	19 050.0	16 199.0	1.1679	0.693 00E-01
56	19 000.0	16 240.0	1.2237	0.107 30
57	19 440.0	15 980.0	1.2770	0.565 00E-01
58	20 430.0	16 020.0	1.3216	0.428 00E-01
59	21 970.0	16 153.0	1.3523	0.377 00E-01
60	23 170.0	16 364.0	1.3766	0.332 00E-01
61	24 280.0	16 840.0	1.4059	0.306 00E-01
62	24 950.0	16 884.0	1.4363	0.518 00E-01
63	25 920.0	17 249.0	1.4619	0.675 00E-01
64	26 920.0	17 254.0	1.4910	0.857 00E-01
65	27 520.0	17 396.0	1.5357	0.103 70
66	28 030.0	18 315.0	1.5796	0.993 00E-01
67	28 840.0	17 816.0	1.6808	0.115 00
68	29 360.0	18 072.0	1.7419	0.131 40
69	29 260.0	18 120.0	1.8109	0.150 00
70	29 880.0	17 729.0	1.8823	0.150 00
71	29 660.0	17 831.0	1.9246	0.140 50
72	30 550.0	17 870.0	1.9717	0.131 00
73	31 810.0	18 040.0	2.0154	0.109 40
74	32 870.0	17 926.0	2.0867	0.900 00E-01
75	33 210.0	17 934.0	2.1343	0.943 00E-01
76	33 760.0	17 971.0	2.1838	0.133 00
77	36 720.0	17 927.0	2.2177	0.114 20
78	37 590.0	17 998.0	2.2673	0.100 10
79	38 140.0	18 242.0	2.2919	0.829 00E-01
80	40 220.0	18 543.0	2.3076	0.624 00E-01

**Additional references**

Granger (1982); Griliches (1985); Richard (1980).

## CHAPTER 18

---

### The Gauss linear model

---

#### 18.1 Specification

In the context of the Gauss linear model the only random variable involved is the variable whose behaviour is of interest. Denoting this random variable by  $y_t$ , we assume that the stochastic process  $\{y_t, t \in \mathbb{T}\}$  is a normal, independent process with  $E(y_t) = \mu_t$  and a time-homogeneous variance  $\sigma^2$  for  $t \in \mathbb{T}$  ( $\mathbb{T}$  being some index set, not necessarily time),

$$y_t \sim N(\mu_t, \sigma^2), \quad t \in \mathbb{T}, \tag{18.1}$$

defined on the probability space  $(S, \mathcal{F}, P(\cdot))$ .

In terms of the general statistical GM (17.15) the relevant conditioning set is the trivial  $\sigma$ -field  $\mathcal{D}_0 = \{S, \emptyset\}$  which implies that

$$\mu_t = E(y_t | \mathcal{D}_0) = E(y_t).$$

That is, the statistical GM is

$$y_t = E(y_t) + u_t, \quad t \in \mathbb{T}, \tag{18.2}$$

with  $\mu_t$  assumed to be related to a set of  $k$  non-stochastic (or controlled) variables  $x_{1t}, x_{2t}, \dots, x_{kt}$ , in the form of the linear function

$$\mu_t = \sum_{i=1}^k b_i x_{it} \equiv \mathbf{b}' \mathbf{x}_t, \quad t \in \mathbb{T} \tag{18.3}$$

is an obvious notation. Defining the non-systematic component by

$$u_t = y_t - E(y_t), \quad t \in \mathbb{T}, \tag{18.4}$$

the statistical GM (2) takes the particular form

$$y_t = \mathbf{b}' \mathbf{x}_t + u_t, \quad t \in \mathbb{T}. \quad (18.5)$$

The underlying probability is naturally defined in terms of the marginal distribution of  $y_t$ , say,  $D(y_t; \boldsymbol{\theta})$ , where  $\boldsymbol{\theta} \equiv (\mathbf{b}, \sigma_1^2)$  are the *statistical parameters of interest*, being the parameters in terms of which the statistical GM (5) is defined. The *probability model* is defined by

$$\Phi = \left\{ D(y_t; \boldsymbol{\theta}) = \frac{1}{\sigma \sqrt{(2\pi)}} \exp \left\{ -\frac{1}{2\sigma^2} (y_t - \mathbf{b}' \mathbf{x}_t)^2 \right\}, \right. \\ \left. \boldsymbol{\theta} \in \mathbb{R}^k \times \mathbb{R}_+, t \in \mathbb{T} \right\} \quad (18.6)$$

In view of the assumption of independence of  $\{y_t, t \in \mathbb{T}\}$  the *sampling model*, providing the link between the observed data and the statistical GM, is defined as follows:

$$\mathbf{y} \equiv (y_1, y_2, \dots, y_T)'$$

is an independent sample from  $D(y_t; \boldsymbol{\theta})$ ,  $t = 1, 2, \dots, T$ , respectively. It could not be a random sample in view of the fact that each  $y_t$  has a different mean.

By construction the systematic and non-systematic components satisfy the following properties:

- (i)       $E(u_t) = E(y_t - E(y_t)) = 0;$
- (ii)      $E(\mu_t u_t) = \mu_t E(u_t) = 0;$
- (iii)     $E(u_t u_s) = \begin{cases} \sigma^2, & t = s \\ 0, & t \neq s, \quad t, s \in \mathbb{T}. \end{cases}$

Properties (i) and (iii) show that  $\{u_t, t \in \mathbb{T}\}$  is a *normal white-noise process* and (ii) establishes the orthogonality of the two components. It is important to note that the distribution in terms of which the above expectation operator  $E(\cdot)$  is defined is none other than  $D(y_t; \boldsymbol{\theta}_0)$ , the distribution underlying the probability model with  $\boldsymbol{\theta}_0$  the ‘true’ value of  $\boldsymbol{\theta}$ .

The Gauss linear model is specified by the statistical GM (5), the probability model (6) and the sampling model defined above. Looking at this statistical model we can see that it purports to model an ‘experimental-like’ situation where the  $x_{it}$ s are either fixed or controlled by the experimenter and the chosen values determine the systematic component of  $y_t$  via (3). This renders this statistical model of limited applicability in econometrics where controlled experimentation is rather rare. At the outset the modeller adopting the Gauss linear model discriminates between  $y_t$  and

the  $x_{it}$ s on probabilistic grounds by assuming  $y_t$  is a random variable and the  $x_{it}$ s non-stochastic or controlled variables. In econometric modelling, however, apart from a time trend variable, say  $x_t = t$ ,  $t \in \mathbb{T}$ , and dummy variables taking the value zero or one by design, it is very difficult to think of non-stochastic or controlled variables.

The Gauss linear model is of interest in econometrics mainly because it enhances our understanding of the linear regression model (see Chapter 19) when the two are compared. The two models seem to be almost identical notation-wise, thus causing some confusion; but a closer comparison reveals important differences rendering the two models applicable to very different situations. This will be pursued further in the next chapter.

## 18.2 Estimation

For expositional purposes let us consider the simplest case where there are only two non-stochastic variables ( $k=2$ ) and the statistical GM of the Gauss linear model takes the simple form

$$y_t = b_1 + b_2 x_{1t} + u_t, \quad t \in \mathbb{T}. \quad (18.7)$$

The reason for choosing this simple case is to utilise the similarity of the mathematical manipulations between the Gauss linear and linear regression models in order to enhance the reader's understanding of the matrix notation used in the context of the latter (see Chapter 19). The first variable in (7) takes the value one for all  $t$ , commonly called the *constant* (or intercept).

In view of the probability model (6) and the sampling model assumption of independence we can deduce that the distribution of the sample (see Chapter 11) takes the form

$$D(y_1, y_2, \dots, y_T; \boldsymbol{\theta}) = \prod_{t=1}^T D(y_t; \boldsymbol{\theta}). \quad (18.8)$$

Hence, the likelihood function, ignoring the constant of proportionality, can be defined by

$$\begin{aligned} L(\boldsymbol{\theta}; \mathbf{y}) &= \prod_{t=1}^T D(y_t; \boldsymbol{\theta}) \\ &= \prod_{t=1}^T \left( \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp \left\{ \frac{1}{2\sigma^2} (y_t - b_1 - b_2 x_t)^2 \right\} \right) \\ &= (2\pi)^{-T/2} (\sigma^2)^{-T/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_t (y_t - b_1 - b_2 x_t)^2 \right\}. \end{aligned} \quad (18.9)$$

(see Section 13.3). The log likelihood takes the form:

$$\begin{aligned}\log L(\theta; \mathbf{y}) &= -\frac{T}{2} \log 2\pi - \frac{T}{2} \log \sigma^2 \\ &\quad - \frac{1}{2\sigma^2} \sum_t (y_t - b_1 - b_2 x_t)^2.\end{aligned}\tag{18.10}$$

The first-order conditions for the derivation of the *maximum likelihood estimators* (MLE's) are:

$$\frac{\partial \log L}{\partial b_1} = -\frac{1}{2\sigma^2} (-2) \sum_t (y_t - b_1 - b_2 x_t) = 0,\tag{18.11}$$

$$\frac{\partial \log L}{\partial b_2} = -\frac{1}{2\sigma^2} (-2) \sum_t (y_t - b_1 - b_2 x_t) x_t = 0,\tag{18.12}$$

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_t (y_t - b_1 - b_2 x_t)^2 = 0.\tag{18.13}$$

Solving (11)–(13) simultaneously we get the MLE's

$$\hat{b}_1 = \bar{y} - \hat{b}_2 \bar{x},\tag{18.14}$$

$$\hat{b}_2 = \frac{\sum_t (y_t - \bar{y})(x_t - \bar{x})}{\sum_t (x_t - \bar{x})^2},\tag{18.15}$$

where

$$\bar{y} = \frac{1}{T} \sum_t y_t, \quad \bar{x} = \frac{1}{T} \sum_t x_t, \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{T} \sum_t \hat{u}_t,\tag{18.16}$$

where  $\hat{u}_t \equiv y_t - \hat{b}_1 - \hat{b}_2 x_t$  represents the estimated *residuals*; the natural 'estimator' of the error term  $u_t$ . Taking second derivatives

$$\begin{aligned}\frac{\partial^2 \log L}{\partial b_1^2} &= -\frac{T}{\sigma^2}, \quad \frac{\partial^2 \log L}{\partial b_2^2} = -\frac{1}{\sigma^2} \sum_t x_t^2, \\ \frac{\partial^2 \log L}{\partial \sigma^4} &= \frac{T}{2\sigma^4} - \frac{1}{\sigma^6} \sum_t u_t^2,\end{aligned}\tag{18.17}$$

$$\begin{aligned}\frac{\partial^2 \log L}{\partial b_1 \partial b_2} &= -\frac{1}{\sigma^2} \sum_t x_t, \quad \frac{\partial^2 \log L}{\partial b_1 \partial \sigma^2} = -\frac{1}{\sigma^4} \sum_t u_t, \\ \frac{\partial^2 \log L}{\partial b_2 \partial \sigma^2} &= -\frac{1}{\sigma^4} \sum_t x_t u_t.\end{aligned}\tag{18.18}$$

For  $\theta = (b_1, b_2, \sigma^2)$ , the sample information matrix  $\mathbf{I}_T(\theta)$  and its inverse are

$$\mathbf{I}_T(\theta) = \begin{bmatrix} \left(\frac{T}{\sigma^2}\right) & \frac{\sum x_t}{\sigma^2} & 0 \\ \frac{\sum x_t}{\sigma^2} & \frac{\sum x_t^2}{\sigma^2} & 0 \\ 0 & 0 & \left(\frac{T}{2\sigma^4}\right) \end{bmatrix} \quad (18.19)$$

and

$$[\mathbf{I}_T(\theta)]^{-1} = \begin{bmatrix} \frac{\sigma^2 \sum x_t^2}{T \sum_t (x_t - \bar{x})^2} & \frac{-\sigma^2 \sum x_t}{T \sum_t (x_t - \bar{x})^2} & 0 \\ \frac{-\sigma^2 \sum x_t}{T \sum_t (x_t - \bar{x})^2} & \frac{\sigma^2}{\sum_t (x_t - \bar{x})^2} & 0 \\ 0 & 0 & \left(\frac{2\sigma^4}{T}\right) \end{bmatrix}. \quad (18.20)$$

Note that  $I_T(\theta)$  is positive definite ( $I_T(\theta) > 0$ ) if  $\sum_t (x_t - \bar{x})^2 \neq 0$ , i.e. there must be at least two distinct values for  $x_t$ . This condition also ensures the existence of  $\hat{b}_2$  as defined by (15).

### **Properties of $\hat{\theta} = (\hat{b}_1, \hat{b}_2, \hat{\sigma}^2)$**

#### (1) Asymptotic properties

The fact that  $\hat{\theta}$  is a MLE enables us to conclude that if the asymptotic information (matrix) defined by  $I_\infty(\theta) = \lim_{T \rightarrow \infty} [(1/T)\mathbf{I}_T(\theta)]$  is positive definite (see Chapters 12–13) then:

- (i)  $\hat{\theta} \xrightarrow{P} \theta$ , i.e.  $\hat{\theta}$  is a consistent estimator of  $\theta$  (if  $\sum_{t=1}^T x_t \rightarrow \infty$  as  $T \rightarrow \infty$ );
- (ii)  $\sqrt{T}(\hat{\theta} - \theta) \xrightarrow{D} N(\mathbf{0}, [\mathbf{I}_\infty(\theta)]^{-1})$ , i.e.  $\hat{\theta}$  is asymptotically normal;
- (iii)  $E_z(\hat{\theta}) = \theta$ , i.e. asymptotically unbiased (the asymptotic mean of  $\hat{\theta}$  is  $\theta$ );
- (iv)  $\text{Var}_z(\hat{\theta}) = [\mathbf{I}_\infty(\theta)]^{-1}$ , i.e.  $\hat{\theta}$  is asymptotically efficient.

$\mathbf{I}_\infty(\boldsymbol{\theta})$  is positive definite if  $\det(\mathbf{I}_\infty(\boldsymbol{\theta})) > 0$ ; this is the case if

$$\lim_{T \rightarrow \infty} \left( \frac{1}{T} \sum_t (x_t - \bar{x})^2 \right) = q_{xx} > 0. \quad (18.21)$$

(2) *Finite sample properties*

$\hat{\boldsymbol{\theta}}$  being a MLE we can deduce that:

(v)  $\hat{\boldsymbol{\theta}}$  is a function of the set of minimal sufficient statistics

$$\tau(\mathbf{y}) = \left( \sum_{t=1}^T y_t^2, \sum_{t=1}^T y_t, \sum_{t=1}^T x_t y_t \right); \quad (18.22)$$

(vi)  $\hat{\boldsymbol{\theta}}$  is invariant with respect to Borel functions, i.e. if  $\mathbf{h}(\boldsymbol{\theta}): \Theta \rightarrow \Theta$  then the MLE of  $\mathbf{h}(\boldsymbol{\theta})$  is  $\mathbf{h}(\hat{\boldsymbol{\theta}})$ ; see Section 13.3.

In order to consider any other small (finite) sample properties of  $\hat{\boldsymbol{\theta}}$  we need to derive its distribution. Because the mathematical manipulations are rather involved in the present case no such manipulations are attempted. It turns out that these manipulations are much easier in matrix notation and will be done in the next chapter for the linear regression model which when reinterpreted applies to the present case unaltered.

$$(vii) \quad \begin{pmatrix} \hat{b}_1 \\ \hat{b}_2 \end{pmatrix} \sim N \left( \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \begin{pmatrix} \text{Var}(\hat{b}_1) & \text{Cov}(\hat{b}_1, \hat{b}_2) \\ \text{Cov}(\hat{b}_1, \hat{b}_2) & \text{Var}(\hat{b}_2) \end{pmatrix} \right), \quad (18.23)$$

where

$$\begin{aligned} \text{Var}(\hat{b}_1) &= \frac{\sigma^2 \sum_t x_t^2}{T \sum_t (x_t - \bar{x})^2}, \\ \text{Cov}(\hat{b}_1, \hat{b}_2) &= \frac{-\sigma^2 \bar{x}}{\sum_t (x_t - \bar{x})^2}, \\ \text{Var}(\hat{b}_2) &= \frac{\sigma^2}{\sum_t (x_t - \bar{x})^2}. \end{aligned}$$

This result follows from the fact that  $\hat{b}_1 = \bar{y} - \hat{b}_2 \bar{x}$ ,  $\hat{b}_2 = \sum_t \lambda_t (y_t - \bar{y})$ , where  $\lambda_t = [(x_t - \bar{x})]/[\sum_t (x_t - \bar{x})^2]$ , and linear functions of normally distributed random variables and thus themselves normally distributed (see Section 6.3). The distribution of  $\hat{\sigma}^2$  takes the form

$$\frac{T \hat{\sigma}^2}{\sigma^2} \sim \chi^2(T-2), \quad (18.24)$$

where  $\chi^2(T-2)$  stands for the chi-square distribution with  $T-2$  degrees of freedom. (2) follows from the fact that  $T\hat{\sigma}^2/\sigma^2 = \sum_{t=1}^T (\hat{u}_t/\sigma)^2$  involves  $T-2$  independent squared standard normally distributed random variables.

- (viii) From (vii) it follows that  $E(\hat{b}_1) = b_1$ ,  $E(\hat{b}_2) = b_2$ , i.e.  $\hat{b}_1$  and  $\hat{b}_2$  are unbiased estimators of  $b_1$  and  $b_2$  respectively. On the other hand, since the mean of a chi-square random variable equals its degrees of freedom (see Appendix 6.1)

$$E\left(\frac{T\hat{\sigma}^2}{\sigma^2}\right) = T-2 \Rightarrow E(\hat{\sigma}^2) = \frac{(T-2)}{T} \sigma^2 \neq \sigma^2,$$

i.e.  $\hat{\sigma}^2$  is a biased estimator of  $\sigma^2$ , but the estimator  $s^2 = [1/(T-2)] \sum_i \hat{u}_i^2$  is unbiased and

$$\frac{(T-2)s^2}{\sigma^2} \sim \chi^2(T-2). \quad (18.25)$$

- (ix)  $(\hat{b}_1, \hat{b}_2)$  are independent of  $s^2$  (or  $\hat{\sigma}^2$ ).

This can be verified by considering the covariance between them.

- (x) Comparing (23) with (20) we can see that  $(\hat{b}_1, \hat{b}_2)$  achieve the Cramer–Rao lower bound and hence we can deduce that they are fully efficient. Given that  $\hat{\sigma}^2$  is biased the Cramer–Rao given by (20) is not applicable, but for  $s^2$  we know that

$$\begin{aligned} \text{Var}\left(\frac{(T-2)s^2}{\sigma^2}\right) &= 2(T-2) \\ \Rightarrow \text{Var}(s^2) &= \frac{2\sigma^4}{T-2} \geq \frac{2\sigma^4}{T} - \text{the Cramer–Rao bound.} \end{aligned}$$

Thus, although  $s^2$  does not achieve the Cramer–Rao lower bound, no other unbiased estimator of  $\sigma^2$  achieves this bound.

### 18.3 Hypothesis testing and confidence intervals

In setting up tests and confidence intervals the distribution of  $\hat{\theta}$  and any pivotal quantities thereof are of paramount importance. Consider the null hypothesis

$$H_0: b_1 = \bar{b}_1 \quad \text{against} \quad H_1: b_1 \neq \bar{b}_1, \quad \bar{b}_1 \text{ being a constant.}$$

Intuition suggests that the distance  $|\hat{b}_1 - \bar{b}_1|$ , scaled by its standard deviation (to avoid any units of measurement problems), might provide the

basis for a ‘good’ test statistic. Given that

$$\frac{(\hat{b}_1 - \bar{b}_1)^2}{\text{Var}(\hat{b}_1)} = \frac{(\hat{b}_1 - \bar{b}_1)^2}{\left( \frac{\sigma^2 \sum_t x_t^2}{T \sum_t (x_t - \bar{x})^2} \right)} \sim \chi^2(1), \quad (18.26)$$

this is not a pivotal quantity unless  $\sigma^2$  is known. Otherwise we must find an alternative pivotal quantity. Taking (25) and (26) together and using the independence between  $\hat{b}_1$  and  $s^2$  we can set up the pivotal quantity

$$\frac{\left( \frac{(\hat{b}_1 - \bar{b}_1)^2}{\sigma^2 \sum_t x_t^2} \right)}{\left( \frac{T \sum_t (x_t - \bar{x})^2}{(T-2)s^2} \right)} = \frac{(\hat{b}_1 - \bar{b}_1)^2}{\left( \frac{s^2 \sum_t x_t^2}{T \sum_t (x_t - \bar{x})^2} \right)} = \frac{(\hat{b}_1 - \bar{b}_1)^2}{\text{Var}(\hat{b}_1)} \sim F(1, T-2), \quad (18.27)$$

$$\Rightarrow \frac{|\hat{b}_1 - \bar{b}_1|}{\sqrt{[\text{Var}(\hat{b}_1)]}} \sim t(T-2). \quad (18.28)$$

The rejection region for a size  $\alpha$  test is

$$C_1 = \left\{ \mathbf{y}: \frac{|\hat{b}_1 - \bar{b}_1|}{\sqrt{[\text{Var}(\hat{b}_1)]}} \geq c_\alpha \right\}, \quad \text{where } 1-\alpha = \int_{-c_\alpha}^{c_\alpha} dt(T-2). \quad (18.29)$$

Using the duality between hypothesis testing and confidence intervals (see Section 14.5) we can construct an  $(1-\alpha)$  level confidence interval for  $b_1$  based on the acceptance region,

$$C_0(\bar{b}_1) = \left\{ \mathbf{y}: \frac{|\hat{b}_1 - \bar{b}_1|}{\sqrt{[\text{Var}(\hat{b}_1)]}} \leq c_\alpha \right\} \Rightarrow Pr \left( -c_\alpha \leq \frac{(\hat{b}_1 - \bar{b}_1)}{\sqrt{[\text{Var}(\hat{b}_1)]}} \leq c_\alpha \right) = 1-\alpha, \quad (18.30)$$

$$C(\mathbf{y}) = \left\{ b_1: \hat{b}_1 - c_\alpha s \sqrt{\frac{\sum_t x_t^2}{T \sum_t (x_t - \bar{x})^2}} \leq b_1 \leq \hat{b}_1 + c_\alpha s \sqrt{\frac{\sum_t x_t^2}{T \sum_t (x_t - \bar{x})^2}} \right\}. \quad (18.31)$$

Similarly, for  $H_0: \hat{b}_2 = \bar{b}_2$  against  $\hat{b}_2 \neq \bar{b}_2$  the rejection region of a size  $\alpha$  test is

$$C_1(\bar{b}_2) = \left\{ \mathbf{y}: \frac{|\hat{b}_2 - \bar{b}_2|}{\sqrt{[\text{Var}(\hat{b}_2)]}} \geq c_\alpha \right\}. \quad (18.32)$$

A  $(1-\alpha)$  confidence interval is

$$C(\mathbf{y}) = \left\{ b_2: \hat{b}_2 - c_\alpha s \sqrt{\left( \frac{1}{\sum_t (x_t - \bar{x})^2} \right)} \leq b_2 \leq \hat{b}_2 + c_\alpha s \sqrt{\left( \frac{1}{\sum_t (x_t - \bar{x})^2} \right)} \right\}. \quad (18.33)$$

Consider  $H_0: \sigma^2 = \bar{\sigma}^2$  against  $H_1: \sigma^2 \neq \bar{\sigma}^2$ . The pivotal quantity (25) can be used directly to set up the acceptance region

$$C_0 = \left\{ \mathbf{y}: a \leq \frac{(T-2)s^2}{\sigma^2} \leq b \right\}, \quad \Pr(C_0) = 1 - \alpha, \quad (18.34)$$

such that

$$\Pr(\chi_{(T-2)}^2 < a) = \Pr(\chi_{(T-2)}^2 > b) = \frac{\alpha}{2}. \quad (18.35)$$

A  $(1-\alpha)$  level confidence interval is

$$C(\mathbf{y}) = \left\{ \sigma^2: \frac{(T-2)s^2}{b} \leq \sigma^2 \leq \frac{(T-2)s^2}{a} \right\}. \quad (18.36)$$

*Remark:* One-sided tests can be easily constructed by modifying the above two-sided results; see Chapter 14.

Consider the question of constructing a  $(1-\alpha)$  level confidence interval for

$$\mu_t = b_1 + b_2 x_t.$$

A natural estimator of  $\mu_t$  is  $\hat{\mu}_t = \hat{b}_1 + \hat{b}_2 x_t$ , with  $E(\hat{\mu}_t) = \mu_t$ , and

$$\begin{aligned} \text{Var}(\hat{\mu}_t) &= \text{Var}(\hat{b}_1 + \hat{b}_2 x_t) \\ &= \text{Var}(\hat{b}_1) + 2x_t \text{Cov}(\hat{b}_1, \hat{b}_2) + x_t^2 \text{Var}(\hat{b}_2) \\ &= \sigma^2 \left( \frac{1}{T} + \frac{(x_t - \bar{x})^2}{\sum_t (x_t - \bar{x})^2} \right). \end{aligned} \quad (18.37)$$

These results imply that the distribution of  $\hat{\mu}_t$  is normal and

$$(i) \quad \frac{(\hat{\mu}_t - \mu_t)}{\sqrt{[\text{Var}(\hat{\mu}_t)]}} \sim N(0, 1); \quad (18.38)$$

$$(ii) \quad t(\mathbf{y}) = \frac{(\hat{\mu}_t - \mu_t)}{\sqrt{[\hat{\text{Var}}(\hat{\mu}_t)]}} \sim t(T-2). \quad (18.39)$$

For  $c_\alpha$  such that  $\int_{c_\alpha}^{\infty} dt(T-k) = 1-\alpha$ , we could construct a  $(1-\alpha)$  confidence interval of the form

$$\begin{aligned} C(\mathbf{y}) = \left\{ \mu_t; \hat{\mu}_t - c_\alpha s \sqrt{\left[ \frac{1}{T} + \frac{(x_t - \bar{x})^2}{\sum_t (x_t - \bar{x})^2} \right]} \leq \mu_t \leq \hat{\mu}_t + c_\alpha s \sqrt{\left[ \frac{1}{T} + \frac{(x_t - \bar{x})^2}{\sum_t (x_t - \bar{x})^2} \right]} \right\}. \end{aligned} \quad (18.40)$$

This confidence interval can be extended to  $t > T$  in order to provide us with a *prediction* confidence interval for  $\mu_{T+l}$ ,  $l \geq 1$

$$\begin{aligned} C(\mathbf{y}) = \left\{ \mu_{T+l}; \hat{\mu}_{T+l} - c_\alpha s \sqrt{\left[ \frac{1}{T} + \frac{(x_t - \bar{x})^2}{\sum_t (x_t - \bar{x})^2} \right]} \leq \mu_{T+l} \right. \\ \left. < \hat{\mu}_{T+l} + c_\alpha s \sqrt{\left[ \frac{1}{T} + \frac{(x_t - \bar{x})^2}{\sum_t (x_t - \bar{x})^2} \right]} \right\}. \quad (18.41) \end{aligned}$$

(see Chapters 12 and 14 on prediction).

In concluding this section it is important to note that the hypothesis testing and confidence interval results derived above, as well as the estimation results of Section 18.2, are crucially dependent on the validity of the assumptions underlying the Gauss linear model. If any of these assumptions are in fact invalid the above results are unwarranted to a greater or lesser degree (see Chapters 20–22 for *misspecification analysis* in the context of the linear regression model).

## 18.4 Experimental design

In Section 18.2 above we have seen that the MLE's  $\hat{b}_1$  and  $\hat{b}_2$  of  $b_1$  and  $b_2$

respectively are distributed as bivariate normal as shown in (23).

The fact that the  $x_t$ s are often controlled variables enables us to consider the question of 'designing' the statistical GM (5) so as to ensure that it satisfies certain desirable properties such as robustness and parsimony. These can be achieved by choosing the  $x_t$ s and their values appropriately. Looking at their variances and covariances we can see that we could make  $\hat{b}_1$  and  $\hat{b}_2$  more 'accurate' by choosing the values of  $x_t$  in a certain way. Firstly, if  $\bar{x}=0$  then

$$\text{Cov}(\hat{b}_1, \hat{b}_2) = 0 \quad (18.42)$$

and  $\hat{b}_1$  and  $\hat{b}_2$  are now independent. This implies that if we were to make a change of origin in  $x_t$  we could ensure that  $\hat{b}_1$  and  $\hat{b}_2$  are independent. Secondly, the variances of  $\hat{b}_1$  and  $\hat{b}_2$  are minimised when  $\sum_t x_t^2$  (given  $\bar{x}=0$ ) is as large as possible. This can be easily achieved by choosing the value of  $x_t$  to be on either side of zero (to achieve  $\bar{x}=0$ ) and as large as possible. For example, we could choose the  $x_t$ s so that

$$\begin{aligned} x_1 &= x_2 = \dots = x_{\frac{1}{2}T} = n, \\ x_{\frac{1}{2}T+1} &= \dots = x_T = -n \end{aligned} \quad (18.43)$$

( $T$  even) and  $n$  is as large as possible; see Kendall and Stuart (1968).

Another important feature of the Gauss linear model is that *repeated observations* on  $y$  can be generated for some specified values of the  $x_t$ s by repeating the experiment represented by the statistical GM (7).

## 18.5 Looking ahead

From the econometric viewpoint the linear control knob model can be seen to have two questionable features. Firstly, the fact that the  $x_{it}$ s are assumed to be non-stochastic reduces the applicability of the model. Secondly, the independent sample assumption can be called into question for most economic data series. In other disciplines where experimentation is possible the Gauss linear model is a very important statistical model. The purpose of the next chapter is to develop a similar statistical model where the first questionable feature is substituted by a more realistic formulation of the systematic component. The variables involved are all assumed to be random variables at the outset.

### ***Important concepts***

Non-stochastic or controlled variables, residuals, experimental design, repeated observations.

***Questions***

1. Explain the statistical GM of the Gauss linear model.
2. Derive the MLE's of  $\mathbf{b}$  and  $\sigma^2$  in the case of the general Gauss linear model where  $y_t = \mathbf{b}'\mathbf{x}_t + u_t$ ,  $t = 1, 2, \dots, T$ ,  $\mathbf{x}_t$  being a  $k \times 1$  vector of non-stochastic variables, and state their asymptotic properties.
3. Explain under what circumstances the MLE's  $\hat{b}_1$  and  $\hat{b}_2$  of  $b_1$  and  $b_2$  respectively are independent. Can we design the values of the non-stochastic variables so as to get independence?
4. Explain why the statistic  $|\hat{b}_1|/\hat{\text{Var}}(\hat{b}_1)$  is distributed as  $t(T-2)$  and use it to set up a test for

$$H_0: b_1 = 0 \quad \text{against} \quad H_1: b_1 \neq 0,$$

as well as a confidence interval for  $b_1$ .

5. Verify that  $\hat{\mathbf{b}}$  and  $\hat{\sigma}^2$  are independent.

***Additional references***

Chow (1983); Dhrymes (1978); Johnston (1984); Judge *et al.* (1982); Kmenta (1971); Koutsoyiannis (1977); Maddala (1977); Pindyck and Rubinfeld (1981).

## CHAPTER 19

---

### The linear regression model I – specification, estimation and testing

---

#### 19.1 Introduction

The linear regression model forms the backbone of most other statistical models of particular interest in econometrics. A sound understanding of the specification, estimation, testing and prediction in the linear regression model holds the key to a better understanding of the other statistical models discussed in the present book.

In relation to the Gauss linear model discussed in Chapter 18, apart from some apparent similarity in the notation and the mathematical manipulations involved in the statistical analysis, the linear regression model purports to model a very different situation from the one envisaged by the former. In particular the Gauss linear model could be considered to be the appropriate statistical model for analysing estimable models of the form

$$M_t = \alpha_0 + \alpha_1 t + \sum_{i=1}^3 c_i Q_{it} + \sum_{i=1}^3 d_i Q_{it}, \quad (19.1)$$

$$M_t = \alpha_0 + \sum_{i=1}^k d_i t^i, \quad (19.2)$$

where  $M_t$  refers to money and  $Q_{it}$ ,  $i = 1, 2, 3$  to quarterly dummy variables, in view of the non-stochastic nature of the  $x_{it}$ s involved. On the other hand, estimable models such as

$$M = AY^{x_1}P^{x_2}I^{x_3}, \quad (19.3)$$

referring to a demand for money function ( $M$  – money,  $Y$  – income,  $P$  – price level,  $I$  – interest rate), could not be analysed in the context of the Gauss

linear model. This is because it is rather arbitrary to discriminate on probabilistic grounds between the variable giving rise to the observed data chosen for  $M$  and those for  $Y, P$  and  $I$ . For estimable models such as (3) the linear regression model as sketched in Chapter 17 seems more appropriate, especially if the observed data chosen do not exhibit time dependence. This will become clearer in the present chapter after the specification of the linear regression model in Section 19.2. The money demand function (3) is used to illustrate the various concepts and results introduced throughout this chapter.

## 19.2 Specification

Let  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  be a vector stochastic process on the probability space  $(S, \mathcal{F}, P(\cdot))$  where  $\mathbf{Z}_t = (y_t, \mathbf{X}'_t)'$  represents the vector of random variables giving rise to the observed data chosen, with  $y_t$  being the variable whose behaviour we are aiming to explain. The stochastic process  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is assumed to be *normal, independent and identically distributed* (NIID) with  $E(\mathbf{Z}_t) = \mathbf{m}$  and  $\text{Cov}(\mathbf{Z}_t) = \Sigma$ . i.e.

$$\begin{pmatrix} y_t \\ \mathbf{X}_t \end{pmatrix} \sim N\left(\begin{pmatrix} m_y \\ \mathbf{m}_x \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \Sigma_{22} \end{pmatrix}\right), \quad t \in \mathbb{T} \quad (19.4)$$

in an obvious notation (see Chapter 15). It is interesting to note at this stage that these assumptions seem rather restrictive for most economic data in general and time-series in particular.

On the basis of the assumption that  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is a NIID vector stochastic process we can proceed to reduce the joint distribution  $D(\mathbf{Z}_1, \dots, \mathbf{Z}_T; \psi)$  in order to define the *statistical GM* of the linear regression model using the general form

$$y_t = \mu_t + u_t, \quad t \in \mathbb{T}, \quad (19.5)$$

where

$\mu_t = E(y_t | \mathbf{X}_t = \mathbf{x}_t)$  is the *systematic component*,

and

$u_t = y_t - E(y_t | \mathbf{X}_t = \mathbf{x}_t)$  the *non-systematic component*

(see Chapter 17). In view of the normality of  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  we deduce that

$$\mu_t = E(y_t | \mathbf{X}_t = \mathbf{x}_t) = \beta_0 + \boldsymbol{\beta}' \mathbf{x}_t \quad (\text{linear in } \mathbf{x}_t), \quad (19.6)$$

where

$$\beta_0 = m_y - \sigma_{12} \Sigma_{22}^{-1} \mathbf{m}_x, \quad \boldsymbol{\beta} = \Sigma_{22}^{-1} \boldsymbol{\sigma}_{21}$$

and

$$\text{Var}(u_t | \mathbf{X}_t = \mathbf{x}_t) = \text{Var}(y_t | \mathbf{X}_t = \mathbf{x}_t) = \sigma^2 \quad (\text{homoskedastic}), \quad (19.7)$$

where  $\sigma^2 = \sigma_{11} - \sigma_{12}\Sigma_{22}^{-1}\sigma_{21}$  (see Chapter 15). The *time invariance* of the parameters  $\beta_0$ ,  $\boldsymbol{\beta}$  and  $\sigma^2$  stems from the identically distributed (ID) assumption related to  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$ . It is important, however, to note that the ID assumption provides only a sufficient condition for the time invariance of the statistical parameters.

In order to simplify the notation let us assume the  $\mathbf{m} = \mathbf{0}$  without any loss of generality given that we can easily transform the original variables in mean derivation form  $(y_t - m_y)$  and  $(\mathbf{X}_t - \mathbf{m}_x)$ . This implies that  $\beta_0$ , the coefficient of the constant, is zero and the systematic component becomes

$$\mu_t = E(y_t / \mathbf{X}_t = \mathbf{x}_t) = \boldsymbol{\beta}' \mathbf{x}_t. \quad (19.8)$$

In practice, however, unless the observed data are in mean deviation form the constant should never be dropped because the estimates derived otherwise are not estimates of the regression coefficients  $\boldsymbol{\beta} = \Sigma_{22}^{-1}\sigma_{21}$  but of  $\boldsymbol{\beta}^* = E(\mathbf{X}_t \mathbf{X}_t')^{-1} E(\mathbf{X}_t' y_t)$ ; see Appendix 19.1 on the role of the constant.

The statistical GM of the linear regression model takes the particular form

$$y_t = \boldsymbol{\beta}' \mathbf{x}_t + u_t, \quad t \in \mathbb{T}, \quad (19.9)$$

with  $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}, \sigma^2)$  being the *statistical parameters of interest*; the parameters in terms of which the statistical GM is defined. By construction the systematic and non-systematic components of (9) satisfy the following properties:

- (i)  $E(u_t / \mathbf{X}_t = \mathbf{x}_t) = E[(y_t - E(y_t / \mathbf{X}_t = \mathbf{x}_t)) / \mathbf{X}_t = \mathbf{x}_t]$   
 $= E(y_t / \mathbf{X}_t = \mathbf{x}_t) - E(y_t / \mathbf{X}_t = \mathbf{x}_t) = 0;$
- (ii)  $E(u_t u_s / \mathbf{X}_t = \mathbf{x}_t)$   
 $= E[(y_t - E(y_t / \mathbf{X}_t = \mathbf{x}_t))(y_s - E(y_s / \mathbf{X}_t = \mathbf{x}_t)) / \mathbf{X}_t = \mathbf{x}_t]$   
 $= \begin{cases} \sigma^2, & t = s \\ 0, & t \neq s; \end{cases}$
- (iii)  $E(\mu_t u_t / \mathbf{X}_t = \mathbf{x}_t) = \mu_t E(u_t / \mathbf{X}_t = \mathbf{x}_t) = 0, \quad t, s \in \mathbb{T}.$

The first two properties define  $\{u_t, t \in \mathbb{T}\}$  to be a white-noise process and (iii) establishes the orthogonality of the two components. It is important to note that the above expectation operator  $E(\cdot / \mathbf{X}_t = \mathbf{x}_t)$  is defined in terms of  $D(y_t / \mathbf{X}_t; \boldsymbol{\theta})$ , which is the distribution underlying the probability model for (9). However, the above properties hold for  $E(\cdot)$  defined in terms of  $D(\mathbf{Z}_t; \psi)$  as well, given that:

- (i)'  $E(u_t) = E\{E(u_t / \mathbf{X}_t = \mathbf{x}_t)\} = 0;$
- (ii)'  $E(u_t u_s) = E\{E(u_t u_s / \mathbf{X}_t = \mathbf{x}_t)\} = \begin{cases} \sigma^2, & t = s \\ 0, & t \neq s; \end{cases}$

and

$$(iii)' \quad E(\mu_t u_t) = E\{E(\mu_t u_t | \mathbf{X}_t = \mathbf{x}_t)\} = 0, \quad t, s \in \mathbb{T}$$

(see Section 7.2 on conditional expectation).

The conditional distribution  $D(y_t | \mathbf{X}_t; \boldsymbol{\theta})$  is related to the joint distribution  $D(y_t, \mathbf{X}_t; \boldsymbol{\psi})$  via the decomposition

$$D(y_t, \mathbf{X}_t; \boldsymbol{\psi}) = D(y_t | \mathbf{X}_t; \boldsymbol{\psi}_1) \cdot D(\mathbf{X}_t; \boldsymbol{\psi}_2) \quad (19.10)$$

(see Chapter 5). Given that in defining the probability model of the linear regression model as based on  $D(y_t | \mathbf{X}_t; \boldsymbol{\theta})$  we choose to ignore  $D(\mathbf{X}_t; \boldsymbol{\psi}_2)$  for the estimation of the statistical parameters of interest  $\boldsymbol{\theta}$ . For this to be possible we need to ensure that  $\mathbf{X}_t$  is *weakly exogenous* with respect to  $\boldsymbol{\theta}$  for the sample period  $t = 1, 2, \dots, T$  (see Section 19.3, below).

For the statistical parameters of interest  $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}, \sigma^2)$  to be well defined we need to ensure that  $\Sigma_{22}$  is non-singular, in view of the formulae  $\boldsymbol{\beta} = \Sigma_{22}^{-1} \boldsymbol{\sigma}_{21}$ ,  $\sigma^2 = \boldsymbol{\sigma}_{11} - \boldsymbol{\sigma}_{12} \Sigma_{22}^{-1} \boldsymbol{\sigma}_{21}$ , at least for the sample period  $t = 1, 2, \dots, T$ . This requires that the sample equivalent of  $\Sigma_{22}, (1/T)(\mathbf{X}'\mathbf{X})$  where  $\mathbf{X} \equiv (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T)'$  is indeed non-singular, i.e.

$$\text{rank}(\mathbf{X}'\mathbf{X}) = \text{rank}(\mathbf{X}) = k, \quad (19.11)$$

$\mathbf{X}_t$  being a  $k \times 1$  vector.

As argued in Chapter 17, the statistical parameters of interest do not necessarily coincide with the theoretical parameters of interest  $\xi$ . We need, however, to ensure that  $\xi$  is uniquely defined in terms of  $\boldsymbol{\theta}$  for  $\xi$  to be *identifiable*. In constructing empirical econometric models we proceed from a well-defined estimated statistical GM (see Chapter 22) to reparametrise it in terms of the theoretical parameters of interest. Any restrictions induced by the reparametrisation, however, should be tested for their validity. For this reason *no a priori restrictions* are imposed on  $\boldsymbol{\theta}$  at the outset to make such restrictions testable at a later stage.

As argued above, the *probability model* underlying (9) is defined in terms of  $D(y_t | \mathbf{X}_t; \boldsymbol{\theta})$  and takes the form

$$\Phi = \left\{ D(y_t | \mathbf{X}_t; \boldsymbol{\theta}) = \frac{1}{\sigma \sqrt{(2\pi)}} \exp \left\{ -\frac{1}{2\sigma^2} (y_t - \boldsymbol{\beta}' \mathbf{x}_t)^2 \right\}, \right. \\ \left. \boldsymbol{\theta} \in \mathbb{R}^k \times \mathbb{R}_+, t \in \mathbb{T} \right\}. \quad (19.12)$$

Moreover, in view of the independence of  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  the *sampling model* takes the form of an independent sample,  $\mathbf{y} \equiv (y_1, \dots, y_T)'$ , sequentially drawn from  $D(y_t | \mathbf{X}_t; \boldsymbol{\theta})$ ,  $t = 1, 2, \dots, T$ , respectively.

Having defined all three components of the linear regression model let us

collect all the assumptions together and specify the statistical model properly.

### *The linear regression model: specification*

(I) **Statistical GM**,  $y_t = \beta' \mathbf{x}_t + u_t$ ,  $t \in \mathbb{T}$

- [1]  $\mu_t = E(y_t | \mathbf{X}_t = \mathbf{x}_t)$  – the systematic component;  $u_t = y_t - E(y_t | \mathbf{X}_t = \mathbf{x}_t)$  – the non-systematic component.
- [2]  $\theta \equiv (\beta, \sigma^2)$ ,  $\beta = \Sigma_{22}^{-1} \sigma_{21}$ ,  $\sigma^2 = \sigma_{11} - \sigma_{12} \Sigma_{22}^{-1} \sigma_{21}$  are the statistical parameters of interest. (Note:  $\Sigma_{22} = \text{Cov}(\mathbf{X}_t)$ ,  $\sigma_{21} = \text{Cov}(\mathbf{X}_t, \mathbf{y}_t)$ ,  $\sigma_{11} = \text{Var}(y_t)$ .)
- [3]  $\mathbf{X}_t$  is weakly exogenous with respect to  $\theta$ ,  $t = 1, 2, \dots, T$ .
- [4] No a priori information on  $\theta$ .
- [5]  $\text{Rank}(\mathbf{X}) = k$ ,  $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T)'$ ;  $T \times k$  data matrix, ( $T > k$ ).

(II) **Probability model**

$$\Phi = \left\{ D(y_t | \mathbf{X}_t; \theta) = \frac{1}{\sigma \sqrt{(2\pi)}} \exp \left[ \frac{1}{2\sigma^2} (y_t - \beta' \mathbf{x}_t)^2 \right], \theta \equiv (\beta, \sigma^2) \in \mathbb{R}^k \times \mathbb{R}_+, t \in \mathbb{T} \right\}.$$

- [6]
  - (i)  $D(y_t | \mathbf{X}_t; \theta)$  is normal;
  - (ii)  $E(y_t | \mathbf{X}_t = \mathbf{x}_t) = \beta' \mathbf{x}_t$  – linear in  $\mathbf{x}_t$ ;
  - (iii)  $\text{Var}(y_t | \mathbf{X}_t = \mathbf{x}_t) = \sigma^2$  – homoskedastic (free of  $\mathbf{x}_t$ );
- [7]  $\theta$  is time invariant.

(III) **Sampling model**

- [8]  $\mathbf{y} \equiv (y_1, \dots, y_T)'$  represents an independent sample sequentially drawn from  $D(y_t | \mathbf{X}_t; \theta)$ ,  $t = 1, 2, \dots, T$ .

An important point to note about the above specification is that the model is specified directly in terms of  $D(y_t | \mathbf{X}_t; \theta)$  making no assumptions about  $D(\mathbf{Z}_t; \psi)$ . For the specification of the linear regression model there is no need to make any assumptions related to  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$ . The problem, however, is that the additional generality gained by going directly to  $D(y_t | \mathbf{X}_t; \theta)$  is more apparent than real. Despite the fact that the assumption that  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is a NIID process is only sufficient (not necessary) for [6] to [8] above, it considerably enhances our understanding of econometric modelling in the context of the linear regression model. This is, firstly, because it is commonly easier in practice to judge the appropriateness of

probabilistic assumptions related to  $\mathbf{Z}_t$  rather than  $(y_t/\mathbf{X}_t = \mathbf{x}_t)$ ; and, secondly, in the context of misspecification analysis possible sources for the departures from the underlying assumptions are of paramount importance. Such sources can commonly be traced to departures from the assumptions postulated for  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  (see Chapters 21–22).

Before we discuss the above assumptions underlying the linear regression it is of some interest to compare the above specification with the standard textbook approach where the probabilistic assumptions are made in terms of the error term.

*Standard textbook specification of the linear regression model*

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

- (1)  $(\mathbf{u}/\mathbf{X}) \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_T)$ ;
- (2) no a priori information on  $(\boldsymbol{\beta}, \sigma^2)$ ;
- (3) rank  $(\mathbf{X}) = k$ .

Assumption (1) implies the orthogonality  $E(\mathbf{X}'_t u_t / \mathbf{X}_t = \mathbf{x}_t) = 0$ ,  $t = 1, 2, \dots, T$ , and assumptions [6] to [8] the probability and the sampling models respectively. This is because  $(\mathbf{y}/\mathbf{X})$  is a *linear function* of  $\mathbf{u}$  and thus normally distributed (see Chapter 15), i.e.

$$(\mathbf{y}/\mathbf{X}) \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_T). \quad (19.13)$$

As we can see, the sampling model assumption of independence is ‘hidden’ behind the form of the conditional covariance  $\sigma^2 I$ . Because of this the independence assumption and its implications are not clearly recognised in certain cases when the linear regression model is used in econometric modelling. As argued in Chapter 17, the sampling model of an independent sample is usually inappropriate when the observed data come in the form of aggregate economic time series. Assumptions (2) and (3) are identical to [4] and [5] above. The assumptions related to the parameters of interest  $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}, \sigma^2)$  and the weak exogeneity of  $\mathbf{X}$ , with respect to  $\boldsymbol{\theta}$  ([2] and [3] above) are not made in the context of the standard textbook specification. These assumptions related to the parametrisation of the statistical GM play a very important role in the context of the methodology proposed in Chapter 1 (see also Chapter 26). Several concepts such as weak exogeneity (see Section 19.3, below) and collinearity (see Sections 20.5–6) are only definable with respect to a given parametrisation. Moreover, the statistical GM is turned into an econometric model by reparametrisation, going from the statistical to the theoretical parameters of interest.

The most important difference between the specification [1]–[8] and (1)–(3), however, is the role attributed to the error term. In the context of the

latter the probabilistic and sampling model assumptions are made in terms of the error term not in terms of the observable random variables involved as in [1]–[8]. This difference has important implications in the context of misspecification testing (testing the underlying assumptions) and action thereof. The error term in the context of a statistical model as specified in the present book is by construction white-noise relative to a given information set  $\mathcal{D}_t \subseteq \mathcal{F}$ .

### 19.3 Discussion of the assumptions

#### [1] The systematic and non-systematic components

As argued in Chapter 17 (see also Chapter 26) the specification of a statistical model is based on the joint distribution of  $\mathbf{Z}_t$ ,  $t = 1, 2, \dots, T$ , i.e.

$$D(\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_T; \boldsymbol{\psi}) \equiv D(\mathbf{Z}; \boldsymbol{\psi}) \quad (19.14)$$

which includes the relevant sample and measurement information.

The specification of the linear regression model can be viewed as directly related to (14) and derived by ‘reduction’ using the assumptions of normality and IID. The *independence* assumption enables us to reduce  $D(\mathbf{Z}; \boldsymbol{\psi})$  into the product of the marginal distributions  $D(\mathbf{Z}_t; \boldsymbol{\psi}_t)$ ,  $t = 1, 2, \dots, T$ , i.e.

$$D(\mathbf{Z}; \boldsymbol{\psi}) = \prod_{t=1}^T D(\mathbf{Z}_t; \boldsymbol{\psi}_t) \quad (19.15)$$

The identical distribution enables us to deduce that  $\boldsymbol{\psi}_t = \boldsymbol{\psi}$  for  $t = 1, 2, \dots, T$ . The next step in the reduction is the following decomposition of  $D(\mathbf{Z}_t; \boldsymbol{\psi})$ :

$$D(\mathbf{Z}_t; \boldsymbol{\psi}) = D(y_t/\mathbf{X}_t; \boldsymbol{\psi}_1) \cdot D(\mathbf{X}_t; \boldsymbol{\psi}_2). \quad (19.16)$$

The normality assumption with  $\Sigma > 0$  and unrestricted enable us to deduce the weak exogeneity of  $\mathbf{X}_t$  relative to  $\boldsymbol{\theta}$ .

The choice of the relevant information set  $\mathcal{D}_t = \{\mathbf{X}_t = \mathbf{x}_t\}$  depends crucially on the NIID assumptions; if these assumptions are invalid the choice of  $\mathcal{D}_t$  will in general be inappropriate. Given this choice of  $\mathcal{D}_t$  the systematic and non-systematic components are defined by:

$$\mu_t = E(y_t/\mathbf{X}_t = \mathbf{x}_t), \quad u_t = y_t - E(y_t/\mathbf{X}_t = \mathbf{x}_t). \quad (19.17)$$

Under the NIID assumptions  $\mu_t$  and  $u_t$  take the particular forms:

$$\mu_t^* = \boldsymbol{\beta}' \mathbf{x}_t, \quad u_t^* = y_t - \boldsymbol{\beta}' \mathbf{x}_t. \quad (19.18)$$

Again, if the NIID assumptions are invalid then

$$\mu_t \neq \mu_t^* \quad \text{and} \quad E(\mu_t^* u_t^* / \mathbf{X}_t = \mathbf{x}_t) \neq 0 \quad (19.19)$$

(see Chapters 21–22).

## [2] The parameters of interest

As discussed in Chapter 17, the parameters in terms of which the statistical GM is defined constitute by definition the statistical parameters of interest and they represent a particular parametrisation of the unknown parameters of the underlying probability model. In the case of the linear regression model the parameters of interest come in the form of  $\theta \equiv (\beta, \sigma^2)$  where  $\beta \equiv \Sigma_{22}^{-1} \sigma_{21}$ ,  $\sigma^2 = \sigma_{11} - \sigma_{12} \Sigma_{22}^{-1} \sigma_{21}$ . As argued above the parametrisation  $\theta$  depends not only on  $D(\mathbf{Z}; \psi)$  but also on the assumptions of NIID. Any changes in  $\mathbf{Z}_t$  or/and the NIID assumptions will in general change the parametrisation.

## [3] Exogeneity

In the linear regression model we begin with  $D(y_t, \mathbf{X}_t; \psi)$  and then we concentrate exclusively on  $D(y_t / \mathbf{X}_t; \psi_1)$  where

$$D(y_t, \mathbf{X}_t / \psi) = D(y_t / \mathbf{X}_t; \psi_1) \cdot D(\mathbf{X}_t; \psi_2), \quad (19.20)$$

which implies that we choose to ignore the marginal distribution  $D(\mathbf{X}_t; \psi_2)$ . In order to be able to do that, this distribution must contain no information relevant for the estimation of the parameters of interest,  $\theta \equiv (\beta, \sigma^2)$ , i.e. the stochastic structure of  $\mathbf{X}_t$  must be irrelevant for any inference on  $\theta$ . Formalising this intuitive idea we say that:  $\mathbf{X}_t$  is *weakly exogenous* over the sample period for  $\theta$  if there exists a reparametrisation with  $\psi \equiv (\psi_1, \psi_2)$  such that:

- (i)  $\theta$  is a function of  $\psi_1$  ( $\theta = h(\psi_1)$ );
- (ii)  $\psi_1$  and  $\psi_2$  are variation free ( $(\psi_1, \psi_2) \in \Psi_1 \times \Psi_2$ ).

Variation free means that for any specific value  $\psi_2$  in  $\Psi_2$ ,  $\psi_1$  can take any other value in  $\Psi_1$  and vice versa. For more details on exogeneity see Engle, Hendry and Richard (1983). When the above conditions are not satisfied the marginal distribution of  $\mathbf{X}_t$  cannot be ignored because it contains relevant information for any inference on  $\theta$ .

[4] **No a priori information on  $\theta \equiv (\beta, \sigma^2)$**

This assumption is made at the outset in order to avoid imposing invalid testable restrictions on  $\theta$ . At this stage the only relevant interpretation of  $\theta$  is as statistical parameters, directly related to  $\psi_1$  in  $D(y_t/X_t; \psi_1)$ . As such no a priori information seems likely to be available for  $\theta$ . Such information is commonly related to the theoretical parameters of interest  $\xi$ . Before  $\theta$  is used to define  $\xi$ , however, we need to ensure that the underlying statistical model is well defined (no misspecification) in terms of the observed data chosen.

[5] **The observed data matrix  $X$  is of full rank**

For the data matrix  $X \equiv (x_1, x_2, \dots, x_T)', T \times k$ , we need to assume that

$$\text{rank}(X) = k, \quad k < T.$$

The need for this assumption is not at all obvious at this stage except perhaps as a sample equivalent to the assumption

$$\text{rank}(\Sigma_{22}) = k,$$

needed to enable us to define the parameters of interest  $\theta$ . This is because  $\text{rank}(X) = \text{rank}(X'X)$ , and

$$\frac{1}{T} \sum_t x_t x_t' = \frac{1}{T} (X'X)$$

can be seen as the sample moment equivalent to  $\Sigma_{22}$ .

[6] **Normality, linearity, homoskedasticity**

The assumption of normality of  $D(y_t, X_t; \psi)$  plays an important role in the specification as well as statistical analysis of the linear regression model. As far as specification is concerned, normality of  $D(y_t, X_t; \psi)$  implies

- (i)  $D(y_t/X_t; \theta)$  is normal (see Chapter 15);
- (ii)  $E(y_t/X_t = x_t) = \beta' x_t$ , a linear function of the observed value  $x_t$  of  $X_t$ ;
- (iii)  $\text{Var}(y_t/X_t = x_t) = \sigma^2$ , the conditional variance is free of  $x_t$ , i.e. homoskedastic.

Moreover, (i)–(iii) come very close to implying that  $D(y_t, X_t; \psi)$  is normal as well (see Chapter 24.2).

## [7] Parameter time-invariance

As far as the parameter invariance assumption is concerned we can see that it stems from the time invariance of the parameters of the distribution  $D(y_t; \mathbf{X}_t; \boldsymbol{\psi})$ ; that is, from the identically distributed (ID) component of the normal IID assumption related to  $\mathbf{Z}_t$ .

## [8] Independent sample

The assumption that  $\mathbf{y}$  is an independent sample from  $D(y_t; \mathbf{X}_t; \boldsymbol{\theta})$ ,  $t = 1, 2, \dots, T$ , is one of the most crucial assumptions underlying the linear regression model. In econometrics this assumption should be looked at very closely because most economic time series have a distinct time dimension (dependency) which cannot be modelled exclusively in terms of exogenous random variables  $\mathbf{X}_t$ . In such cases the non-random sample assumption (see Chapter 23) might be more appropriate.

## 19.4 Estimation

## (1) Maximum likelihood estimators

Let us consider the estimation of the linear regression model as specified by the assumptions [1]–[8] discussed above. Using the assumptions [6] to [8] we can deduce that the likelihood function for the model takes the form

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}, X) &= k(\mathbf{y}) \prod_{t=1}^T \left\{ \frac{1}{\sigma \sqrt{(2\pi)}} \exp \left\{ -\frac{1}{2\sigma^2} (y_t - \boldsymbol{\beta}' \mathbf{x}_t)^2 \right\} \right\} \\ &= k(\mathbf{y}) (2\pi)^{-T/2} (\sigma^2)^{-T/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{t=1}^T (y_t - \boldsymbol{\beta}' \mathbf{x}_t)^2 \right\}, \end{aligned} \quad (19.21)$$

$$\log L = c - \frac{T}{2} \log 2\pi - \frac{T}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=1}^T (y_t - \boldsymbol{\beta}' \mathbf{x}_t)^2, \quad (19.22)$$

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = -\frac{1}{2\sigma^2} (-2) \sum_{t=1}^T (y_t - \boldsymbol{\beta}' \mathbf{x}_t) \mathbf{x}'_t = 0, \quad (19.23)$$

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{T}{2} \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{t=1}^T (y_t - \boldsymbol{\beta}' \mathbf{x}_t)^2 = 0, \quad (19.24)$$

$$(3) \Rightarrow \sum_{t=1}^T y_t \mathbf{x}'_t - \boldsymbol{\beta}' \sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t = 0, \quad \text{i.e. } \hat{\boldsymbol{\beta}} = \left( \sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t \right)^{-1} \sum_{t=1}^T \mathbf{x}_t y_t, \quad (19.25)$$

$$(4) \Rightarrow \sigma^2 = \frac{1}{T} \sum_{t=1}^T (y_t - \hat{\beta}' \mathbf{x}_t)^2 \equiv \frac{1}{T} \sum_{t=1}^T u_t^2, \quad \text{in an obvious notation,} \quad (19.26)$$

are the maximum likelihood estimators (MLE's) of  $\beta$  and  $\sigma^2$ , respectively. If we were to write the statistical GM,  $y_t = \beta' \mathbf{x}_t + u$ ,  $t = 1, 2, \dots, T$ , in the matrix notation form

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \quad (19.27)$$

where  $\mathbf{y} \equiv (y_1, \dots, y_T)'$ ,  $T \times 1$ ,  $\mathbf{X} \equiv (\mathbf{x}_1, \dots, \mathbf{x}_T)'$ ,  $T \times k$ , and  $\mathbf{u} \equiv (u_1, \dots, u_T)'$ ,  $T \times 1$ , the MLE's take the more suggestive form

$$\blacktriangleright \quad \text{and for } \hat{\mathbf{u}} \equiv \mathbf{y} - \mathbf{X}\hat{\beta}, \quad \hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \quad \hat{\sigma}^2 = \frac{1}{T} \hat{\mathbf{u}}'\hat{\mathbf{u}}. \quad (19.28)$$

The information matrix  $\mathbf{I}_T(\theta)$  is defined by

$$\mathbf{I}_T(\theta) = E\left(\left(\frac{\partial \log L}{\partial \theta}\right)\left(\frac{\partial \log L}{\partial \theta}\right)'\right) = E\left(-\frac{\partial^2 \log L}{\partial \theta \partial \theta'}\right),$$

where the last equality holds under the assumption that  $D(y_t/\mathbf{X}_t; \theta)$  represents the 'true' probability model. In the above case

$$\begin{aligned} \frac{\partial^2 \log L}{\partial \beta \partial \beta} &= -\frac{1}{\sigma^2} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t' \equiv -\frac{1}{\sigma^2} (\mathbf{X}'\mathbf{X}), \quad \frac{\partial^2 \log L}{\partial \beta \partial \sigma^2} = -\frac{1}{\sigma^4} \sum_{t=1}^T \mathbf{x}_t u_t, \\ \frac{\partial^2 \log L}{\partial \sigma^4} &= \frac{T}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{t=1}^T u_t^2. \end{aligned} \quad (19.29)$$

Hence

$$\mathbf{I}_T(\theta) = \begin{pmatrix} \frac{\mathbf{X}'\mathbf{X}}{\sigma^2} & 0 \\ 0 & \frac{T}{2\sigma^4} \end{pmatrix} \quad \text{and} \quad [\mathbf{I}_T(\theta)]^{-1} = \begin{pmatrix} \sigma^2(\mathbf{X}'\mathbf{X})^{-1} & 0 \\ 0 & \frac{2\sigma^4}{T} \end{pmatrix}. \quad (19.30)$$

It is very important to remember that the expectation operator above is defined relative to the probability model  $D(y_t/\mathbf{X}_t; \theta)$ .

In order to get some idea as to what the above matrix notation formulae

look like let us consider these formulae for the simple model:

$$y_t = \beta_1 + \beta_2 x_t + u_t, \quad t = 1, 2, \dots, T. \quad (19.31)$$

$$\mathbf{y} \equiv \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_T \end{pmatrix}, \quad \mathbf{u} \equiv \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{pmatrix}, \quad \boldsymbol{\beta} \equiv \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix},$$

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} T & \sum_t x_t \\ \sum_t x_t & \sum_t x_t^2 \end{pmatrix}, \quad \mathbf{X}'\mathbf{y} \equiv \begin{pmatrix} \sum_t y_t \\ \sum_t x_t y_t \end{pmatrix},$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{\left( T \sum_t x_t^2 - \left( \sum_t x_t \right)^2 \right)} \begin{pmatrix} \sum_t x_t^2 & -\sum_t x_t \\ -\sum_t x_t & T \end{pmatrix},$$

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \bar{y} - \hat{\beta}_2 \bar{X} \\ \frac{\sum_t (x_t - \bar{x})(y_t - \bar{y})}{\sum_t (x_t - \bar{x})^2} \end{pmatrix},$$

$$\hat{\sigma}^2 = \frac{1}{T} \left( \sum_t (y_t - \bar{y})^2 - \frac{\left[ \sum_t (x_t - \bar{x})(y_t - \bar{y}) \right]^2}{\sum_t (x_t - \bar{x})^2} \right).$$

Compare these formulae with those of Chapter 18.

One very important feature of the MLE  $\hat{\boldsymbol{\beta}}$  above is that it preserves the original orthogonality between the systematic and non-systematic components

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{u}, \quad \boldsymbol{\mu} \perp \mathbf{u} \quad (19.32)$$

between the estimated systematic and non-systematic components in the form

$$\mathbf{y} = \hat{\boldsymbol{\mu}} + \hat{\mathbf{u}}, \quad \hat{\boldsymbol{\mu}} \perp \hat{\mathbf{u}}, \quad (19.33)$$

$\hat{\mu} = \mathbf{X}\hat{\beta}$ ,  $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\beta}$ , respectively. This is because

$$\hat{\mu} = \mathbf{P}_x \mathbf{y} \quad \text{and} \quad \hat{\mathbf{u}} = (\mathbf{I} - \mathbf{P}_x) \mathbf{y}, \quad (19.34)$$

where  $\mathbf{P}_x = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  is a symmetric ( $\mathbf{P}_x' = \mathbf{P}_x$ ), idempotent ( $\mathbf{P}_x^2 = \mathbf{P}_x$ ) matrix (i.e. it represents an orthogonal projection) and

$$\begin{aligned} E(\hat{\mu}\hat{\mathbf{u}}') &= E(\mathbf{P}_x \mathbf{y} \mathbf{y}' (\mathbf{I} - \mathbf{P}_x)) \\ &= E(\mathbf{P}_x \mathbf{y} \mathbf{u}' (\mathbf{I} - \mathbf{P}_x)), \quad \text{since } (\mathbf{I} - \mathbf{P}_x)\mathbf{y} = (\mathbf{I} - \mathbf{P}_x)\mathbf{u} \\ &= \mathbf{P}_x (\mathbf{I} - \mathbf{P}_x) \sigma^2, \quad \text{since } E(\mathbf{y}\mathbf{u}') = \sigma^2 \mathbf{I}_T \\ &= \mathbf{0}, \quad \text{since } \mathbf{P}_x (\mathbf{I} - \mathbf{P}_x) = \mathbf{0}. \end{aligned}$$

In other words, the systematic and non-systematic components were estimated in such a way so as to preserve the original orthogonality. Geometrically  $\mathbf{P}_x$  and  $(\mathbf{I} - \mathbf{P}_x)$  represent orthogonal projectors onto the subspace spanned by the columns of  $\mathbf{X}$ , say  $\mathcal{M}(\mathbf{X})$ , and into its orthogonal complement  $\mathcal{M}(\mathbf{X})^\perp$ , respectively. The systematic component was estimated by projecting  $\mathbf{y}$  onto  $\mathcal{M}(\mathbf{X})$  and the non-systematic component by projecting  $\mathbf{y}$  into  $\mathcal{M}(\mathbf{X})^\perp$ , i.e.

$$\mathbf{y} = \mathbf{P}_x \mathbf{y} + (\mathbf{I} - \mathbf{P}_x) \mathbf{y}. \quad (19.35)$$

Moreover, this orthogonality, which is equivalent to independence in this context, is passed over to the MLE's  $\hat{\beta}$  and  $\hat{\sigma}^2$  since  $\hat{\mu}$  is independent of  $\hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{y}'(\mathbf{I} - \mathbf{P}_x)\mathbf{y}$ , the residual sums of squares, because  $\mathbf{P}_x(\mathbf{I} - \mathbf{P}_x) = \mathbf{0}$  (see Q6, Chapter 15). Given that  $\hat{\mu} = \mathbf{X}\hat{\beta}$  and  $\hat{\sigma}^2 = (1/T)\hat{\mathbf{u}}'\hat{\mathbf{u}}$  we can deduce that  $\hat{\beta}$  and  $\hat{\sigma}^2$  are independent; see (E2) of Section 7.1.

Another feature of the MLE's  $\hat{\beta}$  and  $\hat{\sigma}^2$  worth noting is the suggestive similarity between these estimators and the parameters  $\beta, \sigma^2$ :

$$\beta \equiv \Sigma_{22}^{-1} \sigma_{21}, \quad \hat{\beta} = \left( \frac{\mathbf{X}'\mathbf{X}}{T} \right)^{-1} \left( \frac{\mathbf{X}'\mathbf{y}}{T} \right), \quad (19.36)$$

$$\begin{aligned} \sigma^2 &\equiv \sigma_{11} - \sigma_{12} \Sigma_{22}^{-1} \sigma_{21}, \\ \hat{\sigma}^2 &= \left( \frac{\mathbf{y}'\mathbf{y}}{T} \right) - \left( \frac{\mathbf{y}'\mathbf{X}}{T} \right) \left( \frac{\mathbf{X}'\mathbf{X}}{T} \right)^{-1} \left( \frac{\mathbf{X}'\mathbf{y}}{T} \right). \end{aligned} \quad (19.37)$$

Looking at these formulae we can see that the MLE's of  $\beta$  and  $\sigma^2$  can be derived by substituting the sample moment equivalents to the population moments:

$$\Sigma_{22} \equiv \frac{1}{T} (\mathbf{X}'\mathbf{X}), \quad \sigma_{21} \equiv \frac{1}{T} \mathbf{X}'\mathbf{y}, \quad \sigma_{11} \equiv \frac{1}{T} \mathbf{y}'\mathbf{y}. \quad (19.38)$$

Using the orthogonality of the estimated components  $\hat{\mu}$  and  $\hat{\mathbf{u}}$  we could

decompose the variation in  $\mathbf{y}$  as measured by  $\mathbf{y}'\mathbf{y}$  into

$$\mathbf{y}'\mathbf{y} = \hat{\boldsymbol{\mu}}'\hat{\boldsymbol{\mu}} + \hat{\mathbf{u}}'\hat{\mathbf{u}} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}'\hat{\mathbf{u}}. \quad (19.39)$$

Using this decomposition we could define the sample equivalent to the *multiple correlation coefficient* (see Chapter 15) to be

$$\tilde{R}^2 = \frac{\hat{\boldsymbol{\mu}}'\hat{\boldsymbol{\mu}}}{\mathbf{y}'\mathbf{y}} = \frac{\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}{\mathbf{y}'\mathbf{y}} = 1 - \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{\mathbf{y}'\mathbf{y}}. \quad (19.40)$$

This represents the ratio of the variation ‘explained’ by  $\hat{\boldsymbol{\mu}}$  over the total variation and can be used as a *measure of goodness of fit* for the linear regression model. A similar measure of fit can be constructed using the decomposition of  $\mathbf{y}$  around its mean  $\bar{\mathbf{y}}$ , that is

$$(\mathbf{y}'\mathbf{y} - T\bar{\mathbf{y}}^2) = (\hat{\boldsymbol{\mu}}'\hat{\boldsymbol{\mu}} - T\bar{\mathbf{y}}^2) + \hat{\mathbf{u}}'\hat{\mathbf{u}}, \quad (19.41)$$

denoted as

$$\text{TSS} = \underset{\text{(total)}}{\text{ESS}} + \underset{\text{(explained)}}{\text{RSS}} + \underset{\text{(residual)}}{\text{RSS}}, \quad (19.42)$$

where SS stands for sums of squares. The multiple correlation coefficient in this case takes the form

$$\hat{R}^2 = \frac{\hat{\boldsymbol{\mu}}'\hat{\boldsymbol{\mu}} - T\bar{\mathbf{y}}^2}{(\mathbf{y}'\mathbf{y} - T\bar{\mathbf{y}}^2)} = 1 - \frac{\text{RSS}}{\text{TSS}}. \quad (19.43)$$

Note that  $R^2$  was used in Chapter 15 to denote the population multiple correlation coefficient but in the econometrics literature  $R^2$  is also used to denote  $\tilde{R}^2$  and  $\hat{R}^2$ .

Both of the above measures of ‘goodness of fit’,  $\tilde{R}^2$  and  $\hat{R}^2$ , have variously been defined to be the sample multiple correlation coefficient in the econometric literature. Caution should be exercised when reading different textbooks because  $\tilde{R}^2$  and  $\hat{R}^2$  have different properties. For example,  $0 < \tilde{R}^2 < 1$ , but no such restriction exists for  $\hat{R}^2$  unless one of the regressors in  $\mathbf{X}_t$  is the constant term. On the role of the constant term see Appendix 19.1.

One serious objection to the use of  $\hat{R}^2$  as a goodness-of-fit measure is the fact that as the number  $k$  of regressors increases,  $\hat{R}^2$  increases as well irrespective of whether the regressors are relevant or not. For this reason a ‘corrected’ goodness-of-fit measure is defined by

$$\bar{R}^2 = 1 - \frac{\left( \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{T-k} \right)}{\left( \frac{\mathbf{y}'\mathbf{y} - T\bar{\mathbf{y}}^2}{T-1} \right)} = 1 - \left( \frac{T-1}{T-k} \right) (1 - \hat{R}^2). \quad (19.44)$$

The correction is the division of the statistics involved by their corresponding degrees of freedom; see Theil (1971).

## (2) An empirical example

In order to illustrate some of the concepts and results introduced so far let us consider estimating a transactions demand for money. Using the simplest form of a demand function we can postulate the theoretical model:

$$M^D = h(Y, P, I), \quad (19.45)$$

where  $M^D$  is the transactions demand for money,  $Y$  is income,  $P$  is the price level and  $I$  is the short-run interest rate referring to the opportunity cost of holding transactions money. Assuming a multiplicative form for  $h(\cdot)$  the demand function takes the form

$$M^D = AY^{\alpha_1}P^{\alpha_2}I^{\alpha_3} \quad (19.46)$$

or

$$\ln M^D = \alpha_0 + \alpha_1 \ln Y + \alpha_2 \ln P + \alpha_3 \ln I, \quad (19.47)$$

where  $\ln$  stands for  $\log_e$  and  $\alpha_0 = \ln A$ .

For expositional purposes let us adopt the commonly accepted approach to econometric modelling (see Chapter 1) in an attempt to highlight some of the problems associated with it. If we were to ignore the discussion on econometric modelling in Chapter 1 and proceed by using the usual ‘textbook’ approach the next step is to transform the theoretical model to an econometric model by adding an error term, i.e. the econometric model is

$$m_t = \alpha_0 + \alpha_1 y_t + \alpha_2 p_t + \alpha_3 i_t + u_t, \quad (19.48)$$

where  $m_t = \ln M_t$ ,  $y_t = \ln Y_t$ ,  $p_t = \ln P_t$ ,  $i_t = \ln I_t$  and  $u_t \sim NI(0, \sigma^2)$ . Choosing some observed data series corresponding to the theoretical variables,  $M$ ,  $Y$ ,  $P$  and  $I$ , say:

$\tilde{M}_t$  –  $M$  1 money stock;

$\tilde{Y}_t$  – real consumers’ expenditure;

$\tilde{P}_t$  – implicit price deflator of  $\tilde{Y}_t$ ;

$\tilde{i}_t$  – interest rate on 7 days’ deposit account (see Chapter 17 and its appendix for these data series),

respectively, the above equation can be transformed into the linear regression statistical GM:

$$\tilde{m}_t = \beta_0 + \beta_1 \tilde{y}_t + \beta_2 \tilde{p}_t + \beta_3 \tilde{i}_t + u_t. \quad (19.49)$$

Estimation of this equation for the period 1963*i*–1982*iv* ( $T=80$ ) using

quarterly seasonally adjusted (for convenience) data yields

$$\hat{\beta} = \begin{pmatrix} 2.896 \\ 0.690 \\ 0.865 \\ -0.055 \end{pmatrix}$$

$$s^2 = 0.00155, \quad \hat{R}^2 = 0.9953, \quad \bar{R}^2 = 0.9951,$$

$$TSS = 24.954, \quad ESS = 24.836, \quad RSS = 0.118.$$

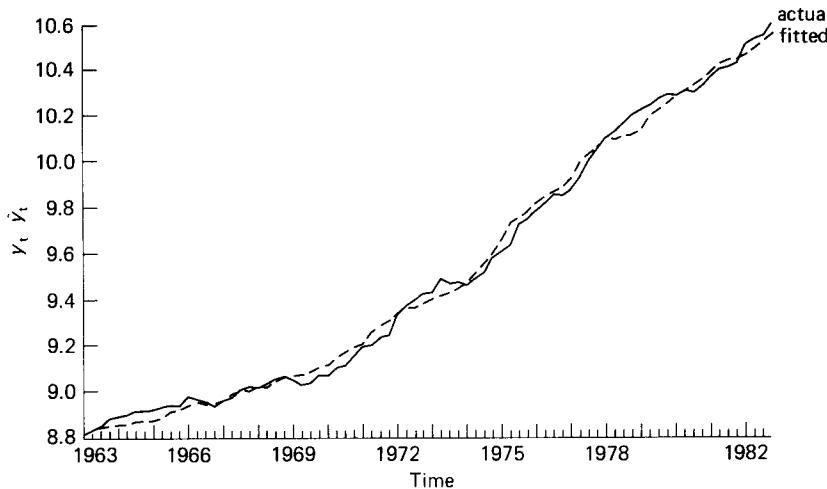
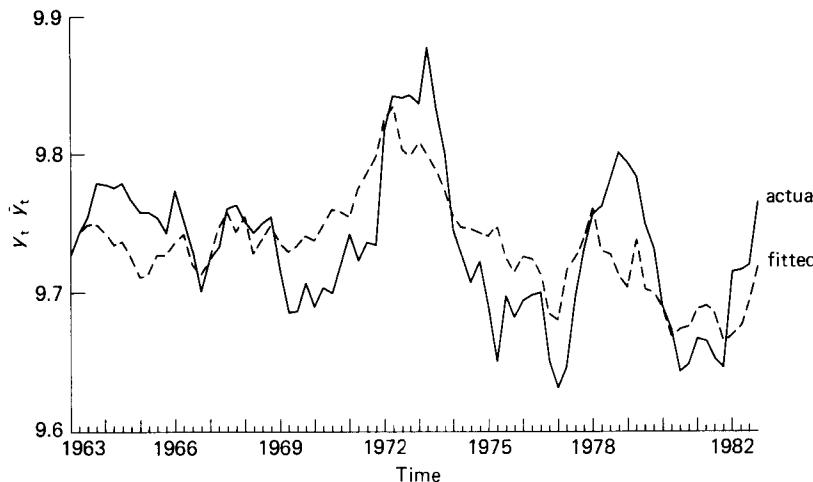
That is, the estimated equation takes the form

$$\hat{m}_t = 2.896 + 0.690\hat{y}_t + 0.865\hat{p}_t - 0.055\hat{i}_t + \hat{u}_t. \quad (19.50)$$

The danger at this point is to get carried away and start discussing the plausibility of the sign and size of the estimated 'elasticities' (?). For example, we might be tempted to argue that the estimated 'elasticities' have both a 'correct' sign and the size assumed on a priori grounds. Moreover, the 'goodness of fit' measures show that we explain 99.5% of the variation. Taken together these results 'indicate' that (50) is a good empirical model for the transactions demand for money. This, however, will be rather premature in view of the fact that before any discussion of a priori economic theory information we need to have a *well-defined* estimated statistical model which at least summarises the sample information adequately. Well defined in the present context refers to ensuring that the assumptions underlying the statistical model adopted are valid. This is because any formal testing of a priori restrictions could only be based on the underlying assumptions which when invalid render the testing procedures incorrect.

Looking at the above estimated equation in view of the discussion of econometric modelling in Chapter 1 several objections might be raised:

- (i) The observed data chosen do not correspond one-to-one to the theoretical variables and thus the estimable model might be different from the theoretical model (see Chapter 23).
- (ii) The sampling model of an independent sample seems questionable in view of the time paths of the observed data (see Fig. 17.1).
- (iii) The high  $\hat{R}^2$  (and  $\bar{R}^2$ ) is due to the fact that the data series for  $M_t$  and  $P_t$  have a very similar time trend (see Fig. 17.1(a) and (c)). If we look at the time path of the actual ( $y_t$ ) and fitted ( $\hat{y}_t$ ) values we notice that  $\hat{y}_t$  'tracks' (explains) largely the trend and very little else (see Fig. 19.1). An obvious way to get some idea of the trend's contribution in  $\hat{R}^2$  is to subtract  $p_t$  from both sides of the money equation in an attempt to 'detrend' the dependent variable.

Fig. 19.1. Actual  $y_t = \ln M_t$  and fitted  $\hat{y}_t$  from (19.50).Fig. 19.2. Actual  $y_t = \ln (M/P)_t$  and fitted  $\hat{y}_t$  from (19.51).

In Fig. 19.2 the actual and fitted values of the 'largely' detrended dependent variable  $(m_t - p_t)$  are shown to emphasise the point. The new regression equation yielded

$$(m_t - p_t) = 2.896 + 0.690y_t - 0.135p_t - 0.055i_t + \hat{u}_t,$$

$$\hat{R}^2 = 0.468, \quad \bar{R}^2 = 0.447, \quad s^2 = 0.00155. \quad (19.51)$$

Looking at this estimated equation we can see that the coefficients of the constant,  $y_t$  and  $i_t$ , are identical in value to the previous estimated equation. The estimated coefficient of  $p_t$  is, as expected, one minus the original estimate and the  $s^2$  is identical for both estimated equations. These suggest that the two estimated equations are identical as far as the estimated coefficients are concerned. This is a special case of a more general result related to arbitrary linear combinations of the  $x_{it}$ 's subtracted from both sides of the statistical GM. In order to see this let us subtract  $\gamma' \mathbf{x}_t$  from both sides of the statistical GM:

$$y_t - \gamma' \mathbf{x}_t = (\beta' - \gamma') \mathbf{x}_t + u_t \quad (19.52)$$

or

$$y_t^* = \beta'^* \mathbf{x}_t + u_t,$$

in an obvious notation. It is easy to see that the non-systematic component as well as  $\sigma^2$  remain unchanged. Moreover, in view of the equality

$$\hat{\mathbf{u}}^* = \hat{\mathbf{u}}, \quad (19.53)$$

where

$$\hat{\mathbf{u}}^* = \mathbf{y}^* - \mathbf{X}\hat{\beta}^*, \quad \hat{\beta}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^* = \hat{\beta} - \gamma,$$

we can deduce that

$$s^2 = \frac{1}{T-k} \hat{\mathbf{u}}^{*'} \hat{\mathbf{u}}^* = \frac{1}{T-k} \hat{\mathbf{u}}' \hat{\mathbf{u}}. \quad (19.54)$$

On the other hand,  $\hat{R}^2$  is not invariant to this transformation because

$$\hat{R}^{*2} = 1 - \frac{\hat{\mathbf{u}}' \hat{\mathbf{u}}}{\mathbf{y}^{*'} \mathbf{y}^* - T\bar{y}^{*2}} \neq \hat{R}^2. \quad (19.55)$$

As we can see, the  $\hat{R}^2$  of the 'detrended' dependent variable equation is less than half of the original. This confirms the suggestion that the trend in  $p_t$  contributes significantly to the high value of the original  $\hat{R}^2$ . It is important to note at this stage that trending data series can be a problem when the asymptotic properties of the MLE's are used uncritically (see sub-section (4) below).

### (3) Properties of the MLE $\hat{\theta} \equiv (\hat{\beta}, \hat{\sigma}^2)$ – finite sample

In order to decide whether the MLE  $\hat{\theta}$  is a 'good' estimator of  $\theta$  we need to consider its properties. The finite sample properties (see Chapters 12 and 13) will be considered first and then the asymptotic properties.

$\hat{\theta}$  being a MLE satisfies certain properties by definition:

- (1) For a Borel function  $h(\cdot)$  the MLE of  $h(\theta)$  is  $h(\hat{\theta})$ . For example, the MLE of  $\log(\beta' \beta)$  is  $\log(\hat{\beta}' \hat{\beta})$ .

- (2) If a minimal sufficient statistic  $\tau(\mathbf{y})$  exists, then  $\hat{\theta}$  must be a function of it.

Using the Lehmann–Scheffe theorem (see Chapter 12) we can deduce that the values of  $\mathbf{y}_0$  for which the ratio

$$\frac{D(\mathbf{y}/\mathbf{X}; \boldsymbol{\theta})}{D(\mathbf{y}_0/\mathbf{X}; \boldsymbol{\theta})} = \frac{(2\pi\sigma^2)^{-T/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}}{(2\pi\sigma^2)^{-T/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}_0 - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y}_0 - \mathbf{X}\boldsymbol{\beta})\right\}} \quad (19.56)$$

is independent of  $\boldsymbol{\theta}$ , are  $\mathbf{y}'_0\mathbf{y}_0 = \mathbf{y}'\mathbf{y}$  and  $\mathbf{X}'\mathbf{y}_0 = \mathbf{X}'\mathbf{y}$ . Hence, the minimal sufficient statistic is  $\tau(\mathbf{y}) \equiv (\tau_1(\mathbf{y}), \tau_2(\mathbf{y})) = (\mathbf{y}'\mathbf{y}, \mathbf{X}'\mathbf{y})$  and  $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\tau_2(\mathbf{y})$ ,  $\hat{\sigma}^2 = (1/T)(\tau_1(\mathbf{y}) - \tau'_2(\mathbf{y})(\mathbf{X}'\mathbf{X})^{-1}\tau_2(\mathbf{y}))$  are indeed functions of  $\tau(\mathbf{y})$ .

In order to discuss any other properties of the MLE  $\hat{\theta}$  of  $\boldsymbol{\theta}$  we need to derive the sampling distribution of  $\hat{\theta}$ . Given that  $\hat{\boldsymbol{\beta}}$  and  $\hat{\sigma}^2$  are independent we can consider them separately.

*The distribution of  $\hat{\boldsymbol{\beta}}$*

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \equiv \mathbf{L}\mathbf{y}, \quad (19.57)$$

where  $\mathbf{L} \equiv (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  is a  $k \times T$  matrix of known constants. That is,  $\hat{\boldsymbol{\beta}}$  is a *linear function* of the normally distributed random vector  $\mathbf{y}$ . Hence

$$\hat{\boldsymbol{\beta}} \sim N(\mathbf{L}\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{L}\mathbf{L}') \quad \text{from N1, Chapter 15,}$$

or

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}). \quad (19.58)$$

From the sampling distribution (58) we can deduce the following properties for  $\hat{\boldsymbol{\beta}}$ :

- (3(i))  $\hat{\boldsymbol{\beta}}$  is an *unbiased estimator* of  $\boldsymbol{\beta}$  since  $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ , i.e. the sampling distribution of  $\hat{\boldsymbol{\beta}}$  has mean equal to  $\boldsymbol{\beta}$ .
- (4(i))  $\hat{\boldsymbol{\beta}}$  is a fully *efficient estimator* of  $\boldsymbol{\beta}$  since  $\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ , i.e.  $\text{Cov}(\hat{\boldsymbol{\beta}})$  achieves the Cramer–Rao lower bound; see (30) above.

*The distribution of  $\hat{\sigma}^2$*

$$\hat{\sigma}^2 = \frac{1}{T}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \frac{1}{T}\hat{\mathbf{u}}'\hat{\mathbf{u}} = \frac{1}{T}\mathbf{u}'\mathbf{M}_x\mathbf{u}, \quad (19.59)$$

where  $\mathbf{M}_x = \mathbf{I} - \mathbf{P}_x$ . From (Q2) of Chapter 15 we can deduce that

$$\left(\frac{T\hat{\sigma}^2}{\sigma^2}\right) \sim \chi^2(\text{tr } \mathbf{M}_x), \quad (19.60)$$

where  $\text{tr } \mathbf{M}_x$  refers to the trace of  $\mathbf{M}_x$  ( $\text{tr } \mathbf{A} = \sum_{i=1}^n a_{ii}$ ,  $\mathbf{A}: n \times n$ ),

$$\begin{aligned}\text{tr } \mathbf{M}_x &= \text{tr } \mathbf{I} - \text{tr } \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \quad (\text{since } \text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr } \mathbf{A} + \text{tr } \mathbf{B}) \\ &= T - \text{tr}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X}) \quad (\text{since } \text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})) \\ &= T - k.\end{aligned}$$

Hence, we can deduce that

$$\left( \frac{T\hat{\sigma}^2}{\sigma^2} \right) \sim \chi^2(T - k). \quad (19.61)$$

Intuitively we can explain this result as saying that  $(\mathbf{u}'\mathbf{M}_x\mathbf{u})/\sigma^2$  represents the summation of the squares of  $T - k$  independent standard normal components.

Using (61) we can deduce that

$$E\left(\frac{T\hat{\sigma}^2}{\sigma^2}\right) = T - k \quad \text{and} \quad \text{Var}\left(\frac{T\hat{\sigma}^2}{\sigma^2}\right) = 2(T - k)$$

(see Appendix 6.1). These results imply that

$$E(\hat{\sigma}^2) = \frac{T - k}{T} \sigma^2 \neq \sigma^2,$$

$$\text{Var}(\hat{\sigma}^2) = \frac{2(T - k)}{T^2} \sigma^4 > \frac{2(T - k)^2 \sigma^4}{T^3} = \text{Cramer-Rao lower bound.}$$

That is:

- (3(ii))  $\hat{\sigma}^2$  is a *biased* estimator of  $\sigma^2$ ; and
- (4(ii))  $\hat{\sigma}^2$  is *not a fully efficient* estimator of  $\sigma^2$ .

However, 3(ii) implies that for

$$s^2 = \frac{1}{T - k} \hat{\mathbf{u}}' \hat{\mathbf{u}} \quad (19.62)$$

$$(T - k) \frac{s^2}{\sigma^2} \sim \chi^2(T - k) \quad (19.63)$$

and  $E(S^2) = \sigma^2$ ,  $\text{Var}(s^2) = (2\sigma^4)/(T - k) > (2\sigma^4)/T$  – Cramer–Rao bound. That is,  $s^2$  is an unbiased estimator of  $\sigma^2$ , although it does not quite achieve the Cramer–Rao lower bound given by the information matrix (30) above. It turns out, however, that no other unbiased estimator of  $\sigma^2$  achieves that bound and among such estimators  $s^2$  has minimum variance. In statistical inference relating to the linear regression model  $s^2$  is preferred to  $\hat{\sigma}^2$  as an estimator of  $\sigma^2$ .

The sampling distributions of the estimators  $\hat{\beta}$  and  $s^2$  involve the

unknown parameters  $\beta$  and  $\sigma^2$ . In practice the covariance of  $\hat{\beta}$  is needed to assess the ‘accuracy’ of the estimates. From the above analysis it is known that

$$\text{Cov}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}, \quad (19.64)$$

which involves the unknown parameter  $\sigma^2$ . The obvious way to proceed in such a case is to use the estimated covariance

$$\hat{\text{Cov}}(\hat{\beta}) = s^2(\mathbf{X}'\mathbf{X})^{-1}. \quad (19.65)$$

The diagonal elements of  $\hat{\text{Cov}}(\hat{\beta})$  refer to the estimated variances of the coefficient estimates and they are usually reported in standard deviation form underneath the coefficient estimates. In the case of the above example results are usually reported in the form

$$\begin{aligned} m_t &= 2.896 + 0.690y_t + 0.865p_t - 0.055i_t + \hat{u}_t, \\ &(1.034) \quad (0.105) \quad (0.020) \quad (0.013) \quad (0.039) \end{aligned} \quad (19.66)$$

$$R^2 = 0.9953, \quad \bar{R}^2 = 0.9951, \quad s = 0.0393, \quad \log L = 147.412, \quad T = 80.$$

Note that having made the distinction between theoretical variables and observed data the upper tildas denoting observed data have been dropped for notational convenience and  $R^2$  is used instead of  $\bar{R}^2$  in order to comply with the traditional econometric notation.

#### (4) Properties of the MLE $\hat{\theta}_T \equiv (\hat{\beta}, \hat{\sigma}^2)$ – asymptotic

An obvious advantage of MLE’s is the fact that under certain regularity conditions they satisfy a number of desirable asymptotic properties (see Chapter 13).

$$(1) \quad \text{Consistency } (\hat{\theta}_T \xrightarrow{P} \theta)$$

Looking at the information matrix (30) we can deduce that:

$$(i) \quad \hat{\sigma}^2 \text{ is a consistent estimator of } \sigma^2, \text{ i.e.}$$

$$\lim_{T \rightarrow \infty} \Pr(|\hat{\sigma}^2 - \sigma^2| < \varepsilon) = 1,$$

since  $\text{MSE}(\hat{\sigma}^2) \rightarrow 0$  as  $T \rightarrow \infty$ ; and

$$(ii) \quad \text{if} \quad \lim_{T \rightarrow \infty} (\mathbf{X}'\mathbf{X})^{-1} \equiv \lim_{T \rightarrow \infty} \left( \sum_t \mathbf{x}_t \mathbf{x}_t' \right)^{-1} = 0, \quad (19.67)$$

$$\lim_{T \rightarrow \infty} \Pr(|\hat{\beta} - \beta| < \varepsilon) = 1, \text{ i.e. } \hat{\beta} \text{ is a consistent estimator of } \beta.$$

Note that the above restriction is equivalent to assuming that  $\mathbf{c}'(\mathbf{X}'\mathbf{X})\mathbf{c} \rightarrow \infty$  for any non-zero vector  $\mathbf{c}$  (see Anderson and Taylor (1979)).

The above condition is needed because it ensures that

$$\text{Cov}(\hat{\boldsymbol{\beta}}) \rightarrow 0 \quad \text{as } T \rightarrow \infty \quad (\text{see Chapter 12}).$$

$$(2) \quad \text{Asymptotic normality } (\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}) \xrightarrow{d} N(0, \mathbf{I}_\infty(\boldsymbol{\theta})^{-1}))$$

In order of  $\hat{\boldsymbol{\theta}}$  to be asymptotically normal we need to ensure that

$$\mathbf{I}_\infty(\boldsymbol{\theta}) = \lim_{T \rightarrow \infty} \left( \frac{1}{T} \mathbf{I}_T(\boldsymbol{\theta}) \right)$$

exists and is non-singular. Given that  $I_\infty(\sigma^2) = 1/2\sigma^4$  we can deduce that

$$(i) \quad \sqrt{T}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{d} N(0, 2\sigma^4). \quad (19.68)$$

Moreover, if  $\lim_{T \rightarrow \infty} (\mathbf{X}'\mathbf{X}/T) = \mathbf{Q}_x$  is bounded and non-singular then

$$(ii) \quad \sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(0, \sigma^2 \mathbf{Q}_x^{-1}). \quad (19.69)$$

From the asymptotic normal distribution of  $\hat{\boldsymbol{\beta}}$  and  $\hat{\sigma}^2$  we can deduce asymptotic unbiasedness as well as asymptotic efficiency (see Chapter 13).

$$(3)* \quad \text{Strong consistency } (\hat{\boldsymbol{\theta}}_T \xrightarrow{\text{a.s.}} \boldsymbol{\theta})$$

$$(i) \quad \hat{\sigma}^2 \text{ is a strongly consistent estimator of } \sigma^2 \quad (\hat{\sigma}^2 \xrightarrow{\text{a.s.}} \sigma^2), \text{ i.e.}$$

$$Pr\left(\lim_{T \rightarrow \infty} \hat{\sigma}^2 = \sigma^2\right) = 1. \quad (19.70)$$

Let us prove this result for  $s^2$  and then extend it to  $\hat{\sigma}^2$ ,

$$(s^2 - \sigma^2) = \mathbf{u}' \left( \frac{\mathbf{I} - \mathbf{P}_x}{T-k} \right) \mathbf{u} - \sigma^2 = \frac{1}{T-k} \sum_{t=1}^{T-k} (w_t^2 - \sigma^2), \quad (19.71)$$

where  $\mathbf{w} \equiv (w_1, w_2, \dots, w_{T-k})$ ,  $\mathbf{w} = \mathbf{H}'\mathbf{u}$ ,  $\mathbf{H}$  being an orthogonal matrix, such that

$$\mathbf{H}'(\mathbf{I} - \mathbf{P}_x)\mathbf{H} = \text{diag}\left(\overbrace{1, 1, \dots, 1}^{T-k}, 0, \dots, 0\right).$$

Note that  $\mathbf{w} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{T-k})$  because  $\mathbf{H}$  is orthogonal. Since  $E(w_t^2 - \sigma^2) = 0$  and  $E(w_t^2 - \sigma^2)^2 = 2\sigma^4 < \infty$ , we can apply Kolmogorov's SLLN (see

Chapter 9) to deduce that

$$\frac{1}{T} \sum_t (w_t^2 - \sigma^2) \xrightarrow{\text{a.s.}} 0, \quad \text{or} \quad s^2 \xrightarrow{\text{a.s.}} \sigma^2. \quad (19.72)$$

Using the fact that

$$(\hat{\sigma}^2 - \sigma^2) = \frac{1}{T} \left( \sum_{t=1}^{T-k} (w_t^2 - \sigma^2) \right) + \frac{k}{T} \sigma^2$$

and the last term goes to zero as  $T \rightarrow \infty$ ,  $\hat{\sigma}^2 \xrightarrow{\text{a.s.}} \sigma^2$ .

(ii)  $\hat{\beta}$  is a strongly consistent estimator of  $\beta$  ( $\hat{\beta} \xrightarrow{\text{a.s.}} \beta$ ) if

- (1)  $|x_{it}| < C$ ,  $i = 1, 2, \dots, k$ ,  $t = 1, 2, \dots, T$ ,  $C$ -constant; and
- (2)  $\left( \frac{\mathbf{X}'\mathbf{X}}{T} \right)$  is non-singular for all  $T$ . (19.73)

$$(\hat{\beta} - \beta) = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{u} = \left( \sum_t \frac{\mathbf{x}_t \mathbf{x}_t'}{T} \right)^{-1} \frac{1}{T} \sum_t \mathbf{x}_t u_t. \quad (19.74)$$

Since  $E(\mathbf{x}_t u_t) = \mathbf{0}$  and  $E(x_{it} u_t)^2 = x_{it}^2 \sigma^2 < \infty$ , we can apply Kolmogorov's SLLN to deduce that

$$\left( \sum_t \frac{\mathbf{x}_t \mathbf{x}_t'}{T} \right)^{-1} \frac{1}{T} \sum_t \mathbf{x}_t u_t \xrightarrow{\text{a.s.}} 0. \quad (19.75)$$

Note that (1) implies that  $|x_{it} x_{js}| < C^*$  for  $i = 1, 2, \dots, k$ ,  $t, s = 1, 2, \dots, T$ ,  $C^*$  being a constant.

It is important to note that the assumption

$$\lim_{T \rightarrow \infty} \left( \frac{1}{T} \sum_t \mathbf{x}_t \mathbf{x}_t' \right) = \mathbf{Q}_x < \infty \quad \text{and non-singular,}$$

needed for the asymptotic normality of  $\hat{\beta}$ , is a rather restrictive assumption because it excludes regressors such as  $x_{it} = t$ ,  $t = 1, 2, \dots, T$ , since

$$\sum_t x_{it}^2 = \frac{1}{6} T(T+1)(2T+1) = O(T^3) \quad (19.76)$$

(see Chapter 10), and  $\lim_{T \rightarrow \infty} [(1/T) \sum_t x_{it}^2] = \infty$ .

The problem arises because the order of magnitude of  $\sum_t x_{it}^2$  is higher than  $O(T)$  and hence it goes to infinity much quicker than  $T$ . The obvious way out is to change the factor  $\sqrt{T}$  in  $\sqrt{T}(\hat{\beta} - \beta)$  so as to achieve the same rate of convergence. For example, in the above case of the regressor  $x_{it} = t$  we need to use  $\sqrt{T^3}$  in order to ensure that  $\lim_{T \rightarrow \infty} [(1/T^3) \sum_t x_{it}^2] < \infty$ . The question which naturally arises is how can we generalise this particular result? One of the most important results in relation to orders of magnitude

is that every random variable with bounded variance is ‘as big as its standard deviation’, i.e. if  $\text{Var}(Z_i) = \sigma_i^2 < \infty$  then  $Z_i = O_p(\sigma_i)$  (see Chapter 10). Using this result we can weaken the above asymptotic normality result (69) to the following:

*Lemma*

*For the linear regression model as specified in Section 19.2 above let*

$$\mathbf{A}_T = \sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t' \quad \text{and} \quad \mathbf{Q}_T = \mathbf{D}_T^{-1} \mathbf{A}_T \mathbf{D}_T^{-1}$$

where

$$\mathbf{D}_T = \text{diag}(\sqrt{(a_{11}^{(T)})}, \sqrt{(a_{22}^{(T)})}, \dots, \sqrt{(a_{kk}^{(T)})})$$

$$\mathbf{A}_T = [a_{ij}^{(T)}], \quad i, j = 1, 2, \dots, k,$$

if

$$(i) \quad a_{ii}^{(T)} \rightarrow \infty \quad \text{as } T \rightarrow \infty$$

(information increases with  $T$ );

$$(ii) \quad \frac{x_{iT+1}^2}{a_{ii}^{(T)}} \rightarrow 0, \quad i = 1, 2, \dots, k \quad \text{as } T \rightarrow \infty$$

(no individual observation dominates the summation);

$$(iii) \quad \lim_{T \rightarrow \infty} \mathbf{Q}_T = \mathbf{Q} < \infty \quad \text{and non-singular,}$$

then

$$\mathbf{D}_T(\hat{\beta} - \beta) \xrightarrow{z} N(\mathbf{0}, \sigma^2 \mathbf{Q}^{-1})$$

(see Anderson (1971)).

## 19.5 Specification testing

*Specification testing* refers to tests based on the assumption of correct specification. That is, tests within the framework specified by the statistical model in question. On the other hand, *misspecification testing* refers to testing outside this specified framework (see Mizon (1977)).

Logically, misspecification tests precede specification tests because, unless we ensure that the assumptions underlying the statistical model in question are valid (misspecification tests), specification tests (based on the validity of these assumptions) can be very misleading. For expositional purposes the estimated money equation of Section 19.4 will be used to illustrate some of the specification tests discussed below. It must be

emphasised, however, that the results of these tests should not be taken seriously in view of the fact that various misspecifications are suspected (indeed, confirmed in Chapters 20–22). In practice, misspecification tests are used first to ensure that the estimated equation represents a well-defined estimated statistical GM and then we go on to apply specification tests. This is because specification tests are based on the assumption of ‘correct specification’.

Within the Neyman–Pearson hypothesis-testing framework a test is defined when the following components (see Chapter 14) are specified:

- (i) the test statistic  $\tau(\mathbf{y})$ ;
- (ii) the size  $\alpha$  of the test;
- (iii) the distribution of  $\tau(\mathbf{y})$  under  $H_0$ ;
- (iv) the rejection (or acceptance) region;
- (v) the distribution of  $\tau(\mathbf{y})$  under  $H_1$ .

### (1) Tests relating to $\sigma^2$

As argued in Chapter 14, the problem of setting up ‘good’ tests for unknown parameters is largely an exercise in finding an appropriate pivot related to the unknown parameter(s) in question. In the case of  $\sigma^2$  the likeliest candidate must be the quantity

$$(T-k) \frac{s^2}{\sigma^2} \sim \chi^2(T-k). \quad (19.77)$$

Let us consider the null hypothesis  $H_0: \sigma^2 = \sigma_0^2$  ( $\sigma_0^2$ -known) against the alternative hypothesis  $H_1: \sigma^2 > \sigma_0^2$ . Common-sense suggests that if the estimated  $\sigma^2$  is much bigger than  $\sigma_0^2$  we will be inclined to reject  $H_0$  in favour of  $H_1$ , i.e. for  $s^2 > c$  where  $c$  is some constant considered to be ‘big enough’, we reject  $H_0$ . In order to determine  $c$  we have to relate this to a probabilistic statement which involves the above pivot and decide on the size of the test  $\alpha$ . That is, define the rejection region to be

$$C_1 = \left\{ \mathbf{y}: (T-k) \frac{s^2}{\sigma^2} > c_x \right\}, \quad (19.78)$$

where  $c_x$  is determined by the distribution of

$$\tau(\mathbf{y}) = (T-k) \frac{s^2}{\sigma^2} \quad \text{under } H_0, \quad (19.79)$$

i.e.

$$Pr(\tau(\mathbf{y}) > c_x; \sigma^2 = \sigma_0^2) = Pr\left(\frac{(T-k)s^2}{\sigma^2} > c_x\right) = \alpha$$

or

$$\int_{c_\alpha}^{\infty} d\chi^2(T-k) = \alpha.$$

In the case of the money example let us assume  $\sigma_0^2 = 0.001$ . This implies that since  $s^2 = 0.00155$ ,  $c_s = 85.94$  for  $\alpha = 0.05$  and the rejection region takes the form

$$C_1 = \left\{ y: \frac{(T-k)s^2}{\sigma_0^2} > 85.94 \right\}.$$

Now,

$$\frac{(T-k)s^2}{\sigma_0^2} = 117.8,$$

and hence  $H_0$  is rejected.

In order to decide whether this is an ‘optimum’ test or not we need to consider its power function for which we need the distribution of  $\tau(y)$  under  $H_1$ . In this case we know that

$$\tau_0(y) = \frac{(T-k)s^2|_{H_1}}{\sigma_0^2} \sim \chi^2(T-k) \quad (19.80)$$

( $\sim$  reads ‘distributed under  $H_1$ ’), and thus

$$\tau(y) = \tau_0(y) \left( \frac{\sigma_0^2}{\sigma^2} \right)^{H_1} \sim \left( \frac{\sigma_0^2}{\sigma^2} \right) \chi^2(T-k), \quad (19.81)$$

because an affine function of a chi-square distributed random variable is also chi-square distributed (see Appendix 6.1). Hence, the power function takes the form

$$\mathcal{P}(\sigma^2) = Pr\left(\tau(y) > c_\alpha \left( \frac{\sigma_0^2}{\sigma^2} \right); \sigma^2 \geq \sigma_0^2\right) = \int_{c_\alpha(\sigma_0^2/\sigma^2)}^{\infty} d\chi^2(T-k). \quad (19.82)$$

The above test can be shown to be uniformly most powerful (UMP); see Chapter 14. Using the same procedure we could construct tests for:

- (i)  $H_0: \sigma^2 = \sigma_0^2$  against  $H_1: \sigma^2 < \sigma_0^2$  (one-sided) with

$$C_1^* = \{y: \tau(y) < c_\alpha^*\}, \quad \alpha = \int_{-\infty}^{c_\alpha^*} d\chi^2(T-k) \quad (19.83)$$

or

- (ii)  $H_0: \sigma^2 = \sigma_0^2$  against  $H_1: \sigma^2 \neq \sigma_0^2$  (two-sided) with

$$C_1^{**} = \{y: \tau(y) < a \text{ or } \tau(y) > b\},$$

$$\int_0^a d\chi^2(T-k) = \int_b^\infty d\chi^2(T-k) = \frac{\alpha}{2}. \quad (19.84)$$

The test defined by  $C_1^*$  is also UMP but the two-sided test defined by  $C_1^{**}$  is UMP unbiased. All these tests can be derived via the likelihood ratio test procedure.

Let us consider another test related to  $\sigma^2$  which will be used extensively in Section 21.6. The sample period is divided into two sub-periods, say,

$$t \in \mathbb{T}_1 = \{1, 2, \dots, T_1\}$$

and

$$t \in \mathbb{T}_2 = \{T_1 + 1, \dots, T\},$$

where  $T - T_1 = T_2$ , and the parameters of interest  $\theta \equiv (\beta, \sigma^2)$  are allowed to be different. That is, the statistical GM is postulated to be

$$y_t = \beta'_1 \mathbf{x}_t + u_t, \quad \text{Var}(y_t | \mathbf{X}_t = \mathbf{x}_t) = \sigma_1^2 \quad \text{for } t \in \mathbb{T}_1 \quad (19.85)$$

and

$$y_t = \beta'_2 \mathbf{x}_t + u_t, \quad \text{Var}(y_t | \mathbf{X}_t = \mathbf{x}_t) = \sigma_2^2 \quad \text{for } t \in \mathbb{T}_2. \quad (19.86)$$

An important hypothesis in this context is

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} = c_0 \quad \text{against} \quad H_1: \frac{\sigma_1^2}{\sigma_2^2} > c_0,$$

where  $c_0$  is a known constant (usually  $c = 1$ ).

Intuition suggests that an obvious way to proceed in order to construct a test for this hypothesis is to estimate the statistical GM for the two sub-periods separately and use

$$s_1^2 = \frac{1}{T_1 - k} \sum_{t=1}^{T_1} \hat{u}_t^2$$

and

$$s_2^2 = \frac{1}{T_2 - k} \sum_{t=T_1+1}^T \hat{u}_t^2$$

to define the statistic  $\tau(\mathbf{y}) = s_1^2/s_2^2$ . Given that

$$\frac{(T_i - k)s_i^2}{\sigma_i^2} \sim \chi^2(T_i - k), \quad i = 1, 2, \quad (19.87)$$

from (77), and  $s_1^2$  is independent of  $s_2^2$  (due to the sampling model assumption), we can deduce that

$$\begin{aligned} & \left( \frac{(T_1 - k)s_1^2 / (\sigma_1^2 / (T_1 - k))}{(T_2 - k)s_2^2 / (\sigma_2^2 / (T_2 - k))} \right) \\ &= \frac{s_1^2}{s_2^2} \left( \frac{\sigma_2^2}{\sigma_1^2} \right) \sim F(T_1 - k, T_2 - k; \delta), \end{aligned} \quad (19.88)$$

Hence,

$$\tau(\mathbf{y}) = \left( \frac{s_1^2}{c_0 s_2^2} \right)^{H_0} \sim F(T_1 - k, T_2 - k).$$

This can be used to define a test based on the rejection region  $C_1 = \{\mathbf{y}: \tau(\mathbf{y}) > c_s\}$  where the critical value  $c_s$  is determined via  $\int_{c_s}^{\infty} dF(T_1 - k, T_2 - k) = \alpha$ ,  $\alpha$  being the size of the test chosen a priori. It turns out that this defines a UMP unbiased test (see Lehmann (1959)). Of particular interest is the case where  $c_0 = 1$ , i.e.  $H_0: \sigma_1^2 = \sigma_2^2$ . Note that the alternative  $H_1: \sigma_1^2 / \sigma_2^2 < 1$  can be easily accommodated by defining it as  $H_1: \sigma_2^2 / \sigma_1^2 > 1$ , i.e. have the greater of the two variances on the numerator.

## (2) Tests relating to $\beta$

The first question usually asked in relation to the coefficient parameters  $\beta$  is whether they are ‘statistically significant’. Statistical significance is formalised in the form of the null hypothesis:

$$H_0: \beta_i = 0$$

against

$$H_1: \beta_i \neq 0 \quad \text{for some } i = 1, 2, \dots, k.$$

Common sense suggests that a natural way to proceed in order to construct a test for these hypotheses is to consider how far from zero  $\hat{\beta}_i$  is. The problem with this, however, is that the estimate of  $\beta$  depends crucially on the units of measurements used for  $y_t$  and  $\mathbf{X}_t$ . The obvious way to avoid this problem is to divide the  $\hat{\beta}_i$ s by their standard deviation. Since  $\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$ , the standard deviation of  $\hat{\beta}_i$  is

$$\sqrt{[\text{Var}(\hat{\beta}_i)]} = \sqrt{[\sigma^2(\mathbf{X}'\mathbf{X})_{ii}^{-1}]},$$

where  $(\mathbf{X}'\mathbf{X})_{ii}^{-1}$  refers to the  $i$ th diagonal element of  $(\mathbf{X}'\mathbf{X})^{-1}$ . Hence we can deduce that a likely pivot for the above hypotheses might be

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{[\sigma^2(\mathbf{X}'\mathbf{X})_{ii}^{-1}]}} \sim N(0, 1). \quad (19.89)$$

The problem with this suggestion, however, is that this is not a pivot given that  $\sigma^2$  is unknown. The natural way to ‘solve’ this problem is to substitute its estimator  $s^2$  in such a way so as to end up with a quantity for which we know the distribution. This is achieved by dividing the above quantity with the square root of

$$\left( (T-k)s^2 \left/ \left( \frac{\sigma^2}{T-k} \right) \right. \right),$$

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{[\sigma^2(\mathbf{X}'\mathbf{X})_{ii}^{-1}]}} = \frac{\hat{\beta}_i - \beta_i}{\sqrt{[s^2(\mathbf{X}'\mathbf{X})_{ii}^{-1}]}} \sim t(T-k), \quad (19.90)$$

$$\sqrt{\left[ \frac{(T-k)s^2}{(T-k)\sigma^2} \right]} =$$

which is a very convenient pivot. For  $H_0$  above

$$\tau(\mathbf{y}) = \frac{\hat{\beta}_i}{s\sqrt{[(\mathbf{X}'\mathbf{X})_{ii}^{-1}]}} \stackrel{H_0}{\sim} t(T-k).$$

Using this we can define the rejection region

$$C_1 = \{\mathbf{y}: |\tau(\mathbf{y})| > c_\alpha\},$$

where  $c_\alpha$  is determined from the  $t$  tables for a given size  $\alpha$ . That is,

$$\int_{-c_\alpha}^{c_\alpha} dt(T-k) = 1 - \alpha.$$

The decision on ‘how optimal’ the above test is can only be considered using its power function. For this we need the distribution of  $\tau(\mathbf{y})$  under  $H_1$ , say  $\beta_i = \beta_i^0$ ,  $\beta_i^0 \neq 0$ . Given that

$$\tau_0(\mathbf{y}) = \frac{\hat{\beta}_i - \beta_i^0}{\sqrt{[s^2(\mathbf{X}'\mathbf{X})_{ii}^{-1}]}} \sim t(T-k), \quad (19.91)$$

$$\tau(\mathbf{y}) = \left( \tau_0(\mathbf{y}) + \frac{\beta_i^0}{\sqrt{[s^2(\mathbf{X}'\mathbf{X})_{ii}^{-1}]}} \right) \sim t(T-k; \delta), \quad (19.92)$$

a non-central  $t$  with non-centrality parameter

$$\delta = \frac{\beta_i^0}{\sigma\sqrt{[(\mathbf{X}'\mathbf{X})_{ii}^{-1}]}}$$

(see Appendix 6.1). This test can be shown to be UMP unbiased.

In the case of the money example above for  $H_0^i$ :  $\beta_i = 0$ ,  $i = 1, 2, 3, 4$ ,

$$\frac{\hat{\beta}_1}{s\sqrt{[(\mathbf{X}'\mathbf{X})_{11}^{-1}]}} = 2.8, \quad \frac{\hat{\beta}_2}{s\sqrt{[(\mathbf{X}'\mathbf{X})_{22}^{-1}]}} = 6.5,$$

$$\frac{\hat{\beta}_3}{s\sqrt{[(\mathbf{X}'\mathbf{X})_{33}^{-1}]}} = 42.9, \quad \frac{\hat{\beta}_4}{s\sqrt{[(\mathbf{X}'\mathbf{X})_{44}^{-1}]}} = -4.1.$$

For a size  $\alpha = 0.05$  two-sided test the critical value is  $c_\alpha = 1.993$ , given that we have  $T-k=76$  degrees of freedom. Assuming that the underlying assumptions of the linear regression model are valid (as it happens they are not!) we can proceed to argue that all the above hypotheses of significance

(the coefficients are zero) are rejected. That is, the coefficients are indeed significantly different from zero. It must be emphasised that the above  $t$ -tests on each coefficient are separate tests and should not be confused with the joint test:  $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ , which will be developed next.

The null hypothesis considered above provides an example of linear hypotheses, i.e. hypotheses specified in the form of linear functions of  $\beta$ . Instead of considering the various forms such linear hypotheses can take we will consider constructing a test for a general formulation.

Restrictions among the parameters  $\beta$ , such as:

- (i)  $\beta_4 = 0;$
- (ii)  $\beta_2 = \beta_3;$
- (iii)  $\beta_2 = 1; \beta_3 + \beta_5 = 1;$
- (iv)  $\beta_2 + \beta_3 + \beta_4 + \beta_5 = 1;$

can be accommodated within the linear formulation

$$\mathbf{R}\beta = \mathbf{r}, \quad \text{rank}(\mathbf{R}) = m \quad (19.93)$$

where  $\mathbf{R}$  and  $\mathbf{r}$  are  $m \times k$  ( $k > m$ ) and  $m \times 1$  known matrices. For example, in the case of (iii),

$$\mathbf{R} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix}, \quad \mathbf{r} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \quad (19.94)$$

This suggests that linear hypotheses related to  $\beta$  can be considered as special cases of the null hypothesis

$$H_0: \mathbf{R}\beta = \mathbf{r} \quad \text{against the alternative} \quad H_1: \mathbf{R}\beta \neq \mathbf{r}.$$

In Chapter 20 a test for this hypothesis will be derived via the likelihood ratio test procedure. In what follows, however, the same test will be derived using the common-sense approach which served us so well in deriving optimal tests for  $\sigma^2$  and  $\beta_i$ .

The problem we face is to construct a test in order to decide whether  $\beta$  satisfies the restrictions  $\mathbf{R}\beta = \mathbf{r}$  or not. Since  $\beta$  is unknown, the next best thing to do is use  $\hat{\beta}$  (knowing that it is a ‘good’ estimator of  $\beta$ ) and check whether the discrepancies

$$\|\mathbf{R}\hat{\beta} - \mathbf{r}\| \quad (19.95)$$

are ‘close to zero’ or not. Since when deriving  $\hat{\beta}$  no such restrictions were taken into consideration, if  $H_0$  is true,  $\hat{\beta}$  should come very close to satisfying these restrictions. The question now is ‘how close is close?’ . The answer to this question can only be given in relation to the distribution of some test

statistic related to (95). We could not use this as a test statistic for two reasons:

- (i) it depends crucially on the units of measurement used for  $y_t$  and  $\mathbf{X}_t$ ;
- and
- (ii) the absolute value feature of (95) makes it very awkward to manipulate.

The units of measurement problem in such a context is usually solved by dividing the quantity by its standard deviation as we did in (89) above. The absolute value difficulty is commonly avoided by squaring the quantity in question. If we apply these in the case of (95) the end result will be the quadratic form

$$(\mathbf{R}\hat{\beta} - \mathbf{r})' [\text{Var}(\mathbf{R}\hat{\beta} - \mathbf{r})]^{-1} (\mathbf{R}\hat{\beta} - \mathbf{r}), \quad (19.96)$$

which is a direct matrix generalisation of (89). Now, the problem is to determine the form of  $\text{Var}(\mathbf{R}\hat{\beta} - \mathbf{r})$ . Since  $\mathbf{R}\hat{\beta} - \mathbf{r}$  is a linear function of a normally distributed random vector,  $\hat{\beta}$ ,

$$(\mathbf{R}\hat{\beta} - \mathbf{r}) \sim N(\mathbf{R}\beta - \mathbf{r}, \sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}') \quad (19.97)$$

(from N1 in Chapter 15). Hence, (96) becomes

$$(\mathbf{R}\hat{\beta} - \mathbf{r})' [\sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{R}\hat{\beta} - \mathbf{r}). \quad (19.98)$$

This being a quadratic form in normally distributed random variables, it must be distributed as a chi-square. Using Q1 of Chapter 15 we can deduce that

$$(\mathbf{R}\hat{\beta} - \mathbf{r})' [\sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{R}\hat{\beta} - \mathbf{r}) \sim \chi^2(m; \delta), \quad (19.99)$$

i.e. (16) is distributed as a non-central chi-square with  $m$  (rank  $(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}')$ ) degrees of freedom and non-centrality parameter

$$\delta = \frac{(\mathbf{R}\beta - \mathbf{r})' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{R}\beta - \mathbf{r})}{\sigma^2} \quad (19.100)$$

(see Appendix 6.1). Looking at (99) we can see that it is not a test statistic as yet because it involves the unknown parameter  $\sigma^2$ . Intuition suggests that if we were to substitute  $s^2$  in the place at  $\sigma^2$  we might get a test statistic. The problem with this is that we end up with

$$\frac{(\mathbf{R}\hat{\beta} - \mathbf{r})' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}'] (\mathbf{R}\hat{\beta} - \mathbf{r})}{s^2}, \quad (19.101)$$

for which we do not know the distribution. An equivalent way to proceed which ensures that we end up with a pivot (a test statistic whose distribution

is known) is the following. Since,

$$(T-k) \frac{s^2}{\sigma^2} \sim \chi^2(T-k), \quad (19.102)$$

if we could show that this is independent of (99) we could take their ratio (divided by the respective degrees of freedom) to end up with an  $F$ -distributed test statistic; see Q5 and Q7 of Chapter 15. In order to prove independence we need to express both quantities in quadratic forms which involve the same normally distributed random vector. From (102) we know that

$$(T-k) \frac{s^2}{\sigma^2} = \frac{\mathbf{u}'(\mathbf{I} - \mathbf{P}_x)\mathbf{u}}{\sigma^2}, \quad \mathbf{P}_x = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'. \quad (19.103)$$

After some manipulation we can express (99) in the form

$$\mathbf{u}'\mathbf{Q}_x\mathbf{u} + \frac{(\mathbf{R}\beta - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\beta - \mathbf{r})}{\sigma^2}, \quad (19.104)$$

where

$$\mathbf{Q}_x = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

In view of  $\mathbf{Q}_x(\mathbf{I} - \mathbf{P}_x) = \mathbf{0}$ , (99) and (102) are independent. This implies that

$$\tau(y) = \frac{\frac{\mathbf{u}'\mathbf{Q}\mathbf{u} + (\mathbf{R}\beta - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\beta - \mathbf{r})}{m\sigma^2}}{\frac{\mathbf{u}'(\mathbf{I} - \mathbf{P}_x)\mathbf{u}}{(T-k)\sigma^2}} \sim F(m, T-k; \delta). \quad (19.105)$$

A more convenient form for  $\tau(y)$  is

$$\tau(y) = \frac{1}{m} \frac{(\mathbf{R}\hat{\beta} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta} - \mathbf{r})}{s^2}, \quad (19.106)$$

which apart from the factor  $(1/m)$  is identical to (101), the quantity derived by our intuitive argument. It is important to note that under  $H_0$ ,  $\mathbf{R}\beta = \mathbf{r}$  and  $\delta = 0$ , i.e.

$$\tau(y) \stackrel{H_0}{\sim} F(m, T-k). \quad (19.107)$$

Using this we can define the rejection region to be

$$C_1 = \{y: \tau(y) > c_\alpha\} \quad \text{where } \alpha = \int_{c_\alpha}^\infty dF(m, T-k). \quad (19.108)$$

The power of this test depends crucially on the non-centrality parameter  $\delta$

as given in (100). In view of the fact that  $E(\tau(\mathbf{y})) = [(T-k)(m+\delta)]/[m(T-k-2)]$  (see Appendix 6.1) we can see that the larger  $\delta$  is the greater the power of the test (ensure that you understand why). The non-centrality parameter is larger the greater the distance  $\|\mathbf{R}\beta - \mathbf{r}\|$  (a very desirable feature) and the smaller the conditional variance  $\sigma^2$ . The power also depends on the degrees of freedom  $v_1 = m$ ,  $v_2 = T-k$ . In order to show this explicitly let us use the well-known relationship between the  $F$  and beta distributions which enables us to deduce that the statistic

$$\tau^*(\mathbf{y}) = [v_1 \tau(\mathbf{y})]/[v_1 \tau(\mathbf{y}) + v_2] \quad (19.109)$$

is distributed as non-central beta (see Section 21.5). The power function in terms of  $\tau^*(\mathbf{y})$  is

$$\begin{aligned} \mathcal{P}(\beta) &= Pr(\tau^*(\mathbf{y}) > c_x^*) \\ &= e^{-(\delta/2)} \sum_{l=0}^{\infty} \frac{(\delta/2)^l}{l!} \int_{c_x^*}^1 \frac{\tau^{*[v_1/2-1+l]}(1-\tau^*)^{(v_2/2)-1}}{B(\frac{1}{2}v_1+l, \frac{1}{2}v_2)} d\tau^* \end{aligned} \quad (19.110)$$

(see Johnson and Kotz (1970)). From (110) we can see that the power of the test, *ceteris paribus*, increases as  $T-k$  increases and  $m$  decreases. It can be shown that the  $F$  test of size  $\alpha$  is UMP unbiased and invariant to transformations of the form:

- (i)  $\mathbf{y}^* = c\mathbf{y}$  ( $c > 0$ )
- (ii)  $\mathbf{y}^* = \mathbf{y} + \boldsymbol{\mu}_0$ , where  $\boldsymbol{\mu}_0 \in \Theta_0$ .

(for further details see Seber (1980)).

One important ‘disadvantage’ of the  $F$ -test is that it provides a joint test for the  $m$  hypotheses  $\mathbf{R}\beta = \mathbf{r}$ . This implies that when the  $F$ -test leads us to reject  $H_0$  any one or any combination of these  $m$  separate hypotheses might be responsible for the rejection and the  $F$ -test throws no light on the matter. In order to be able to decide on this matter we need to consider simultaneous hypothesis testing; see Savin (1984).

As argued in Chapter 14, there exists a duality relationship between hypothesis testing and confidence regions which enables us to transform an optimal test to an optimal confidence region and vice versa. For example, the acceptance region of the  $F$ -test with  $R = \mathbf{h}$ ;  $k \times 1$ ,  $r = \mathbf{h}'\boldsymbol{\beta}^0$ ,  $\boldsymbol{\beta}^0$  known,

$$C_0(\boldsymbol{\beta}^0) = \left\{ \mathbf{y}: \frac{|\mathbf{h}'\hat{\boldsymbol{\beta}} - \mathbf{h}'\boldsymbol{\beta}^0|}{s\sqrt{[\mathbf{h}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{h}]}} \leq c_\alpha \right\}, \quad \int_{-c_\alpha}^{c_\alpha} dt(T-k) = 1 - \alpha, \quad (19.112)$$

can be transformed into a  $(1 - \alpha)$  confidence interval

$$C(\mathbf{y}) = \{\boldsymbol{\beta}: \mathbf{h}'\hat{\boldsymbol{\beta}} - s\sqrt{[\mathbf{h}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{h}]}c_\alpha \leq \mathbf{h}'\boldsymbol{\beta} < \mathbf{h}'\hat{\boldsymbol{\beta}} + c_\alpha s\sqrt{[\mathbf{h}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{h}]}\}. \quad (19.113)$$

Note that the above result is based on the fact that if

$$V \sim F(1, T-k) \text{ then } \sqrt{V} \sim t(T-k). \quad (19.114)$$

A special case of the linear restrictions  $R\boldsymbol{\beta} = \mathbf{r}$  which is of interest in econometric modelling is when the null hypothesis is

$$H_0: \boldsymbol{\beta}_{(1)} = \mathbf{0} \text{ against } H_1: \boldsymbol{\beta}_{(1)} \neq \mathbf{0},$$

where  $\boldsymbol{\beta}_{(1)}$  represents all the coefficients apart from the coefficient of the constant. In this case  $R = \mathbf{I}_{k-1}$  and  $\mathbf{r} = \mathbf{0}$ . Applying this test to the money equation estimated in Section 19.4 we get

$$\tau(\mathbf{y}) = \left(\frac{1}{3}\right) \frac{248.362}{0.0015463} = 5353.904, \quad c_\alpha = 2.76, \quad \alpha = 0.05.$$

This suggests that the null hypothesis is strongly rejected. Caution, however, should be exercised in interpreting this result in view of the discussion of possible misspecification in Section 19.4. Moreover, being a joint test its value can be easily 'inflated' by any one of the coefficients. In the present case the coefficient of  $p_t$  is largely responsible for the high value of the test statistic. If real money stock is used, thus detrending  $M_t$  by dividing it with  $P_t$  which has a very similar trend (see Fig. 17.1(a) and 17.1(c)), the test statistic for the significance of the coefficients takes the value 22.279; a great deal smaller than the one above. This is clearly exemplified in Fig. 19.2 where the goodness of fit looks rather poor.

## 19.6 Prediction

The objective so far has been to estimate or construct tests for hypotheses related to the parameters of the statistical GM

$$y_t = \boldsymbol{\beta}' \mathbf{x}_t + u_t, \quad t \in \mathbb{T}, \quad (19.115)$$

using the observed data for the observation period  $t = 1, 2, \dots, T$ . The question which naturally arises is to what extent we can use (115) together with the estimated parameters of interest in order to predict values of  $y$  beyond the observation period, say

$$y_{T+l}, \quad l = 1, 2, \dots$$

From Section 12.3 we know that the best predictor for  $y_{T+l}$ ,  $l=1,2,\dots$  is its conditional expectation given the relevant information set. In the present case this information set comes in the form of  $\mathcal{D}_{T+l} = \{\mathbf{X}_{T+l} = \mathbf{x}_{T+l}\}$ . This suggests that in order to be able to predict beyond the sample period we need to ensure that  $\mathcal{D}_{T+l}$ ,  $l>0$ , is available. Assuming that  $\mathbf{X}_{T+l} = \mathbf{x}_{T+l}$  is available for some  $l \geq 1$  and knowing that  $E(u_{T+l}/\mathbf{X}_{T+l} = \mathbf{x}_{T+l}) = 0$  and

$$\mu_{T+l} \equiv E(y_{T+l}/\mathbf{X}_{T+l} = \mathbf{x}_{T+l}) = \boldsymbol{\beta}' \mathbf{x}_{T+l}, \quad (19.116)$$

a natural predictor for  $\mu_{T+l}$  must be

$$\hat{\mu}_{T+l} = \hat{\boldsymbol{\beta}} \mathbf{x}_{T+l}. \quad (19.117)$$

In order to assess how good this predictor is we need to compare it with the actual value of  $y$ ,  $y_{T+l}$ . The prediction error is defined as

$$e_{T+l} \equiv y_{T+l} - \hat{\mu}_{T+l} = u_{T+l} + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_T)' \mathbf{x}_{T+l} \quad (19.118)$$

and

$$e_{T+l} \sim N(0, \sigma^2(1 + \mathbf{x}'_{T+l}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_{T+l})), \quad (19.119)$$

since  $e_{T+l}$  is a linear function of normally distributed random variables, and the two quantities involved are independent. The optimal properties of  $\hat{\boldsymbol{\beta}}$  make  $\hat{\mu}_{T+l}$  an 'optimal' predictor and  $e_{T+l}$  has the smallest variance among linear predictors (see Section 12.3 and Harvey (1981)).

In order to construct a confidence interval for  $y_{T+l}$  a pivot is needed. The obvious quantity

$$\frac{e_{T+l}}{\sigma \sqrt{[1 + \mathbf{x}'_{T+l}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_{T+l}]} } \sim N(0, 1) \quad (19.120)$$

is not a pivot because it involves the unknown parameter  $\sigma^2$ . We could reduce it to a pivot, however, by dividing it by  $\sqrt{[(T-k)s^2]/[(T-k)\sigma^2]}$ , since  $s^2$  is independent of  $e_{T+l}$ , to obtain

$$\frac{\frac{e_{T+l}}{\sigma \sqrt{[1 + \mathbf{x}'_{T+l}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_{T+l}]} }}{\sqrt{\left(\frac{s^2}{\sigma^2}\right)}} = \frac{e_{T+l}}{s \sqrt{[1 + \mathbf{x}'_{T+l}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_{T+l}]} } \sim t(T-k), \quad (19.121)$$

see Section 6.3. (Using (121) we can set up the *prediction interval*

$$\begin{aligned} Pr(\hat{\boldsymbol{\beta}} \mathbf{x}_{T+l} - c_\alpha s \sqrt{[1 + \mathbf{x}'_{T+l}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_{T+l}]} \leq y_{T+l} \leq \hat{\boldsymbol{\beta}}' \mathbf{x}_{T+l} \\ + c_\alpha s \sqrt{[1 + \mathbf{x}'_{T+l}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_{T+l}]} ) = 1 - \alpha, \end{aligned} \quad (19.122)$$

where  $c_\alpha$  is determined from the  $t$  tables for a given  $\alpha$  via

$$\int_{-c_\alpha}^{c_\alpha} dt(T-k) = 1 - \alpha.$$

As in the case of specification testing, prediction is based on the assumption that the estimated equation represents a well-defined estimated statistical GM; the underlying assumptions are valid. If this is not the case then using the estimated equation to predict  $y_{T+l}$  can be very misleading. Prediction, however, can be used for misspecification testing purposes if additional observations beyond the sample period are available. It seems obvious that if it is assumed to represent a well-defined statistical GM and the sample observations  $t = 1, 2, \dots, T$ , are used for the estimation of  $\beta$ , then the predictions based on  $\mu_{T+l} = \hat{\beta}' \mathbf{x}_{T+l}, l = 1, 2, \dots, m$ , when compared with  $y_{T+l}, l = 1, 2, \dots, m$ , should give us some idea about the validity of the ‘correct specification’ assumption.

Let us re-estimate the money equation estimated in Section 19.4 for the sub-period 1963*i*–1980*ii* and use the rest of the observed data to get some idea about the predictive ability of the estimated equation. Estimation for the period 1963*i*–1980*ii* yielded

$$m_t = 3.029 + 0.678 y_t + 0.863 p_t - 0.049 i_t + \hat{u}_t, \quad (19.123)$$

(1.116) (0.113) (0.024) (0.014) (0.040)

$$R^2 = 0.993, \quad \bar{R}^2 = 0.993, \quad s = 0.0402,$$

$$RSS = 0.10667, \quad T = 70, \quad \log L = 127.703.$$

Using this estimated equation to predict for the period 1980*iii*–1982*iv* the following prediction errors resulted:

$$\begin{aligned} e_1 &= -0.0317, & e_2 &= -0.0279, & e_3 &= -0.0217, & e_4 &= -0.0243, \\ e_5 &= -0.0314, & e_6 &= -0.0193, & e_7 &= 0.0457, & e_8 &= 0.0408, \\ e_9 &= 0.0276, & e_{10} &= 0.0497. \end{aligned}$$

As can be seen, the estimated equation *underpredicts* for the first six periods and *overpredicts* for the rest. This clearly indicates that the estimated equation leaves a lot to be desired on prediction grounds and re-enforces the initial claim that some misspecification is indeed present.

Several measures of predictive ability have been suggested in the econometric literature. The most commonly used statistics are:

$$MSE = \frac{1}{m} \sum_{t=T+1}^{T+m} e_t^2 \quad (\text{mean square error}); \quad (19.124)$$

$$MAE = \frac{1}{m} \sum_{t=T+1}^{T+m} |e_t| \quad (\text{mean absolute error}); \quad (19.125)$$

$$U = \frac{\left( \frac{1}{m} \sum_{t=T+1}^{T+m} e_t^2 \right)^{\frac{1}{2}}}{\left( \frac{1}{m} \sum_{t=T+1}^{T+m} y_t^2 \right)^{\frac{1}{2}} + \left( \frac{1}{m} \sum_{t=T+1}^{T+m} \hat{\mu}_t^2 \right)^{\frac{1}{2}}} \quad (\text{Theil's inequality coefficient}) \quad (19.126)$$

(see Pindyck and Rubinfeld (1981) for a more extensive discussion). For the above example  $MSE = 0.00112$ ,  $MAE = 0.03201$ ,  $U = 0.835$ . The relatively high value of  $U$  indicates a weakness in the predictive ability of the estimated equation.

The above form of prediction is sometimes called *ex-post* prediction because the actual observations for  $y_t$  and  $\mathbf{X}_t$  are available for the prediction period. *Ex-ante* prediction, on the other hand, refers to prediction where this is not the case and the values of  $\mathbf{X}_t$  for the post sample period are ‘guessed’ (in some way). In ex-ante prediction ensuring that the underlying assumptions of the statistical GM in question are valid is of paramount importance. As with specification testing, ex-ante prediction should be preceded by misspecification testing which is discussed in Chapters 20–22. Having accepted the assumptions underlying the linear regression model as valid we can proceed to ‘guessestimate’  $\mathbf{x}_{T+l}$  by  $\hat{\mathbf{x}}_{T+l}$  and use

$$\hat{\mu}_{T+l} = \hat{\beta}_T \hat{\mathbf{x}}_{T+l}, \quad l = 1, 2, \dots \quad (19.127)$$

as the predictor of  $y_{T+l}$ . In such a case the prediction error defined by  $\hat{e}_{T+l} = y_{T+l} - \hat{\mu}_{T+l}$  can be decomposed into three sources of errors:

$$\hat{e}_{T+l} = u_{T+l} + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_T)' \mathbf{x}_{T+l} + (\mathbf{x}_{T+l} - \hat{\mathbf{x}}_{T+l})' \hat{\boldsymbol{\beta}}_T, \quad (19.128)$$

one additional to  $e_{T+l}$  (see (118)).

## 19.7 The residuals

The residuals for the linear regression model are defined by

$$\hat{u}_t \equiv y_t - \hat{\boldsymbol{\beta}}' \mathbf{x}_t, \quad t = 1, 2, \dots, T \quad (19.129)$$

or

$$\hat{\mathbf{u}} \equiv \mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}} \quad \text{in matrix notation.} \quad (19.130)$$

From the definition of the residuals we can deduce that they should play a very important role in the misspecification analysis (testing the assumptions of the linear regression model) because any test related to

$\{y_t/\mathbf{X}_t = \mathbf{x}_t, t \in \mathbb{T}\}$  can only be tested via  $\hat{u}_t$ . For example, if we were to test any one of the assumptions relating to the probability model

$$(y_t/\mathbf{X}_t = \mathbf{x}_t) \sim N(\beta' \mathbf{x}_t, \sigma^2) \quad (19.131)$$

we cannot use  $y_t$  because (131) refers to its conditional distribution (not the marginal) and we cannot use

$$(y_t - \beta' \mathbf{x}_t) \sim N(0, \sigma^2) \quad (19.132)$$

because  $\beta$  is unknown. The natural thing to do is to use  $(y_t - \hat{\beta}' \mathbf{x}_t)$ , i.e. the residuals  $\hat{u}_t$ ,  $t = 1, 2, \dots, T$ . It must be stressed, however, that in the same way as  $u_t$  does not have a ‘life of its own’ (it stands for  $y_t - \beta' \mathbf{x}_t$ ),  $\hat{u}_t$  stands for  $y_t - \hat{\beta}' \mathbf{x}_t$  and should not be interpreted as an ‘autonomous’ random variable but as the observable form of  $(y_t/\mathbf{X}_t = \mathbf{x}_t)$  in mean deviation form.

The distribution of  $\hat{\mathbf{u}} \equiv \mathbf{y} - \mathbf{X} \hat{\beta} = (\mathbf{I} - \mathbf{P}_x) \mathbf{y} = (\mathbf{I} - \mathbf{P}_x) \mathbf{u}$  takes the form

$$\hat{\mathbf{u}} \sim N(\mathbf{0}, \sigma^2 (\mathbf{I} - \mathbf{P}_x)), \quad (19.133)$$

where  $\mathbf{P}_x$  is the idempotent matrix discussed in Section 19.4 above. Given, however, that  $\text{rank } (\mathbf{I} - \mathbf{P}_x) = \text{tr}(\mathbf{I} - \mathbf{P}_x) = T - k$  we can deduce that the distribution of  $\hat{\mathbf{u}}$  is a singular multivariate normal. Hence, the distributions of  $\hat{\mathbf{u}}$  and  $\mathbf{u}$  can coincide only asymptotically if

$$v(\mathbf{X}' \mathbf{X})^{-1} \rightarrow 0 \quad \text{as } T \rightarrow \infty, \quad (19.134)$$

where  $v(\mathbf{A}) = \max_{i,j} |a_{ij}|$ ,  $\mathbf{A} = [a_{ij}]_{i,j}$ . This condition plays an important role in relation to the asymptotic results related to  $s^2$ . Without the condition (134) the asymptotic distribution of  $s^2$  will depend on the fourth central moment of its finite sample distribution (see Section 21.2). What is more, the condition  $\lim_{T \rightarrow \infty} (\mathbf{X}' \mathbf{X})^{-1} = \mathbf{0}$  or equivalently

$$v(\mathbf{X}' \mathbf{X})^{-1} \rightarrow 0 \quad \text{as } T \rightarrow \infty \quad (19.135)$$

does not imply (134) as can be verified for  $x_{ii} = \sqrt{2^i}$ .

Looking at (133) we can see that the finite sampling distribution of  $\hat{\mathbf{u}}$  is inextricably bound up with the observed values of  $\mathbf{X}_t$  and thus any finite sample test based on  $\hat{\mathbf{u}}$  will be bound up with the particular  $\mathbf{X}$  matrix in hand. This, together with the singularity of (133), prompt us to ask whether we could transform  $\hat{\mathbf{u}}$  in such a way so as to sidestep both problems. One way we could do that is to find a  $T \times (T-k)$  matrix  $\mathbf{H}$  such that

$$\mathbf{H}'(\mathbf{I} - \mathbf{P}_x)\mathbf{H} = \Lambda, \quad (19.136)$$

where  $\Lambda$  takes the form

$$\Lambda = \begin{pmatrix} \mathbf{I}_{T-k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

being the matrix of the eigenvalues of  $\mathbf{I} - \mathbf{P}_x$ . This enables us to define the transformed residuals, known as *BLUS residuals*, to be

$$\hat{\mathbf{v}} = \mathbf{H}'\hat{\mathbf{u}} \sim N(\mathbf{0}, \sigma^2, \mathbf{I}_{T-k}), \quad (19.137)$$

$\hat{\mathbf{v}}$  being a  $(T-k) \times 1$  vector (see Theil (1971)). These residuals can be used in misspecification tests instead of  $\hat{\mathbf{u}}$  but their interpretation becomes rather difficult. This is because  $\hat{v}_t$  is a linear combination of all  $\hat{u}_s$ s and cannot be related to the observation date  $t$ . Another form of transformed residuals which emphasises the time relationship with the observation date is the *recursive residuals*.

The *recursive residuals* are defined by

$$\tilde{\mathbf{v}}_t = \begin{cases} 0 & \text{for } t = 1, 2, \dots, k \\ y_t - \hat{\beta}'_{t-1} \mathbf{x}_t, & t = k+1, \dots, T, \end{cases} \quad (19.138)$$

where

$$\hat{\beta}_{t-1} = (\mathbf{X}_{t-1}^{0'} \mathbf{X}_{t-1}^0)^{-1} \mathbf{X}_{t-1}^0 \mathbf{y}_{t-1}^0, \quad t \geq k+1 \quad (19.139)$$

is the *recursive least-squares estimator* of  $\beta$  with

$$\mathbf{X}_{t-1}^0 \equiv (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{t-1})', \quad \mathbf{y}_{t-1}^0 \equiv (y_1, y_2, \dots, y_{t-1})'.$$

This estimator of  $\beta$  uses information up to  $t-1$  only and it can be of considerable value in the context of theoretical models which involve expectations in various forms.

$$\tilde{v}_t \sim N(0, \sigma^2 (1 + \mathbf{x}_t' (\mathbf{X}_{t-1}^{0'} \mathbf{X}_{t-1}^0)^{-1} \mathbf{x}_t)), \quad (19.140)$$

and for

$$\begin{aligned} \tilde{v}_t^* &= \frac{\tilde{v}_t}{\sqrt{[1 + \mathbf{x}_t' (\mathbf{X}_{t-1}^{0'} \mathbf{X}_{t-1}^0)^{-1} \mathbf{x}_t]}}, \\ \tilde{\mathbf{v}}^* &\equiv (v_{k+1}^*, v_{k+2}^*, \dots, v_T^*)', \\ \tilde{\mathbf{v}}^* &\sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{T-k}). \end{aligned} \quad (19.141)$$

The recursive residuals have certain distinct advantages over  $\hat{\mathbf{u}}$  in misspecification tests related to the time dependency of the  $y_t$ s (see Section 21.5 and Harvey (1981)).

The residuals play a very important role in misspecification testing, as shown in Chapters 20–22, because they provide us with a way to test the underlying assumptions relating to the process  $\{y_t/\mathbf{X}_t, t \in \mathbb{T}\}$ . It is interesting at this stage to have a look at the time path of the residuals for the money equation estimated in Section 19.4, shown in Fig. 19.3. The residuals exemplify a certain tendency to increase once they start increasing and to decrease once they start decreasing. This indicates the presence of strong positive correlation in successive residuals (or serial correlation) and

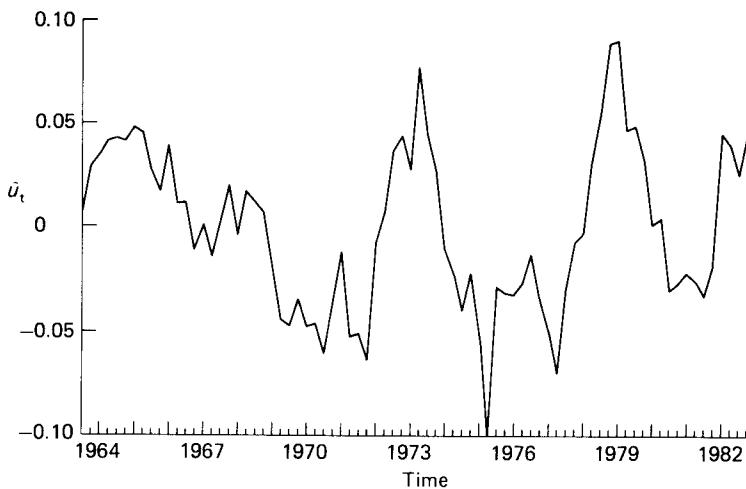


Fig. 19.3. The residuals from (19.66).

therefore the sampling model *assumption of independence* seems rather suspect (invalid). The time path of the residuals indicates the presence of systematic temporal information which was 'ignored' by the postulated systematic component (see Chapter 22).

### 19.8 Summary and conclusion

The linear regression model is undoubtedly the most widely used statistical model in econometric modelling. Moreover, the model provides the foundation for various extensions which give rise to several statistical models of interest in econometric modelling. The main purpose of this chapter has been to discuss the underlying assumptions in order to enable the econometric modeller to decide upon its appropriateness for the particular case in question as well as derive statistical inference results related to the model. These included estimation, specification testing and prediction. The statistical inference results are derived based on the presupposition that the assumptions underlying the linear regression model are valid. When this is not the case these results are not only inappropriate, they can be very misleading.

The money equation estimated in Section 19.4 was chosen to highlight the importance of choosing the most appropriate statistical model by taking into consideration not just the theoretical information but the information related to the observed data chosen. The latter form of information should be taken into consideration in postulating the

probability and sampling models as well as the statistical GM. The estimation and testing results related to the money equation might encourage premature pronouncements of a well-specified transactions demand for money. Such conclusions, however, are not warranted in view of several indications of misspecification discussed briefly above. Before any estimation, specification testing or prediction results can be considered appropriate we need to test the underlying assumptions on the validity of which they are based. This is the task of misspecification testing considered in the next three chapters.

Once the underlying assumptions [1]–[8] are tested and their validity established with the data chosen, the estimated statistical GM is said to constitute *a well-defined estimated statistical model*. This, however, does not necessarily coincide with the empirical econometric model because the statistical and theoretical parameters of interest are commonly different. The well-defined estimated statistical model is transformed into an empirical econometric model by reparametrising it in terms of the theoretical parameters of interest.

### Appendix 19.1 – A note on measurement systems

A measurement system refers to the range of the variables involved and the associated mathematical structure. This system is normally selected so that the mathematical structure of the range reflects the structural properties of the phenomena being measured. Different measurement systems such as nominal, ordinal, interval and ratio are used in practice. Let us consider them one by one very briefly.

- (i) *Nominal*: In a nominal system the only relationships possible are whether the quantities involved belong to a group or not without any ordering among the groups.
- (ii) *Ordinal*: In this system we add to the nominal system the ordering of the various groups (e.g. social class, ordinal utility).
- (iii) *Interval*: In this system we add to the ordinal system the possibility of comparing interpoint distances (e.g. measures of temperature). Note that any linear transformation of the values taken is legitimate (Celsius and Fahrenheit scales).
- (iv) *Ratio*: In the ratio system we add to the interval system a natural origin for the values. In such a system the ratio of two values is a meaningful value.

In the case of the linear regression model caution should be exercised when using different measurement systems for the variables involved. In general, most economic variables are of the ratio type and in such a case the

## 410 Specification, estimation and testing

constant term among the regressors is of paramount importance since it estimates a linear function of the means of the variables involved. For this reason *the constant term should always be included in a regression among ratio scaled variables* because it represents the origin of the estimated empirical relationship.

### **Important concepts**

White-noise error term, exogeneity, parametrisation, residuals,  $\hat{R}^2$ ,  $\bar{R}^2$ ,  $\tilde{R}^2$ ,  $R^2$ , specification tests,  $F$ -test, recursive residuals, BLUS residuals, recursive least-squares, measurement system.

### **Questions**

1. Compare the linear Gauss linear and the linear regression models (statistical GM, probability and sampling models).
2. Explain the concept of exogeneity in the context of the linear regression model.
3. Discuss the differences and similarities between

$$\mu_t \equiv E(y_t | \mathbf{X}_t = \mathbf{x}_t), \quad \mu_t^* \equiv E(y_t / \sigma(\mathbf{X}_t)).$$

4. Explain the role of conditioning in the parametrisation of the statistical model.
5. Explain the relationship between normality, linearity and homoskedasticity.
6. Discuss the orthogonality of the estimated systematic and non-systematic components and their relationship to the goodness-of-fit measure  $\tilde{R}^2$ .
7. Compare the goodness-of-fit measures  $\tilde{R}^2$ ,  $\hat{R}^2$  and  $\bar{R}^2$ .
8. State the finite sample properties of the MLE's  $\hat{\theta} \equiv (\hat{\beta}, \hat{\sigma}^2)'$ .
9. State the asymptotic properties of the MLE's  $\hat{\theta} \equiv (\hat{\beta}, \hat{\sigma}^2)'$ .
10. Consider the model  $y_t = \beta t^\alpha + u_t$  and discuss the consistency and asymptotic normality of  $\hat{\beta}$  for  $\alpha = 1$ ,  $\alpha = \frac{1}{2}$ ,  $\alpha = -1$ . (Note:  $\sum_{t=1}^T t = \frac{1}{2}T(T+1)$ ,  $\lim_{T \rightarrow \infty} (\sum_{t=1}^T 1/t^2) = \pi^2/6$ .)
11. Explain why we need to assume that  $\lim_{\tau \rightarrow \infty} (\mathbf{X}'\mathbf{X})^{-1} = \mathbf{0}$  for the consistency of  $\hat{\beta}$  and no such assumption is needed for the consistency of  $\hat{\sigma}^2$ .

### **Exercises**

1. Derive the MLE's of  $\theta \equiv (\beta, \sigma^2)'$ .

2. Derive the information matrix  $\mathbf{I}_T(\theta)$  and discuss the question of full efficiency of  $\hat{\beta}$  and  $\hat{\sigma}^2$ .
3. Show that
  - (i)  $E(\hat{\mu}'\hat{\mathbf{u}}) = 0$ ;
  - (ii)  $\mathbf{P}_x^2 = \mathbf{P}_x$  ( $\mathbf{P}_x = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ );
  - (iii)  $\mathbf{P}'_x = \mathbf{P}_x$ ;
  - (iv)  $(\mathbf{I} - \mathbf{P}_x)\mathbf{X} = \mathbf{0}$ .
4. Derive the distribution of  $\hat{\beta}$  and  $\hat{\sigma}^2$ .
5. Check whether the variables  $X_{1t} = \lambda^t$ ,  $0 < \lambda < 1$  and  $X_{2t} = t$  satisfy the conditions:
  - (i)  $\lim_{T \rightarrow \infty} \left( \frac{1}{T} (\mathbf{X}'\mathbf{X}) \right) = \mathbf{Q}_x < \infty$  non-singular;
  - (ii)  $\lim_{T \rightarrow \infty} (\mathbf{X}'\mathbf{X})^{-1} = \mathbf{0}$ .

Compare these two conditions.

6. Construct a test for  $H_0: \sigma^2 = \sigma_0^2$  against  $H_1: \sigma^2 \neq \sigma_0^2$ .
7. Explain how the following linear restrictions can be accommodated within the general form  $\mathbf{R}\beta = \mathbf{r}$ .
  - (i)  $\beta_1 = \beta_2$ ;
  - (ii)  $\beta_1 + \beta_2 + \beta_3 = 1$ ;
  - (iii)  $\beta_1 = \beta_2 = \beta_3 = 0$ .
8. Show that
 
$$\begin{aligned} (\mathbf{R}\hat{\beta} - \mathbf{r})'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{R}\hat{\beta} - \mathbf{r}) \\ = \mathbf{u}'\mathbf{Q}\mathbf{u} + (\mathbf{R}\beta - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\beta - \mathbf{r}), \end{aligned}$$
 where  $\mathbf{Q} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ . Show that  $\mathbf{Q}^2 = \mathbf{Q}$ .
9. Derive the distribution of  $\tilde{\mu}_{T+1} = \hat{\beta}'\mathbf{x}_{T+1}$  and use it to construct a prediction interval for  $y_{T+1}$ .
10. Compare the prediction errors with the recursive residuals.
11. Derive the distribution of  $\hat{\mathbf{u}}$  and explain how  $\hat{\mathbf{u}}$  can be transformed to define the BLUS residual vector  $\hat{\mathbf{v}} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{T-k})$ .

#### Additional references

Dhrymes (1978); Goldberger (1968); Judge *et al.* (1982); Madansky (1976); Maddala (1977); Malinvaud (1970); Schmidt (1976); Theil (1983).

## CHAPTER 20

---

### The linear regression model II – departures from the assumptions underlying the statistical GM

---

In the previous chapter we discussed the specification of the linear regression model as well as its statistical analysis based on the underlying eight *standard* assumptions. In the next three chapters several departures from [1]–[8] and their implications will be discussed. The discussion differs somewhat from the usual textbook discussion (see Judge *et al.* (1982)) because of the differences in emphasis in the specification of the model.

In Section 20.1 the implications of having  $E(y_i/\sigma(X_i))$  instead of  $E(y_i/X_i = \mathbf{x}_i)$  as the systematic component are discussed. Such a change gives rise to the stochastic regressors model which as a statistical model shares some features with the linear regression model, but the statistical inference related to the statistical parameters of interest  $\theta$  is somewhat different. The statistical parameters of interest and their role in the context of the statistical GM is the subject of Section 20.2. In this section the so-called omitted variables bias problem is reinterpreted as a parameters of interest issue. In Section 20.3 the assumption of exogeneity is briefly considered. The cases where a priori exact linear and non-linear restrictions on  $\theta$  exist are discussed in Section 20.4. Estimation as well as testing when such information is available are considered. Section 20.5 considers the concept of the rank deficiency of  $X$  known as collinearity and its implications. The potentially more serious problem of ‘near collinearity’ is the subject of Section 20.6. Both problems of collinearity and near collinearity are interpreted as insufficient data information for the analysis of the parameters of interest. It is crucially important to emphasise at the outset that the discussion of the various departures from the assumptions underlying the statistical GM which follows assumes that the probability

and sampling models remain valid and unchanged. This assumption is needed because when the probability and/or the sampling model change the whole statistical model requires respecifying.

## 20.1 The stochastic linear regression model

The first assumption underlying the statistical GM is that the systematic component is defined as

$$\mu_t \equiv E(y_t | \mathbf{X}_t = \mathbf{x}_t). \quad (20.1)$$

An alternative but related form of conditioning is the one with respect to the  $\sigma$ -field generated by  $\mathbf{X}_t$ , which defines the systematic component to be

$$\mu_t^* \equiv E(y_t / \sigma(\mathbf{X}_t)). \quad (20.2)$$

The similarities and differences between (1) and (2) were discussed in Section 7.2. In this section we will consider the meaning and intuition underlying (2) as compared with (1). Let  $X_{1t}$  be the first random variable in  $\mathbf{X}_t$ , which is assumed to be defined on the probability space  $(S, \mathcal{F}, P(\cdot))$ .  $\sigma(X_{1t})$  represents the  $\sigma$ -field generated by  $X_{1t}$ , i.e. the minimal  $\sigma$ -field with respect to which  $X_{1t}$  is a random variable. By construction  $\sigma(X_{1t}) \subset \mathcal{F}$ . The  $\sigma$ -field generated by  $\mathbf{X}_t \equiv (X_{1t}, X_{2t}, \dots, X_{kt})$  is defined to be

$$\sigma(\mathbf{X}_t) = \bigcup_{i=1}^k \sigma(X_{it}) \subset \mathcal{F}. \quad (20.3)$$

Let  $y_t$  be also defined on  $(S, \mathcal{F}, P(\cdot))$ . The conditional expectation  $E(y_t / \sigma(\mathbf{X}_t)) : (S, \sigma(\mathbf{X}_t)) \rightarrow (\mathbb{R}, \mathcal{B})$  is defined via

$$\int_{\sigma(\mathbf{X}_t)} y_t dP = \int_{\sigma(\mathbf{X}_t)} E(y_t / \sigma(\mathbf{X}_t)) dP. \quad (20.4)$$

This shows that  $E(y_t / \sigma(\mathbf{X}_t))$  is a *random variable* with respect to  $\sigma(\mathbf{X}_t)$ . Intuitively, conditioning  $y_t$  on  $\sigma(\mathbf{X}_t)$  amounts to considering the part of the random variable  $y_t$  associated with all the events generated by  $\mathbf{X}_t$ . Conditioning on  $\mathbf{X}_t = \mathbf{x}_t$  can be seen as a special case of this where only the event  $\mathbf{X}_t = \mathbf{x}_t$  is considered. Because of this relationship it should come as no surprise to learn that in the case where  $D(y_t, \mathbf{X}_t; \theta)$  is jointly normal the conditional expectations take the form

$$\mu_t \equiv E(y_t | \mathbf{X}_t = \mathbf{x}_t) = \boldsymbol{\sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \mathbf{x}_t, \quad (20.5)$$

$$\mu_t^* \equiv E(y_t / \sigma(\mathbf{X}_t)) = \boldsymbol{\sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \mathbf{X}_t \quad (20.6)$$

(see Chapter 15). Using (6) we can define the *statistical GM*:

$$y_t = \boldsymbol{\beta}' \mathbf{X}_t + u_t, \quad t \in \mathbb{T}, \quad (20.7)$$

where the parameters of interest are  $\theta \equiv (\beta, \sigma^2)$ ,  $\beta \equiv \Sigma_{22}^{-1} \sigma_{21}$ ,  $\sigma^2 = \sigma_{11} - \sigma_{12} \Sigma_{22}^{-1} \sigma_{21}$ . The random vectors  $(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_T)' \equiv \mathcal{X}$  are assumed to satisfy the rank condition,  $\text{rank}(\mathcal{X}) = k$  for any observed value of  $\mathcal{X}$ . The error term defined by

$$u_t \equiv y_t - E(y_t/\sigma(\mathbf{X}_t)), \quad t \in \mathbb{T} \quad (20.8)$$

satisfies the following properties:

$$E(u_t) = E\{E(u_t/\sigma(\mathbf{X}_t))\} = 0, \quad (20.9)$$

$$E(u_t u_t^*) = E\{E(u_t^* u_t/\sigma(\mathbf{X}_t))\} = 0, \quad (20.10)$$

$$E(u_t u_s) = E\{E(u_t u_s/\sigma(\mathbf{X}_t))\} = \begin{cases} \sigma^2, & t = s, \\ 0, & t \neq s, \end{cases} \quad t, s \in \mathbb{T}. \quad (20.11)$$

The statistical GM (7) represents a situation where the systematic component of  $y_t$  is defined as the part of  $y_t$  associated with the events  $\sigma(\mathbf{X}_t)$  and the observed value of  $\mathbf{X}_t$  does not contain all the relevant information. That is, the stochastic structure of  $\mathbf{X}_t$  is of interest as far as it is related to  $y_t$ . This should be contrasted with the statistical GM of the Gauss linear and linear regression models.

Given that  $\mathbf{X}_t$  in the statistical GM is a random vector, intuition suggests that the *probability model* underlying (7) should come in the form of the joint distribution  $D(y_t, \mathbf{X}_t; \psi)$ . We need, however, a form of this distribution which involves the parameters of interest directly. Such a form is readily available using the equality

$$D(y_t, \mathbf{X}_t; \psi) = D(y_t/\mathbf{X}_t; \psi_1) \cdot D(\mathbf{X}_t; \psi_2), \quad (20.12)$$

with  $\theta \equiv (\beta, \sigma^2)$  being a parametrisation of  $\psi_1$ . This suggests that the probability model underlying (7) should take the form

$$\Phi = \{D(y_t/\mathbf{X}_t; \theta) \cdot D(\mathbf{X}_t; \psi_2), \theta \equiv (\beta, \sigma^2) \in \mathbb{R}^k \times \mathbb{R}_+, t \in \mathbb{T}\}, \quad (20.13)$$

where

$$D(y_t/\mathbf{X}_t; \theta) = \frac{1}{\sigma \sqrt{(2\pi)}} \exp \left\{ -\frac{1}{2\sigma^2} (y_t - \beta' \mathbf{X}_t)^2 \right\}, \quad (20.14)$$

$$D(\mathbf{X}_t; \psi_2) = \frac{(\det \Sigma_{22})^{-\frac{1}{2}}}{(2\pi)^{k/2}} \exp \left\{ -\frac{1}{2} \mathbf{X}_t' \Sigma_{22}^{-1} \mathbf{X}_t \right\} \quad (20.15)$$

(see Chapter 15) and  $\psi_2$  are said to be *nuisance parameters* (not of interest).

The random vectors  $\mathbf{X}_1, \dots, \mathbf{X}_T$  become part of the sampling model which is defined as follows:

$(\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_T)$  is a random sample from  $D(\mathbf{Z}_t; \psi)$ ,  $t = 1, 2, \dots, T$ ,

respectively, where as usual

$$\mathbf{Z}_t \equiv \begin{pmatrix} y_t \\ \mathbf{X}_T \end{pmatrix}.$$

If we collect all the above components together we can specify the stochastic linear regression model as follows:

**The statistical GM:**  $y_t = \boldsymbol{\beta}' \mathbf{X}_t + u_t, t \in \mathbb{T}$

- [1]  $\mu_t = E(y_t/\sigma(\mathbf{X}_t))$  and  $u_t = y_t - E(y_t/\sigma(\mathbf{X}_t))$ .
- [2]  $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}, \sigma^2)$ ,  $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\sigma}_{21}$  and  $\sigma^2 = \sigma_{11} - \boldsymbol{\sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\sigma}_{21}$  are the parameters of interest.
- [3]  $\mathbf{X}_t$  is assumed to be weakly exogenous wrt  $\boldsymbol{\theta}$ , for  $t = 1, 2, \dots, T$ .
- [4] No a priori information on  $\boldsymbol{\theta}$ .
- [5] For  $\mathcal{X} \equiv (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_T)'$ ,  $\text{rank}(\mathcal{X}) = k$  for all observable values of  $\mathcal{X}$ ,  $T > k$ .

**The probability model**

$$\Phi = \left\{ D(\mathbf{Z}_t; \boldsymbol{\psi}) = \left[ \frac{1}{\sigma \sqrt{(2\pi)}} \exp \left\{ -\frac{1}{2\sigma^2} (y_t - \boldsymbol{\beta}' \mathbf{X}_t)^2 \right\} \right] \times \left[ \frac{(\det \boldsymbol{\Sigma}_{22})^{-\frac{1}{2}}}{(2\pi)^{k/2}} \exp \left\{ -\frac{1}{2} (\mathbf{X}_t' \boldsymbol{\Sigma}_{22}^{-1} \mathbf{X}_t) \right\} \right], \boldsymbol{\theta} \in (\mathbb{R}^k \times \mathbb{R}_+) \right\}.$$

- [6] (i)  $D(y/\mathbf{X}_t; \boldsymbol{\theta})$  is normal;
- (ii)  $E(y_t/\sigma(\mathbf{X}_t)) = \boldsymbol{\beta}' \mathbf{X}_t$  – linear in  $\mathbf{X}_t$ ;
- (iii)  $\text{Var}(y_t/\sigma(\mathbf{X}_t)) = \sigma^2$  – homoskedastic;
- [7]  $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}, \sigma^2)$  are time-invariant.

**The sampling model**

- [8]  $(\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_T)$  is a random sample from  $D(\mathbf{Z}_t; \boldsymbol{\psi})$ ,  $t = 1, 2, \dots, T$ , respectively.

The assumption related to the weak exogeneity of  $\mathbf{X}_t$  with respect to  $\boldsymbol{\theta}$  for  $t = 1, 2, \dots, T$  shows clearly that the concept is related only to inference on  $\boldsymbol{\theta}$  and not to the statistical inference based on the estimator of  $\boldsymbol{\theta}$ . As shown below, the distribution of the MLE of  $\boldsymbol{\theta}$  depends crucially on the marginal distribution of  $\mathbf{X}_t$ . Hence, for prediction purposes this marginal distribution has a role to play although for efficient estimation and testing on  $\boldsymbol{\theta}$  it is not needed. This shows clearly that the weak exogeneity concept is about statistical parameters of interest and not so much about distributions.

The probability and sampling models taken together imply that for  $\mathbf{y} \equiv (y_1, y_2, \dots, y_T)'$  and  $\mathcal{X} \equiv (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_T)'$  the likelihood function is

$$L(\boldsymbol{\theta}, \mathbf{y}, \mathcal{X}) = \prod_{t=1}^T D(y_t; \mathbf{X}_t; \boldsymbol{\theta}) \cdot D(\mathbf{X}_t; \boldsymbol{\psi}_2). \quad (20.16)$$

The log likelihood takes the form

$$\log L(\boldsymbol{\theta}) = \text{const} - \frac{T}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_t (y_t - \boldsymbol{\beta}' \mathbf{X}_t)^2 + \sum_t \log D(\mathbf{X}_t; \boldsymbol{\psi}_2). \quad (20.17)$$

The last component in (17) can be treated as a constant as far as the differentiation with respect to the parameters of interest  $\boldsymbol{\theta}$  is concerned. Because of the apparent similarity between (17) and the log likelihood function of the linear regression model (see Section 19.4), it should come as no surprise to learn that maximisation with respect to  $\boldsymbol{\theta}$  yields

$$\hat{\boldsymbol{\beta}}^* = \left( \sum_t \mathbf{X}_t \mathbf{X}_t' \right)^{-1} \sum_t \mathbf{X}_t' y_t \equiv (\mathcal{X}' \mathcal{X})^{-1} \mathcal{X}' \mathbf{y} \quad (20.18)$$

and

$$\hat{\sigma}^{*2} = \frac{1}{T} \sum_t (y_t - \hat{\boldsymbol{\beta}}^* \mathbf{X}_t)^2 \equiv \frac{1}{T} \hat{\mathbf{u}}'^* \hat{\mathbf{u}}^* \quad (20.19)$$

in an obvious notation. The same results can be derived by defining the likelihood function in terms of  $D(y_t, \mathbf{X}_t; \boldsymbol{\psi})$  and, after estimating  $\boldsymbol{\sigma}_{11}$ ,  $\boldsymbol{\sigma}_{12}$ ,  $\boldsymbol{\Sigma}_{22}$ , using these estimators and the invariance property of MLE's to construct the corresponding estimators for  $\boldsymbol{\beta} \equiv \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\sigma}_{21}$  and  $\sigma^2 \equiv \boldsymbol{\sigma}_{11} - \boldsymbol{\sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\sigma}_{21}$ . Looking at (18) and (19) we can see that these MLE's of  $\boldsymbol{\beta}$  and  $\sigma^2$  differ from the corresponding estimators for the linear regression model  $\hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}$ ,  $\hat{\sigma}^2 = (1/T) \mathbf{u}' \mathbf{u}$  in so far as the latter include the observed value  $\mathbf{x}_t$  instead of the random vector  $\mathbf{X}_t$ , as above. This difference, however, implies that  $\hat{\boldsymbol{\beta}}^*$  and  $\hat{\sigma}^{*2}$  are no longer a linear and a quadratic function of  $\mathbf{y}$  and thus the distributional results in relation to  $\hat{\boldsymbol{\beta}}$  and  $\hat{\sigma}^2$  cannot be extended to  $\hat{\boldsymbol{\beta}}^*$  and  $\hat{\sigma}^{*2}$ . That is,  $\hat{\boldsymbol{\beta}}^*$  and  $\hat{\sigma}^*$  are no longer normally and chi-square distributed, respectively. In fact, the distributions of these estimators are not analytically tractable at present. As can be seen from (18) and (19) they are very complicated functions of normally distributed random variables.

The question which naturally arises is whether we can derive any properties of  $\hat{\boldsymbol{\beta}}^*$  and  $\hat{\sigma}^{*2}$  without knowing their distributions. Using the properties SCE1–SCE5 (especially SCE3) on conditional expectations with respect to some  $\sigma$ -field (see Section 7.2) we can deduce the following:

$$\hat{\boldsymbol{\beta}}^* = (\mathcal{X}' \mathcal{X})^{-1} \mathcal{X}' (\mathcal{X} \boldsymbol{\beta} + \mathbf{u}) = \boldsymbol{\beta} + (\mathcal{X}' \mathcal{X})^{-1} \mathcal{X}' \mathbf{u} \quad (20.20)$$

$$\begin{aligned} E(\hat{\beta}^*) &= E[E(\hat{\beta}^*/\sigma(\mathcal{X}))] = \beta + E[(\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'E(\mathbf{u}/\sigma(\mathcal{X}))] \\ &= \beta, \end{aligned} \quad (20.21)$$

if  $E[(\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}] < \infty$ , since by the construction of the statistical GM (7),  $E(\mathbf{u}/\sigma(\mathcal{X})) = 0$ . That is,  $\hat{\beta}^*$  is an *unbiased* estimator of  $\beta$ .

$$\begin{aligned} \text{Cov}(\hat{\beta}^*) &\equiv E(\hat{\beta}^* - \beta)(\hat{\beta}^* - \beta)' = E[E(\hat{\beta}^* - \beta)(\hat{\beta}^* - \beta)'/\sigma(\mathcal{X})] \\ &= E[(\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'E(\mathbf{u}\mathbf{u}'/\sigma(\mathcal{X}))\mathcal{X}(\mathcal{X}'\mathcal{X})^{-1}] \\ &= \sigma^2 E(\mathcal{X}'\mathcal{X})^{-1} \end{aligned} \quad (20.22)$$

if  $E(\mathcal{X}'\mathcal{X})^{-1}$  exists, since  $E(\mathbf{u}\mathbf{u}'/\sigma(\mathcal{X})) = \sigma^2 \mathbf{I}_T$ . Using the similarity between the log likelihood function (17) and that of the linear regression model we can deduce that the information matrix in the present case should be of the form

$$\mathbf{I}_T^*(\theta) = \begin{pmatrix} \frac{E(\mathcal{X}'\mathcal{X})}{\sigma^2} & \mathbf{0} \\ \mathbf{0} & \frac{T}{2\sigma^4} \end{pmatrix}. \quad (20.23)$$

This shows that  $\hat{\beta}^*$  is an *efficient* estimator of  $\beta$ .

Using the same properties for the conditional expectation operator we can show that for the MLE  $\hat{\sigma}^{*2}$  of  $\sigma^2$

$$\begin{aligned} E(\hat{\sigma}^{*2}) &= E[E(\hat{\sigma}^{*2}/\sigma(\mathcal{X}))] = \frac{1}{T} E[E(\hat{\mathbf{u}}^{*'}\hat{\mathbf{u}}^*/\sigma(\mathcal{X}))] \\ &= \frac{1}{T} E[E(\mathbf{u}'\mathbf{M}_X\mathbf{u}/\sigma(\mathcal{X}))], \quad \text{where } \mathbf{M}_X = \mathbf{I}_T - \mathcal{X}(\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}' \\ &= \frac{1}{T} E[E(\text{tr } \mathbf{M}_X\mathbf{u}\mathbf{u}'/\sigma(\mathcal{X}))] = \frac{1}{T} E[\text{tr } \mathbf{M}_X E(\mathbf{u}\mathbf{u}'/\sigma(\mathcal{X}))] \\ &= \frac{1}{T} \sigma^2 E(\text{tr } \mathbf{M}_X) = \frac{1}{T} \sigma^2 E(\text{tr } \mathbf{I}_T - \text{tr}(\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\mathcal{X}) \\ &= \left( \frac{T-k}{T} \right) \sigma^2, \quad \text{for all observable values of } \mathcal{X}. \end{aligned} \quad (20.24)$$

This implies that although  $\hat{\sigma}^{*2}$  is a *biased* estimator of  $\sigma^2$  the estimator defined by

$$s^{*2} = \frac{1}{T-k} \hat{\mathbf{u}}^{*'}\hat{\mathbf{u}}^* \quad (20.25)$$

is unbiased.

Using the Lehmann–Scheffe theorem (see Chapter 12) we can show that  $\tau(\mathbf{y}, \mathcal{X}) \equiv (\mathbf{y}'\mathbf{y}, \mathcal{X}'\mathcal{X}, \mathcal{X}'\mathbf{y})$  is a minimal *sufficient statistic* and, as can be seen from (18) and (19), both estimators are functions of this statistic.

Although we were able to derive certain finite properties of the MLE's  $\hat{\beta}^*$  and  $\hat{\sigma}^{*2}$  without having their distribution, no testing or confidence regions are possible without it. For this reason we usually resort to asymptotic theory. Under the assumption

$$\lim_{T \rightarrow \infty} \left( \frac{E(\mathcal{X}'\mathcal{X})}{T} \right) = \mathbf{Q}_{XX} < \infty \quad \text{and non-singular,} \quad (20.26)$$

we deduce that

$$\sqrt{T}(\hat{\beta}^* - \beta) \underset{\alpha}{\sim} N(\mathbf{0}, \sigma^2 \mathbf{Q}_{XX}^{-1}), \quad (20.27)$$

$$\sqrt{T}(\hat{\sigma}^{*2} - \sigma^2) \underset{\alpha}{\sim} N(0, 2\sigma^4). \quad (20.28)$$

These asymptotic distributions can be used to test hypotheses and set up confidence regions when  $T$  is large.

The above discussion of the stochastic linear regression model as a separate statistical model will be of considerable value in the discussion of the dynamic linear regression model in Chapter 23. In that chapter it is argued that the dynamic linear regression model can be profitably viewed as a hybrid of the linear and stochastic linear regression models.

## 20.2 The statistical parameters of interest

The statistical parameters which define the statistical GM are said to be the statistical parameters of interest. In the case of the linear regression model these are  $\beta = \Sigma_{22}^{-1} \sigma_{21}$  and  $\sigma^2 = \sigma_{11} - \sigma_{12} \Sigma_{22}^{-1} \sigma_{21}$ . Estimation of these statistical parameters provides us with an estimated data generating mechanism assumed to have given rise to the observed data in question. The notion of the statistical parameters of interest is of paramount importance because the whole statistical analysis ‘revolves’ around these parameters. A cursory look at assumptions [1]–[8] defining the linear regression model reveals that all the assumptions are directly or indirectly related to the statistical parameters of interest  $\theta$ . Assumption [1] defines the systematic and non-systematic component in terms of  $\theta$ . The assumption of weak exogeneity [3] is defined relative to  $\theta$ . Any a priori information is introduced into the statistical model via  $\theta$ . Assumption [5] referring to the rank of  $\mathbf{X}$  is indirectly related to  $\theta$  because the condition

$$\text{rank}(\mathbf{X}'\mathbf{X}) = k \quad (20.29)$$

is the sample equivalent to the condition

$$\text{rank}(\Sigma_{22}) = k, \quad (20.30)$$

required to ensure that  $\Sigma_{22}$  is invertible and thus the statistical parameters of interest  $\boldsymbol{\theta}$  can be defined. Note that for  $T > k$ ,  $\text{rank}(\mathbf{X}) = \text{rank}(\mathbf{X}'\mathbf{X})$ . Assumptions [6] to [8] are directly related to  $\boldsymbol{\theta}$  in view of the fact that they are all defined in terms of  $D(y_t/\mathbf{X}_t; \boldsymbol{\theta})$ .

The statistical parameters of interest  $\boldsymbol{\theta}$  do not necessarily coincide with the theoretical parameters of interest, say  $\xi$ . The two sets of parameters, however, should be related in such a way as to ensure that  $\xi$  is uniquely defined in terms of  $\boldsymbol{\theta}$ . Only then the theoretical parameters of interest can be given statistical meaning. In such a case  $\xi$  is said to be *identifiable* (see Chapter 25). Empirical econometric models represent reparametrised statistical GM's in terms of  $\xi$ . Their statistical meaning is derived from  $\boldsymbol{\theta}$  and their theoretical meaning through  $\xi$ . As it stands, the statistical GM,

$$y_t = \boldsymbol{\beta}'\mathbf{x}_t + u_t, \quad t \in \mathbb{T}, \quad (20.31)$$

might or might not have any theoretical meaning depending on the mapping

$$\mathbf{G}(\xi, \boldsymbol{\theta}) = \mathbf{0}, \quad (20.32)$$

relating the two sets of parameters. It does, however, have statistical meaning irrespective of the mapping (32). Moreover, the statistical parameters of interest  $\boldsymbol{\theta}$  are not restricted unduly at the outset in order to enable the modeller to test any such testable restrictions. That is, the statistical GM is not restricted to coincide with any theoretical model at the outset. Before any such restrictions are imposed we need to ensure that the estimated statistical GM is well defined statistically; the underlying assumptions [1]–[8] are valid for the data in hand.

The statistical parametrisation  $\boldsymbol{\theta}$  depends crucially on the choice of  $\mathbf{Z}_t$  and its underlying probabilistic structure as summarised in  $D(\mathbf{Z}; \boldsymbol{\psi})$ . Any changes in  $\mathbf{Z}_t$  or/and  $D(\mathbf{Z}; \boldsymbol{\psi})$  changes  $\boldsymbol{\theta}$  as well as the statistical model in question. Hence, caution should be exercised in postulating arguments which depend on different parametrisations, especially when the parametrisations involved are not directly comparable. In order to illustrate this let us consider the so-called omitted variables bias problem.

The textbook discussion of the omitted variables bias argument can be summarised as follows:

The true specification is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\gamma + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_\varepsilon^2 I_T), \quad (20.33)$$

but instead

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} \sim N(\mathbf{0}, \sigma^2 I_T) \quad (20.34)$$

was estimated by ordinary least-squares (OLS) (see Chapter 21), the OLS estimators being

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (20.35)$$

and

$$\hat{\sigma}^2 = \frac{1}{T-k} \hat{\mathbf{u}}'\hat{\mathbf{u}}, \quad \hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}. \quad (20.36)$$

In view of the fact that a comparison between (33) and (34) reveals that

$$\mathbf{u} = \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \quad (20.37)$$

we can deduce that

$$\mathbf{E}(\mathbf{u}) = \mathbf{W}\boldsymbol{\gamma} \neq \mathbf{0} \quad (20.38)$$

and thus

$$(i) \quad E(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\boldsymbol{\gamma} \neq \mathbf{0}; \quad (20.39)$$

and

$$(ii) \quad E(\hat{\sigma}^2) - \sigma^2 = \frac{1}{(T-k)} \boldsymbol{\gamma}' \mathbf{W}' \mathbf{M}_x \mathbf{W} \boldsymbol{\gamma}. \quad (20.40)$$

$\mathbf{M}_x = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ . That is,  $\hat{\boldsymbol{\beta}}$  and  $\hat{\sigma}^2$  suffer from omitted variables bias unless  $\mathbf{W}'\mathbf{X} = \mathbf{0}$  and  $\boldsymbol{\gamma} = \mathbf{0}$ , respectively; see Maddala (1977), Johnston (1984), Schmidt (1976), *inter alia*.

From the textbook specification approach viewpoint, where the statistical model is derived by attaching an error term to the theoretical model, it is impossible to question the validity of the above argument. On the other hand, looking at it from the specification viewpoint proposed in Chapter 19 we can see a number of serious weaknesses in the argument.

The most obvious weakness of the argument is that it depends on two statistical models with different parametrisations. In particular  $\boldsymbol{\beta}$  in (33) and (34) is very different. If we denote the coefficient of  $\mathbf{X}$  in (33) by  $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\sigma}_{21}$ , the same coefficient in  $\mathbf{X}$  takes the form:

$$\boldsymbol{\beta}_0 = \boldsymbol{\Sigma}_{2,3}^{-1} \boldsymbol{\sigma}_{21} - \boldsymbol{\Sigma}_{2,3}^{-1} \boldsymbol{\Sigma}_{23} \boldsymbol{\Sigma}_{33}^{-1} \boldsymbol{\sigma}_{31},$$

where  $\boldsymbol{\Sigma}_{2,3} = (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{23} \boldsymbol{\Sigma}_{33}^{-1} \boldsymbol{\Sigma}_{32})$ ,  $\boldsymbol{\Sigma}_{33} = \text{Cov}(\mathbf{W}_t)$ ,  $\boldsymbol{\Sigma}_{23} = \text{Cov}(\mathbf{X}_t, \mathbf{W}_t)$ ,  $\boldsymbol{\sigma}_{31} = \text{Cov}(\mathbf{W}_t, y_t)$  (see Chapter 15). Moreover, the probability models underlying (33) and (34) are  $D(y_t/\mathbf{X}_t; \boldsymbol{\theta})$  and  $D(y_t/\mathbf{X}_t, \mathbf{W}_t; \boldsymbol{\theta}_0)$  respectively. Once this is realised we can see that the omitted variables bias (39) should be written as

$$E_{y/X, W}(\hat{\beta}) - \beta_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\gamma \neq \mathbf{0}, \quad (20.41)$$

since

$$E_{y/X, W}(\mathbf{u}) = \mathbf{W}\gamma \neq \mathbf{0}, \quad (20.42)$$

where  $E_{y/X, W}(\cdot)$  refers to the expectation operator defined in terms of  $D(y_t/\mathbf{X}_t, \mathbf{W}_t; \theta_0)$ . Looking at (41) we can see that the omitted variables bias arises when we try to estimate  $\beta_0$  in

$$\mathbf{y} = \mathbf{X}\beta_0 + \mathbf{W}\gamma + \boldsymbol{\epsilon} \quad (20.43)$$

by estimating  $\beta$  in (34) where, by construction,  $\beta \neq \beta_0$ . On the other hand, in the context of the same statistical model,

$$E_{y/X}(\hat{\beta}) - \beta = \mathbf{0} \quad (20.44)$$

since

$$E_{y/X}(\mathbf{u}) = \mathbf{0} \quad (20.45)$$

and no omitted variables problem arises. A similar argument can be made for  $\hat{\sigma}^2$ . From this viewpoint the question of estimating the statistical parameters of interest  $\theta_0 \equiv (\beta_0, \gamma, \sigma_e^2)$  by estimating  $\theta \equiv (\beta, \sigma^2)$  never arises since the two parameter sets  $\theta_0$  and  $\theta$  depend on different sample information,  $\mathcal{F}_0 = \sigma(y_t, \mathbf{X}_t, \mathbf{W}_t, t = 1, 2, \dots, T)$  and  $\mathcal{F} = \sigma(y_t, \mathbf{X}_t, t = 1, 2, \dots, T)$  respectively. This, however, does not imply that the omitted variables argument is useless, quite the opposite. In cases where the sample information is the same ( $\mathcal{F}_0 = \mathcal{F}$ ) the argument can be very useful in deriving misspecification tests (see Chapters 21 and 22). For further discussion of this issue see Spanos (1985b).

The above argument illustrates the dangers of not specifying explicitly the underlying probability model and the statistical parameters of interest. By changing the underlying probability distribution and the parametrisation the results on bias disappear. The two parametrisations are only comparable when they are both derivable from the joint distribution,  $D(\mathbf{Z}_1, \dots, \mathbf{Z}_T; \psi)$  using alternative ‘reduction’ arguments.

### 20.3 Weak exogeneity

When the random vector  $\mathbf{X}_t$  is assumed to be weakly exogenous in the context of the linear regression model it amounts to postulating that the stochastic structure of  $\mathbf{X}_t$ , as specified by its marginal distribution  $D(\mathbf{X}_t; \psi_2)$ , is not relevant as far as the statistical inference on the parameters of interest

$\theta \equiv (\beta, \sigma^2)$  is concerned. That is, although at the outset we postulate  $D(y_t, \mathbf{X}_t; \psi)$  as far as the parameters of interest are concerned,  $D(y_t/\mathbf{X}_t; \psi_1)$  suffices; note that

$$D(y_t, \mathbf{X}_t; \psi) = D(y_t/\mathbf{X}_t; \psi_1) \cdot D(\mathbf{X}_t; \psi_2) \quad (20.46)$$

is true for any joint distribution (see Chapter 5). If we want to test the exogeneity assumption we need to specify  $D(\mathbf{X}_t; \psi_2)$  and consider it in relation to  $D(y_t/\mathbf{X}_t; \psi_1)$  (see Wu (1973), Engle (1984); *inter alia*). These exogeneity tests usually test certain implications of the exogeneity assumption and this can present various problems. The implications of exogeneity tested depend crucially on the other assumptions of the model as well as the appropriate specification of the statistical GM giving rise to  $\mathbf{x}_t$ ,  $t = 1, 2, \dots, T$ ; see Engle *et al.* (1983).

Exogeneity in this context will be treated as a non-directly testable assumption and no exogeneity tests will be considered. It will be argued in Chapter 21 that exogeneity assumptions can be tested indirectly by testing the assumptions [6]–[8]. The argument in a nutshell is that when inappropriate marginalisation and conditioning are used in defining the parameters of interest the assumptions [6]–[8] are unlikely to be valid (see Engle *et al.* (1983), Richard (1982)). For example a way to ‘test’ the weak exogeneity assumption indirectly is to test for departures from the normality of  $D(y_t, \mathbf{X}_t; \psi)$  using the implied normality of  $D(y_t/\mathbf{X}_t; \theta)$  and homoskedasticity of  $\text{Var}(y_t/\mathbf{X}_t = \mathbf{x}_t)$ . For instance, in the case where  $D(y_t, \mathbf{X}_t; \psi)$  is multivariate Student’s *t*, the parameters  $\psi_1$  and  $\psi_2$  above are no longer variation free (see Section 21.4). Testing for departures from normality in the directions implied by  $D(y_t, \mathbf{X}_t; \psi)$  being multivariate *t* can be viewed as an indirect test for the variation free assumption underlying weak exogeneity.

## 20.4 Restrictions on the statistical parameters of interest $\theta$

The statistical inference results on the linear regression model derived in Chapter 19 are based on the assumption that no a priori information on  $\theta \equiv (\beta, \sigma^2)$  is available. Such a priori information, when available, can take various forms such as linear, non-linear, exact, inexact or stochastic. In this section only exact a priori information on  $\beta$  and its implications will be considered; a priori information on  $\sigma^2$  is rather scarce.

### (1) Linear a priori restrictions on $\beta$

Let us assume that a priori information in the form of  $m$  linear restrictions

$$\mathbf{R}\beta = \mathbf{r} \quad (20.47)$$

is also available at the outset, where  $\mathbf{R}$  and  $\mathbf{r}$  are  $m \times k$  and  $m \times 1$  known matrices,  $\text{rank}(\mathbf{R})=m$ . Such restrictions imply that the parameter space where  $\boldsymbol{\beta}$  takes values in no longer  $\mathbb{R}^k$  but some subset of it as determined by (47). These restrictions represent information relevant for the statistical analysis of the linear regression model and can be taken into consideration.

In the estimation of  $\boldsymbol{\theta}$  these restrictions can be taken into consideration by extending the concept of the log likelihood function to include such restrictions. This is achieved by defining the Lagrangian function to be

$$l(\boldsymbol{\theta}, \boldsymbol{\mu}; \mathbf{y}, \mathbf{X}) = \text{const} - \frac{T}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \boldsymbol{\mu}'(\mathbf{R}\boldsymbol{\beta} - \mathbf{r}), \quad (20.48)$$

where  $\boldsymbol{\mu}$  represents an  $m \times 1$  vector of *Lagrange multipliers*. Optimisation of (48) with respect to  $\boldsymbol{\beta}$ ,  $\sigma^2$  and  $\boldsymbol{\mu}$  gives rise to the first-order conditions:

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2} (\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta}) - \mathbf{R}'\boldsymbol{\mu} = \mathbf{0}, \quad (20.49)$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0, \quad (20.50)$$

$$\frac{\partial l}{\partial \boldsymbol{\mu}} = -(\mathbf{R}\boldsymbol{\beta} - \mathbf{r}) = \mathbf{0}. \quad (20.51)$$

Premultiplying (49) with  $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}$  (to make the second term invertible) and solving for  $\boldsymbol{\mu}$  we get

$$\tilde{\boldsymbol{\mu}} = [\sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}), \quad (20.52)$$

where  $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$  is the *unconstrained MLE* of  $\boldsymbol{\beta}$ . This in turn implies that the *constrained MLE* of  $\boldsymbol{\beta}$  is

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) \quad (20.53)$$

and

$$\mathbf{R}\tilde{\boldsymbol{\beta}} - \mathbf{r} = \mathbf{0}. \quad (20.54)$$

From (50) we can deduce that the *constrained MLE* of  $\sigma^2$  is

$$\begin{aligned} \tilde{\sigma}^2 &= \frac{1}{T} (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \\ &= \frac{1}{T} \tilde{\mathbf{u}}'\tilde{\mathbf{u}}, \quad \tilde{\mathbf{u}} \equiv \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}. \end{aligned} \quad (20.55)$$

*Properties of  $\tilde{\theta} \equiv (\tilde{\beta}, \tilde{\mu}, \tilde{\sigma}^2)$*

Using these formulae we can derive the distributions of the constrained MLE's of  $\beta$ ,  $\sigma^2$  and  $\mu$ .  $\tilde{\beta}$  and  $\tilde{\mu}$  being linear functions of  $\hat{\beta}$  we can deduce that

$$\begin{pmatrix} \tilde{\beta} \\ \tilde{\mu} \end{pmatrix} \sim N \left( \begin{pmatrix} \beta - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\beta - \mathbf{r}) \\ [\sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\beta - \mathbf{r}) \end{pmatrix}, \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix} \right), \quad (20.56)$$

where

$$\mathbf{C}_{11} = \sigma^2 [(\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}] \equiv \text{Cov}(\hat{\beta}),$$

$$\mathbf{C}_{12} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} \equiv \text{Cov}(\tilde{\beta}, \tilde{\mu}) = \mathbf{C}'_{21},$$

$$\mathbf{C}_{22} = [\sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} \equiv \text{Cov}(\tilde{\mu}).$$

Using (56) we can deduce that

- (i) When  $\mathbf{R}\beta = \mathbf{r}$ ,  $E(\tilde{\beta}) = \beta$  and  $E(\tilde{\mu}) = \mathbf{0}$ , i.e.  $\tilde{\beta}$  and  $\tilde{\mu}$  are unbiased estimators of  $\beta$  and  $\mathbf{0}$ , respectively.
- (ii)  $\tilde{\beta}$  and  $\tilde{\mu}$  are fully efficient estimators of  $\beta$  and  $\mu$  since their variances achieve the Cramer–Rao lower bounds as can be verified directly using the extended information matrix:

$$\mathbf{I}_T(\beta, \mu, \sigma^2) = \begin{pmatrix} \frac{\mathbf{X}'\mathbf{X}}{\sigma^2} & \mathbf{R}' & \mathbf{0} \\ \mathbf{R} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{T}{2\sigma^4} \end{pmatrix} \quad (20.57)$$

(see exercises 1 and 2).

- (iii)  $[\text{Cov}(\tilde{\beta}) - \text{Cov}(\hat{\beta})] \leq 0$ , i.e. the covariance of the constrained MLE  $\tilde{\beta}$  is always less than or equal to the covariance of the unconstrained MLE  $\hat{\beta}$ , irrespective of whether  $\mathbf{R}\beta = \mathbf{r}$  holds or not; but  $[\text{MSE}(\tilde{\beta}) - \text{MSE}(\hat{\beta})] \geq 0$  where MSE stands for mean square error (see Chapter 12).

The constrained MLE of  $\sigma^2$  can be written in the form

$$\tilde{\sigma}^2 = \hat{\sigma}^2 + \frac{1}{T} (\mathbf{R}\hat{\beta} - \mathbf{r})' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} (\mathbf{R}\hat{\beta} - \mathbf{r}), \quad (20.58)$$

since

$$\tilde{\mathbf{u}} = \hat{\mathbf{u}} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta} - \mathbf{r}). \quad (20.59)$$

Given that

$$\frac{T\hat{\sigma}^2}{\sigma^2} \sim \chi^2(T-k)$$

and

$$(\mathbf{R}\hat{\beta} - \mathbf{r})' \frac{[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}}{\sigma^2} (\mathbf{R}\hat{\beta} - \mathbf{r}) \sim \chi^2(m; \delta), \quad (20.60)$$

where

$$\delta = \frac{(\mathbf{R}\beta - \mathbf{r})' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} (\mathbf{R}\beta - \mathbf{r})}{\sigma^2}, \quad (20.61)$$

we can deduce that

$$\frac{T\tilde{\sigma}^2}{\sigma^2} \sim \chi^2(T+m-k; \delta), \quad (20.62)$$

using the reproductive property of the chi-square (see Appendix 6.1) and the independence of the two components in (59). This implies

$$E(\tilde{\sigma}^2) \neq \sigma^2. \quad (20.63)$$

But for  $\tilde{s}^2 = [1/(T+m-k)]\tilde{\mathbf{u}}'\tilde{\mathbf{u}}$ ,  $E(\tilde{s}^2) = \sigma^2$  when  $\mathbf{R}\beta = \mathbf{r}$ , since  $\delta = 0$ .

### The F-test revisited

In Section 19.5 above we derived the F-test based on the test statistic

$$\tau(\mathbf{y}) = \frac{(\mathbf{R}\hat{\beta} - \mathbf{r})' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} (\mathbf{R}\hat{\beta} - \mathbf{r})}{ms^2} \quad (20.64)$$

for the null hypothesis

$$H_0: \mathbf{R}\beta = \mathbf{r} \text{ against } H_1: \mathbf{R}\beta \neq \mathbf{r},$$

using the intuitive argument that when  $H_0$  is valid  $\|\mathbf{R}\hat{\beta} - \mathbf{r}\|$  must be close to zero. We can derive the same test using various other intuitive arguments similar to this in relation to quantities like  $\|\hat{\beta} - \beta\|$  and  $\|\tilde{\mu}\|$  being close to zero when  $H_0$  is valid (see question 5). A more formal derivation of the F-test can be based on the likelihood ratio test procedure (see Chapter 14).

The above null and alternative hypotheses in the language of Chapter 14 can be written as

$$H_0: \theta \in \Theta_0, \quad H_1: \theta \in \Theta_1 \equiv \Theta - \Theta_0,$$

where

$$\theta \equiv (\beta, \sigma^2), \quad \Theta = \{(\beta, \sigma^2): \beta \in \mathbb{R}^k, \sigma^2 \in \mathbb{R}_+\},$$

$$\Theta_0 = \{(\beta, \sigma^2): \beta \in \mathbb{R}^k, \mathbf{R}\beta = \mathbf{r}, \sigma^2 \in \mathbb{R}_+\}.$$

The likelihood ratio takes the form

$$\lambda(\mathbf{y}) = \frac{\max_{\theta \in \Theta_0} L(\theta; \mathbf{y})}{\max_{\theta \in \Theta} L(\theta; \mathbf{y})} = \frac{L(\tilde{\theta}; \mathbf{y})}{L(\hat{\theta}; \mathbf{y})} = \frac{(2\pi)^{-T/2}(\tilde{\sigma}^2)^{-T/2} e^{-T/2}}{(2\pi)^{-T/2}(\hat{\sigma}^2)^{-T/2} e^{-T/2}} = \left( \frac{\tilde{\sigma}^2}{\hat{\sigma}^2} \right)^{-T/2}. \quad (20.65)$$

The problem we have to face at this stage is to determine the distribution of  $\lambda(y)$  or some monotonic function of it. Using (58) we can write  $\lambda(y)$  in the form

$$\lambda(y) = \left[ 1 + \frac{(\mathbf{R}\hat{\beta} - \mathbf{r})' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} (\mathbf{R}\hat{\beta} - \mathbf{r})}{(T-k)s^2} \right]^{-T/2}. \quad (20.66)$$

Looking at (66) we can see that it is directly related to (64) whose distribution we know. Hence,  $\lambda(y)$  can be transformed into the  $F$ -test using the monotonic transformation

$$\tau(y) = (\lambda(y)^{-2/T} - 1) \left( \frac{T-k}{m} \right). \quad (20.67)$$

This transformation provides us with an alternative way to calculate the value of the test statistic  $\tau(y)$  using the estimates of the restricted and unrestricted MLE's of  $\sigma^2$ . An even simpler operational form of  $\tau(y)$  can be specified using the equality (58). From this equality we can deduce that

$$(\mathbf{R}\hat{\beta} - \mathbf{r})' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} (\mathbf{R}\hat{\beta} - \mathbf{r}) = \tilde{\mathbf{u}}'\tilde{\mathbf{u}} - \hat{\mathbf{u}}'\hat{\mathbf{u}} \quad (20.68)$$

(see exercise 4). This implies that  $\tau(y)$  can be written in the form

$$\tau(y) = \frac{\tilde{\mathbf{u}}'\tilde{\mathbf{u}} - \hat{\mathbf{u}}'\hat{\mathbf{u}}}{\hat{\mathbf{u}}'\hat{\mathbf{u}}} \left( \frac{T-k}{m} \right) \quad (20.69)$$

or

$$\tau(y) = \frac{\text{RRSS} - \text{URSS}}{\text{URSS}} \left( \frac{T-k}{m} \right), \quad (20.70)$$

where RRSS and URSS stand for *restricted* and *unrestricted residual sums of squares*, respectively. This is a more convenient form because most computer packages report the RSS and instead of going through the calculation needed for (64) we estimate the regression equation with and without the restrictions and use the RSS in the two cases to calculate  $\tau(y)$  as in (70).

### Example

Let us return to the money equation estimated in Chapter 19:

$$m_t = 2.896 + 0.690y_t + 0.865p_t - 0.055i_t + \hat{u}_t, \quad (20.71) \\ (1.034) \ (0.105) \ (0.020) \ (0.013) \ (0.04)$$

$$R^2 = 0.995, \bar{R}^2 = 0.995, s = 0.0393,$$

$$\log L = 147.4, \text{ RSS} = 0.11752, T = 80.$$

Assuming that this is a well-defined estimated statistical GM (a very questionable assumption) we can proceed to consider specification tests related to a priori restrictions on the parameters of interest. One set of such a priori restrictions which is interesting from the economic theory viewpoint is

$$H_0: \beta_2 = 1 \text{ and } \beta_3 = 1 \text{ against } H_1: \beta_2 \neq 1 \text{ or } \beta_3 \neq 1.$$

Interpreting  $\beta_2$  and  $\beta_3$  as income and price elasticities, respectively, we can view  $H_0$  as a unit elasticity hypothesis.

In order to use the form of the  $F$ -test (linear restrictions) as specified in (70) we need to re-estimate (71) imposing the restrictions. This estimation yielded

$$(m_t - p_t - y_t) = -0.529 - 0.219i_t + \hat{u}_t, \quad (20.72)$$

(0.055) (0.019) (0.087)

$$R^2 = 0.629, \quad \bar{R}^2 = 0.624, \quad s = 0.0866,$$

$$\log L = 83.18, \quad \text{RSS} = 0.58552, \quad T = 80.$$

Given that  $\text{RRSS} = 0.58552$  and  $\text{URSS} = 0.11752$  we can deduce that

$$\tau(\mathbf{y}) = \left( \frac{0.58552 - 0.11752}{0.11752} \right) \left( \frac{76}{2} \right) = 151.327. \quad (20.73)$$

For a size  $\alpha = 0.05$  test the rejection region is

$$C_1 = \{\mathbf{y}: \tau(\mathbf{y}) \geq 3.12\}. \quad (20.74)$$

Hence we can conclude that  $H_0$  is strongly rejected. It must be stressed, however, that this is a specification test and is based on the presupposition that all the assumptions underlying the linear regression model are valid. From the limited analysis of this estimated equation in Chapter 19 there are clear signs such as its predictive ability and the residual's time pattern that some of the underlying assumptions might be invalid. In such a case the above conclusion based on the  $F$ -test might be very misleading.

The above form of the  $F$ -test will play a very important role in the context of misspecification testing to be considered in Chapters 21–23.

## (2) *Exact non-linear restrictions on $\beta$* <sup>†</sup>

Having considered the estimation and testing of the linear regression model when a priori information in the form of exact linear restrictions on  $\beta$  we turn to exact non-linear restrictions.

<sup>†</sup> This section relies heavily on Chapter 16.

Consider the case where a priori information comes in the form of  $m$  non-linear restrictions (e.g.  $\beta_1 = \beta_2/\beta_3$ ,  $\beta_1 = -\beta_2^2$ ):

$$h_i(\boldsymbol{\beta}) = 0, \quad i = 1, 2, \dots, m,$$

or, in matrix form:

$$\mathbf{H}(\boldsymbol{\beta}) = \mathbf{0}. \quad (20.75)$$

In order to ensure independence between the  $m$  restrictions we assume that

$$\text{rank}\left(\frac{\partial \mathbf{H}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right) = m. \quad (20.76)$$

As in the case of the linear restrictions, let us consider first the question of constructing a test for the null hypothesis

$$H_0: \mathbf{H}(\boldsymbol{\beta}) = \mathbf{0} \quad \text{against} \quad H_1: \mathbf{H}(\boldsymbol{\beta}) \neq \mathbf{0}. \quad (20.77)$$

Using the same intuitive argument as the one which served us so well in constructing the  $F$ -test (see Section 19.5) we expect that when  $H_0$  is valid,  $\mathbf{H}(\hat{\boldsymbol{\beta}}) \approx \mathbf{0}$ . The problem then becomes one of constructing a test statistic based on the distance

$$\|\mathbf{H}(\hat{\boldsymbol{\beta}}) - \mathbf{0}\|. \quad (20.78)$$

Following the same arguments in relation to the units of measurement and the absolute value we might transform (78) into

$$\mathbf{H}(\hat{\boldsymbol{\beta}})' [\text{Cov}(\mathbf{H}(\hat{\boldsymbol{\beta}}))]^{-1} \mathbf{H}(\hat{\boldsymbol{\beta}}). \quad (20.79)$$

Unlike the case of the linear restrictions, however, we do not know the distribution of (79) and thus it is of no use as a test statistic. This is because although we know that  $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$  the distribution of  $\mathbf{H}(\hat{\boldsymbol{\beta}})$  is no longer normal, being a non-linear function of  $\hat{\boldsymbol{\beta}}$ . Although, for some non-linear functions  $h_i(\hat{\boldsymbol{\beta}})$  we might be able to derive their distribution, it is of little value because we need general results which can be used for any non-linear restrictions. The construction of the  $F$ -test suggests that if we could linearise  $\mathbf{H}(\hat{\boldsymbol{\beta}})$  such a general test could be derived along similar lines as the  $F$ -test. *Linearisation of  $\mathbf{H}(\hat{\boldsymbol{\beta}})$*  can be achieved by taking its first-order Taylor's expansion at  $\boldsymbol{\beta}$ , i.e.

$$\mathbf{H}(\hat{\boldsymbol{\beta}}) = \mathbf{H}(\boldsymbol{\beta}) + \frac{\partial \mathbf{H}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + o_p(1), \quad (20.80)$$

where  $o_p(1)$  stands for ‘asymptotically negligible terms’ (see Chapter 10); (80) provides us with a linearised version of  $\mathbf{H}(\hat{\boldsymbol{\beta}})$  at the expense of the approximation. The fact that we chose to ignore all the higher terms as

asymptotically negligible implies that any result based on (80) can only be justified *asymptotically*. Hence, any tests based on (80) can only be asymptotic. What we could not get in finite sample theory (linearity), we get by going asymptotic. Given that

$$\sqrt{\alpha} T(\hat{\beta} - \beta) \sim N(\mathbf{0}, \sigma^2 \mathbf{Q}_X^{-1}), \quad (20.81)$$

we can deduce that

$$\sqrt{\alpha} T(\mathbf{H}(\hat{\beta}) - \mathbf{H}(\beta)) \sim N\left(\mathbf{0}, \sigma^2 \left[ \left( \frac{\partial \mathbf{H}(\beta)}{\partial \beta} \right) \mathbf{Q}_X^{-1} \left( \frac{\partial \mathbf{H}(\beta)}{\partial \beta} \right)' \right] \right) \quad (20.82)$$

(see N1 in Chapter 15). This result implies that if we substitute the asymptotic covariance ( $\text{Cov}_{\alpha}(\mathbf{H}(\hat{\beta}))$ ) in (79) we could get an asymptotic test because

$$\mathbf{H}(\hat{\beta})' [\text{Cov}_{\alpha}(\mathbf{H}(\hat{\beta})}]^{-1} \mathbf{H}(\hat{\beta}) \sim \chi^2(m; \delta), \quad (20.83)$$

where

$$\text{Cov}_{\alpha}(\hat{\beta}) = \sigma^2 \left[ \left( \frac{\partial \mathbf{H}(\beta)}{\partial \beta} \right) \mathbf{Q}_X^{-1} \left( \frac{\partial \mathbf{H}(\beta)}{\partial \beta} \right)' \right]$$

and

$$\delta = \mathbf{H}(\beta)' [\text{Cov}_{\alpha}(\hat{\beta})]^{-1} \mathbf{H}(\beta). \quad (20.84)$$

As it stands (83) cannot be used as a test statistic because  $\beta$  and  $\sigma^2$  are unknown and  $\mathbf{Q}_X$  is not available. Given, however, that  $s^2 \xrightarrow{P} \sigma^2$  and  $\lim_{T \rightarrow \infty} (\mathbf{X}'\mathbf{X})/T = \mathbf{Q}_X$  we can deduce that

$$s^2 \left[ \left( \frac{\partial \mathbf{H}(\hat{\beta})}{\partial \beta} \right) \left( \frac{\mathbf{X}'\mathbf{X}}{T} \right)^{-1} \left( \frac{\partial \mathbf{H}(\hat{\beta})}{\partial \beta} \right)' \right] \xrightarrow{\alpha} \text{Cov}(\hat{\beta}), \quad (20.85)$$

where

$$\frac{\partial \mathbf{H}(\hat{\beta})}{\partial \beta} = \frac{\partial \mathbf{H}(\beta)}{\partial \beta} \Big|_{\beta=\hat{\beta}}.$$

This enables us to construct the test statistic

$$W(\mathbf{y}) = \mathbf{H}(\hat{\beta})' \left[ s^2 \left( \frac{\partial \mathbf{H}(\hat{\beta})}{\partial \beta} \right) (\mathbf{X}'\mathbf{X})^{-1} \left( \frac{\partial \mathbf{H}(\hat{\beta})}{\partial \beta} \right)' \right]^{-1} \mathbf{H}(\hat{\beta}) \sim \chi^2(m; \delta), \quad (20.86)$$

which can be used to define an  $\alpha$  size asymptotic test based on the rejection region

$$C_1 = \{ \mathbf{y}: W(\mathbf{y}) > c_{\alpha} \}, \quad (20.87)$$

where

$$\int_{c\alpha}^{\infty} d\chi^2(m) = \alpha.$$

This is because

$$W(\mathbf{y}) \stackrel{H_0}{\sim} \chi^2(m). \quad (20.88)$$

This test is known as the *Wald test* whose general form was discussed in Chapter 16. In the same chapter two other asymptotic test procedures which give rise to asymptotically equivalent tests were also discussed. These are the Lagrange multiplier and the likelihood ratio asymptotic test procedures.

The Lagrange multiplier test procedure is based on the *constrained MLE*  $\tilde{\beta}$  of  $\beta$  instead of the unconstrained MLE  $\hat{\beta}$ . That is,  $\tilde{\beta}$  is derived from the optimisation of the Lagrangian function

$$l(\theta, \mu; \mathbf{y}) = \log L(\theta; \mathbf{y}) - \mu' \mathbf{H}(\beta) \quad (20.89)$$

via

$$\frac{\partial l}{\partial \beta} = \frac{\partial \log L}{\partial \beta} - \frac{\partial \mathbf{H}(\beta)}{\partial \beta} \mu(\tilde{\beta}) = 0$$

and

$$\frac{\partial l}{\partial \mu} = -\mathbf{H}(\tilde{\beta}) = 0, \quad (20.90)$$

$$\Rightarrow \frac{\partial \mathbf{H}(\beta)}{\partial \beta} \mu(\tilde{\beta}) = \frac{\partial \log L}{\partial \beta} (\tilde{\theta}; \mathbf{y}) \quad \text{and} \quad \mathbf{H}(\tilde{\beta}) = 0, \quad (20.91)$$

$\tilde{\theta} \equiv (\tilde{\beta}, \tilde{\sigma}^2)$  is the constrained MLE of  $\theta \equiv (\beta, \sigma^2)$ . In order to understand what is involved in (89)–(91), it is advisable to compare these with (48), (49) and (52) above for the linear restrictions case. In direct analogy to the linear restrictions case the distance

$$\|\mu(\tilde{\beta}) - \mathbf{0}\| \quad (20.92)$$

can be used as a measure of whether  $H_0$  is true or not. Using the same intuitive argument as for  $\|\mathbf{H}(\hat{\beta}) - \mathbf{0}\|$  above we can construct the quantity

$$\frac{1}{T} \mu(\tilde{\beta})' [\operatorname{Cov}_{\alpha}(\mu(\tilde{\beta}))]^{-1} \mu(\tilde{\beta}) \quad (20.93)$$

as the basis of a possible test statistic. It can be shown that

$$\frac{1}{\sqrt{T}} (\mu(\tilde{\beta}) - \mu(\beta)) \sim N \left( \mathbf{0}, \sigma^2 \left[ \left( \frac{\partial \mathbf{H}(\beta)}{\partial \beta} \right) \mathbf{Q}_X^{-1} \left( \frac{\partial \mathbf{H}(\beta)}{\partial \beta} \right)' \right]^{-1} \right) \quad (20.94)$$

(see Chapter 16). This suggests that the statistic

$$LM(\mathbf{y}) = \boldsymbol{\mu}(\tilde{\boldsymbol{\beta}})' \left[ \hat{\sigma}^2 \left( \frac{\partial \mathbf{H}(\tilde{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} \right) (\mathbf{X}'\mathbf{X})^{-1} \left( \frac{\partial \mathbf{H}(\tilde{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} \right)' \right] \boldsymbol{\mu}(\tilde{\boldsymbol{\beta}}) \sim \chi^2(m; \delta) \quad (20.95)$$

can be used to construct an  $\alpha$  size test with rejection region

$$C_1 = \{ \mathbf{y}: LM(\mathbf{y}) > c_\alpha \} \quad (20.96)$$

and power function

$$\mathcal{P}(\boldsymbol{\beta}) = Pr(LM(\mathbf{y}) > c_\alpha) = \int_{c_\alpha}^\infty d\chi^2(m; \delta). \quad (20.97)$$

In Chapter 16 it was shown that the Lagrange multiplier test can take an alternative formulation based on the *score function*  $[\partial \log L(\tilde{\boldsymbol{\theta}})]/\partial \boldsymbol{\theta}$ . In view of the relationship between the score function and the Lagrange multipliers given in (91) we can deduce that we can construct a test for  $H_0$  against  $H_1$  based on the distance

$$\left| \left| \frac{\partial \log L(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} - 0 \right| \right|. \quad (20.98)$$

Following the same argument as in the case of the construction of  $W(\mathbf{y})$  and  $LM(\mathbf{y})$  we can suggest that the quantity

$$\frac{\partial \log L(\tilde{\boldsymbol{\theta}})'}{\partial \boldsymbol{\theta}} \left( \text{Cov}_{\alpha} \left( \frac{\partial \log L(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \right) \right)^{-1} \frac{\partial \log L(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \quad (20.99)$$

should form the basis of another reasonable test statistic. Given that

$$\frac{1}{\sqrt{T}} \frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \underset{\alpha}{\sim} N(\mathbf{0}, \mathbf{I}_\infty(\boldsymbol{\theta})) \quad (20.100)$$

(see Chapters 13 and 16) we can deduce the following test statistic:

$$ES(\mathbf{y}) = \frac{1}{T} \left( \frac{\partial \log L(\tilde{\boldsymbol{\theta}})'}{\partial \boldsymbol{\theta}} \right) I_\infty(\tilde{\boldsymbol{\theta}})^{-1} \left( \frac{\partial \log L(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \right) \underset{\alpha}{\sim} \chi^2(m). \quad (20.101)$$

This test statistic can be simplified further by noting that  $H_0$  involves only a subset of  $\boldsymbol{\theta}$  (just  $\boldsymbol{\beta}$ ) and  $I_\infty(\boldsymbol{\theta})$  is block diagonal. Using the results of Chapter 16 we can deduce that

$$ES(\mathbf{y}) = \left( \frac{\partial \log L(\tilde{\boldsymbol{\beta}}, 0)}{\partial \boldsymbol{\beta}} \right)' [\hat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1}] \left( \frac{\partial \log L(\tilde{\boldsymbol{\beta}}, 0)}{\partial \boldsymbol{\beta}} \right). \quad (20.102)$$

The test statistic  $ES(\mathbf{y})$  constitutes what is sometimes called the *efficient score* form of the Lagrange multiplier test.

*The likelihood ratio test* is based on the test statistic (see Chapter 16):

$$LR(\mathbf{y}) = -2(\log L(\tilde{\boldsymbol{\theta}}; \mathbf{y}) - \log L(\hat{\boldsymbol{\theta}}; \mathbf{y})) \underset{x}{\sim} \chi^2(m; \delta), \quad (20.103)$$

with a rejection region

$$C_1 = \{\mathbf{y}: LR(\mathbf{y}) > c_\alpha\}. \quad (20.104)$$

Using the first-order Taylor's expansion we can approximate  $LR(\mathbf{y})$  as

$$LR(\mathbf{y}) \simeq T(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})' I_\infty(\boldsymbol{\theta})(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}). \quad (20.105)$$

This approximation is very suggestive because it shows clearly an important feature shared by all four test statistics  $W(\mathbf{y})$ ,  $LM(\mathbf{y})$ ,  $ES(\mathbf{y})$  and  $LR(\mathbf{y})$ . All four are based on the intuitive argument that some distance  $\|\mathbf{H}(\hat{\boldsymbol{\beta}})\|$ ,  $\|\boldsymbol{\mu}(\tilde{\boldsymbol{\beta}})\|$ ,  $\|[\hat{\epsilon} \log L(\tilde{\boldsymbol{\beta}}, 0)]/\hat{\epsilon}\boldsymbol{\beta}\|$  and  $\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\|$ , respectively, is 'close to zero' when  $H_0$  is valid.

Another important feature shared by all four test statistics is that their asymptotic distribution ( $\chi^2(m)$  under  $H_0$ ) depends crucially on the *asymptotic normality* of a certain quantity involved in defining these distances,  $\sqrt{T}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$ ,  $(1/\sqrt{T})(\boldsymbol{\mu}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{\mu}(\boldsymbol{\theta}))$ ,  $(1/\sqrt{T})[\hat{\epsilon} \log L(\tilde{\boldsymbol{\beta}}, 0)]/\hat{\epsilon}\boldsymbol{\beta}$  and  $\sqrt{T}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})$  respectively. All three tests, the Wald ( $W$ ), Lagrange multiplier ( $LM$ ) and likelihood ratio ( $LR$ ) are asymptotically equivalent, in the sense that they have the same asymptotic power characteristics. On practical grounds the only difference between the three test statistics is purely computational,  $W(\mathbf{y})$  is based only on the unconstrained MLE  $\hat{\boldsymbol{\beta}}$ ,  $LM(\mathbf{y})$  is based on the constrained MLE  $\tilde{\boldsymbol{\beta}}$  and  $LR(\mathbf{y})$  on both. For a given example size  $T$ , however, the three test statistics can lead to different decisions as far as rejection of  $H_0$  is concerned. For example, if we were to apply the above procedures to the case where  $\mathbf{H}(\boldsymbol{\beta}) = \mathbf{R}\boldsymbol{\beta} - \mathbf{r}$  we could show that  $W(\mathbf{y}) \geq LR(\mathbf{y}) \geq LM(\mathbf{y})$  (see exercises 5 and 6).

## 20.5 Collinearity

As argued in Section 19.3 above, assumption [5] stating that

$$\text{rank}(\mathbf{X}) = k, \quad T > k, \quad (20.106)$$

is directly related to the statistical parameters of interest  $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$  ( $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\sigma}_{21}$ ,  $\sigma^2 = \sigma_{11} - \boldsymbol{\sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\sigma}_{21}$ ) via  $\boldsymbol{\Sigma}_{22}$ . This is because

$$\text{rank}(\mathbf{X}) = \text{rank}(\mathbf{X}'\mathbf{X}) \quad (20.107)$$

and (106) represents the sample equivalent to

$$\text{rank}(\Sigma_{22}) = k. \quad (20.108)$$

When (108) is invalid and  $\Sigma_{22}$  is singular  $\beta$  and  $\sigma^2$  cannot even be defined. Condition (108), however, cannot be verified directly and thus we need to rely on (106) which ensures that the estimators

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{T}\mathbf{y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} \quad (20.109)$$

of  $\beta$  and  $\sigma^2$  can be defined. In the case where  $\mathbf{X}'\mathbf{X}$  is singular  $\hat{\beta}$  and  $\hat{\sigma}^2$  cannot be derived.

The problem we face is that the singularity of  $(\mathbf{X}'\mathbf{X})$  does not necessarily imply the singularity of  $\Sigma_{22}$ . This is because the singularity of  $(\mathbf{X}'\mathbf{X})$  might be a problem with the observed data in hand and not a population problem. For example, in the case where  $T < k$   $\text{rank}(\mathbf{X}'\mathbf{X}) < k$ , irrespective of  $\Sigma_{22}$  because of the inadequacy of the observed data information. The only clear conclusion to be drawn from the failure of the condition (106) is that the *sample information in  $\mathbf{X}$  is inadequate for the estimation of the statistical parameters of interest  $\beta$  and  $\sigma^2$* . The source of the problem is rather more difficult to establish (sometimes impossible).

In econometric modelling the problem of collinearity is rather rare and when it occurs the reason is commonly because the modeller has ignored relevant measurement information (see Chapter 26) related to the data chosen. For example, in the case where an *accounting identity* holds among some of the  $x_{it}$ s.

It is important to note that the problem of collinearity is defined relative to a given parametrisation. The presence of collinearity among the columns of  $\mathbf{X}$ , however, does not preclude the possibility of estimating another parametrisation/restriction of the statistical GM. One such parametrisation is provided by a particular linear combination of the columns of  $\mathbf{X}$  based on the eigenvalues and eigenvectors of  $(\mathbf{X}'\mathbf{X})$ .

Let

$$\mathbf{P}(\mathbf{X}'\mathbf{X})\mathbf{P}' = \mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_m, 0, 0, \dots, 0) \quad (20.110)$$

and  $\mathbf{P}'\mathbf{P} = \mathbf{P}\mathbf{P}' = \mathbf{I}$ , where  $\mathbf{P}$  represents a  $k \times k$  orthogonal matrix whose columns are the eigenvectors of  $(\mathbf{X}'\mathbf{X})$  and  $\lambda_1, \lambda_2, \dots, \lambda_m$  its non-zero eigenvalues (see Householder (1974)). If we define the new observed data matrix to be  $\mathbf{X}^* = \mathbf{X}\mathbf{P}$  and  $\beta^* = \mathbf{P}'\beta$  the associated coefficient parameters we could reparametrise the statistical GM into

$$y_t = \beta^{*'} \mathbf{x}_t^* + u_t, \quad t = 1, 2, \dots, T. \quad (20.111)$$

The new data matrix  $\mathbf{X}^*$  can be viewed as referring to the values of the

artificial random variable  $\mathbf{X}_i^* = \mathbf{X}'_i \mathbf{p}_i$ ,  $i = 1, 2, \dots, k$ . The columns of  $\mathbf{X}^*$  defined by  $\mathbf{X}_i^* = \mathbf{X} \mathbf{p}_i$  are known as *principal components* of  $\mathbf{X}$  and in view of

$$(\mathbf{X}'\mathbf{X})\mathbf{p}_i = \mathbf{0} \quad \text{for } i = m+1, \dots, k \quad (20.112)$$

(see 110), where  $\mathbf{p}_i$ ,  $i = 1, 2, \dots, k$ , are the columns of  $\mathbf{P}$ , we can deduce that

$$\mathbf{X}^* \equiv (\mathbf{X}_1^* : \mathbf{X}_2^*), \quad \mathbf{X}_2^* = \mathbf{0}.$$

Decomposing  $\boldsymbol{\beta}^* = \mathbf{P}'\boldsymbol{\beta}$  conformably in the form  $\boldsymbol{\beta}^* = (\boldsymbol{\alpha}', \boldsymbol{\gamma}')$  we can rewrite (iii) for

$$t = 1, 2, \dots, T \quad \text{as } \mathbf{y} = \mathbf{X}^* \boldsymbol{\alpha} + \mathbf{u}, \quad (20.114)$$

where  $\text{rank}(\mathbf{X}_1) = m$ , with  $\boldsymbol{\alpha}$  and  $\boldsymbol{\tau}^2$  being the new parameters which are now *data specific*. These parameters can be estimated via

$$\hat{\boldsymbol{\alpha}} = (\mathbf{X}_1^* \mathbf{X}_1^*)^{-1} \mathbf{X}_1^* \mathbf{y} \quad \text{and} \quad \tilde{\tau}^2 = \frac{1}{T} \mathbf{y}' (\mathbf{I} - \mathbf{X}_1^* (\mathbf{X}_1^* \mathbf{X}_1^*)^{-1} \mathbf{X}_1^*) \mathbf{y}. \quad (20.115)$$

Moreover, in view of the relationship

$$\boldsymbol{\beta} = \mathbf{P}\boldsymbol{\beta}^* = \mathbf{P}_1\boldsymbol{\alpha} + \mathbf{P}_2\boldsymbol{\gamma}$$

any linear combination  $\mathbf{c}'\boldsymbol{\beta}$  of  $\boldsymbol{\beta}$  is estimable if  $\mathbf{c}'\mathbf{P}_2 = \mathbf{0}$  since

$$\mathbf{c}'\boldsymbol{\beta} = \mathbf{c}'\mathbf{P}_1\boldsymbol{\alpha} + \mathbf{c}'\mathbf{P}_2\boldsymbol{\gamma} = \mathbf{c}'\mathbf{P}_1\boldsymbol{\alpha}, \quad (20.117)$$

Using the principal components as the new columns of  $\mathbf{X}$ , however, does not constitute a solution to the original collinearity problem because the estimators (115) refer to a new parametrisation. This shows clearly how the collinearity problem is relative to a given parametrisation and not just a problem of data matrices. The same is also true for a potentially more serious problem, that of ‘near collinearity’, to be considered in the next section.

## 20.6 ‘Near’ collinearity

If we define collinearity as the situation where the rank of  $(\mathbf{X}'\mathbf{X})$  is less than  $k$ , ‘near’ collinearity refers to the situation where  $(\mathbf{X}'\mathbf{X})$  is ‘nearly’ singular or ill-conditioned as is known in numerical analysis. The effect of this near singularity is that the solution of the system

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \quad (20.118)$$

to derive  $\hat{\boldsymbol{\beta}}$  is highly sensitive to small changes in  $(\mathbf{X}'\mathbf{X})$  and  $\mathbf{X}'\mathbf{y}$ . That is, small changes in  $(\mathbf{X}'\mathbf{X})$  or  $\mathbf{X}'\mathbf{y}$  can lead to big changes in  $\hat{\boldsymbol{\beta}}$ . As with

collinearity, this is a problem of *insufficient data information* relative to a given *parametrisation*. In the present context the information in  $\mathbf{X}$  is not quite adequate for the estimation of the statistical parameters of interest  $\beta$  and  $\sigma^2$ . This might be due to insufficient sample information or to the choice of the variables involved ( $\Sigma_{22}$  is nearly singular). For example, in cases where there is not enough variability in some of the observed data series the sample information is inadequate for the task of determining  $\beta$  and  $\sigma^2$ . In such cases ‘near’ collinearity is a problem to be tackled. On the other hand, when the problem is inherent in the choice of  $\mathbf{X}_t$  (i.e. a population problem) then near collinearity is not really a problem to be tackled. In practice, however, there is no way to distinguish between these two sources of ‘near’ collinearity because we do not know  $\Sigma_{22}$ , unless we are in a Monte Carlo experimental situation (see Hendry (1984)). This suggests that any assessment of whether ‘near’ collinearity relative to a given parametrisation is a problem to be tackled will depend on assumptions about the ‘true’ values of the statistical parameters of interest.

Some of the commonly used criteria for detecting ‘near’ collinearity suggested by the textbook econometric literature are motivated solely by its effect on the ‘accuracy’ of  $\hat{\beta}$  as measured by

$$\text{Cov}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \quad (20.119)$$

Such criteria include:

(a) *Simple correlations*

These refer to the transformation of the  $(\mathbf{X}'\mathbf{X})$  matrix into a correlation matrix by standardising the regressors using

$$\tilde{x}_{it} = \frac{(x_{it} - \bar{x}_i)}{\sqrt{\sum_{t=1}^T (x_{it} - \bar{x}_i)^2}}, \quad i = 1, 2, \dots, k. \quad (20.120)$$

The standardisation is used in order to eliminate the units of measurement problem. High simple correlations are sometimes interpreted as indicators of near collinearity.

(b) *Auxiliary regressions*

Auxiliary regressions are estimated between each regressor and all the others, say  $x_{ht}$  and  $x_{1t}, x_{2t}, \dots, x_{h-1t}, x_{h+1t}, \dots, x_{kt}$ ,

$$x_{ht} = \alpha_0 + \alpha_1 x_{1t} + \cdots + \alpha_{h-1} x_{h-1t} + \alpha_{h+1} x_{h+1t} + \cdots + \alpha_k x_{kt} + \varepsilon_t, \quad (20.121)$$

and a high value of the multiple correlation coefficient from this regression,  $R_h^2$ , is used as a criterion for ‘near’ collinearity. This is motivated by the following form of the covariance of  $\hat{\beta}_h$ :

$$\text{Cov}(\hat{\beta}_h) = \sigma^2 (\mathbf{X}'\mathbf{X})_{hh}^{-1} = \sigma^2 [(1 - R_h^2) \mathbf{x}'_h \mathbf{x}_h]^{-1} \quad (20.122)$$

(see Theil (1971)). A high value for  $R_h^2$  (everything else assumed fixed) leads to a high value for  $\text{Cov}(\hat{\beta}_h)$ , viewed as the uncertainty related to  $\hat{\beta}_h$ . By the same token, however, a small value for  $\mathbf{x}'_h \mathbf{x}_h$ , interpreted as the variability of the  $h$ th regressor, will have the same effect. Note that  $R_h^2$  refers to

$$R_h^2 = \frac{\mathbf{x}'_h \mathbf{X}_{-h} (\mathbf{X}'_{-h} \mathbf{X}_{-h})^{-1} \mathbf{X}'_{-h} \mathbf{x}_h}{\mathbf{x}'_h \mathbf{x}_h}. \quad (20.123)$$

### (c) Condition numbers

Using the spectral decomposition of  $(\mathbf{X}'\mathbf{X})$  in (110) we can express (119) in the form

$$\text{Cov}(\hat{\beta}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 (\mathbf{P} \Lambda^{-1} \mathbf{P}') = \sigma^2 \sum_{i=1}^k \left( \frac{\mathbf{p}_i \mathbf{p}'_i}{\lambda_i} \right), \quad (20.124)$$

and thus

$$\text{Var}(\hat{\beta}_i) = \sigma^2 \sum_{j=1}^n \left( \frac{p_{ij}^2}{\lambda_j} \right), \quad i = 1, 2, \dots, k \quad (20.125)$$

(see Silvey (1969)). This suggests that the variance of each estimated coefficient  $\hat{\beta}_i$  depends on all the eigenvalues of  $(\mathbf{X}'\mathbf{X})$ . The presence of a relatively small eigenvalue  $\lambda_i$  will ‘dominate’ these variances. For this reason we look at the condition numbers

$$\kappa_i(\mathbf{X}'\mathbf{X}) = \frac{\lambda_{\max}}{\lambda_i}, \quad i = 1, 2, \dots, k, \quad (20.126)$$

for large values indicating ‘near’ collinearity, where  $\lambda_{\max}$  refers to the largest eigenvalue (see Belsley *et al.* (1980)). How large a condition number is large enough to indicate the presence of near collinearity is an open question in view of the fact that the eigenvalues  $\lambda_1, \dots, \lambda_k$  are not invariant to scale changes. For further discussion see Belsley (1984).

Several other criteria for detecting ‘near’ collinearity have been suggested in the econometric literature (see Judge *et al.* (1985) for a survey) but all these criteria, together with (a)–(c) above, suffer from two major weaknesses:

- (i) they do not seem to take account of the fact that ‘near’ collinearity should be assessed relative to a given parametrisation; and

- (ii) none of these criteria is invariant to linear transformations of the data (changes of origin and scale).

Ideally, we would like the matrix  $\mathbf{X}'\mathbf{X}$  to be diagonal, reflecting the orthogonality of the regressors, because in such a case the statistical GM takes a form where the effect of each regressor can be assessed separately given that  $\Sigma_{22} = \text{diag}(\sigma_{22}, \sigma_{33}, \dots, \sigma_{kk})$  and

$$\beta_i = \frac{\sigma_{1i}}{\sigma_{ii}}, \quad i = 2, 3, \dots, k, \quad (20.127)$$

$$\sigma_{1i} = \text{Cov}(X_{it}y_t), \quad \beta_1 = m_y - \sum_{i=2}^k \frac{\sigma_{1i}}{\sigma_{ii}} m_{x_i},$$

$$m_y = E(y_t), \quad m_{x_i} = E(X_{it}), \quad i = 2, 3, \dots, k. \quad (20.128)$$

In such a case

$$\hat{\beta}_i = (\mathbf{x}'_i \mathbf{x}_i)^{-1} \mathbf{x}'_i \mathbf{y} \quad \text{and} \quad \text{Var}(\hat{\beta}_i) = \sigma^2 (\mathbf{x}'_i \mathbf{x}_i)^{-1}, \quad (20.129)$$

with the estimator as well as its variance being effected only by the  $i$ th regressor. This represents a very robust estimated statistical model because changes in the behaviour of one regressor over time will affect no other coefficient estimator but its own. Moreover, by increasing the number of regressors the model remains unchanged. This situation can be reached by design in the case of the Gauss linear model (see Chapter 18). Although such an option is not directly available in the context of the linear regression model because we do not control the values of  $\mathbf{X}_t$ , we can achieve a similar effect via reparametrisation. Given that the statistical parameters of interest  $\boldsymbol{\theta}$  rarely coincide with the theoretical parameters of interest  $\boldsymbol{\xi}$  we can tackle the problem at the stage of reparametrising the estimated statistical GM in order to derive an empirical econometric model. It is important to note that statistical GM is postulated only as a crude approximation of the actual DGP purporting to summarise the sample information in a particular way, as suggested by the estimable model. Issues of efficient estimation utilising all a priori information do not arise at this stage. Hence, the presence of ‘near’ collinearity in the context of the statistical model should be viewed as providing us with important information relating to the adequacy of the sample information for the estimation of  $\boldsymbol{\theta}$ . This information will help us to construct a much more robust empirical econometric model by reparametrising the estimated statistical GM in ways which provide us with transformed regressors which are close to being orthogonal without, however, sacrificing its theoretical meaning. This is possible because there is no unique way to reparametrise a statistical GM into an empirical econometric model and one of the objectives in constructing the latter is to

ensure that the sample information is sufficient for the accurate determination of its parameters. It is important to emphasise at this stage that ‘good’ empirical econometric models are not, as some econometric and time-series textbooks would have us believe, given to us from outside by ‘sophisticated’ theories or by observed data regularities, but constructed by econometric modellers using their ingenuity and craftsmanship.

In view of the above discussion it seems preferable to put more emphasis in constructing robust empirical econometric models with ‘nearly’ orthogonal regressors without, however, sacrificing either their statistical or economic meaning. Hence, instead of worrying how to detect ‘near’ collinearity (which cannot be defined precisely anyway) by some battery of suspect criteria, it is preferable to turn the question on its head and consider the problem as one of constructing empirical econometric models with ‘nearly’ orthogonal regressors among its other desirable properties. To that end we need some criterion which assesses the contribution of each regressor separately and is invariant to linear transformations of the observed data. That is, a criterion whose value remains unchanged when

$$y_t \text{ is mapped into } y_t^* = a_1 y_t + c_1 \quad (20.130)$$

and

$$\mathbf{X}_t \text{ is mapped into } \mathbf{X}_t^* = A_2 \mathbf{X}_t + \mathbf{c}_2, \quad (20.131)$$

where  $a_1 \neq 0$  and  $A_2$  is a  $(k-1) \times (k-1)$  non-singular matrix.

We know that under the normality assumption

$$\mathbf{Z}_t \sim N(\mathbf{m}, \Sigma), \quad t \in \mathbb{T} \quad (20.132)$$

where

$$\mathbf{Z}_t \equiv \begin{pmatrix} y_t \\ \mathbf{X}_t \end{pmatrix}, \quad \mathbf{m} = \begin{pmatrix} m_1 \\ \mathbf{m}_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \Sigma_{22} \end{pmatrix}, \quad (20.133)$$

the statistic

$$(\bar{\mathbf{Z}}, \mathbf{S}), \quad \bar{\mathbf{Z}} = \frac{1}{T} \sum_{t=1}^T \mathbf{Z}_t \quad \text{and} \quad \mathbf{S} = \sum_{t=1}^T (\mathbf{Z}_t - \bar{\mathbf{Z}})(\mathbf{Z}_t - \bar{\mathbf{Z}})' \quad (20.134)$$

is a sufficient statistic (see Chapter 15). Under the data transformation (130) and (131) the sufficient statistic is transformed as

$$\bar{\mathbf{Z}} \rightarrow \mathbf{A}\bar{\mathbf{Z}} + \mathbf{c} \quad (20.135)$$

and

$$\mathbf{S} \rightarrow \mathbf{A}\mathbf{S}\mathbf{A}', \quad (20.136)$$

where

$$\mathbf{A} = \begin{pmatrix} a_1 & \mathbf{0} \\ \mathbf{0} & A_2 \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} c_1 \\ \mathbf{c}_2 \end{pmatrix}$$

are  $k \times k$  and  $k \times 1$  matrices. The corresponding transformations on the parameters are:

$$\begin{aligned}\mathbf{m} &\rightarrow \mathbf{A}\mathbf{m} + \mathbf{c}, \\ \boldsymbol{\Sigma} &\rightarrow \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'.\end{aligned}\tag{20.137}$$

What we are seeking is a criterion which remains unchanged under this group of transformations. The obvious candidates as likely measures for assessing the separate contributions of the regressors involved are the *multiple* and *partial correlation coefficients* (see Chapter 15). The sample partial correlation coefficient of  $y_t$  with one of the regressors, say  $X_{1t}$ , given the rest  $\mathbf{X}_{3t}$ , is given in equation (15.48). This represents a measure of the correlation of  $y_t$  and  $X_{1t}$  when the effect of all the other variables have been ‘partialled out’. On the other hand, if we want the incremental contribution of  $X_{1t}$  to the regression of  $y_t$  on  $\mathbf{X}_t$  we need to take the difference between the sample multiple correlation coefficient, between  $y_t$  and  $\mathbf{X}_t$ , and that between  $y_t$  and  $\mathbf{X}_{-1t}$  ( $\mathbf{X}_t$  with  $X_{1t}$  excluded), denoted by  $\hat{R}^2$  and  $\hat{R}_{-1}^2$ , i.e. use

$$(\hat{R}^2 - \hat{R}_{-1}^2)\tag{20.138}$$

(see equation (15.39) for  $\hat{R}^2$ ). It is not very surprising that both of these measures are directly related via

$$(\hat{R}^2 - \hat{R}_{-1}^2) = \hat{\rho}_{12.3}^2(1 - \hat{R}_{-1}^2)\tag{20.139}$$

(see Theil (1971)), where  $\hat{\rho}_{12.3}$  denotes the sample partial correlation coefficient of  $y_t$  and  $X_{1t}$  given the rest  $\mathbf{X}_{3t}$ .

Let

$$\hat{R}^2 = g(\bar{\mathbf{Z}}, \mathbf{S}) \quad \text{and} \quad \hat{\rho}_{12.3}^2 = g_1(\bar{\mathbf{Z}}, \mathbf{S}).$$

We can verify directly that

$$g(\bar{\mathbf{Z}}, \mathbf{S}) = g(\mathbf{A}\bar{\mathbf{Z}} + \mathbf{c}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')\tag{20.140}$$

and

$$g_1(\bar{\mathbf{Z}}, \mathbf{S}) = g_1(\mathbf{A}\bar{\mathbf{Z}} + \mathbf{c}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}').\tag{20.141}$$

That is, the sample multiple and partial correlation coefficients are *invariant* to the group of *linear transformations* on the data. Indeed it can be shown that  $\hat{R}^2$  is a maximal invariant to this group of transformations (see Muirhead (1982)). That is, any invariant statistic is a function of  $\hat{R}^2$ .

The multiple and partial correlation coefficients together with the simple correlation coefficients can be used as ‘guides’ to the construction of empirical models with nearly orthogonal regressors. The multiple correlation in particular can be used to get an overall picture of the relative contribution of the various regressors (in both the statistical GM and the

empirical econometric model) using the incremental contributions

$$(\hat{R}^2 - \hat{R}_{-i}^2), \quad i = 1, 2, \dots, k-1, \quad (20.142)$$

in conjunction with

$$\hat{R}^2 - \sum_{i=1}^{k-1} (\hat{R}^2 - \hat{R}_i^2), \quad (20.143)$$

which Theil (1971) called the *multicollinearity effect*. In the present context such an interpretation should be viewed as coincidental to the main aim of constructing robust empirical econometric models. It is important to remember that the computation of the multiple correlation coefficient differs from one computer package to another and it is rarely the one used above.

To conclude, note that the various so-called ‘solutions’ to the problem of near collinearity such as dropping or adding regressors or supplementing the model with a priori information are simply ways to introduce *alternative reparametrisations*, not solutions to the original problem.

### ***Important concepts***

Stochastic linear regression model, statistical versus theoretical parameters of interest, omitted variables bias, reparametrisation, constrained and unconstrained MLE’s, a priori linear and non-linear restrictions, restricted and unrestricted residual sum of squares, collinearity, ‘near’ collinearity, orthogonal regressors, incremental contributions, partial correlation coefficient, invariant to linear transformations.

### ***Questions***

1. Compare and contrast

- (i)  $E(y_t | \mathbf{X}_t = \mathbf{x}_t)$ ,  $E(y_t / \sigma(\mathbf{X}_t))$ ;
- (ii)  $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ ,  $\hat{\beta}^* = (\mathbf{X}''\mathbf{X}')^{-1}\mathbf{X}''\mathbf{y}$ ;
- (iii)  $\hat{\sigma}^2 = \frac{1}{T} \hat{\mathbf{u}}'\hat{\mathbf{u}}$ ,  $\hat{\sigma}^{*2} = \frac{1}{T} \hat{\mathbf{u}}'^*\hat{\mathbf{u}}^*$ .

2. Compare and contrast the statistical GM’s, the probability and sampling models for the linear regression and stochastic linear regression statistical models.

3. ‘Let the “true” model be

$$y_t = \boldsymbol{\beta}'\mathbf{x}_t + \gamma'\mathbf{w}_t + \varepsilon_t$$

and the one used be  $y_t = \beta' x_t + u_t$ . It can be shown that for  $\hat{\beta} = (X'X)^{-1}X'y$

- (i)  $E(\hat{\beta}) \neq \beta$ , i.e.  $\hat{\beta}$  suffers from omitted variables bias; and
- (ii)  $E(u_t) = \gamma' w_t$ .

Discuss.

4. Explain informally how you would go about constructing an exogeneity test. Discuss the difficulties associated with such a test.
5. Compare the constrained and unconstrained MLE's of  $\beta$  and  $\sigma^2$ :

$$\tilde{\beta} = \hat{\beta} - (X'X)^{-1}R' [R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r),$$

$$\hat{\beta} = (X'X)^{-1}X'y,$$

$$\tilde{\sigma}^2 = \frac{1}{T} (y - X\tilde{\beta})'(y - X\tilde{\beta}),$$

$$\hat{\sigma}^2 = \frac{1}{T} (y - X\hat{\beta})'(y - X\hat{\beta}).$$

6. Explain how you would go about constructing a test for

$$H_0: R\beta = r \text{ against } H_1: R\beta \neq r$$

based on the intuitive argument that when  $H_0$  is true  $\tilde{\mu}$  is close to zero. ‘Why don't we use the distance  $\|R\tilde{\beta} - r\|$ ?’ Compare the resulting test with the  $F$ -test.

7. Explain intuitively the derivation of the Wald test for

$$H_0: H(\beta) = 0 \text{ against } H_1: H(\beta) \neq 0.$$

‘Why don't we use  $\|H(\tilde{\beta})\|$  instead of  $\|H(\hat{\beta})\|$  as the basis of the argument for the derivation?’

8. What do we mean by the statement that the Wald, Lagrange multiplier and likelihood ratio test procedures give rise to three *asymptotically equivalent tests*?
9. When do we prefer an asymptotic to a finite sample test?
10. Explain the role of the assumption that  $\text{rank}(X) = k$  in the context of the linear regression model.
11. Discuss the concepts of ‘collinearity’ and ‘near-collinearity’ and their implications as far as the MLE's  $\hat{\beta}$  and  $\hat{\sigma}^2$  are concerned.
12. How do we interpret the fact that  $(X'X)$  is ‘ill-conditioned’ in the context of the linear regression model? How do we tackle the problem?

***Exercises***

1. Using the first-order conditions (49)–(51) of Section 20.4 derive the information matrix (57).
2. Using the partitioned inverse formula,

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{F}\mathbf{E}^{-1}\mathbf{F}' & -\mathbf{F}\mathbf{E}^{-1} \\ -\mathbf{E}^{-1}\mathbf{F}' & \mathbf{E}^{-1} \end{pmatrix}$$

$$\mathbf{E} = \mathbf{D} - \mathbf{B}'\mathbf{A}^{-1}\mathbf{B}, \quad \mathbf{F} = \mathbf{A}^{-1}\mathbf{B},$$

derive  $[\mathbf{I}_T(\boldsymbol{\beta}, \boldsymbol{\mu}, \sigma^2)]^{-1}$  and compare its various elements with  $\mathbf{C}_{11}$ ,  $\mathbf{C}_{12}$  and  $\mathbf{C}_{22}$  in (56).

3. Verify the distribution of

$$\begin{pmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{\mu}} \end{pmatrix}$$

in (56).

4. Verify the equality  $(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) = \tilde{\mathbf{u}}'\tilde{\mathbf{u}} - \hat{\mathbf{u}}'\hat{\mathbf{u}}$ .
5. For the null hypothesis  $H_0: \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$  against  $H_1: \mathbf{R}\boldsymbol{\beta} \neq \mathbf{r}$  use the Wald, Lagrange multiplier and likelihood ratio test procedures to derive the following test statistics:

$$W(\mathbf{y}) = \left( \frac{T}{T-k} \right) m\tau(\mathbf{y}), \quad LM(\mathbf{y}) = \frac{Tm\tau(\mathbf{y})}{(T-k) + m\tau(\mathbf{y})},$$

$$LR(\mathbf{y}) = T \log \left( 1 + \frac{m\tau(\mathbf{y})}{T-k} \right),$$

respectively, where  $\tau(\mathbf{y})$  is the test statistic of the  $F$ -test.

6. Using  $W(\mathbf{y})$ ,  $LM(\mathbf{y})$  and  $LR(\mathbf{y})$  from exercise 5 show that

$$W(\mathbf{y}) \geq LR(\mathbf{y}) \geq LM(\mathbf{y}).$$

(Note that  $\log(1+z) \geq z/(1+z)$ ,  $z \geq \log(1+z)$ ,  $z \geq 0$ ) (see Evans and Savin (1982).)

**Additional references**

Aitchison and Silvey (1958); Judge *et al.* (1982); Leamer (1983).

## CHAPTER 21

---

### The linear regression model III – departures from the assumptions underlying the probability model

---

The purpose of this chapter is to consider various forms of departures from the assumptions of the probability model:

- [6] (i)  $D(y_t/X_t; \theta)$  is normal,
  - (ii)  $E(y_t/X_t = x_t) = \beta' x_t$ , linear in  $x_t$ ,
  - (iii)  $\text{Var}(y_t/X_t = x_t) = \sigma^2$ , homoskedastic,
- [7]  $\theta \equiv (\beta_1 \sigma^2)$  are time-invariant.

In each of the Sections 2–5 the above assumptions will be relaxed one at a time, retaining the others, and the following interrelated questions will be discussed:

- (a) what are the implications of the departures considered?
- (b) how do we detect such departures?, and
- (c) how do we proceed if departures are detected?

It is important to note at the outset that the following discussion which considers individual assumptions being relaxed separately limits the scope of misspecification analysis because it is rather rare to encounter such conditions in practice. More often than not various assumptions are invalid simultaneously. This is considered in more detail in Section 1. Section 6 discusses the problem of structural change which constitutes a particularly important form of departure from [7].

#### 21.1 Misspecification testing and auxiliary regressions

Misspecification testing refers to the testing of the assumptions underlying a statistical model. In its context the null hypothesis is uniquely defined as the assumption(s) in question being valid. The alternative takes a particular form of departure from the null which is invariably non-unique. This is

because departures from a given assumption can take numerous forms with the specified alternative being only one such form. Moreover, most misspecification tests are based on the questionable presupposition that the other assumptions of the model are valid. This is because joint misspecification testing is considerably more involved. For these reasons the choice in a misspecification test is between rejecting and not rejecting the null; accepting the alternative should be excluded at this stage.

An important implication for the question on how to proceed if the null is rejected is that before any action is taken the results of the other misspecification tests should also be considered. It is often the case that a particular form of departure from one assumption might also affect other assumptions. For example when the assumption of sample independence [8] is invalid the other misspecification tests are influenced (see Chapter 22).

In general the way to proceed when any of the assumptions [6]–[8] are invalid is first to narrow down the source of the departures by relating them back to the NIID assumption of  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  and then respecify the model taking into account the departure from NIID. The respecification of the model involves a reconsideration of the reduction from  $D(\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_T; \psi)$  to  $D(y_t/\mathbf{X}_t; \theta)$  so as to account for the departures from the assumptions involved. As argued in Chapters 19–20 this reduction coming in the form of:

$$D(\mathbf{Z}_1, \dots, \mathbf{Z}_T; \psi) = \prod_{t=1}^T D(\mathbf{Z}_t; \psi) \quad (21.1)$$

$$= \prod_{t=1}^T D(y_t/\mathbf{X}_t, \psi_1) \cdot D(\mathbf{X}_t; \psi_2) \quad (21.2)$$

involves the independence and the identically distributed assumptions in (1). The normality assumption plays an important role in defining the parametrisation of interest  $\theta \equiv (\beta, \sigma^2)$  as well as the weak exogeneity condition. Once the source of the detected departure is related to one or more of the NIID assumptions the respecification takes the form of an alternative form of reduction. This is illustrated most vividly in Chapter 22 where assumption [8] is discussed. It turns out that when [8] is invalid not only the results in Chapter 19 are invalid but the other misspecification tests are ‘largely’ inappropriate as well. For this reason it is advisable in practice to test assumption [8] first and then proceed with the other assumptions if [8] is not rejected. The sequence of misspecification tests considered in what follows is chosen only for expositional purposes.

With the above discussion in mind let us consider the question of general procedures for the derivation of misspecification tests. In cases where the alternative in a misspecification test is given a specific parametric form the various procedures encountered in specification testing ( $F$ -type tests, Wald,

Lagrange multiplier and likelihood ratio) can be easily adapted to apply in the present context. In addition to these procedures several specific misspecification test procedures have been proposed in the literature (see White (1982), Bierens (1982), *inter alia*). Of particular interest in the present book are the procedures based on the ‘omitted variables’ argument which lead to auxiliary regressions (see Ramsey (1969), (1974), Pagan and Hall (1983), Pagan (1984), *inter alia*). This particular procedure is given a prominent role in what follows because it is easy to implement in practice and it provides a common-sense interpretation of most other misspecification tests.

The ‘omitted variables’ argument was criticised in Section 20.2 because it was based on the comparison of two ‘non-comparable’ statistical GM’s. This was because the information sets underlying the latter were different. It was argued, however, that the argument could be reformulated by postulating the same sample information sets. In particular if both parametrisations can be derived from  $D(\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_T; \psi)$  by using alternative reduction arguments then the two statistical GM’s can be made comparable.

Let  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  be a vector stochastic process defined on the probability space  $(S, \mathcal{F}, P(\cdot))$  which includes the stochastic variables of interest. In Chapter 17 it was argued that for a given  $\mathcal{L}_t \subseteq \mathcal{F}$

$$y_t = E(y_t | \mathcal{L}_t) + u_t, \quad t \in \mathbb{T} \quad (21.3)$$

defines a general statistical GM with

$$\mu_t = E(y_t | \mathcal{L}_t), \quad u_t = y_t - E(y_t | \mathcal{L}_t) \quad (21.4)$$

satisfying some desirable properties by construction including the orthogonality condition:

$$E(\mu_t u_t) = 0, \quad t \in \mathbb{T}. \quad (21.5)$$

It is important to note, however, that (3)–(4) as defined above are just ‘empty boxes’. These are filled when  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is given a specific probabilistic structure such as NIID. In the latter case (3)–(4) take the specific forms:

$$y_t = \beta' \mathbf{x}_t + u_t, \quad t \in \mathbb{T} \quad (21.6)$$

$$\mu_t^* = \beta' \mathbf{x}_t \quad \text{and} \quad u_t^* = y_t - \beta' \mathbf{x}_t, \quad (21.7)$$

with the conditioning information set being

$$\mathcal{L}_t = \{\mathbf{X}_t = \mathbf{x}_t\}. \quad (21.8)$$

When any of the assumptions in NIID are invalid, however, the various properties of  $\mu_t$  and  $u_t$  no longer hold for  $\mu_t^*$  and  $u_t^*$ . In particular the

orthogonality condition (5) is invalid. The non-orthogonality

$$E(\mu_t^* u_t^*) \neq 0, \quad t \in \mathbb{T} \quad (21.9)$$

can be used to derive various misspecification tests. If we specify the alternative in a parametric form which includes the null as a special case (9) could be used to derive misspecification tests based on certain auxiliary regressions.

In order to illustrate this procedure let us consider two important parametric forms which can provide the basis of several misspecification tests:

$$(a) \quad g^*(\mathbf{x}_t) = \sum_{i=1}^m \gamma_i (\mu_t^*)^i \quad (21.0)$$

$$(b) \quad g(\mathbf{x}_t) = a + \sum_{i=1}^k b_i x_{it} + \sum_{i=1}^k \sum_{j \geq i}^k c_{ij} x_{it} x_{jt} \\ + \sum_{i=1}^k \sum_{j \geq i}^k \sum_{l \geq j}^k d_{ijl} x_{it} x_{jt} x_{lt} + \dots \quad (21.11)$$

The polynomial  $g^*(\mathbf{x}_t)$  is related to RESET type tests (see Ramsey (1969)) and  $g(\mathbf{x}_t)$  is known as the Kolmogorov–Gabor polynomial (see Ivakhnenko (1984)). Both of these polynomials can be used to specify a general parametric form for the alternative systematic component:

$$\mu_t = \boldsymbol{\beta}'_0 \mathbf{x}_t + \gamma'_0 \mathbf{z}_t^* \quad (21.12)$$

where  $\mathbf{z}_t^*$  represents known functions of the variables  $\mathbf{Z}_{t-1}, \dots, \mathbf{Z}_1, \mathbf{X}_t$ . This gives rise to the alternative statistical GM

$$y_t = \boldsymbol{\beta}'_0 \mathbf{x}_t + \gamma'_0 \mathbf{z}_t^* + \varepsilon_t, \quad t \in \mathbb{T} \quad (21.13)$$

which includes (6) as a special case under

$$H_0: \gamma_0 = \mathbf{0}, \quad \text{with } H_1: \gamma_0 \neq \mathbf{0}. \quad (21.14)$$

A direct comparison between (13) and (6) gives rise to the auxiliary regression

$$u_t = (\boldsymbol{\beta}_0 - \boldsymbol{\beta})' \mathbf{x}_t + \gamma'_0 \mathbf{z}_t^* + \varepsilon_t, \quad (21.15)$$

whose operational form

$$\hat{u}_t = (\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})' \mathbf{x}_t + \gamma'_0 \mathbf{z}_t^* + \varepsilon_t \quad (21.16)$$

can be used to test (14) directly. The most obvious test is the  $F$ -type test discussed in Sections 19.5 and 20.3. The  $F$ -test will take the general form

$$FT(y) = \frac{RRSS - URSS}{URSS} \left( \frac{T-k^*}{m} \right) \quad (21.17)$$

where  $RRSS$  and  $URSS$  refer to the residuals sum of squares from (6) and (16) (or (13)), respectively;  $k^*$  being the number of parameters in (13) and  $m$  the number of restrictions.

This procedure could be easily extended to the higher central moments of  $y_t/\mathbf{X}_t$ ,

$$E(u_t^r/\mathbf{X}_t = \mathbf{x}_t), \quad r \geq 2. \quad (21.18)$$

For further discussion see Spanos (1985b).

## 21.2 Normality

As argued above, the assumptions underlying the probability model are all interrelated and they stem from the fact that  $D(y_t, \mathbf{X}_t; \psi)$  is assumed to be multivariate normal. When  $D(y_t, \mathbf{X}_t; \psi)$  is assumed to be some other multivariate distribution the regression function takes a more general form (not necessarily linear),

$$E(y_t/\mathbf{X}_t = \mathbf{x}_t) = h(\psi, \mathbf{x}_t), \quad (21.19)$$

and the skedasticity function is not necessarily free of  $\mathbf{x}_t$ ,

$$\text{Var}(y_t/\mathbf{X}_t = \mathbf{x}_t) = g(\psi, \mathbf{x}_t). \quad (21.20)$$

Several examples of regression and skedasticity functions in the bivariate case were considered in Chapter 7. In this section, however, we are going to consider relaxing the assumption of normality only, keeping linearity and homoskedasticity. In particular we will consider the *consequences* of assuming

$$(y_t/\mathbf{X}_t = \mathbf{x}_t) \sim D(\beta' \mathbf{x}_t, \sigma^2), \quad (21.21)$$

where  $D(\cdot)$  is an unknown distribution, and discuss the problem of *testing* whether  $D(\cdot)$  is in fact *normal or not*.

### (1) Consequences of non-normality

Let us consider the effect of the non-normality assumption in (21) on the specification, estimation and testing in the context of the linear regression model discussed in Chapter 19.

As far as specification (see Section 19.2) is concerned only marginal changes are needed. After removing assumption [6](i) the other assumptions can be reinterpreted in terms of  $D(\beta' \mathbf{x}_t, \sigma^2)$ . This suggests that relaxing normality but retaining linearity and homoskedasticity might not constitute a major break from the linear regression framework.

The first casualty of (21) as far as estimation (see Section 19.4) is concerned is the method of maximum likelihood itself which cannot be used unless the form of  $D(\cdot)$  is known. We could, however, use the least-squares method of estimation briefly discussed in Section 13.1, where the form of the underlying distribution is ‘apparently’ not needed.

Least-squares is an alternative method of estimation which is historically much older than the maximum likelihood or the method of moments. The least-squares method estimates the unknown parameters  $\theta$  by minimising the squares of the distance between the observable random variables  $y_t$ ,  $t \in \mathbb{T}$ , and  $h_t(\theta)$  (a function of  $\theta$  purporting to approximate the mechanism giving rise to the observed values  $y_t$ ), weighted by a precision factor  $1/\kappa_t$  which is assumed known, i.e.

$$\min_{\theta \in \Theta} \sum_t \left( \frac{y_t - h_t(\theta)}{\kappa_t} \right)^2. \quad (21.22)$$

It is interesting to note that this method was first suggested by Gauss in 1794 as an alternative to maximising what we, nowadays, call the log-likelihood function under the normality assumption (see Section 13.1 for more details). In an attempt to motivate the least-squares method he argued that:

the most probable value of the desired parameters will be that in which the sum of the squares of differences between the actually observed and computed values multiplied by numbers that measure the degree of precision, is a minimum . . .

This clearly shows a direct relationship between the normality assumption and the least-squares method of estimation. It can be argued, however, that the least-squares method can be applied to estimation problems without assuming normality. In relation to such an argument Pearson (1920) warned that:

we can only assert that the least-squares methods are theoretically accurate on the assumption that our observations . . . obey the normal law. . . . Hence in disregarding normal distributions and claiming great generality . . . by merely using the principle of least-squares . . . the apparent generalisation has been gained merely at the expense of theoretical validity . . .

Despite this forceful argument let us consider the estimation of the linear regression model without assuming normality, but retaining linearity and homoskedasticity as in (21).

The least-squares method suggests minimising

$$l(\beta) = \sum_{t=1}^T \frac{(y_t - \beta' \mathbf{x}_t)^2}{\sigma^2}, \quad (21.23)$$

or, equivalently:

$$l(\beta) = \sum_{t=1}^T (y_t - \beta' x_t)^2 = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta), \quad (21.24)$$

$$\frac{\partial l}{\partial \beta} = -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta) = 0. \quad (21.25)$$

Solving the system of normal equations (25) (assuming that  $\text{rank}(\mathbf{X}) = k$ ) we get the ordinary least-squares (OLS) estimator of  $\beta$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \quad (21.26)$$

The OLS estimator of  $\sigma^2$  is

$$\hat{s}^2 = \frac{1}{T-k} l(\mathbf{b}) = \frac{1}{T-k} (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}). \quad (21.27)$$

Let us consider the properties of the OLS estimators  $\mathbf{b}$  and  $\hat{s}^2$  in view of the fact that the form of  $D(\beta' x_t, \sigma^2)$  is not known.

### *Finite sample properties of $\mathbf{b}$ and $\hat{s}^2$*

Although  $\mathbf{b}$  is identical to  $\hat{\beta}$  (the MLE of  $\beta$ ) the similarity does not extend to the properties unless  $D(y_t/\mathbf{X}_t; \theta)$  is normal.

(a) Since  $\mathbf{b} = \mathbf{L}\mathbf{y}$ , the OLS estimator is *linear in  $\mathbf{y}$* .

Using the properties of the expectation operator  $E(\cdot)$  we can deduce:

(b)  $E(\mathbf{b}) = E(\mathbf{b} + \mathbf{L}\mathbf{u}) = \beta + \mathbf{L}E(\mathbf{u}) = \beta$ , i.e.  $\mathbf{b}$  is an unbiased estimator of  $\beta$ .

(c)  $E(\mathbf{b} - \beta)(\mathbf{b} - \beta)' = E(\mathbf{L}\mathbf{u}\mathbf{u}'\mathbf{L}') = \sigma^2 \mathbf{L}\mathbf{L}' = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$ .

Given that we have the mean and variance of  $\mathbf{b}$  but not its distribution, what other properties can we deduce?

Clearly, we cannot say anything about *sufficiency* or *full efficiency* without knowing  $D(y_t/\mathbf{X}_t; \theta)$  but hopefully we could discuss *relative efficiency* within the class of estimators satisfying (a) and (b). The Gauss–Markov theorem provides us with such a result.

### *Gauss–Markov theorem*

Under the assumption (21),  $\mathbf{b}$ , the OLS estimator of  $\beta$ , has *minimum variance* among the class of linear and unbiased estimators (for a proof see Judge *et al.* (1982)).

As far as  $\hat{s}^2$  is concerned, we can show that

(d)  $E(\hat{s}^2) = \sigma^2$ , i.e.  $\hat{s}^2$  is an unbiased estimator of  $\sigma^2$ ,

using only the properties of the expectation operator relative to  $D(\beta' x_t, \sigma^2)$ .

In order to test any hypotheses or set up confidence intervals for

$\theta = (\beta, \sigma^2)$  we need the distribution of the OLS estimators  $\mathbf{b}$  and  $\hat{s}^2$ . Thus, unless we specify the form of  $D(\beta' \mathbf{x}_t, \sigma^2)$ , no test or/and confidence interval statistics can be derived. The question which naturally arises is to what extent ‘asymptotic theory’ can at least provide us with large sample results.

### Asymptotic distribution of $\mathbf{b}$ and $\hat{s}^2$

*Lemma 21.1*

*Under assumption (21),*

$$\sqrt{T}(\mathbf{b} - \beta) \underset{z}{\sim} N(\mathbf{0}, \sigma^2 \mathbf{Q}_x^{-1}) \quad (21.28)$$

*if*

$$\lim_{T \rightarrow \infty} \left( \frac{\mathbf{X}' \mathbf{X}}{T} \right) = \mathbf{Q}_x \quad (21.29)$$

*is finite and non-singular.*

*Lemma 21.2*

*Under (21) we can deduce that*

$$\sqrt{T}(\hat{s}^2 - \sigma^2) \underset{z}{\sim} N\left(0, \left(\frac{\mu_4}{\sigma_4} - 1\right)\sigma^4\right), \quad (21.30)$$

*where  $\mu_4$  refers to the fourth central moment of  $D(y_t/\mathbf{X}_t; \theta)$  assumed to be finite (see Schmidt (1976)).*

*Note that in the case where  $D(y_t/\mathbf{X}_t; \theta)$  is normal*

$$\frac{\mu_4}{\sigma^4} = 3 \Rightarrow \sqrt{T}(\hat{s}^2 - \sigma^2) \underset{z}{\sim} N(0, 2\sigma^4). \quad (21.31)$$

*Lemma 21.3*

*Under (21)*

$$\mathbf{b} \xrightarrow{P} \beta \quad (21.32)$$

$$(\text{if } \lim_{T \rightarrow \infty} (\mathbf{X}' \mathbf{X}) = \mathbf{0}) \quad (21.33)$$

*and*

$$\hat{s}^2 \xrightarrow{P} \sigma^2. \quad (21.34)$$

From the above lemmas we can see that although the asymptotic distribution of  $\mathbf{b}$  coincides with the asymptotic distribution of the MLE this is not the case with  $\hat{s}^2$ . The asymptotic distribution of  $\mathbf{b}$  does not depend on

$D(y_t/\mathbf{X}_t; \boldsymbol{\theta})$  but that of  $\hat{s}^2$  does via  $\mu_4$ . The question which naturally arises is to what extent the various results related to tests about  $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}, \sigma^2)$  (see Section 19.5) are at least asymptotically justifiable. Let us consider the  $F$ -test for  $H_0: \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$  against  $H_1: \mathbf{R}\boldsymbol{\beta} \neq \mathbf{r}$ . From lemma 21.1 we can deduce that under  $H_0: \sqrt{x} T(\mathbf{R}\mathbf{b} - \mathbf{r}) \sim N(0, \sigma^2(\mathbf{R}\mathbf{Q}_x^{-1}\mathbf{R}')^{-1})$ , which implies that

$$(\mathbf{R}\mathbf{b} - \mathbf{r})' \frac{[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}}{\sigma^2} (\mathbf{R}\mathbf{b} - \mathbf{r}) \stackrel{H_0}{\underset{x}{\sim}} \chi^2(m). \quad (21.35)$$

Using this result in conjunction with lemma 21.3 we can deduce that

$$\tau_T(\mathbf{y}) = (\mathbf{R}\mathbf{b} - \mathbf{r})' \frac{[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}}{m\hat{s}^2} (\mathbf{R}\mathbf{b} - \mathbf{r}) \stackrel{H_0}{\underset{x}{\sim}} \frac{1}{m} \chi^2(m) \quad (21.36)$$

under  $H_0$ , and thus the  $F$ -test is robust with respect to the non-normality assumption (21) above. Although the asymptotic distribution of  $\tau_T(\mathbf{y})$  is chi-square, in practice the  $F$ -distribution provides a better approximation for a small  $T$  (see Section 19.5)). This is particularly true when  $D(\boldsymbol{\beta}'\mathbf{x}_t, \sigma^2)$  has heavy tails. The significance  $t$ -test being a special case of the  $F$ -test,

$$\tau(\mathbf{y}) = \frac{b_i}{\hat{s}\sqrt{[(\mathbf{X}'\mathbf{X})]_{ii}^{-1}}} \stackrel{H_0}{\underset{x}{\sim}} N(0, 1) \quad \text{under } H_0: \beta_i = 0 \quad (21.37)$$

is also asymptotically justifiable and robust relative to the non-normality assumption (21) above.

Because of lemma 21.2 intuition suggests that the testing results in relation to  $\sigma^2$  will not be robust relative to the non-normality assumption. Given that the asymptotic distribution of  $\hat{s}^2$  depends on  $\mu_4$  or  $\alpha_4 = \mu_4/\sigma^4$  the kurtosis coefficient, any departures from normality (where  $\alpha_4 = 3$ ) will seriously affect the results based on the normality assumption. In particular the size  $\alpha$  and power of these tests can be very different from the ones based on the postulated value of  $\alpha$ . This can seriously affect all tests which depend on the distribution of  $\hat{s}^2$  such as some heteroskedasticity and structural change tests (see Sections 21.4–21.6 below). In order to get non-normality robust tests in such cases we need to modify them to take account of  $\hat{\mu}_4$ .

## (2) Testing for departures from normality

Tests for normality can be divided into parametric and non-parametric tests depending on whether the alternative is given a parametric form or not.

452      **Departures from assumptions – probability model**

(a)      *Non-parametric tests*

*The Kolmogorov–Smirnov test*

Based on the assumption that  $\{u_t/X_t, t \in \mathbb{T}\}$  is an IID process we can use the results of Appendix 11.1 to construct test with rejection region

$$C_1 = \{\mathbf{y}: \sqrt{T} \hat{D}_T^* > c_\alpha\} \quad (21.38)$$

where  $\hat{D}_T^*$  refers to the Kolmogorov–Smirnov test statistic in terms of the residuals. Typical values of  $c_\alpha$  are:

$$\begin{array}{cccc} \alpha & .01 & .05 & .01 \\ c_\alpha & 1.23 & 1.36 & 1.67 \end{array} \quad (21.39)$$

For a most illuminating discussion of this and similar tests see Durbin (1973).

*The Shapiro–Wilk test*

This test is based on the ratio of two different estimators of the variance  $\sigma^2$ .

$$W = \left[ \sum_{t=1}^n a_{tT} (\hat{u}_{(T-t+1)} - \hat{u}_{(1)}) \right]^2 / \left[ \sum_{t=1}^T \hat{u}_t^2 \right] \quad (21.40)$$

where  $\hat{u}_{(1)} \leq \hat{u}_{(2)} \leq \dots \leq \hat{u}_{(T)}$  are the ordered residuals,

$$n = \frac{T}{2} \quad \text{if } T \text{ is even} \quad \text{or} \quad n = \frac{T-1}{2} \quad \text{if } T \text{ is odd},$$

and  $a_{tT}$  is a weight coefficient tabulated by Shapiro and Wilk (1965) for sample sizes  $2 < T \leq 50$ . The rejection region takes the form:

$$C_1 = \{\mathbf{y}: W < c_\alpha\} \quad (21.41)$$

where  $c_\alpha$  are tabulated in the above paper.

(b)      *Parametric tests*

*The skewness–kurtosis test*

The most widely used parametric test for normality is the skewness–kurtosis. The parametric alternative in this test comes in the form of the Pearson family of densities.

The Pearson family of distributions is based on the differential equation

$$\frac{d \ln f(z)}{dz} = \frac{(z-a)}{c_0 + c_1 z + c_2 z^2}, \quad (21.42)$$

where solution for different values of  $(a, c_0, c_1, c_2)$  generates a large number of interesting distributions such as the gamma, beta and Student's  $t$ . It can be shown that knowledge of  $\sigma^2$ ,  $\alpha_3$  and  $\alpha_4$  can be used to determine the distribution of  $Z$  within the Pearson family. In particular:

$$a = c_1 = (\alpha_4 + 3)(\alpha_3)^{\frac{1}{2}}\sigma, \quad (21.43)$$

$$c_0 = (4\alpha_4 - 3\alpha_3)\sigma^2/d, \quad (21.44)$$

$$c_2 = (2\alpha_4 - 3\alpha_3 - 6)/d, \quad d = (10\alpha_4 - 12\alpha_3 - 18) \quad (21.45)$$

(see Kendall and Stuart (1969)). These parameters can be easily estimated using  $\hat{\sigma}$ ,  $\hat{\alpha}_3$  and  $\hat{\alpha}_4$  and then used to give us some idea about the nature of the departure from non-normality. Such information will be of considerable interest in tackling non-normality (see subsection (3)). In the case of normality  $c_1 = c_2 = 0 \Rightarrow \alpha_3 = 0, \alpha_4 = 3$ . Departures from normality within the Pearson family of particular interest are the following cases:

- (a)  $c_2 = 0, c_1 \neq 0$ . This gives rise to *gamma-type distributions* with the chi-square an important member of this class of distributions. For

$$Z \sim \chi^2(m), \quad \alpha_3 = \left(\frac{2^3}{m}\right)^{\frac{1}{2}}, \quad \alpha_4 = 3 + \frac{12}{m}, \quad m \geq 1. \quad (21.46)$$

- (b)  $c_1 = 0, c_0 > 0, c_2 > 0$ . An important member of this class of distributions is the *Student's t*. For  $Z \sim t(m)$ ,  $\alpha_3 = 0, \alpha_4 = 3 + 6/(m-4)$ , ( $m \geq 4$ ).
- (c)  $c_1 < 0 < c_2$ . This gives rise to *beta-type distributions* which are directly related to the chi-square and *F*-distributions. In particular if  $Z_i \sim \chi^2(m_i)$ ,  $i = 1, 2$ , and  $Z_1, Z_2$  are independent, then

$$Z = \left( \frac{Z_1}{Z_1 + Z_2} \right) \sim B\left(\frac{m_1}{2}, \frac{m_2}{2}\right), \quad (21.47)$$

where  $B(m_1/2, m_2/2)$  denotes the beta distribution with parameters  $m_1/2$  and  $m_2/2$ .

As argued above normality within the Pearson family is characterised by

$$\alpha_3 = (\mu_3/\sigma^3) = 0 \quad \text{and} \quad \alpha_4 = (\mu_4/\sigma^4) = 3. \quad (21.48)$$

It is interesting to note that (48) also characterises normality within the ‘short’ (first four moments) Gram–Charlier expansion:

$$g(z) = [1 - \frac{1}{6}\alpha_3(z^3 - 3z) + \frac{1}{24}(\alpha_4 - 3)(z^4 - 6z^2 + 3)]\Phi(z) \quad (21.49)$$

(see Section 10.6).

Bera and Jarque (1982) using the Pearson family as the parametric alternative derived the following skewness–kurtosis test as a Lagrange

multiplier test:

$$\tau_N^*(y) = \left[ \frac{T}{6} \hat{\alpha}_3^2 + \frac{T}{24} (\hat{\alpha}_4 - 3)^2 \right]_{\alpha}^{H_0} \sim \chi^2(2) \quad (21.50)$$

where

$$\hat{\alpha}_3 = \left[ \left( \frac{1}{T} \sum_{t=1}^T \hat{u}_t^3 \right) / \left( \frac{1}{T} \sum_{t=1}^T \hat{u}_t^2 \right)^{\frac{3}{2}} \right] \quad (21.51)$$

$$\hat{\alpha}_4 = \left[ \left( \frac{1}{T} \sum_{t=1}^T \hat{u}_t^4 \right) / \left( \frac{1}{T} \sum_{t=1}^T \hat{u}_t^2 \right)^2 \right]. \quad (21.51)$$

The rejection region is defined by

$$C_1 = \{y: \tau_N^*(y) > c_\alpha\}, \quad \int_{c_\alpha}^{\infty} d\chi^2(2) = \alpha. \quad (21.53)$$

A less formal derivation of the test can be based on the asymptotic distributions of  $\hat{\alpha}_3$  and  $\hat{\alpha}_4$ :

$$\sqrt{T} \hat{\alpha}_3 \stackrel[H_0]{\alpha}{\sim} N(0, 6) \quad (21.54)$$

$$\sqrt{T}(\hat{\alpha}_4 - 3) \stackrel[H_0]{\alpha}{\sim} N(0, 24). \quad (21.55)$$

With  $\hat{\alpha}_3$  and  $\hat{\alpha}_4$  being asymptotically independent (see Kendall and Stuart (1969)) we can add the squares of their standardised forms to derive (50); see Section 6.3.

Let us consider the skewness–kurtosis test for the money equation

$$m_t = 2.896 + 0.690 y_t + 0.865 p_t - 0.055 i_t + \hat{u}_t, \quad (21.56)$$

(1.034)	(0.105)	(0.020)	(0.013)	(0.039)
---------	---------	---------	---------	---------

$$R^2 = 0.995, \quad \bar{R}^2 = 0.995, \quad s = 0.0393, \quad \log L = 147.4,$$

$$T = 80, \quad \hat{\alpha}_3^2 = 0.005, \quad (\hat{\alpha}_4 - 3)^2 = 0.145.$$

Thus,  $\tau_N^*(y) = 0.55$  and since  $c_\alpha = 5.99$  for  $\alpha = 0.5$  we can deduce that under the assumption that the other assumptions underlying the linear regression model are *valid* the null hypothesis  $H_0: \alpha_3 = 0$  and  $\alpha_4 = 3$  is not rejected for  $\alpha = 0.05$ .

There are several things to note about the above skewness–kurtosis test. Firstly, it is an asymptotic test and caution should be exercised when the sample size  $T$  is small. For higher-order approximations of the finite sample distribution of  $\hat{\alpha}_3$  and  $\hat{\alpha}_4$  see Pearson, D'Agostino and Bowman (1977), Bowman and Shenton (1975), *inter alia*. Secondly, the test is sensitive to 'outliers' ('unusually large' deviations). This can be both a blessing and a

hindrance. The first reaction of a practitioner whose residuals fail this normality test is to look for such outliers. When the apparent non-normality can be explained by the presence of these outliers the problem can be solved when the presence of the outliers can itself be explained. Otherwise, alternative forms of tackling non-normality need to be considered as discussed below. Thirdly, in the case where the standard error of the regression  $\hat{\sigma}$  is relatively large (because very little of the variation in  $y_t$  is actually explained), it can dominate the test statistic  $\tau_N^*(y)$ . It will be suggested in Chapter 23 that the acceptance of normality in the case of the money equation above is largely due to this. Fourthly, rejection of normality using the skewness–kurtosis test gives us no information as to the nature of the departures from normality unless it is due to the presence of outliers.

A natural way to extend the skewness–kurtosis test is to include *cumulants* of order higher than four which are zero under normality (see Appendix 6.1).

### (3) *Tackling non-normality*

When the normality assumption is invalid there are two possible ways to proceed. One is to postulate a more appropriate distribution for  $D(y_t/X_t; \theta)$  and respecify the linear regression model accordingly. This option is rarely considered, however, because most of the results in this context are developed under the normality assumption. For this reason the second way to proceed, based on normalising transformations, is by far the most commonly used way to tackle non-normality. This approach amounts to applying a transformation to  $y_t$  or/and  $X_t$  so as to induce normality. Because of the relationship between normality, linearity and homoskedasticity these transformations commonly induce linearity and homoskedasticity as well.

One of the most interesting family of transformations in this context is the Box–Cox (1964) transformation. For an arbitrary random variable  $Z$  the Box–Cox transformation takes the form

$$Z^* = \frac{Z^\delta - 1}{\delta}, \quad 0 \leq \delta \leq 1. \quad (21.57)$$

Of particular interest are the three cases:

$$(i) \quad \delta = -1, \quad Z^* = Z^{-1} - \text{reciprocal}; \quad (21.58)$$

$$(ii) \quad \delta = 0.5, \quad Z^* = (Z)^{\frac{1}{2}} - \text{square root}; \quad (21.59)$$

$$(iii) \quad \delta=0, \quad Z^* = \log_e Z - \text{logarithmic} \quad (21.60)$$

(note:  $\lim_{\delta \rightarrow 0} Z^* = \log_e Z$ ).

The first two cases are not commonly used in econometric modelling because of the difficulties involved in interpreting  $Z^*$  in the context of an empirical econometric model. Often, however, the square-root transformation might be convenient as a homoskedasticity inducing transformation. This is because certain economic time-series exhibit variances which change with its trending mean ( $m_t$ ), i.e.  $\text{Var}(Z_t) = m_t \sigma^2$ ,  $t = 1, 2, \dots, T$ . In such cases the square-root transformation can be used as a variance-stabilising one (see Appendix 21.1) since  $\text{Var}(Z_t^*) \approx \sigma^2$ .

The logarithmic transformation is of considerable interest in econometric modelling for a variety of reasons. Firstly, for a random variable  $Z_t$  whose distribution is closer to the log normal, gamma or chi-square (i.e. positively skewed), the distribution of  $\log_e Z_t$  is approximately normal (see Johnson and Kotz (1970)). The  $\log_e$  transformation induces ‘near symmetry’ to the original skewed distribution and allows  $Z_t^*$  to take negative values even though  $Z$  could not. For economic data which take only positive values this can be a useful transformation to achieve near normality. Secondly, the  $\log_e$  transformation can be used as a variance-stabilising transformation in the case where the heteroskedasticity takes the form

$$\text{Var}(y_t / \mathbf{X}_t = \mathbf{x}_t) = \sigma_t^2 = (\mu_t)^{\frac{1}{2}} \sigma^2, \quad t = 1, 2, \dots, T. \quad (21.61)$$

For  $y_t^* = \log_e y_t$ ,  $\text{Var}(y_t^* / \mathbf{X}_t = \mathbf{x}_t) = \sigma^2$ ,  $t = 1, 2, \dots, T$ . Thirdly, the log transformation can be used to define useful economic concepts such as elasticities and growth rates. For example, in the case of the money equation considered above the variables are all in logarithmic form and the estimated coefficients can be interpreted as elasticities (assuming that the estimated equation constitutes a well-defined statistical model; a doubtful assumption). Moreover, the growth rate of  $Z_t$  defined by  $Z^+ = (Z_t - Z_{t-1}) / Z_{t-1}$  can be approximated by  $\Delta \log_e Z_t \equiv \log_e Z_t - \log_e Z_{t-1}$  because  $\Delta \log_e Z_t \approx \log(1 + Z_t^+) \approx Z_t^+$ .

In practice the Box–Cox transformation can be used with  $\delta$  unspecified and let the data determine its value (see Zarembka (1974)). For the money equation the original variables  $M_t$ ,  $Y_t$ ,  $P_t$  and  $I_t$  were used in the Box–Cox transformed equation:

$$\left( \frac{M_t^\delta - 1}{\delta} \right) = \beta_1 + \beta_2 \left( \frac{Y_t^\delta - 1}{\delta} \right) + \beta_3 \left( \frac{P_t^\delta - 1}{\delta} \right) + \beta_4 \left( \frac{I_t^\delta - 1}{\delta} \right) + u_t \quad (21.62)$$

and allowed the data to determine the value of  $\delta$ . The estimated  $\delta$  value chosen was  $\delta = 0.530$  and

$$\begin{aligned}\hat{\beta}_1 &= 0.252, & \hat{\beta}_2 &= 0.865, & \hat{\beta}_3 &= 0.005, & \hat{\beta}_4 &= -0.000\ 07. \\ (0.223) & & (0.119) & & (0.0001) & & (0.000\ 02)\end{aligned}$$

'Does this mean that the original logarithmic transformation is inappropriate?' The answer is, not necessarily. This is because the estimated value of  $\delta$  depends on the estimated equation being a well-defined statistical GM (no misspecification). In the money equation example there is enough evidence to suggest that various forms of misspecification are indeed present (see also Sections 21.3–7 and Chapter 22).

The alternative way to tackle non-linearity by postulating a more appropriate form for the distribution of  $Z_t$  remains largely unexplored. Most of the results in this direction are limited to multivariate distributions closely related to the normal such as the elliptical family of distributions (see Section 21.3 below). On the question of robust estimation see Amemiya (1985).

### 21.3 Linearity

As argued above, the assumption

$$E(y_t/X_t = x_t) = \beta' x_t, \quad (21.63)$$

where  $\beta = \Sigma_{22}^{-1} \sigma_{21}$  can be viewed as a consequence of the assumption that  $Z_t \sim N(\mathbf{0}, \Sigma)$ ,  $t \in \mathbb{T}$  ( $Z_t$  is a normal IID sequence of r.v.'s). The form of (63) is not as restrictive as it seems at first sight because  $E(y_t/X_t^* = x_t^*)$  can be non-linear in  $x_t^*$  but linear in  $x_t = l(x_t^*)$  where  $l(\cdot)$  is a well-behaved transformation such as  $x_t = \log x_t^*$  and  $x_t = (x_t^*)^{1/2}$ . Moreover, terms such as

$$c_0 + c_1 t + c_2 t^2 + \cdots + c_m t^m$$

and

$$c_0 + \sum_{i=1}^h \left( \alpha_i \cos\left(\frac{2\pi}{c_i} t\right) + \gamma_i \sin\left(\frac{2\pi}{c_i} t\right) \right), \quad (21.64)$$

purporting to model a time trend and seasonal effects respectively, can be easily accommodated as part of the constant. This can be justified in the context of the above analysis by extending  $Z_t \sim N(\mathbf{0}, \Sigma)$ ,  $t \in \mathbb{T}$ , to  $Z_t \sim N(\mathbf{m}_t, \Sigma)$ ,  $t \in \mathbb{T}$ , being an independent sequence of random vectors where the mean is a function of time and the covariance matrix is the same for all  $t \in \mathbb{T}$ . The sequence of random vectors  $\{Z_t, t \in \mathbb{T}\}$  in this case constitutes a non-stationary sequence (see Section 21.5 below). The non-linearities of interest in this section are the ones which cannot be accommodated into a linear conditional mean after transformation.

It is important to note that postulating (63), without assuming normality of  $D(y_t, \mathbf{X}_t; \psi)$ , we limit the class of symmetric distributions in which  $D(y_t, \mathbf{X}_t; \psi)$  could belong to that of elliptical distributions, denoted by  $EL(\mu, \Sigma)$  (see Kelker (1970)). These distributions provide an extension of the multivariate normal distribution which preserve its bell-like shape and symmetry. Assuming that

$$\begin{pmatrix} y_t \\ \mathbf{x}_t \end{pmatrix} \sim EL\left(\begin{pmatrix} 0 \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \Sigma_{22} \end{pmatrix}\right) \quad (21.65)$$

implies that

$$E(y_t | \mathbf{X}_t = \mathbf{x}_t) = \sigma_{12} \Sigma_{22}^{-1} \mathbf{x}_t \quad (21.66)$$

and

$$\text{Var}(y_t | \mathbf{X}_t = \mathbf{x}_t) = g(\mathbf{x}_t)(\sigma_{11} - \sigma_{12} \Sigma_{22}^{-1} \sigma_{21}). \quad (21.67)$$

This shows that the assumption of linearity is not as sensitive to some departures from normality as the homoskedasticity assumption. Indeed, homoskedasticity of the conditional variance characterises the normal distribution within the class of elliptical distributions (see Chmielewski (1981)).

### (1) Implications of non-linearity

Let us consider the implications of non-linearity for the results of Chapter 19 related to the estimation, testing and prediction in the context of the linear regression model. In particular, ‘what are the implications of assuming that  $D(\mathbf{Z}_t; \psi)$  is not normal and

$$E(y_t | \mathbf{X}_t = \mathbf{x}_t) = h(\mathbf{x}_t), \quad (21.68)$$

where  $h(\mathbf{x}_t) \neq \beta' \mathbf{x}_t$ ?’

In Chapter 19 the statistical GM for the linear regression model was defined to be

$$y_t = \beta' \mathbf{x}_t + u_t, \quad (21.69)$$

thinking that  $\mu_t^* = E(y_t | \mathbf{X}_t = \mathbf{x}_t) = \beta' \mathbf{x}_t$  and  $u_t^* = y_t - \mu_t^*$  with  $E(u_t | \mathbf{X}_t = \mathbf{x}_t) = 0$ ,  $E(\mu_t^* u_t^* | \mathbf{X}_t = \mathbf{x}_t) = 0$  and  $E(u_t^{*2} | \mathbf{X}_t = \mathbf{x}_t) = \sigma^2$ . The ‘true’ statistical GM, however, is

$$y_t = h(\mathbf{x}_t) + \varepsilon_t, \quad (21.70)$$

where  $\mu_t = E(y_t | \mathbf{X}_t = \mathbf{x}_t) = h(\mathbf{x}_t)$  and  $\varepsilon_t = y_t - E(y_t | \mathbf{X}_t = \mathbf{x}_t)$ . Comparing (69) and (70) we can see that the error term in the former is no longer white noise but  $u_t = y_t - \beta' \mathbf{x}_t = h(\mathbf{x}_t) - \beta' \mathbf{x}_t + \varepsilon_t \equiv g(\mathbf{x}_t) + \varepsilon_t$ . Moreover,

$E(u_t^*/\mathbf{X}_t = \mathbf{x}_t) = g(\mathbf{x}_t)$ ,  $E(\mu_t^* u_t^*) \neq 0$  and

$$E(u_t^2/\mathbf{X}_t = \mathbf{x}_t) = g(\mathbf{x}_t)^2 + \sigma_e^2. \quad (21.71)$$

In view of these properties of  $u_t$  we can deduce that for

$$\mathbf{e} \equiv (g(\mathbf{x}_1), g(\mathbf{x}_2), \dots, g(\mathbf{x}_T))', \quad (21.72)$$

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e} \neq \boldsymbol{\beta}, \quad (21.72)$$

and

$$E(s^2) = \sigma^2 + \mathbf{e}' \frac{\mathbf{M}_x}{T-k} \mathbf{e} \neq \sigma^2, \quad \mathbf{M}_x = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}', \quad (21.73)$$

because  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{e} + \boldsymbol{\varepsilon}$  not  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ . Moreover,  $\hat{\boldsymbol{\beta}}$  and  $s^2$  are also inconsistent estimators of  $\boldsymbol{\beta}$  and  $\sigma^2$  unless the approximation error  $\mathbf{e}$  satisfies  $(1/T)\mathbf{X}'\mathbf{e} \rightarrow 0$  and  $(1/T)\mathbf{e}'\mathbf{M}_x\mathbf{e} \rightarrow 0$  as  $T \rightarrow \infty$  respectively. That is, unless  $h(\mathbf{x}_t)$  is not ‘too’ non-linear and the non-linearity decreases with  $T$ ,  $\hat{\boldsymbol{\beta}}$  and  $s^2$  are inconsistent estimators of  $\boldsymbol{\beta}$  and  $\sigma^2$ .

As we can see, the consequences of non-linearity are quite serious as far as the properties of  $\hat{\boldsymbol{\beta}}$  and  $s^2$  are concerned, being biased and inconsistent estimators of  $\boldsymbol{\beta}$  and  $\sigma^2$ , in general. What is more, the testing and prediction results derived in Chapter 19 are generally invalid in the case of non-linearity. In view of this the question arises as to what is it we are estimating by  $s^2$  and  $\hat{\boldsymbol{\beta}}$  in (70)?

Given that  $u_t = (h(\mathbf{x}_t) - \boldsymbol{\beta}'\mathbf{x}_t) + \varepsilon_t$  we can think of  $\hat{\boldsymbol{\beta}}$  as an estimator of  $\boldsymbol{\beta}^*$  where  $\boldsymbol{\beta}^*$  is the parameter which minimises the mean square error of  $u_t$ , i.e.

$$\boldsymbol{\beta}^* = \min_{\boldsymbol{\beta}} \sigma^2(\boldsymbol{\beta}) \quad \text{where } \sigma^2(\boldsymbol{\beta}) \equiv E(u_t^2). \quad (21.74)$$

This is because  $[\partial\sigma^2(\boldsymbol{\beta})]/\partial\boldsymbol{\beta} = (-2)E[(h(\mathbf{x}_t) - \boldsymbol{\beta}'\mathbf{x}_t)\mathbf{x}_t'] = 0$  (assuming that we can differentiate inside the expectation operator). Hence,  $\boldsymbol{\beta}^* = E(\mathbf{x}_t\mathbf{x}_t')^{-1}E(h(\mathbf{x}_t)\mathbf{x}_t') = \Sigma_{22}^{-1}\boldsymbol{\sigma}_{2h}$ , say. Moreover,  $s^2$  can be viewed as the natural estimator of  $\sigma^2(\boldsymbol{\beta}^*)$ . That is,  $\hat{\boldsymbol{\beta}}$  and  $s^2$  are the natural estimators of a least-squares approximation  $\boldsymbol{\beta}^*\mathbf{x}_t$  to the unknown function  $h(\mathbf{x}_t)$  and the least-squares approximation-error respectively. What is more, we can show that  $\hat{\boldsymbol{\beta}} \xrightarrow{\text{a.s.}} \boldsymbol{\beta}^*$  and  $s^2 \xrightarrow{\text{a.s.}} \sigma^2(\boldsymbol{\beta}^*)$  (see White (1980)).

## (2) Testing for non-linearity

In view of the serious implications of non-linearity for the results of Chapter 19 it is important to be able to test for departures from the linearity assumption. In particular we need to construct tests for

$$H_0: E(y_t/\mathbf{X}_t = \mathbf{x}_t) = \boldsymbol{\beta}'\mathbf{x}_t \quad (21.75)$$

against

$$H_1: E(y_t | \mathbf{X}_t = \mathbf{x}_t) = h(\mathbf{x}_t). \quad (21.76)$$

This, however, raises the question of postulating a particular functional form for  $h(\mathbf{x}_t)$  which is not available unless we are prepared to assume a particular form for  $D(\mathbf{Z}_t; \boldsymbol{\psi})$ . Alternatively, we could use the parametrisation related to the Kolmogorov–Gabor and systematic component polynomials introduced in Section 21.2.

Using, say, a third-order Kolmogorov–Gabor polynomial ( $KG(3)$ ) we can postulate the alternative statistical GM:

$$y_t = \boldsymbol{\beta}'_0 \mathbf{x}_t + \gamma'_2 \boldsymbol{\psi}_{2t} + \gamma'_3 \boldsymbol{\psi}_{3t} + \varepsilon_t \quad (21.77)$$

where  $\boldsymbol{\psi}_{2t}$  includes the second-order terms

$$x_{it} x_{jt}, \quad i \geq j, \quad i, j = 2, 3, \dots, k, \quad (21.78)$$

and  $\boldsymbol{\psi}_{3t}$  the third-order terms

$$x_{it} x_{jt} x_{lt}, \quad i \geq j \geq l, \quad i, j, l = 2, 3, \dots, k. \quad (21.79)$$

Note that  $x_{1t}$  is assumed to be the constant.

Assuming that  $T$  is large enough to enable us to estimate (77) we can test linearity in the form of:

$$H_0: \gamma_2 = \mathbf{0} \quad \text{and} \quad \gamma_3 = \mathbf{0}, \quad H_1: \gamma_2 \neq \mathbf{0} \quad \text{or} \quad \gamma_3 \neq \mathbf{0}$$

using the usual  $F$ -type test (see Section 21.1). An asymptotically equivalent test can be based on the  $R^2$  of the auxiliary regression:

$$\hat{u}_t = (\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})' \mathbf{x}_t + \gamma'_2 \boldsymbol{\psi}_{2t} + \gamma'_3 \boldsymbol{\psi}_{3t} + \varepsilon_t \quad (21.80)$$

using the Lagrange multiplier test statistic

$$LM(\mathbf{y}) = TR^2 = T \left( \frac{RRSS - URSS}{RRSS} \right)^{H_0} \underset{\alpha}{\sim} \chi^2(q) \quad (21.81)$$

$q$  being the number of restrictions (see Engle (1984)). Its rejection region is

$$C_1 = \{ \mathbf{y}: LM(\mathbf{y}) > c_\alpha \}, \quad \int_{c_\alpha}^\infty d\chi^2(q) = \alpha.$$

For small  $T$  the  $F$ -type test is preferable in practice because of the degrees of freedom adjustment; see Section 19.5.

Using the polynomial in  $\mu_t$  we can postulate the alternative GM of the form:

$$y_t = \boldsymbol{\beta}'_* \mathbf{x}_t + c_2 \mu_t^2 + c_3 \mu_t^3 + \cdots + c_m \mu_t^m + v_t \quad (21.82)$$

where  $\mu_t = \boldsymbol{\beta}' \mathbf{x}_t$ . A direct comparison between (75) and (82) gives rise to a

RESET type test (see Ramsey (1974)) for linearity based on  $H_0: c_4 = c_3 = \dots = c_m = 0, H_1: c_i \neq 0, i = 2, \dots, m$ . Again this can be tested using the  $F$ -type test or the LM test both based on the auxiliary regression:

$$\hat{u}_t = (\boldsymbol{\beta}_* - \hat{\boldsymbol{\beta}})' \mathbf{x}_t + \sum_{i=2}^m c_i \hat{\mu}_t^i + v_t, \quad \hat{\mu}_t = \boldsymbol{\beta}' \mathbf{x}_t. \quad (21.83)$$

Let us apply these tests to the money equation estimated in Section 19.4. The  $F$ -test based on (77) with terms up to third order (but excluding  $\tilde{y}_t^3$  because of collinearity with  $\tilde{y}_t$ ) yielded:

$$FT(\mathbf{y}) = \frac{0.117520 - 0.045477}{0.045477} \left( \frac{67}{9} \right) = 11.72.$$

Given that  $c_x = 2.02$  the null hypothesis of *linearity is strongly rejected*. Similarly, the RESET type test based on (82) with  $m=4$  (excluding  $\hat{\mu}_t^3$  because of collinearity with  $\hat{\mu}_t$ ) yielded:

$$FT(\mathbf{y}) = \frac{0.117520 - 0.06028}{0.06028} \left( \frac{74}{2} \right) = 35.13.$$

Again, with  $c_x = 3.12$  *linearity is strongly rejected*.

It is important to note that although the RESET type test is based on a more restrictive form of the alternative (compare (77) with (82)) it might be the only test available in the case where the degrees of freedom are at a premium (see Chapter 23).

### (3) Tackling non-linearity

As argued in Section 21.1 the results of the various misspecification tests should be considered simultaneously because the assumptions are closely interrelated. For example in the case of the estimated money equation it is highly likely that the linearity assumption was rejected because the independent sample assumption [8] is invalid. In cases, however, where the source of the departure is indeed the normality assumption (leading to non-linearity) we need to consider the question of how to proceed by relaxing the normality of  $\{\mathbf{Z}_{t-1}, t \in \mathbb{T}\}$ . One way to proceed from this is to postulate a general distribution  $D(y_t, \mathbf{X}_t; \psi)$  and derive the specific form of the conditional expectation

$$E(y_t | \mathbf{X}_t = \mathbf{x}_t) = h(\mathbf{x}_t). \quad (21.84)$$

Choosing the form of  $D(y_t, \mathbf{X}_t; \psi)$  will determine both the form of the conditional expectation as well as the conditional variance (see Chapter 7).

An alternative way to proceed is to use some normalising transformation

on the original variables  $y_t$  and  $\mathbf{X}_t$  so as to ensure that the transformed variables  $y_t^*$  and  $\mathbf{X}_t^*$  are indeed jointly normal and hence

$$E(y_t^*/\mathbf{X}_t^* = \mathbf{x}_t^*) = \boldsymbol{\beta}^{*\prime} \mathbf{x}_t^* \quad (21.85)$$

and

$$\text{Var}(y_t^*/\mathbf{X}_t^* = \mathbf{x}_t^*) = \sigma^2. \quad (21.86)$$

The transformations considered in Section 21.2 in relation to normality are also directly related to the problem of non-linearity. The Box–Cox transformation can be used with different values of  $\delta$  for each random variable involved to linearise highly non-linear functional forms. In such a case the transformed r.v.'s take the general form

$$X_{it}^* = \left( \frac{X_{it}^{\delta_i} - 1}{\delta_i} \right), \quad i = 1, 2, \dots, k \quad (21.87)$$

(see Box and Tidwell (1962)).

In practice non-linear regression models are used in conjunction with the normality of the conditional distribution (see Judge *et al.* (1985), *inter alia*). The question which naturally arises is, ‘how can we reconcile the non-linearity of the conditional expectation and the normality of  $D(y_t/\mathbf{X}_t; \boldsymbol{\theta})$ ?’ As mentioned in Section 19.2, the linearity of  $\mu_t \equiv E(y_t/\mathbf{X}_t = \mathbf{x}_t)$  is a direct consequence of the normality of the joint distribution  $D(y_t, \mathbf{X}_t; \boldsymbol{\psi})$ . One way the non-linearity of  $E(y_t/\mathbf{X}_t = \mathbf{x}_t)$  and the normality of  $D(y_t/\mathbf{X}_t; \boldsymbol{\theta})$  can be reconciled is to argue that the conditional distribution is normal in the transformed variables  $\mathbf{X}_t^* = h(\mathbf{X}_t)$ , i.e.  $D(y_t/\mathbf{X}_t^* = \mathbf{x}_t^*)$  linear in  $\mathbf{x}_t^*$  but non-linear in  $\mathbf{x}_t$ , i.e.

$$E(y_t/\mathbf{X}_t = \mathbf{x}_t) = g(\mathbf{x}_t, \gamma). \quad (21.88)$$

Moreover, the *parameters of interest* are not the linear regression parameters  $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}, \sigma^2)$  but  $\boldsymbol{\phi} \equiv (\gamma, \sigma_*^2)$ . It must be emphasised that non-linearity in the present context refers to both non-linearity in parameters ( $\gamma$ ) and variables ( $\mathbf{X}_t$ ).

Non-linear regression models based on the statistical GM:

$$y_t = g(\mathbf{x}_t, \gamma) + u_t \quad (21.89)$$

can be estimated by least-squares based on the minimisation of

$$S_T(\gamma) = \sum_{t=1}^T (y_t - g(\mathbf{x}_t, \gamma))^2. \quad (21.90)$$

This minimisation will give rise to certain non-linear normal equations which can be solved numerically (see Harvey (1981), Judge *et al.* (1985), Malinvaud (1970), *inter alia*) to provide least-squares estimators for  $\gamma$ :  $m \times 1$ .  $\sigma_*^2$  can then be estimated by

$$s^2 = \frac{1}{T-k} \sum_t (y_t - g(\mathbf{x}_t, \hat{\gamma}))^2. \quad (21.91)$$

Statistical analysis of these parameters of interest is based on asymptotic theory (see Amemiya (1983) for an excellent discussion of some of these results).

## 21.4 Homoskedasticity

The assumption that  $\text{Var}(y_t | \mathbf{X}_t = \mathbf{x}_t) = \sigma^2$  is free of  $\mathbf{x}_t$  is a consequence of the assumption that  $D(y_t, \mathbf{X}_t; \psi)$  is multivariate normal. As argued above, the assumption of homoskedasticity is inextricably related to the assumption of normality and we cannot retain one and reject the other uncritically. Indeed, as mentioned above, homoskedasticity of  $\text{Var}(y_t | \mathbf{X}_t = \mathbf{x}_t)$  characterises the normal distribution within the elliptical class. For argument's sake, let us assume that the probability model is in fact based on  $D(\beta' \mathbf{x}_t, \sigma_t^2)$  where  $D(\cdot)$  is some unknown distribution and  $\sigma_t^2 = h(\mathbf{x}_t)$ .

### (1) Implications of heteroskedasticity

As far as the estimators  $\hat{\beta}$  and  $s^2$  are concerned we can show that

- (i)  $E(\hat{\beta}) = \beta$ , i.e.  $\hat{\beta}$  is an unbiased estimator of  $\beta$ .
- (ii)  $\text{Cov}(\hat{\beta}) = (\mathbf{X}' \mathbf{X})^{-1} (\mathbf{X}' \Omega \mathbf{X}) (\mathbf{X}' \mathbf{X})^{-1}$ , (21.92)

where

$$\Omega = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_T^2) = \sigma^2 \Lambda.$$

If  $\lim_{T \rightarrow \infty} \underset{P}{\text{cov}}((1/T) \mathbf{X}' \Omega \mathbf{X})$  is bounded and non-singular then

- (iii)  $\hat{\beta} \rightarrow \beta$ , i.e.  $\hat{\beta}$  is a consistent estimator of  $\beta$ .

These results suggest that  $\hat{\beta} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}$  retains some desirable properties such as unbiasedness and consistency, although it might be inefficient.  $\hat{\beta}$  is usually compared with the so-called *generalised least-squares* (GLS) estimator of  $\beta$ ,  $\bar{\beta}$ , derived by minimising

$$l(\beta) = (\mathbf{y} - \mathbf{X}\beta)' \Omega^{-1} (\mathbf{y} - \mathbf{X}\beta), \quad (21.93)$$

$$\begin{aligned}\frac{\partial l(\beta)}{\partial \beta} = 0 \Rightarrow \bar{\beta} &= (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}\mathbf{y} \\ &\equiv \left( \sum_t \left( \frac{\mathbf{x}_t}{\sigma_t} \right) \left( \frac{\mathbf{x}_t}{\sigma_t} \right)' \right)^{-1} \sum_t \left( \frac{\mathbf{x}_t}{\sigma_t} \right) \left( \frac{\mathbf{y}_t}{\sigma_t} \right).\end{aligned}\quad (21.94)$$

Given that

$$\text{Cov}(\bar{\beta}) = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1} \quad (21.95)$$

and

$$\text{Cov}(\hat{\beta}) \geq \text{Cov}(\bar{\beta}) \quad (21.96)$$

(see Dhrymes (1978)),  $\hat{\beta}$  is said to be relatively inefficient. It must be emphasised, however, that this efficiency comparison is based on the presupposition that  $\Lambda$  is known a priori and thus the above efficiency comparison is largely irrelevant. It should surprise nobody to ‘discover’ that supplementing the statistical model with additional information we can get a more efficient estimator. Moreover, when  $\Lambda$  is known there is no need for GLS because we can transform the original variables in order to return to a homoskedastic conditional variance of the form

$$\text{Var}(y_t^*/\mathbf{X}_t^* = \mathbf{x}_t^*) = \sigma^2, \quad t = 1, \dots, T. \quad (21.97)$$

This can be achieved by transforming  $\mathbf{y}$  and  $\mathbf{X}$  into

$$\mathbf{y}^* = \mathbf{H}\mathbf{y} \quad \text{and} \quad \mathbf{X}^* = \mathbf{H}\mathbf{X} \quad \text{where} \quad \mathbf{H}'\mathbf{H} = \Lambda^{-1}. \quad (21.98)$$

In terms of the transformed variables the statistical GM takes the form

$$\mathbf{y}^* = \mathbf{X}^*\beta + \mathbf{u}^* \quad (21.99)$$

and the linear regression assumptions are valid for  $y_t^*$  and  $\mathbf{X}_t^*$ . Indeed, it can be verified that

$$\hat{\beta} = (\mathbf{X}^{*'}\mathbf{X}^*)^{-1}\mathbf{X}^{*'}\mathbf{y}^* = (\mathbf{X}'\Lambda^{-1}\mathbf{X})^{-1}\mathbf{X}'\Lambda^{-1}\mathbf{y} = \bar{\beta}. \quad (21.100)$$

Hence, the GLS estimator is rather unnecessary in the case where  $\Lambda$  is known a priori.

The question which naturally arises at this stage is, ‘what happens when  $\Omega$  is unknown?’ The conventional wisdom has been that since  $\Omega$  involves  $T$  unknown incidental parameters and increases with the sample size it is clearly out of the question to estimate  $T+k$  parameters from  $T$  observations. Moreover, although  $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$  is both unbiased and consistent  $s^2(\mathbf{X}'\mathbf{X})^{-1}$  is an inconsistent estimator of  $\text{Cov}(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$  and the difference

$$\sigma^2(\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\Omega\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1} \quad (21.101)$$

can be positive or negative. Hence, no inference on  $\beta$ , based on  $\hat{\beta}$ , is possible since for a consistent estimator of  $\text{Cov}(\hat{\beta})$  we need to know  $\Omega$  (or estimate it consistently). So, the only way to proceed is to model  $\sigma_t^2$  so as to 'solve' the incidental parameters problem.

Although there is an element of truth in the above viewpoint White (1980) pointed out that for consistent inference based on  $\hat{\beta}$  we do not need to estimate  $\Omega$  by itself but  $(\mathbf{X}'\Omega\mathbf{X})$ , and the two problems are not equivalent. The natural estimator  $\sigma_t^2$  is  $\hat{u}_t^2 = (y_t - \hat{\beta}'\mathbf{x}_t)^2$ , which is clearly unsatisfactory because it is based on only one observation and no further information accrues by increasing the sample size. On the other hand, there is a perfectly acceptable estimator for  $(\mathbf{X}'\Omega\mathbf{X})$  coming in the form of

$$\hat{\mathbf{W}}_T = \frac{1}{T} \sum_{t=1}^T \hat{u}_t^2 \mathbf{x}_t \mathbf{x}_t', \quad (21.102)$$

for which information accrues as  $T \rightarrow \infty$ . White (1980) showed that under certain regularity restrictions

$$\hat{\beta} \xrightarrow{\text{a.s.}} \beta, \quad (21.103)$$

and

$$\hat{\mathbf{W}}_T \xrightarrow{\text{a.s.}} (\mathbf{X}'\Omega\mathbf{X}). \quad (21.104)$$

The most important implication of this is that consistent inference, such as the  $F$ -test, is asymptotically justifiable, although the loss in efficiency should be kept in mind. In particular a test for heteroskedasticity could be based on the difference

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} - \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \quad (21.105)$$

Before we consider this test it is important to summarise the argument so far.

Under the assumption that the probability model is based on the distribution  $D(\beta'\mathbf{x}_t, \sigma_t^2)$ , although no estimator of  $\Omega = \text{diag}(\sigma_1^2, \dots, \sigma_T^2)$  is possible,  $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$  is both unbiased and consistent (under certain conditions) and a consistent estimator of  $\text{Cov}(\hat{\beta})$  is available in the form of  $\hat{\mathbf{W}}_T$ . This enables us to use  $\hat{\beta}$  for hypothesis testing related to  $\beta$ . The argument of 'modelling'  $\sigma_t^2$  will be taken up after we consider the question of testing for departures from homoskedasticity.

## (2) Testing departures from homoskedasticity

White (1980), after proposing a consistent estimator for  $\mathbf{X}'\Omega\mathbf{X}$ , went on to

use the difference (equivalent to (105)):

$$(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}) - \sigma^2(\mathbf{X}'\mathbf{X}) \quad (21.106)$$

to construct a test for departures from homoskedasticity. (106) can be expressed in the form

$$\sum_{t=1}^T (E(u_t^2) - \sigma^2) \mathbf{x}_t \mathbf{x}_t' \quad (21.107)$$

and a test for heteroskedasticity could be based on the statistic

$$\frac{1}{T} \sum_{t=1}^T (\hat{u}_t^2 - \hat{\sigma}^2) \mathbf{x}_t \mathbf{x}_t', \quad (21.108)$$

the natural estimator of (107). Given that (108) is symmetric we can express the  $\frac{1}{2}k(k-1)$  different elements in the form

$$\frac{1}{T} \sum_{t=1}^T (\hat{u}_t^2 - \hat{\sigma}^2) \boldsymbol{\psi} \quad (21.109)$$

where

$$\begin{aligned} \boldsymbol{\psi}_t &= (\psi_{1t}, \psi_{2t}, \dots, \psi_{mt})', \quad \psi_{lt} = x_{it} x_{jt}, \\ i &\geq j, \quad i, j = 1, 2, \dots, k, \quad l = 1, 2, \dots, m, \quad m = \frac{1}{2}k(k-1). \end{aligned}$$

Note the similarity between  $\boldsymbol{\psi}_t$  above and the second-order term of the Kolmogorov–Gabor polynomial (11). Using (109), White (1980) went on to suggest the test statistic

$$\tau_T(\mathbf{y}) = \left( \frac{1}{T} \sum_{t=1}^T (\hat{u}_t^2 - \hat{\sigma}^2) \boldsymbol{\psi}_t \right) \hat{\mathbf{D}}_T^{-1} \left( \frac{1}{T} \sum_{t=1}^T (\hat{u}_t^2 - \hat{\sigma}^2) \boldsymbol{\psi}_t \right), \quad (21.110)$$

where

$$\begin{aligned} \hat{\mathbf{D}}_T &= \frac{1}{T} \sum_{t=1}^T (\hat{u}_t^2 - \hat{\sigma}^2)^2 (\bar{\boldsymbol{\psi}}_t - \bar{\boldsymbol{\psi}}_T) (\bar{\boldsymbol{\psi}}_t - \bar{\boldsymbol{\psi}}_T)', \\ \bar{\boldsymbol{\psi}}_T &= \frac{1}{T} \sum_{t=1}^T \boldsymbol{\psi}_t. \end{aligned} \quad (21.111)$$

Under the assumptions of homoskedasticity  $\tau_T(\mathbf{y}) \sim \chi^2(m)$  and a size  $\alpha$  test can be based on the rejection region

$$C_1 = \{\mathbf{y}: \tau_T(\mathbf{y}) > c_\alpha\}, \quad \text{where } \int_{c_\alpha}^\infty d\chi^2(m) = \alpha. \quad (21.112)$$

Because of the difficulty in deriving the test statistic (109) White went on to suggest an asymptotically equivalent test based on the  $R^2$  of the auxiliary

regression equation

$$\hat{u}_t^2 = \alpha_0 + \alpha_1 \psi_{1t} + \alpha_2 \psi_{2t} + \cdots + \alpha_m \psi_{mt}. \quad (21.113)$$

Under the assumption of homoskedasticity,

$$TR^2 \underset{\chi}{\sim} \chi^2(m), \quad (21.114)$$

and  $TR^2$  could replace  $\tau_T(\mathbf{y})$  in (112) to define an asymptotically equivalent test. It is important to note that the constant in the original regression should not be involved in defining the  $\psi_{it}$ s but the auxiliary regression should have a constant added.

*Example*

For the money equation estimated above the estimated auxiliary equation of the form

$$\hat{u}_t^2 = c_0 + \gamma' \psi_t + v_t$$

yielded  $R^2 = 0.190$ ,  $FT(\mathbf{y}) = 2.8$  and  $TR^2 = 15.2$ . In view of the fact that  $F(6.73) = 2.73$  and  $\chi^2(6) = 12.6$  for  $\alpha = 0.05$  the null hypothesis of homoskedasticity is rejected by both tests.

The most important feature of the above White heteroskedasticity test is that ‘apparently’ no particular form of heteroskedasticity is postulated. In subsection (3) below, however, it is demonstrated that the White test is an exact test in the case where  $D(\mathbf{Z}_t; \boldsymbol{\psi})$  is assumed to be multivariate  $t$ . In this case the conditional mean is  $\mu_t = \boldsymbol{\beta}' \mathbf{x}_t$  but the variance takes the form:

$$\sigma_t^2 = \sigma^2 + \mathbf{x}_t' \mathbf{Q} \mathbf{x}_t. \quad (21.115)$$

Using the ‘omitted variables’ argument for  $u_t^2 = E(u_t^2 | \mathbf{X}_t = \mathbf{x}_t) + v_t$  we can derive the above auxiliary regression (see Spanos (1985b)). This suggests that although the test is likely to have positive power for various forms of heteroskedasticity it will have highest power for alternatives in the multivariate  $t$  direction. That is, multivariate distributions for  $D(\mathbf{Z}_t; \boldsymbol{\psi})$  which are symmetric but have heavier tails than the normal.

In practice it is advisable to use the White test in conjunction with other tests based on particular forms of heteroskedasticity. In particular, tests which allow first and higher-order terms to enter the auxiliary regression, such as the Breusch–Pagan test (see (128) below).

Important examples of heteroskedasticity considered in the econometric literature (see Judge *et al.* (1985), Harvey (1981)) are:

$$(1) \quad \sigma_t^2 = \sigma^2 \boldsymbol{\alpha}' \mathbf{x}_t^*; \quad (21.116)$$

$$(ii) \quad \sigma_t^2 = \sigma^2(\alpha' \mathbf{x}_t^*)^2; \quad (21.117)$$

$$(iii) \quad \sigma_t^2 = \exp(\alpha' \mathbf{x}_t^*); \quad (21.118)$$

where  $\sigma_t^2 = \text{Var}(y_t/\mathbf{X}_t = \mathbf{x}_t)$  and  $\mathbf{x}_t^*$  is an  $m \times 1$  vector which includes known transformations of  $\mathbf{x}_t$  and its first element is the constant term. It must be noted that in the econometric literature these forms of heteroskedasticity are expressed in terms of  $\mathbf{w}_t$  which might include observations from ‘other’ weakly exogenous variables not included in the statistical GM. This form of heteroskedasticity is excluded in the present context because, as argued in Chapter 17, the specification of a statistical model is based on all the observable random variables comprising the sample information. It seems very arbitrary to exclude a subset of such variables from the definition of the systematic component  $E(y_t/\mathbf{X}_t = \mathbf{x}_t)$  and include them only in the conditional variance. In such a case it seems logical to respecify the systematic component as well in order to take this information into consideration. Inappropriate conditioning in defining the systematic component can lead to heteroskedastic errors if the ignored information affects the conditional variance. A very important example of this case is when the sampling model assumption of independence is inappropriate, a non-random sample is the appropriate assumption. In this case the systematic component should be defined in such a way so as to take the temporal dependence among the random variables involved into consideration (see Chapter 22 for an extensive discussion). If, however, the systematic component is defined as  $\mu_t \equiv E(y_t/\mathbf{X}_t = \mathbf{x}_t)$  then this will lead to autocorrelated and heteroskedastic residuals because important temporal information was left out from  $\mu_t$ . A similar problem arises in the case where  $y_t$  and  $\mathbf{X}_t$  are non-stationary stochastic processes (see Chapter 8) with distinct *time trends*. These problems raise the same issues as in the case of non-linearity being detected by heteroskedasticity misspecification tests discussed in the previous section.

Let us consider constructing misspecification tests for the particular forms of heteroskedasticity (i)–(iii). It can be easily verified that (i)–(iii) are special cases of the general form

$$(iv) \quad \sigma_t^2 = h(\alpha' \mathbf{x}_t^*), \quad (21.119)$$

for which we will consider a Lagrange multiplier misspecification test. Breusch and Pagan (1979) argued that the homoskedasticity assumption is equivalent to the hypothesis

$$H_0: \alpha_2 = \alpha_3 = \dots = \alpha_m = 0,$$

given that the first element of  $\mathbf{x}_t^*$  is the constant and  $h(\alpha_1) = \sigma^2$ . The log

likelihood function (retaining normality, see discussion above) is

$$\log L(\boldsymbol{\beta}, \boldsymbol{\alpha}; \mathbf{x}) = \text{const} - \frac{1}{2} \sum_{t=1}^T \log \sigma_t^2 - \frac{1}{2} \sum_{t=1}^T \sigma_t^{-2} (y_t - \boldsymbol{\beta}' \mathbf{x}_t)^2, \quad (21.120)$$

where  $\sigma_t^2 = h(\boldsymbol{\alpha}' \mathbf{x}_t^*)$ . Under  $H_0$ ,  $\sigma_t^2 = \sigma^2$  and the Lagrange multiplier test statistic based on the score takes the general form

$$LM = \left( \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\tilde{\boldsymbol{\theta}}) \mathbf{I}(\tilde{\boldsymbol{\theta}})^{-1} \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\tilde{\boldsymbol{\theta}}) \right), \quad (21.121)$$

where  $\tilde{\boldsymbol{\theta}}$  refers to the constrained MLE of  $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}, \boldsymbol{\alpha})$ . Given that only a subset of the parameters  $\boldsymbol{\theta}$  is constrained the above form reduces to

$$LM = \left( \frac{\partial}{\partial \boldsymbol{\alpha}} \log L(\mathbf{0}, \tilde{\boldsymbol{\alpha}}) \right)' \left( \tilde{\mathbf{I}}_{22} - \tilde{\mathbf{I}}_{21} \tilde{\mathbf{I}}_{11}^{-1} \tilde{\mathbf{I}}_{21} \right)^{-1} \left( \frac{\partial}{\partial \boldsymbol{\alpha}} \log L(\mathbf{0}, \tilde{\boldsymbol{\alpha}}) \right) \quad (21.122)$$

(see Chapter 16). In the above case the score and the information matrix evaluated under  $H_0$  take the forms

$$\frac{\partial}{\partial \boldsymbol{\alpha}} \log L(\mathbf{0}, \tilde{\boldsymbol{\alpha}}) = \frac{1}{2} \left\{ \hat{\sigma}^{-2} \frac{\partial}{\partial \boldsymbol{\alpha}} h(\hat{\alpha}_1) \right\} \sum_{t=1}^T \mathbf{x}_t^* (\hat{\sigma}^{-2} \hat{u}_t^2 - 1), \quad (21.123)$$

$$\tilde{\mathbf{I}}_{22} = \frac{1}{2} \left[ \hat{\sigma}^{-2} \frac{\partial}{\partial \boldsymbol{\alpha}} h(\hat{\alpha}_1) \right]^2 \sum_{t=1}^T \mathbf{x}_t^* \mathbf{x}_t^{*\prime}, \quad (21.124)$$

and  $\tilde{\mathbf{I}}_{21} = 0$ , where  $\hat{\sigma}^2 = (1/T) \sum_{t=1}^T \hat{u}_t^2$  is the MLE of  $\sigma^2$  under  $H_0$ . Hence, the  $LM$  test statistic is

$$LM = \frac{1}{2} \left( \sum_t \mathbf{x}_t^* w_t \right)' \left( \sum_t \mathbf{x}_t^* \mathbf{x}_t^{*\prime} \right)^{-1} \left( \sum_t \mathbf{x}_t^* w_t \right) \stackrel{H_0}{\underset{\alpha}{\sim}} \chi^2(m-1), \quad (21.125)$$

where  $w_t = [(\hat{u}_t^2 / \hat{\sigma}^2) - 1]$ . Given that  $R^2$  in the linear regression model is

$$R^2 = \frac{\mathbf{y}' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y} - T \bar{y}^2}{\mathbf{y}' \mathbf{y} - T \bar{y}^2} \quad (21.126)$$

(see Chapter 19), the  $LM$  test statistic expressed in the form

$$LM^* = \frac{T \left( \sum_t \mathbf{x}_t^* w_t \right)' \left( \sum_t \mathbf{x}_t^* \mathbf{x}_t^{*\prime} \right)^{-1} \left( \sum_t \mathbf{x}_t^* w_t \right)}{\sum_t w_t^2} \quad (21.127)$$

is asymptotically equivalent to  $TR^2$  from the auxiliary regression

$$\frac{\hat{u}_t^2}{\hat{\sigma}^2} = \alpha_1 + \alpha_2 x_{2t}^* + \cdots + \alpha_m x_{mt}^*, \quad (21.128)$$

that is,  $TR^2 \stackrel{H_0}{\sim} \chi^2(m-1)$  (see Breusch and Pagan (1979), Harvey (1981)).

If we apply this test to the estimated money equation with  $\mathbf{x}_t^* \equiv (\mathbf{x}_t, \psi_{2t}, \psi_{3t})$  (see (78) and (79))  $x_{2t}^3, x_{4t}^3$  excluded because of collinearity) the auxiliary regression

$$\frac{\hat{u}_t^2}{\hat{\sigma}^2} = \gamma'_1 \mathbf{x}_t + \gamma'_2 \psi_{2t} + \gamma'_3 \psi_{3t} + v_t \quad (21.129)$$

yielded  $R^2 = 0.250, FT(\mathbf{y}) = 2.055$ . Given that  $TR^2 = 20, \chi^2(11) = 19.675$  and  $F(11, 68) = 1.94$ , the null hypothesis of homoskedasticity is rejected by both test statistics.

### (3) Tackling heteroskedasticity

When the assumption of homoskedasticity is rejected using some misspecification test the question which arises is, 'how do we proceed?' The first thing we should do when residual heteroskedasticity is detected is to diagnose the likeliest source giving rise to it and respecify the statistical model in view of the diagnosis.

In the case where heteroskedasticity is accompanied by non-normality or/and non-linearity the obvious way to proceed is to seek an appropriate normalising, variance-stabilising transformation. The inverse and  $\log_e$  transformations discussed above can be used in such a case after the form of heteroscedasticity has been diagnosed. This is similar to the GLS procedure where  $\boldsymbol{\Lambda}$  is known and the initial variables transformed to

$$\mathbf{y}^* = \mathbf{Hy}, \quad \mathbf{X}^* = \mathbf{HX} \quad \text{for} \quad \mathbf{H}'\mathbf{H} = \boldsymbol{\Lambda}^{-1}.$$

In the case of the estimated money equation considered in Section 21.3 above the normality assumption was not rejected but the linearity and homoskedasticity assumptions were both rejected. In view of the time paths of the observed data involved (see Fig. 17.1) and the residuals (see Fig. 19.3) it seems that the likeliest source of non-linearity and heteroskedasticity might be the inappropriate conditioning which led to dynamic misspecification (see Chapter 22). This 'apparent' non-linearity, heteroskedasticity can be tackled by respecifying the statistical model.

An alternative to the normalising, variance-stabilising transformation is to postulate a non-normal distribution for  $D(y_t, \mathbf{X}_t; \boldsymbol{\theta})$  and proceed to derive  $E(y_t | \mathbf{X}_t = \mathbf{x}_t)$  and  $\text{Var}(y_t | \mathbf{X}_t = \mathbf{x}_t)$  which hopefully provide a more appropriate statistical for the actual DGP being modelled. The results in this direction, however, are very limited, possibly because of the complexity of the approach. In order to illustrate these difficulties let us consider the

case where  $D(y_t, \mathbf{X}_t; \boldsymbol{\theta})$  is multivariate  $t$  with  $n$  degrees of freedom, denoted by

$$\begin{pmatrix} y_t \\ \mathbf{X}_t \end{pmatrix} \sim S_n \left( \begin{pmatrix} 0 \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\sigma}_{11} & \boldsymbol{\sigma}_{12} \\ \boldsymbol{\sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right). \quad (21.131)$$

It turns out that the conditional mean is identical to the case of normality (largely because of the similarity of the shape with the normal) but the conditional variance is heteroskedastic, i.e.

$$E(y_t / \mathbf{X}_t = \mathbf{x}_t) = \boldsymbol{\sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \mathbf{x}_t \quad (21.132)$$

and

$$\text{Var}(y_t / \mathbf{X}_t = \mathbf{x}_t) = \frac{n}{(n+k-2)} (1 + \mathbf{x}_t' \boldsymbol{\Sigma}_{22}^{-1} \mathbf{x}_t) (\boldsymbol{\sigma}_{11} - \boldsymbol{\sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\sigma}_{21}) \quad \text{for } n+k>2 \quad (21.133)$$

(see Zellner (1971)). As we can see, the conditional mean is identical to the one under normality but the conditional variance is heteroskedastic. In particular the conditional variance is a quadratic function of the observed values of  $\mathbf{X}_t$ . In cases where linearity is a valid assumption and some form of heteroskedasticity is present the multivariate  $t$ -assumption seems an obvious choice. Moreover, testing for heteroskedasticity based on

$$H_0: \sigma_t^2 = \sigma^2, \quad t = 1, 2, \dots, T$$

against  $H_1: \sigma_t^2 = (\mathbf{x}'_t \mathbf{Q} \mathbf{x}_t) + \sigma^2, t = 1, 2, \dots, T$ ,  $\mathbf{Q}$  being a  $k \times k$  matrix, will lead directly to a test identical to the White test.

The main problem associated with a multivariate  $t$ -based linear regression model is that in view of (133) the weak exogeneity assumption of  $\mathbf{X}_t$  with respect to  $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}, \sigma^2)$  no longer holds. This is because the parameters  $\psi_1$  and  $\psi_2$  in the decomposition

$$D(y_t, \mathbf{X}_t; \boldsymbol{\psi}) = D(y_t / \mathbf{X}_t; \boldsymbol{\psi}_1) \cdot D(\mathbf{X}_t; \boldsymbol{\psi}_2) \quad (21.134)$$

are no longer variation free (see Chapter 19 and Engle *et al.* (1983) for more details) because  $\boldsymbol{\psi}_1 \equiv (\boldsymbol{\sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1}, \boldsymbol{\sigma}_{11} - \boldsymbol{\sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1}, \boldsymbol{\sigma}_{21}, \boldsymbol{\Sigma}_{22}^{-1})$  and  $\boldsymbol{\psi}_2 \equiv (\boldsymbol{\Sigma}_{22})$  and the constant in the conditional variance depends on the dimensionality of  $\mathbf{X}_t$ . This shows that  $\boldsymbol{\psi}_1$  and  $\boldsymbol{\psi}_2$  are no longer variation free.

The linear regression model based on a multivariate  $t$ -distribution but with homoskedastic conditional variance of the form

$$\text{Var}(y_t / \mathbf{X}_t = \mathbf{x}_t) = \frac{v_0 \sigma^2}{(v_0 - 2)}, \quad v_0 > 2 \quad (21.135)$$

was discussed by Zellner (1976). He showed that in this case  $\hat{\boldsymbol{\beta}}$  and  $\hat{\sigma}^2$  are indeed the MLE's of  $\boldsymbol{\beta}$  and  $\sigma^2$  as in the case of normality.

### 21.5 Parameter time invariance

#### (1) Parameter time dependence

An important assumption underlying the linear regression statistical GM

$$y_t = \beta' \mathbf{x}_t + u_t, \quad t \in \mathbb{T} \quad (21.136)$$

is that the parameters of interest  $\theta \equiv (\beta, \sigma^2)$  are *time invariant*, where  $\beta \equiv \Sigma_{22}^{-1} \sigma_{21}$  and  $\sigma^2 = \sigma_{11} - \sigma_{12} \Sigma_{22}^{-1} \sigma_{21}$ . The time invariance of these parameters is a consequence of the identically distributed component of the assumption

$$\mathbf{Z}_t \sim N(\mathbf{0}, \Sigma), \quad t \in \mathbb{T}, \quad \text{i.e. } \{\mathbf{Z}_t, t \in \mathbb{T}\} \text{ is NIID.} \quad (21.137)$$

This assumption, however, seems rather unrealistic for most economic time-series data. An obvious generalisation is to retain the independence assumption but relax the *identically distributed* restriction. That is, assume that  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is an independent stochastic process (see Chapter 8). This introduces some time-heterogeneity in the process by allowing its parameters to be different at each point in time, i.e.

$$\mathbf{Z}_t \sim N(\mathbf{m}(t), \Sigma(t)), \quad t \in \mathbb{T}, \quad (21.138)$$

where  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  represents a vector stochastic process.

A cursory look at Fig. 17.1 representing the time path of several economic time series for the period 1963*i*–1982*iv* confirms that the assumption (137) is rather unrealistic. The time paths exhibit very distinct time trends which could conceivably be modelled by linear or exponential type trends such as:

$$(i) \quad m_t = \alpha_0 + \alpha_1 t; \quad (21.139)$$

$$(ii) \quad m_t = \exp\{\alpha_0 + \alpha_1 t\}; \quad (21.140)$$

$$(iii) \quad m_t = \alpha_0 + \alpha_1(1 - e^{-t/\tau}), \quad \tau > 0. \quad (21.141)$$

The extension to a general stochastic process where time dependence is also allowed will be considered in Chapters 22 and 23. For the purposes of this chapter independence will be assumed throughout.

In the specification of the linear regression model we argued that (137) is equivalent to  $\mathbf{Z}_t \sim N(\mathbf{m}, \Sigma)$  because we could always define  $\mathbf{Z}_t$  in mean deviation or add a constant term to the statistical GM; the constant is defined by  $\beta_1 = m_1 - \sigma_{12} \Sigma_{22}^{-1} \mathbf{m}_2$  (see Chapter 19). In the case where (138) is valid, however, using mean deviation is not possible because the mean varies with  $t$ . Assuming that

$$\begin{pmatrix} y_t \\ \mathbf{X}_t \end{pmatrix} \sim N\left(\begin{pmatrix} m_1(t) \\ \mathbf{m}_2(t) \end{pmatrix} \begin{pmatrix} \sigma_{11}(t), \sigma_{12}(t) \\ \sigma_{21}(t), \Sigma_{22}(t) \end{pmatrix}\right), \quad (21.142)$$

we can deduce that the conditional mean and variance take the form

$$E(y_t/X_t = \mathbf{x}_t) = \boldsymbol{\beta}'_t \mathbf{x}_t^*, \quad (21.143)$$

$$\text{Var}(y_t/X_t = \mathbf{x}_t) = \sigma_t^2, \quad (21.144)$$

where

$$\boldsymbol{\beta}'_t = (\beta_{1t}, \boldsymbol{\beta}'_{(1)t}), \quad \beta_{1t} = m_1(t) - \boldsymbol{\sigma}_{12}(t)\boldsymbol{\Sigma}_{22}(t)^{-1}\boldsymbol{\sigma}_{21}(t),$$

$$\boldsymbol{\beta}_{(1)t} = \boldsymbol{\Sigma}_{22}(t)^{-1}\boldsymbol{\sigma}_{21}(t), \quad \mathbf{x}_t^* = (1, \mathbf{x}'_t)$$

and

$$\sigma_t^2 = \sigma_{11}(t) - \boldsymbol{\sigma}_{12}(t)\boldsymbol{\Sigma}_{22}(t)^{-1}\boldsymbol{\sigma}_{21}(t).$$

Several comments are in order. Firstly, for notational convenience the star in  $\mathbf{x}_t^*$  will be dropped and the conditional mean written as  $\boldsymbol{\beta}'_t \mathbf{x}_t$ . Secondly, the sequence  $\mathbf{Z}_t$  under (142) defines a non-stationary independent stochastic process (see Chapter 8). Without further restrictions on the time heterogeneity of  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  the parameters of interest  $\boldsymbol{\theta}_t = (\boldsymbol{\beta}_t, \sigma_t^2)$  cannot be estimated because they increase with the sample size  $T$ . This gives us a fair warning that testing for departures from parameter time invariance will not be easy. Thirdly, (142) is only a sufficient condition for (143) and (144), it is not necessary. We could conceive of parametrisations of (142) which could lead to time invariant  $\boldsymbol{\beta}$  and  $\sigma^2$ . Fourthly, it is important to distinguish between time invariance and homoskedasticity of  $\text{Var}(y_t/X_t = \mathbf{x}_t)$ , at least at the theoretical level. Homoskedasticity as a property of the conditional variance refers to the state where it is free of the conditioning variables (see Chapter 7). In the context of the linear regression model homoskedasticity of  $\text{Var}(y_t/X_t = \mathbf{x}_t)$  stems from the normality of  $\mathbf{Z}_t$ . On the other hand, time invariance refers to the time-homogeneity of  $\text{Var}(y_t/X_t = \mathbf{x}_t)$  and follows from the assumption that  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is an identically distributed process. In principle, heteroskedasticity and time dependence need to be distinguished because they arise by relaxing different assumptions relating to the stochastic process  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$ . In practice, however, it will not be easy to discriminate between the two on the basis of a misspecification test for either. Moreover, heteroskedasticity and time dependence can be both present as in the case where (142) is a multivariate  $t$ -distribution (see Section 21.4 above). Finally, the form of  $\beta_{1t}$  above suggests that, in the case of economic time series exemplifying a very distinct trend, even if the variance is constant over time, the coefficient of the constant term will be time dependent, in general. In cases where the non-stationarity is homogeneous (restricted to a local trend, see Section 8.4) and can be 'eliminated' by differencing, its main effect will be on  $\boldsymbol{\beta}_{1t}$ , leaving  $\boldsymbol{\beta}_{(1)t}$  'largely' time-invariant. This might explain why in regressions with time series data the coefficient of the constant seems highly volatile although the other coefficients appear to be relatively constant.

## (2) Testing for parameter time dependence

Assuming that  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is a non-stationary, independent normal process, and defining the systematic and non-systematic components by

$$\mu_t = E(y_t | \mathbf{X}_t = \mathbf{x}_t) \quad \text{and} \quad u_t = y_t - E(y_t | \mathbf{X}_t = \mathbf{x}_t), \quad (21.145)$$

the implied statistical GM takes the form

$$y_t = \boldsymbol{\beta}'_t \mathbf{x}_t + u_t, \quad t \in \mathbb{T}, \quad (21.146)$$

with  $\boldsymbol{\theta}_t \equiv (\boldsymbol{\beta}_t, \sigma_t^2)$  being the statistical parameters of interest. If we compare (146) with (136) we can see that the null hypothesis for parameter time invariance for a sample of size  $T$  is

$$H_0: \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_T = \boldsymbol{\beta} \quad \text{and} \quad \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_T^2 = \sigma^2$$

against

$$H_1: \boldsymbol{\beta}_t \neq \boldsymbol{\beta} \quad \text{or} \quad \sigma_t^2 \neq \sigma^2 \quad \text{for any } t = 1, 2, \dots, T.$$

Given that the number of parameters to be estimated is  $T(k+1) + T$  and we only have  $T$  observations it is obvious that  $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_T$  are not estimable. It is instructive, however, to ignore this and go ahead to attempt estimation of these parameters by maximum likelihood.

Differentiation of the log likelihood function:

$$\log L = \text{const} - \frac{1}{2} \sum_{t=1}^T \log \sigma_t^2 - \frac{1}{2} \sum_{t=1}^T \sigma_t^{-2} (y_t - \boldsymbol{\beta}'_t \mathbf{x}_t)^2 \quad (21.147)$$

yields the following first-order conditions:

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}_t} = \sigma_t^{-2} (y_t - \boldsymbol{\beta}'_t \mathbf{x}_t) \mathbf{x}'_t = \mathbf{0}, \quad \frac{\partial \log L}{\partial \sigma_t^2} = -\frac{1}{2\sigma_t^2} + \frac{1}{2\sigma_t^4} u_t^2 = 0 \quad (21.148)$$

These equations cannot be solved for  $\boldsymbol{\beta}_t$  and  $\sigma_t^2$  because  $\text{rank}(\mathbf{x}_t) = \text{rank}(\mathbf{x}_t, \mathbf{x}'_t) = 1$ , which suggests that  $\mathbf{x}_t \mathbf{x}'_t$  cannot be inverted; no MLE of  $\boldsymbol{\theta}_t$  exists. Knowing the ‘source’ of the problem, however, might give us ideas on how we might ‘solve’ it. In this case intuition suggests that a possible invertible form of  $\mathbf{x}_t \mathbf{x}'_t$  is  $(\sum_{t=1}^k \mathbf{x}_t \mathbf{x}'_t) = (\mathbf{X}_k^0 \mathbf{X}_k^0)$ , where  $\mathbf{X}_k^0 \equiv (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k)$ . That is, use the observations  $t = 1, 2, \dots, k$ , in order to get  $\text{rank}(\mathbf{X}_k^0) = k$  and invert it to estimate  $\boldsymbol{\beta}_k$  via

$$\hat{\boldsymbol{\beta}}_k = (\mathbf{X}_k^0 \mathbf{X}_k^0)^{-1} \mathbf{X}_k^0 \mathbf{y}_k^0 = (\mathbf{X}_k^0)^{-1} \mathbf{y}_k^0, \quad (21.149)$$

$\mathbf{y}_k^0 \equiv (y_1, \dots, y_k)$ . Moreover, for  $t = k+1, k+2, \dots, T$ , the corresponding  $\boldsymbol{\beta}_t$ s could conceivably be estimated by

$$\hat{\boldsymbol{\beta}}_t = (\mathbf{X}_t^0 \mathbf{X}_t^0)^{-1} \mathbf{X}_t^0 \mathbf{y}_t^0, \quad t = k+1, \dots, T. \quad (21.150)$$

In turn the residuals  $\hat{u}_t = (y_t - \hat{\boldsymbol{\beta}}_t' \mathbf{x}_t)$ ,  $t = k+1, \dots, T$ , can be derived which,

however, cannot be used to estimate  $\sigma_t^2$  because the estimator implied by the above first-order conditions is

$$\hat{\sigma}_t^2 = \hat{u}_t^2. \quad (21.151)$$

This is clearly unsatisfactory given that we only have one observation for each  $\sigma_t^2$  and the  $\hat{u}_t$ s are not even independent. An alternative form of residuals which are at least independent are the *recursive residuals* (see Section 19.7). These constitute the one-step-ahead prediction errors

$$\tilde{v}_t = (y_t - \hat{\beta}_{t-1}' \mathbf{x}_t) = u_t + \mathbf{x}_t' (\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_{t-1}), \quad t = k+1, \dots, T, \quad (21.152)$$

and they can be used to update the recursive estimators of  $\boldsymbol{\beta}_t$ s as each new observation becomes available using the relationship

$$\hat{\boldsymbol{\beta}}_t = \hat{\boldsymbol{\beta}}_{t-1} + (\mathbf{X}_{t-1}^{0'} \mathbf{X}_{t-1}^0)^{-1} \mathbf{x}_t \left( \frac{\tilde{v}_t}{d_t^2} \right), \quad t = k+1, \dots, T, \quad (21.153)$$

(see exercises 5 and 6) where

$$d_t = (1 + \mathbf{x}_t' (\mathbf{X}_{t-1}^{0'} \mathbf{X}_{t-1}^0)^{-1} \mathbf{x}_t)^{\frac{1}{2}}. \quad (21.154)$$

As we can see from (153), the new information at time  $t$  comes in the form of  $\tilde{v}_t$  and  $\boldsymbol{\beta}_t$  is estimated by updating  $\hat{\boldsymbol{\beta}}_{t-1}$ .

Substituting  $\hat{\boldsymbol{\beta}}_{t-1}$  in (152) yields

$$\begin{aligned} \tilde{v}_t &= u_t + \mathbf{x}' \left[ \boldsymbol{\beta}_t - (\mathbf{X}_{t-1}^{0'} \mathbf{X}_{t-1}^0)^{-1} \sum_{i=1}^{t-1} \mathbf{x}_i \mathbf{x}_i' \boldsymbol{\beta}_i \right] - \mathbf{x}_t' (\mathbf{X}_{t-1}^{0'} \mathbf{X}_{t-1}^0)^{-1} \sum_{i=1}^{t-1} \mathbf{x}_i u_i \\ &\quad (21.155) \end{aligned}$$

(see exercise 7). Hence, under  $H_0$ ,  $E(\tilde{v}_t) = 0$ ,  $E(\tilde{v}_t^2) = \sigma^2 d_t^2$ ,  $t = k+1, \dots, T$ . This implies that the standardised recursive residuals

$$w_t = \frac{\tilde{v}_t}{d_t}, \quad t = k+1, \dots, T \quad (21.156)$$

suggest themselves as the natural candidates on which a test for  $H_0$  might be devised. Indeed, for  $\mathbf{w} \equiv (w_{k+1}, w_{k+2}, \dots, w_T)'$ ,

$$\mathbf{w} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{T-k}) \quad (21.157)$$

and

$$\mathbf{w} \sim N(\boldsymbol{\delta}, \mathbf{C}), \quad (21.158)$$

where

$$\boldsymbol{\delta} \equiv (\delta_{k+1}, \delta_{k+2}, \dots, \delta_T), \quad \mathbf{C} = [C_{ts}], \quad t, s = k+1, \dots, T,$$

$$\delta_t = \frac{1}{d_t} \left[ \mathbf{x}_t' \left( \boldsymbol{\beta}_t - (\mathbf{X}_{t-1}^{0'} \mathbf{X}_{t-1}^0)^{-1} \sum_{i=1}^{t-1} \mathbf{x}_i \mathbf{x}_i' \boldsymbol{\beta}_i \right) \right], \quad t = k+1, \dots, T, \quad (21.159)$$

$$c_n = \frac{1}{d_t^2} \left[ \sigma_t^2 + \mathbf{x}_t' (\mathbf{X}_{t-1}^{0'} \mathbf{X}_{t-1}^0)^{-1} \left\{ \sum_{i=1}^{t-1} \mathbf{x}_i \mathbf{x}_i' \sigma_i^2 \right\} (\mathbf{X}_{t-1}^{0'} \mathbf{X}_{t-1}^0)^{-1} \mathbf{x}_t \right],$$

$$t = k+1, \dots, T, \quad (21.160)$$

$$c_{ts} = \frac{1}{d_t d_s} \left[ \mathbf{x}_t' (\mathbf{X}_{t-1}^{0'} \mathbf{X}_{t-1}^0)^{-1} \left\{ \sum_{i=1}^{t-1} \mathbf{x}_i \mathbf{x}_i' \sigma_i^2 \right\} \right.$$

$$\left. \times (\mathbf{X}_{s-1}^{0'} \mathbf{X}_{s-1}^0)^{-1} \mathbf{x}_s - \mathbf{x}_s' (\mathbf{X}_{s-1}^{0'} \mathbf{X}_{s-1}^0) \mathbf{x}_t \sigma_t^2 \right] \quad (21.161)$$

for  $t < s$ ,  $t = k+1, \dots, T$  (see exercise 8).

If we separate  $H_0$  into

$$H_0^{(1)}: \boldsymbol{\beta}_t = \boldsymbol{\beta}, \quad \text{for all } t = 1, 2, \dots, T,$$

$$H_0^{(2)}: \sigma_t^2 = \sigma^2, \quad \text{for all } t = 1, 2, \dots, T,$$

we can see that

$$\mathbf{w} \stackrel{H_0^{(1)}}{\sim} N(\mathbf{0}, \mathbf{C}), \quad (21.162)$$

but

$$\mathbf{w} \stackrel{H_0^{(2)}}{\sim} N(\boldsymbol{\delta}, \sigma^2 \mathbf{I}_{T-k}). \quad (21.163)$$

This shows that coefficient time dependence only affects the mean of  $\mathbf{w}$  and variance time dependence affects its covariance. The implication from these results is that we could construct a test for  $H_0^{(1)}$  given that  $H_0^{(2)}$  holds against  $H_1^{(1)}: \boldsymbol{\beta}_t \neq \boldsymbol{\beta}$  for any  $t = 1, 2, \dots, T$ , based on the chi-square distribution. In view of (163) we can deduce that

$$\left( \sum_{t=k+1}^T w_t \right) \stackrel{H_0^{(2)}}{\sim} N \left( \sum_{t=k+1}^T \delta_t, \sigma^2 \right). \quad (21.164)$$

This result implies that testing for  $H_0^{(1)}$ , given  $H_0^{(2)}$  is valid, is equivalent to testing for  $E(w_t) = 0$  against  $E(w_t) \neq 0$ . Before we can use (164) as the basis of a test statistic we need to estimate  $\sigma^2$ . A natural estimator for  $\sigma^2$  is

$$s_w^2 = \frac{1}{T-k-1} \sum_{t=k+1}^T (w_t - \bar{w})^2, \quad (21.165)$$

where  $\bar{w} = [1/(T-k)] \sum_{t=k+1}^T w_t$ . Note that  $\bar{w} \neq 0$  when  $H_0^{(1)}$  is not valid. This enables us to construct the test statistic

$$\tau_1(\mathbf{y}) = (T-k)^{\frac{1}{2}} \left( \frac{\bar{w}}{s_w} \right)^{H_0} \sim t(T-k-1). \quad (21.166)$$

Using this we can construct a size  $\alpha$  test based on the rejection region

$$C_1 = \{\mathbf{y}: |\tau_1(\mathbf{y})| \geq c_\alpha\}, \quad 1 - \alpha = \int_{-\infty}^{c_\alpha} dt (T - k - 1) \quad (21.167)$$

(see Harvey (1981)). Under  $H_0$  the above test based on (166) and (167) is UMP unbiased (see Lehmann (1959)). On the other hand, when  $H_0^{(2)}$  does not hold  $E(s_w^2) > \sigma^2$  and this can reduce the power of the test significantly (see Dufour (1982)).

Another test related to  $H_0^{(1)}$  conditional on  $H_0^{(2)}$  being valid was suggested by Brown, Durbin and Evans (1975). The CUSUM-test is based on the test statistic

$$W_t = \sum_{i=k+1}^t \frac{w_i}{s}, \quad t = k + 1, \dots, T, \quad (21.168)$$

where  $s^2 = [1/(T-k)] \sum_{t=1}^T \hat{u}_t^2$ . They showed that under  $H_0$  the distribution of  $W_t$  can be approximated by  $N(0, t-k)$  ( $W_t$  being an approximate Brownian motion). This led to the rejection region

$$C_1 = \{\mathbf{y}: |W_t| > c_\alpha\}, \quad c_\alpha = a(T-k)^{\frac{1}{2}} + 2a(t-k)(T-k)^{-\frac{1}{2}}, \quad (21.169)$$

with  $a$  depending on the size  $\alpha$  of the test. For  $\alpha = 0.01, 0.05, 0.10, a = 1.143, 0.948, 0.850$ , respectively. The underlying intuition of this test is that if  $H_0$  is invalid there will be some systematic changes in the  $\beta_i$ 's which will give rise to a disproportionate number of  $w_i$ 's having the same sign. Hopefully, these will be detected via their cumulative effects  $W_t$ ,  $t = k + 1, \dots, T$ .

Brown *et al.* (1975) suggested a second test based on the test statistic

$$V_t = \left( \frac{\sum_{i=k+1}^t w_i^2}{\sum_{i=k+1}^T w_i^2} \right), \quad t = k + 1, \dots, T. \quad (21.170)$$

Under  $H_0$ :  $V_t$ , known as the CUSUMSQ-statistic, has a beta distribution with parameters  $\frac{1}{2}(T-t), \frac{1}{2}(t-k)$ , i.e.

$$V_t \stackrel{H_0}{\sim} B(\frac{1}{2}(T-t), \frac{1}{2}(t-k)). \quad (21.171)$$

In view of the relationship between the beta and  $F$ -distributions (see Johnson and Kotz (1970)) we can deduce that

$$V_t^* = \frac{(t-k)}{(T-t)} \left( \frac{V_t}{1-V_t} \right)^{\frac{1}{2}} \sim F((T-t), (t-k)). \quad (21.172)$$

This enables us to use the  $F$ -test rejection region whose tabulated values are more commonly available.

Looking at the three test statistics (166), (168) and (170) we can see that one way to construct a test for  $H_0^{(2)}$  is to compare the prediction error squared ( $w_t^2$ ) with the average over the previous periods, i.e. use the test statistics

$$\tau^2(\mathbf{y}_t^0) = \frac{w_t^2}{\frac{1}{(t-k)} \sum_{i=k}^{t-1} w_i^2}, \quad t = k+1, \dots, T. \quad (21.173)$$

The intuition underlying (173) is that the denominator can be viewed as the natural estimator of  $\sigma_{t-1}^2$  which is compared with the new information at time  $t$ . Note that

$$\left( \frac{w_t^2}{\sigma^2} \right) \stackrel{H_0}{\sim} \chi^2(1) \quad \text{and} \quad \left( \frac{1}{\sigma^2} \sum_{i=1}^{t-1} w_i^2 \right) \stackrel{H_0}{\sim} \chi^2(t-k), \quad (21.174)$$

and the two random variables are independent. These imply that under  $H_0$ ,

$$\tau^2(\mathbf{y}_t^0) \sim F(1, t-k) \quad \text{or} \quad \tau(\mathbf{y}_t^0) \sim t(t-k), \quad t = k+1, \dots, T. \quad (21.175)$$

It must be noted that  $\tau(\mathbf{y}_t^0)$  provides a test statistic for  $\beta_t = \beta_{t-1}$  assuming that  $\sigma_t^2 = \sigma_{t-1}^2$ ; see Section 21.6 for some additional comments. For an overall test of  $H_0$  we should use a multiple comparison procedure based on the Bonferroni and related inequalities (see Savin (1984)).

One important point related to all the above tests based on the standardised recursive residuals is that the *implicit null* hypothesis is not quite  $H_0$  but a closely related hypothesis. If we return to equation (156) we can see that

$$E(w_t) = 0 \quad \text{if } \mathbf{x}'_t(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1}) = 0, \quad (21.176)$$

which is not the same as  $(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1}) = \mathbf{0}$ .

In practice, the above tests should be used in conjunction with the *time paths* of the recursive estimators  $\hat{\boldsymbol{\beta}}_{it}$ ,  $i = 1, 2, \dots, k$ , and the standardised recursive residuals  $w_t$ ,  $t = k+1, \dots, T$ . If we ignore the first few values of these series their time paths can give us a lot of information relating to the time invariance of the parameters of interest.

In the case of the estimated money equation discussed above, the time paths of  $\beta_{1t}$ ,  $\beta_{2t}$ ,  $\beta_{3t}$ ,  $\beta_{4t}$  are shown in Fig. 21.1(a)–(d) for the period  $t = 20, \dots, 80$ . As we can see, the graphs suggest the presence of some time dependence in the estimated coefficients re-enforcing the variance time dependence detected in Section 21.4 using the heteroskedasticity tests.

(3) **Tackling parameter time dependence**

When parameter time invariance is rejected by some misspecification test the question which naturally arises is, 'how do we proceed?' The answer to this question depends crucially on the likely source of time dependence. If time dependence is due to the behaviour of the agents behind the actual DGP we should try to model this behaviour in such a way so as to take this additional information. *Random coefficient models* (see Pagan (1980)) or *state space models* (see Anderson and Moore (1979)) can be used in such cases. If, however, time dependence is due to inappropriate conditioning or  $Z_t$  is a non-stationary stochastic process then the way to proceed is to respecify the systematic component or transform the original time series in order to induce stationarity.

In the case of the estimated money equation considered above it is highly likely that the coefficient time dependency exemplified is due to the non-stationarity of the data involved as their time paths (see Fig. 17.1) indicate. One way to tackle the problem in this case is to transform the stochastic processes involved so as to induce some stationarity. For example, if  $\{M_t, t \geq 1\}$  shows an exponential-like time path, transforming it to  $\{\Delta \ln(M/P)_t, t \geq 1\}$  can reduce it to a near stationary process. The time path of  $\Delta \ln(M/P)_t = \ln(M/P)_t - \ln(M/P)_{t-1}$  as shown in Fig. 21.2, suggests that this transformation has induced near stationarity to the original time series. It is interesting to note that if  $\Delta \ln(M/P)_t$  is stationary then

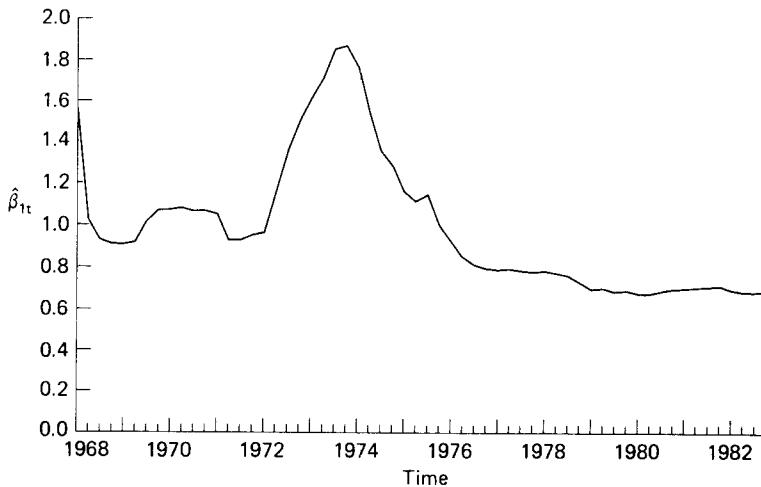
$$\Delta^2 \ln\left(\frac{M}{P}\right)_t = \ln\left(\frac{M}{P}\right)_t - 2 \ln\left(\frac{M}{P}\right)_{t-1} + \ln\left(\frac{M}{P}\right)_{t-2} \quad (21.177)$$

is also stationary (see Fig. 21.3); any linear combination of a stationary process is stationary. Caution, however, should be exercised in using stationarity inducing transformation, because *overdifferencing*, for example, can increase the variance of the process unnecessarily (see Tintner (1940)). In the present example this is indeed the case given that the variance of  $\Delta^2 \ln(M/P)_t$  is more than twice the variance of  $\Delta \ln(M/P)_t$ ,

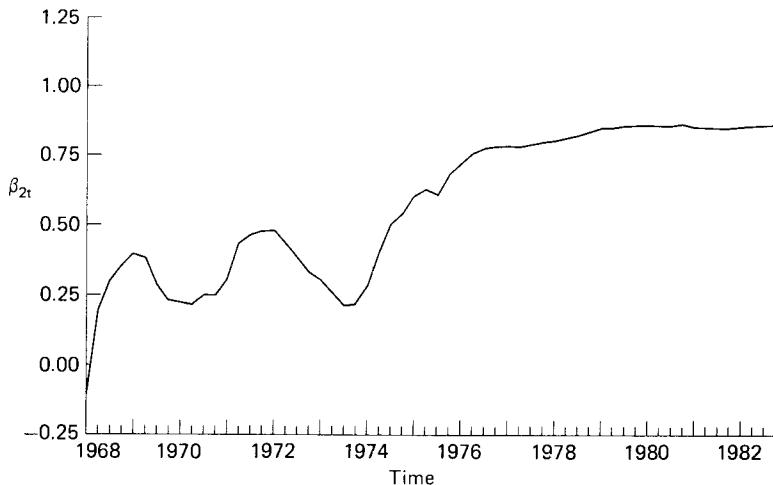
$$\text{Var}\left(\Delta \ln\left(\frac{M}{P}\right)_t\right) = 0.000574, \quad \text{Var}\left(\Delta^2 \ln\left(\frac{M}{P}\right)_t\right) = 0.001354. \quad (21.178)$$

In econometric modelling differencing to achieve near stationarity should not be used at the expense of theoretical interpretation. It is often possible to 'model' the non-stationarity using appropriate explanatory variables.

Note that it is the stationarity of  $\{y_t/X_t, t \in \mathbb{T}\}$  which is important, not that of  $\{Z_t, t \in \mathbb{T}\}$ .



(a)



(b)

Fig. 21.1(a). The time path of the recursive estimate of  $\beta_{1t}$  – the coefficient of the constant. (b) The time path of the recursive estimate of  $\beta_{2t}$  – the coefficient of  $y_t$ .

In view of our initial discussion related to the possible inappropriateness of the sampling model assumption of independence (see Chapter 19) we can argue that the above detected parameter time dependence might also be due to invalid conditioning (see Chapters 22–23 for a detailed discussion). In

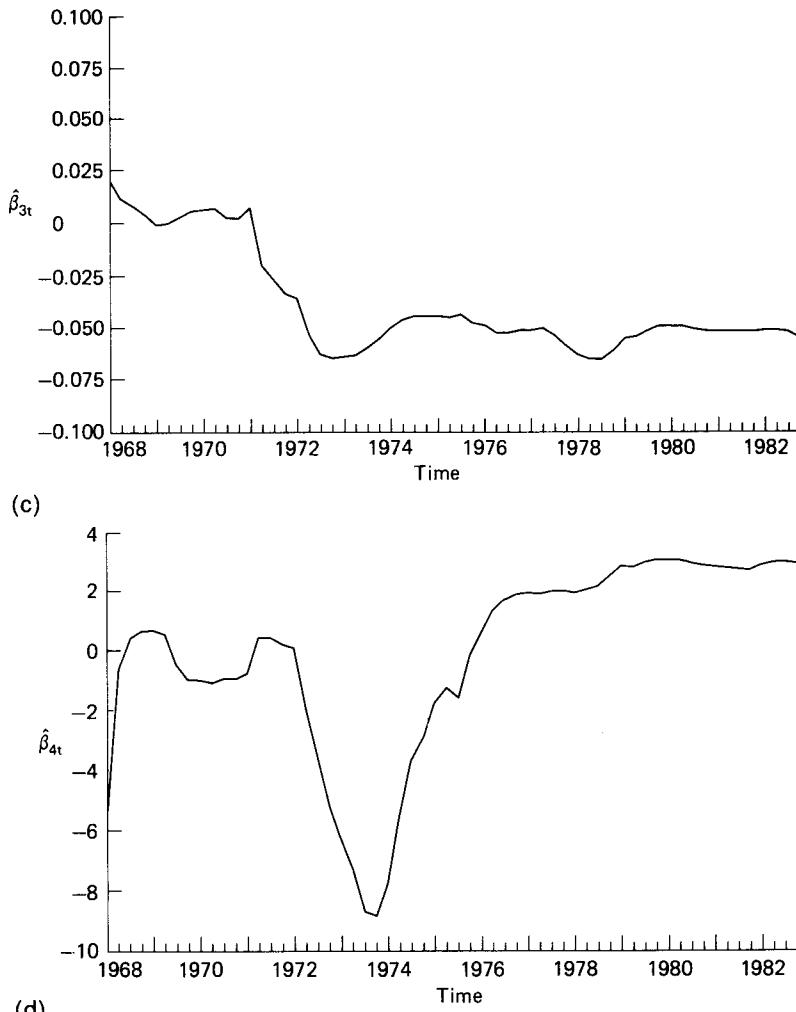
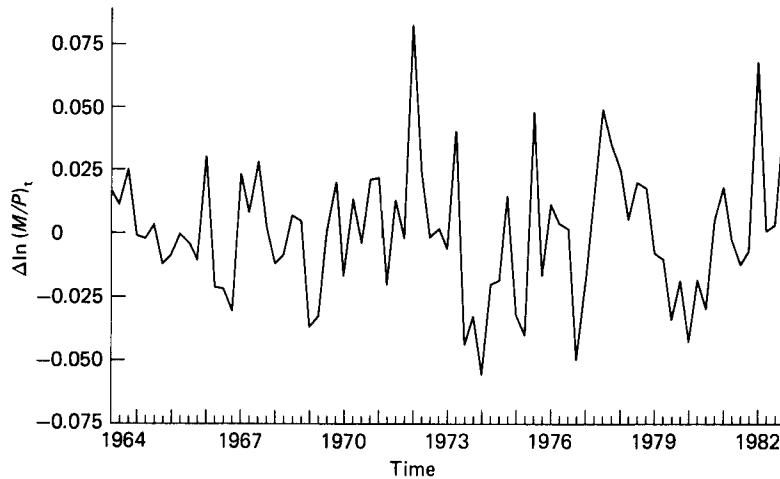
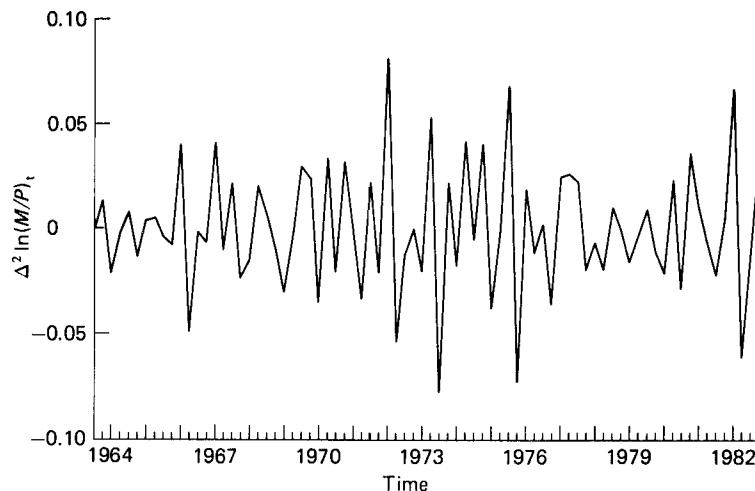


Fig. 21.1(c). The time path of the recursive estimate of  $\beta_{3t}$  – the coefficient of  $p_t$ . (d) The time path of the recursive estimate of  $\beta_{4t}$  – the coefficient of  $i_t$ .

in this case the way to tackle time dependence is to respecify  $\mu_t$  in order to take account of the additional information present.

## 21.6 Parameter structural change

Parameter structural change is interpreted as a special case of time dependence, considered in Section 21.5, where some a priori information

Fig. 21.2. The time path of  $\Delta \ln (M/P)_t$ .Fig. 21.3. The time path of  $\Delta^2 \ln (M/P)_t$ .

related to the point of change is available. For example, in the case of the money equation estimated in Chapter 19 and discussed in the previous sections we know that some change in monetary policy has occurred in 1971 which might have induced a structural change.

The statistical GM for the case where only one structural change has occurred at  $t = T_1$  ( $T_1 > k$ ), takes the general form

$$y_t = \beta'_1 \mathbf{x}_t + u_{1t}, \quad t \in \mathbb{T}_1 \quad (21.179)$$

$$y_t = \beta'_2 \mathbf{x}_t + u_{2t}, \quad t \in \mathbb{T}_2 \quad (21.180)$$

where  $\mathbb{T}_1 = \{1, 2, \dots, T_1\}$ ,  $\mathbb{T}_2 = \{T_1 + 1, \dots, T\}$  with  $\theta_1 \equiv (\beta_1, \sigma_1^2)$  and  $\theta_2 \equiv (\beta_2, \sigma_2^2)$  being the underlying parameters of interest respectively. For the case where the sample period is  $t = 1, 2, \dots, T_1, T_1 + 1, \dots, T$ , the distribution of the sample takes the form

$$\begin{pmatrix} \mathbf{y}_1 / \mathbf{X}_1 \\ \mathbf{y}_2 / \mathbf{X}_2 \end{pmatrix} \sim N \left( \begin{pmatrix} \mathbf{X}_1 \beta_1 \\ \mathbf{X}_2 \beta_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 \mathbf{I}_{T_1} & \mathbf{0} \\ \mathbf{0} & \sigma_2^2 \mathbf{I}_{T_2} \end{pmatrix} \right), \quad (21.181)$$

where  $T_2 \equiv T - T_1$ . The hypothesis of interest in this case is

$$H_0: \beta_1 = \beta_2 \quad \text{and} \quad \sigma_1^2 = \sigma_2^2,$$

which can be separated into

$$H_0^{(1)}: \beta_1 = \beta_2, \quad H_0^{(2)}: \sigma_1^2 = \sigma_2^2; \quad (H_0 = H_0^{(1)} \cap H_0^{(2)}).$$

The alternative  $H_1$  is specified as in (181). These hypotheses are very similar to the ones we considered in the previous section and it should come as no surprise to learn that constructing optimal tests for the present case raises similar problems. Moreover, the test statistics enjoy more than just a passing resemblance with some of the statistics derived in Section 21.5. The main advantage in the present case is that the point of structural change  $T_1$  is assumed known a priori. This, however, raises the issue of whether  $T_2 > k$  ( $\theta_2$  is estimable) or  $T_2 < k$  ( $\theta_2$  is not estimable). Let us consider the two cases separately.

### Case 1 ( $T_2 > k$ )

Chow (1960) proposed the  $F$ -type (see Chapter 19) test statistic

$$\tau_2(\mathbf{y}) = \left( \frac{RSS_T - RSS_1 - RSS_2}{RSS_1 + RSS_2} \right) \left( \frac{T - 2k}{k} \right), \quad (21.182)$$

where  $RSS_T$ ,  $RSS_1$  and  $RSS_2$  refer to the residual sums of squares for the whole sample, subperiod 1 ( $t = 1, 2, \dots, T_1$ ) and subperiod 2 ( $t = T_1 + 1, \dots, T$ ) respectively. The test statistic (182) can be used to construct a UMP invariant test (see Lehmann (1959)) for  $H_0^{(1)}$  against  $H_1 \cap H_0^{(2)}$ . That is, it can be used to construct an 'optimal' test for  $\beta_1 = \beta_2$  against  $H_1$  with  $\sigma_1^2 = \sigma_2^2$ . The test statistic is distributed as follows:

$$\tau_2(\mathbf{y}) \sim F(k, T - 2k) \quad \text{under } H_0^{(1)}, \quad (21.183)$$

$$\tau_2(\mathbf{y}) \sim F(k, T - 2k; \delta) \quad \text{under } H_1 \cap H_0^{(2)}, \quad (21.184)$$

where

$$\delta = (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)' \frac{[(\mathbf{X}'_1 \mathbf{X}_1)^{-1} + (\mathbf{X}'_2 \mathbf{X}_2)^{-1}]^{-1}}{\sigma^2} (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2). \quad (21.185)$$

The distribution of  $\tau_2(\mathbf{y})$  under  $H_0^{(1)}$  can be used to define a size  $\alpha$  test whose rejection region is

$$C_1 = \{\mathbf{y}: \tau_2(\mathbf{y}) > c_\alpha\}, \quad \alpha = \int_{c_\alpha}^\infty dF(k, T-2k). \quad (21.186)$$

This test raises the same issues as the time invariance tests considered in Section 21.6 where  $\sigma_1^2 = \sigma_2^2$  had to be explicitly (or implicitly) assumed in constructing a coefficient time invariant test. There is no reason to believe, however, that  $H_0^{(2)}$  is valid when  $H_0^{(1)}$  might not be. A test for  $H_0^{(2)}$  against  $H_1$  can be based on the test statistic comparing the two estimated variances (see Chapter 19):

$$\tau_3(\mathbf{y}) = \frac{s_2^2}{s_1^2} \equiv \left( \frac{RSS_2}{RSS_1} \right) \left( \frac{T_1 - k}{T_2 - k} \right), \quad (21.187)$$

$$\tau_3(\mathbf{y}) \sim F(T_2 - k, T_1 - k) \quad \text{under } H_0^{(2)}, \quad (21.188)$$

$$\tau_3(\mathbf{y}) \sim F(T_2 - k, T_1 - k; \delta) \quad \text{under } H_1, \quad (21.189)$$

where  $\delta = (\sigma_2^2 / \sigma_1^2)$  is the non-centrality parameter which ‘fortunately’ does not depend on  $\boldsymbol{\beta}_1$  or  $\boldsymbol{\beta}_2$ . This turns out to be critical because the test for  $H_0^{(1)}$  against  $H_1$  defined by the rejection region,

$$C_1 \{\mathbf{y}: \tau_3(\mathbf{y}) > c_\alpha\}, \quad \alpha = \int_{c_\alpha}^\infty dF(T_1 - k, T_2 - k), \quad (21.190)$$

is *independent* of the test defined by  $\tau_2(\mathbf{y})$  (see Phillips and McCabe (1983)). This implies that a test for  $H_0$  against  $H_1$  can be implemented by testing for  $H_0^{(2)}$  first using  $\tau_3(\mathbf{y})$  and, if accepted, testing for  $H_0^{(1)}$  using  $\tau_2(\mathbf{y})$  in that sequence.

Let us apply the above tests to the money equation considered in the previous sections. As mentioned above, a structural change is suspected to have occurred in 1971 because of important changes in monetary policy. Estimation of the money equation for the period 1963*i*–1971*iv* yielded

$$m_t = -0.793 + 1.050y_t + 0.305p_t - 0.007i_t + \hat{u}_t, \quad (21.191)$$

(2.055)	(0.208)	(0.103)	(0.014)	(0.015)
---------	---------	---------	---------	---------

$$R^2 = 0.968, \quad \bar{R}^2 = 0.964, \quad s_1 = 0.0155,$$

$$\log L = 90.08, \quad RSS_1 = 0.00672, \quad T = 32.$$

Estimation of the same equation for the period 1972*i*–1982*iv* yielded

$$m_t = -6.397 + 1.641y_t + 0.784p_t - 0.076i_t + \hat{u}_t, \quad (21.192)$$

(1.875) (0.191) (0.022) (0.014) (0.035)

$$R^2 = 0.994, \quad \bar{R}^2 = 0.993, \quad s_2 = 0.0346,$$

$$\log L = 95.37, \quad RSS_2 = 0.05284, \quad T = 48.$$

Testing for  $H_0^{(2)}$  against  $H_1$  the test statistic  $\tau_3(y)$  is

$$\tau_3(y) = \frac{(0.05284)}{(0.00672)} \left( \frac{28}{44} \right) = 5.004. \quad (21.193)$$

For a size  $\alpha = 0.05$ ,  $c_\alpha = 1.81$  and  $H_0^{(2)}$  is rejected. At this stage there is not much point in proceeding with testing  $H_0^{(1)}$  against  $H_1 \cap H_0^{(2)}$  given that  $H_0^{(2)}$  has been rejected. For illustration purposes, however, let us consider the test regardless.

The residual sum of squares for the whole period is  $RSS = 0.11752$  (see Chapter 19). Hence, testing for  $H_0^{(1)}$  against  $H_1 \cap H_0^{(2)}$  the test statistic is

$$\tau_2(y) = \frac{(0.11752) - (0.00672) - (0.05284)}{(0.00672) + (0.05284)} \left( \frac{72}{4} \right) = 17.516. \quad (21.194)$$

Given that  $c_\alpha = 2.5$  for a size  $\alpha = 0.05$  test we can deduce that  $H_0^{(1)}$  is also strongly rejected.

It is important to note that in the case of the test for  $H_0$  the size is not  $\alpha$  but  $1 - (1 - \alpha)^2$ , i.e. for the above example the overall size is 0.0975. This is because  $H_0$  was tested as two independent hypotheses in a multiple hypothesis testing context (see Savin (1984)).

### **Case 2 ( $T_2 < k$ )**

In this case  $\theta_2$  is not estimable since  $\text{rank}(\mathbf{X}'_2 \mathbf{X}_2) \leq T_2 < k$ . This is very similar to the time invariance test where  $T_2 = 1$ . If we were to express the test statistic

$$\tau(y_t^0) = \frac{w_t^2}{\left( \frac{1}{t-k} \sum_{i=k}^{t-1} w_i^2 \right)} \quad (21.195)$$

(see (173)) in terms of the residual sum of squares with  $t \in \mathbb{T}_2$  the following test statistic emerges:

$$\begin{aligned} CH &= (\mathbf{y}_2 - \mathbf{X}_2 \hat{\boldsymbol{\beta}}_1)' \frac{[\mathbf{I} + \mathbf{X}_2(\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_2]^{-1}}{\hat{\mathbf{u}}'_1 \hat{\mathbf{u}}_1} (\mathbf{y}_2 - \mathbf{X}_2 \hat{\boldsymbol{\beta}}_1) \left( \frac{T_1 - k}{T_2} \right) \\ &\equiv \left( \frac{RSS_T - RSS_1}{RSS_1} \right) \left( \frac{T_1 - k}{T_2} \right). \end{aligned} \quad (21.196)$$

This is based on the equalities  $RSS_t = RSS_{t-1} + w_t^2$  and  $\sum_{i=k}^{t-1} w_i^2 = RSS_{t-1}$  (see exercise 9). The test statistic  $CH$ , known as *Chow test* (see Chow (1960)), as in the case of  $\tau(\mathbf{y}_t^0)$  in Section 21.5, can be used to construct a UMP invariant test not of  $H_0$  against  $H_1$  but of  $H_0^*$  against  $H_1 \cap H_0^{(2)}$  where  $H_0^* : \mathbf{X}_2(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)^0 = \mathbf{0}$ . This is not surprising given that  $\boldsymbol{\beta}_2$  is not estimable and thus we need to rely on  $(\mathbf{y}_2 - \mathbf{X}_2 \hat{\boldsymbol{\beta}}_1) = \mathbf{u}_2 - \mathbf{X}_2(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_2)$  (a direct analogue to the recursive residuals  $\tilde{v}_t$  of Section 21.5) in order to construct a test for  $H_0^{(1)}$ . The  $CH$ -test statistic is distributed as follows:

$$CH \sim F(T_2, T_1 - k) \quad \text{under } H_0^* \cap H_0^{(2)}, \quad (21.197)$$

$$CH \sim F(T_2, T_1 - k; \delta) \quad \text{under } H_1 \cap H_0^{(2)}, \quad (21.198)$$

where

$$\delta = (\mathbf{y}_2 - \mathbf{X}_2 \boldsymbol{\beta}_1)' \frac{[\mathbf{I} + \mathbf{X}_2(\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_2]^{-1}}{\sigma^2} (\mathbf{y}_2 - \mathbf{X}_2 \boldsymbol{\beta}_1). \quad (21.199)$$

Given that parameter time dependence can be viewed as an extension of structural change at every observation  $t = k+1, \dots, T$ , it is not surprising that the Chow test statistic is directly related to the *CUSUMSQ* statistic discussed in Section 21.5 above (see McAleer and Fisher (1982)).

This test can also be used for  $H_0$  against  $H_1 \cap H_0^{(2)}$  but caution should be exercised because although the size is correct the test can be inconsistent when  $H_0^*$  is valid but  $H_0^{(1)}$  is not (see Rea (1978), Mizon and Richard (1986)). Moreover, for  $H_0^*$  against  $H_1$  the above test does not have the correct size against alternatives when  $\sigma_1^2 \neq \sigma_2^2$ , given that the distribution under  $H_0^*$  when  $H_0^{(2)}$  is invalid is not  $F(T_2, T_1 - k)$ . Similarly for testing  $H_0$  against  $H_1$  the Chow test has the correct size but low power for alternatives when  $H_0^*$  and  $H_0^{(2)}$  are valid but  $H_0^{(1)}$  is not, given that the implicit alternative is in fact  $H_0^* \cap H_0^{(2)}$ . These comments can also be made for the test based on  $\tau(\mathbf{y}_t^0)$  in Section 21.5.

As in case 1 ( $T_2 < k$ ), we need to test  $H_0^{(2)}$  before we can safely apply the Chow test. Given that  $s_2^2$  cannot be estimated, intuition suggests using the prediction sum of squares

$$\sum_{t=T_1+1}^T (y_{2t} - \hat{\boldsymbol{\beta}}'_1 \mathbf{x}_{2t})^2 = (\mathbf{y}_2 - \mathbf{X}_2 \hat{\boldsymbol{\beta}}_1)' (\mathbf{y}_2 - \mathbf{X}_2 \hat{\boldsymbol{\beta}}_1) \quad (21.200)$$

instead. This gives rise to the test statistic

$$\tau_4(\mathbf{y}) = \frac{(\mathbf{y}_2 - \mathbf{X}_2 \hat{\boldsymbol{\beta}}_1)'(\mathbf{y}_2 - \mathbf{X}_2 \hat{\boldsymbol{\beta}}_1)}{s_1^2}, \quad \text{where } s_1^2 = \frac{RSS_1}{T_1 - k}. \quad (21.201)$$

It is not very difficult to see, however, that the numerator and denominator of this statistic are not independent and hence the ratio is not  $F$ -distributed.

Asymptotically, however,  $s_1^2 \xrightarrow{P} \sigma^2$  under  $H_0^{(2)}$  and thus

$$\tau_4(\mathbf{y}) \underset{\alpha}{\sim} \chi^2(T_2) \quad \text{under } H_0^{(2)}. \quad (21.202)$$

Using this we can construct an *asymptotic size  $\alpha$  test* for  $H_0^{(2)}$  against  $H_1$  based on the rejection region

$$C_1 = \{\mathbf{y}: \tau_4(\mathbf{y}) > c_\alpha\}, \quad \alpha = \int_{c_\alpha}^\infty d\chi^2(T_2). \quad (21.203)$$

Let us apply the above tests to the money equation discussed above. Estimation of this equation for the period 1963*i*–1980*iii* yielded:

$$m_t = 2.685 + 0.713y_t + 0.852p_t - 0.052i_t + \hat{u}_t, \quad (21.204)$$

(1.055)	(0.107)	(0.022)	(0.014)	(0.039)
---------	---------	---------	---------	---------

$$R^2 = 0.994, \quad \bar{R}^2 = 0.994, \quad s_1 = 0.0392,$$

$$\log L = 138.52, \quad RSS_1 = 0.10923, \quad T = 75.$$

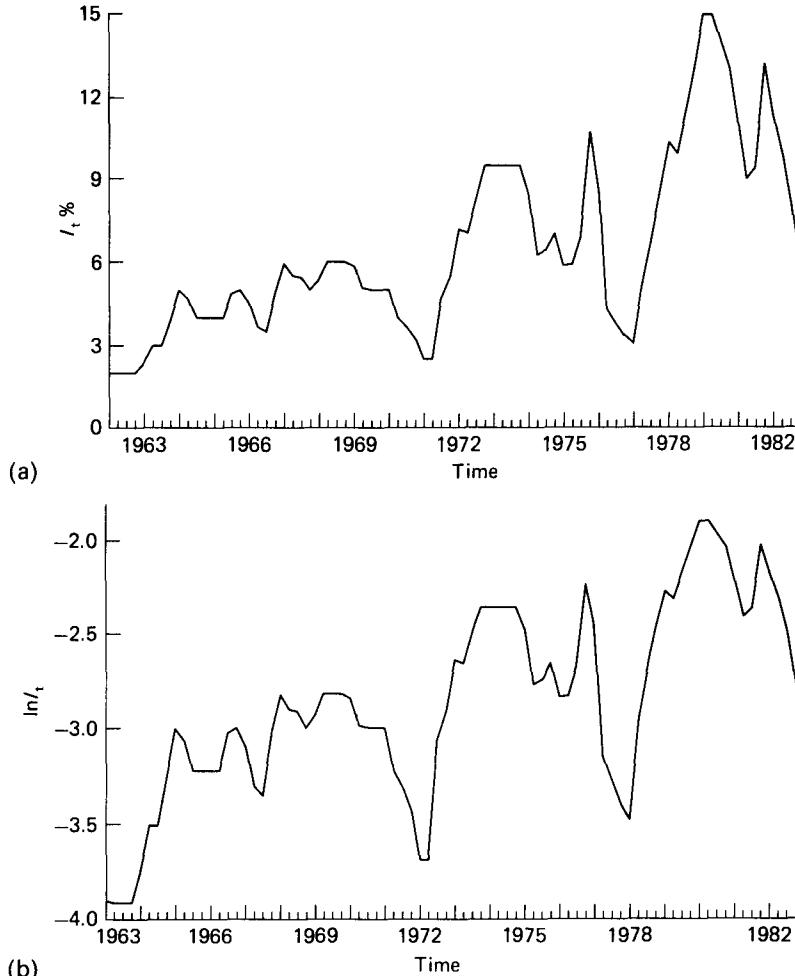
Testing for  $H_0^{(2)}$  against  $H_1: \tau_4(\mathbf{y}) = 6.484, c_\alpha = 11.07, \alpha = 0.05$ . Testing for  $H_0^*$  against  $H_1 \cap H_0^{(2)}$ :  $CH = 1.078, c_\alpha = 2.34, \alpha = 0.05$ . These results imply that both hypotheses cannot be rejected. This is not very surprising given that the post-sample period used for the tests was rather small ( $T_2 = 5$ ). In cases where  $T_2$  is larger we expect the two hypotheses to be rejected on the basis of the results using  $CH$ . In the present case this was not attempted because when  $T_2 > k$ ,  $CH$  is no longer the ‘best’ test to use;  $\tau_2(\mathbf{y})$  is the optimal test.

### Appendix 21.1 – variance stabilising transformations

Consider the case where  $\text{Var}(y_t | \mathbf{X}_t = \mathbf{x}_t) = \sigma^2 g(\mu_t)$  and  $g(\cdot)$  is a known function. Our aim is to find a transformation  $h(\cdot)$  such that  $\text{Var}(y_t^* | \mathbf{X}_t = \mathbf{x}_t) = \sigma^2$ . Let us assume that we can take the first-order Taylor expansion of  $h(y_t)$  at  $\mu_t$ :

$$h(y_t) \approx h(\mu_t) + (y_t - \mu_t)h'(\mu_t),$$

$h'(\mu_t)$  being the first derivative of  $h(y_t)$  evaluated at  $y_t = \mu_t$ . Then, we can use

Fig. 21.4(a). Time graph of  $I_t$ . (b) Time graph of  $\ln I_t$ .

this approximation in order to approximate the variance of  $h(y_t)$  by

$$\begin{aligned}\text{Var}(h(y_t)/\mathbf{X}_t = \mathbf{x}_t) &\simeq \text{Var}[\{h(\mu_t) + (y_t - \mu_t)h'(\mu_t)\}/\mathbf{X}_t = \mathbf{x}_t] \\ &= \text{Var}[y_t/\mathbf{X}_t = \mathbf{x}_t](h'(\mu_t))^2.\end{aligned}$$

This implies that when we choose  $h(\cdot)$  such that

$$h'(\mu_t) = \frac{1}{[g(\mu_t)]},$$

then  $\text{Var}(y_t^*/\mathbf{X}_t = \mathbf{x}_t) \simeq \sigma^2$ .

The variance stabilising transformation can be used in a general case where the variance of a random variable depends on some unwanted parameters  $\psi_t$ . In the case where  $\text{Var}(y_t/X_t=x_t)=\mu_t\sigma^2$  the transformation  $h(y_t)=\log_e y_t$  is called for because

$$h'(\mu_t) = \frac{1}{\mu_t}.$$

In Fig. 21.4(a) the time path of 7 days' interest rate is shown which exhibits a variance increasing with the level  $\mu_t$ . Its  $\log_e$  transformation is shown in Fig. 21.4(b).

### **Important concepts**

Auxiliary regression misspecification test procedure, Kolmogorov–Gabor polynomial, Gauss–Markov theorem, OLS estimators, elliptical distributions, non-linear conditional expectations, normalising transformations, GLS estimators, variance stabilising transformations, time dependent parameters, recursive estimator, structural change.

### **Questions**

1. Explain the relationship between normality, linearity and homoskedasticity.
2. Explain the intuition underlying the ‘omitted variables’ type misspecification test.
3. State the finite sample properties of the OLS estimator

$$\mathbf{b}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

based on the assumptions that  $(y_t/X_t=x_t)\sim D(\beta'\mathbf{x}_t, \sigma^2)$  and  $\mathbf{y}$  is an independent sample from  $D(y_t/X_t; \psi), t=1, 2, \dots, T$ , where the form of  $D(\cdot)$  is not known.

4. Explain the statement: ‘The OLS estimator  $\mathbf{b}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$  of  $\beta$  has minimum variance among the class of linear and unbiased estimators.’
5. ‘The Gauss–Markov theorem shows that  $\mathbf{b}$  is a fully efficient estimator.’ Discuss.
6. ‘The normality assumption is largely unnecessary for the linear regression model because without it we can use the OLS estimators  $\mathbf{b}$  and  $s^2$  of  $\beta$  and  $\sigma^2$ , respectively. Moreover, all the hypothesis testing results about  $\beta$  and  $\sigma^2$  derived using the MLE’s  $\hat{\beta}$  and  $\sigma^2$  are asymptotically valid, anyway.’ Discuss.

490      **Departures from assumptions – probability model**

7. Explain the difference in the asymptotic distributions of  $s^2$  and  $\hat{s}^2$ , the MLE and OLS estimators of  $\sigma^2$ .
8. Explain the intuition underlying the skewness–kurtosis test for departures from normality.
9. How do we proceed when the normality assumption is rejected in the linear regression model?
10. Discuss the implications of non-linearity in the context of the linear regression model.
11. How can we test for non-linearity?
12. ‘When linearity is rejected by some misspecification test we should adopt a non-linear specification  $h(\theta, \mathbf{x}_t)$  instead of  $\beta' \mathbf{x}_t$  and retaining the assumption of normality for  $D(y_t | \mathbf{X}_t; \psi_1)$  proceed to derive MLE’s for  $\theta$  and  $\sigma^2$ .’ Discuss.
13. Discuss the implications of heteroskedasticity in the context of the linear regression model.
14. ‘The comparison between  $\bar{\beta}$  and  $\hat{\beta}$  is largely irrelevant since the derivation of  $\bar{\beta}$  is based on the assumption that  $\Omega$  is known up to a scalar multiple.’ Discuss.
15. Explain the intuition underlying the derivation of the GLS estimator  $\bar{\beta}$  of  $\beta$ .
16. Discuss the following matrix inequalities:

$$(\mathbf{X}' \Omega^{-1} \mathbf{X})^{-1} - (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \Omega \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \leq \mathbf{0},$$

$$\sigma^2 (\mathbf{X}' \mathbf{X})^{-1} - (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \Omega \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \geq \mathbf{0}.$$

17. ‘Although  $\hat{\beta} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}$  is both unbiased and consistent estimator of  $\beta$ , no consistent inference about  $\beta$  is possible since no consistent estimator of  $\Omega$  is possible unless  $\sigma_t^2$  is modelled.’ Discuss.
18. Explain the intuition underlying the White heteroskedasticity test.
19. How is the White heteroskedasticity related to non-linearity as well?
20. ‘The way to proceed in the case where homoskedasticity is rejected by some misspecification test is to model  $\sigma_t^2$  by relating it to  $\mathbf{x}_t$  or some other exogenous variable  $\mathbf{z}_t$  and retaining the linearity and normality assumptions proceed to derive MLE’s of  $\beta, \sigma^2$  and the unknown parameters in the postulated model of  $\sigma_t^2$ .’ Discuss.
21. Discuss the linear regression model based on the assumption that  $D(y_t, \mathbf{X}_t; \psi)$  is multivariate  $t$ . Explain why  $\mathbf{X}_t$  cannot be weakly exogenous for  $\theta \equiv (\beta, \sigma^2)$  in this model.
22. Explain how we can reconcile the heteroskedasticity of  $\text{Var}(y_t | \mathbf{X}_t = \mathbf{x}_t)$  with the normality of  $D(y_t | \mathbf{X}_t; \theta)$ .
23. What do we mean by time invariance of  $\theta \equiv (\beta, \sigma^2)$ ?
24. Explain how non-stationarity of

$$\mathbf{Z}_t \equiv \begin{pmatrix} y_t \\ \mathbf{X}_t \end{pmatrix}$$

can lead to time dependence parameters of interest.

25. Explain the following form of the recursive estimator of  $\beta_t$ :

$$\hat{\beta}_t = \hat{\beta}_{t-1} + (\mathbf{X}_t' \mathbf{X}_t^0)^{-1} \mathbf{x}_t (y_t - \hat{\beta}_{t-1}' \mathbf{x}_t), \quad t = k+1, \dots, T.$$

26. How can we test for parameter time-invariance in the context of the linear regression model?
27. Explain the intuition underlying the *CUSUM* test.
28. ‘The implicit null in testing for coefficient time invariance using the recursive residuals is not  $H_0^{(1)}$ :  $\beta_t = \beta_{t-1}$  but  $H_0^*: \mathbf{x}_t(\beta_t - \beta_{t-1}) = 0$ .’ Discuss.
29. How do we tackle the problem of time dependence of  $\theta$ ?
30. How do we test for coefficient structural stability (constancy) in the case where  $T_2 > k$ ? How is this different from the case  $T_2 < k$ ?
31. ‘In the case where  $T_2 < k$  the implicit null for coefficient constancy is not  $H_0^{(1)}$ :  $(\beta_1 - \beta_2) = \mathbf{0}$  but  $H_0^*: \mathbf{X}_2(\beta_1 - \beta_2) = \mathbf{0}$ .’ Discuss.

### Exercises

1. Let  $D(v_t; \theta)$  be univariate normal of the form  $v_t \sim N(0, \sigma^2)$ . Derive

$$E\left(\frac{\partial^2 \ln D(v_t; \theta)}{\partial \sigma}\right) \text{ and } E\left(\frac{\partial \ln D(v_t; \theta)}{\partial \sigma}\right)^2;$$

when  $v_t \sim N(0, \sigma^2)$  and compare them with the same quantitites when  $v_t \sim D(0, \sigma_t^2)$  and the form of  $D(\cdot)$  is unknown.

2. Show that  $\mathbf{b} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}$  has minimum variance among the class of unbiased estimators of the form

$$\mathbf{b}^* = (\mathbf{L} + \mathbf{C}) \mathbf{y}, \quad \mathbf{L} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}',$$

$\mathbf{C}$  is a  $k \times T$  arbitrary matrix.

3. Show that under the assumption  $(y_t | \mathbf{X}_t = \mathbf{x}_t) \sim D(h(\mathbf{x}_t), \sigma^2)$ ,

$$E(\hat{\beta}) \neq \beta \quad \text{and} \quad E(s^2) \neq \sigma^2.$$

4. Show that under the assumption  $(y_t | \mathbf{X}_t = \mathbf{x}_t) \sim D(\beta' \mathbf{x}_t, \sigma_t^2)$

$$\text{Cov}(\hat{\beta}) = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \Omega \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1}.$$

5. Derive the GLS estimator of  $\beta$  under the assumption of exercise 3. Show that knowing  $\Omega$  or  $\Lambda$  where  $\Omega = \sigma^2 \Lambda$  is essentially the same as far as estimation of  $\beta$  is concerned.

492      **Departures from assumptions – probability model**

6. Using the formula  $\mathbf{B}^{-1} = \mathbf{A}^{-1} - c\mathbf{A}^{-1}\mathbf{a}\boldsymbol{\beta}'\mathbf{A}^{-1}$  for  $\mathbf{B} = \mathbf{A} + \mathbf{a}\boldsymbol{\beta}'$ , where  $c = 1/(1 + \boldsymbol{\beta}'\mathbf{A}^{-1}\mathbf{a})$ , show that

$$(\mathbf{X}_t^{0'} \mathbf{X}_t^0)^{-1} = (\mathbf{X}_{t-1}^{0'} \mathbf{X}_{t-1}^0)^{-1} - \frac{(\mathbf{X}_{t-1}^{0'} \mathbf{X}_{t-1}^0)^{-1} \mathbf{x}_t \mathbf{x}_t' (\mathbf{X}_{t-1}^{0'} \mathbf{X}_{t-1}^0)^{-1}}{1 + \mathbf{x}_t' (\mathbf{X}_{t-1}^{0'} \mathbf{X}_{t-1}^0)^{-1} \mathbf{x}_t}.$$

7. Show that

$$\hat{\boldsymbol{\beta}}_t = \hat{\boldsymbol{\beta}}_{t-1} + \frac{(\mathbf{X}_{t-1}^{0'} \mathbf{X}_{t-1}^0)^{-1} \mathbf{x}_t (y_t - \hat{\boldsymbol{\beta}}_{t-1}' \mathbf{x}_t)}{1 + \mathbf{x}_t' (\mathbf{X}_{t-1}^{0'} \mathbf{X}_{t-1}^0)^{-1} \mathbf{x}_t}.$$

8. Show that  $(\mathbf{X}_{t-1}^{0'} \mathbf{X}_{t-1}^0)^{-1} = \sum_{i=1}^{t-1} \mathbf{x}_i \mathbf{x}_i'$ .  
 9. Verify the expressions for  $\text{Var}(w_t)$  and  $\text{Cov}(w_t w_s)$  of Section 21.5.  
 10. Show that  $w_t^2 = RSS_t - RSS_{t-1}$ ,  $t \geq k+1$ , where  $RSS_t = \sum_{i=1}^t (y_i - \hat{\boldsymbol{\beta}}_t' \mathbf{x}_i)^2$ .

**Additional references**

Bickel and Doksum (1981); Gourieroux *et al.* (1984); Hinkley (1975); Lau (1985); Ramsey (1974); White and MacDonald (1980); Zarembka (1974).

## CHAPTER 22

---

### The linear regression model IV – departures from the sampling model assumption

---

One of the most crucial assumptions underlying the linear regression model is the sampling model assumption that  $\mathbf{y} \equiv (y_1, y_2, \dots, y_T)'$  constitutes an independent sample sequentially drawn from  $D(y_t/\mathbf{X}_t; \boldsymbol{\theta})$ ,  $t = 1, 2, \dots, T$ , respectively. This assumption enables us to define the likelihood function to be

$$L(\boldsymbol{\theta}; \mathbf{y}) = c(\mathbf{y}) \prod_{t=1}^T D(y_t/\mathbf{X}_t; \boldsymbol{\theta}). \quad (22.1)$$

Intuitively, this assumption amounts to postulating that the ordering of the observations in  $y_t$  and  $\mathbf{X}_t$  plays no role in the statistical analysis of the model. That is, in the case where the data on  $y_t$  and  $\mathbf{X}_t$  are punched observation by observation a reshuffling of the cards will change none of the results in Chapter 19. This is a very restrictive assumption for most economic time-series data where some temporal dependence between successive values seems apparent. As argued in Chapter 17, for most economic time series the non-random sampling model seems more appropriate.

In Section 22.1 we consider the implications of a non-independent sample for the statistical results derived in Chapter 19. It is argued that these implications depend crucially on how non-independence is modelled and two alternative modelling strategies are discussed. These strategies give rise to two alternative approaches, the *respecification* and *autocorrelation approaches* to the misspecification testing and ways to tackle the dependence in the sample. In the context of the autocorrelation approach the dependence is interpreted as due to error temporal correlation. On the other hand, in the context of the misspecification approach the error term's

role as the non-systematic component of the statistical GM is retained and the dependence in the samples is modelled from first principles in terms of the observable random variables involved. Sections 22.2 and 22.3 consider various ways to proceed with a non-random sample and the misspecification testing for the independent sample assumptions respectively. In Section 22.4 the discussion of misspecification analysis in Chapters 20–22 is put into perspective.

## 22.1 Implications of a non-random sample

### (1) Defining the concept of a non-random sample

It is no exaggeration to say that the sampling model assumption of independence is by far the most crucial assumption underlying the linear regression model. As shown below, when this assumption is invalid no estimation, testing or prediction result derived in Chapter 19 is valid in general. In order to understand the circumstances under which the independence assumption might be inappropriate in practice, it is instructive to return to the linear regression model and consider the reduction from  $D(\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_T; \boldsymbol{\psi})$  to  $D(y_t/\mathbf{X}_t; \boldsymbol{\theta}), t = 1, 2, \dots, T$ , in order to understand the role of the assumption in the reduction process and its relation to the NIID assumption for  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$ ,  $\mathbf{Z}_t \equiv (y_t, \mathbf{X}_t)'$ .

As argued in Chapter 19, the linear regression model could be based directly on the conditional distribution  $D(y_t/\mathbf{X}_t; \boldsymbol{\psi}_1)$  and no need to define  $D(\mathbf{Z}_t; \boldsymbol{\psi})$  arises. This was not the approach adopted for a very good reason. In practice it is much easier to judge the appropriateness of assumptions related to  $D(\mathbf{Z}_t; \boldsymbol{\psi})$ , on the basis of the observed data, rather than assumptions related to  $D(y_t/\mathbf{X}_t; \boldsymbol{\psi}_1)$ . What is more, the nature of the latter distribution is largely determined by that of the former.

In Chapter 19 we assumed that  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is a normal, independent and identically distributed (NIID) stochastic process (see Chapter 8). On the basis of this assumption we were able to build the linear regression model defined by assumptions [1]–[8]. In particular the probability and sampling model assumptions can be viewed as consequences of  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  being NIID. The *normality* of  $D(y_t/\mathbf{X}_t; \boldsymbol{\psi}_1)$ , the *linearity* of  $E(y_t/\mathbf{X}_t = \mathbf{x}_t)$  and the *homoskedasticity* of  $\text{Var}(y_t/\mathbf{X}_t = \mathbf{x}_t)$  stem directly from the normality of  $\mathbf{Z}_t$ . The time invariance of  $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}, \sigma^2)$  and the *independent sample* assumptions stem from the *identically distributed* and *independent* component of NIID. Note that in the present context we distinguish between homoskedasticity and time invariance of  $\text{Var}(y_t/\mathbf{X}_t = \mathbf{x}_t)$  (see Section 21.5). This suggests that

the obvious way to make the independent sample assumption inappropriate is to assume that  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is a (dependent) stochastic process. In particular (because we do not want to lose the convenience of normality) we assume that

$$\mathbf{Z}_t \sim N(\mathbf{m}(t), \Sigma(t, t)), \quad \text{Cov}(\mathbf{Z}_t \mathbf{Z}_s) = \Sigma(t, s), \quad t, s \in \mathbb{T}. \quad (22.2)$$

If we return to the money equation estimated in Chapter 19, a cursory look at the realisation of  $\mathbf{Z}_t$ ,  $t = 1, 2, \dots, T$  (see Fig. 17.1(a)–(d)), would convince us that the above assumption seems much more appropriate than the IID assumption for such data. The realisations of the process exhibit a very distinct time trend (the mean changes systematically over time) and for at least two of them the variance seems to change as well.

The question which naturally arises at this stage is to what extent the assumptions underlying the linear regression model will be affected by relaxing the IID assumption for  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$ . One obvious change will come in the form of a non-independent sample  $\mathbf{y} \equiv (y_1, y_2, \dots, y_T)'$ . What is not so obvious is the distribution which will replace  $D(y_t | \mathbf{X}_t; \psi_1)$ . Table 22.1 summarises the important steps in the reduction of the linear regression model from the initial assumption that  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is a NIID stochastic process and contrasts these with the case where  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is a non-IID process. The only minor difference between the ‘construction’ of the linear regression model as summarised in this table and that of Chapter 19 is that the mean of  $\mathbf{Z}_t$  is intentionally given a non-zero value because the mean plays an important role in the case of a non-IID stochastic process.

As we can see from Table 22.1, the first important difference between the IID and non-IID case is that distribution of  $\mathbf{Z}_1, \dots, \mathbf{Z}_T$  is complicated considerably in the latter case with the presence of the temporal covariances and the fact that all the parameters are changing with  $t$ . The presence of the temporal covariance implies that the decomposition of the joint distribution of  $\mathbf{Z}_1, \dots, \mathbf{Z}_T$  in terms of the marginal distributions is no longer valid. The dependence among the  $\mathbf{Z}_t$ s implies that the only decomposition possible is the sequential conditioning decomposition (see Chapter 6) where the conditioning is relative to the past history of the process denoted by  $\mathbf{Z}_{t-1}^0 \equiv (\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_{t-1})$ ; with  $t$  representing the ‘present’. This in turn implies that the role played by  $D(\mathbf{Z}_t; \psi)$  in the IID case will now be taken over by  $D(\mathbf{Z}_t / \mathbf{Z}_{t-1}^0; \psi^*(t))$ . In particular the corresponding decomposition into the conditional ( $D(y_t / \mathbf{X}_t; \psi_1)$ ) and marginal ( $D(\mathbf{X}_t; \psi_2)$ ) distribution will now be

$$D(\mathbf{Z}_t / \mathbf{Z}_{t-1}^0; \psi^*(t)) = D(y_t / \mathbf{Z}_{t-1}^0, \mathbf{X}_t; \psi_1^*(t)) \cdot D(\mathbf{X}_t / \mathbf{Z}_{t-1}^0; \psi_2^*(t)), \quad (22.3)$$

where the *past history* of the process comes into both distributions because it contains relevant information for  $y_t$  as well as  $\mathbf{X}_t$ . Although the algebra in

Table 22.1. Relaxing the assumptions of independence and identical distribution

	(1) Normal (2) Independent (3) Identically distributed	(1)' Normal (2)' Dependent (3)' Non-identically distributed
$\{\mathbf{Z}_t, t \in \mathbb{T}\}$	$\begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \\ \vdots \\ \mathbf{Z}_T \end{pmatrix} \sim N \left( \begin{pmatrix} \mathbf{m} \\ \mathbf{m} \\ \vdots \\ \mathbf{m} \end{pmatrix}, \begin{pmatrix} \Sigma & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \Sigma & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \Sigma \end{pmatrix} \right)$	$\begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \\ \vdots \\ \mathbf{Z}_T \end{pmatrix} \sim N \left( \begin{pmatrix} \mathbf{m}(1) \\ \mathbf{m}(2) \\ \vdots \\ \mathbf{m}(T) \end{pmatrix}, \begin{pmatrix} \Sigma(1, 1), \Sigma(1, 2) & \cdots & \Sigma(1, T) \\ \Sigma(2, 1), \Sigma(2, 2) & \cdots & \Sigma(2, T) \\ \vdots & \ddots & \vdots \\ \Sigma(T, 1) & \cdots & \Sigma(T, T) \end{pmatrix} \right)$
Decomposition	$D(\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_T; \psi) = \prod_{i=1}^T D(\mathbf{Z}_i; \psi)$ $= \prod_{t=1}^T D(y_t/\mathbf{X}_t; \psi_1) D(\mathbf{X}_t; \psi_2)$ $\mathbf{Z}_t \equiv \begin{pmatrix} y_t \\ \mathbf{X}_t \end{pmatrix} \sim N \left( \begin{pmatrix} m_y \\ \mathbf{m}_x \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$	$D(\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_T; \psi^*) = \prod_{t=1}^T D(\mathbf{Z}_t/\mathbf{Z}_{t-1}^0; \psi^*(t))$ $= \prod_{t=1}^T D(y_t/\mathbf{Z}_{t-1}^0, \mathbf{X}_t; \psi_1^*(t)) \cdot D(\mathbf{X}_t/\mathbf{Z}_{t-1}^0; \psi_2^*(t))$ $(\mathbf{Z}_t/\mathbf{Z}_{t-1}^0) \sim N \left( m_y(t) + \sum_{i=1}^{t-1} [\mathbf{a}_{11}(i, t)y_{t-i} + \mathbf{a}_{12}(i, t)\mathbf{x}_{t-i}] \mathbf{m}_x(t) + \sum_{i=1}^{t-1} [\mathbf{a}_{21}(i, t)y_{t-i} + \mathbf{a}_{22}(i, t)\mathbf{x}_{t-i}], \begin{pmatrix} \omega_{11}(t) & \omega_{12}(t) \\ \omega_{21}(t) & \Omega_{22}(t) \end{pmatrix} \right)$
Probability model	<ul style="list-style-type: none"> <li>(i) <math>(y_t/\mathbf{X}_t) \sim N(c_0 + \beta' \mathbf{x}_t, \sigma^2)</math>  <math>c_0 = m_y - \sigma_{12} \Sigma_{22}^{-1} \sigma_{21} \mathbf{m}_x, \quad \beta = \Sigma_{22}^{-1} \sigma_{21},</math>  <math>\sigma^2 = \sigma_{11} - \sigma_{12} \Sigma_{22}^{-1} \sigma_{21}</math></li> <li>(ii) <math>E(y_t/\mathbf{X}_t = \mathbf{x}_t)</math> – linear in <math>\mathbf{x}_t</math></li> <li>(iii) <math>\text{Var}(y_t/\mathbf{X}_t = \mathbf{x}_t)</math> – homoskedastic</li> <li>(iv) <math>\theta \equiv (c_0, \beta, \sigma^2)</math> – time independent</li> </ul>	<ul style="list-style-type: none"> <li>(i) <math>(y_t/\mathbf{Z}_{t-1}^0, \mathbf{X}_t) \sim N(c_0(t) + \beta'_0(t)\mathbf{x}_t + \sum_{i=1}^{t-1} [\alpha_i(t)y_{t-i} + \beta'_i(t)\mathbf{x}_{t-i}], \sigma_0^2(t))</math> (for the definition of the parameters involved see Appendix I)</li> <li>(ii) <math>E(y_t/\sigma(\mathbf{Y}_{t-1}^0), \mathbf{X}_t^0 = \mathbf{x}_t^0)</math> – linear in <math>\mathbf{Z}_{t-1}^0</math></li> <li>(iii) <math>\text{Var}(y_t/\sigma(\mathbf{Y}_{t-1}^0), \mathbf{X}_t^0 = \mathbf{x}_t)</math> – homoskedastic (free of <math>\mathbf{Z}_{t-1}^0</math>)</li> <li>(iv) <math>\theta_t^* \equiv (c_0(t), \beta_0(t), \beta_i(t), \alpha_i(t), i = 1, 2, \dots, t-1, \sigma_0^2(t))</math> – time dependent</li> </ul>
Sampling model	$\mathbf{y} \equiv (y_1, \dots, y_T)'$ is an independent sample sequentially drawn from $D(y_t/\mathbf{X}_t; \psi_1)$ , $t = 1, 2, \dots, T$ , respectively	$\mathbf{y} \equiv (y_1, \dots, y_T)'$ is a non-random sample sequentially drawn from $D(y_t/\mathbf{Z}_{t-1}^0, \mathbf{X}_t; \psi_1^*(t))$ , $t = 1, 2, \dots, T$ , respectively

the non-IID case is rather involved, the underlying argument is the same as in the IID case. The role of  $D(y_t/\mathbf{X}_t; \boldsymbol{\psi}_1)$  is taken over by  $D(y_t/\mathbf{Z}_{t-1}^0, \mathbf{X}_t; \boldsymbol{\psi}_1^*(t))$  and the probability and sampling models in the non-IID case need to be defined in terms of the latter conditional distribution. A closer look at this distribution, however, reveals the following:

$$(i) \quad E(y_t/\sigma(\mathbf{Y}_{t-1}^0), \mathbf{X}_t^0 = \mathbf{x}_t^0) = c_0(t) + \boldsymbol{\beta}'_0(t)\mathbf{x}_t + \sum_{i=1}^{t-1} [\alpha_i(t)y_{t-i} + \boldsymbol{\beta}'_i(t)\mathbf{x}_{t-i}]; \quad (22.4)$$

$$(ii) \quad \text{Var}(y_t/\sigma(\mathbf{Y}_{t-1}^0), \mathbf{X}_t^0 = \mathbf{x}_t^0) = \sigma_0^2(t); \quad (22.5)$$

$$(iii) \quad \boldsymbol{\theta}_t^* \equiv (c_0(t), \boldsymbol{\beta}_0(t), \boldsymbol{\beta}_i(t), \alpha_i(t), i = 1, 2, \dots, t-1, \sigma_0^2(t)), \quad (22.6)$$

where

$$\mathbf{Y}_{t-1}^0 \equiv (y_1, y_2, \dots, y_{t-1}), \quad \mathbf{X}_t^0 \equiv (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_t).$$

Firstly, although the conditional mean is linear in the conditioning variables, the number of variables increases with  $t$ . Secondly, even though the conditional variance is homoskedastic (free of the conditioning variables) it is time dependent along with other parameters of the distribution. This renders the conditional distribution  $D(y_t/\mathbf{Z}_{t-1}^0, \mathbf{X}_t; \boldsymbol{\psi}_1^*(t))$ , as defined above, non-operational as a basis of an alternative specification to be contrasted with the linear regression model. It is in a sense much too general to be operational in practice. Theoretically, however, we can define the new sampling model as:

$\mathbf{y} \equiv (y_1, \dots, y_T)'$  represents a non-random sample sequentially drawn, from  $D(y_t/\mathbf{Z}_{t-1}^0, \mathbf{X}_t; \boldsymbol{\psi}_1^*(t))$ ,  $t = 1, 2, \dots, T$ , respectively.

In view of this it is clear that what we need to do is to restrict the generality of the underlying assumptions related to  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  in order to render  $D(y_t/\mathbf{Z}_{t-1}^0, \mathbf{X}_t; \boldsymbol{\psi}_1^*(t))$  operational. In particular, we need to impose certain restrictions on the form of the *time-heterogeneity* and *dependence* of the stochastic process  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$ . This is the purpose of the next two sub-sections. In the next sub-section the dependence of  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is restricted to *asymptotic independence* and the complete time-heterogeneity (non-identically distributed) is restricted to *stationarity* (see Chapter 8) in order to get an operational model. In sub-section (3) the time-dependence assumption is restricted even further in an attempt to retain the systematic component of the linear regression model and relegate the dependence to the error term.

(2) *The respecification approach*

As seen above, when  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is assumed to be normal, and non-IID, the conditional distribution we are interested in,  $D(y_t | \mathbf{Z}_{t-1}^0, \mathbf{X}_t; \psi_1^*(2))$ , has a mean whose number of terms increases with  $t$ , and its parameters  $\psi_1^*(t)$  are time dependent (see (4)–(6)). The question which naturally arises at this stage is to what extent we need to restrict the non-IID assumption in order to ‘solve’ the *incidental parameters problem*. In order to understand the role of the dependence and non-identically distributed assumptions let us consider restricting the latter first.

Let us restrict the non-identically distributed assumption to that of *stationarity*, i.e. assume that  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is a normal, stationary stochastic process (without any restrictions on the form of dependence). Stationarity (see Section 8.2) implies that

$$E(\mathbf{Z}_t) = \mathbf{m}, \quad \text{Cov}(\mathbf{Z}_t \mathbf{Z}_s) = \Sigma(|t-s|), \quad t, s \in \mathbb{T}, \quad (22.7)$$

i.e.

$$\begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \\ \vdots \\ \mathbf{Z}_T \end{pmatrix} \sim N \left( \begin{pmatrix} \mathbf{m} \\ \mathbf{m} \\ \vdots \\ \mathbf{m} \end{pmatrix}, \begin{pmatrix} \Sigma_0 \Sigma_1 & \cdots & \Sigma_{T-1} \\ \Sigma_1 \Sigma_0 & \ddots & \Sigma_{T-2} \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \Sigma_1 \\ \Sigma_{T-1} & \cdots & \Sigma_1 \Sigma_0 \end{pmatrix} \right),$$

$$\Sigma_i \equiv \Sigma(|t-s|), \quad |t-s|=i, \quad i=1, 2, \dots, T-1. \quad (22.8)$$

That is, the  $\mathbf{Z}_t$ s have identical means and variances and their temporal covariances depend only on the absolute value of the distance between them. This reduces the above sample  $[T \times (k+1)] \times [T \times (k+1)]$  covariance matrix to a block Toeplitz matrix (see Akaike (1974)). This restricts the original covariance matrix considerably by inducing symmetry and reducing the number of ‘different’  $(k+1) \times (k+1)$  matrices making up these covariances from  $T^2$  to  $T-1$ . A closer look at stationarity reveals that it is a direct extension of the identically distributed assumption to the case of a dependent sequence of random variables. In terms of observed data the sample realisation of a stationary process exemplifies no systematic changes either in mean or variance and any  $\tau$ -period section of the realisation should look like any other  $\tau$ -period section. That is, if we slide a  $\tau$ -period ‘window’ over the realisation along the time axis the ‘picture’ should not differ systematically. Examples of such realisations are given in Figs. 21.2 and 21.3.

Assuming that  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is a normal stationary process implies that as far

as  $D(y_t/\mathbf{Z}_{t-1}^0, \mathbf{X}_t; \psi_2^*(t))$  is concerned (4)–(6) take the form:

$$(i) \quad E(y_t/\sigma(\mathbf{Y}_{t-1}^0), \mathbf{X}_t^0 = \mathbf{x}_t^0) = c_0 + \boldsymbol{\beta}'_0 \mathbf{x}_t + \sum_{i=1}^{t-1} \alpha_i y_{t-i} + \sum_{i=1}^{t-1} \boldsymbol{\beta}'_i \mathbf{x}_{t-i}, \quad (22.9)$$

$$(ii) \quad \text{Var}(y_t/\sigma(\mathbf{Y}_{t-1}^0), \mathbf{X}_t^0 = \mathbf{x}_t^0) = \sigma_0^2; \quad (22.10)$$

$$(iii) \quad \boldsymbol{\theta}^* \equiv (c_0, \boldsymbol{\beta}_0, \boldsymbol{\beta}_i, \alpha_i, i = 1, 2, \dots, t-1, \sigma_0^2). \quad (22.11)$$

As we can see, stationarity enables us to ‘solve’ the parameter time-dependence problem but the incidental parameters problem remains largely unresolved because the number of terms (and unknown parameters) in the conditional mean (9) increases with  $t$ . Hence, the time-homogeneity introduced by stationarity is not sufficient for an operational model. We also need to restrict the form of dependence of  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$ . In particular we need to restrict the ‘memory’ of the process by imposing restrictions such as ergodicity, mixing or asymptotic independence. In the present case the most convenient memory restriction is that of *asymptotic independence*. This restricts the conditional memory of the normal process so as to enable us to approximate the conditional mean of the process using an *mth-order Markov* process (see Chapter 8). A stochastic vector process  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is said to be *mth-order Markov* if

$$E(\mathbf{Z}_t/\mathbf{Z}_{t-1}^0) = E(\mathbf{Z}_t/\sigma(\mathbf{Z}_{t-1}, \mathbf{Z}_{t-2}, \dots, \mathbf{Z}_{t-m})), \quad t > m. \quad (22.12)$$

Assuming that  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is:

- (i) normal;
- (ii) stationary; and
- (iii) asymptotically independent

enables us to deduce that for large enough  $m$  (hopefully  $m < T$ ) the conditional mean takes the form

$$\mu_t^* \equiv E(y_t/\sigma(\mathbf{Y}_{t-1}^0), \mathbf{X}_t^0 = \mathbf{x}_t^0) = c_0 + \boldsymbol{\beta}'_0 \mathbf{x}_t + \sum_{i=1}^m \alpha_i y_{t-i} + \sum_{i=1}^m \boldsymbol{\beta}'_i \mathbf{x}_{t-i}. \quad (22.13)$$

This form provides us with an operational form for the systematic component for  $t > m$ . Now  $\boldsymbol{\theta}^* = (c_0, \boldsymbol{\beta}_0, \boldsymbol{\beta}_i, \alpha_i, i = 1, 2, \dots, m, \sigma_0^2)$  is both time invariant and its dimensionality remains fixed as  $T$  increases. Indeed,  $D(y_t/\mathbf{Z}_{t-1}^0, \mathbf{X}_t; \psi_1^*)$  yields a mean linear in  $\mathbf{x}_t, y_{t-i}, \mathbf{x}_{t-i}$ ,  $i = 1, 2, \dots, m$ , a homoskedastic variance and  $\psi_1^*$  is time invariant. Hence, defining the non-systematic component by

$$u_t^* = y_t - E(y_t/\sigma(\mathbf{Y}_{t-1}^0)\mathbf{X}_t^0 = \mathbf{x}_t^0), \quad t > m, \quad (22.14)$$

we can define a new statistical GM based on  $D(y_t/\mathbf{Z}_{t-1}^0, \mathbf{X}_t; \boldsymbol{\psi}_1^*)$  to be

$$y_t = \boldsymbol{\beta}'_0 \mathbf{x}_t + \sum_{i=1}^m (\alpha_i y_{t-i} + \boldsymbol{\beta}'_i \mathbf{x}_{t-i}) + u_t^*, \quad t > m. \quad (22.15)$$

In matrix form for the sample period  $t = m+1, \dots, T$ , it can be expressed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{Z}^*\boldsymbol{\gamma} + \mathbf{u}^* \quad (22.16)$$

in an obvious notation. Note that  $c_0$  has been dropped for notational convenience (implicitly included in  $\boldsymbol{\beta}_0$ ).

It must be noted at this stage that the assumption of stationarity and asymptotic independence for  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  are not the *least restrictive* assumptions for the results which follow. For example asymptotic independence can be weakened to ergodicity (see Section 8.3) without affecting any of the asymptotic results which follow. Moreover, by ‘strengthening’ the memory restriction to that of  $\varphi$ -mixing some time-heterogeneity might be allowed without affecting the asymptotic results (see White (1984)).

It is also important to note that the maximum lag  $m$  in (15) does not represent the maximum memory lag of the process  $\{y_t/\mathbf{Z}_{t-1}^0, \mathbf{X}_t, t \in \mathbb{T}\}$  as in the case of an  $m$ -dependence (see Section 8.3). Although there is a duality result relating an  $m$ -dependent with an  $m$ th-order Markov process, in the case of the latter the memory is considerably longer than  $m$  (see Chapter 8). This is one of the reasons why the AR representation is preferred in the present context. The maximum memory lag is determined by the solution of the lag polynomial:

$$\alpha(L) = \left(1 - \sum_{i=1}^m \alpha_i L^i\right) = 0. \quad (22.17)$$

In view of the statistical GM (15) let us consider the implications of the non-random sample (i)–(iii) for the estimation, testing and prediction results in Chapter 19. Starting with estimation we can deduce that for

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (22.18)$$

and

$$s^2 = (1/T - k)\hat{\mathbf{u}}'\hat{\mathbf{u}} \quad (22.19)$$

the following results hold:

- (i)  $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}_0 + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Z}^*\boldsymbol{\gamma}) \neq \boldsymbol{\beta}$ ;  $\hat{\boldsymbol{\beta}}$  is a biased estimator of  $\boldsymbol{\beta}$ ;
- (ii)  $\hat{\boldsymbol{\beta}} \not\rightarrow \boldsymbol{\beta}$ ;  $\hat{\boldsymbol{\beta}}$  is an *inconsistent* estimator of  $\boldsymbol{\beta}$ ;
- (iii)  $MSE(\hat{\boldsymbol{\beta}}) \equiv \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Z}^*\boldsymbol{\gamma}\boldsymbol{\gamma}'\mathbf{Z}^*)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \neq \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ ;
- (iv)  $E(s^2) = \sigma^2 + \boldsymbol{\gamma}'E(\mathbf{Z}^*\mathbf{M}_x\mathbf{Z}^*)\boldsymbol{\gamma} \neq \sigma^2$ ;  $s^2$  is a *biased* estimator of  $\sigma^2$ ;

- (v)  $s^2 \xrightarrow{P} \sigma^2$ ;  $s^2$  is an *inconsistent* estimator of  $\sigma^2$ ;  
 (vi)  $s^2(\mathbf{X}'\mathbf{X})^{-1} \not\xrightarrow{P} \text{MSE}(\hat{\boldsymbol{\beta}})$ ;

where  $\mathbf{M}_x = \mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ . These results show clearly that the implications of the non-random sample assumption are very serious for the appropriateness of  $\hat{\boldsymbol{\theta}} \equiv (\hat{\boldsymbol{\beta}}, s^2)$  as estimators of  $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}, \sigma^2)$ . Moreover, (i)(vi) taken together imply that none of the testing or prediction results derived in Chapter 19, under the assumption of an independent sample, are valid. In particular the  $t$ -statistics for the significance of the coefficients in the estimated money equation are invalid together with the tests for the coefficients of  $y_t$  and  $p_t$  being equal to one as well as the prediction intervals.

At first sight the argument underlying the derivation of (i)–(vi) seems to be identical to the ‘omitted variables problem’ criticised in Section 20.2 as being rather uninteresting in that context. A direct comparison, however, between (15) and

$$y_t = \boldsymbol{\beta}' \mathbf{x}_t + u_t, \quad t \in \mathbb{T} \quad (22.20)$$

reveals that both statistical GM’s are special cases of the general statistical GM

$$y_t = E(y_t | \sigma(\mathbf{Y}_{t-1}^0), \mathbf{X}_t^0 = \mathbf{x}_t^0) + \varepsilon_t \quad (22.21)$$

under alternative assumptions on  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$ . In this sense (20) and (15) constitute ‘reductions’ from the same joint distribution

$$D(\mathbf{Z}_1, \dots, \mathbf{Z}_T; \psi) \quad (22.22)$$

which makes them directly comparable: they are based on the same conditioning set

$$\mathcal{Q}_t = \{\sigma(\mathbf{Y}_{t-1}^0), \mathbf{X}_t^0 = \mathbf{x}_t^0\}. \quad (22.23)$$

### (3) The autocorrelation approach

The main difference between the respecification approach considered above and the autocorrelation approach lies with the systematic component. In the respecification approach we need to respecify the systematic component in order to take account of (model) the temporal systematic information from the sample. In the autocorrelation approach the systematic component remains the same and hence the temporal systematic information will be relegated to the error term which will no longer be non-systematic. The term *autocorrelation* stems from the fact that dependence in this context is interpreted as due to the correlation among

the error terms. This is contrary to the logic of the approach of statistical model specification propounded in the present book (see Chapter 17). The approach, however, is important for various reasons. Firstly, the comparison with the respecification approach is very illuminating for both approaches. Secondly, the autocorrelation approach dominates the textbook econometric literature and as a consequence it provides the basis for most misspecification tests of the independent sample assumption.

The systematic component for the autocorrelation approach is exactly the same as the one under the independent sample assumption. That is, assuming a certain form of temporal dependence (see (41)).

$$\mu_t \equiv E(y_t/\mathcal{D}_t) = \beta' \mathbf{x}_t. \quad (22.24)$$

This implies that the temporal dependence in the sample will be left in the error term:

$$\varepsilon_t = y_t - E(y_t/\mathcal{D}_t). \quad (22.25)$$

$\mathcal{D}_t$  as defined in (23). In view of this the error term will satisfy the following properties:

$$(i) \quad E(\varepsilon_t/\mathcal{D}_t) = 0 \quad (22.26)$$

$$(ii) \quad E(\varepsilon_t \varepsilon_s / \mathcal{D}_t) = \begin{cases} \sigma^2(t), & t=s \\ v(t, s), & t \neq s. \end{cases} \quad (22.27)$$

These assumptions in terms of the observable r.v.'s are often expressed in the rather misleading notation:

$$(y/\mathbf{X}) \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{V}_T), \quad \mathbf{V}_T > 0. \quad (22.28)$$

The question which naturally arises at this stage is, 'what are the implications of this formulation for the results related to the linear regression model derived in Chapter 19 under the independence assumption?' As far as the estimation results related to  $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$  and  $s^2 = [1/(T-k)]\hat{\mathbf{u}}'\hat{\mathbf{u}}$ ,  $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\beta}$  are concerned we can show that:

- (i)'  $E(\hat{\beta}) = \beta$ ,  $\hat{\beta}$  is an unbiased estimator of  $\beta$ ;
- (ii)'  $\hat{\beta} \xrightarrow{P} \beta$  if  $\lim_{T \rightarrow \infty} (\mathbf{X}'\mathbf{V}_T\mathbf{X})/T < \infty$  and non-singular;
- (iii)'  $\text{Cov}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_T\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ ;
- (iv)'  $E(s^2) = [\sigma^2/(T-k)] \text{tr}(1 - \mathbf{P}_X)\mathbf{V}_T \neq \sigma^2$ ;
- (v)'  $s^2 \xrightarrow{P} \sigma^2$ ,  $s^2$  is an inconsistent estimator of  $\sigma^2$ ;
- (vi)'  $s^2(\mathbf{X}'\mathbf{X})^{-1} \not\xrightarrow{P} \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_T\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ ;
- (vii)  $\hat{\beta}$  and  $s^2$  are *not* independent.

In view of (iii)', (v)', (vi)' and (vii)' we can conclude that the testing results derived in Chapter 19 are also invalid.

The important difference between the results (i)–(vi) and (i)'–(vi)' is that  $\hat{\beta}$  is not such a ‘bad’ estimator in the latter case. This is not surprising, however, given that we retained the systematic component of the linear regression model. On the other hand, the results based on  $\text{Cov}(\mathbf{y}/\mathcal{X} = \mathbf{X}) = \sigma^2 \mathbf{I}_T$  are inappropriate. The only undesirable property of  $\hat{\beta}$  in the present context is said to be its inefficiency relative to the proper MLE of  $\beta$  when  $\mathbf{V}_T$  is assumed known. That is,  $\hat{\beta}$  is said to be an inefficient estimator relative to the GLS estimator

$$\hat{\beta} = (\mathbf{X}' \mathbf{V}_T^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}_T^{-1} \mathbf{y} \quad (22.29)$$

(see Judge *et al.* (1985)). A very similar situation was encountered in the case of heteroskedasticity and the same comment applies here as well. This efficiency comparison is largely irrelevant. In order to be able to make justifiable efficiency comparisons we should be able to compare  $(\hat{\beta}, \hat{\sigma}^2)$  with estimators based on the same information set. It is well known, however, that in the case where  $\mathbf{V}_T$  is unknown no consistent estimator of the parameters of interest exist and the information matrix could not be used to define a full efficiency lower bound.

## 22.2 Tackling temporal dependence

The question: ‘How do we proceed when the independent sample assumption is invalid?’ will be considered before the testing of this assumption because in the autocorrelation approach the two are inextricably bound up and the testing becomes easier to understand when the above question is considered first. This is because most misspecification tests of sample independence in the autocorrelation approach consider particular forms of departure from the independence assumption which we will discuss in the present section. This approach, however, will be considered after the respecification approach because, as mentioned above, the former is a special case of the latter. Moreover, the respecification approach provides a most illuminating framework in the context of which the autocorrelation approach can be thoroughly discussed.

### (1) *The respecification approach*

In Section 22.1 we considered the question of respecifying the components of the linear regression model in view of the dependence in the sampling model. It was argued that in the case where  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is assumed to be a stationary and asymptotically independent process the systematic

component takes the form

$$\mu_t^* = E(y_t/\sigma(Y_{t-1}^0), \mathbf{X}_t^0 = \mathbf{x}_t^0) = \boldsymbol{\beta}'_0 \mathbf{x}_t + \sum_{i=1}^m (\alpha_i y_{t-i} + \boldsymbol{\beta}'_i \mathbf{x}_{t-i}), \quad t > m. \quad (22.30)$$

The non-systematic component is defined by

$$u_t = y_t - E(y_t/\sigma(Y_{t-1}^0), \mathbf{X}_t^0 = \mathbf{x}_t^0), \quad t > m. \quad (22.31)$$

This suggests that  $\{u_t, t > m\}$  defines a *martingale difference process* relative to the sequence  $\mathcal{D}_t = \sigma(\mathbf{Z}_t^0, \mathbf{X}_{t+1})$ ,  $t > m$ , of  $\sigma$ -fields (see Chapter 8). That is,

$$E(u_t/\mathcal{D}_{t-1}) = 0, \quad t > m. \quad (22.32)$$

Moreover,

$$E(u_t u_s) = \begin{cases} \sigma_0^2, & t = s \\ 0, & t > s. \end{cases} \quad (22.33)$$

i.e. it is an *innovation process*. These properties will play a very important role in the statistical analysis of the implied statistical GM:

$$y_t = \boldsymbol{\beta}'_0 \mathbf{x}_t + \sum_{i=1}^m \alpha_i y_{t-i} + \sum_{i=1}^m \boldsymbol{\beta}'_i \mathbf{x}_{t-i} + u_t, \quad t > m. \quad (22.34)$$

If we compare (34) with the statistical GM of the linear regression model we can see that the error term, say  $\varepsilon_t$ , in

$$y_t = \boldsymbol{\beta}' \mathbf{x}_t + \varepsilon_t, \quad t \in \mathbb{T}, \quad (22.35)$$

is no longer white noise relative to the information set  $(\sigma(Y_{t-1}^0), \mathbf{X}_t^0 = \mathbf{x}_t^0)$  given that  $\varepsilon_t$  is largely predictable from this information set (see Granger (1980)). Moreover, in view of the behaviour of  $\varepsilon_t$  the recursive estimator

$$\hat{\boldsymbol{\beta}}_t = \hat{\boldsymbol{\beta}}_{t-1} + (\mathbf{X}_t^0 \mathbf{X}_t^0)^{-1} \mathbf{x}_t (y_t - \hat{\boldsymbol{\beta}}_{t-1} \mathbf{x}_t) \quad (22.36)$$

(see Chapter 21) might exemplify parameter time dependence given that  $(y_t - \hat{\boldsymbol{\beta}}_{t-1} \mathbf{x}_t)$ ,  $t > k$ , is no longer a mean innovation process but varies systematically with  $t$ . Hence, detecting parameter dependence using the recursive estimator  $\hat{\boldsymbol{\beta}}_t$ ,  $t > k$ , should be interpreted with caution when invalid conditioning might be the cause of the erratic behaviour of

$$E(\hat{\boldsymbol{\beta}}_t - \hat{\boldsymbol{\beta}}_{t-1}), \quad t > k. \quad (22.37)$$

The probability distribution underlying (34) comes in the form of  $D(y_t/\mathbf{Z}_{t-1}^0, \mathbf{X}_t; \psi_1)$ , which is related to the original sequential decomposition

$$D(\mathbf{Z}^*; \psi) = \prod_{t=1}^T D(\mathbf{Z}_t/\mathbf{Z}_{t-1}, \dots, \mathbf{Z}_1; \psi), \quad (22.38)$$

$$D(\mathbf{Z}_t/\mathbf{Z}_{t-1}^0; \boldsymbol{\psi}) = D(y_t/\mathbf{Z}_{t-1}^0, \mathbf{X}_t; \boldsymbol{\psi}_1) \cdot D(\mathbf{X}_t/\mathbf{Z}_{t-1}^0; \boldsymbol{\psi}_2). \quad (22.39)$$

The parameters of interest  $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}_i, \alpha_{i+1}, i=0, 1, \dots, m, \sigma^2)$  are functions of  $\boldsymbol{\psi}_1$  and  $\mathbf{X}_t$  remains weakly exogenous with respect to  $\boldsymbol{\theta}$  because  $\boldsymbol{\psi}_1$  and  $\boldsymbol{\psi}_2$  are variation free (see Chapter 19). Hence, the estimation of  $\boldsymbol{\theta}$  can be based on  $D(y_t/\mathbf{Z}_{t-1}^0, \mathbf{X}_t; \boldsymbol{\psi}_1)$  only. For prediction purposes, however, the fact that  $D(\mathbf{X}_t/\mathbf{Z}_{t-1}^0; \boldsymbol{\psi}_2)$  involves  $\mathbf{Y}_{t-1}^0$  cannot be ignored because of the feedback between these and  $\mathbf{X}_t$ . Hence, for prediction purposes in order to be able to concentrate exclusively on  $D(y_t/\mathbf{Z}_{t-1}^0, \mathbf{X}_t; \boldsymbol{\psi}_1)$  we need to assume that

$$D(\mathbf{X}_t/\mathbf{Z}_{t-1}^0; \boldsymbol{\psi}_2) = D(\mathbf{X}_t/\mathbf{X}_{t-1}^0; \boldsymbol{\psi}_2), \quad (22.40)$$

i.e.  $\mathbf{Y}_{t-1}^0$  does not *Granger cause*  $\mathbf{X}_t$  (see Engle *et al.* (1983)). When weak exogeneity is supplemented with Granger non-causality we say that  $\mathbf{X}_t$  is *strongly exogenous* with respect to  $\boldsymbol{\theta}$ .

The above changes to the linear regression model due to the non-random sample taken together amount to specifying a new statistical model which we call the *dynamic linear regression model*. Because of its importance in econometric modelling the specification, estimation, testing and prediction in the context of the dynamic linear regression model will not be considered here but in a separate chapter (see Chapter 23).

## (2) The autocorrelation approach

As argued above in the case of the autocorrelation approach the stochastic process  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is restricted even further than just stationary and asymptotic independent. In order to ensure (24) we need to restrict the temporal dependence among the components of  $\mathbf{Z}_t$  to be 'largely' identical, i.e.

$$\text{Cov}(Z_{ti} Z_{sj}) = \sigma_{ij} r(|t-s|), \quad i, j = 1, 2, \dots, k+1, \quad t, s \in \mathbb{T} \quad (22.41)$$

(see Spanos (1985a) for further details).

The question of restricting the time-heterogeneity and memory of the process arises in the context of the autocorrelation approach as restrictions on  $\sigma(t)$  and  $r(t, s)$  (see (27)). Indeed, looking at (28) we can see that  $\sigma^2(t)$  has already been restricted (implicitly),  $\sigma^2(t) = \sigma^2, t \in \mathbb{T}$ .

Assuming that  $\{\varepsilon_t, t \geq 1\}$  is a stationary process  $\mathbf{V}_T$  becomes a Toeplitz matrix (see Durbin (1960)) of the form  $v_{ts} = v_{|t-s|}$ ,  $t, s = 1, 2, \dots, T$  and although the number of unknown parameters is reduced the incidental parameters problem is not entirely solved unless some restrictions on the 'memory' of the process, such as ergodicity or mixing, are imposed. Hence, the same sort of restrictions as in the respecification approach are needed here as well. In practice, however, this problem is solved by postulating a

generating mechanism for  $\varepsilon_t$  which ensures both stationarity as well as some form of asymptotic independence (see Chapter 8). This mechanism (or model) is postulated to complement the statistical GM of the linear regression model under the independence assumption in order to take account of the non-independence. The most commonly used models for  $\varepsilon_t$  are the AR( $m$ ), MA( $m$ ) and ARMA( $p, q$ ) models discussed in Chapter 8. By far the most widely used autocorrelated error mechanism is the AR(1) model where  $\varepsilon_t = \rho \varepsilon_{t-1} + u_t$ , and  $u_t$  is a white-noise process. Taking this as a typical example, the statistical GM under the non-random sample assumption in the context of the autocorrelation approach becomes

$$y_t = \beta' \mathbf{x}_t + \varepsilon_t, \quad (22.42)$$

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t, \quad 0 < \rho < 1, \quad t \in \mathbb{T}. \quad (22.43)$$

The effect of postulating (43) in order to supplement (42) is to reduce the number of unknown parameters of  $\mathbf{V}_T$  to just one,  $\rho$ , which is time invariant as well. The temporal covariance matrix  $\mathbf{V}_T$  takes the form

$$\mathbf{V}_T(\rho) = \left( \frac{1}{1 - \rho^2} \right) \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & & & \rho^{T-2} \\ \rho^2 & & \ddots & & \vdots \\ \vdots & & & & \vdots \\ \rho^{T-1} & \rho^{T-2} & \cdots & & 1 \end{pmatrix} \quad (22.44)$$

(see Chapter 8). On the assumption that (43) represents an appropriate model for the dependency in the sample we can proceed to estimate the parameters of interest  $\theta \equiv (\beta, \rho, \sigma^2)$  where  $\sigma^2 = E(\varepsilon_t^2)$ .

The likelihood function based on the joint distribution under (42)–(43) is

$$\begin{aligned} L(\theta; \mathbf{y}) &= (2\pi\sigma^2)^{-T/2} (\det \mathbf{V}_T(\rho))^{-1/2} \\ &\times \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\beta)' \mathbf{V}_T(\rho)^{-1} (\mathbf{y} - \mathbf{X}\beta) \right\}. \end{aligned} \quad (22.45)$$

Using the fact that  $\det \mathbf{V}_T(\rho) = (1 - \rho^2)^{-T}$  the log likelihood is

$$\begin{aligned} \log L(\theta; \mathbf{y}) &= \text{const} - \frac{T}{2} \log \sigma^2 + \frac{1}{2} \log (1 - \rho^2) \\ &- \frac{1}{2\sigma^2} \left\{ (1 - \rho^2)\varepsilon_1 + \sum_{t=2}^T (\varepsilon_t - \rho\varepsilon_{t-1})^2 \right\}. \end{aligned} \quad (22.46)$$

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2} \mathbf{X}' \mathbf{V}_T(\rho)^{-1} \boldsymbol{\varepsilon} = \mathbf{0}, \quad (22.47)$$

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4} \boldsymbol{\varepsilon}' \mathbf{V}_T(\rho)^{-1} \boldsymbol{\varepsilon} = \mathbf{0}, \quad (22.48)$$

$$\frac{\partial \log L}{\partial \rho} = \frac{\rho}{(1-\rho^2)} + \frac{\rho \varepsilon_1}{\sigma^2} + \frac{1}{\sigma^2} \sum_{t=2}^T (\varepsilon_t - \rho \varepsilon_{t-1}) \varepsilon_{t-1} = 0, \quad (22.49)$$

where  $\boldsymbol{\varepsilon} \equiv \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ . Unfortunately, the first-order conditions  $(\partial \log L)/\partial \boldsymbol{\theta} = \mathbf{0}$  cannot be solved explicitly for  $\tilde{\boldsymbol{\theta}}$  the MLE of  $\boldsymbol{\theta}$  because they are not linear in  $\boldsymbol{\theta}$ . To derive  $\tilde{\boldsymbol{\theta}}$  we need to use some numerical optimisation procedure (see Beach and McKinnon (1978), Harvey (1981), *inter alia*). For a more extensive discussion of the estimation and testing in the context of the autocorrelation approach, see Judge *et al.* (1985).

### (3) The two approaches compared – the common factor restrictions

As mentioned in Section 22.1, the autocorrelation approach is a special case of the respecification approach. In order to show this let us consider the example of the statistical GM in the autocorrelation approach when the error mechanism is postulated to be an AR( $m$ ) process

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \cdots + \rho_m \varepsilon_{t-m} + u_t. \quad (22.50)$$

When this is substituted in  $y_t = \boldsymbol{\beta}' \mathbf{x}_t + \varepsilon_t$  the resulting hybrid statistical GM is

$$y_t = \boldsymbol{\beta}' \mathbf{x}_t + \sum_{i=1}^m \rho_i (y_{t-i} - \boldsymbol{\beta}' \mathbf{x}_{t-i}) + u_t. \quad (22.51)$$

If we compare this with the statistical GM (34) in the respecification approach we can see that the two are closely related. Indeed, (34) is identical to (51) under

$$H_0: \boldsymbol{\beta} \rho_i = -\boldsymbol{\beta}_i, \quad i = 1, 2, \dots, m. \quad (22.52)$$

These are called *common factor restrictions* (see Sargan (1964), Hendry and Mizon (1978), Sargan (1980)). This implies that in this case the model suggested by the autocorrelation approach is a special case of (34) with the common factors imposed a priori. The natural question which arises at this stage is whether this is a general result or is only true for this particular example. In Spanos (1985a) it was shown that in the case of a temporal covariance matrix  $\mathbf{V}_T$  based on a stationary ergodic process the hybrid statistical GM in the context of the autocorrelation approach takes the

general form

$$y_t = \beta' \mathbf{x}_t + \sum_{i=1}^m a_i(y_{t-i} - \beta' \mathbf{x}_{t-i}) + u_t. \quad (22.53)$$

Hence, the autocorrelation approach can be viewed as a special case of the respecification approach with the common factors restrictions imposed a priori.

Having established this relationship between the two approaches it is interesting to consider the question: ‘Why do the common factors restrictions arise naturally in the autocorrelation approach?’ The answer lies with the way the two approaches take account of the temporal dependence in the formulation of the non-random sample assumption.

In the respecification approach based on the statistical model specification procedure proposed in Chapter 17 any systematic information in the statistical model should be incorporated directly into the definition of the systematic component. The error term has to be non-systematic relative to the information set of the statistical model and represents the ‘unmodelled’ part of  $y_t$ . Hence, the systematic component needs to be redefined when temporal systematic information is present in the sampling model. On the other hand, the autocorrelation approach attributes the dependence in the sample to the covariance of the error terms retaining the same systematic component. The common factors arise because by modelling the error term the implicit conditioning is of the form

$$E(\varepsilon_t / \sigma(\mathbf{E}_{t-1}^0)) = \sum_{i=1}^m a_i \varepsilon_{t-i}, \quad (22.54)$$

where  $\mathbf{E}_{t-1}^0 \equiv (\varepsilon_{t-1}, \varepsilon_{t-2}, \dots, \varepsilon_0)$ . The implied statistical GM is

$$\varepsilon_t = E(\varepsilon_t / \sigma(\mathbf{E}_{t-1}^0)) + u_t \quad (22.55)$$

or

$$y_t - \beta' \mathbf{x}_t = \sum_{i=1}^m a_i(y_{t-i} - \beta' \mathbf{x}_{t-i}) + u_t, \quad (22.56)$$

which is identical to (53). Hence, the common factors are the result of ‘modelling’ the dependence in the sample in terms of the error term and not in terms of the observable random variables directly. The result of both approaches is a more general statistical GM which ‘models’ the dependence in the sample in two different but related ways. The restrictiveness of the autocorrelation approach can be seen by relating the common factor restrictions (52) to the parameters of the original AR( $m$ ) representation of

$\{\mathbf{Z}_t, t \in \mathbb{T}\}$  based on the following sequential conditional distribution:

$$(\mathbf{Z}_t / \mathbf{Z}_{t-1}^0) \sim N \left( \sum_{i=1}^m \begin{pmatrix} a_{11}(i) & \mathbf{a}_{12}(i) \\ \mathbf{a}_{21}(i) & \mathbf{A}_{22}(i) \end{pmatrix} \begin{pmatrix} y_{t-i} \\ \mathbf{X}_{t-i} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\omega}_{11} & \boldsymbol{\omega}_{12} \\ \boldsymbol{\omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix} \right). \quad (22.57)$$

As shown in the appendix, the parameters of the statistical GM (4) are related to the above parameters via the following equations:

$$\boldsymbol{\beta}_0 = \boldsymbol{\Omega}_{22}^{-1} \boldsymbol{\omega}_{21}, \quad \boldsymbol{\beta}_i = \{\mathbf{a}_{12}(i) + \boldsymbol{\omega}_{12} \boldsymbol{\Omega}_{22}^{-1} \mathbf{A}_{22}(i)\} \quad (22.58)$$

and

$$\alpha_i = \{a_{11}(i) + \boldsymbol{\omega}_{11} \boldsymbol{\Omega}_{22}^{-1} \mathbf{a}_{21}(i)\} \quad \text{for } i = 1, 2, \dots, m.$$

Hence the common factor restrictions hold when

$$\mathbf{a}_{12}(i) = \mathbf{a}'_{21}(i) = \mathbf{0} \quad \text{and} \quad \mathbf{A}_{22}(i) = a_{11}(i) \mathbf{I} \quad \text{for all } i = 1, \dots, m. \quad (22.59)$$

That is, the common factor restrictions hold when Granger non-causality holds among all  $Z_{it}$ s and an identical form of temporal self-dependence exists for all  $Z_{it}$ s,  $i = 1, \dots, m$ ,  $t > m$  (see Spanos (1985a)). These are very unrealistic restrictions to impose a priori. In principle the common factor restrictions can be tested indirectly by testing (59) in the context of the general AR( $m$ ) representation.

A direct test for these restrictions can be formulated as a specification test in the context of the respecification approach. In order to illustrate how the common factor restrictions can be tested let us return to the money equation estimated in Chapter 19 and consider the case where  $m = 1$ . The statistical GM of the money equation for the respecification and autocorrelation approaches is

$$m_t = \beta_1 + \beta_2 y_t + \beta_3 p_t + \beta_4 i_t + \alpha_1 m_{t-1} + \alpha_2 y_{t-1} + \alpha_3 p_{t-1} + \alpha_4 i_{t-1} + u_t \quad (22.60)$$

and

$$m_t = \beta_1^* + \beta_2^* y_t + \beta_3^* p_t + \beta_4^* i_t + e_t, \quad e_t = \alpha_1 e_{t-1} + u_t, \quad |\alpha_1| < 1 \quad (22.61)$$

respectively. If we use the lag operator  $L z_t = z_{t-i}$ ,  $i = 1, 2, \dots$ , we can express (60) in the form:

$$(1 - \alpha_1 L) m_t = \beta_1 + \beta_2 \left( 1 + \frac{\alpha_2}{\beta_2} L \right) y_t + \beta_3 \left( 1 + \frac{\alpha_3}{\beta_3} L \right) p_t + \beta_4 \left( 1 + \frac{\alpha_4}{\beta_4} L \right) i_t + u_t. \quad (22.62)$$

Under

$$H_0: \alpha_1 = -\frac{\alpha_2}{\beta_2}, \quad \alpha_1 = -\frac{\alpha_3}{\beta_3}, \quad \alpha_1 = -\frac{\alpha_4}{\beta_4}$$

we can see that the two sides of (62) have the *common factor*  $(1 - \alpha_1 L)$  which can be eliminated by dividing both sides by the common factor. This will give rise to (61).

The null hypothesis  $H_0$  is tested against

$$H_1: \alpha_1 \neq -\frac{\alpha_2}{\beta_2} \quad \text{or} \quad \alpha_1 \neq -\frac{\alpha_3}{\beta_3} \quad \text{or} \quad \alpha_1 \neq -\frac{\alpha_4}{\beta_4}.$$

Although the Wald test procedure (see Chapter 16) is theoretically much more attractive, given that estimation under  $H_1$  is considerably easier (see Mizon (1977), Sargan (1980) on the Wald tests), in our example the likelihood ratio test is more convenient because most computer packages provide the log likelihood. The rejection region based on the asymptotic likelihood ratio test statistic (see Chapter 16),

$$-2 \log_e \lambda(\mathbf{y}) = 2 \left\{ \log_e L(\hat{\boldsymbol{\theta}}; \mathbf{y}) - \log_e L(\tilde{\boldsymbol{\theta}}; \mathbf{y}) \right\} \underset{\alpha}{\sim} \chi^2(k-1), \quad (22.63)$$

where  $\hat{\boldsymbol{\theta}}$  and  $\tilde{\boldsymbol{\theta}}$  refer to the MLE's of  $\boldsymbol{\theta}$  under  $H_1$  and  $H_0$ , respectively, takes the form

$$C_1 = \{ \mathbf{y}: -2 \log_e \lambda(\mathbf{y}) \geq c_\alpha \}, \quad \alpha = \int_{c_\alpha}^\infty d\chi^2(k-1). \quad (22.64)$$

Estimation of (60) for the period 1963ii–1982iv yielded

$$\begin{aligned} m_t &= -0.766 + 0.793m_{t-1} + 0.038y_t + 0.240y_{t-1} + 0.023p_t \\ &\quad (0.582) (0.060) \quad (0.169) \quad (0.182) \quad (0.208) \\ &\quad + 0.160p_{t-1} - 0.041i_t + 0.006i_{t-1} + \hat{u}_t, \\ &\quad (0.220) \quad (0.012) \quad (0.013) \quad (0.018) \end{aligned} \quad (22.65)$$

$$R^2 = 0.999, \quad \bar{R}^2 = 0.999, \quad s = 0.0181,$$

$$\log L = 209.25, \quad T = 79.$$

Estimation of (61) for the same period yielded

$$m_t = 4.196 + 0.561y_t + 0.884p_t - 0.040i_t + \hat{e}_t, \quad \hat{e}_t = 0.819\hat{e}_{t-1} + \hat{u}_t, \quad (1.53) \quad (0.158) \quad (0.037) \quad (0.013) \quad (0.064) \quad (0.022)$$

$$R^2 = 0.998, \quad \bar{R}^2 = 0.998, \quad s = 0.0223, \quad \log L = 187.73, \quad T = 79. \quad (22.66)$$

Hence,  $-2 \log_e \lambda(\mathbf{y}) = 43.04$ . Given that  $c_\alpha = 7.815$  for  $\alpha = 0.05$  and three degrees of freedom,  $H_0$  is strongly rejected.

As mentioned above, the validity of the result of a common factor test

depends on the appropriateness of the statistical GM postulated for the general model as part of a well-defined estimated statistical model. The question of ensuring this is extensively discussed in the next chapter.

### 22.3 Testing the independent sample assumption

#### (1) *The respecification approach*

In Section 22.1 it was argued that the statistical results related to the linear regression model (see Chapter 19) under the independent sample assumptions are invalidated by the non-independence of the sample. For this reason it is of paramount importance to be able to test for independence.

As argued in Section 22.2, the statistical GM takes different forms when the sampling model is independent or asymptotically independent, that is,

$$y_t = \beta' x_t + u_t, \quad t = 1, 2, \dots, T \quad (22.67)$$

and

$$y_t = \beta_0 x_t + \sum_{i=1}^m (\alpha_i y_{t-i} + \beta'_i x_{t-i}) + u_t, \quad t = 1, 2, \dots, T \quad (22.68)$$

respectively. This implies that a test of independence can be constructed based on the significance of the parameters  $\alpha \equiv (\alpha_1, \dots, \alpha_m)', \beta^* \equiv (\beta'_1, \beta'_2, \dots, \beta'_m)',$  i.e.

$$H_0: \alpha = 0 \quad \text{and} \quad \beta^* = 0$$

against

$$H_1: \alpha \neq 0 \quad \text{or} \quad \beta^* \neq 0.$$

In view of the linearity of the restrictions an *F*-test type procedure (see Chapters 19 and 20) should provide us with a reasonable test. The discussion of testing linear hypotheses in Chapter 20 suggests the statistic:

$$\tau(y) = \frac{RSS - URSS}{URSS} \left( \frac{T - k(m+1)}{mk} \right). \dagger \quad (22.69)$$

$mkt(y)$  has an asymptotic chi-square distribution ( $\chi^2(mk)$ ) under  $H_0.$  In small samples, however, it might be preferable to use the *F* distribution approximation ( $F(mk, T - k(m+1))$ ) even though it is only asymptotically

<sup>†</sup> Note that one of the  $k$  regressors is always the constant term.

## 512 Departures from the sampling model assumption

justifiable. This is because the statistic  $\tau^*(y) = mk\tau(y)$  increases with the number of regressors; a feature which is particularly problematical in the present case because of the choice of  $m$ . On the other hand, the test statistic (69) does not necessarily increase when  $m$  increases. In practice we use the test based on the rejection region  $C_1 = \{y: \tau(y) > c_\alpha\}$ , where  $c_\alpha$  is defined by

$$\alpha = \int_{c_\alpha}^{\infty} dF(mk, T - k(m + 1)) \quad (22.70)$$

'as if' it were a finite sample test.

Let us consider this test for the money equation estimated in Chapter 19 with  $m=4$  using the estimation period 1964*i*-1982*iv*,

$$m_t = 2.763 + 0.705y_t + 0.862p_t - 0.053i_t + \hat{u}_t, \quad (22.71)$$

(1.10)	(0.112)	(0.022)	(0.014)	(0.040)
--------	---------	---------	---------	---------

$$R^2 = 0.995, \bar{R}^2 = 0.995, s = 0.04022,$$

$$\log L = 138.425, \quad RSS = 0.1165, \quad T = 76,$$

$$m_t = 0.706 + 0.589m_{t-1} - 0.018m_{t-2} - 0.046m_{t-3} + 0.214m_{t-4} \quad (22.72)$$

(0.815)	(0.132)	(0.152)	(0.166)	(0.129)
---------	---------	---------	---------	---------

$$+ 0.191y_t + 0.518y_{t-1} - 0.253y_{t-2} - 0.116y_{t-3} - 0.022y_{t-4} \quad (0.199) \quad (0.261) \quad (0.255) \quad (0.260) \quad (0.223)$$

$$- 0.060p_t + 0.606p_{t-1} - 0.381p_{t-2} + 0.558p_{t-3} - 0.479p_{t-4} \quad (0.348) \quad (0.670) \quad (0.642) \quad (0.630) \quad (0.348)$$

$$- 0.047i_t + 0.017i_{t-1} - 0.025i_{t-2} + 0.006i_{t-3} - 0.018i_{t-4} \quad (0.014) \quad (0.020) \quad (0.022) \quad (0.021) \quad (0.014)$$

$$+ \hat{u}_t, \quad (0.018) \quad (22.72)$$

$$R^2 = 0.999, \bar{R}^2 = 0.999, s = 0.01778,$$

$$\log L = 210.033, \quad RSS = 0.017697, \quad T = 76.$$

The test statistic (69) takes the value  $\tau(y) = 19.25$ . Given that  $c_\alpha = 1.812$  for  $\alpha = 0.05$  and degrees of freedom (16, 56) we can deduce that  $H_0$  is strongly rejected. That is, the independence assumption is invalid. This confirms our initial reaction to the time path of the residuals  $\hat{u}_t = y_t - \hat{\beta}'x_t$  (see Fig. 19.3) that some time dependency was apparently present.

An asymptotically equivalent test to the  $F$ -test considered above which corresponds to the *Lagrange multiplier* (see Chapters 16 and 20) test can be based on the statistic

$$LM(y) = TR^2 \stackrel{H_0}{\sim} \chi^2(mk), \quad (22.73)$$

where the  $R^2$  comes from the auxiliary regression

$$\hat{u}_t = (\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})' \mathbf{x}_t + \sum_{i=1}^m (\alpha_i y_{t-i} + \boldsymbol{\beta}'_i \mathbf{x}_{t-i}) + \varepsilon_t, \quad t = m+1, \dots, T. \quad (22.74)$$

In the case of the money equation the  $R^2$  for this auxiliary regression is 0.848 which implies that  $LM(\mathbf{y}) = 64.45$ . Given that  $c_\chi = 26.926$  for  $\alpha = 0.05$  and 16 degrees of freedom,  $h_0$  is again strongly rejected.

It is interesting to note that  $LM(\mathbf{y})$  can be expressed in the form:

$$LM(\mathbf{y}) = TR^2 = T \left( \frac{RRSS - URSS}{RRSS} \right) \quad (22.75)$$

(see exercise 4). This form of the test statistic suggests that the test suffers from the same problem as the one based on the statistic  $\tau^*(\mathbf{y})$ , given that  $R^2$  increases with the number of regressors (see Section 19.4).

## (2) The autocorrelation approach

As argued in Section 22.3 above, tackling the non-independence of the sample in the context of the autocorrelation approach before testing the appropriateness of the implied common factors is not the correct strategy to adopt. In testing the common factors, implied by adopting an error autocorrelation formulation, however, we need to refer back to the respecification approach. Hence, the question arises, ‘how useful is a test of the independence assumption in the context of the autocorrelation approach given that the test is based on an assumption which is likely to be erroneous?’ In order to answer this question it is instructive to compare the statistical GM’s of the two approaches:

$$y_t = \boldsymbol{\beta}'_0 \mathbf{x}_t + \sum_{i=1}^m \alpha_i y_{t-i} + \sum_{i=1}^m \boldsymbol{\beta}'_i \mathbf{x}_{t-i} + u_t, \quad t > m, \quad (22.76)$$

and

$$y_t = \boldsymbol{\beta}' \mathbf{x}_t + \varepsilon_t, \quad a(L)\varepsilon_t = b(L)u_t, \quad (22.77)$$

where  $a(L)$  and  $b(L)$  are  $p$ th- and  $q$ th-order polynomials in  $L$ . That is, the postulated model for the error term is an ARMA( $p, q$ ) time series formulation (see Section 8.4).

The error term  $\varepsilon_t$  interpreted in the context of (76) takes the form

$$\varepsilon_t = (\boldsymbol{\beta}_0 - \boldsymbol{\beta})' \mathbf{x}_t + \sum_{i=1}^m [\alpha_i y_{t-i} + \boldsymbol{\beta}'_i \mathbf{x}_{t-i}] + u_t, \quad t > m. \quad (22.78)$$

That is,  $\varepsilon_t$  is a linear function of the normal, stationary and asymptotically independent process  $\{\mathbf{Z}_{t-i}, i = 1, 2, \dots, m\}$ . This suggests that  $\{\varepsilon_t, t > m\}$  as

defined in (78) is itself a normal, stationary and asymptotically independent process. Such a process, however, can always be approximated to any degree of approximation by an ARMA( $p, q$ ) stationary process with ‘large enough’  $p$  and  $q$  (see Hannan (1970), Rosanov (1967), *inter alia*). In view of this we can see that testing for departures from the independent sample assumption in the context of the autocorrelation approach is not unreasonable. What could be very misleading is to interpret the rejection of the independence assumption as an ‘endorsement’ of the error autocorrelation model postulated by the misspecification test. Such an interpretation is totally unwarranted.

An interesting feature of the comparison between (76) and (77) is that of the role of the coefficients of  $\mathbf{x}_t$ ,  $\boldsymbol{\beta}_0$  and  $\boldsymbol{\beta}$ , are equal only when the common factor restrictions are valid. In such a case estimation of  $\boldsymbol{\beta}$  in  $y_t = \boldsymbol{\beta}'\mathbf{x}_t + \varepsilon_t$  should yield the same estimate as the estimate of  $\boldsymbol{\beta}$  in

$$y_t = \boldsymbol{\beta}'\mathbf{x}_t + \varepsilon_t, \quad a(L)\varepsilon_t = b(L)u_t. \quad (22.79)$$

Hence, a crude ‘test’ of the appropriateness of the implicitly imposed common factor restrictions might be to compare the two estimates of  $\boldsymbol{\beta}$  in the context of the autocorrelation approach. Such a comparison might be quite useful in cases where one is reading somebody else’s published work and there is no possibility of testing the common factor restrictions directly.

In view of the above discussion we can conclude that tests of the sample independence assumption in the context of the autocorrelation approach do have a role to play in misspecification testing related to the linear regression model, in so far as they indicate that the residuals  $\hat{\varepsilon}_t$ ,  $t = 1, \dots, T$ , do not constitute a realisation of a white-noise process. For this reason we are going to consider some of these tests in the light of the above discussion. In particular the emphasis will be placed on the non-parametric aspects of these tests. That is, the particular form of the error autocorrelation (AR(1), MA(1), ARMA( $p, q$ )) will be less crucial in the discussion. In a certain sense these autocorrelation based tests will be viewed in the context of the auxiliary regression

$$\hat{u}_t = \boldsymbol{\delta}'\mathbf{x}_t + \sum_{i=1}^m \rho_i \hat{u}_{t-i} + v_t, \quad t > m \quad (22.80)$$

which constitutes a special case of (74).

The particular aspect of the process  $\{\varepsilon_t, t \in \mathbb{T}\}$  we are interested in is its temporal structure. The null hypothesis of interest in the present context is that  $\{\varepsilon_t, t \in \mathbb{T}\}$  is a white-noise process (uncorrelated over time) and the alternative is that it is a dependent stochastic process. The natural way to

proceed in order to construct tests for such a hypothesis is to choose a ‘measure’ of temporal association among the  $\varepsilon_t$ s and devise a procedure to determine whether the estimated temporal associations are significant or not.

The most obvious measure of temporal dependence is the correlation between  $\varepsilon_t$  and  $\varepsilon_{t+l}$ , what we call *l*th-order *autocorrelation* defined by

$$r_l = \frac{\text{Cov}(\varepsilon_t \varepsilon_{t+l})}{[\text{Var}(\varepsilon_t) \text{Var}(\varepsilon_{t+l})]^{1/2}}, \quad l \geq 1. \quad (22.81)$$

In the case where  $\{\varepsilon_t, t \in \mathbb{T}\}$  is also assumed to be stationary ( $\text{Var}(\varepsilon_t) = \text{Var}(\varepsilon_{t+l})$ ) this becomes

$$r_l = \frac{\text{Cov}(\varepsilon_t \varepsilon_{t+l})}{\text{Var}(\varepsilon_t)}, \quad l \geq 1. \quad (22.82)$$

Note that  $-1 \leq r_l \leq 1$ . The natural estimator of  $r_l$  in the present context is

$$\hat{r}_l = \left\{ \left( \sum_{t=l+1}^T \hat{\varepsilon}_t \hat{\varepsilon}_{t+l} \right) \Big/ \left( \sum_{t=1}^T \hat{\varepsilon}_t^2 \right) \right\}, \quad l \geq 1. \quad (22.83)$$

Intuition suggests that a test for the sample independence assumption in the present context should consider whether the values of  $\hat{r}_l$ , for some  $l = 1, 2, \dots, m$ , say, are significantly different from zero. In the next subsection we consider tests based on  $l = 1$  and then generalise the results to  $l = m > 1$ .

### *The Durbin–Watson test ( $l = 1$ )*

The most widely used (and misused) misspecification test for the independence assumption in the context of the autocorrelation approach is the so-called Durbin–Watson test. The postulated statistical GM for the purposes of this test is

$$y_t = \beta' \mathbf{x}_t + \varepsilon_t, \quad \varepsilon_t = \rho \varepsilon_{t-1} + u_t, \quad |\rho| < 1, \quad t \in \mathbb{T}. \quad (22.84)$$

The null hypothesis of interest is  $H_0: \rho = 0$  (i.e.  $\varepsilon_t = u_t$ , white noise) against the alternative  $H_1: \rho \neq 0$ . Building on the work of Anderson (1948), Durbin and Watson (1950, 1951) proposed the test statistic,

$$\tau_1(\mathbf{y}) = \frac{\hat{\varepsilon}' \mathbf{A}_1 \hat{\varepsilon}}{\hat{\varepsilon}' \hat{\varepsilon}} = \frac{\sum_{t=2}^T (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^T \hat{\varepsilon}_t^2}, \quad (22.85)$$

where

$$\mathbf{A}_1 = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ 0 & -1 & 2 & -1 & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & \cdots & -1 & 1 \end{bmatrix}. \quad (22.86)$$

The relationship between  $\mathbf{A}_1$  and  $\mathbf{V}_T$  is given by

$$\mathbf{V}_T(\rho)^{-1} = (1-\rho)^2 \mathbf{I} + \rho \mathbf{A}_1 + \rho(1-\rho) \mathbf{C}, \quad (22.87)$$

where  $\mathbf{C} = \text{diag}(1, 0, \dots, 0, 1)$ . Durbin and Watson used the approximation

$$\mathbf{V}_T(\rho)^{-1} \approx (1-\rho)^2 \mathbf{I} + \rho \mathbf{A}_1, \quad (22.88)$$

which enabled them to use Anderson's result based on a known temporal covariance matrix. As we can see from (88), when  $\rho=0$ ,  $\mathbf{V}_T(\rho)^{-1} = \mathbf{I}_T$ . The rejection region for the null hypothesis

$$H_0: \rho=0, \quad \text{against} \quad H_1: \rho \neq 0$$

takes the form

$$C_1 = \{y: \tau_1(y) \leq c_z\}, \quad (22.89)$$

where  $c_z$  refers to the critical value for a size  $\alpha$  test, determined by the distribution of the test statistic (85) under  $H_0$ . This distribution, however, is inextricably bound up with the observed data matrix  $\mathbf{X}$  given that  $\hat{\boldsymbol{\epsilon}} = \mathbf{M}_x \boldsymbol{\epsilon}$ ,  $\mathbf{M}_x = I_T - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  (see Chapter 19) and

$$\tau_1(y) = \frac{\boldsymbol{\epsilon}' \mathbf{M}_x \mathbf{A}_1 \mathbf{M}_x \boldsymbol{\epsilon}}{\boldsymbol{\epsilon}' \mathbf{M}_x \boldsymbol{\epsilon}}. \quad (22.90)$$

This implies that the Durbin–Watson test is *data specific* and the distribution of  $\tau_1(y)$  needs to be evaluated for each  $\mathbf{X}$ . In order to make the evaluation of this distribution easier to handle Durbin and Watson, using the fact that  $\mathbf{M}_x$  and  $\mathbf{M}_x \mathbf{A}_1 \mathbf{M}_x$  commute (i.e.  $\mathbf{M}_x (\mathbf{M}_x \mathbf{A}_1 \mathbf{M}_x) = (\mathbf{M}_x \mathbf{A}_1 \mathbf{M}_x) \mathbf{M}_x$ ) and  $\mathbf{M}_x$  is an idempotent matrix, suggested using an orthogonal matrix  $\mathbf{H}$  which diagonalises both quadratic forms in (90) simultaneously, i.e.

$$\tau_1(y) = \frac{\sum_{i=1}^{T-k} v_i \xi_i^2}{\sum_{i=1}^{T-k} \xi_i^2}, \quad (22.91)$$

where  $v_1, v_2, \dots, v_{T-k}$  are the non-zero eigenvalues of  $\mathbf{M}_X \mathbf{A}_1$  and

$$\xi = \mathbf{H}' \varepsilon \sim N(\mathbf{0}, \sigma^2 I_T). \quad (22.92)$$

Hence the value  $c_x$  can be evaluated by ‘solving’ the following probabilistic statement based on (91) for  $c_x$ :

$$Pr(\tau_1(\mathbf{y}) \leq c_x) = Pr\left(\sum_{i=1}^{T-k} (v_i - c_x) \xi_i^2 \leq 0\right) = \alpha \quad (22.93)$$

for a given size  $\alpha$ .

Hannan (1970), however, suggested that in practice there was no need to evaluate  $c_x$ . Instead we could evaluate

$$p(\mathbf{y}) = Pr\left(\sum_{i=1}^{T-k} (v_i - \tau_1(\mathbf{y})) \xi_i^2 \leq 0\right), \quad (22.94)$$

and if this probability exceeds  $\alpha$  we accept  $H_0$ , otherwise we reject  $H_0$  in favour of  $H_1^+ : \rho > 0$ . For  $H_1^- : \rho < 0$  as the alternative  $4 - \tau_1(\mathbf{y})$  should be used in place of  $\tau_1(\mathbf{y})$ . The value 4 stems from the fact that

$$\tau_1(\mathbf{y}) \simeq 2(1 - \hat{r}_1) \quad \text{where } \hat{r}_1 = \frac{\sum_{t=2}^T \hat{\varepsilon}_t \hat{\varepsilon}_{t-1}}{\sum_{t=1}^T \hat{\varepsilon}_t^2} \quad (22.95)$$

is the first-order residual correlation coefficient and takes values between  $-1$  and  $1$ ; hence  $0 \leq \tau_1(\mathbf{y}) < 4$ .

In the case of the estimated money equation discussed above the Durbin–Watson test statistic takes the value

$$\tau_1(\mathbf{y}) = 0.376. \quad (22.96)$$

For this value  $p(\mathbf{y}) = 0.000$  and hence  $H_0: \rho = 0$  is strongly rejected for any size  $\alpha$  test. The question which arises is, ‘how do we interpret the rejection of  $H_0$  in this case?’ ‘Can we use it as an indication that the appropriate statistical GM is not  $y_t = \beta' \mathbf{x}_t + u_t$  but (84)?’ The answer is, certainly not. In no circumstances should we interpret the rejection of  $H_0$  against some specific form of departure from independence as a confirmation of the validity of the alternative in the context of the autocorrelation approach. This is because in a misspecification testing framework rejection of  $H_0$  should never be interpreted as equivalent to acceptance of  $H_1$ ; see Davidson and McKinnon (1985).

In the case where the evaluation of (93) is not possible, Durbin and Watson (1950, 1951) using the eigenvalues of  $\mathbf{A}_1$ , proposed a bounds test which does not depend on the particular observed data matrix  $\mathbf{X}$ . That is,

they derived  $d_L$  and  $d_U$  such that for any  $\mathbf{X}$ ,

$$d_L \leq \tau_1(\mathbf{y}) \leq d_U, \quad (22.97)$$

where  $d_L$  and  $d_U$  are *independent of  $\mathbf{X}$* . This led them to propose the bounds test for

$$H_0: \rho = 0 \quad \text{against} \quad H_1^+: \rho > 0 \quad (22.98)$$

based on

$$C_1 = \{\mathbf{y}: \tau_1(\mathbf{y}) \leq d_L\} \quad \text{and} \quad C_0 = \{\mathbf{y}: \tau_1(\mathbf{y}) \geq d_U\}. \quad (22.99)$$

In the case where  $d_L \leq \tau_1(\mathbf{y}) \leq d_U$  the test is inconclusive (see Maddala (1977) for a detailed discussion of the inconclusive region). For the case  $H_0: \rho = 0$  against  $H_1: \rho < 0$  the test statistic  $4 - \tau_1(\mathbf{y})$  should be used in (99).

In view of the discussion at the beginning of this section the Durbin–Watson (DW) test as a test of the independent sample assumption is useful in so far as it is based on the first-order autocorrelation coefficient  $r_1$ . Because of the relationship (95) it is reasonable to assume that the test will have adequate power against other forms of first-order dependence such as MA(1) (see King (1983) for an excellent survey of the DW and related tests). Hence, in practice the test should be used not as a test related to an AR(1) error autocorrelation only but as a general first-order dependence test. Moreover, the DW-test is likely to have power against higher-order dependence in so far as the first order autocorrelation coefficient ‘captures’ part of this temporal dependence.

### *Higher-order tests*

Given that estimation of the linear regression model is easier to handle under the independent sample assumption rather than when supplemented with an autocorrelation error model, it should come as no surprise to discover that the Lagrange multiplier test procedure (see Chapter 16) provides the most convenient method for constructing asymptotic misspecification tests for sample independence.

In order to illustrate the derivation of misspecification tests for higher-order dependencies among the  $y_t$ s let us consider the *LM* test procedure for the simple AR(1) case which generalises directly to an AR( $m$ ) as well as a moving-average (MA( $m$ )). Postulating an AR(1) form of dependency among the  $y_t$ s is equivalent to changing the statistical GM for the linear regression model to

$$y_t = \boldsymbol{\beta}' \mathbf{x}_t + \rho y_{t-1} - \rho \boldsymbol{\beta}' \mathbf{x}_{t-1} + u_t, \quad (22.100)$$

as can be easily verified from (84). The null hypothesis is  $H_0: \rho = 0$  against  $H_1: \rho \neq 0$ . Because the estimation of (100) under  $H_0$  is much easier than its estimation under  $H_1$  the *LM*-test procedure is computationally preferable to both the Wald and the likelihood ratio test procedures.

The efficient score form of the *LM*-test statistic is

$$LM(\mathbf{y}) = \left( \frac{\partial \log L(\hat{\boldsymbol{\theta}}; \mathbf{y})}{\partial \boldsymbol{\theta}} \right), \quad \mathbf{I}_T(\tilde{\boldsymbol{\theta}})^{-1} \left( \frac{\partial \log L(\tilde{\boldsymbol{\theta}}; \mathbf{y})}{\partial \boldsymbol{\theta}} \right). \quad (22.101)$$

In the case where  $H_0$  involves only a subset of  $\boldsymbol{\theta}$ , say  $H_0: \boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^0$ , where  $\boldsymbol{\theta} \equiv (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$  then the statistic takes the form

$$LM(\mathbf{y}) = \left( \frac{\partial \log L(\boldsymbol{\theta}_1^0, \tilde{\boldsymbol{\theta}}_2)}{\partial \boldsymbol{\theta}_1} \right)' (\tilde{\mathbf{I}}_{11} - \tilde{\mathbf{I}}_{12}\tilde{\mathbf{I}}_{22}^{-1}\tilde{\mathbf{I}}_{21})^{-1} \left( \frac{\partial \log L(\boldsymbol{\theta}_1^0, \tilde{\boldsymbol{\theta}}_2)}{\partial \boldsymbol{\theta}_1} \right) \quad (22.102)$$

(see Chapter 16).

In the present case the log likelihood function was given in equation (46) and for  $\boldsymbol{\theta}_1 = \rho$  and  $\boldsymbol{\theta}_2 \equiv (\boldsymbol{\beta}, \sigma^2)$  it can be shown that:

$$\mathbf{I}_T(\boldsymbol{\theta}) = \begin{pmatrix} \frac{T}{(1-\rho^2)} & \mathbf{0} & 0 \\ \mathbf{0} & \frac{(\mathbf{X}'\mathbf{V}_T^{-1}\mathbf{X})}{\sigma^2} & \mathbf{0} \\ 0 & \mathbf{0} & \frac{T}{2\sigma^2} \end{pmatrix}, \quad (22.103)$$

$$\left. \frac{\partial \log L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_1} \right|_{\substack{\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}} \\ \sigma^2 = \tilde{\sigma}^2 \\ \rho = 0}} = T \begin{pmatrix} \sum_{t=2}^T \hat{e}_t \hat{e}_{t-1} \\ \frac{\sum_{t=1}^{T-1} \hat{e}_t^2}{T} \end{pmatrix} = T \hat{r}_1, \quad (22.104)$$

$$\left. (\mathbf{I}_{11} - \mathbf{I}_{12}\mathbf{I}_{22}^{-1}\mathbf{I}_{21})^{-1} \right|_{\substack{\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}} \\ \sigma^2 = \tilde{\sigma}^2 \\ \rho = 0}} = \frac{1}{T} \quad (22.105)$$

and

$$LM(\mathbf{y}) = T \hat{r}_1^2 \sim \chi^2(1) \quad (22.106)$$

(see exercise 5). Hence, the *Lagrange multiplier test* is defined by the

rejection region

$$C_1 = \{\mathbf{y}: LM(\mathbf{y}) > c_x\} \quad \text{where } x = \int_{c_x}^{\infty} d\chi^2(1). \quad (22.107)$$

The form of the test statistic in (106) makes a lot of intuitive sense given that the first-order residual correlation coefficient is the best measure of first-order dependence among the residuals. Thus it should come as no surprise to learn that for  $\tau$ th-order temporal dependence, say

$$\varepsilon_t = \rho \varepsilon_{t-\tau} + u_t, \quad \tau \geq 1,$$

the *LM*-test statistic takes the form

$$LM(\mathbf{y}) = T \hat{r}_\tau^2 \quad (22.108)$$

where

$$\hat{r}_\tau = \left[ \left( \sum_{t=\tau+1}^T \hat{\varepsilon}_t \hat{\varepsilon}_{t-\tau} \right) / \sum_{t=1}^T \hat{\varepsilon}_t^2 \right], \quad \tau = 1, 2, \dots, m < T \quad (22.109)$$

(see Godfrey (1978), Breusch and Pagan (1980)).

This result generalises directly to higher-order autoregressive and moving-average error models. Indeed both forms of error models give rise to the same *LM*-test statistic. That is, for the statistical GM,  $y_t = \beta' \mathbf{x}_t + \varepsilon_t$ ,  $t = 1, 2, \dots, T$ , supplemented with either an AR( $m$ ) error process:

$$(i) \quad \varepsilon_t = \rho_1 \varepsilon_{t-1} + \dots + \rho_m \varepsilon_{t-m} + u_t, \quad (22.110)$$

or a MA( $m$ ) process:

$$(ii) \quad \varepsilon_t = u_t + a_1 u_{t-1} + \dots + a_m u_{t-m}, \quad (22.111)$$

the *LM*-test statistic for

$$H_0: \rho_i = 0 \quad (a_i = 0) \quad \text{for all } i = 1, 2, \dots, m$$

against

$$H_1: \rho_i \neq 0 \quad (a_i \neq 0) \quad \text{for any } i = 1, 2, \dots, m$$

takes the same form

$$LM(\mathbf{y}) = T \left( \sum_{\tau=1}^m \hat{r}_\tau^2 \right)^{H_0} \sim \chi^2(m), \quad (22.112)$$

with rejection region

$$C_1 = \{\mathbf{y}: LM(\mathbf{y}) \geq c_x\}, \quad x = \int_{c_x}^{\infty} d\chi^2(m) \quad (22.113)$$

(see Godfrey (1978), Godfrey and Wickens (1982)). The intuition underlying this result is that because of the approximations involved in the derivation

of the *LM*-test statistic the two underlying error processes cannot be distinguished asymptotically.

An asymptotically equivalent test, sometimes called the modified *LM*-test, can be based on the  $R^2$  of the auxiliary regression (see (80)):

$$\hat{e}_t = \boldsymbol{\delta}' \mathbf{x}_t + \gamma_1 \hat{e}_{t-1} + \cdots + \gamma_m \hat{e}_{t-m} + u_t. \quad (22.114)$$

That is,

$$TR^2 \stackrel{H_0}{\sim} \chi^2(m) \quad (22.115)$$

with the rejection region

$$C_1 = \{\mathbf{y}: TR^2 \geq c_\alpha\}, \quad \alpha = \int_{c_\alpha}^\infty d\chi^2(m). \quad (22.116)$$

This is a test for the joint significance of  $\gamma_1, \gamma_2, \dots, \gamma_m$ , similar to (112) above.

The auxiliary regression for the estimated money equation with  $m=6$  yielded:

$$TR^2 = 72(0.7149) = 51.47. \quad (22.117)$$

Given that  $c_\alpha = 12.592$  for  $\alpha = 0.05$  the null hypothesis  $H_0: \boldsymbol{\gamma} = \mathbf{0}$  is strongly rejected in favour of  $H_1: \boldsymbol{\gamma} \neq \mathbf{0}$ . Hence, the estimated money equation has failed every single misspecification test for independence showing clearly that the independent sample assumption was grossly inappropriate. The above discussion also suggests that the departures from linearity, homoskedasticity and parameter time invariance detected in Chapter 21 might be related to the dependence in the sample as well.

## 22.4 Looking back

In Chapters 20–22 we discussed the implications of certain departures from the assumptions underlying the linear regression model, how we can test for these assumptions as well as how we should proceed when these assumptions are invalid. In relation to the implications of these departures we have seen that the statistical results (estimation testing and prediction) related to the linear regression model (see Chapter 19) are seriously affected, with some departures, such as a non-random sample, invalidating these results altogether. This makes the testing for these departures particularly crucial for econometric modelling. This is because the first stage in the statistical analysis of an econometric model is the specification and estimation of a well-defined (no misspecifications) statistical model. Unless we start with a well-defined estimated statistical GM any deductions such as specification testing of a priori restrictions, economic interpretation of

estimated parameters, prediction and policy analysis, will be misleading, if not outright unwarranted, given that these conclusions will be based on erroneous statistical foundations.

When some misspecification is detected the general way to proceed is to *respecify* the statistical model in view of the departures from the underlying assumption(s). This sometimes involves the reconsideration of all the assumptions underlying the statistical model such as the case of the independent sample assumption.

One important disadvantage of the misspecification testing discussed in Chapters 20–22 is the fact that most of the departures were considered in isolation. A more appropriate procedure will be to derive joint misspecification tests. For such tests the auxiliary regressions test procedure discussed in Chapter 21 seems to provide the most practical way forward. In particular it is intuitively obvious that the best way to generate a ‘good’ misspecification test is to turn it into a specification test. That is, extend the linear regression model in the directions of possible departures from the underlying assumptions in a way which defines a new statistical model which contains the linear regression model as a special case; under the null that the underlying assumptions of the latter are valid (see Spanos (1985c)).

Once we have ensured that the underlying assumptions [1]–[8] of the linear regression model are valid for the data series chosen we call the estimated statistical GM with the underlying assumptions *a well-defined estimated statistical model*. This provides the starting point for reparametrisation/restriction and model selection in an attempt to construct an empirical econometric model (see Fig. 1.2). Using statistical procedures based on a well-defined estimated statistical model we can proceed to reparametrise the statistical GM in terms of the theoretical parameters of interest  $\xi$  which are directly related to the statistical parameters  $\theta$  via a system of equations of the form

$$\mathbb{G}(\theta, \xi) = \mathbf{0}. \quad (22.118)$$

The reparametrisation in terms of  $\xi$  is possible only when the system of equations provides a unique solution for  $\xi$  in terms of  $\theta$ :

$$\xi = \mathbf{H}(\theta) \quad (22.119)$$

(explicitly or implicitly). In such a case  $\xi$  is said to be *identified*. As we can see, the theoretical parameters of interest derive their statistical meaning from their relationship to  $\theta$ . In the case where there are fewer  $\xi_i$ s than  $\theta_i$ s the extra restrictions implied by (118) can (and should) be tested before being imposed.

It is important to note that there is a multitude of possible reparametrisations in terms of theoretical parameters of interest giving rise

to a number of possible empirical econometric models. This raises the question of *model selection* where various statistical as well as theory-oriented criteria can be used such as *theory consistency, parsimony, encompassing, robustness, and nearly orthogonal explanatory variables* (see Hendry and Richard (1983)), in order to select the ‘best’ empirical econometric model. This reparametrisation/restriction of the estimated statistical GM, however, should not be achieved at the expense of its statistical properties. The empirical econometric model needs to be a well-defined estimated statistical model itself in order to be used for prediction and policy analysis.

### Appendix 22.1 – deriving the conditional expectation

Assuming that

$$\begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \\ \vdots \\ \mathbf{Z}_T \end{pmatrix} \sim N \left( \begin{pmatrix} \mathbf{m}(1) \\ \mathbf{m}(2) \\ \vdots \\ \mathbf{m}(T) \end{pmatrix}, \begin{pmatrix} \Sigma(1, 1), & \Sigma(1, 2), & \dots, & \Sigma(1, T) \\ \Sigma(2, 1), & \Sigma(2, 2), & \dots, & \Sigma(2, T) \\ \vdots & & & \vdots \\ \Sigma(T, 1), & \dots, & \dots, & \Sigma(T, T) \end{pmatrix} \right),$$

we can deduce that

$$(\mathbf{Z}_t / \mathbf{Z}_{t-1}^0) \sim N \left( \mathbf{m}(t) + \sum_{i=1}^{t-1} \mathbf{A}_i(t)[\mathbf{Z}_{t-i} - \mathbf{m}(t-i)], \boldsymbol{\Omega}(t) \right),$$

For  $\mathbf{A}^*(t) = [\mathbf{A}_1(t), \mathbf{A}_2(t), \dots, \mathbf{A}_{t-1}(t)]$

$$\mathbf{A}^*(t) = \boldsymbol{\Sigma}_{t-1}^0 (\boldsymbol{\Sigma}_{t-1|t-1}^{00})^{-1},$$

$$\boldsymbol{\Omega}(t) = \Sigma(t, t) - \boldsymbol{\Sigma}_{t-1}^0 (\boldsymbol{\Sigma}_{t-1|t-1}^{00})^{-1} \boldsymbol{\Sigma}_{t-1|t}^0,$$

$$\boldsymbol{\Sigma}_{t-1}^0 \equiv (\Sigma(t, 1), \Sigma(t, 2), \dots, \Sigma(t, t-1)),$$

$$\boldsymbol{\Sigma}_{t-1|t-1}^{00} = [\Sigma(i, j)]_{i,j}, \quad i, j = 1, 2, \dots, t-1,$$

$$\boldsymbol{\Sigma}_{t-1|t}^0 \equiv (\Sigma(1, t)', \Sigma(2, t)', \dots, \Sigma(t-1, t)'),$$

$$\mathbf{Z}_{t-1}^0 \equiv (\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_{t-1}).$$

For

$$\mathbf{Z}_t \equiv \begin{pmatrix} y_t \\ \mathbf{X}_t \end{pmatrix}, \quad \mathbf{m}(t) = \begin{pmatrix} m_y(t) \\ \mathbf{m}_x(t) \end{pmatrix},$$

$$\mathbf{A}_i(t) = \begin{pmatrix} a_{11}(i, t), \mathbf{a}_{12}(i, t) \\ \mathbf{a}_{21}(i, t), \mathbf{A}_{22}(i, t) \end{pmatrix}, \quad \boldsymbol{\Omega}(t) = \begin{pmatrix} \omega_{11}(t) & \omega_{12}(t) \\ \omega_{21}(t) & \boldsymbol{\Omega}_{22}(t) \end{pmatrix},$$

$$(y_t / \mathbf{Z}_{t-1}^0, \mathbf{X}_t) \sim N \left( c_0(t) + \boldsymbol{\beta}'_0(t) \mathbf{x}_t + \sum_{i=1}^{t-1} [\alpha_i(t) y_{t-i} + \boldsymbol{\beta}'_i(t) \mathbf{x}_{t-i}], \sigma_0^2(t) \right),$$

$$c_0(t) = m_y(t) - \omega_{12}(t)\Omega_{22}^{-1}(t)\mathbf{m}_x(t)$$

$$\beta'_0(t) = \omega_{21}(t)\Omega_{22}(t)^{-1},$$

$$\alpha_i(t) = a_{11}(i, t) - \omega_{12}(t)\Omega_{22}^{-1}(t)\mathbf{a}_{21}(i, t),$$

$$\beta'_i(t) = \mathbf{a}_{12}(i, t) - \omega_{12}(t)\Omega_{22}^{-1}(t)\mathbf{A}_{22}(i, t).$$

### ***Important concepts***

Error autocorrelation, stationary process, ergodicity, mixing, an innovation process, a martingale difference process, strong exogeneity, common factor restrictions, Durbin–Watson test, well-defined estimated statistical GM, reparametrisation/restriction, model selection.

### ***Questions***

1. Compare the interpretation of non-independence in the context of the autocorrelation approach with that of the respecification approach.
2. Compare and contrast the implications of the respecification and autocorrelation approaches for the properties of the estimators  $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ ,  $s^2 = [1/(T-k)]\hat{\mathbf{u}}'\hat{\mathbf{u}}$ ,  $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\beta}$  and the  $F$ -test for  $H_0: \mathbf{R}\beta = \mathbf{r}$  against  $H_1: \mathbf{R}\beta \neq \mathbf{r}$ .
3. Explain the role of stationarity, in the context of the respecification approach, for the form of the statistical GM (34).
4. ‘The concept of a stationary stochastic process constitutes a generalisation of the concept of identically distributed random variables to stochastic processes.’ Discuss.
5. ‘Inappropriate conditioning due to the non-independence of the sample leads to time dependency of the parameters of interest  $\theta \equiv (\beta, \sigma^2)$ .’ Explain.
6. Compare the ways to tackle the problem of a non-random sample in the context of the respecification and autocorrelation approaches.
7. Explain why we do not need strong exogeneity for the estimation of the parameters of interest in (34).
8. Explain how the common factor restrictions arise in the context of the autocorrelation approach.
9. Discuss the merits of testing the independent sample assumption in the context of the respecification approach as compared with that of the autocorrelation approach.
10. ‘Rejecting the null hypothesis in a misspecification test for error independence in the context of the autocorrelation approach should not be interpreted as acceptance of the alternative.’ Discuss.

11. Explain the exact Durbin–Watson test.
12. ‘The Durbin–Watson test as a test of the independence assumption in the context of the autocorrelation approach is useful in so far as it is directly related to  $r_1$ .’ Discuss.
13. Discuss the Lagrange multiplier test for higher-order error autocorrelation (AR( $m$ ) and MA( $m$ )) and explain intuitively why the test is identical for both models.
14. State briefly the changes in the assumptions underlying the linear regression model brought about by the non-independence of the sample.
15. Explain the role of some asymptotic independence restriction on the memory of  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  in defining the statistical GM for a non-random sample.

### ***Exercises***

1. Verify the implications of a non-random sample assumption for  $\hat{\beta}$  and  $s^2$  as given by (i)–(vi) and (i)'–(vi)' of Section 22.1 in the context of the respecification and autocorrelation approaches, respectively.
2. Show that  $[\hat{c} \log L(\boldsymbol{\theta}; \mathbf{y})]/\partial \rho = 0$  where  $\log L(\boldsymbol{\theta}; \mathbf{y})$  given in equation (49) is non-linear in  $\rho$ .
3. Derive the Wald test for the common factor restriction in the case where  $k=2$  and  $m=1$ .
4. Verify the formula  $TR^2 = [(RRSS - URSS)/RRSS]T$  given in equation (75); see Engle (1984).
5. Derive the LM-test statistic for  $H_0: \rho=0$  against  $H_1: \rho \neq 0$  in the case where  $\varepsilon_t = \rho \varepsilon_{t-1} + u_t$  as given in equation (106).

### ***Additional references***

Anderson (1971); Godfrey (1981); Madansky (1976).

## CHAPTER 23

---

### The dynamic linear regression model

---

In Chapter 22 it was argued that the respecification of the linear regression model induced by the inappropriateness of the independent sample assumption led to a new statistical model which we called the dynamic linear regression (DLR) model. The purpose of the present chapter is to consider the statistical analysis (specification, estimation, testing and prediction) of the DLR model.

The dependence in the sample raises the issue of introducing the concept of *dependent random variables* or *stochastic processes*. For this reason the reader is advised to refer back to Chapter 8 where the idea of a stochastic process and related concepts are discussed in some detail before proceeding further with the discussion which follows.

The linear regression model can be viewed as a statistical model derived by reduction from the joint distribution  $D(\mathbf{Z}_1, \dots, \mathbf{Z}_T; \boldsymbol{\psi})$ , where  $\mathbf{Z}_t \equiv (\mathbf{y}_t, \mathbf{X})'$ , and  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is assumed to be a normal, independent and identically distributed (NIID) vector stochastic process. For the purposes of the present chapter we need to extend this to a more general stochastic process in order to take the dependence, which constitutes additional systematic information, into consideration.

In Chapter 21 the identically distributed component was relaxed leading to time varying parameters. The main aim of the present chapter is to relax the independence component but retain the identically distributed assumption in the form of stationarity.

In Section 23.1 the DLR is specified assuming that  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is a stationary, asymptotically independent normal process. Some of the arguments discussed in Chapter 22 in terms of the respecification approach will be considered more formally. Section 23.2 considers the estimation of the DLR using approximate MLE's. Sections 23.3 and 23.4 discuss

misspecification and specification testing in the context of the DLR model, respectively. Section 23.5 considers the problem of prediction. In Section 23.6 the empirical econometric model constructed in Section 23.4 is used to ‘explain’ the misspecification results derived in Chapters 19–22.

### 23.1 Specification

In defining the linear regression model  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  was assumed to be an independent and identically distributed (IID) multivariate normal process. In defining the dynamic linear regression (DLR) model  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is assumed to be a stationary, asymptotically independent normal process. That is, the assumption of identically distributed has been extended to that of stationarity and the independence assumption to that of *asymptotic independence* (see Chapter 8). In particular,  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is assumed to have the following autoregressive representation:

$$\mathbf{Z}_t = \sum_{i=1}^m \mathbf{A}(i)\mathbf{Z}_{t-i} + \mathbf{E}_t, \quad t > m, \quad (23.1)$$

where

$$E(\mathbf{Z}_t / \sigma(\mathbf{Z}_{t-1}^0)) = \sum_{i=1}^m \mathbf{A}(i)\mathbf{Z}_{t-i}, \quad (23.2)$$

$$\mathbf{Z}_{t-1}^0 = (\mathbf{Z}_{t-1}, \mathbf{Z}_{t-2}, \dots, \mathbf{Z}_1),$$

$$\mathbf{A}(i) = \begin{pmatrix} \mathbf{a}_{11}(i) & \mathbf{a}_{12}(i) \\ \mathbf{a}_{21}(i) & \mathbf{A}_{22}(i) \end{pmatrix}$$

and

$$\mathbf{E}_t = \mathbf{Z}_t - E(\mathbf{Z}_t / \sigma(\mathbf{Z}_{t-1}^0)), \quad t > m, \quad (23.3)$$

with  $\{\mathbf{E}_t, \sigma(\mathbf{Z}_{t-1}^0), t > m\}$  defining a vector martingale difference process, which is also an innovation process (see Chapter 8), such that

$$(\mathbf{E}_t / \mathbf{Z}_{t-1}^0) \sim N(\mathbf{0}, \boldsymbol{\Omega}), \quad \boldsymbol{\Omega} > \mathbf{0}, \quad t > m. \quad (23.4)$$

It is important to note that stationarity is not a necessary condition for the existence of the autoregressive representation (1). The existence of (1) with  $\mathbf{A}(i)$ ,  $i = 1, 2, \dots, m$ , and  $\boldsymbol{\Omega}$  being time invariant are the essential conditions for the argument which follows and both are possible without the stationarity of  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$ . For homogeneous non-stationary stochastic process (1) also exists but differencing is needed to achieve stationarity (see Box and Jenkins (1976)). *Homogeneous non-stationary processes* are

characterised by realisations which, apart from the apparent existence of a local trend, the time path seems similar in the various parts of the realisation. In such cases the differencing transformation

$$\Delta^i = (1 - L)^i, \quad \text{where } L^i X_t = X_{t-i}, \quad i = 1, 2, \dots$$

can be used to transform the original series to a stationary one (see Chapter 8).

The distribution underlying (1) is  $D(\mathbf{Z}_t / \mathbf{Z}_{t-1}^0; \psi)$  arising from the sequential decomposition of the joint distribution of  $(\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_T)$ :

$$\begin{aligned} D(\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_T; \psi) \\ = D(\mathbf{Z}_1, \dots, \mathbf{Z}_m; \psi) \prod_{t=m+1}^T D(\mathbf{Z}_t / \mathbf{Z}_{t-1}, \dots, \mathbf{Z}_{t-m}; \psi). \end{aligned} \quad (23.5)$$

$D(\mathbf{Z}_1, \dots, \mathbf{Z}_m; \psi)$  refers to the joint distribution of the first  $m$  observations interpreted as the *initial conditions*. Asymptotic independence enables us to argue that asymptotically the effect of the initial conditions is negligible. One important implication of this is that we can ignore the first  $m$  observations and treat  $t = m + 1, \dots, T$  as being the sample for statistical inference purposes. In what follows this ‘solution’ will be adopted for expositional purposes because ‘proper’ treatment of the initial conditions can complicate the argument without adding to our understanding.

The statistical generating mechanism (GM) of the dynamic linear regression model is based on the conditional distribution  $D(y_t / \mathbf{X}_t, \mathbf{Z}_{t-1}^0; \psi_1)$  related to  $D(\mathbf{Z}_t / \mathbf{Z}_{t-1}^0; \psi)$  via

$$D(\mathbf{Z}_t / \mathbf{Z}_{t-1}^0; \psi) = D(y_t / \mathbf{X}_t, \mathbf{Z}_{t-1}^0; \psi_1) D(\mathbf{X}_t / \mathbf{Z}_{t-1}^0; \psi_2). \quad (23.6)$$

The *systematic component* is defined by

$$\mu_t = E(y_t / \sigma(\mathbf{Y}_{t-1}^0), \mathbf{X}_t^0 = \mathbf{x}_t^0) = \boldsymbol{\beta}'_0 \mathbf{x}_t + \sum_{i=1}^m (\alpha_i y_{t-i} + \boldsymbol{\beta}'_i \mathbf{x}_{t-i}), \quad t > m, \quad (23.7)$$

where

$$\mathbf{Y}_{t-1}^0 \equiv (y_{t-1}, y_{t-2}, \dots, y_1), \quad \mathbf{X}_t^0 \equiv (\mathbf{X}_t, \mathbf{X}_{t-1}, \dots, \mathbf{X}_1),$$

$$\alpha_i \equiv (\alpha_{11}(i) + \boldsymbol{\omega}_{12} \boldsymbol{\Omega}_{22}^{-1} \mathbf{a}_{21}(i)), \quad \boldsymbol{\beta}_0 \equiv \boldsymbol{\Omega}_{22}^{-1} \boldsymbol{\omega}_{21},$$

$$\boldsymbol{\beta}'_i = (\mathbf{a}_{12}(i) - \boldsymbol{\omega}_{12} \boldsymbol{\Omega}_{22}^{-1} \mathbf{A}_{22}(i)), \quad i = 1, 2, \dots, m$$

(see Spanos (1985a)). The *non-systematic component* is defined by

$$u_t = y_t - E(y_t / \mathcal{F}_{t-1}), \quad (23.8)$$

where  $\mathcal{F}_{t-1} = \sigma(\mathbf{Z}_{t-1}^0, \mathbf{X}_t = \mathbf{x}_t)$ , and satisfies the following properties:

$$(ET1) \quad E(u_t) = E\{E(u_t/\mathcal{F}_{t-1})\} = 0. \quad (23.9)$$

$$(ET2) \quad E(u_t u_s) = E\{E(u_t u_s/\mathcal{F}_{t-1})\} = \begin{cases} \sigma_0^2 & t = s \\ 0 & t > s, \end{cases} \quad (23.10)$$

where

$$\sigma_0^2 = \boldsymbol{\omega}_{11} - \boldsymbol{\omega}_{12} \boldsymbol{\Omega}_{22}^{-1} \boldsymbol{\omega}_{21}.$$

These two properties show that  $u_t$  is a martingale difference process relative to  $\mathcal{F}_t$  with bounded variance, i.e. an innovation process (see Section 8). Moreover, the non-systematic component is also orthogonal to the systematic component, i.e.

$$(ET3) \quad E(\mu_t u_t) = E\{E(\mu_t u_t/\mathcal{F}_{t-1})\} = 0. \quad (23.11)$$

The properties ET1–ET3 can be verified directly using the properties of the conditional expectation discussed in Section 7.2. In view of the equality

$$\sigma(u_t, u_{t-1}, \dots, u_1) = \mathcal{F}_t, \quad (23.12)$$

we can deduce that

$$(ET4) \quad E(u_t/\sigma(\mathbf{U}_{t-1}^0)) = 0, \quad t > m, \quad (23.13)$$

for  $\mathbf{U}_{t-1}^0 = (u_{t-1}, u_{t-2}, \dots, u_1)$ ,  
i.e.  $u_t$  is not predictable from its own past.

This property extends the notion of a white-noise process encountered so far (see Granger (1980)).

As argued in Chapter 17, the parameters of interest are the parameters in terms of which the statistical GM is defined unless stated otherwise. These parameters should be defined more precisely as *statistical* parameters of interest with the theoretical parameters of interest being functions of the former. In the present context the statistical parameters of interest are  $\theta^* = \mathbf{H}(\psi_1)$  where  $\theta^* \equiv (\beta_0, \beta_1, \dots, \beta_m, \alpha_1, \dots, \alpha_m, \sigma_0^2)$ .

The normality of  $D(\mathbf{Z}_t/\mathbf{Z}_t^0; \psi)$  implies that  $\psi_1$  and  $\psi_2$  are variation free and thus  $\mathbf{X}_t$  is *weakly exogenous* with respect to  $\theta$ . This suggests that  $\theta^*$  can be estimated efficiently without any reference to the marginal distribution  $D(\mathbf{X}_t/\mathbf{Z}_{t-1}^0; \psi_2)$ . The presence of  $\mathbf{Y}_{t-1}^0$  in this marginal distribution, however, raises questions in the context of prediction because of the feedback from the lagged  $y_t$ s. In order to be able to treat the  $x_t$ s as given when predicting  $y_t$  we need to ensure that no such feedback exists. For this purpose we need to assume that

$$D(\mathbf{X}_t/\mathbf{Z}_{t-1}^0; \psi_2) = D(\mathbf{X}_t/\mathbf{X}_{t-1}^0; \psi_2), \quad t = m+1, \dots, T, \quad (23.14)$$

i.e.  $y_t$  does not Granger cause  $\mathbf{X}_t$  (see Engle *et al.* (1983) for a more detailed

discussion). Weak exogeneity of  $\mathbf{X}_t$  with respect to  $\boldsymbol{\theta}^*$  when supplemented with Granger non-causality, as defined in (14), is called *strong exogeneity*. Note that in the present context Granger non-causality is equivalent to

$$\mathbf{a}_{21}(i) = 0, \quad i = 1, 2, \dots, m \quad (23.15)$$

in (1), which suggests that the assumption is testable.

In the case of the linear regression model it was argued that, although the joint distribution  $D(\mathbf{Z}_t; \psi)$  was used to motivate the statistical model, its specification can be based exclusively on the conditional distribution  $D(y_t | \mathbf{X}_t; \psi_1)$ . The same applies to the specification of the dynamic linear regression model which can be based exclusively on  $D(y_t | \mathbf{Z}_{t-1}^0, \mathbf{X}_t; \psi_1)$ . In such a case, however, certain restrictions need to be imposed on the parameters of the statistical generating mechanism (GM):

$$y_t = \boldsymbol{\beta}'_0 \mathbf{x}_t + \sum_{i=1}^m \alpha_i y_{t-i} + \sum_{i=1}^m \boldsymbol{\beta}'_i \mathbf{x}_{t-i} + u_t, \quad t > m. \quad (23.16)$$

In particular we need to assume that the parameters  $(\alpha_1, \alpha_2, \dots, \alpha_m)$  satisfy the restriction that all the roots of the polynomial

$$\left( \lambda^m - \sum_{i=1}^{m-1} \alpha_i \lambda^{m-i} \right) = 0 \quad (23.17)$$

lie inside the unit circle, i.e.  $|\lambda_i| < 1, i = 1, 2, \dots, m$  (see Dhrymes (1978)). This restriction is necessary to ensure that  $\{y_t, t \in \mathbb{T}\}$  as generated by (16) is indeed an asymptotically independent stationary stochastic process. In the case where  $m = 1$  the restriction is  $|\alpha| < 1$  which ensures that

$$\text{Cov}(y_t y_{t+\tau}) = \sigma^2 \alpha^\tau \left( \frac{1 - \alpha^{2\tau}}{1 - \alpha^2} \right) \rightarrow 0 \quad \text{as } \tau \rightarrow \infty$$

(see Chapter 8). It is important to note that in the case where  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is assumed to be both stationary and asymptotically independent the above restriction on the roots of the polynomial in (17) is satisfied automatically.

For notational convenience let us rewrite the statistical GM (16) in the more concise form

$$y_t = \boldsymbol{\beta}^* \mathbf{X}_t^* + u_t, \quad t > m, \quad (23.18)$$

where

$$\boldsymbol{\beta}^* \equiv (\alpha_1, \alpha_2, \dots, \alpha_m, \boldsymbol{\beta}'_0, \boldsymbol{\beta}'_1, \dots, \boldsymbol{\beta}'_m)' : m(k+1) \times 1,$$

$$\mathbf{X}_t^* \equiv (y_{t-1}, y_{t-2}, \dots, y_{t-m}, \mathbf{x}'_t, \mathbf{x}'_{t-1}, \dots, \mathbf{x}'_{t-m})' : m(k+1) \times 1.$$

For the sample period  $t = m+1, \dots, T$ , (18) can be written in the following matrix form:

$$\mathbf{y} = \mathbf{X}^* \boldsymbol{\beta}^* + \mathbf{u}, \quad (23.19)$$

where  $\mathbf{y}: (T-m) \times 1$ ,  $\mathbf{X}^*: (T-m) \times m(k+1)$ . Note that  $\mathbf{x}_t$  is  $k \times 1$  because it includes the constant as well but  $\mathbf{x}_{t-i}$ ,  $i=1, 2, \dots, m$ , are  $(k-1) \times 1$  vectors; this convention is adopted to simplify the notation. Looking at (18) and (19) the discerning reader will have noticed a purposeful attempt to use notation which relates the dynamic linear regression model to the linear and stochastic linear regression models. Indeed, the statistical GM in (18) and (19) is a hybrid of the statistical GM's of these models. The part  $\sum_{i=1}^m \alpha_i y_{t-i}$  is directly related to the stochastic linear regression model in view of the conditioning on the  $\sigma$ -field  $\sigma(\mathbf{Y}_{t-1}^0)$  and the rest of the systematic component being a direct extension of that of the linear regression model. This relationship will prove very important in the statistical analysis of the parameters of the dynamic linear regression model discussed in what follows.

In direct analogy to the linear and stochastic linear regression models we need to assume that  $\mathbf{X}^*$  as defined above is of full rank, i.e.  $\text{rank}(\mathbf{X}^*) = m(k+1)$  for all the observable values of  $\mathbf{Y}_{T-1}^0 \equiv (y_m, y_{m+1}, \dots, y_{T-1})'$ .

The probability model underlying (16) comes in the form of the product of the sequentially conditional normal distribution  $D(y_t / \mathbf{Z}_{t-1}^0, \mathbf{X}_t; \boldsymbol{\psi}_1)$ ,  $t > m$ . For the sample period  $t = m+1, \dots, T$  the distribution of the sample is

$$D^*(\mathbf{y}; \boldsymbol{\psi}_1) \equiv \prod_{t=m+1}^T D(y_t / \mathbf{Z}_{t-1}^0, \mathbf{X}_t; \boldsymbol{\psi}_1). \quad (23.20)$$

The sampling model is specified to be a non-random sample from  $D^*(\mathbf{y}; \boldsymbol{\psi}_1)$ . Equivalently,  $\mathbf{y} \equiv (y_{m+1}, y_{m+2}, \dots, y_T)'$  can be viewed as a non-random sample sequentially drawn from  $D(y_t / \mathbf{Z}_{t-1}^0, \mathbf{X}_t; \boldsymbol{\psi}_1)$ ,  $t = m+1, \dots, T$ , respectively. As argued above, asymptotically, the effect of the initial conditions summarised by

$$D(\mathbf{Z}_1; \boldsymbol{\psi}) \prod_{t=1}^m D(y_t | \mathbf{Z}_{t-1}^0, \mathbf{X}_t; \boldsymbol{\psi}) \quad (23.21)$$

can be ignored: a strategy adopted in Section 23.2 for expositional purposes. The interested reader should consult Priestley (1981) for a readable discussion of how the initial conditions can be treated. Further discussion of these conditions is given in Section 23.3 below.

### *The dynamic linear regression model – specification*

#### **(I) The statistical GM**

$$y_t = \boldsymbol{\beta}'_0 \mathbf{x}_t + \sum_{i=1}^m \alpha_i y_{t-i} + \sum_{i=1}^m \boldsymbol{\beta}'_i \mathbf{x}_{t-i} + u_t, \quad t > m. \quad (23.22)$$

$$[1] \quad \mu_t = E(y_t/\sigma(\mathbf{Y}_{t-1}^0), \mathbf{X}_t^0 = \mathbf{x}_t^0) = \boldsymbol{\beta}'_0 \mathbf{x}_t + \sum_{i=1}^m (\alpha_i y_{t-i} + \boldsymbol{\beta}_i \mathbf{x}_{t-i}). \quad (23.23)$$

[2] The (statistical) parameters of interest are

$$\boldsymbol{\theta}^* \equiv (\alpha_1, \alpha_2, \dots, \alpha_m, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_m, \sigma_0^2);$$

see Appendix 22.1 for the form of the mapping  $\boldsymbol{\theta}^* = \mathbf{H}(\psi_1)$ .

[3]  $\mathbf{X}_t$  is strongly exogenous with respect to  $\boldsymbol{\theta}^*$ .

[4] The parameters  $\boldsymbol{\alpha} \equiv (\alpha_1, \alpha_2, \dots, \alpha_m)'$  satisfy the restriction that all the roots of the polynomial

$$\left( \lambda^m - \sum_{i=1}^{m-1} \alpha_i \lambda^{m-i} \right) = 0$$

are less than one in absolute value.

[5]  $\text{rank}(\mathbf{X}^*) = m(k+1)$  for all observable values of  $\mathbf{Y}_{T-m}^0$ ,  $T > m(k+1)$ .

## (II) The probability model

$$\Phi = \left\{ D(y_t/\mathbf{Z}_{t-1}^0, \mathbf{X}_t; \boldsymbol{\theta}^*) = \frac{1}{\sigma_0 \sqrt{(2\pi)}} \exp \left\{ -\frac{1}{2\sigma_0^2} (y_t - \boldsymbol{\beta}^{*\prime} \mathbf{X}_t^*)^2 \right\}, \right. \\ \left. \boldsymbol{\theta}^* \in \mathbb{R}^{m(k+1)} \times \mathbb{R}_+, t > m \right\}. \quad (23.24)$$

- [6] (i)  $D(y_t/\mathbf{Z}_{t-1}^0, \mathbf{X}_t; \boldsymbol{\theta}^*)$  is normal;
- (ii)  $E(y_t/\sigma(\mathbf{Y}_{t-1}^0), \mathbf{X}_t^0 = \mathbf{x}_t^0) = \boldsymbol{\beta}^{*\prime} \mathbf{X}_t^*$  – linear in  $\mathbf{X}_t^*$ ;
- (iii)  $\text{Var}(y_t/\sigma(\mathbf{Y}_{t-1}^0), \mathbf{X}_t^0 = \mathbf{x}_t^0) = \sigma_0^2$  – homoskedastic (free of  $\mathbf{X}_t^*$ );

[7]  $\boldsymbol{\theta}^*$  is time invariant.

## (III) The sampling model

- [8]  $\mathbf{y} \equiv (y_{m+1}, y_{m+2}, \dots, y_T)'$  is a stationary, asymptotically independent sample sequentially drawn from  $D(y_t/\mathbf{Z}_{t-1}^0, \mathbf{X}_t; \boldsymbol{\theta}^*)$ ,  $t = m+1, m+2, \dots, T$ , respectively.

Note that the above specification is based on  $D(y_t/\mathbf{Z}_{t-1}^0, \mathbf{X}_t; \boldsymbol{\theta}^*)$ ,  $t = m+1, \dots, T$ , directly and not on  $D(\mathbf{Z}_t/\mathbf{Z}_{t-1}^0; \psi)$ . This is the reason why we need assumption [4] in order to ensure the asymptotic independence of  $\{(y_t/\mathbf{Z}_{t-1}^0, \mathbf{X}_t), t > m\}$ .

## 23.2 Estimation

In view of the probability and sampling model assumptions [6] to [8] the likelihood function is defined by

$$L(\theta^*; \mathbf{y}) = \prod_{t=m+1}^T D(y_t / \mathbf{Z}_{t-1}^0, \mathbf{X}_t; \theta^*) \quad (23.25)$$

and

$$\begin{aligned} \log L(\theta^*; \mathbf{y}) &= \text{const} - \frac{(T-m)}{2} \log \sigma_0^2 \\ &\quad - \frac{1}{2\sigma_0^2} (\mathbf{y} - \mathbf{X}^* \tilde{\beta}^*)' (\mathbf{y} - \mathbf{X}^* \tilde{\beta}^*), \end{aligned} \quad (23.26)$$

$$\begin{aligned} \frac{\partial \log L}{\partial \beta} &= + \frac{1}{\sigma_0^2} (\mathbf{X}^{*\prime} \mathbf{y} - \mathbf{X}^{*\prime} \mathbf{X}^* \tilde{\beta}^*) = 0 \\ \Rightarrow \quad \tilde{\beta}^* &= (\mathbf{X}^{*\prime} \mathbf{X}^*)^{-1} \mathbf{X}^{*\prime} \mathbf{y}, \end{aligned} \quad (23.27)$$

$$\begin{aligned} \frac{\partial \log L}{\partial \sigma^2} &= - \frac{(T-m)}{2\sigma_0^2} + \frac{1}{2\sigma_0^4} \mathbf{u}' \mathbf{u} = 0 \\ \Rightarrow \quad \hat{\sigma}_0^2 &= \frac{1}{(T-m)} \mathbf{u}' \mathbf{u,} \end{aligned} \quad (23.28)$$

where  $\tilde{\mathbf{u}} = \mathbf{y} - \mathbf{X}^* \tilde{\beta}^*$ . The estimators  $\tilde{\beta}^*$  and  $\hat{\sigma}_0^2$  are said to be *approximate maximum likelihood estimators* (MLE's) of  $\beta^*$  and  $\sigma_0^2$ , respectively, because the initial conditions have been ignored. The formulae for these estimators bring out the similarity between the dynamic, linear and stochastic linear regression models. Moreover, the similarity does not end with the formulae. Given that the statistical GM for the dynamic linear regression model can be viewed as a hybrid of the other two models we can deduce that in direct analogy to the stochastic linear regression model the finite sample distributions of  $\tilde{\beta}^*$  and  $\hat{\sigma}_0^2$  are likely to be largely intractable. One important difference between the dynamic and stochastic linear regression models, however, is that in the former case, although the orthogonality between  $\mu_t$  and  $u_t$  ( $\mu_t \perp u_t$ ) holds in the decomposition

$$y_t = \mu_t + u_t, \quad \text{for each } t > m, \quad (23.29)$$

it does not extend to the sample period formulation

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{u} \quad (23.30)$$

because  $u_t$  is not independent of future  $y_t$ s, i.e.

$$E(u_t u_{t+\tau}) = \begin{cases} = 0 & \text{for } \tau < 0 \\ \neq 0 & \text{for } \tau \geq 0, \quad t > m. \end{cases} \quad (23.31)$$

One important implication of this is that  $\tilde{\beta}^*$  is a biased estimator of  $\beta^*$ , i.e.

$$E(\tilde{\beta}^* - \beta^*) = E[(\mathbf{X}^{*'}\mathbf{X}^*)^{-1}\mathbf{X}^{*'}\mathbf{u}] \neq 0. \quad (23.32)$$

This, however, is only a problem for a small  $T$  because asymptotically  $\tilde{\theta}^* \equiv (\tilde{\beta}^*, \tilde{\sigma}_0^2)$  enjoys certain attractive properties under fairly general conditions, including asymptotic unbiasedness.

### *Asymptotic properties of $\tilde{\theta}_T^*$*

Using the analogy between the dynamic and stochastic linear regression models we can argue that the asymptotic properties of  $\tilde{\theta}_T^*$  depend crucially on the order of magnitude (see Chapter 10) of the information matrix  $I_T(\theta^*)$  defined by

$$I_T(\theta^*) = \begin{pmatrix} \frac{E(\mathbf{X}^{*'}\mathbf{X}^*)}{\sigma_0^2} & 0 \\ 0 & \frac{(T-m)}{2\sigma_0^4} \end{pmatrix}. \quad (23.33)$$

This can be verified directly from (27) and (28). If  $E(\mathbf{X}^{*'}\mathbf{X}^*)$  is of order  $O_p(T)$  then  $I_T(\theta^*) = O_p(T)$  and thus the asymptotic information matrix  $\mathbf{I}_\infty(\theta^*) = \lim_{T \rightarrow \infty} [(1/T)\mathbf{I}_T(\theta^*)] < \infty$ . Moreover, if  $\mathbf{G}_T \equiv E(\mathbf{X}^{*'}\mathbf{X}^*/T)$  is also non-singular for all  $T$  ‘sufficiently large’, then the asymptotic properties of  $\tilde{\theta}_T^*$  as a MLE of  $\theta^*$  can be deduced. Let us consider the argument in more detail.

If we define the information set  $\mathcal{F}_{t-1} = (\sigma(Y_{t-1}^0), \mathbf{X}_t^0 = \mathbf{x}_t^0)$  we can deduce that the non-systematic component as defined above takes the form

$$u_t = y_t - E(y_t | \mathcal{F}_{t-1}), \quad t > m, \quad (23.34)$$

and the sequence  $\{u_t, \mathcal{F}_t, t > m\}$  represents a martingale difference, i.e.  $E(u_t | \mathcal{F}_{t-1}) = 0, t > m$  (see Section 8.4). Using the limit theorems related to martingales (see Section 9.3) we can then proceed to derive the asymptotic properties of  $\tilde{\theta}_T^*$ .

In view of the following relationship:

$$(\tilde{\beta}^* - \beta^*) = \left( \frac{\mathbf{X}^{*'}\mathbf{X}^*}{T} \right)^{-1} \left( \frac{\mathbf{X}^{*'}\mathbf{u}}{T} \right), \quad (23.35)$$

if we can ensure that

$$\left\{ \left( \frac{\mathbf{X}^{*'}\mathbf{X}^*}{T} \right) - \mathbf{G}_T \right\}_{T=1}^{\infty} \xrightarrow{\text{a.s.}} \mathbf{0}, \quad (23.36)$$

then we can use the strong law of large numbers for martingales to show

that

$$\left( \frac{\mathbf{X}^{*'} \mathbf{u}}{T} \right) \xrightarrow{\text{a.s.}} 0. \quad (23.37)$$

Hence,

$$\tilde{\beta}^* \xrightarrow{\text{a.s.}} \beta^*. \quad (23.38)$$

The convergence in (37) stems from the fact that  $\{\mathbf{X}_t^* u_t, \mathcal{F}_t\}$  defines a martingale difference given that

$$E(\mathbf{X}_t^* u_t / \mathcal{F}_{t-1}) = \mathbf{X}_t^* E(u_t / \mathcal{F}_{t-1}) = \mathbf{0}, \quad i = 1, 2, \dots, m(k+1). \quad (23.38)$$

This suggests that the main assumption underlying the strong consistency of  $\tilde{\beta}^*$  is the order of magnitude of  $E(\mathbf{X}^* \mathbf{X}^*)$  and the non-singularity of  $E(\mathbf{X}^* \mathbf{X}^* / T)$  for  $T > m(k+1)$ . Given that  $\mathbf{X}_t^* \equiv (y_{t-1}, y_{t-2}, \dots, y_{t-m}, \mathbf{x}_t, \mathbf{x}_{t-1}, \dots, \mathbf{x}_{t-m})$  we can ensure that  $E(\mathbf{X}^* \mathbf{X}^*)$  satisfies these conditions if the cross-products involved satisfy the restrictions. As far as the cross-products which involve the  $y_{t-i}$ s are concerned the restrictions are ensured by assumption [4] on the roots of  $(\lambda^m - \sum_{i=1}^{m-1} \alpha_i \lambda^{m-i}) = 0$ . For the  $x_t$ s we need to assume that:

- (i)  $|x_{ti}| < C$ ,  $i = 1, 2, \dots, k$ ,  $t \in \mathbb{T}$ ,  $C$  being a constant;
- (ii)  $\lim_{T \rightarrow \infty} [1/(T-\tau)] \sum_{t=0}^{T-\tau} \mathbf{x}_t \mathbf{x}'_{t-\tau} = \mathbf{Q}_\tau$  exists for  $\tau \geq 1$  and in particular  $\mathbf{Q}_0$  is also non-singular.

These assumptions ensure that  $E(\mathbf{X}^* \mathbf{X}^*) = O(T)$ , i.e.  $\mathbf{G}_T = O(1)$  and  $\mathbf{G}_T \rightarrow \mathbf{G}$  where  $\mathbf{G}$  is non-singular. These in turn imply that  $\mathbf{I}_T(\theta^*) = O(T)$  and  $\mathbf{I}_\infty(\theta^*) = O(1)$  which ensures that not only (37) holds but

$$\hat{\sigma}_0^2 \xrightarrow{\text{a.s.}} \sigma_0^2. \quad (23.39)$$

Moreover, by multiplying  $(\tilde{\theta}_T^* - \theta^*)$  with  $\sqrt{T}$  (the order of its standard deviation) *asymptotic normality* can be deduced:

$$\sqrt{T}(\tilde{\theta}_T^* - \theta^*) \underset{x}{\sim} N(\mathbf{0}, \mathbf{I}_\infty(\theta^*)^{-1}). \quad (23.40)$$

That is,

$$\sqrt{T}(\tilde{\beta}^* - \beta^*) \underset{x}{\sim} N(\mathbf{0}, \sigma_0^2 \mathbf{G}^{-1}), \quad (23.41)$$

$$\sqrt{T}(\hat{\sigma}_0^2 - \sigma_0^2) \underset{x}{\sim} N(\mathbf{0}, 2\sigma_0^4). \quad (23.42)$$

The (*weak*) consistency of  $\tilde{\theta}_T^*$  as an estimator of  $\theta^*$ , i.e.

$$\tilde{\theta}_T^* \xrightarrow{P} \theta^*, \quad (23.43)$$

follows from the strong consistency (see Chapter 10). Moreover, *asymptotic efficiency* follows from (40).

If we compare the above asymptotic properties of  $\tilde{\theta}_T^*$  with those of  $\hat{\theta}_T = (\hat{\beta}, \hat{\sigma}^2)$  in the linear regression model we can see that their asymptotic properties are almost identical. This suggests that the statistical testing and prediction results derived in the context of the linear regression model which are based on the asymptotic properties of  $\hat{\theta}_T$  are likely to apply to the present context of the DLR model. Moreover, any finite sample based result, which is also justifiable on asymptotic grounds, is likely to apply to the present case with minor modifications. The purpose of the next two sections is to consider this in more detail.

### *Example*

The tests for independence applied to the money equation in Chapter 22 showed that the assumption is invalid. Moreover, the erratic behaviour of the recursive estimators and the rejection of the linearity and homoskedasticity assumptions in Chapter 21 confirmed the invalidity of the conditioning on the current observed values of  $\mathbf{X}_t$  only. In such a case the natural way to proceed is to respecify the appropriate statistical model for the modelling of the money equation so as to take into consideration the time dependence in the sample.

In view of the discussion of the assumption of stationarity as well as economic theoretical reasons the dependent variable chosen for the postulated statistical GM is  $m_t^* = \ln(M_t/P_t)$  (see Fig. 19.2). The value of the maximum lag postulated is  $m=4$ , mainly because previous studies demonstrated the optimum lag for 'memory' restriction adequate to characterise similar economic time series (see Hendry (1980)). The postulated statistical GM is of the form

$$m_t^* = \beta_0 + \sum_{i=1}^4 \alpha_i m_{t-i}^* + \sum_{i=0}^4 (\beta_{1i} y_{t-i} + \beta_{2i} p_{t-i} + \beta_{3i} l_{t-i}) + c_1 Q_{1t} + c_2 Q_{2t} + c_3 Q_{3t} + u_t, \quad (23.44)$$

where  $Q_{it}$ ,  $i=1, 2, 3$ , are three dummy variables which refer to important monetary policy changes which led to short-run 'unusual' changes:  
 $(Q_{1t}-1971iv)$  *The introduction of competition and credit control.*

Table 23.1. Estimated statistical GM

	$j=0$	$j=1$	$j=2$	$j=3$	$j=4$
$m_{t-j}^*$		0.593 (0.106)	0.004 (0.122)	-0.060 (0.133)	0.231 (0.104)
$y_{t-j}$	0.359 (0.161)	0.311 (0.217)	-0.053 (0.205)	-0.311 (0.212)	0.020 (0.177)
$p_{t-j}$	-0.831 (0.287)	0.428 (0.594)	0.509 (0.570)	-0.032 (0.572)	-0.092 (0.306)
$i_{t-j}$	-0.054 (0.011)	0.023 (0.017)	-0.032 (0.018)	0.019 (0.017)	-0.024 (0.012)
$\hat{\beta}_0 = -1.065 \quad \hat{c}_1 = -0.049 \quad \hat{c}_2 = 0.064 \quad \hat{c}_3 = 0.050$					
$(0.651) \quad (0.016) \quad (0.017) \quad (0.016)$					
$R^2 = 0.952, \bar{R}^2 = 0.932, s = 0.0141, \log L = 229.764, T = 76$					

(Q<sub>2i</sub>-1975iv) The suspension of the corset and the Bank of England asked the banks to channel the new lending away from personal loans.

(Q<sub>3i</sub>-1982i) The introduction of M1 as a monetary target.

The estimated coefficients for the period 1964i-1982iv are shown in Table 23.1. Estimation of (44) with  $\Delta m_t^*$  as the dependent variable changed only the goodness-of-fit measure as given by  $R^2$  and  $\bar{R}^2$  to  $R^2 = 0.796$  and  $\bar{R}^2 = 0.711$ . The change measures the loss of goodness of fit due to the presence of a trend (compare Fig. 19.2 with 21.2). The parameters  $\theta \equiv (\beta_0, \beta_{1i}, \beta_{2i}, \beta_{3i}, z_{i+1}, i=0, 1, 2, 3, 4, c_1, c_2, c_3, \sigma_0^2)$  in terms of which the statistical GM is defined are the *statistical* and not the (economic) theoretical parameters of interest. In order to be able to determine the latter (using specification testing), we need to ensure first that the estimated statistical GM is well defined. That is, that the assumptions [1]-[8] underlying the statistical model are indeed valid. Testing for these assumptions is the task of misspecification testing in the context of the dynamic linear regression model, considered in the next section.

Looking at the time graph of the actual ( $y_t$ ) and fitted ( $\hat{y}_t$ ) (see Fig. 23.1) values of the dependent variable we can see that  $\hat{y}_t$  'tracks'  $y_t$  quite closely for the estimation period. This is partly confirmed by the value of the variance of the regression which is nearly one-tenth of that of the money equation estimated in Chapter 19; see Fig. 23.2 showing the two sets of residuals. This reduction is even more impressive given that it has been achieved using the same sample information set. In terms of the  $R^2$  we can see that this also confirms the improvement in the goodness of fit being

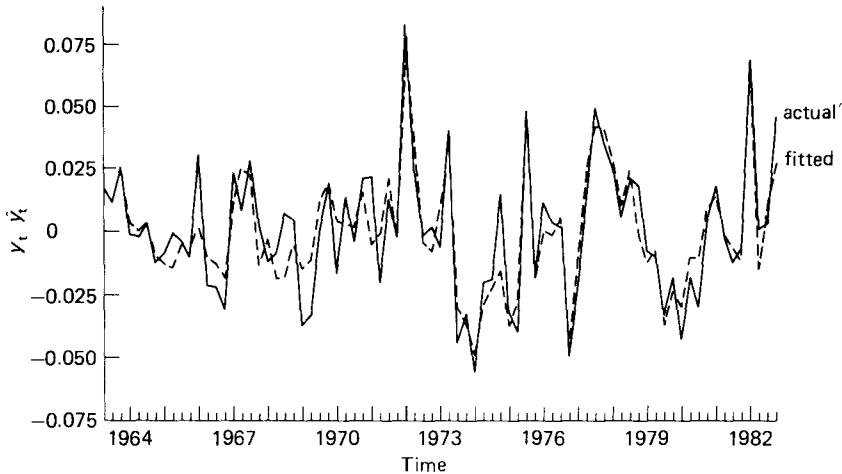


Fig. 23.1. The time graph of actual  $y_t = \Delta \ln(M/P)_t$  and fitted  $\hat{y}_t$  from the estimated regression in Table 23.1.

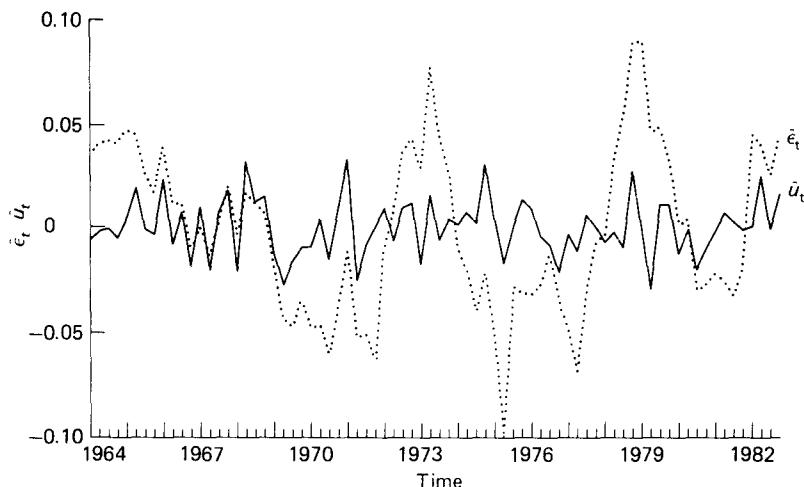


Fig. 23.2. Comparing the residuals from the estimated regressions (19.66) and (23.44) (see Table 23.1).

more than double the original one with  $m_t^*$  as the dependent variable and  $y_t$ ,  $p_t$  and  $i_t$  only the regressors (see Chapter 19).

Before we proceed with the misspecification testing of the above estimated statistical GM it is important to comment on the presence of the

three dummy variables. Their presence brings out the importance of the sample period economic ‘history’ background in econometric modelling. Unless the modeller is aware of this background the modelling can very easily go astray. In the above case, leaving the three dummies out, the normality assumption will certainly be affected. It is also important to remember that such dummy variables are only employed when the modeller believes that certain events had only a temporary effect on the relationship without any lasting changes in the relationship. In the case of longer-term effects we need to model them, not just pick them up using dummies. Moreover, the number of such dummies should be restricted to be relatively small. A liberal use of dummy variables can certainly achieve wonders in terms of goodness of fit but very little else. Indeed, a dummy variable for each observation which yields a perfect fit but no ‘explanation’ of any kind. In a certain sense the coefficients of dummy variables represent a measure of our ‘ignorance’.

### 23.3 Misspecification testing

As argued above, specification testing is based on the assumption of *correct specification*, that is, the assumptions underlying the statistical model in question are valid. This is because departures from these assumptions can invalidate the testing procedures. For this reason it is important to test for the validity of these assumptions before we can proceed to determine an empirical econometric model on the sound basis of a well-defined estimated statistical GM.

#### (1) Assumption underlying the statistical GM

*Assumption [1]* refers to the definition of the systematic and non-systematic components of the statistical GM. The most important restriction in defining the systematic component as

$$\mu_t = E(y_t/\sigma(\mathbf{Y}_{t-1}^0), \mathbf{X}_t^0 = \mathbf{x}_t^0) = \boldsymbol{\beta}'_0 \mathbf{x}_t + \sum_{i=1}^m (\alpha_i y_{t-i} + \boldsymbol{\beta}'_i \mathbf{x}_{t-i}) \quad (23.45)$$

is the choice of the *maximum lag m*. As argued in what follows an inappropriate choice of *m* can have some serious consequences for the estimation results derived in Section 23.2 as well as the specification testing results to be considered in Section 23.4 below.

##### (i) *m chosen ‘too large’*

If *m* is chosen larger than its optimum but unknown value *m\** then ‘near

collinearity' (insufficient data information) or even exact collinearity will 'creep' into the statistical GM (see Section 20.6). This is because as  $m$  is increased the same observed data are 'asked' to provide further and further information about an increasing number of unknown parameters. The implications of insufficient data information discussed in the context of the linear regression problem can be applied to the DLR model with some reinterpretation due to the presence of lagged  $y_t$ s in  $\mathbf{X}_t^*$ .

(ii)  *$m$  chosen 'too small'*

If  $m$  is chosen 'too small' then the omitted lagged  $\mathbf{Z}_t$ s will form part of the unmodelled part of  $y_t$  and the error term

$$u_t^* = y_t - \boldsymbol{\beta}'_0 \mathbf{x}_t - \sum_{i=1}^m (\alpha_i y_{t-i} + \boldsymbol{\beta}'_i \mathbf{x}_{t-i}) \quad (23.46)$$

is no longer non-systematic relative to the information set

$$\mathcal{F}_{t-1} = (\sigma(\mathbf{Y}_{t-1}^0), \mathbf{X}_t^0 = \mathbf{x}_t^0). \quad (23.47)$$

That is,  $\{u_t^*, \mathcal{F}_t, t > m\}$  will no longer be a martingale difference, having very serious consequences for the properties of  $\hat{\theta}^*$  discussed in Section 23.2. In particular the consistency and asymptotic normality of  $\hat{\theta}^*$  are no longer valid, as can be verified using the results of Section 22.1; see also Section 23.2 above. Because of these implications it is important to be able to test for  $m < m^*$ . Given that the 'true' statistical GM is given by

$$y_t = \boldsymbol{\beta}'_0 \mathbf{x}_t + \sum_{i=1}^{m_*} (\alpha_i y_{t-i} + \boldsymbol{\beta}'_i \mathbf{x}_{t-i}) + u_t, \quad t > m, \quad (23.48)$$

the error term  $u_t^*$  can be written in the form

$$u_t^* = u_t + \sum_{i=m}^{m_*} (\alpha_i y_{t-i} + \boldsymbol{\beta}'_i \mathbf{x}_{t-i}), \quad t > m. \quad (23.49)$$

This implies that  $m < m^*$  can be tested using the null hypothesis

$H_0: \boldsymbol{\alpha}_m^* = \mathbf{0}$  and  $\boldsymbol{\beta}_m^* = \mathbf{0}$  against  $H_1: \boldsymbol{\alpha}_m^* \neq \mathbf{0}$  or  $\boldsymbol{\beta}_m^* \neq \mathbf{0}$ ,  
where

$$\boldsymbol{\alpha}_m^* \equiv (\alpha_{m+1}, \dots, \alpha_m)', \quad \boldsymbol{\beta}_m^* \equiv (\boldsymbol{\beta}_{m+1}, \dots, \boldsymbol{\beta}_m').$$

The obvious test for such a hypothesis will be analogous to the  $F$ -type test for independence suggested in Chapter 22. Alternatively, we could use an asymptotically equivalent test based on  $TR^2$  from the auxiliary regression

$$\hat{u}_t^* = (\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})' \mathbf{x}_t + \sum_{i=1}^{m_*} (\alpha_i y_{t-i} + \boldsymbol{\beta}'_i \mathbf{x}_{t-i}) + \varepsilon_t, \quad t > m, \quad (23.50)$$

where  $\hat{u}_t^*$  refers to the residuals from the estimation of

$$y_t = \beta_0' \mathbf{x}_t + \sum_{i=1}^m (\alpha_i y_{t-i} + \beta_i' \mathbf{x}_{t-i}) + u_t^*$$

The rejection region is defined by

$$C_1 = \{\mathbf{y}: TR^2 > c_\alpha\}, \quad \alpha = \int_{c_\alpha}^\infty d\chi^2[(m^* - m)k] \quad (23.52)$$

(see Chapter 22 for more details).

The *F*-type test for the money equation estimated in Section 23.3 with  $m^* = 6$  yielded

$$FT(y) = \left( \frac{0.01049 - 0.00965}{0.00965} \right) \left( \frac{43}{8} \right) = 0.467. \quad (23.53)$$

Given that  $c_\alpha = 2.14$  for  $\alpha = 0.05$ ,  $H_0$  is strongly accepted and the value of  $m = 4$  seems appropriate.

As argued in Chapter 22, the various tests for residual autocorrelation can be used in the context of the respecification approach as tests for independence assumption especially when the degrees of freedom are at a premium. In the present context the same misspecification tests can be used with certain modifications as indirect tests of  $m < m^*$ . As far as the asymptotic Lagrange multiplier (*LM*) tests for AR( $p$ ) or MA( $p$ ), where  $p \geq 1$ , based on the auxiliary regressions are concerned, can be applied in the present context without any modifications. The reason is that the presence of the lagged  $y_t$ s in the systematic component of the statistical GM (51) makes no difference asymptotically. On the other hand, the Durbin–Watson (DW) test is *not applicable* in the present context because the test depends crucially on the non-stochastic nature of the matrix  $\mathbf{X}$ . Durbin (1970) proposed a test for AR(1) ( $u_t^* = \rho u_{t-1}^* + u_t$ ) in the context of the DLR model based on the so-called *h*-test statistic defined by

$$h = (1 - \frac{1}{2} DW(\mathbf{y})) \left( \frac{T-m}{1 - (T-m) \text{Var}(\hat{\alpha}_i)} \right)^{\frac{1}{2}} \sim N(0, 1) \quad (23.54)$$

under  $H_0: \rho = 0$ . Its rejection region takes the form

$$C_1 = \{\mathbf{y}: |h| > c_\alpha\}, \quad \frac{1}{2}\alpha = \int_{c_\alpha}^\infty \phi(h) dh, \quad (23.55)$$

where  $\phi(\cdot)$  is the density function of a standard normal distribution. It is interesting to note that Durbin's *h*-test can be rationalised as a Lagrange multiplier test. The Lagrange multiplier test statistic for a first-order

dependency (AR(1) or MA(1)) takes the form

$$\tau(y) = \tilde{r}_1^2 \left\{ \frac{T^*}{1 - T^* \text{Var}(\tilde{\alpha}_1)} \right\} \quad (23.56)$$

(see Harvey (1981)), and  $\tau(y) \underset{\alpha}{\sim} \chi^2(1)$ . Noting that

$$\tilde{r}_1 = (1 - \frac{1}{2}DW), \quad (23.57)$$

we can see that the Durbin  $h$ -test can be interpreted as a Lagrange multiplier ( $LM$ ) test based on the first-order temporal correlation coefficient

As in the case of the linear regression model the above Durbin's  $h$ -test and the  $LM(l)$  ( $l \geq 1$ ) test can be viewed as tests of significance in the context of the auxiliary regression

$$\hat{u}_t = \delta' \mathbf{x}_t^* + \sum_{i=1}^l \rho_i \hat{u}_{t-i} + v_t, \quad t > m+l. \quad (23.58)$$

The obvious way to test

$$H_0: \rho_1 = \rho_2 = \dots = \rho_l = 0, \quad H_1: \rho_i \neq 0, \quad \text{for any } i = 1, 2, \dots, l \quad (23.59)$$

is to use the  $F$ -test approximation which includes the degrees of freedom correction term instead of its asymptotic chi-square form. The autocorrelation error tests will be particularly useful in cases where the  $F$ -test based on (48) cannot be applied because the degrees of freedom are at a premium.

In the case of the estimated money equation in Table 23.1 the above test statistics for  $\alpha = 0.05$  yielded:

$$h = 1.19, \quad c_\alpha = 1.96,$$

$$LM(2): FT(y) = \left( \frac{0.010487 - 0.010339}{0.010339} \right) \left( \frac{49}{2} \right) = 0.350, \quad c_\alpha = 3.18, \quad (23.60)$$

$$LM(3): FT(y) = \left( \frac{0.010500 - 0.010291}{0.010291} \right) \left( \frac{47}{3} \right) = 0.318, \quad c_\alpha = 2.80, \quad (23.61)$$

$$LM(4): FT(y) = \left( \frac{0.010466 - 0.010255}{0.010255} \right) \left( \frac{45}{4} \right) = 0.231, \quad c_\alpha = 2.5\%. \quad (23.62)$$

As we can see from the above results in all cases the null hypothesis is accepted confirming (53) above.

The above tests can be viewed as indirect ways to test the assumption postulating the adequacy of the maximum lag  $m$ . The question which naturally arises is whether  $m$  can be determined directly by the data. In the statistical time-series literature this question has been considered extensively and various formal procedures have been suggested such as Akaike's AIC and BIC or Parzen's CAT criteria (see Priestley (1981) for a readable summary of these procedures). In econometric practice, however, it might be preferable to postulate  $m$  on a priori grounds and then use the above indirect tests for its adequacy.

*Assumption [2]* specifies the statistical parameters of interest as being  $\theta^* \equiv (\alpha_1, \dots, \alpha_m, \beta_0, \beta_1, \dots, \beta_m, \sigma_0^2)$ . These parameters provide us with an opportunity to consider two issues we only discussed in passing. The first issue is related to the distinction made in Chapter 17 between the statistical and (economic) theoretical parameters of interest. In the present context  $\theta^*$  as defined above has very little, if any, economic interpretation. Hence,  $\theta^*$  represents the statistical parameters of interest. These parameters enable us to specify a well-defined statistical model which can provide the basis of the 'design' for an empirical econometric model. As argued in Chapter 17, the estimated statistical GM could be viewed as a sufficient statistic for the theoretical parameters of interest. The statistical parameters of interest provide only a statistically 'adequate' (sufficient) parametrisation with the theoretical parameters of interest being defined as functions of the former. This is because a theoretical parameter is well defined (statistically) only when it is directly related to a well-defined statistical parameter. The determination of the theoretical parameters of interest will be considered in Section 23.4 on specification testing. The second related issue is concerned with the presence of 'near' collinearity. In Section 20.6 it was argued that 'near' collinearity is defined relative to a given parametrisation and information set. In the present context it is likely that 'near' collinearity (or insufficient data information) might be a problem relative to the parametrisation based on  $\theta^*$ . The problem, however, can be easily overcome in determining the theoretical parameters of interest so as to 'design' a parsimonious as well as 'robust' theoretical parametrisation. Both issues will be considered in Section 23.6 below in relation to the statistical GM estimated in Section 23.2.

*Assumption [3]* postulates the strong exogeneity of  $\mathbf{X}_t$  with respect to the parameters  $\theta^*$  for  $t = m+1, \dots, T$ . As far as the weak exogeneity component of this assumption is concerned it will be treated as a non-testable presupposition as in the context of the linear regression model (see Section 20.3). The Granger non-causality component, however, is testable in the context of the general autoregressive representation

$$\mathbf{Z}_t = \sum_{i=1}^m \mathbf{A}(i) \mathbf{Z}_{t-i} + \mathbf{E}_t \quad (23.63)$$

(see Appendix 22.1). In particular,  $y_t$  does not ‘Granger cause’  $\mathbf{X}_t$  if  $\alpha_{21}(i) = \mathbf{0}$  for  $i = 1, 2, \dots, m$ . This suggests that a test of Granger non-causality can be based on  $H_0: \alpha_{21}(i) = \mathbf{0}$  for all  $i = 1, 2, \dots, m$  against  $H_1: \alpha_{21}(i) \neq \mathbf{0}$  for any  $i = 1, 2, \dots, m$ . For the case where  $k = 1$ , Granger (1969) suggested the Wald test statistic

$$W = (T - m) \left( \frac{RRSS - URSS}{URSS} \right)^{\frac{H_0}{2}} \sim \chi^2(m), \quad (23.64)$$

where URSS and RRSS refer to the residuals sums of squares of the regressions with and without the  $y_{t-i}$ s,  $i = 1, 2, \dots, m$ , respectively. The rejection region takes the form

$$C_1 = \{\mathbf{X}: W > c_x\}, \quad c_x = \int_{c_y}^{\infty} d\chi^2(m). \quad (23.65)$$

The Wald test statistic can be viewed as an  $F$ -type test statistic and thus a natural way to generalise it to the case where  $k > 1$  is to use an  $F$ -type test of significance in the context of multivariate linear regression (see Chapter 24). For a comprehensive survey of Granger non-causality tests see Geweke (1984).

*Assumption [4]* refers to the restrictions on  $\alpha$  needed to ensure that  $\{y_t, t \in \mathbb{T}\}$  as generated by the statistical GM:

$$y_t = \beta'_0 \mathbf{x}_t + \sum_{i=1}^m \alpha_i y_t + \sum_{i=1}^m \beta'_i \mathbf{x}_{t-i} + u_t, \quad t > m \quad (23.66)$$

is an asymptotically independent stochastic process. If we rewrite (66) in the form

$$\alpha(L)y_t = w_t + u_t, \quad (23.67)$$

where

$$\alpha(L) = (1 - \alpha_1 L - \cdots - \alpha_m L^m), \quad w_t = \sum_{i=0}^m \beta'_i \mathbf{x}_{t-i}$$

and treat it as a difference equation, then its solution (assuming that  $\alpha(L) = 0$  has  $m$  distinct roots) takes the general form

$$y_t = g(t) + \alpha^{-1}(L)(w_t + u_t), \quad (23.68)$$

where  $g(t) = c_1 \lambda_1^t + c_2 \lambda_2^t + \cdots + c_m \lambda_m^t$  (see Dhrymes (1978)). The component

$g(t)$ , called the complementary function, is the solution of the homogeneous difference equation  $\alpha(L)y_t=0$  and  $\mathbf{c}=(c_1, c_2, \dots, c_m)'$  are constants determined by the *initial* conditions  $y_1, y_2, \dots, y_m$  via

$$\left. \begin{aligned} y_1 &= c_1 + c_2 + \cdots + c_m \\ y_2 &= c_1 \lambda_1 + c_2 \lambda_2 + \cdots + c_m \lambda_m \\ &\vdots \\ y_m &= c_1 \lambda_1^{m-1} + c_2 \lambda_2^{m-1} + \cdots + c_m \lambda_m^{m-1} \end{aligned} \right\}. \quad (23.69)$$

In order to ensure the asymptotic independence (stationarity) of  $\{y_t, t \in \mathbb{T}\}$  we need this component to decay to zero as  $t \rightarrow \infty$  in order for  $\{y_t, t \in \mathbb{T}\}$  to 'forget' the initial conditions (see Priestley (1981)). For this to be the case ( $\lambda_1, \lambda_2, \dots, \lambda_m$ ), which are the roots of the polynomial

$$\alpha(\lambda) \equiv (\lambda^m - \alpha_1 \lambda^{m-1} - \cdots - \alpha_m) = 0. \quad (23.70)$$

should satisfy the restrictions

$$|\lambda_i| < 1, \quad i = 1, 2, \dots, m. \quad (23.71)$$

Note that the roots of  $\alpha(L)$  are  $(1/\lambda_i)$ ,  $i = 1, 2, \dots, m$ .

When the restrictions hold

$$\lambda_i^t \rightarrow 0 \quad \text{as } t \rightarrow \infty \quad \text{and} \quad \lim_{t \rightarrow \infty} g(t) = 0. \quad (23.72)$$

As argued in Section 23.1 in the case where  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is assumed to be a stationary, asymptotically independent process at the outset, the time invariance of the statistical parameters of interest and the restrictions  $|\lambda_i| < 1$ ,  $i = 1, 2, \dots, m$ , are automatically satisfied.

In order to get some idea as to what happens in the case where the restrictions  $|\lambda_i| < 1$ ,  $i = 1, 2, \dots, m$ , are not satisfied let us consider the simplest case where  $m = 1$  and  $\lambda_1 = 1$  ( $\alpha_1 = 1$ ), i.e.

$$y_t = y_{t-1} + w_t + u_t = \sum_{s=1}^t (w_s + u_s), \quad (23.73)$$

$$E(y_t) = w_t t \quad \text{and} \quad \text{Cov}(y_t y_{t+\tau}) = \sigma_0^2 t, \quad t > 1. \quad (23.74)$$

These suggest that both the mean and covariance of  $\{y_t, t \in \mathbb{T}\}$  increase with  $t$  and thus as  $t \rightarrow \infty$  they become unbounded. Thus  $\{y_t, t \in \mathbb{T}\}$ , as generated by (73), is not only non-stationary (with its mean and covariance varying with  $t$ ) but also its 'memory' remains constant.

In the case where  $|\alpha_1| > 1$  again  $\{y_t, t \in \mathbb{T}\}$  is non-stationary since

$$E(y_t) = w_t \left( \frac{1 - \alpha_1^t}{1 - \alpha_1} \right) \quad \text{and} \quad \text{Cov}(y_t, y_{t+\tau}) = \sigma_0^2 \left( \frac{1 - \alpha_1^{2t}}{1 - \alpha_1} \right) \alpha_1^\tau, \quad (23.75)$$

and thus  $E(y_t) \rightarrow \infty$  and  $\text{Cov}(y_t, y_{t+\tau}) \rightarrow \infty$  as  $t \rightarrow \infty$ . Moreover, the ‘memory’ of the process increases as the gap  $\tau \rightarrow \infty$ !

In the simplest case where  $m=1$  when the restriction  $|\lambda_1| < 1$  is not satisfied we run into problems which invalidate some of the results of Section 23.2. In particular the asymptotic properties of the approximate MLE’s of  $\theta^*$  need to be modified (see Fuller (1976) for a more detailed discussion). For the general case where  $m > 1$  the situation is even more complicated and most of the questions related to the asymptotic properties of  $\tilde{\theta}^*$  are as yet unresolved (see Rao (1984)).

*Assumption [5]*, relating to the rank of  $\mathbf{X}^* \equiv (y_{t-1}, \dots, y_{t-m}, \mathbf{x}'_t, \mathbf{x}'_{t-1}, \dots, \mathbf{x}'_{t-m})$ , has already been discussed in relation to assumption [1]. In the case where for the observed sample  $y$  the rank condition falls, i.e.

$$\text{rank}(\mathbf{X}^*) = n < m(k+1), \quad (23.76)$$

a unique  $\tilde{\theta}^*$  does not exist. If this is due to the fact that the postulated  $m$  is much larger than the optimum maximum lag  $m^*$  then the way to proceed is to reduce  $m$ . If, however, the problem is due to the rank of the submatrix  $\mathbf{X} \equiv (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T)'$  then we need to reparametrise (see Chapter 20). In either case the problem is relatively easy to detect. What is more difficult to detect is ‘near’ collinearity which might be particularly relevant in the present context. As argued above, however, the problem is relative to a given parametrisation and thus can be tackled alongside the reparametrisation of (48) in our attempt to ‘design’ an empirical econometric model based on the estimated form of (48) (see Section 20.6).

## (2) Assumptions underlying the probability model

The assumptions underlying the probability model constitute a hybrid of those of the linear and stochastic linear regression models considered in Chapters 19 and 20 respectively. The only new feature in the present context is the presence of the initial conditions coming in the form of the following distribution:

$$D(\mathbf{Z}_1; \boldsymbol{\phi}) \prod_{t=2}^m D(y_t / \mathbf{Z}_{t-1}^0, \mathbf{X}_t; \boldsymbol{\phi}), \quad (23.77)$$

where  $\boldsymbol{\phi} = \mathbf{H}(\psi)$ . For expositional purposes we chose to ignore these initial conditions in Sections 23.1 and 23.2. This enabled us to ignore the problem

of having to estimate the statistical GM

$$y_t = \beta'_0 \mathbf{x}_t + \sum_{i=1}^{t-1} \alpha_i y_{t-i} + \sum_{i=1}^{t-1} \beta'_i \mathbf{x}_{t-i} + u_t \quad (23.78)$$

for the period  $t = 1, 2, \dots, m$ . The easiest way we can take the initial conditions into consideration is to assume that  $\phi$  coincides with  $\theta^*$ . If this is adopted the approximate MLE's  $\tilde{\theta}^*$  will be modified in so far as the various summations involved will no longer be of the form  $\sum_{t=m+1}^T$  uniformly but  $\sum_{t=1+i}^T \mathbf{X}_t^* \mathbf{X}_{t-i}^*$  and  $\sum_{t=2+i}^T \mathbf{X}_{t-i}^* y_t$ ,  $i = 1, 2, \dots, m$ . That is, start summing from the point where observations become available for each individual component.

As far as the assumptions of normality, linearity and homoskedasticity are concerned the same comments made in Chapter 21 in the context of the linear regression model apply here with minor modifications. In particular the results based on asymptotic theory arguments carry over to the present case. The implications of non-normality, as defined in the context of the linear regression model (see Chapter 21), can be extended directly to the DLR model unchanged. The OLS estimators of  $\beta^*$  and  $\sigma_0^2$  have more or less the same asymptotic properties and any testing based on their asymptotic distributions remains valid. In relation to misspecification testing for departures from normality the asymptotic test based on the skewness and kurtosis coefficients remains valid without any changes. Let us apply this test to the money equation estimated in Section 23.3. The skewness–kurtosis test statistic is  $\tau(y) = 2.347$  which implies that for  $\alpha = 0.05$  the null hypothesis of normality is not rejected since  $c_\alpha = 5.99$ .

Testing linearity in the present context presents us with additional problems in so far as the test based on the Kolmogorov–Gabor polynomial (see (21.11) and (21.77)) will not be operational. The RESET type test, however, based on (21.10) and (21.83) is likely to be operational in the present context. Applying this test with  $\hat{y}_t^4$  ( $\hat{y}_t^2$  and  $\hat{y}_t^3$  were excluded because of collinearity) in the auxiliary regression

$$y_t = \delta' \mathbf{x}_t^* + \gamma \hat{y}_t^4 + v_t \quad (23.79)$$

yielded:  $FT(y) = 2.867$ ,  $c_\alpha = 4.03$  for  $\alpha = 0.05$ , which does not reject  $H_0$ .

Testing homoskedasticity in the present context using the Kolmogorov–Gabor parametrisation or the White test (see Nicholls and Pagan (1983)) presents us with the same ‘lack of degrees of freedom’ problem. On the other hand, it might be interesting to use only the cross-products of the  $x_{it}$ s only as in the linear regression case. Such a test can be used to refute the conjecture made in Chapter 21 about the likely source of the detected heteroskedasticity. It was conjectured that invalid conditioning (ignoring

the dependence in the sample) was the main reason for rejecting the homoskedasticity assumption based on the results of the White test. An obvious way to refute such a conjecture is to use the same regressors  $\psi_{1t}, \dots, \psi_{6t}$  in an auxiliary regression where  $\hat{u}_t$  refers to the residuals of the dynamic money equation estimated in Section 23.3 above. This auxiliary regression yielded:

$$TR^2 = 5.11, \quad FT(y) = 0.830. \quad (23.80)$$

The values of both test statistics for the significance of the coefficients of the  $\psi_{it}$ s reject the alternative (heteroskedasticity) most strongly; their critical values being  $c_x = 12.6$  and  $c_x = 2.2$  respectively for  $\alpha = 0.05$ . The time path of the residuals shown in Fig. 23.3(a) exemplifies no obvious systematic variation.

In the present context heteroskedasticity takes the general form

$$\text{Var}(y_t/\sigma(\mathbf{Y}_{t-1}^0), \mathbf{X}_t^0 = \mathbf{x}_t^0) = h(\mathbf{Y}_{t-1}^0, \mathbf{x}_t^0). \quad (23.81)$$

This suggests that an obvious way to ‘solve the problem of applying the White test’ is to use the lagged residuals as proxies for  $\mathbf{Z}_{t-1}^0$ . That is, use  $\hat{u}_{t-1}^2, \hat{u}_{t-2}^2, \dots, \hat{u}_{t-p}^2$  as proxy regressors in the auxiliary regression:

$$\hat{u}_t^2 = c_0 + c_1 \hat{u}_{t-1}^2 + c_2 \hat{u}_{t-2}^2 + \dots + c_p \hat{u}_{t-p}^2 + e_t \quad (23.82)$$

and test the significance of  $c_1, \dots, c_p$ . This is the so-called ARCH (autoregressive conditional heteroskedasticity) test and was suggested by Engle (1982). In the case of the above money equation the ARCH test statistic for  $p=4$  takes the value  $FT(y) = 0.074$ , with critical value  $c_x = 2.51$  for  $\alpha = 0.05$ . Hence, no heteroskedasticity is detected, confirming the ‘short’ White test given above.

In Chapter 22 it was argued that the main reason for the detected parameter time dependence related to the money equation was the invalid conditioning. That is, the fact that the temporal dependence in the sample was not taken into consideration. Having conditioned on all past information to define the dynamic linear regression model the question arises as to what extent the newly defined parameters are time invariant. In Fig. 23.3(b)–(d) we can see that the time paths of the estimated recursive coefficients of  $y_t$ ,  $p_t$  and  $i_t$  exemplify remarkable time invariance for the last 40 periods. It might be instructive to compare them with Fig. 21.1(b)–(d) of the static money equation.

### 23.4 Specification testing

Specification testing refers to testing of hypotheses related to the statistical

parameters of interest assuming that the assumptions underlying the statistical model in question are valid. In the context of the dynamic linear regression (DLR) model specification testing is particularly important because the estimated statistical GM as it stands has very little, if any, economic interpretation. The estimated statistical GM when tested for any misspecifications and none of the underlying assumptions is rejected can only be interpreted as providing a convenient summarisation of the sample

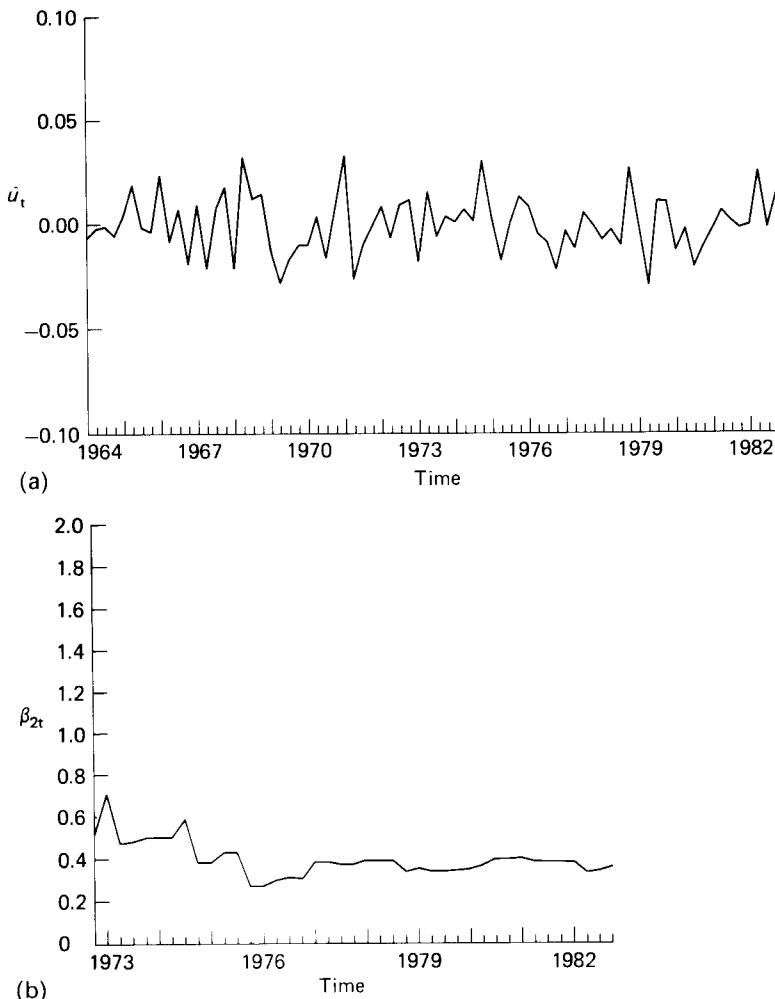
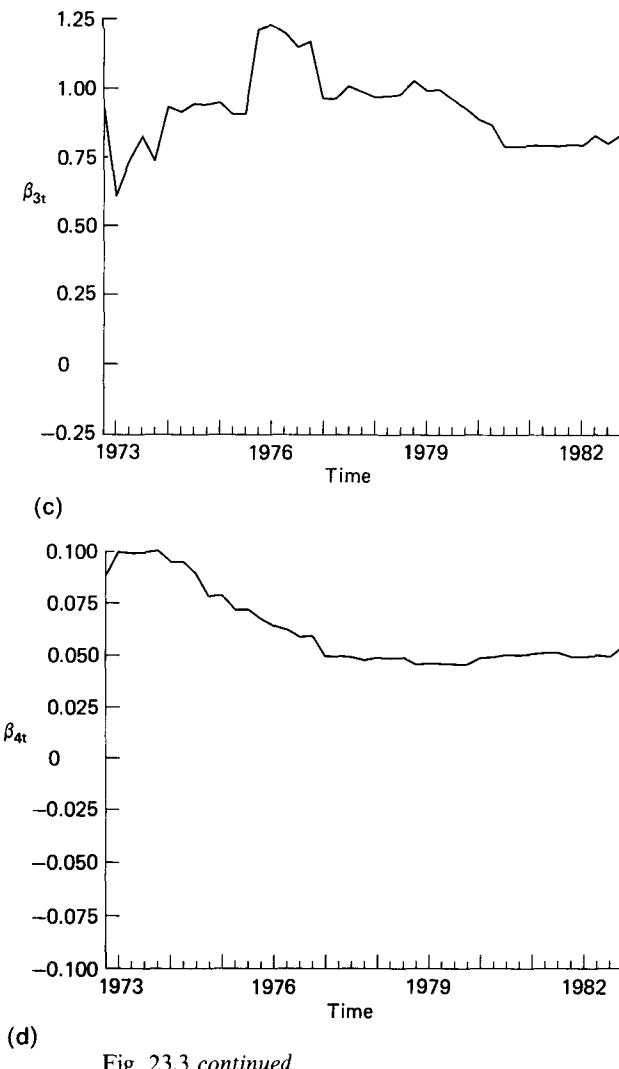


Fig. 23.3(a). The residuals from (23.44). (b)–(d) The time paths of the recursive estimates of  $\beta_{2t}$ ,  $\beta_{3t}$  and  $\beta_{4t}$ , the coefficients of  $\ln y_t$ ,  $\ln P_t$  and  $\ln I_t$  respectively from (23.44).

Fig. 23.3 *continued*

information. The (economic) theoretical parameters of interest are assumed to be simple functions of the statistical parameters  $\theta^*$ . These theoretical parameters will be determined using specification testing.

The well-defined estimated statistical GM provides the firm foundation on which the empirical econometric model will be constructed. There are two important considerations which must be taken into account in going from the estimated statistical GM to the empirical econometric model.

Firstly, any theoretical restrictions needed to determine the theoretical from the statistical parameters of interest must be tested before being imposed. Secondly, when these restrictions are imposed we should ensure that none of the statistical properties defining the original statistical model has been invalidated. That is, we should ensure that the empirical econometric model constructed is as well specified (statistically) as the original statistical GM on which it was based.

An important class of restrictions motivated by economic theory considerations are the exact linear restrictions related to  $\beta^*$ :

$$H_0: \mathbf{R}\beta^* = \mathbf{r} \text{ against } H_1: \mathbf{R}\beta^* \neq \mathbf{r}, \quad (23.83)$$

where  $\mathbf{R}$  and  $\mathbf{r}$  are  $q \times k^*$  and  $q \times 1$  known matrices with  $\text{rank}(\mathbf{R}) = q$ ,  $k^* = m(k+1)$ . Using the analogy with the linear regression model the  $F$ -type test statistic suggests itself:

$$\begin{aligned} FT^*(\mathbf{y}) &= \frac{(\mathbf{R}\tilde{\beta}^* - \mathbf{r})' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} (\mathbf{R}\tilde{\beta}^* - \mathbf{r})}{q\tilde{\sigma}_0^2} \\ &= \frac{RRSS - URSS}{URSS} \left( \frac{T - k^*}{q} \right) \end{aligned} \quad (23.84)$$

(see Chapter 20). The problem, however, is that since the distribution of  $\tilde{\beta}^*$  is no longer normal we cannot deduce the distribution of  $FT^*(\mathbf{y})$  as  $F(q, T - k^*)$  under  $H_0$ . On the other hand, we could use the asymptotic distribution of  $\beta^*$ , i.e.

$$\sqrt{T}(\tilde{\beta}^* - \beta^*) \underset{\alpha}{\sim} N(0, \sigma^2 \mathbf{G}^{-1}) \quad (23.85)$$

in order to deduce that

$$qFT^*(\mathbf{y}) \underset{\alpha}{\overset{H_0}{\sim}} \chi^2(q).$$

Using this result we could justify the use of the  $F$ -type test statistic (84) as an approximation of the chi-square test based on (85). Indeed, Kiviet (1981) has shown that in practice the  $F$ -type approximate test might be preferable in the present context because of the presence of a large number of regressors.

In order to illustrate the wide applicability of the above  $F$ -type specification test let us consider the simplest form of the DLR model's statistical GM where  $k=1$  and  $m=1$ :

$$y_t = \beta_0 x_t + \beta_1 x_{t-1} + \alpha_1 y_{t-1} + u_t. \quad (23.86)$$

Despite its simplicity (86) incorporates a large proportion of empirical

552      **The dynamic linear regression model**

econometric models in the applied econometric literature. In particular, following Hendry and Richard (1983), we can consider at least nine special cases of (86) where certain restrictions among the coefficients  $\beta_0$ ,  $\beta_1$  and  $\alpha_1$  are imposed:

Case 1. *Static regression* ( $\beta_1 = \alpha_1 = 0$ ):

$$y_t = \beta_0 x_t + u_t. \quad (23.87)$$

Case 2. *Autoregressive of order one* (AR(1)) ( $\beta_0 = \beta_1 = 0$ ):

$$y_t = \alpha_1 y_{t-1} + u_t. \quad (23.88)$$

Case 3. *Growth rate model* ( $\alpha_1 = 1$ ,  $\beta_0 = -\beta_1$ ):

$$\Delta y_t = \beta_0 \Delta x_t + u_t. \quad (23.89)$$

Case 4. *Leading indicator model* ( $\beta_0 = \alpha_2 = 0$ ):

$$y_t = \beta_1 x_{t-1} + u_t. \quad (23.90)$$

Case 5. *Finite distributed lag model* ( $\alpha_1 = 0$ ):

$$y_t = \beta_0 x_t + \beta_1 x_{t-1} + u_t. \quad (23.91)$$

Case 6. *Partial adjustment model* ( $\beta_1 = 0$ ):

$$y_t = \beta_0 x_t + \alpha_1 y_{t-1} + u_t. \quad (23.92)$$

Case 7. *Error-correction model* ( $\beta_0 + \beta_1 + \alpha_1 = 1$ ):

$$\Delta y_t = \beta_0 \Delta x_t + (1 - \alpha_1)(x_{t-1} - y_{t-1}) + u_t. \quad (23.93)$$

Case 8. ‘*Dead-start*’ model ( $\beta_0 = 0$ ):

$$y_t = \beta_1 x_{t-1} + \alpha_1 y_{t-1} + u_t. \quad (23.94)$$

For the above eight cases the restrictions imposed are all linear restrictions which can be tested using the test statistic (84) in conjunction with the rejection region

$$C_1 = \{y: FT(y) > c_\alpha\}$$

where  $c_\alpha$  is determined by

$$\alpha = \int_{c_\alpha}^{\infty} dF(q, T - k^*).$$

An important family of restrictions considered extensively in Chapter 22 are the common factor restrictions. In relation to (86) the common factor restriction is  $\alpha_1 \beta_0 + \beta_1 = 0$ , which gives rise to the special case of an autoregressive error model.

Case 9. Autoregressive error model ( $\alpha_1\beta_0 + \beta_1 = 0$ ):

$$y_t = \beta_0 x_t + \varepsilon_t, \quad \varepsilon_t = \alpha_1 \varepsilon_{t-1} + u_t. \quad (23.95)$$

For further discussion on all nine cases see Hendry and Richard (1983), and Hendry, Pagan and Sargan (1984).

In practice, the construction ('design') of an empirical econometric model taxes the ingenuity and craftsmanship of the applied econometrician more than any other part of econometric modelling. There are no rules or established procedures which automatically reduce any well-defined ('correctly' specified) estimated statistical GM to a 'proper' empirical econometric model. This is mainly because both economic theory as well as the properties of the sample data play a role in the choice ('design') of the latter. In order to illustrate this, let us return to the statistical GM for a money equation estimated in Section 23.2. In Section 23.3 this estimated equation was tested for any possible misspecifications and none of the underlying assumptions tested was rejected. The natural question to ask at this stage is 'assuming that this estimated statistical GM constitutes a well-defined statistical model, how do we proceed to specify (choose) an empirical econometric model?'

As it stands, the money equation estimated in Section 23.2 does not have any direct economic interpretation. The estimated parameters can only be viewed as well-defined statistical parameters. In order to be able to proceed with the 'design' of an empirical econometric model we need to consider the question of the estimable form of the theoretical model, in view of the observed data used to estimate the statistical GM (see Chapter 1). In the case of the money equation estimated in Section 23.2 we need to decide whether the theoretical model of a transactions demand for money considered in Chapter 19 could coincide with the estimable model. Demand in the context of a theoretical model is a theoretical concept which refers to the intentions of economic agents corresponding to a range of hypothetical values for the variables affecting their intentions. On the other hand, the observed data chosen refer to actual money stock M1 and there is no reason why the two should coincide for all time periods. Moreover, the other variables used in the context of the theoretical model are again theoretical constructs and should not be uncritically assumed to coincide with the observed data chosen. In view of these comments one should be careful in 'searching' for a demand for money function. In particular the assumption that the theory accounts for all the information in the data apart from a white-noise term is highly questionable in the present case.

In the case of the estimated money equation the special case of a static demand for money equation can be easily tested by testing for the significance of the coefficients of all the lagged variables using the *F*-type

test considered above. That is, test the null hypothesis  $H_0: \alpha=0$  and  $\beta_i=0$ ,  $i=1, 2, 3, 4$ , against  $H_1: \alpha \neq 0$  or  $\beta_i \neq 0$ ,  $i=1, 2, 3, 4$ . The test statistic for this hypothesis is

$$FT(y) = \left( \frac{0.11650 - 0.01053}{0.01053} \right) \left( \frac{53}{19} \right) = 28.072. \quad (23.96)$$

Given that  $c_\alpha = 1.76$  for  $\alpha = 0.05$  we can deduce that  $H_0$  is strongly rejected. A moment's reflection suggests that this is hardly surprising given that what the observed data refer to are not intentions or hypothetical range of values, but realisations. That is, what we observe is in effect the actual *adjustment process* for money stock M1 and not the original intentions. Hence, without any further information the estimable form of the model could only be a money adjustment equation which can be dominated by the demand, supply or even institutional factors. The latter question can only be decided by the data in conjunction with further a priori information.

Having decided that the estimable model is likely to be an adjustment process rather than a demand function we could proceed with the 'design' of the empirical econometric model. Using previous studies related to adjustment equations, without actually calling them as such (see Davidson *et al.* (1978), Hendry (1980), Hendry and Ungern-Sternberg (1981), Hendry (1983), Hendry and Richard (1983)), the following empirical econometric model was chosen:

$$\begin{aligned} \Delta \ln \left( \frac{M}{P} \right)_t &= -0.134 - 0.474 \left( \frac{1}{3} \sum_{i=1}^4 \Delta \ln \left( \frac{M}{P} \right)_{t-i} \right) \\ &\quad (0.02) \quad (0.130) \\ &- 0.196 \left( \ln \left( \frac{M}{P} \right)_{t-1} - \ln Y_{t-1} \right) + 1.239 \left( \frac{1}{3} \sum_{i=0}^3 \Delta \ln Y_{t-i} \right) \\ &\quad (0.022) \quad (0.314) \\ &- 0.801 \Delta \ln P_t - 0.059 \ln I_t - 0.025 \sum_{i=1}^4 (-1)_i \ln I_{t-i} \\ &\quad (0.145) \quad (0.007) \quad (0.008) \\ &- 0.045 Q_{1t} + 0.059 Q_{2t} + 0.053 Q_{3t} + \hat{u}_t, \end{aligned} \quad (23.97)$$

$$R^2 = 0.758, \quad \bar{R}^2 = 0.725, \quad s = 0.0137,$$

$$\log L = 223.318, \quad RSS = 0.01247, \quad T = 76.$$

All the restrictions imposed on the estimated statistical GM in order to reduce it to the above empirical econometric model are linear and the relevant *F*-type test statistic is

$$FT(\mathbf{y}) = \left( \frac{0.012\ 476 - 0.010\ 530}{0.010\ 530} \right) \left( \frac{53}{13} \right) = 0.753. \quad (23.98)$$

Given that  $c_\alpha = 1.88$  for a size  $\alpha = 0.05$  test we can deduce that these restrictions are not rejected. This test, however, does not suffice by itself to establish the validity of (97) as a well-defined statistical model as well. For this we need to ensure that the misspecification test results of the original estimated statistical GM are maintained by (97).

- (i) *Choice of m* – testing for  $m = 4$  against  $m = 6$ . Using the *F*-type test with  $RRSS = 0.012\ 379$  and  $URSS = 0.011\ 342$  the value of the test statistic is

$$FT(\mathbf{y}) = \frac{0.012\ 379 - 0.011\ 342}{0.011\ 342} \left( \frac{56}{8} \right) = 0.640$$

with

$$c_\alpha = 2.11, \alpha = 0.05,$$

the null hypothesis  $m = 4$  is not rejected. The Lagrange multiplier error autocorrelation test for  $l = 2$  in its *F*-type form yielded:

$$FT(\mathbf{y}) = \frac{0.012\ 379 - 0.012\ 177}{0.012\ 177} \left( \frac{62}{2} \right) = 0.514 \quad (23.99)$$

with  $c_\alpha = 3.15, \alpha = 0.05$ .

- (ii) *Misspecification test for normality – skewness–kurtosis test*. With  $\hat{\alpha}_3 = 0.3433$  and  $\hat{\alpha}_4 - 3 = -0.2788$  the test statistic is

$$\tau(\mathbf{y}) = \frac{T}{6} \hat{\alpha}_3^2 + \frac{T}{24} (\hat{\alpha}_4 - 3)^2 = 1.801 \quad (23.100)$$

and with  $c_\alpha = 5.99$  for  $\alpha = 0.05$  the assumption of normality is not rejected.

- (iii) *Misspecification tests for linearity and homoskedasticity*

- (a) RESET test with  $\hat{y}_t^2, \hat{y}_t^3, \hat{y}_t^4$ :

$$FT(\mathbf{y}) = 0.685, \quad c_\alpha = 2.76.$$

- (b) White test:

$$FT(\mathbf{y}) = 1.18, \quad c_\alpha = 1.75.$$

- (c) ARCH test:

$$FT(\mathbf{y}) = 0.531, \quad c_\alpha = 2.51.$$

These tests indicate no misspecification at  $\alpha = 0.05$ .

556      **The dynamic linear regression model**

- (iv) *Structural change tests* (see Section 21.6). It might be interesting to test the hypothesis that the changes ‘picked up’ by the dummy variables are indeed temporary changes without any lasting effects as assumed and not important structural changes. The first dummy variable was defined by

$$D_1 = 1, \quad t = 36, \quad D_1 = 0 \quad \text{for } t \neq 36, \quad t = 5, 6, \dots, 80.$$

Using  $T_2 = 37$  the *F*-type test for  $H_0^{(2)}$ :  $\sigma_1^2 = \sigma_2^2$  yielded

$$FT_1(y) = \frac{0.016\ 00}{0.014\ 549} = 1.1, \quad (23.103)$$

which for  $\alpha = 0.05$ ,  $c_\alpha = 1.85$ , implies that  $H_0^{(2)}$  is not rejected. Given this result we can proceed to test  $H_0^{(1)}$ :  $\beta_1 = \beta_2$ . The *F*-type test statistic is

$$FT_2(y) = \frac{0.013\ 749 - 0.013\ 678\ 3}{0.013\ 678\ 3} \left( \frac{60}{6} \right) = 0.0517, \quad (23.104)$$

given that  $c_\alpha = 2.254$  for  $\alpha = 0.05$ ,  $H_0^{(1)}$  is strongly accepted. Using the same procedure for  $T_2 = 51$ ,

$$FT_1(y) = \frac{0.016\ 278}{0.013\ 356} = 1.22, \quad (23.105)$$

$$c_\alpha = 1.83, \quad \alpha = 0.05$$

and

$$FT_2(y) = \frac{0.013\ 749 - 0.012\ 867}{0.012\ 867} \left( \frac{60}{6} \right) = 0.685,$$

$$c_\alpha = 2.254, \quad \alpha = 0.05. \quad (23.106)$$

Hence,  $H_0$ :  $\beta_1 = \beta_2$  and  $\sigma_1^2 = \sigma_2^2$  is accepted at  $\alpha^* = 1 - (1 - \alpha)^2 = 0.0975$ .

In Chapter 21 we used the distinction between structural change and parameter invariance with the former referring to the case where the point of change is known *a priori*. Let us consider time invariance in relation to (97) as well.

A very important property for empirical econometric models when needed for prediction or policy analysis is the time invariance of the estimated coefficients. For this reason it is necessary to consider the time invariance of the model in order to check whether the original time invariance exemplified by the estimated coefficients of the statistical GM has been lost or not. In ‘designing’ the empirical econometric model we try to capture the invariant features of the observable phenomena in order for

the model to have a certain value for prediction and policy analysis purposes. Hence, if the model has been designed at the expense of the time invariance the estimated statistical GM will be of very little value.

The recursive estimates of  $[\sum_{j=1}^4 \Delta(m_{t-j} - p_{t-j})]$ ,  $(m_{t-1} - p_{t-1} - y_{t-1})$ ,  $(\sum_{j=0}^3 \Delta y_{t-j})$ ,  $\Delta p_t$ ,  $i_t$ , and  $\sum_{j=1}^4 (-1)^j i_{t-j}$  are shown in Fig. 23.4(a)–(f) respectively for the period 1969*i*–1982*iv*. Apart from some initial volatility due to insufficient sample information these estimates show remarkable time constancy. The estimated theoretical parameters of interest have indeed preserved the time invariance exemplified by statistical parameters of interest in Section 23.3.

It is important to note that the above misspecification tests are not ‘proper’ tests in the same sense as in the context of the statistical GM. They should be interpreted as ‘diagnostic checks’ in order to ensure that the determination of the empirical econometric model was not achieved at the expense of the correct specification assumption. This is because in ‘designing’ the empirical econometric model from the estimated statistical GM we need to maintain the statistical properties which ensure that the end product is not just an economically meaningful estimated equation but a well-defined statistical model as well. Having satisfied ourselves that (97) is indeed well defined statistically we can proceed with its economic meaning as a money adjustment equation.

The main terms of the estimated adjustment equation are:

- (i)  $(\frac{1}{3} \sum_{i=1}^4 \Delta \ln(M/P)_{t-i})$  – the average annual rate of growth of real money stock;
- (ii)  $(\ln(M/P)_{t-1} - \ln Y_{t-1})$  – the error-correction term (see Hendry (1980));
- (iii)  $(\frac{1}{3} \sum_{i=1}^4 \Delta \ln Y_{t-i})$  – the average annual rate of growth of real consumers’ expenditure;
- (iv)  $\Delta \ln P_t$  – inflation rate;
- (v)  $\ln I_t$  – interest rate (7 days’ deposit account);
- (vi)  $\sum_{i=1}^4 (-1)^i \ln I_{t-i}$  – annual polynomial lag for interest rate.

Interpreting (97) as a money adjustment equation we can see that both the rate of interest and consumers’ expenditure play an important role in the determination of the changes in real money stock. As far as inflation is concerned we can see that the restriction for its coefficient to be equal to minus one against being less than minus one (one-sided test) is not rejected at  $\alpha=0.05$  given that  $c_2 = -1.67$  and the test statistic is

$$\tau(y) = \frac{-0.801 + 1.00}{0.145} = -0.137. \quad (23.107)$$

This implies that in effect the inflation rate term cancels out from both sides

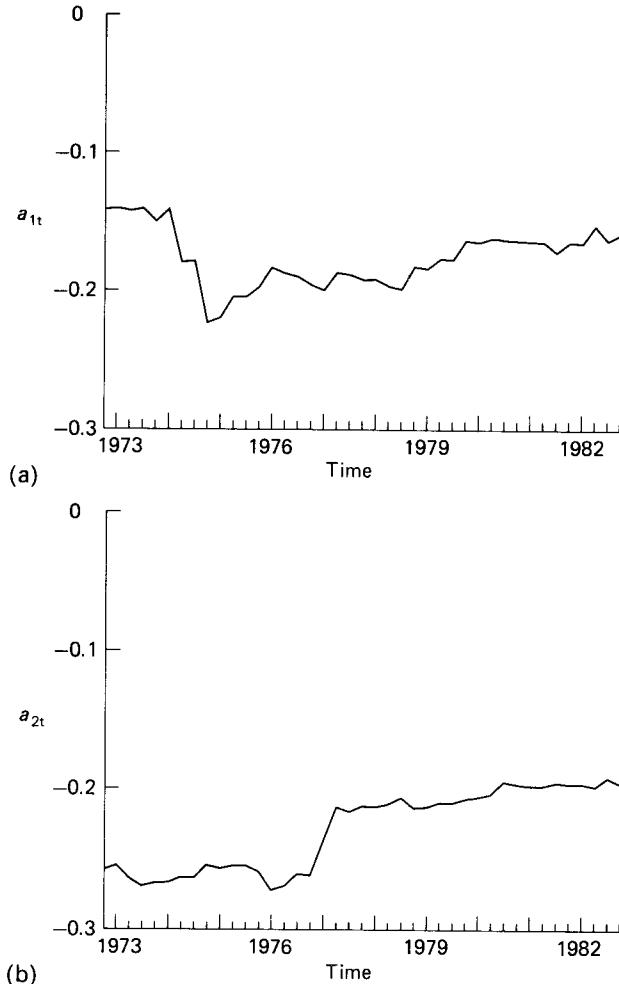
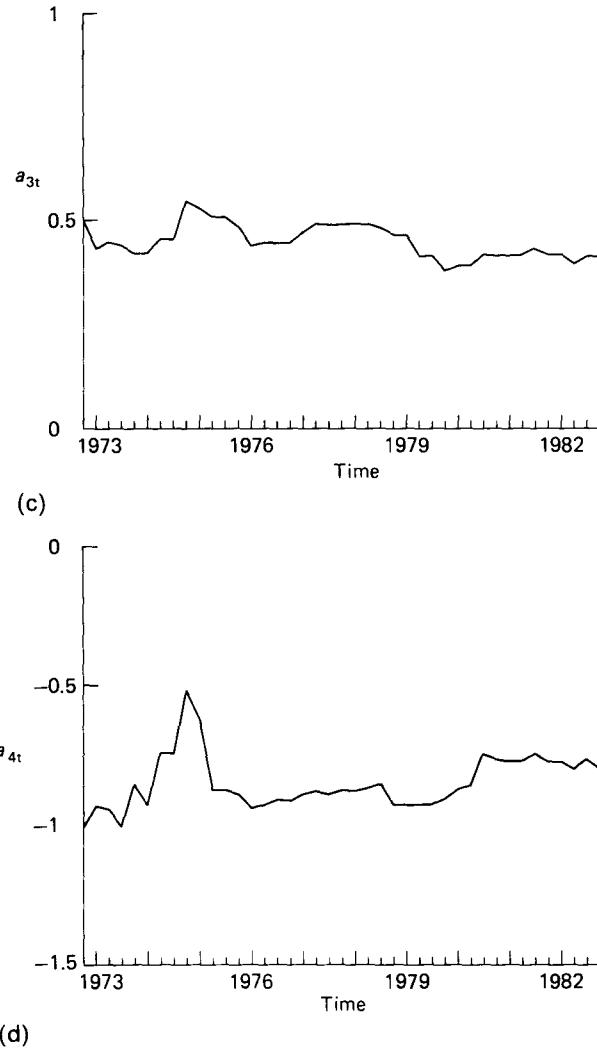


Fig. 23.4(a)–(f). The time paths of the recursive estimates of the coefficients of  $(\sum_{j=1}^4 \Delta \ln (M/P)_{t-j})$ ,  $(\ln (M/PY)_{t-1})$ ,  $(\sum_{j=1}^3 \Delta \ln Y_{t-j})$ ,  $\ln I_t$  and  $(\sum_{j=1}^4 (-1)^j \ln I_{t-j})$  respectively (from (97)).

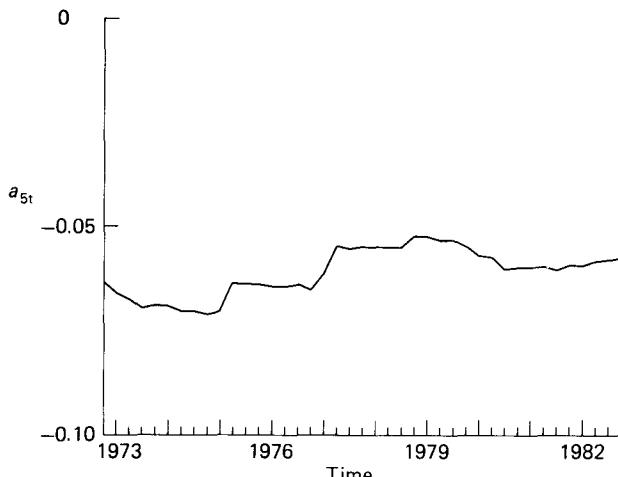
with short-run behaviour being largely determined in nominal terms; not an unreasonable proposition. The long-run represented by the static solution (assuming that  $y_t = y_{t-i}, \mathbf{x}_t = \mathbf{x}_{t-i}, i = 1, 2, \dots$ ) of the adjustment equation is

$$\left( \frac{M}{PY} \right) = AI^{-0.301}(1+p)^{-4.087}, \quad (23.108)$$

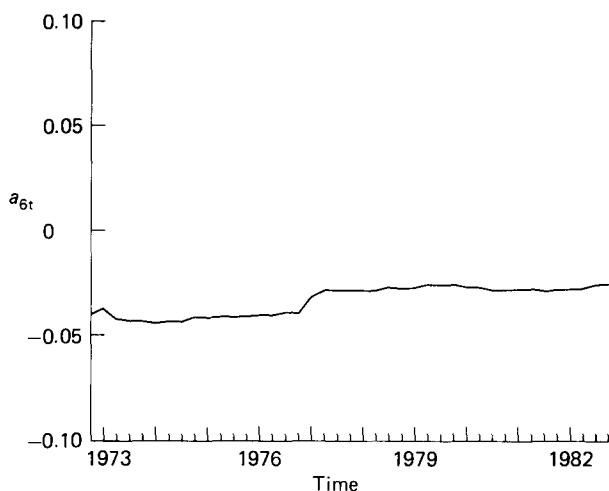
Fig. 23.4. *continued*

where  $A$  is a constant (see Hendry (1980)). This suggests that the long-run behaviour differs from the short-run in so far as the former is related to real money stock.

The question which arises at this stage is, 'how is the estimated adjustment equation related to the theoretical model of a transactions demand for money?' From (97) we can see that the adjustment equation is dominated by the demand side in view of the signs of  $\sum_{i=1}^4 \Delta \ln Y_{t-i}, \ln I_t$



(e)



(f)

Fig. 23.4. *continued*

and  $\sum_{i=1}^4 (-1)^i \ln I_{t-i}$ . This suggests that if we were to assume that the supply side is perfectly elastic then the equilibrium state, where 'no inherent tendency to change' exists, can be related directly to (108). Hence, in view of the perfect elasticity of supply (108) can be interpreted as a transactions demand for money equation. For a more extensive discussion of adjustment processes, equilibrium and demand, supply functions, see Hendry and Spanos (1980).

Re-estimation of (97) with  $\Delta p_t$  excluded from both sides yielded the following more parsimonious empirical model:

$$\begin{aligned} \Delta m_t = & -0.124 - 0.485 \left( \frac{1}{3} \sum_{j=1}^4 \Delta(m_{t-j} - p_{t-j}) \right) - 0.202(m_{t-1} - p_{t-1} - y_{t-1}) \\ & + 1.197 \left( \frac{1}{3} \sum_{j=0}^3 \Delta y_{t-j} \right) - 0.056 i_t - 0.025 \sum_{j=1}^4 (-1)^j i_{t-j} \\ & + \text{dummies} + \hat{u}_t, \end{aligned} \quad (23.109)$$

(0.014)

$$R^2 = 0.709, \quad \bar{R}^2 = 0.676, \quad s = 0.01384,$$

$$\log L = 222.25, \quad T = 76.$$

The above estimated coefficients can be interpreted as estimates of the theoretical parameters of interest defining the money adjustment equation. These parameters are simple functions of the statistical parameters of interest defining the statistical GM. An interesting issue in the context of this distinction between theoretical and statistical parameters of interest is related to the problem of 'near' collinearity or/and short data (collectively called insufficient data information) raised in Chapter 20.

In view of the large number of estimated parameters involved in the money statistical GM one might be forgiven for suspecting that insufficient data information problems might be affecting the statistical parametrisation estimated in Section 23.3. One of the aims of econometric modelling, however, is to 'design' robust estimated coefficients which are directly related to the theoretical parameters of interest. For this reason it will be interesting to consider the correlation matrix of the estimated coefficients as a rough guide to such robustness, see Table 23.2.

The correlations among the coefficients of the regressors are relatively small; none is greater than 0.81 with only one greater than 0.68. These correlations suggest that most of the estimated (theoretical) parameters of interest are nearly orthogonal; an important criterion of a 'good design'. The first column of Table 23.2 shows the partial correlation coefficients (see Section 20.6) between  $\Delta m_t$  and the regressors with the numbers in parentheses underneath referring to the simple correlation coefficients.

The values of the partial correlation coefficients show that every regressor contributes substantially to the explanation of  $\Delta m_t$  with the error-correction term and the interest rate playing a particularly important role.

Another important feature of the empirical model 'designed' above (see

Table 23.2. Orthogonality of the explanatory variables

$\Delta m_t$	Partial correlation	Correlation matrix of coefficients			
$\left( \frac{1}{3} \sum_{j=1}^4 \Delta(m_{t-j} - p_{t-j}) \right)$	-0.413 (-0.053)				
$(m_{t-1} - p_{t-1} - y_{t-1})$	-0.754 (-0.418)	0.078			
$\left( \frac{1}{3} \sum_{j=0}^3 \Delta y_{t-j} \right)$	0.422 (0.056)	-0.625	-0.035		
$i_t$	-0.701 (-0.080)	0.161	0.809	0.053	
$\sum_{j=1}^4 (-1)^j i_{t-j}$	-0.342 (0.264)	-0.077	0.611	0.093	0.647

Table 23.3. Restricted coefficient estimates based on (97)

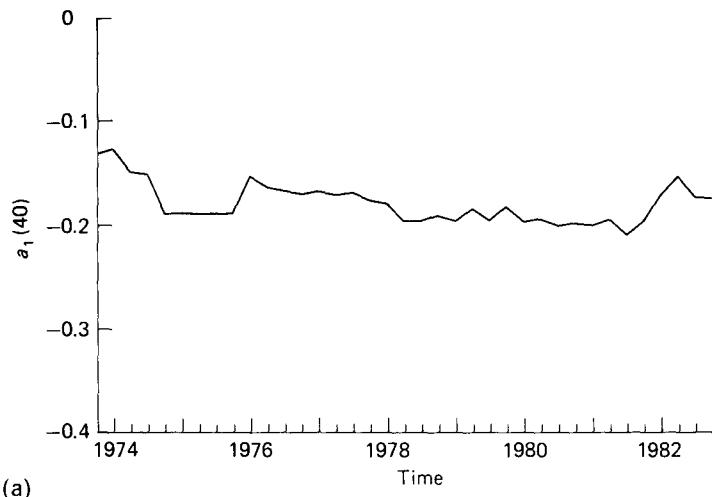
	$j=0$	$j=1$	$j=2$	$j=3$	$j=4$
$m_{t-j}^*$		0.354	0	0	0.158
$y_{t-j}$	0.413	0.196	0	-0.413	0
$p_{t-j}$	-0.801	0.801	0	0	0
$i_{t-j}$	-0.059	0.025	-0.025	0.025	-0.025

(109)) is that the restricted coefficient estimates do not differ significantly from the unrestricted estimates as can be seen from Table 23.3.

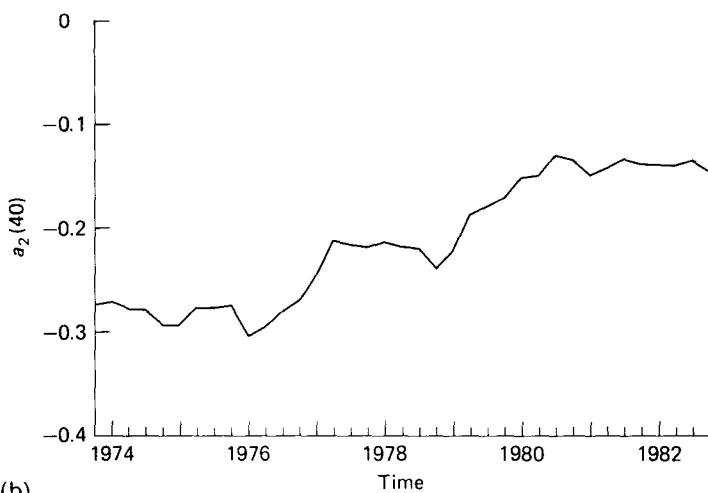
Further evidence of the constancy of the estimated coefficients is given in Fig. 23.5(a)–(f) where the 40-observation ‘window’ estimates are plotted. These estimates, based on a fixed sample size of 40 observations, run through the whole sample, that is,  $\beta_1$  based on observations 1–40,  $\beta_2$  on 2–41,  $\beta_3$  on 3–42, etc.

### 23.5 Prediction

An important question to consider in the context of the dynamic linear regression (DLR) model is to predict the value of  $y_{T+1}$  given the sample information for the period  $t = 1, 2, \dots, T$ . As argued in Chapter 13, the best predictor (in mean square error (MSE) sense) is given by the conditional expectation of  $y_{T+1}$  given the relevant information set. For simplicity let

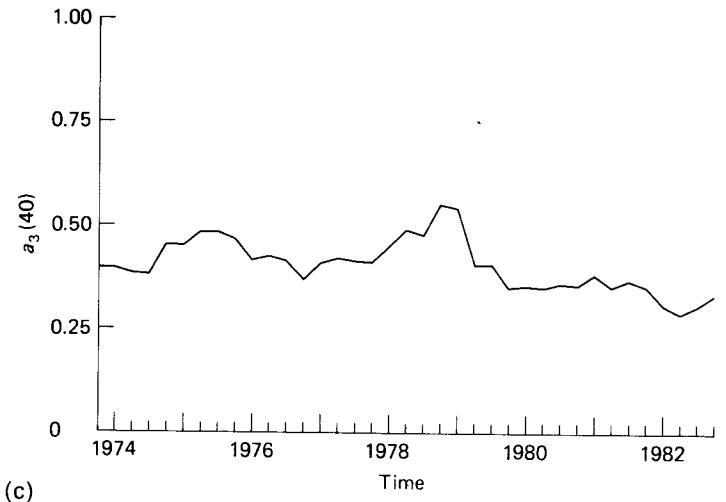


(a)

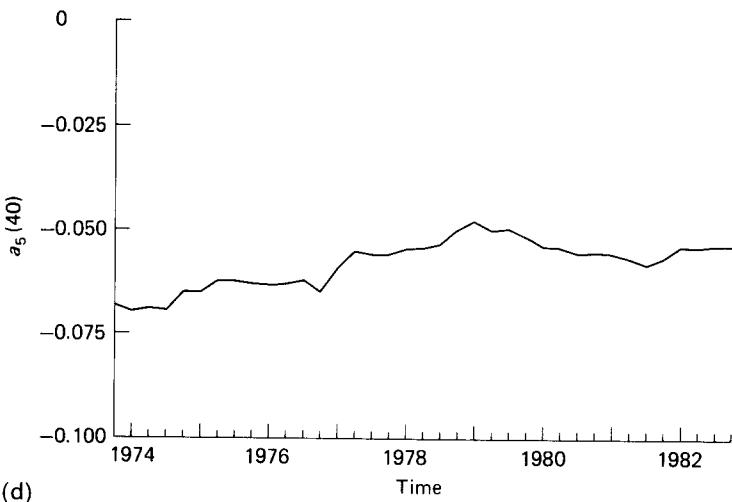


(b)

Fig. 23.5(a)–(e). The time paths of 40 observation window estimates of the coefficients of (109) apart from the constant.

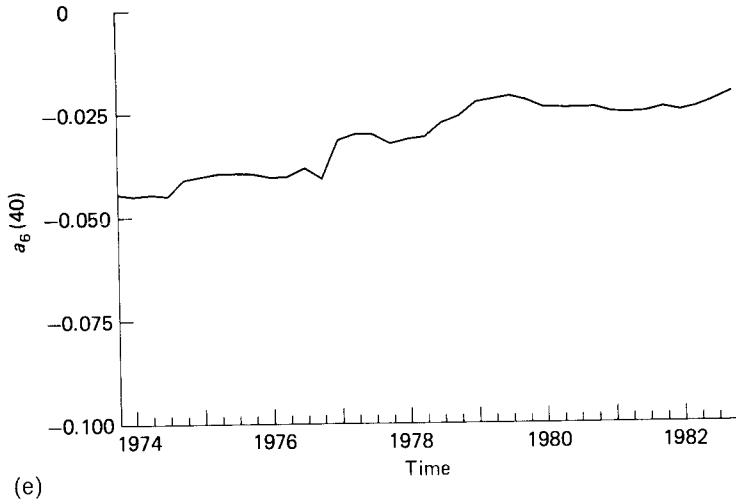


(c)



(d)

Fig. 23.5. *continued*

Fig. 23.5. *continued*

us assume that  $\mathbf{x}_{T+1}^0 \equiv (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T, \mathbf{x}_{T+1})'$  is given and the parameters  $\theta^* \equiv (\alpha_1, \alpha_2, \dots, \alpha_m, \beta_0, \beta_1, \dots, \beta_m, \sigma_0^2)$  are known. In such a case the best predictor of  $y_{T+1}$  is given by its systematic component, i.e.

$$\begin{aligned}\mu_{T+1} &\equiv E(y_{T+1}/\sigma(\mathbf{Y}_T^0), \mathbf{X}_{T+1}^0 = \mathbf{x}_{T+1}^0) \\ &= \sum_{i=1}^m \alpha_i y_{T+1-i} + \sum_{i=0}^m \beta'_i \mathbf{x}_{T+1-i}.\end{aligned}\quad (23.110)$$

Moreover, the prediction error is

$$y_{T+1} - \mu_{T+1} = u_{T+1}, \quad (23.111)$$

$$\text{MSE}(u_{T+1}) = E(u_{T+1}^2/\sigma(\mathbf{Y}_T^0), \mathbf{X}_{T+1}^0 = \mathbf{x}_{T+1}^0) = \sigma_0^2, \quad (23.112)$$

and

$$E(u_{T+1}/\sigma(\mathbf{Y}_T^0), \mathbf{X}_{T+1}^0 = \mathbf{x}_{T+1}^0) = 0.$$

Similarly the best predictor of  $y_{T+2}$  given a similar information set is

$$\mu_{T+2} = \alpha_1 \mu_{T+1} + \sum_{i=2}^m \alpha_i y_{T+2-i} + \sum_{i=0}^m \beta'_i \mathbf{x}_{T+2-i}. \quad (23.113)$$

In general we can construct predictors for  $l$  periods ahead by substituting

the predictors  $\mu_{T+i}$ ,  $i = 1, 2, \dots, l-1$ , into (110), i.e.

$$\mu_{T+l} = \sum_{i=1}^{l-1} \alpha_i \mu_{T+l-i} + \sum_{i=l}^m \alpha_i y_{T+l-i} + \sum_{i=0}^m \beta'_i \mathbf{x}_{T+l-i}, \quad l=2, 3, \dots, m \quad (23.114)$$

and

$$\mu_{T+l} = \sum_{i=1}^m \alpha_i \mu_{T+l-i} + \sum_{i=0}^m \tilde{\beta}'_i \mathbf{x}_{T+l-i}, \quad l=m+1, m+2, \dots \quad (23.115)$$

The above predictors were derived on the assumption that  $\theta^*$  was known a priori, a grossly unrealistic assumption. In the case where  $\theta^*$  is not known we use  $\tilde{\theta}^*$ , the approximate MLE and the predictor of  $y_{T+l}$  takes the form

$$\tilde{\mu}_{T+l} = \sum_{i=1}^{l-1} \tilde{\alpha}_i \tilde{\mu}_{T+l-i} + \sum_{i=l}^m \tilde{\alpha}_i y_{T+l-i} + \sum_{i=0}^m \tilde{\beta}'_i \mathbf{x}_{T+l-i}, \quad l=2, 3, \dots, m \quad (23.116)$$

and

$$\tilde{\mu}_{T+l} = \sum_{i=1}^m \tilde{\alpha}_i \mu_{T+l-i} + \sum_{i=0}^m \tilde{\beta}'_i \mathbf{x}_{T+l-i}, \quad l=m+1, \dots \quad (23.117)$$

In the case where  $\mathbf{x}_{T+l}$ ,  $l=1, 2, \dots$ , is not known it has to be ‘guessed’ and substituted in the above formulae.

Returning to (110), we can see that the prediction error variance is given by (111), i.e.

$$E[(y_{T+1} - \mu_{T+1})^2 / \sigma(\mathbf{Y}_T^0), \mathbf{X}_{T-1}^0 = \mathbf{x}_{T-1}^0] = \sigma_0^2. \quad (23.118)$$

Similarly,

$$E[(y_{T+2} - \mu_{T+2})^2 / \sigma(\mathbf{Y}_T^0), \mathbf{X}_{T+1}^0 = \mathbf{x}_{T+1}^0] = \alpha_1^2 \sigma_0^2 + \sigma_0^2 \quad (23.119)$$

and

$$E[(y_{T+l} - \mu_{T+l})^2 / \sigma(\mathbf{Y}_T^0), \mathbf{X}_{T+l}^0 = \mathbf{x}_{T+l}^0] = \sum_{i=1}^{l-1} \alpha_i^2 \sigma_0^2 + \sigma_0^2, \quad l=2, 3, \dots \quad (23.120)$$

$$= \sum_{i=1}^m \alpha_i^2 \sigma_0^2 + \sum_{i=m+1}^l \sigma_0^2, \quad l=m+1, \dots \quad (23.121)$$

The corresponding formulae for the case where  $\tilde{\theta}^*$  is used instead of  $\theta^*$ , when the latter is unknown, are rather complicated. In practice, however, since  $\tilde{\theta}^*$  is a consistent estimator of  $\theta^*$  the formulae (118)–(121) are also used (but only asymptotically valid) in the case where the estimated coefficients are used.

For hypothesis testing and interval estimation related to prediction

errors the asymptotic distribution is given by

$$(y_{T+l} - \tilde{\mu}_{T+l}) \underset{x}{\sim} N\left(0, \sum_{i=1}^l \alpha_i^2 \sigma_0^2\right), \quad l = 1, 2, \dots, m, \quad (23.122)$$

where  $\alpha_l = 1$ . Note that the predictors  $\tilde{\mu}_{T+1}, \tilde{\mu}_{T+2}, \dots, \tilde{\mu}_{T+l}$  are correlated and any joint testing or confidence estimation needs to take the covariance into consideration as well.

### 23.6 Looking back

In Section 23.4 we constructed an empirical econometric model of a money adjustment equation based on a well-defined statistical GM estimated in Section 23.2. It will be useful at this stage to return to the money equation estimated in the context of the linear regression model (see Chapter 19) and attempt to ‘explain’ the various statistical inference ‘results’ derived in Chapters 19–22. For convenience let us reproduce both equations estimated for the same sample period 1964*i*–1982*iv*:

$$\begin{aligned} \Delta m_t &= -0.124 - 0.485 \left( \frac{1}{3} \sum_{j=1}^4 \Delta(m_{t-j} - p_{t-j}) \right) \\ &\quad (0.018) \quad (0.130) \\ &\quad - 0.202(m_{t-1} - p_{t-1} - y_{t-1}) + 1.197 \left( \frac{1}{3} \sum_{j=0}^3 \Delta y_{t-j} \right) \\ &\quad (0.022) \quad (0.314) \\ &\quad - 0.056i_t - 0.025 \sum_{j=1}^4 (-1)^j i_{t-j} + \text{dummies} + \hat{u}_t, \\ &\quad (0.007) \quad (0.008) \quad (0.014) \end{aligned} \quad (23.123)$$

$$R^2 = 0.709, \quad \bar{R}^2 = 0.675, \quad s = 0.01384,$$

$$\log L = 222.25, \quad RSS = 0.012832, \quad T = 76.$$

Note that small letters are used to denote the natural logs of the variables represented by the capital letters.

$$m_t = 2.763 + 0.705y_t + 0.862p_t - 0.053i_t + \hat{e}_t, \quad (23.124)$$

$$(1.100) \quad (0.112) \quad (0.022) \quad (0.014) \quad (0.040)$$

$$R_2 = 0.995, \quad \bar{R}^2 = 0.995, \quad s = 0.04022,$$

$$\log L = 138.425, \quad RSS = 0.1165, \quad T = 76.$$

Looking at these estimated equations we can see how grossly misspecified equation (124) is as a linear regression statistical GM. The inappropriate

assumption of an independent sample invalidates all the statistical inference results based on (124) derived in Chapter 19. Moreover, interpreting the estimated coefficients as elasticities and using these in any way is again unwarranted given that these are not well-defined statistical parameters and any arguments based on the estimated coefficients can be very misleading.

In terms of goodness of fit the estimated variance of (123) is almost a tenth of that of (124).

From the statistical viewpoint the main difference between (123) and (124) comes in the form of the unmodelled part of  $y_t$ . The residuals for the two estimated equations behave very differently. In Fig. 23.6,  $\hat{u}_t$  and  $\hat{\epsilon}_t$  are plotted over time and as we can see they differ greatly. Firstly, the standard deviation of  $\hat{u}_t$  is three times smaller than that of  $\hat{\epsilon}_t$ . Secondly, the white-noise assumption seems much more appropriate for  $u_t$  rather than  $\epsilon_t$ . From Chapters 19–22 we know that  $\hat{\epsilon}_t$  exhibit not only time dependency but heteroskedasticity as well. Hence, in no circumstances should (124) be interpreted as an empirical econometric model.

The main problem associated with (124) is that of invalid conditioning. That is, in defining the systematic component we should have conditioned not only on the ‘present’ ( $\mathbf{X}_t = \mathbf{x}_t$ ) but the past as well. This invalid conditioning induced time dependency in  $\hat{\epsilon}_t$ ,  $\hat{\sigma}_{\epsilon}^2$  as well as  $\hat{\beta}$ . Looking at (123) we can see how this problem can arise. Using the terminology

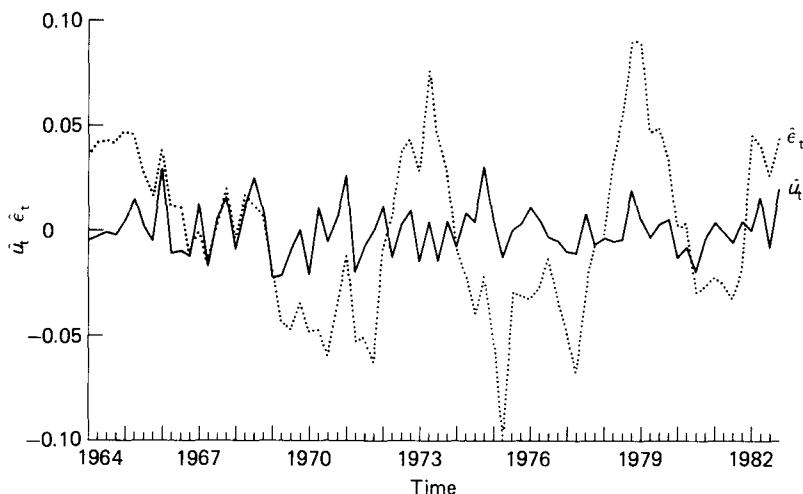


Fig. 23.6. Comparing the residuals from (123) and (124).

introduced by Hendry and Richard (1982) the empirical econometric model (123) ‘encompasses’ (124). Encompassing refers to the ability of a particular estimated statistical model to explain the results of rival models (see Mizon (1984) for an extensive discussion). The comparison of (123) and (124) above was in effect a simple exercise in encompassing. In this case, however, a formal discussion of encompassing seems rather unnecessary given that (124) could not be seriously entertained as an empirical econometric model given that it has failed almost all the misspecification tests applied.

Although the statistical foundations of (123) seem reasonably sound, its theoretical underpinnings are less obvious. This, however, is not surprising given that economic theory is rather embarrassingly silent on estimable adjustment equations. At this early stage it might be advisable to rely more heavily on data-based specifications which might provide the basis for more realistic theoretical formulations of estimable models. Muellbauer (1986) provides an excellent example of how economic theoretical arguments can be used to ‘rationalise’ and interpret successful data-based empirical econometric models. This seems a most promising way forward with economic theory and data-based specifications complementing each other; see Pagan (1985), Nickell (1985), Granger and Engle (1985).

The above discussion exemplifies the dangers of using an inappropriate statistical model in econometric modelling. At the outset it must have been obvious that the sampling model assumption of independence associated with the linear regression model was highly suspect for the kind of data chosen. Despite that, we adopted an inappropriate statistical model in order to illustrate the importance of the decision to adopt one in preference of other statistical models. Throughout the discussion in Chapters 17–23 every attempt has been made to persuade the reader that the nature and statistical properties of the observed data chosen have an important role to play in econometric modelling. These should be taken into consideration when the decision to adopt a particular statistical model in preference to the others is made. In a certain sense the choice of the statistical model to be used for the particular case of econometric modelling under consideration is one of the most important decisions to be taken by the modeller. Once an inappropriate choice is made quite a few misleading conclusions can be drawn unless the modeller is knowledgeable enough to put the estimated statistical GM through a battery of misspecification tests before embarking on specification testing and prediction. If, however, the modeller follows the naive methodological maxim that ‘the theory accounts for all the information in the data (irrespective of the choice of the data) apart from a white-noise error term’ or ‘only theoretical information is real information’, then the misspecification testing seems only of secondary importance and misleading conclusions are more than likely.

***Important concepts***

Autoregressive representation, homogeneous non-stationary processes, initial conditions, asymptotic independence, Granger non-causality, strong exogeneity, unit circle restrictions, approximate MLE, Durbin's  $h$  test, statistical versus theoretical parameters of interest, error correction term, long-run solution, partial correlation coefficients, encompassing.

***Questions***

1. Explain the role of the stationarity and asymptotic independence of the stochastic process  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  in the context of the specification of the DLR model.
2. Is the stationarity of  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  necessary for the autoregressive representation of the process?
3. Define the concept of strong exogeneity and explain its role in the context of the DLR model.
4. ‘The statistical GM of the DLR model is a hybrid of those for the linear and stochastic linear regression models.’ Explain.
5. Discuss the difference between the exact and approximate MLE’s of  $\theta^*$ . ‘Do they have the same asymptotic properties?’ Explain your answer.
6. State the asymptotic properties of the approximate MLE  $\tilde{\theta}^*$  of  $\theta^*$ .
7. ‘How do we test whether the maximum lag postulated in the specification of the statistical GM is too large?’
8. ‘How do we interpret residual autocorrelation in the context of an estimated statistical GM in the DLR model?’
9. ‘Why is the Durbin–Watson test inappropriate in testing for an AR(1) error term in the context of the DLR model?’

***Additional references***

Anderson (1959); Crowder (1980); Durbin (1960); Harvey (1981); Granger and Weiss (1983); Mann and Wald (1943a); Priestley (1981).

## CHAPTER 24

---

### The multivariate linear regression model

---

#### 24.1 Introduction

The multivariate linear regression model is a direct extension of the linear regression model to the case where the dependent variable is an  $m \times 1$  random vector  $\mathbf{y}_t$ . That is, the statistical GM takes the form

$$\mathbf{y}_t = \mathbf{B}'\mathbf{x}_t + \mathbf{u}_t, \quad t \in \mathbb{T}, \quad (24.1)$$

where  $\mathbf{y}_t: m \times 1$ ,  $\mathbf{B}: k \times m$ ,  $\mathbf{x}_t: k \times 1$ ,  $\mathbf{u}_t: m \times 1$ . The system (1) is effectively a system of  $m$  linear regression equations:

$$y_{it} = \beta_i' \mathbf{x}_t + u_{it}, \quad i = 1, 2, \dots, m, \quad t \in \mathbb{T}, \quad (24.2)$$

with  $\mathbf{B} = (\beta_1, \beta_2, \dots, \beta_m)$ .

In direct analogy with the  $m=1$  case (see Chapter 19) the multivariate linear regression model will be derived from first principles based on the joint distribution of the observable random variables involved,  $D(\mathbf{Z}_t; \psi)$  where  $\mathbf{Z}_t \equiv (\mathbf{y}_t', \mathbf{X}_t')'$ ,  $(m+k) \times 1$ . Assuming that  $\mathbf{Z}_t$  is an IID normally distributed vector, i.e.

$$\begin{pmatrix} \mathbf{y}_t \\ \mathbf{X}_t \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right) \quad \text{for all } t \in \mathbb{T}, \quad (24.3)$$

we can proceed to define the systematic and non-systematic components by:

$$\mu_t = E(\mathbf{y}_t / \mathbf{X}_t = \mathbf{x}_t) = \mathbf{B}'\mathbf{x}_t, \quad \mathbf{B} = \Sigma_{22}^{-1}\Sigma_{21} \quad (24.4)$$

and

$$\mathbf{u}_t = \mathbf{y}_t - E(\mathbf{y}_t / \mathbf{X}_t = \mathbf{x}_t), \quad t \in \mathbb{T}. \quad (24.5)$$

572      **The multivariate linear regression model**

Moreover, by construction,  $\mathbf{u}_t$  and  $\boldsymbol{\mu}_t$  satisfy the following properties:

$$(i) \quad E(\mathbf{u}_t) = E[E(\mathbf{u}_t / \mathbf{X}_t = \mathbf{x}_t)] = 0;$$

$$(ii) \quad E(\mathbf{u}_t \mathbf{u}'_s) = \mathbf{E}[E(\mathbf{u}_t \mathbf{u}'_s / \mathbf{X}_t = \mathbf{x}_t)] = \begin{cases} \boldsymbol{\Omega} & t = s \\ \mathbf{0} & t \neq s; \end{cases}$$

$$(iii) \quad E(\boldsymbol{\mu}_t \mathbf{u}'_t) = E[E(\boldsymbol{\mu}_t \mathbf{u}'_t / \mathbf{X}_t = \mathbf{x}_t)] = E[\boldsymbol{\mu}_t E(\mathbf{u}'_t / \mathbf{X}_t = \mathbf{x}_t)] = \mathbf{0}, \quad t \in \mathbb{T},$$

where  $\boldsymbol{\Omega} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}$  (compare these with the results in Section 19.2).

The similarity between the  $m = 1$  case and the general case allows us to consider several loose ends left in Chapter 19. The first is the use of the joint distribution  $D(\mathbf{Z}_t; \boldsymbol{\psi})$  in defining the model instead of concentrating exclusively on  $D(\mathbf{y}_t / \mathbf{X}_t; \boldsymbol{\psi}_1)$ . The loss of generality in postulating the form of the joint distribution is more than compensated for by the additional insight provided. In practice it is often easier to ‘judge’ the plausibility of assumptions relating to the nature of  $D(\mathbf{Z}_t; \boldsymbol{\psi})$  rather than  $D(\mathbf{y}_t / \mathbf{X}_t; \boldsymbol{\psi}_1)$ . Moreover, in misspecification analysis the relationship between the assumptions underlying the model and those underlying the random vector process  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  enhances our understanding of the nature of the possible departures. An interesting example of this is the relationship of the assumption that  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is a

- (1) normal ( $N$ );
- (2) independent ( $I$ ); and
- (3) identically distributed (ID) process; and
- [6] (i)  $D(\mathbf{y}_t / \mathbf{X}_t; \boldsymbol{\psi}_1)$  is normal;
- (ii)  $E(\mathbf{y}_t / \mathbf{X}_t = \mathbf{x}_t)$  is linear in  $\mathbf{x}_t$ ;
- (iii)  $\text{Cov}(\mathbf{y}_t / \mathbf{X}_t = \mathbf{x}_t)$  is homoskedastic (free of  $\mathbf{x}_t$ ); .
- [7]  $\boldsymbol{\theta} \equiv (\mathbf{B}, \boldsymbol{\Omega})$  are time-invariant;
- [8]  $\{\mathbf{y}_t / \mathbf{X}_t, t \in \mathbb{T}\}$  is an independent process.

The relationship between these components is shown diagrammatically below:

$$(N) \rightarrow \begin{cases} (i) \\ (ii), \quad (ID) \rightarrow [7], \quad (I) \rightarrow [8] \\ (iii) \end{cases}$$

The question which naturally arises is whether (i)–(iii) imply ( $N$ ) or not. The following lemma shows that if (i)–(iii) are supplemented by the assumption

that  $\mathbf{X}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{22})$ ,  $\det(\boldsymbol{\Sigma}_{22}) \neq 0$ , the reverse implication holds.

*Lemma 24.1*

$\mathbf{Z}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma})$  for  $t \in \mathbb{T}$  if and only if

- (i)  $\mathbf{X}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{22})$ ,  $\det(\boldsymbol{\Sigma}_{22}) \neq 0$ ;
- (ii)  $E(\mathbf{y}_t / \mathbf{X}_t = \mathbf{x}_t) = \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \mathbf{x}_t$ ;
- (iii)  $\text{Cov}(\mathbf{y}_t / \mathbf{X}_t = \mathbf{x}_t) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}$
- (iv)  $(\mathbf{y}_t / \mathbf{X}_t) \sim N(\mathbf{B}' \mathbf{X}_t, \boldsymbol{\Omega})$

(see Barra (1981)).

The statistical GM (1) for the sample period  $t = 1, 2, \dots, T$  is written as

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U}, \quad (24.6)$$

where  $\mathbf{Y}: T \times m$ ,  $\mathbf{X}: T \times k$ ,  $\mathbf{B}: k \times m$ ,  $\mathbf{U}: T \times m$ . The system in (1) can be viewed as the  $t$ th row of (6). The  $i$ th row taking the form

$$\mathbf{y}_i = \mathbf{X}\beta_i + \mathbf{u}_i, \quad i = 1, 2, \dots, m \quad (24.7)$$

represents all  $T$  observations on the  $i$ th regression in (2). In order to define the conditional distribution  $D(\mathbf{Y}/\mathbf{X}; \boldsymbol{\psi}_1)$  we need the special notation of Kronecker products (see Appendix 2). Using this notation the matrix distribution can be written in the form

$$(\mathbf{Y}/\mathbf{X} = \mathbf{X}) \sim N(\mathbf{XB}, \boldsymbol{\Omega} \otimes \mathbf{I}_T), \quad (24.8)$$

where  $\boldsymbol{\Omega} \otimes \mathbf{I}_T$  represents the covariance of

$$\text{vec}(\mathbf{Y}) = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_m \end{pmatrix} : Tm \times 1.$$

The vectoring operator  $\text{vec}(\cdot)$  transforms a matrix into a column vector by stacking the columns of the matrix one beneath the other. Using the vectoring operator we can express (6) in the form

$$\text{vec}(\mathbf{Y}) = (\mathbf{I}_m \otimes \mathbf{X}) \text{vec}(\mathbf{B}) + \text{vec}(\mathbf{U}) \quad (24.9)$$

or

$$\mathbf{y}^* = \mathbf{X}_* \beta_* + \mathbf{u}_* \quad (24.10)$$

in an obvious notation.

The multivariate linear regression (MLR) model is of considerable interest in econometrics because of its direct relationship with the simultaneous equations formulation to be considered in Chapter 25. In particular, the latter formulation can be viewed as a reparametrisation of the MLR model where the *statistical parameters* of interest  $\boldsymbol{\theta} \equiv (\mathbf{B}, \boldsymbol{\Omega})$  do not coincide with the *theoretical parameters* of interest  $\boldsymbol{\xi}$ . Instead, the two sets of parameters are related by some system of implicit equations of the form:

$$h_i(\boldsymbol{\theta}, \boldsymbol{\xi}) = 0, \quad i = 1, 2, \dots, p. \quad (24.11)$$

These equations can be interpreted as providing an alternative parametrisation for the statistical GM in terms of the theoretical parameters of interest. In view of this relationship between the two statistical models a sound understanding of the MLR model will pave the way for the simultaneous equations formulation in Chapter 25.

## 24.2 Specification and estimation

In direct analogy to the linear regression model ( $m = 1$ ) the multivariate linear regression model is specified as follows:

(I) **Statistical GM:**  $\mathbf{y}_t = \mathbf{B}'\mathbf{x}_t + \mathbf{u}_t, \quad t \in \mathbb{T}$

$$\mathbf{y}_t: m \times 1, \quad \mathbf{x}_t: k \times 1, \quad \mathbf{B}: k \times m.$$

[1] The systematic and non-systematic components are:

$$\boldsymbol{\mu}_t = E(\mathbf{y}_t / \mathbf{X}_t = \mathbf{x}_t) = \mathbf{B}'\mathbf{x}_t, \quad \mathbf{u}_t = \mathbf{y}_t - E(\mathbf{y}_t / \mathbf{X}_t = \mathbf{x}_t),$$

and by construction

$$E(\mathbf{u}_t) = E[E(\mathbf{u}_t / \mathbf{X}_t = \mathbf{x}_t)] = \mathbf{0},$$

$$E(\boldsymbol{\mu}_t' \mathbf{u}_t) = E[E(\boldsymbol{\mu}_t' \mathbf{u}_t / \mathbf{X}_t = \mathbf{x}_t)] = 0, \quad t \in \mathbb{T}.$$

- [2] The statistical parameters of interest are  $\boldsymbol{\theta} \equiv (\mathbf{B}, \boldsymbol{\Omega})$  where  $\mathbf{B} = \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}$ ,  $\boldsymbol{\Omega} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}$ .
- [3]  $\mathbf{X}_t$  is assumed to be weakly exogenous with respect to  $\boldsymbol{\theta}$ .
- [4] No a priori information on  $\boldsymbol{\theta}$ .
- [5]  $\text{Rank}(\mathbf{X}) = k$ ,  $\mathbf{X} \equiv (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T)': T \times k$ , for  $T > k$ .

(II) **Probability model**

$$\Phi = \left\{ D(\mathbf{y}_t / \mathbf{X}_t; \boldsymbol{\theta}) = \frac{(\det \boldsymbol{\Omega})^{-\frac{1}{2}}}{(2\pi)^{m/2}} \exp\left\{-\frac{1}{2}(\mathbf{y}_t - \mathbf{B}'\mathbf{x}_t)' \boldsymbol{\Omega}^{-1} (\mathbf{y}_t - \mathbf{B}'\mathbf{x}_t)\right\}, \right. \\ \left. \boldsymbol{\theta} \in \mathbb{R}^{mk} \times \mathbf{C}^m \dagger, t \in \mathbb{T} \right\},$$

- [6] (i)  $D(\mathbf{y}_t / \mathbf{X}_t; \boldsymbol{\theta})$  – normal;  
(ii)  $E(\mathbf{y}_t / \mathbf{X}_t = \mathbf{x}_t) = \mathbf{B}'\mathbf{x}_t$  – linear in  $\mathbf{x}_t$ ;  
(iii)  $\text{Cov}(\mathbf{y}_t / \mathbf{X}_t = \mathbf{x}_t) = \boldsymbol{\Omega}$  – homoskedastic (free of  $\mathbf{x}_t$ );
- [7]  $\boldsymbol{\theta}$  is time invariant.

(III) **Sampling model**

- [8]  $\mathbf{Y} \equiv (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_T)'$  is an independent sample sequentially drawn from  $D(\mathbf{y}_t / \mathbf{X}_t; \boldsymbol{\theta})$ ,  $t = 1, 2, \dots, T$ , and  $T \geq m+k$ .

The above specification is almost identical with that of  $m=1$  considered in Chapter 19. The discussion of the assumptions in the same chapter applies to [1]–[8] above with only minor modifications due to  $m > 1$ . The only real change brought about by  $m > 1$  is the increase in the number of statistical parameters of interest being  $mk + \frac{1}{2}m(m+1)$ . It should come as no surprise to learn that the similarities between the two statistical models extend to estimation, testing and prediction.

From assumptions [6] to [8] we can deduce that the likelihood function takes the form

$$L(\boldsymbol{\theta}; \mathbf{Y}) = c(\mathbf{Y}) \prod_{t=1}^T D(\mathbf{y}_t / \mathbf{X}_t; \boldsymbol{\theta})$$

and the log likelihood is

$$\log L = \text{const} - \frac{T}{2} \log(\det \boldsymbol{\Omega}) - \frac{1}{2} \sum_t (\mathbf{y}_t - \mathbf{B}'\mathbf{x}_t)' \boldsymbol{\Omega}^{-1} (\mathbf{y}_t - \mathbf{B}'\mathbf{x}_t) \quad (24.12)$$

$$= \text{const} - \frac{1}{2}[T \log(\det \boldsymbol{\Omega}) + \text{tr } \boldsymbol{\Omega}^{-1} (\mathbf{Y} - \mathbf{XB})(\mathbf{Y} - \mathbf{XB})'] \quad (24.13)$$

(see exercise 1). The first-order conditions are

$$\frac{\partial \log L}{\partial \mathbf{B}} = (\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{XB}) \boldsymbol{\Omega}^{-1} = \mathbf{0}, \quad (24.14)$$

$$\frac{\partial \log L}{\partial \boldsymbol{\Omega}^{-1}} = \frac{T}{2} \boldsymbol{\Omega} - \frac{1}{2}(\mathbf{Y} - \mathbf{XB})'(\mathbf{Y} - \mathbf{XB}) = \mathbf{0}. \quad (24.15)$$

<sup>†</sup>  $\mathbf{C}^m$  denotes the space of all real positive definite symmetric matrices of rank  $m$ .

These first-order conditions lead to the following MLE's:

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \quad (24.16)$$

$$\hat{\Omega} = \frac{1}{T} \hat{\mathbf{U}}'\hat{\mathbf{U}}, \quad (24.17)$$

where  $\hat{\mathbf{U}} = \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}$ . For  $\hat{\Omega}$  to be positive definite we need to assume that  $T \geq m+k$  (see Dykstra (1970)). It is interesting to note that (16) amounts to estimating each regression equation separately by

$$\hat{\beta}_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_i, \quad i = 1, 2, \dots, m. \quad (24.18)$$

Moreover, the residuals from these separate regressions  $\hat{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{X}\hat{\beta}_i$  can be used to derive  $\hat{\Omega}$  via  $\hat{\omega}_{ij} = (1/T)\hat{\mathbf{u}}_i'\hat{\mathbf{u}}_j$ ,  $i, j = 1, 2, \dots, m$ .

As in the case of  $\hat{\beta}$  in the linear regression model, the MLE  $\hat{\mathbf{B}}$  preserves the original orthogonality between the systematic and non-systematic components. That is, for  $\hat{\mu}_t = \hat{\mathbf{B}}'\mathbf{x}_t$  and  $\hat{\mathbf{u}}_t = \mathbf{y}_t - \hat{\mathbf{B}}'\mathbf{x}_t$ ,

$$\mathbf{y}_t = \hat{\mu}_t + \hat{\mathbf{u}}_t, \quad t = 1, 2, \dots, T \quad (24.19)$$

and  $\hat{\mu}_t \perp \hat{\mathbf{u}}_t$ . This orthogonality can be used to define a goodness-of-fit measure by extending  $\tilde{R}^2 = 1 - (\hat{\mathbf{u}}'\hat{\mathbf{u}})/(\mathbf{y}'\mathbf{y})$  to

$$\mathbf{G} = \mathbf{I} - (\hat{\mathbf{U}}'\hat{\mathbf{U}})(\mathbf{Y}'\mathbf{Y})^{-1} = (\mathbf{Y}'\mathbf{Y} - \hat{\mathbf{U}}'\hat{\mathbf{U}})(\mathbf{Y}'\mathbf{Y})^{-1}. \quad (24.20)$$

The matrix  $\mathbf{G}$  varies between the identity matrix when  $\hat{\mathbf{U}} = \mathbf{0}$  and zero when  $\mathbf{Y} = \hat{\mathbf{U}}$  (no explanation). In order to reduce this matrix goodness-of-fit measure to a scalar we can use the trace or the determinant

$$d_1 = \frac{1}{m} \text{tr } \mathbf{G}, \quad d_2 = \det(\mathbf{G}) \quad (24.21)$$

(see Hooper (1959)).

In terms of the eigenvalues  $(\lambda_1, \lambda_2, \dots, \lambda_m)$  of  $\mathbf{G}$  the above measures of goodness of fit take the form

$$d_1 = \frac{1}{m} \sum_{i=1}^m \lambda_i \quad \text{and} \quad d_2 = \prod_{i=1}^m \lambda_i. \quad (24.22)$$

The orthogonality extends directly to  $\hat{\mathbf{M}} = \mathbf{X}\hat{\mathbf{B}}$  and  $\hat{\mathbf{U}}$  and can be used to show that  $\hat{\mathbf{B}}$  and  $\hat{\Omega}$  are *independent* random matrices. In the present context this amounts to

$$\text{Cov}(\hat{\mathbf{B}} \otimes \hat{\Omega}) = \mathbf{0}, \quad (24.23)$$

where  $E(\cdot)$  is relative to  $D(\mathbf{Y}/\mathbf{X}; \boldsymbol{\theta})$ .

### Finite sample properties of $\hat{\mathbf{B}}$ and $\hat{\Omega}$

From the fact that  $\hat{\mathbf{B}}$  and  $\hat{\Omega}$  are MLE's we can deduce that they enjoy the *invariance property* of such estimators (see Chapter 13) and they are functions of the *minimal sufficient statistics*, if they exist. Using the Lehman–Scheffe result (see Chapter 12) we can see that the ratio

$$\frac{D(\mathbf{Y}/\mathbf{X}; \boldsymbol{\theta})}{D(\mathbf{Y}_0/\mathbf{X}; \boldsymbol{\theta})} = \exp\left\{-\frac{1}{2} \text{tr } \boldsymbol{\Omega}^{-1} [\mathbf{Y}'\mathbf{Y} - \mathbf{Y}'_0\mathbf{Y}_0 - (\mathbf{Y} - \mathbf{Y}_0)' \mathbf{X} \mathbf{B} - \mathbf{B}' \mathbf{X}' (\mathbf{Y} - \mathbf{Y}_0)]\right\} \quad (24.24)$$

is independent of  $\boldsymbol{\theta}$  if  $\mathbf{Y}'\mathbf{Y} = \mathbf{Y}'\mathbf{Y}_0$  and  $\mathbf{Y}'\mathbf{X} = \mathbf{Y}'_0\mathbf{X}$ . This implies that

$$\tau(\mathbf{Y}) = (\tau_1(\mathbf{Y}), \tau_2(\mathbf{Y})), \quad \text{where } \tau_1(\mathbf{Y}) = \mathbf{Y}'\mathbf{Y}, \tau_2(\mathbf{Y}) = \mathbf{Y}'\mathbf{X}$$

defines the set of minimal sufficient statistics and

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1} \tau_2(\mathbf{Y}'), \quad (24.25)$$

$$\hat{\Omega} = \frac{1}{T} (\tau_1(\mathbf{Y}) - \tau_2(\mathbf{Y})(\mathbf{X}'\mathbf{X})^{-1} \tau_2(\mathbf{Y}')). \quad (24.26)$$

In order to discuss the other properties of  $\hat{\mathbf{B}}$  and  $\hat{\Omega}$  let us derive their distributions.

Since

$$\begin{aligned} \hat{\mathbf{B}} &= \mathbf{B} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{U} \\ &= \mathbf{B} + \mathbf{L} \mathbf{U}, \quad \mathbf{L} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}', \end{aligned}$$

we can deduce that

$$\hat{\mathbf{B}} \sim N(\mathbf{B}, \boldsymbol{\Omega} \otimes (\mathbf{X}'\mathbf{X})^{-1}). \quad (24.27)$$

This is because  $\hat{\mathbf{B}}$  is a linear function of  $\mathbf{Y}$  where

$$(\mathbf{Y}/\mathbf{X}) \sim N(\mathbf{XB}, \boldsymbol{\Omega} \otimes \mathbf{I}). \quad (24.28)$$

Given that  $T\hat{\Omega} = \mathbf{Y}(\mathbf{I} - \mathbf{M}_x)\mathbf{Y}'$ , its distribution is the matrix equivalent to the chi-square, known as the Wishart distribution with  $T-k$  degrees of freedom and scale matrix  $\boldsymbol{\Omega}$  and written as

$$T\hat{\Omega} \sim W_m(\boldsymbol{\Omega}, T-k) \quad (24.29)$$

(see Appendix 1). In the case where  $m=1$ ,  $T\hat{\Omega} = \hat{\mathbf{u}}'\hat{\mathbf{u}}$  and

$$T\hat{\Omega} \sim \sigma^2 \chi^2(T-k), \quad E(T\hat{\Omega}) = \sigma^2(T-k). \quad (24.30)$$

The Wishart distribution enjoys most of the attractive properties of the multivariate normal distribution (see Appendix 1). In direct analogy to (30),

$$E[(T\hat{\Omega}) = (T-k)]\boldsymbol{\Omega}, \quad (24.31)$$

and thus  $\tilde{\Omega} = [1/(T-k)]\hat{U}'\hat{U}$  is an unbiased estimator of  $\Omega$ . In view of (25)–(31) we can summarise the finite sample properties of the MLE's  $\hat{B}$  and  $\hat{\Omega}$  of  $B$  and  $\Omega$  respectively:

- (1)  $\hat{B}$  and  $\hat{\Omega}$  are *invariant* (with respect to Borel functions of the form  $g(\cdot): \Theta \rightarrow \mathbb{R}^r$  ( $1 \leq r \leq (mk + \frac{1}{2}m(m+1))$ )).
- (2)  $\hat{B}$  and  $\hat{\Omega}$  are *functions of the minimal sufficient statistics*  $\tau_1(Y) = Y'Y$  and  $\tau_2(Y) = Y'X$ .
- (3)  $\hat{B}$  is an *unbiased estimator* of  $B$  (i.e.  $E^\dagger(\hat{B}) = B$ ) but  $\hat{\Omega}$  is a biased estimator of  $\Omega$ ;  $\tilde{\Omega} = [1/(T-k)]\hat{U}'\hat{U}$  being unbiased.
- (4)  $\hat{B}$  is a *fully efficient estimator* of  $B$  in view of the fact that  $\text{Cov}(\hat{B}) = \Omega \otimes (X'X)^{-1}$  and the information matrix of  $\theta \equiv (B, \Omega)$  takes the form

$$I_T(\theta) = \begin{pmatrix} \Omega^{-1} \otimes X'X & 0 \\ 0 & \frac{T}{2} (\Omega^{-1} \otimes \Omega^{-1}) \end{pmatrix} \quad (24.32)$$

(see Rothenberg (1973)).

- (5)  $\hat{B}$  and  $\hat{\Omega}$  are independent; in view of the orthogonality in (19).

### Asymptotic properties of $\hat{B}$ and $\hat{\Omega}$

Arguing again by analogy to the  $m=1$  case we can derive the asymptotic properties of the MLE's  $\hat{B}$  and  $\hat{\Omega}$  of  $B$  and  $\Omega$ , respectively.

- (i) *Consistency:*  $(\hat{B} \xrightarrow{P} B, \hat{\Omega} \xrightarrow{P} \Omega)$

In view of the result  $(\hat{B} - B) \sim N(\mathbf{0}, \Omega \otimes (X'X)_T^{-1})$  we can deduce that if  $\lim_{T \rightarrow \infty} (X'X)_T^{-1} = \mathbf{0}$  then  $\text{Cov}(\hat{B}) \rightarrow \mathbf{0}$  and thus  $\hat{B}$  is a consistent estimator of  $B$  (see Chapters 12 and 19). Similarly, given that  $\lim_{T \rightarrow \infty} E(\hat{\Omega}) = \Omega$  and  $\lim_{T \rightarrow \infty} \text{Cov}(\hat{\Omega}) = \mathbf{0}$ ,  $\hat{\Omega} \xrightarrow{P} \Omega$ .

*Note* that the following statements are equivalent:

- (a)  $\lim_{T \rightarrow \infty} (X'X)_T^{-1} = \mathbf{0};$
- (b)  $\lambda_{\min}(X'X)_T \rightarrow \infty \quad \text{as } T \rightarrow \infty;$
- (c)  $\lambda_{\max}(X'X)_T^{-1} \rightarrow 0 \quad \text{as } T \rightarrow \infty;$
- (d)  $\text{tr}(X'X)_T^{-1} \rightarrow 0;$

† Note that the expectation operator  $E(\cdot)$  is relative to the underlying probability model  $D(y_t/X_t; \theta)$ .

where  $\lambda_{\min}(\mathbf{X}'\mathbf{X})_T$  and  $\lambda_{\max}(\mathbf{X}'\mathbf{X})_T^{-1}$  refer to the smallest and largest eigenvalue of  $(\mathbf{X}'\mathbf{X})_T$  and its inverse respectively; see Amemiya (1985).

(ii) *Strong consistency:*  $(\hat{\mathbf{B}} \xrightarrow{\text{a.s.}} B)$

If

$$\lim_{T \rightarrow \infty} (\mathbf{X}'\mathbf{X})_T^{-1} = \mathbf{0} \quad \text{and} \quad \left| \frac{\lambda_{\max}(\mathbf{X}'\mathbf{X})_T}{\lambda_{\min}(\mathbf{X}'\mathbf{X})_T} \right| < C$$

for some arbitrary constant  $C$ , then  $\hat{\mathbf{B}} \xrightarrow{\text{a.s.}} \mathbf{B}$ ; see Anderson and Taylor (1979).

(iii) *Asymptotic normality*

From the theory of maximum likelihood estimation we know that under relatively mild conditions (see Chapter 13) the MLE  $\hat{\boldsymbol{\theta}}$  of  $\boldsymbol{\theta}$   $\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\text{a.s.}} N(\mathbf{0}, \mathbf{I}_\infty(\boldsymbol{\theta})^{-1})$ . For this result to apply, however, we need the boundedness of  $\mathbf{I}_\infty(\boldsymbol{\theta}) = \lim_{T \rightarrow \infty} (1/T)\mathbf{I}_T(\boldsymbol{\theta})$  as well as its non-singularity. In the present case the asymptotic information matrix is bounded and non-singular (full rank) if  $\lim_{T \rightarrow \infty} (\mathbf{X}'\mathbf{X})/T = \mathbf{Q}_x < \infty$  and non-singular. Under this condition we can deduce that

$$\sqrt{T}(\hat{\mathbf{B}} - \mathbf{B}) \xrightarrow{\text{a.s.}} N(\mathbf{0}, \boldsymbol{\Omega} \otimes \mathbf{Q}_x^{-1}) \quad (24.33)$$

and

$$\sqrt{T}(\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}) \xrightarrow{\text{a.s.}} N(\mathbf{0}, 2(\boldsymbol{\Omega} \otimes \boldsymbol{\Omega})) \quad (24.34)$$

(see Rothenberg (1973)).

Note that if  $\{(\mathbf{X}'\mathbf{X})_T, T > k\}$  is a sequence of  $k \times k$  positive definite matrices such that  $(\mathbf{X}'\mathbf{X})_{T-1} - (\mathbf{X}'\mathbf{X})_T$  is positive semi-definite and  $\mathbf{c}'(\mathbf{X}'\mathbf{X})_T \mathbf{c} \rightarrow \infty$  as  $T \rightarrow \infty$  for every  $\mathbf{c} \neq \mathbf{0}$  then  $\lim_{T \rightarrow \infty} (\mathbf{X}'\mathbf{X})_T^{-1} = \mathbf{0}$ .

(iv) In view of (iii) we can deduce that  $\hat{\mathbf{B}}$  and  $\hat{\boldsymbol{\Omega}}$  are both *asymptotically unbiased* and *efficient*.

### 24.3 A priori information

One particularly important departure from the assumptions underlying the multivariate linear regression model is the introduction of a priori restrictions related to  $\boldsymbol{\theta}$ . When such additional information is available assumption [4] no longer applies and the results on estimation derived in Section 24.2 need to be modified. The importance of a priori information in

the present context arises partly because it allows us to derive tests which can be usefully employed in misspecification testing and partly because this will provide the link between the multivariate linear regression model and the simultaneous equations model to be considered in Chapter 25.

### (1) *Linear restrictions 'related' to $X_t$*

The first form of restrictions to be considered is

$$\mathbf{D}_1 \mathbf{B} + \mathbf{C}_1 = \mathbf{0}, \quad (24.35)$$

where  $\mathbf{D}_1 : p \times k$  ( $p < k$ ),  $\text{rank}(\mathbf{D}_1) = p$  and  $\mathbf{C}_1 : p \times m$  are matrices of known constants. A particularly important special case of (35) is when

$$\mathbf{D}_1 \equiv (\mathbf{0}, \mathbf{I}_{k_2}), \quad \mathbf{B} = \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix}, \quad \mathbf{C}_1 = \mathbf{0} \quad (24.36)$$

and (35) takes the form  $\mathbf{B}_2 = \mathbf{0}$ . That is, a subset of the coefficients in  $\mathbf{B}$  is zero. The thing to note about these restrictions is that they are not the same as the form

$$\mathbf{R}\beta = \mathbf{r}, \quad (24.37)$$

discussed in the context of the  $m = 1$  case (see Chapter 20). This is because the  $\mathbf{D}_1$  matrix affects all the columns of  $\mathbf{B}$  and thus the same restrictions, apart from the constants in  $\mathbf{C}_1$ , are imposed on all  $m$  linear regression equations. The form of restrictions comparable to (37) in the present context is

$$\mathbf{R}\beta_* = \mathbf{r}, \quad (24.38)$$

where  $\beta_* = \text{vec}(\mathbf{B}) = (\beta'_1, \beta'_2, \dots, \beta'_m)': mk \times 1$ ,  $\mathbf{R}: p \times mk$ ,  $\mathbf{r}: p \times 1$ . This form of linear restrictions is more general than (35) as well as

$$\mathbf{B}\Gamma_1 + \Delta_1 = \mathbf{0}. \quad (24.39)$$

All three forms, (35), (38) and (39), will be discussed in this section because they are interesting for different reasons.

When the restrictions (35) are interpreted in the context of the statistical GM

$$\mathbf{y}_t = \mathbf{B}'\mathbf{x}_t + \mathbf{u}_t, \quad t \in \mathbb{T}, \quad (24.40)$$

we can see that they are directly related to the regressors  $X_{it}$ ,  $i = 1, 2, \dots, k$ . The easiest way to take (35) into consideration in the estimation of  $\theta \equiv (\mathbf{B}, \Omega)$  is to 'solve' the system (35) for  $\mathbf{B}$  and substitute the 'solution' into (40). In order to do that we define two arbitrary matrices  $\mathbf{D}_1^*: (k-p) \times k$ ,  $\text{rank}(\mathbf{D}_1^*) =$

$k - p$ , and  $\mathbf{C}_1^* : (k - p) \times m$ , and reformulate (35) into

$$\mathbf{DB} + \mathbf{C} = \mathbf{0} \quad (24.41)$$

where

$$\mathbf{D} = (\mathbf{D}_1, \mathbf{D}_1^*), \quad k \times k, \quad \mathbf{C} = \begin{pmatrix} \mathbf{C}_1 \\ \mathbf{C}_1^* \end{pmatrix}, \quad k \times m.$$

The fact that  $\text{rank}(\mathbf{D}) = k$  enables us to solve (41) for  $\mathbf{B}$  to yield

$$\mathbf{B} = -\mathbf{D}^{-1}\mathbf{C} = \mathbf{G}_1\mathbf{C}_1 + \mathbf{G}_1^*\mathbf{C}_1^*, \quad (24.42)$$

where  $\mathbf{G} \equiv (\mathbf{G}_1, \mathbf{G}_1^*) = -\mathbf{D}^{-1}$ . Substituting this into (40) for  $t = 1, 2, \dots, T$  yields

$$\mathbf{Y}^* = \mathbf{X}^*\mathbf{C}_1^* + \mathbf{U}, \quad (24.43)$$

where  $\mathbf{Y}^* \equiv \mathbf{Y} - \mathbf{X}\mathbf{G}_1\mathbf{C}_1$  and  $\mathbf{X}^* = \mathbf{X}\mathbf{G}_1^*$ . The fact that the form of the underlying probability model is unchanged implies that the MLE of  $\mathbf{C}_1^*$  is

$$\begin{aligned} \mathbf{C}_1^* &= (\mathbf{X}^*\mathbf{X}^*)^{-1}\mathbf{X}^*\mathbf{Y}^* = (\mathbf{G}_1^*\mathbf{X}\mathbf{X}\mathbf{G}_1^*)^{-1}\mathbf{G}_1^*\mathbf{X}'(\mathbf{Y} - \mathbf{X}\mathbf{G}_1\mathbf{C}_1) \\ &= (\mathbf{G}_1^*\mathbf{X}\mathbf{X}\mathbf{G}_1^*)^{-1}\mathbf{G}_1^*\mathbf{X}'\mathbf{X}'(\hat{\mathbf{B}} - \mathbf{G}_1\mathbf{C}_1). \end{aligned} \quad (24.44)$$

Hence, from (42) the constrained MLE of  $\mathbf{B}$  is

$$\tilde{\mathbf{B}} = \mathbf{G}_1\mathbf{C}_1 + \mathbf{G}_1^*(\mathbf{G}_1^*\mathbf{X}\mathbf{X}\mathbf{G}_1^*)^{-1}\mathbf{G}_1^*\mathbf{X}'\mathbf{X}'(\hat{\mathbf{B}} - \mathbf{G}_1\mathbf{C}_1) \quad (24.45)$$

$$\begin{aligned} &= \mathbf{G}_1\mathbf{C}_1 + \mathbf{L}(\hat{\mathbf{B}} - \mathbf{G}_1\mathbf{C}_1), \quad \text{where } \mathbf{L} = \mathbf{G}_1^*(\mathbf{G}_1^*\mathbf{X}\mathbf{X}\mathbf{G}_1^*)^{-1}\mathbf{G}_1^*\mathbf{X}'\mathbf{X}' \\ &= \hat{\mathbf{B}} - \mathbf{P}(\hat{\mathbf{B}} - \mathbf{G}_1\mathbf{C}_1), \quad \text{where } \mathbf{P} = \mathbf{I} - \mathbf{L}. \end{aligned} \quad (24.46)$$

Given that  $\mathbf{L}^2 = \mathbf{L}$ ,  $\mathbf{P}^2 = \mathbf{P}$  and  $\mathbf{LP} = \mathbf{0}$  (i.e. they are orthogonal projections) we can deduce that  $\mathbf{P}$  takes the form

$$\mathbf{P} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{D}'_1[\mathbf{D}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{D}'_1]^{-1}\mathbf{D}_1 \quad (24.47)$$

(see exercise 2). This implies that

$$\tilde{\mathbf{B}} = \hat{\mathbf{B}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{D}'_1[\mathbf{D}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{D}'_1]^{-1}(\mathbf{D}_1\hat{\mathbf{B}} + \mathbf{C}_1), \quad (24.48)$$

since  $\mathbf{D}_1\mathbf{G}_1 = \mathbf{I}_p$ . Moreover, the constrained MLE of  $\boldsymbol{\Omega}$  is

$$\hat{\boldsymbol{\Omega}} = \frac{1}{T}\tilde{\mathbf{C}}'\tilde{\mathbf{C}} = \hat{\boldsymbol{\Omega}} + \frac{1}{T}(\tilde{\mathbf{B}} - \hat{\mathbf{B}})'(\mathbf{X}'\mathbf{X})(\tilde{\mathbf{B}} - \hat{\mathbf{B}}). \quad (24.49)$$

Looking at the constrained MLE's of  $\mathbf{B}$  and  $\boldsymbol{\Omega}$  we can see that they are direct extensions of the results in the case of  $m = 1$  in Chapter 20.

Another important special case of (35) is the case where all the coefficients apart from the constant terms, say  $\beta_{.1}$ , are zero. This can be expressed in the

form (35) with

$$\mathbf{D}_1 = (\mathbf{0}, \mathbf{I}_{k-1}), \quad \mathbf{B} = (\boldsymbol{\beta}_{\cdot 1}, \mathbf{B}_{(1)}), \quad \mathbf{C} = \mathbf{0}$$

and  $H_0$  takes the form  $\mathbf{B}_{(1)} = \mathbf{0}$ .

## (2) Linear restrictions 'related' to $y_t$

The second form of restrictions to be considered is

$$\mathbf{B}\Gamma_1 + \Delta_1 = \mathbf{0}, \quad (24.50)$$

where  $\Gamma_1: m \times q$  ( $q \leq m$ ) and  $\Delta_1: k \times q$  are known matrices with  $\text{rank}(\Gamma_1) = q$ . The restrictions in (50) represent linear *between-equations* restrictions because the  $i$ th row of  $\mathbf{B}$  represents the  $i$ th coefficient on all equations. Interpreted in the context of (35) these restrictions are directly related to the  $y_{it}$ s. This implies that if we follow the procedure used for the restrictions in (38) we have to be much more careful because the form of the underlying probability model might be affected. Richard (1979) shows how this procedure can give rise to the restricted MLE's of  $\mathbf{B}$  and  $\Omega$ . For expositional purposes we will adopt the Lagrange multiplier procedure. The Lagrangian function is

$$\begin{aligned} l(\mathbf{B}, \Omega, \mathbf{M}) = & -\frac{T}{2} \log(\det \Omega) - \frac{1}{2} \text{tr} \Omega^{-1} (\mathbf{Y} - \mathbf{XB})(\mathbf{Y} - \mathbf{XB})' \\ & - \text{tr}[\Lambda'(\mathbf{B}\Gamma_1 + \Delta_1)], \end{aligned} \quad (24.51)$$

where  $\Lambda$  is a matrix of Lagrange multipliers.

$$\frac{\partial l}{\partial \mathbf{B}} = (\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{XB})\Omega^{-1} - \Lambda\Gamma_1' = \mathbf{0}, \quad (24.52)$$

$$\frac{\partial l}{\partial \Omega^{-1}} = \frac{T}{2} \Omega - \frac{1}{2} (\mathbf{Y} - \mathbf{XB})(\mathbf{Y} - \mathbf{XB})' = \mathbf{0}, \quad (24.53)$$

$$\frac{\partial l}{\partial \Lambda} = -(\mathbf{B}\Gamma_1 + \Delta_1) = \mathbf{0} \quad (24.54)$$

(see Appendix 2). From (52) we can deduce that

$$(\mathbf{X}'\mathbf{X})(\hat{\mathbf{B}} - \mathbf{B}) = \Lambda\Gamma_1'\Omega. \quad (24.55)$$

Premultiplying by  $\Delta_1$  and solving for  $\Lambda$  yields

$$\Lambda = (\mathbf{X}'\mathbf{X})(\hat{\mathbf{B}}\Gamma_1 - \mathbf{B}\Gamma_1)(\Gamma_1'\Omega\Gamma_1)^{-1}, \quad (24.56)$$

which in view of (54) becomes

$$\Lambda = (\mathbf{X}'\mathbf{X})(\hat{\mathbf{B}}\Gamma_1 + \Delta_1)(\Gamma_1'\Omega\Gamma_1)^{-1}. \quad (24.57)$$

This implies that the constrained MLE's of  $\mathbf{B}$  and  $\boldsymbol{\Omega}$  are

$$\tilde{\mathbf{B}} = \hat{\mathbf{B}} - (\hat{\mathbf{B}}\Gamma_1 + \Delta_1)(\Gamma'_1\hat{\boldsymbol{\Omega}}\Gamma_1)^{-1}\Gamma'_1\hat{\boldsymbol{\Omega}}, \quad (24.58)$$

$$\tilde{\boldsymbol{\Omega}} = \frac{1}{T} \tilde{\mathbf{U}}'\tilde{\mathbf{U}} = \hat{\boldsymbol{\Omega}} + \frac{1}{T}(\tilde{\mathbf{B}} - \hat{\mathbf{B}})'(\mathbf{X}'\mathbf{X})(\tilde{\mathbf{B}} - \hat{\mathbf{B}}) \quad (24.59)$$

(see Richard (1979)). If we compare (58) with (48) we can see that the main difference is that  $\hat{\boldsymbol{\Omega}}$  enters the MLE estimator of  $\mathbf{B}$  in view of the fact that the restrictions (50) affect the form of the probability model. It is interesting to note that if we premultiply (58) by  $\Gamma_1$  it yields (54). The above formulae, (58), (59), will be of considerable value in Chapter 25.

### (3) Linear restrictions 'related' to both $\mathbf{y}_t$ and $\mathbf{X}_t$

A natural way to proceed is to combine the linear restrictions (38) and (50) in the form of

$$\mathbf{D}_1\mathbf{B}\Gamma + \mathbf{C} = \mathbf{0}, \quad (24.60)$$

where  $\mathbf{D}_1: p \times k$ ,  $\Gamma_1: m \times q$ ,  $\mathbf{C}: p \times q$ , are known matrices with  $\text{rank}(\mathbf{D}_1) = p$ ,  $\text{rank}(\Gamma_1) = q$ . Using the Lagrangian function

$$\begin{aligned} l(\mathbf{B}, \boldsymbol{\Omega}, \Lambda) = & -\frac{T}{2} \log(\det \boldsymbol{\Omega}) - \frac{1}{2} \text{tr } \boldsymbol{\Omega}^{-1}(\mathbf{Y} - \mathbf{XB})'(\mathbf{Y} - \mathbf{XB}) \\ & - \text{tr}[\Lambda'(\mathbf{D}_1\mathbf{B}\Gamma_1 + \mathbf{C})], \end{aligned} \quad (24.61)$$

we can show that the restricted MLE's are

$$\tilde{\mathbf{B}} = \hat{\mathbf{B}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{D}'_1[\mathbf{D}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{D}'_1]^{-1}(\mathbf{D}_1\hat{\mathbf{B}}\Gamma_1 + \mathbf{C})(\Gamma'_1\hat{\boldsymbol{\Omega}}\Gamma_1)^{-1}\Gamma'_1\hat{\boldsymbol{\Omega}}, \quad (24.62)$$

$$\tilde{\boldsymbol{\Omega}} = \hat{\boldsymbol{\Omega}} + \frac{1}{T}(\tilde{\mathbf{B}} - \hat{\mathbf{B}})'(\mathbf{X}'\mathbf{X})(\tilde{\mathbf{B}} - \hat{\mathbf{B}}). \quad (24.63)$$

An alternative way to derive (62) and (63) is to consider

$$\mathbf{D}_1\mathbf{B}^* + \mathbf{C} = \mathbf{0} \quad (24.64)$$

for the transformed specification

$$\mathbf{Y}^* = \mathbf{XB}^* + \mathbf{E}. \quad (24.65)$$

where  $\mathbf{Y}^* = \mathbf{Y}\Gamma_1$ ,  $\mathbf{B}^* = \mathbf{B}\Gamma_1$  and  $\mathbf{E} = \mathbf{U}\Gamma_1$ .

The linear restrictions in (60) in vector form can be written as

$$\text{vec}(\mathbf{D}_1\mathbf{B}\Gamma_1 + \mathbf{C}) = (\Gamma'_1 \otimes \mathbf{D}_1) \text{vec}(\mathbf{B}) + \text{vec}(\mathbf{C}) = \mathbf{0} \quad (24.66)$$

or

$$(\Gamma'_1 \otimes \mathbf{D}_1)\beta_* = \mathbf{r}, \quad (24.67)$$

where  $\beta_* = \text{vec } \mathbf{B}$  and  $\mathbf{r} = -\text{vec}(\mathbf{C})$ . This suggests that an obvious way to generalise this is to substitute  $(\Gamma_1 \otimes \mathbf{D}_1)$  with a  $p \times km$  matrix  $\mathbf{R}$  to formulate the restrictions in the form

$$\mathbf{R}\beta_* = \mathbf{r}, \quad (24.68)$$

where  $\text{rank}(\mathbf{R}) = p$  ( $p < km$ ). The restrictions in (68) represent the most general form of linear restrictions in view of the fact that  $\beta_*$  enables us to 'reach' each coefficient of  $\mathbf{B}$  directly and impose within-equation and between-equations restrictions separately.

In the case where only within-equation linear restrictions are available  $\mathbf{R}$  is block-diagonal, i.e.

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_2 & & \\ \vdots & & \ddots & \\ \mathbf{0} & & \mathbf{0} & \mathbf{R}_m \end{pmatrix} \text{ and } \mathbf{r} = \begin{pmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \vdots \\ \mathbf{r}_m \end{pmatrix}, \quad (24.69)$$

$$\mathbf{R}_i: p_i \times k, \quad i = 1, 2, \dots, m, \quad \text{rank}(\mathbf{R}_i) = p_i, \quad \mathbf{r}_i: p_i \times 1.$$

Exclusion restrictions are a special case of within-equation restrictions where  $\mathbf{R}_i$  has a unit sub-matrix, of dimension equal to the number of excluded variables, and zeros everywhere else.

Across-equations linear restrictions can be accommodated in the off block-diagonal submatrices  $\mathbf{R}_{ij}$ ,  $i, j = 1, 2, \dots, m, i \neq j$  of  $\mathbf{R}$  with  $\mathbf{R}_{ij}$  referring to the restrictions between equations  $i$  and  $j$ .

Let us consider the derivation of the constrained MLE's of  $\mathbf{B}$  and  $\Omega$  under the linear restrictions (68). The most convenient form of the statistical GM for the sample period  $t = 1, 2, \dots, T$  is not

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{U}, \quad (24.70)$$

as in the case of (35) and (39), but its vectorised formulation

$$\mathbf{y}_* = \mathbf{X}_*\beta_* + \mathbf{u}_*, \quad (24.71)$$

where

$$\mathbf{y}_* = (\mathbf{y}'_1, \mathbf{y}'_2, \dots, \mathbf{y}'_m)': Tm \times 1, \quad \mathbf{X}_* = (\mathbf{I}_m \otimes \mathbf{X}): Tm \times mk,$$

$$\beta_* = (\beta'_1, \beta'_2, \dots, \beta'_m)': mk \times 1, \quad \mathbf{u}_* = (\mathbf{u}'_1, \mathbf{u}'_2, \dots, \mathbf{u}'_m)': Tm \times 1$$

and

$$\Omega_* = (\Omega \otimes \mathbf{I}_T): Tm \times Tm.$$

The Lagrangian function is defined using the vector  $\lambda: p \times 1$  of multipliers:

$$\begin{aligned} l(\beta_*, \Omega_*, \lambda) = & -\frac{T}{2} \log(\det \Omega_*) - \frac{1}{2} (\mathbf{y}_* - \mathbf{X}_*\beta_*)' \Omega_*^{-1} (\mathbf{y}_* - \mathbf{X}_*\beta_*) \\ & - \lambda' (\mathbf{R}\beta_* - \mathbf{r}). \end{aligned} \quad (24.72)$$

$$\frac{\partial l}{\partial \beta_*} = \mathbf{X}'_* \Omega_*^{-1} (\mathbf{y}_* - \mathbf{X}_* \beta_*) - \mathbf{R}' \lambda = \mathbf{0}, \quad (24.73)$$

$$\frac{\partial l}{\partial \Omega_*} = -\frac{T}{2} \Omega_*^{-1} + \frac{1}{2} \Omega_*^{-1} (\mathbf{y}_* - \mathbf{X}_* \beta_*) (\mathbf{y}_* - \mathbf{X}_* \beta_*)' \Omega_*^{-1} \quad (24.74)$$

$$\frac{\partial l}{\partial \lambda} = (\mathbf{R} \beta_* - \mathbf{r}) = \mathbf{0}. \quad (24.75)$$

Looking at the above first-order conditions (73)–(75) we can see that they constitute a system of non-linear equations which cannot be solved explicitly unless  $\Omega$  is assumed to be known. In the latter case (73) and (75) imply that

$$\tilde{\beta}_* = \bar{\beta}_* - (\mathbf{X}'_* \Omega_*^{-1} \mathbf{X}_*)^{-1} \mathbf{R}' [\mathbf{R} (\mathbf{X}'_* \Omega_*^{-1} \mathbf{X}_*)^{-1} \mathbf{R}^*]^{-1} (\mathbf{R} \bar{\beta}_* - \mathbf{r}), \quad (24.76)$$

$$\tilde{\lambda} = [\mathbf{R} (\mathbf{X}'_* \Omega_*^{-1} \mathbf{X}_*)^{-1} \mathbf{R}']^{-1} (\mathbf{R} \bar{\beta}_* - \mathbf{r}). \quad (24.77)$$

and

$$\bar{\beta}_* = (\mathbf{X}'_* \Omega_*^{-1} \mathbf{X}_*)^{-1} \mathbf{X}'_* \Omega_*^{-1} \mathbf{y}_*. \quad (24.78)$$

If we compare these formulae with those in the  $m=1$  case (see Chapter 20) we can see that the only difference (when  $\Omega$  is known) is the presence of  $\Omega_*$ . This is because in the  $m>1$  case the restrictions  $\mathbf{R}\beta_*=\mathbf{r}$  affect the underlying probability model by restricting  $\mathbf{y}_t$ . In the econometric literature the estimator (78) is known as the *generalised least-squares* (GLS) estimator.

In practice  $\Omega$  is unknown and thus in order to ‘solve’ the conditions (73)–(75) we need to resort to iterative numerical optimisation (see Harvey (1981), Quandt (1983), *inter alia*).

The purpose of the next section is to consider two special cases of (68) where the restrictions can be substituted directly into a reformulated statistical GM. These are the cases of exclusion and across-equations linear homogeneous restrictions. In these two cases the constrained MLE of  $\beta_*$  takes a form similar to (78).

## 24.4 The Zellner and Malinvaud formulations

In econometric modelling two special cases of the general linear restrictions

$$\mathbf{R}\beta_* = \mathbf{r} \quad (24.79)$$

are particularly useful. These are the exclusion and across-equations linear homogeneous restrictions. In order to illustrate these let us consider the

two-equation case

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} \beta_{11} & \beta_{21} & \beta_{31} \\ \beta_{12} & \beta_{22} & \beta_{23} \end{pmatrix} \begin{pmatrix} x_{1t} \\ x_{2t} \\ x_{3t} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix}, \quad t \in \mathbb{T}. \quad (24.80)$$

- (i) *Exclusion restrictions:*  $\beta_{11} = 0, \beta_{23} = 0$ ;
- (ii) *Across-equation linear homogeneous restrictions:*  $\beta_{21} = \beta_{12}$ .

It turns out that in these two cases the restrictions can be accommodated directly into a reformulation of the statistical GM and no constrained optimisation is necessary. The purpose of this section is to discuss the estimation of  $\beta_*$  under these two forms of restrictions and derive explicit formulae which will prove useful in Chapter 25.

Let us consider the exclusion restrictions first. The vectorised form of

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U}, \quad (24.81)$$

as defined in the previous sections, takes the explicit form

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} \mathbf{X} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X} & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{X} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix} + \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_m \end{pmatrix} \quad (24.82)$$

or

$$\mathbf{y}_* = \mathbf{X}_* \boldsymbol{\beta}_* + \mathbf{u}_*, \quad (24.83)$$

in an obvious notation. Exclusion restrictions can be accommodated directly into (82) by allowing the regressor matrix  $\mathbf{X}$  to be different for different regression equations  $y_i = \mathbf{X}\beta_i + \mathbf{u}_i, i = 1, 2, \dots, m$ , and redefining the  $\beta_i$ s accordingly. That is, reformulate (82) into

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{X}_m \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1^* \\ \boldsymbol{\beta}_2^* \\ \vdots \\ \boldsymbol{\beta}_m^* \end{pmatrix} + \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_m \end{pmatrix} \quad (24.84)$$

or

$$\mathbf{y}_* = \mathbf{X}_* \boldsymbol{\beta}_*^* + \mathbf{u}_*, \quad (24.85)$$

where  $\mathbf{X}_i$  refers to the regressor data matrix for the  $i$ th equation and  $\boldsymbol{\beta}_i^*$  the corresponding coefficients vector. In the case of the example in (80) with the

restrictions  $\beta_{11}=0$ ,  $\beta_{23}=0$ , (84) takes the form

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix}, \quad (24.86)$$

where  $\mathbf{X}_1 \equiv (\mathbf{x}_2, \mathbf{x}_3)$ ,  $\mathbf{X}_2 \equiv (\mathbf{x}_1, \mathbf{x}_2)$ ,  $\boldsymbol{\beta}_1 = (\beta_{21} \beta_{31})'$  and  $\boldsymbol{\beta}_2 = (\beta_{12} \beta_{22})'$ .

The formulation (84) is known as the *seemingly unrelated regression equations* (SURE), a term coined by Zellner (1962), because the  $m$  linear regression equations in (84) seem to be unrelated at first sight but this turns out to be false. When different restrictions are placed on different equations the original statistical GM is affected and the various equations become interrelated. In particular the covariance matrix  $\boldsymbol{\Omega}$  enters the estimator of  $\boldsymbol{\beta}_*^*$ . As shown in the previous section, in the case where  $\boldsymbol{\Omega}$  is known the MLE of  $\boldsymbol{\beta}_*^*$  takes the form

$$\bar{\boldsymbol{\beta}}_*^* = (\mathbf{X}_*^* (\boldsymbol{\Omega}^{-1} \otimes \mathbf{I}_T) \mathbf{X}_*^*)^{-1} \mathbf{X}_*^* (\boldsymbol{\Omega}^{-1} \otimes \mathbf{I}_T) \mathbf{y}_*. \quad (24.87)$$

Otherwise, the MLE is derived using some iterative numerical procedure. For this case Zellner (1962) suggested the two-step least-squares estimator

$$\hat{\boldsymbol{\beta}}_*^* = (\mathbf{X}_*^* (\hat{\boldsymbol{\Omega}}^{-1} \otimes \mathbf{I}_T) \mathbf{X}_*^*)^{-1} \mathbf{X}_*^* (\hat{\boldsymbol{\Omega}} \otimes \mathbf{I}_T) \mathbf{y}_*, \quad (24.88)$$

where  $\hat{\boldsymbol{\Omega}} = (1/T) \hat{\mathbf{U}}' \hat{\mathbf{U}}$ ,  $\hat{\mathbf{U}} = \mathbf{Y} - \hat{\mathbf{X}}\hat{\boldsymbol{\beta}}$ . It is not very difficult to see that this estimator can be viewed as an approximation to the MLE defined in the previous section by the first-order conditions (73)–(75) where only two iterations were performed. One to derive  $\hat{\boldsymbol{\Omega}}$  and then substitute into (87). Zellner went on to show that if

$$\lim_{T \rightarrow \infty} \left( \mathbf{X}_*^* \frac{\boldsymbol{\Omega}_*^{-1}}{T} \mathbf{X}_*^* \right) = \mathbf{Q}_* < \infty, \quad (24.89)$$

and non-singular, the asymptotic distribution of (87) and (88) coincide, taking the form

$$\sqrt{T}(\bar{\boldsymbol{\beta}}_*^* - \boldsymbol{\beta}_*^*) \underset{z}{\sim} N(\mathbf{0}, \mathbf{Q}_*^{-1}). \quad (24.90)$$

It is interesting to note that in the cases:

(a)  $\mathbf{X}_1 = \mathbf{X}_2 = \dots = \mathbf{X}_m = \mathbf{X}$ ;

and

(b)  $\boldsymbol{\Omega} = \text{diag}(\omega_{11}, \omega_{22}, \dots, \omega_{mm})$

$$\bar{\boldsymbol{\beta}}_*^* = \hat{\boldsymbol{\beta}}_* = (\mathbf{X}_*^* \mathbf{X}_*^*)^{-1} \mathbf{X}_*^* \mathbf{y}_* \quad (24.91)$$

(see Schmidt (1976)).

Another important special case of the linear restrictions in (79) is the case

of across-equation linear homogeneous restrictions such as  $\beta_{21} = \beta_{12}$  in example (80). Such restrictions can be accommodated into the formulation (82) directly by redefining the regressor matrix as

$$\mathbf{X}_t^* = \begin{pmatrix} \mathbf{x}'_{1t} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}'_{2t} & & \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{x}'_{mt} \end{pmatrix} \quad (24.92)$$

(where  $\mathbf{x}_{it}$  refers to the regressors included in the  $i$ th equation) and the coefficient vector  $\boldsymbol{\beta}_*$ , so as to include only the independent coefficients. The form

$$\mathbf{y}_t = \mathbf{X}_t^* \boldsymbol{\beta}_* + \mathbf{u}_t \quad (24.93)$$

is said to be the *Malinvaud form* (see Malinvaud (1970)). For the above example the restriction  $\beta_{12} = \beta_{21}$  can be accommodated into (80) by defining  $\mathbf{X}_t^*$  and  $\boldsymbol{\beta}_*$  as

$$\mathbf{X}_t^* = \begin{pmatrix} x_{1t} & x_{3t} & 0 \\ x_{1t} & 0 & x_{2t} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta}_* = \begin{pmatrix} \beta_{21} \\ \beta_{31} \\ \beta_{22} \end{pmatrix}. \quad (24.94)$$

The constrained MLE of  $\boldsymbol{\beta}_*$  in the case where  $\boldsymbol{\Omega}$  is known is

$$\hat{\boldsymbol{\beta}}_*^* = \left( \sum_{t=1}^T \mathbf{X}_t^{*\prime} \boldsymbol{\Omega}^{-1} \mathbf{X}_t^* \right)^{-1} \sum_{t=1}^T \mathbf{X}_t^{*\prime} \boldsymbol{\Omega}^{-1} \mathbf{y}_t. \quad (24.95)$$

Given that  $\boldsymbol{\Omega}$  is usually unknown, the MLE of  $\boldsymbol{\beta}_*$  as defined in the previous section by (73)–(75) can be approximated by the GLS estimator based on the iterative formula

$$\hat{\boldsymbol{\beta}}_{i+1}^* = \left( \sum_{t=1}^T \mathbf{X}_t^{*\prime} \hat{\boldsymbol{\Omega}}_i^{-1} \mathbf{X}_t^* \right)^{-1} \sum_{t=1}^T \mathbf{X}_t^{*\prime} \hat{\boldsymbol{\Omega}}_i^{-1} \mathbf{y}_t, \quad i = 1, 2, \dots, l, \quad (24.96)$$

where  $l$  refers to the number of iterations which is either chosen a priori or determined by some convergence criterion such as

$$|\hat{\boldsymbol{\beta}}_{i+1}^* - \hat{\boldsymbol{\beta}}_i^*| < \varepsilon \quad \text{for some } \varepsilon > 0, \quad \text{e.g. } \varepsilon = 0.001. \quad (24.97)$$

In the case where  $l = 2$  the estimator defined by (96) coincides with

$$\hat{\boldsymbol{\beta}}_*^* = \left( \sum_{t=1}^T \mathbf{X}_t^{*\prime} \hat{\boldsymbol{\Omega}}^{-1} \mathbf{X}_t^* \right)^{-1} \sum_{t=1}^T \mathbf{X}_t^{*\prime} \hat{\boldsymbol{\Omega}}^{-1} \mathbf{y}_t, \quad (24.98)$$

where  $\hat{\boldsymbol{\Omega}} = (1/T) \hat{\mathbf{U}}' \hat{\mathbf{U}}$ ,  $\hat{\mathbf{U}} = \mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\beta}}$ .

## 24.5 Specification testing

In the context of the linear regression model the  $F$ -type test proved to be by far the most useful test in both specification as well as misspecification analysis; see Chapters 19–22. The question which naturally arises is whether the  $F$ -type test can be extended to the multivariate linear regression model. The main purpose of this section is to derive an extended  $F$ -type test which serves the same purpose as the  $F$ -test in Chapters 19–22.

From Section 24.2 we know that for the MLE's  $\hat{\mathbf{B}}$  and  $\hat{\Omega}$

$$(i) \quad \hat{\mathbf{B}} \sim N(\mathbf{B}, \Omega \otimes (\mathbf{X}'\mathbf{X})^{-1}); \quad (24.99)$$

and

$$(ii) \quad \mathbf{T}\hat{\Omega} \sim W_m(\Omega, T-k). \quad (24.100)$$

Using these results we can deduce that, in the case where we consider one regression from the system, say the  $i$ th,

$$\mathbf{y}_i = \mathbf{X}\beta_i + \mathbf{u}_i, \quad (24.101)$$

the MLE's of  $\beta_i$  and  $\omega_{ii}$  are

$$\hat{\beta}_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_i \quad \text{and} \quad \hat{\omega}_{ii} = \frac{1}{T} \hat{\mathbf{u}}_i'\hat{\mathbf{u}}_i, \quad \hat{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{X}\hat{\beta}_i \quad (24.102)$$

Moreover, using the properties of the multivariate normal (see Chapter 15) and Wishart (see Appendix 1) distributions we can deduce that

$$\hat{\beta}_i \sim N(\beta_i, \omega_{ii}(\mathbf{X}'\mathbf{X})^{-1}) \quad \text{and} \quad T \left( \frac{\hat{\omega}_{ii}}{\omega_{ii}} \right) \sim \chi^2(T-k). \quad (24.103)$$

These results ensure that, in the case of linear restrictions related to  $\beta_i$  of the form  $H_0: \mathbf{R}_i\beta_i = \mathbf{r}_i$  against  $H_1: \mathbf{R}_i\beta_i \neq \mathbf{r}_i$  where  $\mathbf{R}_i$  and  $\mathbf{r}_i$  are  $p_i \times k$  and  $p_i \times 1$  known matrices and  $\text{rank}(\mathbf{R}_i) = p_i$ , the  $F$ -test based on the test statistic

$$FT(\mathbf{y}_i) = \frac{(\mathbf{R}_i\hat{\beta}_i - \mathbf{r}_i)'[\mathbf{R}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}_i]^{-1}(\mathbf{R}_i\hat{\beta}_i - \mathbf{r}_i)}{\hat{\mathbf{u}}_i'\hat{\mathbf{u}}_i} \left( \frac{T-k}{p_i} \right) \quad (24.104)$$

is applicable without any changes. In particular, tests of significance for individual coefficients based on the test statistic

$$\tau(\mathbf{y}_i) = \frac{\hat{\beta}_{ji} - 0}{\sqrt{\left( \frac{\hat{\mathbf{u}}_i'\hat{\mathbf{u}}_i}{T-k} \right)(\mathbf{X}'\mathbf{X})_{jj}^{-1}}} \quad (24.105)$$

(a special case of (104)) is applicable to the present context without any modifications; see Chapters 19 and 20.

Let us now consider the derivation of a test for the null hypothesis:

$$H_0: \mathbf{DB} - \mathbf{C} = \mathbf{0} \quad \text{against} \quad H_1: \mathbf{DB} - \mathbf{C} \neq \mathbf{0} \quad (24.106)$$

where  $\mathbf{D}$  and  $\mathbf{C}$  are  $p \times k$  and  $p \times m$  known matrices,  $\text{rank}(\mathbf{D}) = p$ . A particularly important special case of (106) is when

$$\mathbf{D} = (\mathbf{0}, \mathbf{I}_{k_2}): k_2 \times k, \quad \mathbf{B} = \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix} \quad \text{and} \quad \mathbf{C} = \mathbf{0}: k_2 \times m,$$

i.e.  $H_0: \mathbf{B}_2 = \mathbf{0}$  against  $H_1: \mathbf{B}_2 \neq \mathbf{0}$ . The constrained MLE's of  $\mathbf{B}$  and  $\boldsymbol{\Omega}$  under  $H_0$  take the form

$$\tilde{\mathbf{B}} = \hat{\mathbf{B}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{D}'[\mathbf{D}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{D}']^{-1}(\mathbf{D}\hat{\mathbf{B}} - \mathbf{C}) \quad (24.107)$$

and

$$\tilde{\boldsymbol{\Omega}} = \hat{\boldsymbol{\Omega}} + \frac{1}{T}(\tilde{\mathbf{B}} - \hat{\mathbf{B}})'(\mathbf{X}'\mathbf{X})(\tilde{\mathbf{B}} - \hat{\mathbf{B}}), \quad (24.108)$$

where  $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ ,  $\hat{\boldsymbol{\Omega}} = (1/T)\hat{\mathbf{U}}'\hat{\mathbf{U}}$ ,  $\hat{\mathbf{U}} = \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}$  are the unconstrained MLE's of  $\mathbf{B}$  and  $\boldsymbol{\Omega}$  (see Section 24.3 above). Using the same intuitive argument as in the  $m=1$  case (see Chapter 20) a test for  $H_0$  could be based on the distance

$$\|\mathbf{D}\hat{\mathbf{B}} - \mathbf{C}\|. \quad (24.109)$$

The closer this distance is to zero the more the support for  $H_0$ . If we normalise this distance by defining the matrix quadratic form

$$\boldsymbol{\Omega}^{-1}(\mathbf{D}\hat{\mathbf{B}} - \mathbf{C})'[\mathbf{D}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{D}']^{-1}(\mathbf{D}\hat{\mathbf{B}} - \mathbf{C}), \quad (24.110)$$

the similarity between (110) and the  $F$ -test statistic (104) is all too apparent. Moreover, in view of the equality

$$\hat{\mathbf{U}}'\hat{\mathbf{U}} = \hat{\mathbf{U}}'\hat{\mathbf{U}} + (\mathbf{D}\hat{\mathbf{B}} - \mathbf{C})'[\mathbf{D}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{D}']^{-1}(\mathbf{D}\hat{\mathbf{B}} - \mathbf{C}) \quad (24.111)$$

stemming from (108), (110) can be written in the form

$$\hat{\mathbf{G}} = (\hat{\mathbf{U}}'\hat{\mathbf{U}} - \hat{\mathbf{U}}'\hat{\mathbf{U}})(\hat{\mathbf{U}}'\hat{\mathbf{U}})^{-1}, \quad (24.112)$$

where  $\tilde{\mathbf{U}} = \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}$ . This form constitutes a direct extension of the  $F$ -test statistic to the general  $m > 1$  case. Continuing the analogy, we can show that

$$\hat{\mathbf{U}}'\hat{\mathbf{U}} \sim W_m(\boldsymbol{\Omega}, T-k), \quad T \geq m+k, \quad (24.113)$$

where  $\hat{\mathbf{U}}'\hat{\mathbf{U}} = \mathbf{U}'\mathbf{M}_X\mathbf{U}$ ,  $\mathbf{M}_X = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ . Moreover, in view of (112) the distribution of  $\tilde{\mathbf{U}}'\tilde{\mathbf{U}} - \hat{\mathbf{U}}'\hat{\mathbf{U}}$  is a direct extension of the non-central chi-square distribution, the non-central Wishart, denoted by

$$(\tilde{\mathbf{U}}'\tilde{\mathbf{U}} - \hat{\mathbf{U}}'\hat{\mathbf{U}}) \sim W_m(\boldsymbol{\Omega}, p; \Delta), \quad T \geq m+k, \quad (24.114)$$

where

$$\Delta = \Omega^{-1}(\mathbf{DB} - \mathbf{C})'[\mathbf{D}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{D}']^{-1}(\mathbf{DB} - \mathbf{C}) \quad (24.115)$$

is the non-centrality parameter. This is because

$$(\tilde{\mathbf{U}}'\tilde{\mathbf{U}} - \hat{\mathbf{U}}'\hat{\mathbf{U}}) = \mathbf{U}'\mathbf{M}_D\mathbf{U}$$

where

$$\mathbf{M}_D = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{D}'[\mathbf{D}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{D}']^{-1}\mathbf{D}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}', \quad (24.116)$$

and  $\mathbf{M}_D$  is a symmetric idempotent matrix ( $\mathbf{M}_D' = \mathbf{M}_D$ ,  $\mathbf{M}_D\mathbf{M}_D = \mathbf{M}_D$ ) with  $\text{rank}(\mathbf{M}_D) = \text{rank}(\mathbf{D}) = p$ . Given that  $\mathbf{M}_D$  and  $\mathbf{M}_X$  are orthogonal, i.e.

$$\mathbf{M}_D\mathbf{M}_X = \mathbf{0}, \quad (24.117)$$

we can also deduce that  $\mathbf{U}'\mathbf{M}_X\mathbf{U}$  and  $\mathbf{U}'\mathbf{M}_D\mathbf{U}$  are independently distributed (see Chapter 15). The analogy between the  $F$ -test statistic,

$$FT(\mathbf{y}) = \frac{\tilde{\mathbf{u}}'\tilde{\mathbf{u}} - \hat{\mathbf{u}}'\hat{\mathbf{u}}}{\hat{\mathbf{u}}'\hat{\mathbf{u}}} \left( \frac{T-k}{P} \right)^{H_0} \sim F(p, T-k), \quad (24.118)$$

in the case  $k=1$ , and  $\hat{\mathbf{G}}$  as defined in (112), is established. The problem, however, is that  $\hat{\mathbf{G}}$  is a random  $m \times m$  matrix, not a random variable as in the case of (118). The obvious way to reduce a matrix to a scalar is to use a matrix real-valued function such as the determinant or the trace of a matrix:

$$\tau_1(\mathbf{Y}) = \det[(\tilde{\mathbf{U}}'\tilde{\mathbf{U}} - \hat{\mathbf{U}}'\hat{\mathbf{U}})(\hat{\mathbf{U}}'\hat{\mathbf{U}})^{-1}], \quad (24.119)$$

$$\tau_2(\mathbf{Y}) = \text{tr}[(\tilde{\mathbf{U}}'\tilde{\mathbf{U}} - \hat{\mathbf{U}}'\hat{\mathbf{U}})(\hat{\mathbf{U}}'\hat{\mathbf{U}})^{-1}]. \quad (24.120)$$

In order to construct tests for  $H_0$  against  $H_1$  using the rejection regions

$$C_i = \{\mathbf{Y}: \tau_i(\mathbf{Y}) > c_i\}, \quad i = 1, 2, \quad (24.121)$$

we need the distribution of the test statistics  $\tau_1(\mathbf{Y})$  and  $\tau_2(\mathbf{Y})$ . These distributions can be derived from the joint distribution of the eigenvalues of  $\hat{\mathbf{G}}$ , say,  $\lambda_1, \lambda_2, \dots, \lambda_l$ , where  $l = \min(m, p)$ , because

$$\tau_1(\mathbf{Y}) = \sum_{i=1}^l \lambda_i \quad \text{and} \quad \tau_2(\mathbf{Y}) = \prod_{i=1}^l \lambda_i. \quad (24.122)$$

The distribution of  $\lambda \equiv (\lambda_1, \lambda_2, \dots, \lambda_l)'$  was derived by Constantine (1963) and James (1964) in terms of a zonal polynomial expansion. Constantine (1966) went on to derive the distribution of  $\tau_2(\mathbf{y})$  in terms of generalised Laguerre polynomials which is rather complicated to be used directly. For this reason several approximations in terms of the non-central chi-square distribution have been suggested in the statistical literature; see Muirhead (1982) for an excellent discussion. Based on such approximations several

592      **The multivariate linear regression model**

tables relating to the upper percentage points of

$$\tau_2^*(\mathbf{y}) = (T - k)\tau_2(\mathbf{y}) \quad (24.123)$$

have been constructed (see Davis (1970)). For large  $T - k$  we can also use the asymptotic result,

$$\tau_2^*(\mathbf{y}) \underset{\chi^2}{\overset{H_0}{\sim}} \chi^2(mp), \quad (24.124)$$

in order to derive  $c_x$  in (121). The test statistic is known as the *Lawley-Hotelling* statistic. Similar results can be derived for the determinental ratio test statistic

$$\tau_1^*(\mathbf{y}) = (T - k)\tau_1(\mathbf{y}). \quad (24.125)$$

The test statistics  $\tau_1(\mathbf{y})$  and  $\tau_2(\mathbf{y})$  can be interpreted as arising from the Wald test procedure discussed in Chapter 16. Using the other two test procedures, the likelihood ratio and Lagrange multiplier procedures, we can construct alternative tests for  $H_0$  against  $H_1$ . The *likelihood ratio* test procedure gives rise to the test statistic

$$LR(\mathbf{Y}) = \left( \frac{L(\hat{\boldsymbol{\theta}}; \mathbf{Y})}{L(\boldsymbol{\theta}; \mathbf{Y})} \right)^{2/T} = \frac{\det(\hat{\boldsymbol{\Omega}})}{\det(\boldsymbol{\Omega})} = \det[(\hat{\mathbf{U}}'\hat{\mathbf{U}})(\tilde{\mathbf{U}}'\tilde{\mathbf{U}})^{-1}]. \quad (24.126)$$

In terms of the eigenvalues of  $\hat{\mathbf{G}}$  this test statistic takes the form

$$LR(\mathbf{Y}) = \prod_{i=1}^l \left( \frac{1}{1 + \lambda_i} \right), \quad (24.127)$$

and thus its rejection region is defined by

$$C_1 = \{ \mathbf{Y} : LR(\mathbf{Y}) \leq c_x \}, \quad (24.128)$$

$c_x$  being determined by the distribution of  $LR(\mathbf{Y})$  under  $H_0$ . This distribution can be expressed as the product of  $p$  independent beta distributed random variables (see Johnson and Kotz (1970), Anderson (1984), *inter alia*). For large  $T$  we might also use the asymptotic result

$$-T^* \log LR(\mathbf{y}) \underset{\chi^2}{\overset{H_0}{\sim}} \chi^2(mp), \quad (24.129)$$

where  $T^* = [T - k - \frac{1}{2}(m - p + 1)]$  ( $p \geq m$ ); see Davis (1979) for tables of upper percentage points  $c_x$ .

The Lagrange multiplier test statistic based on the function

$$l(\boldsymbol{\theta}, \boldsymbol{\lambda}) = -\frac{T}{2} \det(\boldsymbol{\Omega}) - \text{tr}(\boldsymbol{\Omega}^{-1}(\mathbf{Y} - \mathbf{XB})(\mathbf{Y} - \mathbf{XB})') - \text{tr}(\boldsymbol{\Lambda}'(\mathbf{DB} - \mathbf{C})) \quad (24.130)$$

can be expressed in the form:

$$LM(\mathbf{Y}) = \text{tr}(\tilde{\mathbf{G}}), \quad (24.131)$$

where

$$\tilde{\mathbf{G}} = (\tilde{\mathbf{U}}'\tilde{\mathbf{U}} - \hat{\mathbf{U}}'\hat{\mathbf{U}})(\tilde{\mathbf{U}}'\tilde{\mathbf{U}})^{-1}. \quad (24.132)$$

This test statistic is known in the statistical literature as Pillai's trace test statistic because it was suggested by Pillai (1955), but not as a Lagrange multiplier test statistic. In terms of the eigenvalues  $\lambda_1, \lambda_2, \dots, \lambda_m$  this statistic takes the form

$$LM(\mathbf{Y}) = \sum_{i=1}^l \left( \frac{\lambda_i}{1+\lambda_i} \right). \quad (24.133)$$

The distribution of  $LM(\mathbf{Y})$  was obtained by Pillai and Jayachandran (1970) but this is also rather complicated to be used directly and several approximations have been suggested in the literature; see Pillai (1976), (1977), for references. For large  $T-k$  the critical value  $c_\alpha$  for a rejection region identical to (121) can be based on the asymptotic result

$$(T-k)LM(\mathbf{Y}) \underset{\alpha}{\stackrel{H_0}{\sim}} \chi^2(mp). \quad (24.134)$$

A similar test statistic known as *Wilks' ratio* test statistic is defined as the other matrix scalar function of  $\tilde{\mathbf{G}}$ , the determinant

$$\tau_3(\mathbf{Y}) = \det(\tilde{\mathbf{G}}). \quad (24.135)$$

In terms of the eigenvalues of  $\tilde{\mathbf{G}}$  this test statistic is

$$\tau_3(\mathbf{Y}) = \prod_{i=1}^l \left( \frac{\lambda_i}{1+\lambda_i} \right). \quad (24.136)$$

It is interesting to note that  $\tilde{\mathbf{G}}$  as defined above is directly related to multivariate goodness of fit measure  $\mathbf{G}$  as defined in Section 24.2 above.

Note that

$$\mathbf{G} = (\mathbf{Y}'\mathbf{Y} - \hat{\mathbf{U}}'\hat{\mathbf{U}})(\mathbf{Y}'\mathbf{Y})^{-1}. \quad (24.137)$$

In order to see the relationship let us consider the special case where

$$H_0: \mathbf{B}_2 = \mathbf{0}, \quad H_1: \mathbf{B}_2 \neq \mathbf{0}, \quad (24.138)$$

and

$$\mathbf{Y} = \mathbf{X}_1 \mathbf{B}_1 + \mathbf{X}_2 \mathbf{B}_2 + \mathbf{U}. \quad (24.139)$$

Defining the restricted residuals by  $\tilde{\mathbf{U}} = \mathbf{Y} - \mathbf{X}_1 \hat{\mathbf{B}}_1$  where  $\hat{\mathbf{B}}_1 = (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{Y}$  we can view  $\tilde{\mathbf{G}}$  as the multivariate multiple correlation

matrix of the auxiliary multiple regression

$$\hat{\mathbf{U}} = \mathbf{X}_1 \mathbf{B}_1 + \mathbf{X}_2 \mathbf{B}_2 + \mathbf{V}. \quad (24.140)$$

All the tests based on the test statistics mentioned so far are unbiased but no uniformly most powerful test exists for  $H_0$  against  $H_1$ ; see Giri (1977), Hart and Money (1976) for power comparisons.

A particularly important special case of  $H_0: \mathbf{DB} - \mathbf{C} = \mathbf{0}$  against  $H_1: \mathbf{DB} - \mathbf{C} \neq \mathbf{0}$  is the case where the sample period is divided into two subsamples, say,  $\mathbb{T}_1 = (1, 2, \dots, T_1)$  and  $\mathbb{T}_2 = (T_1 + 1, \dots, T)$ , where  $T - T_1 = T_2$  and  $T_1, T_2 > k$ . If we allow the conditional means for the two sub-periods to be different, i.e. for

$$t \in \mathbb{T}_1: \mathbf{Y}_1 = \mathbf{X}_1 \mathbf{B}_1 + \mathbf{U}_1, \quad (24.141)$$

$$t \in \mathbb{T}_2: \mathbf{Y}_2 = \mathbf{X}_2 \mathbf{B}_2 + \mathbf{U}_2, \quad (24.142)$$

but the conditional variances to be the same, i.e.

$$E(\mathbf{Y}_i / \mathcal{X}_i = \mathbf{X}_i) = \boldsymbol{\Omega} \otimes \mathbf{I}_{T_i}, \quad i = 1, 2, \quad (24.143)$$

then the hypothesis:  $H_0: \mathbf{B}_1 = \mathbf{B}_2$  against  $H_1: \mathbf{B}_1 \neq \mathbf{B}_2$  can be accommodated into the above formulation with

$$\mathbf{D} = (\mathbf{I}_k, -\mathbf{I}_k), \quad \mathbf{B} = \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix}, \quad \mathbf{C} = \mathbf{0}$$

and

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & 0 \\ 0 & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix} + \begin{pmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{pmatrix}. \quad (24.144)$$

This is a direct extension of the  $F$ -test for structural change considered in Chapter 21. In the same chapter it was argued that we need to test the equality of the conditional variances before we apply the coefficient constancy test.

A natural way to proceed in order to test the hypothesis  $H_0: \boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2$  against  $H_1: \boldsymbol{\Omega}_1 \neq \boldsymbol{\Omega}_2$  is to generalise the  $F$ -test derived in Chapter 21. That is, use a scalar function of the 'ratio':

$$\mathbf{V} = \hat{\boldsymbol{\Omega}}_2 \hat{\boldsymbol{\Omega}}_1^{-1}, \quad (24.145)$$

where

$$\hat{\boldsymbol{\Omega}}_i = \frac{1}{T_i} \hat{\mathbf{U}}_i' \hat{\mathbf{U}}_i, \quad i = 1, 2.$$

such as

$$(i) \quad d_1 = \det(\hat{\boldsymbol{\Omega}}_2 \hat{\boldsymbol{\Omega}}_1^{-1});$$

$$(ii) \quad d_2 = \text{tr}(\hat{\boldsymbol{\Omega}}_2 \hat{\boldsymbol{\Omega}}_1^{-1}).$$

## 24.6 Misspecification testing

Misspecification testing in the context of the multivariate linear regression model is of considerable interest in econometric modelling because of its relationship to the simultaneous equations model to be discussed in Chapter 25. As argued above, the latter model is a reparametrisation of the former and the reparametrisation can at best be as well defined (statistically) as the statistical parameters  $\theta \equiv (\mathbf{B}, \boldsymbol{\Omega})$ . In practice, before any questions related to the theoretical parameters of interest  $\xi$  can be asked the misspecification testing for  $\theta$  must be successfully completed.

As far as assumptions [1]–[8] are concerned the discussion in Chapters 19 to 22 is directly applicable with minor modifications. Let us consider the probability and sampling model assumptions in the present case where  $m > 1$ .

*Normality.* The assumption that  $D(\mathbf{y}_t/\mathbf{X}_t; \theta)$  is multivariate normal can be tested using a multivariate extension of the skewness–kurtosis test. The skewness and kurtosis coefficients for a random vector  $\mathbf{u}_t$  with mean zero and covariance matrix  $\boldsymbol{\Omega}$  are defined by

$$\alpha_{3,m} = [E(\mathbf{u}_t' \boldsymbol{\Omega}^{-1} \mathbf{u}_t)^3]^{\frac{1}{2}} \quad \text{and} \quad \alpha_{4,m} = E(\mathbf{u}_t' \boldsymbol{\Omega}^{-1} \mathbf{u}_t)^2, \quad (24.157)$$

respectively. In the case where  $\mathbf{u}_t \sim N(\mathbf{0}, \boldsymbol{\Omega})$ ,  $\alpha_{3,m} = 0$  and  $\alpha_{4,m} = m(m+2)$ . These coefficients can be estimated by

$$\hat{\alpha}_{3,m}^2 = \frac{1}{T^2} \sum_{t=1}^T \sum_{s=1}^T g_{ts}^3, \quad \hat{\alpha}_{4,m} = \frac{1}{T} \sum_{t=1}^T g_{tt}^2, \quad (24.158)$$

where

$$g_{ts} = \hat{\mathbf{u}}_t' \hat{\boldsymbol{\Omega}}^{-1} \hat{\mathbf{u}}_s, \quad t, s = 1, 2, \dots, T. \quad (24.159)$$

*Asymptotically*, we can show that

$$(i) \quad \frac{T}{6} \hat{\alpha}_{3,m}^2 \xrightarrow{H_0} \chi^2(l), \quad l = \frac{1}{6}m(m+1)(m+2) \quad (24.160)$$

and

$$(ii) \quad \frac{T}{8m(m+2)} (\hat{\alpha}_{4,m} - m(m+2))^2 \xrightarrow{H_0} \chi^2(1). \quad (24.161)$$

Note that in the case where  $m = 1$ :

$$\frac{T}{6} \hat{\alpha}_{3,m}^2 \xrightarrow{H_0} \chi^2(1) \quad \text{and} \quad \frac{T}{24} (\hat{\alpha}_{4,m} - 3)^2 \xrightarrow{H_0} \chi^2(1). \quad (24.162)$$

Using (160) and (161) we can define separate tests for the hypothesis

$$H_0^{(1)}: \alpha_{3,m} = 0 \quad \text{against} \quad H_1^{(1)}: \alpha_{3,m} \neq 0$$

and

$$H_0^{(2)}: \alpha_{4,m} = m(m+2) \quad H_1^{(2)}: \alpha_{4,m} \neq m(m+2)$$

respectively. When  $D(\mathbf{y}_t, \mathbf{X}_t; \boldsymbol{\theta})$  is normal then  $H_0^{(1)} \cap H_0^{(2)}$  is valid.

As in the case where  $m=1$ , the above tests based on the residuals skewness and kurtosis coefficients are rather sensitive to outliers and should be used with caution.

For further discussion of tests for multivariate normality see Mardia (1980), Small (1980) and Seber (1984), *inter alia*.

*Linearity.* A test for the linearity of the conditional mean can be based on the auxiliary regression

$$\hat{\mathbf{u}}_t = (\mathbf{B}_0 - \hat{\mathbf{B}})' \mathbf{x}_t + \boldsymbol{\Gamma}' \boldsymbol{\psi}_t + \boldsymbol{\varepsilon}_t, \quad t = 1, 2, \dots, T. \quad (24.163)$$

where  $\boldsymbol{\psi}_1 \equiv (\psi_{1t}, \dots, \psi_{pt})'$  are the higher-order terms related to the Kolmogorov–Gabor or RESET type polynomials (see (21.10) and (21.11)). The hypothesis to be tested in the present case takes the form

$$H_0: \boldsymbol{\Gamma} = \mathbf{0} \quad \text{against} \quad H_1: \boldsymbol{\Gamma} \neq \mathbf{0}. \quad (24.164)$$

This hypothesis can be tested using any one of the tests  $\tau_i(\mathbf{Y})$ ,  $i = 1, 2, 3$ ,  $LR(\mathbf{Y})$  or  $LM(\mathbf{Y})$  discussed in Section 24.4.

*Homoskedasticity.* A direct extension of the White test for departures from homoskedasticity is based on the following multivariate linear auxiliary regression:

$$\hat{\boldsymbol{\phi}}_t = c_0 + \mathbf{C}' \boldsymbol{\psi}_t + \boldsymbol{\varepsilon}_t, \quad (24.165)$$

where

$$\hat{\boldsymbol{\phi}}_t \equiv (\hat{\phi}_{1t}, \hat{\phi}_{2t}, \dots, \hat{\phi}_{qt})'$$

and

$$\hat{\phi}_{it} = \hat{u}_{it} \hat{u}_{jt}, \quad i \geq j = 1, 2, \dots, m, \quad q = \frac{1}{2}m(m+1). \quad (24.166)$$

Testing for homoskedasticity can be based on

$$H_0: \mathbf{C} = \mathbf{0} \quad \text{against} \quad H_1: \mathbf{C} \neq \mathbf{0}. \quad (24.167)$$

A linear set of restrictions which can be tested using the tests discussed in Section 24.4 above. The main difference with the  $m=1$  case is that we have the cross-products of the residuals in addition to the cross-products of the regressors.

*Time invariance and structural change.* The discussion of departures from the time invariance of  $\theta \equiv (\beta, \sigma^2)$  in the  $m=1$  case was based on the behaviour of the recursive estimators of  $\theta$ , say  $\theta_t$ ,  $t=k+1, \dots, T$ . This discussion can be generalised directly to the multivariate case where  $\theta \equiv (\mathbf{B}, \Omega)$  without any great difficulty. The same applies to the discussion of structural change whose tests have been considered briefly in Section 24.4.

*Independence.* Using the analogy with the  $m=1$  case we can argue that when  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$ ,  $\mathbf{Z}_t \equiv (\mathbf{y}_t'; \mathbf{X}_t')$  is assumed to be a normal, stationary and  $l$ -th-order Markov process (see Chapter 8), the statistical GM takes the form

$$\mathbf{y}_t = \mathbf{B}'_0 \mathbf{x}_t + \sum_{i=1}^l \mathbf{A}'_i \mathbf{y}_{t-i} + \sum_{i=1}^l \mathbf{B}'_i \mathbf{x}_{t-i} + \varepsilon_t, \quad (24.168)$$

where  $\{\varepsilon_t, t > l\}$  is a vector innovation process. If we compare (168) with the statistical GM under independence

$$\mathbf{y}_t = \mathbf{B}' \mathbf{x}_t + \mathbf{u}_t, \quad (24.169)$$

we can see that the independence assumption can be tested using the auxiliary multivariate regression

$$\hat{\mathbf{u}}_t = (\mathbf{B}_0 - \hat{\mathbf{B}})' \mathbf{x}_t + \sum_{i=1}^l [\mathbf{A}'_i \mathbf{y}_{t-i} + \mathbf{B}'_i \mathbf{x}_{t-i}] + \varepsilon_t. \quad (24.170)$$

In particular, the hypothesis of interest takes the form

$$H_0: \mathbf{A}_i = \mathbf{0} \quad \text{and} \quad \mathbf{B}_i = \mathbf{0} \quad \text{for all } i = 1, 2, \dots, l$$

against

$$H_1: \mathbf{A}_i \neq \mathbf{0} \quad \text{or} \quad \mathbf{B}_i \neq \mathbf{0} \quad \text{for any } i = 1, \dots, l.$$

This hypothesis can be tested using the multivariate  $F$ -type tests discussed in Section 24.4.

The independence test which corresponds to the autocorrelation approach (see Chapter 22) could be based on the auxiliary multivariate regression

$$\hat{\mathbf{u}}_t = \mathbf{D}_0 \mathbf{x}_t + \mathbf{C}'_1 \hat{\mathbf{u}}_{t-1} + \cdots + \mathbf{C}'_l \hat{\mathbf{u}}_{t-l} + \mathbf{v}_t. \quad (24.171)$$

That is, test  $H_0: \mathbf{C}_1 = \mathbf{C}_2 = \cdots = \mathbf{C}_l = \mathbf{0}$  against  $H_1: \mathbf{C}_i \neq \mathbf{0}$  for any  $i = 1, 2, \dots, l$ . This can also be tested using the tests developed in Section 24.4 for linear restrictions.

Testing for departures from the independence assumption is particularly important in econometric modelling with time-series data. When the assumption is inappropriate a respecification of the multivariate linear regression model gives rise to the multivariate dynamic linear regression model which is very briefly considered in Section 20.8.

## 24.7 Prediction

In view of the assumption that

$$\mathbf{y}_t = \mathbf{B}' \mathbf{x}_t + \mathbf{u}_t, \quad t \in \mathbb{T}, \quad (24.172)$$

the best predictor of  $\mathbf{y}_{T+l}$ , given that the observations  $t = 1, 2, \dots, T$  were used to estimate  $\mathbf{B}$  and  $\boldsymbol{\Omega}$ , can only be its conditional expectation

$$\hat{\mathbf{y}}_{T+l} = \hat{\mathbf{B}}' \mathbf{x}_{T+l}, \quad l = 1, 2, \dots, \quad (24.173)$$

where  $\mathbf{x}_{T+l}$  represents the observed value of the random vector  $\mathbf{X}_t$  at  $t = T+l$ . The prediction error is

$$\mathbf{e}_{T+l} \equiv \mathbf{y}_{T+l} - \hat{\mathbf{y}}_{T+l} = (\hat{\mathbf{B}} - \mathbf{B})' \mathbf{x}_{T+l} + \mathbf{u}_{T+l}. \quad (24.174)$$

Given that  $\mathbf{e}_{T+l}$  is a linear function of normally distributed r.v.'s,

$$\mathbf{e}_{T+l} \sim N(\mathbf{0}, \boldsymbol{\Omega}(1 + \mathbf{x}'_{T+l}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_{T+l})) \quad (24.175)$$

(see exercise 7), which is a direct generalisation of the prediction error distribution in the case of the linear regression model. Since  $\boldsymbol{\Omega}$  is unknown its unbiased estimator

$$\mathbf{S} = \frac{1}{T-k} \hat{\mathbf{U}}' \hat{\mathbf{U}} \quad (24.176)$$

is used to construct the prediction test statistic

$$H = (\mathbf{y}_{T+l} - \hat{\mathbf{y}}_{T+l})' \mathbf{S}_F^{-1} (\mathbf{y}_{T+l} - \hat{\mathbf{y}}_{T+l}), \quad (24.177)$$

where

$$\mathbf{S}_F = \mathbf{S}(1 + \mathbf{x}'_{T+l}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_{T+l}). \quad (24.178)$$

Hotelling (1931) showed that

$$H^* = \frac{(T-k-m+1)H}{(T-k)m} \sim F(m, T-k-m+1), \quad (24.179)$$

and this can be used to test hypotheses about the predictions or construct prediction regions.

## 24.8 The multivariate dynamic linear regression (MDLR) model

In direct analogy to the  $m=1$  case the MDLR model is specified as follows:

### I Statistical GM

$$\mathbf{y}_t = \mathbf{B}'_0 \mathbf{x}_t + \sum_{i=1}^l \mathbf{A}'_i \mathbf{y}_{t-i} + \sum_{i=1}^l \mathbf{B}'_i \mathbf{x}_{t-i} + \mathbf{u}_t, \quad t > l. \quad (24.180)$$

600      **The multivariate linear regression model**

- [1]  $\mu_t = E(\mathbf{y}_t/\sigma(\mathbf{Y}_{t-1}^0), \mathbf{X}_t^0 = \mathbf{x}_t)$  and  $\mathbf{u}_t = \mathbf{y}_t - E(\mathbf{y}_t/\sigma(\mathbf{Y}_{t-1}^0), \mathbf{X}_t^0 = \mathbf{x}_t^0)$ .
- [2]  $\theta^* \equiv (\mathbf{A}_1, \dots, \mathbf{A}_l, \mathbf{B}_0, \mathbf{B}_1, \dots, \mathbf{B}_l, \boldsymbol{\Omega}_0)$  are the statistical parameters of interest.
- [3]  $\mathbf{X}_t$  is strongly exogenous with respect to  $\theta^*$ .
- [4] The roots of the matrix polynomial

$$\left( \lambda^l \mathbf{I} - \sum_{i=1}^{l-1} \mathbf{A}_i \lambda^{l-i} \right) = \mathbf{0}$$

lie within the unit circle.

- [5] Rank( $\mathbf{X}^*$ ) =  $k^*$  where  $\mathbf{X}^* \equiv (\mathbf{Y}_{-1}, \dots, \mathbf{Y}_{-l}, \mathbf{X}, \mathbf{X}_{-1}, \dots, \mathbf{X}_{-l})$ ,  $k^* = mk + lm(k-1) + lm^2$ .

**(II) Probability model**

$$\begin{aligned} \Phi = & \left\{ D(\mathbf{y}_t/\mathbf{Z}_{t-1}^0; \theta^*) \right. \\ & = \frac{(\det \boldsymbol{\Omega}_0)^{-\frac{1}{2}}}{(2\pi)^{m/2}} \times \exp\left\{ -\frac{1}{2}(\mathbf{y}_t - \mathbf{B}^{*\prime} \mathbf{X}_t^*)' \boldsymbol{\Omega}_0^{-1} (\mathbf{y}_t - \mathbf{B}^{*\prime} \mathbf{X}_t^*) \right\}, \\ & \quad \left. \theta^* \in \Theta, t \in \mathbb{T} \right\}. \end{aligned} \quad (24.181)$$

- [6] (i)  $D(\mathbf{y}_t/\mathbf{Z}_{t-1}^0; \theta^*)$  – normal;
- (ii)  $E(\mathbf{y}_t/\sigma(\mathbf{Y}_{t-1}^0), \mathbf{X}_t^0 = \mathbf{x}_t^0) = \mathbf{B}^{*\prime} \mathbf{X}_t^*$  – linear in  $\mathbf{X}_t^*$ ;
- (iii)  $\text{Cov}(\mathbf{y}_t/\sigma(\mathbf{Y}_{t-1}^0), \mathbf{X}_t^0 = \mathbf{x}_t^0) = \boldsymbol{\Omega}_0$  – homoskedastic;
- [7]  $\theta^*$  is time invariant.

**(III) Sampling model**

- [8]  $\mathbf{Y} \equiv (\mathbf{Y}_1, \dots, \mathbf{Y}_T)'$  is a non-random sample sequentially drawn from  $D(\mathbf{y}_t/\mathbf{Z}_{t-1}^0; \theta^*)$ ,  $t = 1, 2, \dots, T$ , respectively.

*Note:*

$$\mathbf{B}^{*\prime} \equiv (\mathbf{A}'_1, \mathbf{A}'_2, \dots, \mathbf{A}'_l, \mathbf{B}'_0, \mathbf{B}'_1, \dots, \mathbf{B}'_l),$$

$$\mathbf{X}^* \equiv (\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \dots, \mathbf{y}_{t-l}, \mathbf{x}_t, \mathbf{x}_{t-1}, \dots, \mathbf{x}_{t-l}).$$

The estimation, misspecification and specification testing in the context of this statistical model follows closely that of the multivariate linear regression model considered in Sections 24.2–24.5 above. The modifications to these results needed to apply to the MDLR model are analogous to the ones considered in the context of the  $m=1$  case (see Chapter 23). In particular the approximate MLE's of  $\theta^*$

$$\hat{\mathbf{B}}^* = (\mathbf{X}^{*\prime} \mathbf{X}^*)^{-1} \mathbf{X}^* \mathbf{Y} \quad (24.182)$$

and

$$\hat{\Omega}_0 = \frac{1}{T} \hat{\mathbf{U}}^* \hat{\mathbf{U}}^*, \quad \hat{\mathbf{U}}^* = \mathbf{Y} - \mathbf{X}^* \hat{\mathbf{B}}^* \quad (24.183)$$

behave asymptotically like  $\hat{\mathbf{B}}$  and  $\hat{\Omega}$  (see Anderson and Taylor (1979)). Moreover, the multivariate  $F$ -type tests considered in Section 24.4 are asymptotically justifiable in the context of the MDLR model.

For policy analysis and prediction purposes it is helpful to reformulate the statistical GM (180) in order to express it in the first-order autoregression form

$$\mathbf{y}^* = \mathbf{A}_1^{*\prime} \mathbf{y}_{t-1}^* + \mathbf{B}_1^{*\prime} \mathbf{Z}_t^* + \mathbf{u}_t^*, \quad (24.184)$$

where

$$\mathbf{y}_t^* \equiv \begin{pmatrix} \mathbf{y}_t \\ \mathbf{y}_{t-1} \\ \vdots \\ \mathbf{y}_{t-l+1} \end{pmatrix}, \quad \mathbf{u}_t^* \equiv \begin{pmatrix} \mathbf{u}_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}, \quad \mathbf{A}_1^* \equiv \begin{pmatrix} \mathbf{A}_1 & -\mathbf{I}_m & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{A}_2 & \mathbf{0} & -\mathbf{I}_m & & \vdots \\ \vdots & \vdots & & \ddots & \mathbf{0} \\ \mathbf{A}_l & \mathbf{0} & & & \mathbf{0} \end{pmatrix},$$

$$\mathbf{B}_1^* \equiv \begin{pmatrix} \mathbf{B}_0 \\ \mathbf{B}_1 \\ \vdots \\ \mathbf{B}_l \end{pmatrix}, \quad \mathbf{Z}_t^* \equiv \begin{pmatrix} \mathbf{x}_t \\ \mathbf{x}_{t-1} \\ \vdots \\ \mathbf{x}_{t-l} \end{pmatrix}.$$

This form can be viewed as a first-order non-homogeneous vector difference equation (see Miller (1968)) which becomes

$$\mathbf{y}_t^* = (\mathbf{A}_1^{*\prime})' \mathbf{y}_t^* + \sum_{i=0}^{t-1} (\mathbf{A}_1^{*\prime})^i \mathbf{B}_1^* \mathbf{Z}_{t-i}^* + \sum_{i=0}^{t-1} (\mathbf{A}_1^{*\prime})^i \mathbf{u}_{t-i} \quad (24.185)$$

by repeated substitution. Assuming that  $\lim_{t \rightarrow \infty} (\mathbf{A}_1^*)^t = \mathbf{0}$  (compare with  $|x_1| < 1$  in the  $m=1$  case), then the solution of (184) is

$$\mathbf{y}_t^* = \sum_{i=0}^t (\mathbf{A}_1^{*\prime})^i \mathbf{B}_1^* + \sum_{i=0}^t (\mathbf{A}_1^{*\prime})^i \mathbf{u}_{t-i}. \quad (24.186)$$

This is known in the econometric literature as the *final form*, with

$$\mathbf{M}_0 = \mathbf{B}_1^*, \quad (24.187)$$

$$\mathbf{M}_\tau = \mathbf{B}_1^* \mathbf{A}_1^{*\prime}, \quad \tau = 1, 2, \dots. \quad (24.188)$$

known as the *impact* and *interim multipliers* of delay  $\tau$  respectively. The solution in (186) provides us also with the so-called *long-run equilibrium*

multiplier matrix defined by

$$\mathbf{L} = \sum_{\tau=0}^{\infty} \mathbf{B}_1^* \mathbf{A}_1^{*\tau} = \mathbf{B}_1^* (\mathbf{I} - \mathbf{A}_1^*)^{-1}. \quad (24.189)$$

The elements of this matrix  $l_{ij}$  refer to the total expected response of the  $j$ th endogenous variable to the sustained unit change in the  $i$ th exogenous variable, holding the other exogenous variables constant (see Schmidt (1973), (1979), for a discussion of the statistical analysis of these multipliers).

Returning to the question of prediction we can see that the natural predictor for  $\mathbf{y}_{T+1}$  given by  $\mathbf{y}_1, \dots, \mathbf{y}_T$  and  $\mathbf{x}_1, \dots, \mathbf{x}_{T+1}$  is

$$\hat{\mathbf{y}}_{T+1} = \hat{\mathbf{A}}_1^{*\prime} \mathbf{y}_T + \hat{\mathbf{B}}_1^{*\prime} \mathbf{Z}_{T+1}^*. \quad (24.190)$$

In order to predict  $\mathbf{y}_{T+2}$  we need to know  $\mathbf{x}_{T+1}$ ,  $\mathbf{x}_{T+2}$  as well as  $\mathbf{y}_{T+1}$ . Assuming that  $\mathbf{x}_{T+1}$  and  $\mathbf{x}_{T+2}$  are available we can use the predictor of  $\mathbf{y}_{T+1}$ ,  $\hat{\mathbf{y}}_{T+1}$  in order to get  $\mathbf{y}_{T+2}$  in

$$\begin{aligned} \hat{\mathbf{y}}_{T+2} &= \hat{\mathbf{A}}_1^{*\prime} \mathbf{y}_{T+1} + \hat{\mathbf{B}}_1^{*\prime} \mathbf{Z}_{T+2}^* \\ &= \hat{\mathbf{A}}_1^{*\prime} (\hat{\mathbf{A}}_1^{*\prime} \mathbf{y}_T + \hat{\mathbf{B}}_1^{*\prime} \mathbf{Z}_{T+1}^*) + \hat{\mathbf{B}}_1^{*\prime} \mathbf{Z}_{T+2}^* \\ &= (\hat{\mathbf{A}}_1^{*\prime})^2 \mathbf{y}_T + \hat{\mathbf{A}}_1^{*\prime} \hat{\mathbf{B}}_1^{*\prime} \mathbf{Z}_{T+1}^* + \hat{\mathbf{B}}_1^{*\prime} \mathbf{Z}_{T+2}^*. \end{aligned} \quad (24.191)$$

Hence,

$$\hat{\mathbf{y}}_{T+\tau} = (\hat{\mathbf{A}}_1^{*\prime})^\tau \mathbf{y}_T + \sum_{j=1}^{\tau} (\hat{\mathbf{A}}_1^{*\prime})^{\tau-j} \hat{\mathbf{B}}_1^{*\prime} \mathbf{Z}_{T+j}^*, \quad \tau = 1, 2, \dots \quad (24.192)$$

will provide predictions for future values of  $\mathbf{y}_t$  assuming that the values taken by the regressors are available. For the asymptotic covariance matrix of the prediction error  $(\mathbf{y}_{T+\tau} - \hat{\mathbf{y}}_{T+\tau})$  see Schmidt (1974).

Prediction in the context of the multivariate (dynamic) linear regression model is particularly important in econometric modelling because of its relationship with the simultaneous equation model discussed in the next chapter. The above discussion of prediction carries over to the simultaneous equation model with minor modifications and the concepts of impact, interim and long-run multipliers are useful in simulations and policy analysis.

### Appendix 24.1 – The Wishart distribution

Let  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  be a sequence of  $n \times 1$  independent random vectors such that  $\mathbf{Z}_t \sim N(\mathbf{0}, \boldsymbol{\Omega})$ ,  $t \in \mathbb{T}$ , then for  $T \geq n$ ,  $\mathbf{S} = \sum_{t=1}^T \mathbf{Z}_t \mathbf{Z}_t'$ ,  $\mathbf{S} \sim W_n(\boldsymbol{\Omega}, T)$ , where the density function of  $\mathbf{S}$  is

$$D(\mathbf{S}; \boldsymbol{\theta}) = \frac{c(\det \mathbf{S})^{[(T-n-1)/2]}}{(\det \boldsymbol{\Omega})^{[T/2]}} \exp\left\{-\frac{1}{2} \text{tr } \boldsymbol{\Omega}^{-1} \mathbf{S}\right\}$$

where

$$c = \left[ 2^{Tn/2} \pi^{\{[n(n-1)]/4\}} \prod_{i=1}^n \Gamma\left(\frac{T+1-i}{2}\right) \right]^{-1},$$

$\Gamma(\cdot)$  being the gamma function (see Press (1972)).

### Properties of the Wishart distribution

- (i) If  $S_1, \dots, S_k$  are  $n \times n$  random independent matrices and  $S_i \sim W_n(\Omega, T_i)$ ,  $i = 1, 2, \dots, k$ , then

$$\left( \sum_{i=1}^k S_i \right) \sim W_n(\Omega, T),$$

where

$$T = \sum_{i=1}^k T_i.$$

- (ii) If  $S \sim W(\Omega, T)$  and  $M$  is a  $k \times n$  matrix of rank  $k$  then

$$MSM' \sim W_k(M\Omega M', T)$$

(see Muirhead (1982), Press (1972) *inter alia*).

These results enable us to deduce that if  $S \sim W_n(\Omega, T)$  and  $S$  and  $\Omega$  are partitioned conformally as

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}, \quad \Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix},$$

then

- (a)  $S_{1i} \sim W_{ni}(\Omega_{ii}, T)$ ,  $i = 1, 2, \dots$ , where  $n_1 + n_2 = n$ ,  $S_{11}: n_1 \times n_1$ ,  $S_{22}: n_2 \times n_2$ .
- (b)  $S_{11}$  and  $S_{12}$  are independent if  $\Omega_{12} = 0$ .
- (c)  $(S_{11} - S_{12}S_{22}^{-1}S_{21}) \sim W_{n_1}(\Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}, T - n_2)$  and is independent of  $S_{12}$  and  $S_{22}$ .
- (d)  $(S_{12}/S_{22}) \sim N(\Omega_{12}\Omega_{22}^{-1}S_{22}, (\Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}) \otimes S_{22})$ .

### Appendix 24.2 – Kronecker products and matrix differentiation

The Kronecker product between two arbitrary matrices  $A: m \times n$  and  $B: p \times q$  is defined by

$$A \otimes B = \begin{pmatrix} \alpha_{11}B & \alpha_{12}B & \cdots & \alpha_{1n}B \\ \alpha_{21}B & \alpha_{22}B & & \\ \vdots & & & \\ \alpha_{m1}B & & & \alpha_{mn}B \end{pmatrix}.$$

## 604 The multivariate linear regression model

Let  $\mathbf{A}$ ,  $\mathbf{B}$ ,  $\mathbf{C}$  and  $\mathbf{D}$  be arbitrary matrices, then

- (i)  $\mathbf{A} \otimes (\alpha \mathbf{B}) = \alpha(\mathbf{A} \otimes \mathbf{B})$ ,  $\alpha$  being a scalar;
- (ii)  $(\mathbf{A} + \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes \mathbf{C} + \mathbf{B} \otimes \mathbf{C}$ ,  $\mathbf{A}$  and  $\mathbf{B}$  being of the same order;
- (iii)  $\mathbf{A} \otimes (\mathbf{B} + \mathbf{C}) = \mathbf{A} \otimes \mathbf{B} + \mathbf{A} \otimes \mathbf{C}$ ,  $\mathbf{B}$  and  $\mathbf{C}$  being of the same order;
- (iv)  $\mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C}) = (\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C}$ ;
- (v)  $(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'$ ;
- (vi)  $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}$ ;
- (vii)  $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$ ,  $\mathbf{A}$  and  $\mathbf{B}$  being square non-singular matrices;
- (viii)  $\text{vec}(\mathbf{ACB}) = (\mathbf{B}' \otimes \mathbf{A}) \text{vec}(\mathbf{C})$ ;
- (ix)  $\text{tr}(\mathbf{A} \otimes \mathbf{B}) = (\text{tr } \mathbf{A})(\text{tr } \mathbf{B})$ ,  $\mathbf{A}$  and  $\mathbf{B}$  being square matrices;
- (x)  $\det(\mathbf{A} \otimes \mathbf{B}) = (\det \mathbf{A})^m (\det \mathbf{B})^n$ ,  $\mathbf{A}$  and  $\mathbf{B}$  being  $n \times n$  and  $m \times m$  matrices;
- (xi)  $\text{vec}(\mathbf{A} + \mathbf{B}) = \text{vec}(\mathbf{A}) + \text{vec}(\mathbf{B})$ ,  $\mathbf{A}$  and  $\mathbf{B}$  being of the same order;
- (xii)  $\text{tr}(\mathbf{AB}) = (\text{vec}(\mathbf{B}'))(\text{vec}(\mathbf{A}))$ .

### *Useful derivatives*

- (i)  $\frac{\partial \log(\det \mathbf{A})}{\partial \mathbf{A}} = (\mathbf{A}^{-1})'$ ;
- (ii)  $\frac{\partial \text{tr}(\mathbf{AB})}{\partial \mathbf{B}} = \mathbf{A}'$ ;
- (iii)  $\frac{\partial \text{tr}(\mathbf{A}' \mathbf{B})}{\partial \mathbf{B}} = \mathbf{A}$ ;
- (iv)  $\frac{\partial \text{tr}(\mathbf{X}' \mathbf{AXB})}{\partial \mathbf{X}} = \mathbf{AXB} + \mathbf{A}' \mathbf{XB}'$ ;
- (v)  $\frac{\partial \text{tr}(\mathbf{A}^n)}{\partial \mathbf{A}} = n \mathbf{A}^{-1}$ ;
- (vi)  $\frac{\partial \text{vec}(\mathbf{AXB})}{\partial \text{vec } \mathbf{X}} = \mathbf{B}' \otimes \mathbf{A}$ .

### *Important concepts*

Wishart distribution, trace correlation, coefficient of alienation, iterative numerical optimisation, SURE and Malinvaud formulations, estimated GLS estimator, exclusion restrictions, linear homogeneous restrictions, final form, impact, interim and long-run multipliers.

**Questions**

1. Compare the linear regression and multivariate linear regression statistical models.
2. Explain how linearity and homoskedasticity are related to the normality of  $\mathbf{Z}_t$ .
3. Compare the formulations  $\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{U}$  and  $\mathbf{y}_* = \mathbf{X}_*\beta_* + \mathbf{u}_*$ .
4. ‘How do you explain the fact that, although  $y_{1t}, y_{2t}, \dots, y_{mt}$  are correlated, the MLE estimator of  $\mathbf{B}$ , given by

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

does not involve  $\hat{\Omega}$ ?’ Derive the distribution of  $\hat{\mathbf{B}}$ .

5. Explain why the assumption  $T \geq m+k$  is needed for the existence of the MLE  $\hat{\Omega}$  of  $\Omega$ . Discuss the distribution of  $\hat{\Omega}$ .
6. Discuss the relationship between the goodness-of-fit measures  $d_1 = (1/m) \text{tr } \mathbf{G}$  and  $d_2 = \det(\mathbf{G})$  where

$$\mathbf{G} = \mathbf{I} - (\mathbf{Y}'\mathbf{Y})^{-1}\hat{\mathbf{U}}'\hat{\mathbf{U}}$$

with  $d = [\det(\Sigma)] / [\det(\Sigma_{11}) \det(\Sigma_{22})]$ , known as Hotelling’s *alienation coefficient*.

7. State the distributions of the MLE’s  $\hat{\mathbf{B}}$  and  $\hat{\Omega}$  and discuss the properties which can be deduced from their distributions.
8. Discuss intuitively how the conditions

$$(i) \quad \lim_{T \rightarrow \infty} (\mathbf{X}'\mathbf{X})^{-1} = \mathbf{0}; \quad \text{and}$$

$$(ii) \quad \left| \frac{\lambda_l}{\lambda_s} \right| < K \quad \text{for all } T;$$

a.s.  
imply that  $\hat{\mathbf{B}} \rightarrow \mathbf{B}$ .

9. Give two examples for each of the following forms of restrictions:

  - (i)  $\mathbf{D}_1\mathbf{B} + \mathbf{C}_1 = \mathbf{0}$ ;
  - (ii)  $\mathbf{B}\Gamma_1 + \Delta_1 = \mathbf{0}$ .

Discuss the differences between them.

10. Explain how the linear restrictions formulation  $\mathbf{R}\beta_* = \mathbf{r}$  generalises (i) and (ii) in question 9.
11. Verify the following equality:

$$\begin{aligned} \tilde{\mathbf{B}} &= \hat{\mathbf{B}} - (\hat{\mathbf{B}}\Gamma_1 + \Delta_1)(\Gamma_1'\hat{\Omega}\Gamma_1)^{-1}\Gamma_1'\hat{\Omega} = \hat{\mathbf{B}} - \\ &\quad (\hat{\mathbf{B}}\Gamma_1 + \Delta_1)(\Gamma_1'\tilde{\Omega}\Gamma_1)^{-1}\Gamma_1'\tilde{\Omega}. \end{aligned}$$

12. What is the reason for the interest in the Zellner and Malinvaud formulations?

606      **The multivariate linear regression model**

13. Explain how the GLS-type estimators of  $\beta_*$  for the Zellner and Malinvaud formulations can be derived using some numerical optimisation iterative formula.
14. Discuss the question of constructing a test for

$$H_0: \mathbf{DB} - \mathbf{C} = \mathbf{0} \quad \text{against} \quad H_1: \mathbf{DB} - \mathbf{C} \neq \mathbf{0}.$$

15. Compare the following test statistics defined in Section 24.5:

$$\tau_1(\mathbf{Y}), \tau_2(\mathbf{Y}), LR(\mathbf{Y}), LM(\mathbf{Y}), \tau_3(\mathbf{Y}).$$

16. Discuss the question of testing for departures from normality and compare it with the same test for the  $m=1$  case.
17. Explain how you would go about testing for linearity, homoskedasticity and independence in the context of the multivariate linear regression model.
18. ‘Misspecification testing in the context of the multivariate dynamic linear regression model is related to that of the non-dynamic model in the same way as the dynamic linear regression is related to the linear regression model.’ Discuss.
19. Explain the concepts of impact, interim and long-run multipliers and discuss their usefulness.

**Exercises**

1. Verify the following:
  - (i)  $\text{vec}(\mathbf{B}) \neq \text{vec}(\mathbf{B}')$ ;
  - (ii)  $\text{Cov}(\text{vec}(\mathbf{Y}) \text{ vec}(\mathbf{Y}')') = \mathbf{I}_T \otimes \boldsymbol{\Omega}$ ;
  - (iii)  $\sum_t (\mathbf{y}_t - \mathbf{B}' \mathbf{x}_t)' \boldsymbol{\Omega}^{-1} (\mathbf{y}_t - \mathbf{B}' \mathbf{x}_t) = \text{tr } \boldsymbol{\Omega}^{-1} (\mathbf{Y} - \mathbf{XB})' (\mathbf{Y} - \mathbf{XB})$ .
2. Using the relationships  $\mathbf{L}^2 = \mathbf{L}$ ,  $\mathbf{P}^2 = \mathbf{P}$  and  $\mathbf{LP} = \mathbf{0}$  show that for  $\mathbf{L} = \mathbf{G}_1^* (\mathbf{G}_1^* \mathbf{X}' \mathbf{X} \mathbf{G}_1^*)^{-1} \mathbf{G}_1^* \mathbf{X}' \mathbf{X}$ ,  $\mathbf{P}$  takes the form

$$\mathbf{P} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{D}_1' [\mathbf{D}_1 (\mathbf{X}' \mathbf{X})^{-1} \mathbf{D}_1']^{-1} \mathbf{D}_1,$$

where

$$\mathbf{D} = \begin{pmatrix} \mathbf{D}_1 \\ \mathbf{D}_1^* \end{pmatrix}, \quad (\mathbf{G}_1, \mathbf{G}_1^*) = -\mathbf{D}^{-1}$$

(see Section 24.3).

3. Verify the formulae (58), (59) and (62), (63).
4. Consider the following system of equations ( $m=2$ )

$$y_{1t} = \beta_{11} x_{1t} + \beta_{21} x_{2t} + \beta_{31} x_{3t} + \beta_{41} x_{4t} + u_{1t},$$

$$y_{2t} = \beta_{12} x_{1t} + \beta_{22} x_{2t} + \beta_{32} x_{3t} + \beta_{42} x_{4t} + u_{2t}.$$

Discuss the estimation of this system in the following three cases:

- (i) no a priori restrictions;
- (ii)  $\beta_{31} = 0, \beta_{42} = 0$ ;
- (iii)  $\beta_{31} = \beta_{32}$ .

5. Derive the  $F$ -type test for  $H_0: \mathbf{DB} - \mathbf{C} = \mathbf{0}$  against  $H_1: \mathbf{DB} - \mathbf{C} \neq \mathbf{0}$ .
6. 'In testing for departures from the assumption of independence we can use either of the following auxiliary equations:

$$\begin{aligned}\hat{\mathbf{u}}_t &= \sum_{i=1}^l \mathbf{A}_i \mathbf{y}_{t-i} + \sum_{i=1}^l \mathbf{B}'_i \mathbf{x}_{t-i} + \mathbf{v}_t, \\ \hat{u}_t &= (\mathbf{B}_0 - \hat{\mathbf{B}}_0)' \mathbf{x}_t + \sum_{i=1}^l \mathbf{A}_i \mathbf{y}_{t-i} + \sum_{i=1}^l \mathbf{B}'_i \mathbf{x}_{t-i} + \mathbf{v}_t,\end{aligned}$$

because  $\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}$  and thus both cases should give the same answer.'

Discuss.

7. Construct a  $1 - \alpha$  prediction region for  $\mathbf{y}_{T+\tau}$ .

#### Additional references

Anderson (1984); Kendall and Stuart (1968); Mardia *et al.* (1979); Morrison (1976); Srivastava and Khatri (1979).

## CHAPTER 25

---

### The simultaneous equations model

---

#### 25.1 Introduction

The simultaneous equations model was first proposed by Haavelmo (1943) as a way to provide a statistical framework in the context of which a theoretical model which comprises a system of simultaneous interdependent equations can be analysed. His suggestion was to provide the basis of a research programme undertaken by the Cowles Foundation during the late 1940s and early 50s. Their results, published in two monographs (Koopmans (1950), Hood and Koopmans (1953)), dominated the econometric research agenda for the next three decades.

In order to motivate the simultaneous equations formulation let us consider the following theoretical model:

$$m_t = \alpha_{11} + \alpha_{21}i_t + \alpha_{31}p_t + \alpha_{41}y_t \quad (25.1)$$

$$i_t = \alpha_{12} + \alpha_{22}m_t + \alpha_{32}p_t + \alpha_{42}g_t \quad (25.2)$$

where  $m_t$ ,  $i_t$ ,  $p_t$ ,  $y_t$ ,  $g_t$  refer to (the  $\log_e$  of) the theoretical variables, money, interest rate, price level, income and government budget deficit, respectively. For expositional purposes let us assume that there exist observed data series which correspond one-to-one to these theoretical variables. That is, (1)–(2) is also an estimable model (see Chapter 1). The question which naturally arises is to what extent the estimable model (1)–(2) can be statistically analysed in the context of the multivariate linear regression model discussed in Chapter 24. A moment's reflection suggests that the presence of the so-called *endogenous* variables  $i_t$  and  $m_t$  on the RHS of (1) and (2), respectively, raises new problems. The alternative

formulation:

$$m_t = \beta_{11} + \beta_{21}p_t + \beta_{31}y_t + \beta_{41}g_t, \quad (25.3)$$

$$i_t = \beta_{12} + \beta_{22}p_t + \beta_{32}y_t + \beta_{42}g_t, \quad (25.4)$$

can be analysed in the context of the multivariate linear regression model because the  $\beta_{ij}$ s,  $i=1,2$ ,  $i=1,2,3,4$ ,  $j=1,2$ , are directly related to the statistical parameters  $\mathbf{B}$  of the multivariate linear regression model (see Chapter 24).

This suggests that if we could find a way to relate the two parametrisations in (1)–(2) and (3)–(4) we could interpret the theoretical parameters  $\alpha_{ij}$ s as a reparametrisation of the  $\beta_{ij}$ s. In what follows it is argued that this is possible as long as the  $\alpha_{ij}$ s can be uniquely defined in terms of the  $\beta_{ij}$ s. The way this is achieved is by ‘reparametrising’ (3)–(4) first into the formulation:

$$m_t = a_{11} + a_{21}i_t + a_{31}y_t + a_{41}p_t + a_{51}g_t \quad (25.5)$$

$$i_t = a_{12} + a_{22}m_t + a_{32}y_t + a_{42}p_t + a_{52}g_t \quad (25.6)$$

and then derive the  $\alpha_{ij}$ s by imposing restrictions on the  $a_{ij}$ s such as  $a_{51}=0$ ,  $a_{32}=0$ . In view of this it is important to emphasise at the outset that the simultaneous equations formulation should be interpreted as a theoretical parametrisation of particular interest in econometric modelling because it ‘models’ the co-determination of behaviour, and not as a statistical model.

In Section 25.2 the relationship between the multivariate linear regression and the simultaneous equation formulation is explicitly derived in an attempt to introduce the problem of reparametrisation and overparametrisation. The latter problem raises the issue of identification which is considered in Section 25.3. The specification of the simultaneous equation model as an extension of the multivariate linear regression model where the statistical parameters of interest do not coincide with the theoretical (structural) parameters of interest is discussed in Section 25.4. The estimation of the theoretical parameters of interest by the method of maximum likelihood is considered in Section 25.5. Section 25.6 considers two least-squares estimators in an attempt to enhance our understanding of the problem of simultaneity and its implications. These estimators are related to the instrumental variables method in Section 25.7. In Section 25.8 we consider misspecification testing at three different but interrelated levels. Section 25.9 discusses the issues of specification testing and model selection. In Section 25.10 the problem of prediction is briefly discussed.

It is important to note at the outset that even though the dynamic

simultaneous equations model is not explicitly considered the results which follow can be extended to the more general case in the same way as in the context of the multivariate linear regression model (see Chapter 24). In particular, if we interpret  $\mathbf{X}_t$  as including all the *predetermined variables*, i.e.

$$\mathbf{X}_t = (\mathbf{y}'_{t-1}, \dots, \mathbf{y}'_{t-l}, \mathbf{x}'_t, \mathbf{x}'_{t-1}, \dots, \mathbf{x}'_{t-l})',$$

the following results on estimation, misspecification and specification testing as well as prediction go through asymptotically (see Hendry and Richard (1983)).

## 25.2 The multivariate linear regression and simultaneous equations models

The multivariate linear regression model discussed in Chapter 24 was based on the statistical GM:

$$\mathbf{y}_t = \mathbf{B}'\mathbf{x}_t + \mathbf{u}_t, \quad t \in \mathbb{T}, \quad (25.7)$$

with  $\boldsymbol{\theta} \equiv (\mathbf{B}, \boldsymbol{\Omega})$ ,  $\mathbf{B} = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ ,  $\boldsymbol{\Omega} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ , the statistical parameters of interest. As argued in Section 25.1, for certain estimable models in econometric modelling the theoretical parameters of interest do not coincide with  $\boldsymbol{\theta}$  and we need to reparametrise (7) so as to accommodate such models.

In order to motivate the reparametrisation needed to accommodate estimable models such as (1)–(2) let us separate  $y_{1t}$  from the other endogenous variables  $\mathbf{y}_t^{(1)}$  and decompose  $\mathbf{B}$  and  $\boldsymbol{\Omega}$  conformably in an obvious notation:

$$\left( \begin{pmatrix} y_{1t} \\ \mathbf{y}_t \end{pmatrix} \middle| \mathbf{X}_t = \mathbf{x}_t \right) \sim N \left( \begin{pmatrix} \boldsymbol{\beta}'_1 \mathbf{x}_t \\ \mathbf{B}'_{(1)} \mathbf{x}_t \end{pmatrix} \begin{pmatrix} \boldsymbol{\omega}_{11} & \boldsymbol{\omega}_{12} \\ \boldsymbol{\omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix} \right). \quad (25.8)$$

A natural way to get the endogenous variables  $\mathbf{y}_t^{(1)}$  into the systematic component, purporting to explain  $y_{1t}$ , is to condition on the  $\sigma$ -field generated by  $\mathbf{y}_t^{(1)}$ , say  $\sigma(\mathbf{y}_t^{(1)})$ , i.e. for

$$\mathcal{F}_t^{(1)} \equiv (\sigma(\mathbf{y}_t^{(1)}), \mathbf{X}_t = \mathbf{x}_t) \quad (25.9)$$

$$\mu_{1t} \equiv E(y_{1t} / \mathcal{F}_t^{(1)}) = \boldsymbol{\Gamma}_1^0 \mathbf{y}_t^{(1)} + \boldsymbol{\Delta}'_1 \mathbf{x}_t, \quad (25.10)$$

where  $\boldsymbol{\Gamma}_1^0 = \boldsymbol{\Omega}_{22}^{-1} \boldsymbol{\omega}_{21}$  and  $\boldsymbol{\Delta}_1 = \boldsymbol{\beta}_1 - \mathbf{B}_{(1)} \boldsymbol{\Omega}_{22}^{-1} \boldsymbol{\omega}_{21}$ . The systematic component defined by (10) can be used to construct the statistical GM

$$y_{1t} = \boldsymbol{\Gamma}_1^0 \mathbf{y}_t^{(1)} + \boldsymbol{\Delta}'_1 \mathbf{x}_t + \varepsilon_{1t}, \quad (25.11)$$

where

$$\varepsilon_{1t} = y_{1t} - E(y_{1t}/\mathcal{F}_t^{(1)}).$$

Note:

$$(i) \quad E(\varepsilon_{1t}) = E[E(\varepsilon_{1t}/\mathcal{F}_t^{(1)})] = 0,$$

$$(ii) \quad E(\varepsilon_{1t}\varepsilon_{1s}) = E[E(\varepsilon_{1t}\varepsilon_{1s}/\mathcal{F}_t^{(1)})]$$

$$= \begin{cases} v_{11} & t=s \\ 0 & t \neq s, \quad t, s \in \mathbb{T}, \end{cases} \quad v_{11} = \omega_{11} - \boldsymbol{\omega}_{12}\boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\omega}_{21}.$$

$$(iii) \quad E(\mu_{1t}\varepsilon_{1t}) = E[E(\mu_{1t}\varepsilon_{1t}/\mathcal{F}_t^{(1)})] = 0, \quad t \in \mathbb{T}$$

(see Chapter 7 for the properties of the conditional expectations needed to prove (i)–(iii)).

Looking at (11) we can see that such a statistical GM can easily accommodate any estimable equation of the form:

$$m_t = \alpha_{11} + \alpha_{12}i_t + \alpha_{13}p_t + \alpha_{14}y_t \quad (25.12)$$

or

$$i_t = \alpha_{21} + \alpha_{22}m_t + \alpha_{23}p_t + \alpha_{24}g_t$$

separately. The thing to note about (11) is that its parameters we call structural parameters  $(\alpha_1, v_{11})$  where

$$\alpha_1 = (\boldsymbol{\Gamma}_1^0 \boldsymbol{\Delta}_1')', \quad (25.13)$$

$$v_{11} = E(\varepsilon_{1t}^2), \quad (25.14)$$

are simple functions of  $\mathbf{B}$  and  $\boldsymbol{\Omega}$ . That is, they constitute an alternative parametrisation of  $\theta$ , in  $D(\mathbf{y}_t/\mathbf{X}_t; \theta)$ , based on the decomposition

$$D(\mathbf{y}_t/\mathbf{X}_t; \theta) = D(y_{1t}/y_t^{(1)}, \mathbf{X}_t; \boldsymbol{\eta}_1) \cdot D(y_t^{(1)}/\mathbf{X}_t; \boldsymbol{\eta}_{(1)}). \quad (25.15)$$

Moreover, the normality assumption ensures that  $y_t^{(1)}$  is indeed weakly exogenous with respect to the parameters  $(\alpha_1, v_{11})$ . The statistical GM (11) is a hybrid of the linear and stochastic linear regression models. These comments suggest that the so-called simultaneity problem has nothing to do with the presence of stochastic variables among the regressors.

The decomposition (15) holds for any one endogenous variable  $y_{it}$

$$D(\mathbf{y}_t/\mathbf{X}_t; \theta) = D(y_{it}/y_t^{(i)}, \mathbf{X}_t; \boldsymbol{\eta}_i) \cdot D(y_t^{(i)}/\mathbf{X}_t; \boldsymbol{\eta}_{(i)}), \quad (25.16)$$

giving rise to a statistical GM:

$$y_{it} = \boldsymbol{\Gamma}_i^0 \mathbf{y}_t^{(i)} + \boldsymbol{\Delta}_i' \mathbf{x}_t + \varepsilon_{it}. \quad (25.17)$$

The problem, however, is that no decomposition of  $D(\mathbf{y}_t/\mathbf{X}_t; \boldsymbol{\theta})$  exists which can sustain all  $m$  equations  $i = 1, 2, \dots, m$  in (17). That is, the system

$$\boldsymbol{\Gamma}' \mathbf{y}_t + \boldsymbol{\Delta}' \mathbf{x}_t + \boldsymbol{\varepsilon}_t = \mathbf{0}, \quad (25.18)$$

where

$$\boldsymbol{\Gamma}' \equiv (\boldsymbol{\Gamma}'_1, \boldsymbol{\Gamma}'_2, \dots, \boldsymbol{\Gamma}'_m), \quad \boldsymbol{\Delta}' \equiv (\boldsymbol{\Delta}'_1, \boldsymbol{\Delta}'_2, \dots, \boldsymbol{\Delta}'_m),$$

$$\boldsymbol{\varepsilon}_t \equiv (\varepsilon_{1t}, \varepsilon_{2t}, \dots, \varepsilon_{mt})'$$

( $\boldsymbol{\Gamma}_i$  is essentially  $\boldsymbol{\Gamma}_i^0$  with  $-1$  added as its  $i$ th element), is not a well-defined statistical GM because it constitutes an over-reparametrisation of (7). For one equation, say the first, the reparametrisation

$$\boldsymbol{\eta}_1 \equiv (\boldsymbol{\Gamma}_1^0, \boldsymbol{\Delta}_1, v_{11}) \quad \text{and} \quad \boldsymbol{\eta}_{(1)} \equiv (\mathbf{B}_{(1)}, \boldsymbol{\Omega}_{22})$$

is well defined but

$$\boldsymbol{\eta}_1 \times \boldsymbol{\eta}_2 \times \cdots \times \boldsymbol{\eta}_m \equiv \prod_{i=1}^m \boldsymbol{\eta}_i$$

is not.

A particular case where the cartesian product  $\prod_{i=1}^m \boldsymbol{\eta}_i$  is a proper reparametrisation of  $\boldsymbol{\theta}$  is when there exists a natural ordering of the  $y_{it}$ s such that  $y_{jt}$  depends only on the  $y_{it}$ s up to  $j-1$ , i.e.

$$E(y_{jt}/\mathcal{F}_t^{(j)}) = E(y_{jt}/\sigma(y_{it}, i=1, 2, \dots, j-1), \mathbf{X}_t = \mathbf{x}_t), \quad j = 1, 2, \dots, m. \quad (25.19)$$

In this case the distribution  $D(\mathbf{y}_t/\mathbf{X}_t; \boldsymbol{\theta})$  can be decomposed in the form

$$D(\mathbf{y}_t/\mathbf{X}_t; \boldsymbol{\theta}) = \prod_{i=1}^m D(y_{it}/y_{1t}, y_{2t}, \dots, y_{i-1t}, \mathbf{X}_t; \boldsymbol{\eta}_i). \quad (25.20)$$

This decomposition gives rise to a lower triangular matrix  $\boldsymbol{\Gamma}$  and the system (18) is then a well-defined statistical GM known as a *recursive system* (see below).

In the non-recursive case the parametrisation given in (18) is defined in terms of  $n = m(m-1) + mk + \frac{1}{2}(m+1)$  unknown parameters  $\boldsymbol{\eta} \equiv (\boldsymbol{\Gamma}, \boldsymbol{\Delta}, \mathbf{V})$  where  $\mathbf{V} \equiv E(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t')$ , and there are only  $mk + \frac{1}{2}m(m+1)$  well-defined statistical parameters in  $\boldsymbol{\theta}$ ; a shortfall of  $m(m-1)$  parameters. Given that  $\boldsymbol{\eta}$  constitutes a reparametrisation of  $\boldsymbol{\theta}$  there can only be  $mk + \frac{1}{2}m(m+1)$  well-defined parameters in  $\boldsymbol{\eta}$ . In order to see the relationship between  $\boldsymbol{\theta}$  and  $\boldsymbol{\eta}$  let us premultiply (18) by  $(\boldsymbol{\Gamma}')^{-1}$  (assumed to be non-singular):

$$\mathbf{y}_t + (\boldsymbol{\Gamma}')^{-1} \boldsymbol{\Delta}' \mathbf{x}_t + (\boldsymbol{\Gamma}')^{-1} \boldsymbol{\varepsilon}_t = \mathbf{0}. \quad (25.21)$$

If we compare (21) with (7) we deduce that  $\boldsymbol{\theta}$  and  $\boldsymbol{\eta}$  are related via

$$\mathbf{B}\boldsymbol{\Gamma} + \boldsymbol{\Delta} = \mathbf{0}, \quad \mathbf{V} = \boldsymbol{\Gamma}' \boldsymbol{\Omega} \boldsymbol{\Gamma}. \quad (25.22)$$

The first thing to note about this system of equations is that they are not a priori restrictions of the form considered in Chapter 24 where  $\Gamma$  and  $\Delta$  are assumed known. The system of equations in (22) 'defines' the parameters  $\eta$  in terms of  $\theta$ . As it stands, however, the system allows an infinity of solutions for  $\eta$  given  $\theta$ .

The parameters  $\theta$  and  $\eta$  will be referred to as *statistical* and *structural parameters*, respectively. The system of equations (22) enables us to determine only a subset  $\eta_1$  of  $\eta$  ( $\eta \equiv (\eta_1 : \eta_2)$ ) for any given set of well-defined statistical parameters  $\theta$ . That is, (22) can be solved for  $mk + \frac{1}{2}m(m+1)$  structural parameters  $\eta$ , in the form

$$\eta_1 = \mathbf{G}(\theta, \eta_2). \quad (25.23)$$

Hence, we need additional information to determine  $\eta_2$  elsewhere.

Note that  $\eta_2$  is a  $m(m-1) \times 1$  vector of structural parameters. Without any additional information the structural parameters  $\eta$  are said to be *not identified*. The problem of identifying the structural parameters of interest using additional information will be considered formally in the next section. In the meantime it is sufficient to note that, if we supplement the system (22) with  $m(m-1)$  additional independent restrictions, then a unique solution (implicit or explicit),

$$\xi = \mathbf{G}^*(\theta), \quad (25.24)$$

exists for the structural parameters of interest  $\xi = \mathbf{L}(\eta)$ .

There are several things worth summarising in the above discussion. Firstly, given that the structural formulation (18) is a reparametrisation of the statistical GM for the multivariate linear regression model, the structural parameters of interest  $\xi$  (when identified) are no more well defined than the statistical parameters  $\theta$ . This suggests that before any question about  $\xi$  can be asked we need to ensure that  $\theta$  is well defined (no misspecification test has shown any departures from the assumptions underlying the multivariate linear regression model). Secondly, (18) does not constitute a well-defined statistical GM unless  $\Gamma$  is a triangular matrix; the system of equations is recursive. This is because although each equation separately is well defined, the system as a whole is not because of overparametrisation. In the case of a recursive system  $\Gamma$  is *lower triangular* and  $\mathbf{V}$  is *diagonal* with

$$v_{ll} = \frac{\det(\Sigma_l)}{\det(\Sigma_{l-1})}, \quad (25.25)$$

$\Sigma_l$  being the  $l$ th leading diagonal matrix. This implies that  $\mathbf{X}_{i=1}^m \eta_i$

constitutes a proper reparametrisation of  $\theta$  with  $mk + \frac{1}{2}m(m+1)$  structural parameters in  $(\Gamma, \Delta, V)$ . Using the notation  $y_{i-1,t}^0 \equiv (y_{1t}, y_{2t}, \dots, y_{0-1,t})'$  we can express the *recursive system* in the form

$$y_{it} = \gamma_i^0 y_{i-1,t}^0 + \delta_i \mathbf{x}_t + \varepsilon_{it}, \quad i = 1, 2, \dots, m, \quad t \in \mathbb{T}.$$

We can estimate the structural parameters  $\eta_i \equiv (\gamma_i^0, \delta_i, v_{ii})$  by

$$\begin{pmatrix} \hat{\gamma}_i^0 \\ \hat{\delta}_i \end{pmatrix} = \begin{pmatrix} \mathbf{Y}_{i-1}^{0'} \mathbf{Y}_{i-1}^0 & \mathbf{Y}_{i-1}^{0'} \mathbf{X} \\ \mathbf{X}' \mathbf{Y}_{i-1}^0 & \mathbf{X}' \mathbf{X} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{Y}_{i-1}^{0'} \mathbf{y}_i \\ \mathbf{X}' \mathbf{y}_i \end{pmatrix} \quad (25.27)$$

and

$$\hat{v}_{ii} = \frac{1}{T} (\mathbf{y}_i - \mathbf{Y}_{i-1}^0 \hat{\gamma}_i^0 - \mathbf{X} \hat{\delta}_i)' (\mathbf{y}_i - \mathbf{Y}_{i-1}^0 \hat{\gamma}_i^0 - \mathbf{X} \hat{\delta}_i), \quad i = 1, 2, \dots, m. \quad (25.28)$$

It can be verified that these are indeed the MLE's of  $\eta_i$ .

### 25.3 Identification using linear homogeneous restrictions

As argued in the previous section, the reparametrisation of the statistical GM,

$$\mathbf{y}_t = \mathbf{B}' \mathbf{x}_t + \mathbf{u}_t, \quad (25.29)$$

in the form of the structural system,

$$\Gamma' \mathbf{y}_t + \Delta' \mathbf{x}_t + \varepsilon_t = 0, \quad (25.30)$$

is not well defined because there are only  $mk + \frac{1}{2}m(m+1)$  statistical parameters  $\theta$  defining (29) and  $m(m-1) + mk + \frac{1}{2}m(m+1)$  structural parameters  $\eta$  defining (30). The two sets of parameters  $\theta$  and  $\eta$  are related via

$$\mathbf{B}\Gamma + \Delta = \mathbf{0}, \quad \mathbf{V} = \Gamma' \Omega \Gamma. \quad (25.31)$$

No unique solution for  $\eta$  exists, corresponding to any set of well-defined statistical parameters  $\theta$ , unless the system of equations (31) is supplemented with some additional a priori restrictions.

In order to simplify the discussion of the identification problem let us assume that the system  $\mathbf{V} = \Gamma' \Omega \Gamma$  determines  $\mathbf{V}$  for given  $\Gamma$  and  $\Omega$  and concentrate on the determination of  $\Gamma$  and  $\Delta$ . The system of equations

$$\begin{aligned} \mathbf{B}\Gamma + \Delta &= \mathbf{0} \text{ written in the form} \\ \Pi \Delta &= \mathbf{0}, \end{aligned} \quad (25.32)$$

where

$$\Pi \equiv (\mathbf{B}, \mathbf{I})$$

and

$$\mathbf{A} \equiv \begin{pmatrix} \Gamma \\ \Delta \end{pmatrix}$$

is linear in  $\mathbf{A}$  for given  $\mathbf{B}$ . In Kronecker product notation (see Appendix 24.2), (32) can be expressed as

$$(\mathbf{I}_m \otimes \boldsymbol{\Pi}) \text{vec}(\mathbf{A}) = \mathbf{0} \quad \text{or} \quad \boldsymbol{\Pi}_* \boldsymbol{\alpha} = \mathbf{0}, \quad (25.33)$$

where  $\boldsymbol{\alpha} \equiv (\boldsymbol{\alpha}'_1, \boldsymbol{\alpha}'_2, \dots, \boldsymbol{\alpha}'_m)'$  are the columns of  $\mathbf{A}$  turned into a  $m(m+k-1) \times 1$  vector. The system of equations (33) cannot be solved uniquely for  $\boldsymbol{\alpha}$  because  $\text{rank}(\boldsymbol{\Pi}_*) = mk < m(m+k-1)$  for  $m > 1$ . For a unique solution we need to supplement the system with a priori restrictions such as the *linear homogeneous restrictions*

$$\boldsymbol{\Phi} \boldsymbol{\alpha} = \mathbf{0}, \quad (25.34)$$

ensuring that  $\text{rank}(\boldsymbol{\Phi}) \geq m(m-1)$ . If this is the case then the system

$$\begin{pmatrix} \boldsymbol{\Pi}_* \\ \boldsymbol{\Phi} \end{pmatrix} \boldsymbol{\alpha} = \mathbf{0} \quad (25.35)$$

has a *unique solution for  $\boldsymbol{\alpha}$* .

### *Definition 1*

The structural parameters  $\boldsymbol{\alpha}$  are said to be **identified** if and only if

$$\text{rank} \begin{pmatrix} \boldsymbol{\Pi}_* \\ \boldsymbol{\Phi} \end{pmatrix} = m(m+k-1). \quad (25.36)$$

Using the result that

$$\text{rank} \begin{pmatrix} \boldsymbol{\Pi}_* \\ \boldsymbol{\Phi} \end{pmatrix} = \text{rank}(\boldsymbol{\Phi} \mathbf{A}^*) + mk \quad (25.37)$$

(see Schmidt (1976)), we can state condition (36) in the form

$$\text{rank}(\boldsymbol{\Phi} \mathbf{A}^*) = m(m-1). \quad (25.38)$$

Note that  $\mathbf{A}^*$  denotes the *restricted* structural coefficient parameters. Hence, the structural formulation (30) is said to be identified if and only if we can supplement it with at least  $m(m-1)$  additional restrictions. More general restrictions as well as covariance restrictions are beyond the scope of the present book (see Hsiao (1983) *inter alia*).

The identification problem in econometrics is usually tackled not in terms of the system (30) as a whole but equation by equation using a particular form of linear homogeneous restrictions, the so-called *exclusion* (or zero-one) *restrictions*. In order to motivate the problem let us consider the two equation estimable model introduced in Section 25.1 above. The

*unrestricted structural form* (30) (compare with (5)–(6)) of this model is

$$m_t = \gamma_{21}i_t + \delta_{11} + \delta_{21}y_t + \delta_{31}p_t + \delta_{41}g_t + \varepsilon_{1t}, \quad (25.39)$$

$$i_t = \gamma_{12}m_t + \delta_{12} + \delta_{22}y_t + \delta_{32}p_t + \delta_{42}g_t + \varepsilon_{2t}. \quad (25.40)$$

As can be seen, the two equations are indistinguishable given that they differ only by a normalisation condition. The overparametrisation arises because the statistical GM underlying (39) and (40) is in effect

$$m_t = \beta_{11} + \beta_{21}y_t + \beta_{31}p_t + \beta_{41}g_t + u_{1t}, \quad (25.41)$$

$$i_t = \beta_{12} + \beta_{22}y_t + \beta_{32}p_t + \beta_{42}g_t + u_{2t}, \quad (25.42)$$

and the two parametrisations  $\mathbf{B}$  and  $(\Gamma, \Delta)$  are related via

$$\begin{pmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \\ \beta_{31} & \beta_{32} \\ \beta_{41} & \beta_{42} \end{pmatrix} \begin{pmatrix} -1 & \gamma_{12} \\ \gamma_{21} & -1 \end{pmatrix} + \begin{pmatrix} \delta_{11} & \delta_{12} \\ \delta_{21} & \delta_{22} \\ \delta_{31} & \delta_{32} \\ \delta_{41} & \delta_{42} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad (25.43)$$

which cannot be solved for  $\Gamma$  and  $\Delta$  uniquely, given that there are only eight  $\beta_{ij}$ s and ten  $\gamma_{ij}$ s and  $\delta_{ij}$ s.

A natural way to ‘solve’ the identification problem is to impose exclusion restrictions on (39) and (40) such as  $\delta_{41}=0$ ,  $\delta_{22}=0$ . These restrictions enable us to distinguish between the money and interest rate equations. Equivalently, the restrictions enable us to get a unique solution for  $(\gamma_{12}, \gamma_{21}, \delta_{11}, \delta_{12}, \delta_{13}, \delta_{14}, \delta_{21}, \delta_{31})$  given  $(\beta_{ij}, i=1, \dots, 4, j=1, 2)$ .

The *exclusion restrictions* on the  $i$ th structural equation can be expressed in the form

$$\Phi_i \alpha_i = 0, \quad (25.44)$$

where  $\Phi_i$  is a ‘selector’ matrix of zeros and ones and  $\alpha_i$  is the  $i$ th column of  $\mathbf{A}$ . In the above example the selector matrices are

$$\Phi_1 = (0, 0, 0, 0, 0, 1) \quad \text{and} \quad \Phi_2 = (0, 0, 0, 1, 0, 0). \quad (25.45)$$

The identification condition (36) for each equation separately, known as the *rank condition*, takes the form

$$\text{rank}\left(\frac{\Pi^*}{\Phi_i}\right) = m + k - 1, \quad i = 1, 2, \dots, m \quad (25.46)$$

or

$$\text{rank}(\Phi_i \mathbf{A}^*) = m - 1, \quad i = 1, 2, \dots, m. \quad (25.47)$$

In the special case of exclusion restrictions the order condition can be made even easier to apply. Let us consider the first equation of the system

(31):

$$\mathbf{B}\Gamma_{.1} + \Delta_{.1} = \mathbf{0} \quad (25.48)$$

and impose  $(m - m_1) + (k - k_1)$  exclusion restrictions; omit  $(m - m_1)$  endogenous and  $(k - k_1)$  exogenous variables from the first equation. In this case we do not need to define  $\Phi_1$  explicitly and consider (46) because we can substitute the restrictions directly into (48). Re-arranging the variables so as to have the excluded ones last, the *restricted structural parameters*  $\Gamma_1^*$  and  $\Delta_1^*$  are

$$\Gamma_1^* = \begin{pmatrix} -1 \\ \gamma_1 \\ \mathbf{0} \end{pmatrix}, \quad \Delta_1^* = \begin{pmatrix} \delta_1 \\ \mathbf{0} \end{pmatrix} \quad \text{where } \gamma_1: (m_1 - 1) \times 1, \quad \delta_1: k_1 \times 1. \quad (25.49)$$

Partitioning  $\mathbf{B}$  conformably the system (48) under the exclusion restrictions becomes:

$$\begin{pmatrix} \beta_{11} & \mathbf{B}_{12} & \mathbf{B}_{13} \\ \beta_{21} & \mathbf{B}_{22} & \mathbf{B}_{23} \end{pmatrix} \begin{pmatrix} -1 \\ \gamma_1 \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} \delta_1 \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix} \quad (25.50)$$

or

$$-\beta_{11} + \mathbf{B}_{12}\gamma_1 + \delta_1 = \mathbf{0}, \quad (25.51)$$

$$-\beta_{21} + \mathbf{B}_{22}\gamma_1 = \mathbf{0}, \quad (25.52)$$

where

$$\beta_{11}: k_1 \times 1, \quad \mathbf{B}_{12}: k_1 \times (m_1 - 1), \quad \mathbf{B}_{13}: k_1 \times (m - m_1),$$

$$\beta_{21}: (k - k_1) \times 1, \quad \mathbf{B}_{22}: (k - k_1) \times (m_1 - 1), \quad \mathbf{B}_{23}: (k - k_1) \times (m - m_1).$$

Determining  $\delta_1$  from (51) presents no problems if  $\gamma_1$  is uniquely determined in (52). For this to be the case we need the condition that

$$\text{rank}(\mathbf{B}_{22}) = m_1 - 1. \quad (25.53)$$

In view of the result that the  $\text{rank}(\mathbf{B}_{22}) = \min(k - k_1, m_1 - 1)$  we can deduce that a *necessary condition* for identification under exclusion restrictions is that

$$(k - k_1) \geq m_1 - 1. \quad (25.54)$$

That is, the number of excluded exogenous variables is greater or equal to the number of included endogenous variables minus one. This condition in terms of the selector matrix  $\Phi_i$  takes the form

$$\text{rank}(\Phi_{1j}) = m_1 - 1. \quad (25.55)$$

This is known as the *order condition* for identification which is necessary but not sufficient. This can be easily seen in the example (39), (40), above when the exclusion restrictions are  $\delta_{41}=0$ ,  $\delta_{14}=0$ . The selector matrices for the two equations are  $\Phi_1=(0, 0, 0, 0, 0, 1)$ ,  $\Phi_2=(0, 0, 0, 0, 0, 1)$ . Clearly  $\text{rank}(\Phi_1)=\text{rank}(\Phi_2)=1$  and thus the order condition is satisfied but the rank condition (47) fails because

$$\text{rank}(\Phi_i \mathbf{A}^*) = \text{rank}(0, 0) = 0, \quad i = 1, 2. \quad (25.56)$$

This is because when  $g_t$  is excluded from both equations the restriction is 'phoney'. Such a situation arises when:

- (i) all equations satisfy the same restriction; and
- (ii) some other equation satisfies all the restrictions of the  $i$ th equation (see example below).

Let us introduce some nomenclature related to the identification of particular equations in the system (30).

### *Definition 2*

*A particular equation of the structural form (30), say the  $i$ th, is said to be:*

- (i) **under identified if**

$$\text{rank}(\Phi_i \mathbf{A}^*) < m - 1; \quad (25.57)$$

- (ii) **exactly identified if**

$$\text{rank}(\Phi_i \mathbf{A}^*) = \text{rank}(\Phi_i) = m - 1; \quad (25.58)$$

- (iii) **over identified if**

$$\text{rank}(\Phi_i \mathbf{A}^*) = m - 1, \quad \text{rank}(\Phi_i) > m - 1. \quad (25.59)$$

*The system (30) is said to be identified if every equation is identified.*

### *Example*

Consider the following structural form with the exclusion restrictions imposed:

$$-y_{1t} + \delta_{11}x_{1t} = \varepsilon_{1t}, \quad (25.60)$$

$$\gamma_{12}y_{1t} - y_{2t} + \delta_{12}x_{1t} = \varepsilon_{2t}, \quad (25.61)$$

$$\gamma_{13}y_{1t} - y_{3t} + \delta_{13}x_{1t} + \delta_{23}x_{2t} = \varepsilon_{3t}, \quad (25.62)$$

$$\Phi_1 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad \Phi_2 = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\Phi_3 = (0, 1, 0, 0, 0)$$

$$\Phi_1 A^* = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & \delta_{23} \end{pmatrix}, \quad \Phi_2 A^* = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & \delta_{23} \end{pmatrix},$$

$$\Phi_3 A^* = (0, 0, -1).$$

The first equation is overidentified since  $\text{rank}(\Phi_1 A^*)=2$  but  $\text{rank}(\Phi_1)=3$ . The second equation is underidentified because  $\text{rank}(\Phi_2 A^*)=1$  even though  $\text{rank}(\Phi_2)=2$  (the order condition holds). This is because the first equation satisfied all the restrictions of the second equation rendering these conditions ‘phoney’. The third equation is underidentified as well, because  $\text{rank}(\Phi_3 A^*)=1<2$ .

It is important to note that when certain equations of the structural form (30) are overidentified then not only the structural parameters of interest are uniquely defined by (31) but the statistical parameters  $\theta$  are themselves restricted. That is, overidentifying restrictions imply that  $\theta \in \Theta_1$  where  $\Theta_1$  is a subset of the parameter space  $\Theta$ . An important implication of this is that the identifying restrictions cannot be tested but the overidentifying ones can (see Section 25.9).

The above discussion of the identification problem depends crucially on the assumption that the statistical parameters of interest  $\theta \equiv (\mathbf{B}, \Omega)$  are well defined; the assumptions [1]–[8] underlying the multivariate linear regression model are valid. However, in the case where some assumption is invalid and the parametrisation changes we need to reconsider the identification of the system. For example, in the case where the independence assumption is invalid and the statistical GM takes the form

$$\mathbf{y}_t = \mathbf{B}_0 \mathbf{x}_t + \sum_{i=1}^l \mathbf{A}'_i \mathbf{y}_{t-i} + \sum_{i=1}^l \mathbf{B}'_i \mathbf{x}_{t-i} + \mathbf{u}_t, \quad t > l \quad (25.63)$$

the identification problem has to be related to the statistical parameters  $\theta^* \equiv (\mathbf{A}_1, \dots, \mathbf{A}_l, \mathbf{B}_0, \mathbf{B}_1, \dots, \mathbf{B}_l, \Omega_0)$ , not  $\theta$ . Hence, the identification of the system is not just a matter of a priori information only.

## 25.4 Specification

The simultaneous equation formulation as a statistical model is viewed as a

reparametrisation of the multivariate linear regression model where the theoretical (structural) parameters of interests  $\xi$  do not coincide with the statistical parameters of interest  $\theta$ . In particular the statistical model is specified as follows:

**(I) Statistical GM**

$$\mathbf{y}_t = \mathbf{B}\mathbf{x}_t + \mathbf{u}_t, \quad t \in \mathbb{T}. \quad (25.64)$$

- [1]  $\mu_t = E(\mathbf{y}_t | \mathbf{X}_t = \mathbf{x}_t)$  and  $\mathbf{u}_t = \mathbf{y}_t - E(\mathbf{y}_t | \mathbf{X}_t = \mathbf{x}_t)$  are the systematic and non-systematic components, respectively.
- [2]
  - (i)  $\theta \equiv (\mathbf{B}, \Omega)$  are the statistical parameters of interest where  $\mathbf{B} = \Sigma_{22}^{-1} \Sigma_{21}$ ,  $\Omega = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$ ,  $\theta \in \Theta \equiv \mathbb{R}^{mk} \times \mathbf{C}_m$ ;
  - (ii)  $\xi = \mathbf{L}(\Gamma, \Delta, \mathbf{V})$  are the theoretical parameters of interest where  $\Gamma = \mathbf{H}_1(\theta)$ ,  $\Delta = \mathbf{H}_2(\theta)$ ,  $\mathbf{V} = \mathbf{H}_3(\theta)$ .
- [3]  $\mathbf{X}_t$  is assumed to be weakly exogenous with respect to  $\theta$  (and  $\xi$ ).
- [4] The theoretical parameters of interest  $\xi = H(\theta)$ ,  $\xi \in \Xi$ , are identified.
- [5]  $\text{rank}(\mathbf{X}) = k$  where  $\mathbf{X} \equiv (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T)' : T \times k$  for  $T > k$ .

**(II) Probability model**

$$\Phi = \left\{ D(\mathbf{y}_t | \mathbf{X}_t; \theta) = \frac{(\det \Omega)^{-\frac{1}{2}}}{(2\pi)^{m/2}} \exp\left\{ -\frac{1}{2}(\mathbf{y}_t - \mathbf{B}'\mathbf{x}_t)' \Omega^{-1} (\mathbf{y}_t - \mathbf{B}'\mathbf{x}_t) \right\}, \theta \in \mathbb{R}^{mk} \times \mathbf{C}_m, t \in \mathbb{T} \right\}. \quad (25.65)$$

- [6]
  - (i)  $D(\mathbf{y}_t | \mathbf{X}_t; \theta)$  is normal;
  - (ii)  $E(\mathbf{y}_t | \mathbf{X}_t = \mathbf{x}_t) = \mathbf{B}'\mathbf{x}_t$  – linear in  $\mathbf{x}_t$ ;
  - (iii)  $\text{Cov}(\mathbf{y}_t | \mathbf{X}_t = \mathbf{x}_t) = \Omega$  – homoskedastic (free of  $\mathbf{x}_t$ );
- [7]  $\theta \equiv (\mathbf{B}, \Omega)$  is time invariant.

**(III) Sampling model**

- [8]  $\mathbf{y} \equiv (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_T)'$  is an independent sample sequentially drawn from  $D(\mathbf{y}_t | \mathbf{X}_t; \theta)$ ,  $t = 1, 2, \dots, T$ , respectively.

The above specification suggests most clearly that before we can proceed to consider the theoretical parameters of interest  $\xi$  we need to ensure that the statistical parameters of interest  $\theta$  are well defined. That is, the misspecification testing discussed in Chapter 24 precedes any discussion of either identification or statistical inference related to  $\xi$ . Testing departures from multivariate normality, linearity, homoskedasticity, time invariance

and independence became of paramount importance in econometric modelling in the context of simultaneous equations.

## 25.5 Maximum likelihood estimation

In view of the simultaneous equations model specification considered in the previous section and the discussion of the identification problem in Section 25.3 we can deduce that in the case where the theoretical (structural) parameters of interest  $\xi$  are *just-identified* the mapping  $\mathbf{H}(\cdot): \mathbb{R}^{mk} \times \mathbf{C}_n \rightarrow \Xi$ , where  $\xi = \mathbf{H}(\theta)$ , is one-to-one and onto. This implies that the reparametrisation is invertible,  $\theta = \mathbf{H}^{-1}(\xi)$  and an obvious estimator of  $\xi$  is the indirect maximum likelihood estimator (IMLE).

*Definition 3*

*In the case where  $\xi$  is just-identified its (**indirect**) **maximum likelihood estimator** (IMLE) is defined by*

$$\hat{\xi} = \mathbf{H}(\hat{\theta}), \quad (25.66)$$

where

$$\hat{\theta} = (\hat{\mathbf{B}}, \hat{\Omega}), \quad \hat{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \quad \hat{\Omega} = \frac{1}{T} \hat{\mathbf{U}}'\hat{\mathbf{U}}, \quad \hat{\mathbf{U}} = \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}. \quad (25.67)$$

*This estimator is based on the **invariance** property of MLE's.*

A more explicit formula for the IMLE is given in the case of one equation, say the first, when just-identification is achieved by linear homogeneous restrictions. It is defined as the solution of

$$\begin{pmatrix} \hat{\Pi} \\ \Phi_1 \end{pmatrix} \hat{\alpha}_{.1} = \mathbf{0}, \quad \hat{\Pi} = (\hat{\mathbf{B}} \mathbf{I}_k). \quad (25.68)$$

In the simpler case of exclusion restrictions the system (66) is

$$-\hat{\beta}_{11} + \hat{\mathbf{B}}_{12}\hat{\gamma}_1 + \delta_1 = \mathbf{0}. \quad (25.69)$$

$$-\hat{\beta}_{21} + \hat{\mathbf{B}}_{22}\hat{\gamma}_1 = \mathbf{0}. \quad (25.70)$$

and the IMLE's take the explicit form

$$\hat{\gamma}_1 = \hat{\mathbf{B}}_{22}^{-1}\hat{\mathbf{B}}_{21}, \quad (25.71)$$

$$\hat{\delta}_1 = \hat{\beta}_{11} - \hat{\mathbf{B}}_{12}\hat{\mathbf{B}}_{22}^{-1}\hat{\mathbf{B}}_{21}, \quad (25.72)$$

given that  $\mathbf{B}_{22}$  is a square  $(m_1 - 1) \times (m_1 - 1)$  non-singular matrix.

The IMLE can be viewed as an unconstrained MLE of the theoretical parameters of interest which might provide the 'benchmark' for testing any

## 622 The simultaneous equations model

overidentifying restrictions. As argued above, the identifying restrictions are not testable because they provide the reparametrisation from  $\theta$  to  $\xi$  but the overidentifying restrictions are testable because they imply restrictions for  $\theta$ , the statistical parameters of interest.

In order to simplify the derivation of general MLE's of  $\xi$  let us utilise the formulae derived in the context of constrained MLE's of  $\theta$  under a prior linear restrictions in Chapter 24.

The constrained MLE's of  $\mathbf{B}$  and  $\Omega$ , in the context of the multivariate linear regression model, subject to the linear a priori restrictions

$$\Pi \mathbf{A} \equiv \mathbf{B}\Gamma + \Delta = \mathbf{0}, \quad (25.73)$$

where  $(\Gamma, \Delta)$  are known, take the following form:

$$\tilde{\mathbf{B}} = \tilde{\mathbf{B}} - (\hat{\mathbf{B}}\Gamma + \Delta)(\Gamma'\hat{\Omega}\Gamma)^{-1}\Gamma'\hat{\Omega}, \quad (25.74)$$

$$\hat{\Omega} = \hat{\Omega} + \frac{1}{T}(\tilde{\mathbf{B}} - \hat{\mathbf{B}})'(\mathbf{X}'\mathbf{X})(\tilde{\mathbf{B}} - \hat{\mathbf{B}}) \quad (25.75)$$

(see Chapter 24). In the context of the simultaneous equations formulation these formulae can be reinterpreted as determining the theoretical parameters of interest  $\xi = \mathbf{G}(\Gamma, \Delta, \mathbf{V})$  given  $\tilde{\mathbf{B}}$  and  $\hat{\Omega}$ . That is, determine  $\xi$  via

$$\tilde{\mathbf{B}} = \tilde{\mathbf{B}} - (\hat{\mathbf{B}}\Gamma(\hat{\xi}) + \Delta(\hat{\xi}))(\Gamma(\hat{\xi}')\hat{\Omega}\Gamma(\hat{\xi}))^{-1}\Gamma(\hat{\xi})\hat{\Omega} \quad (25.76)$$

and

$$\mathbf{V}(\hat{\xi}) = \Gamma(\hat{\xi})'\hat{\Omega}\Gamma(\hat{\xi}) = \frac{1}{T}\mathbf{A}(\hat{\xi})'\mathbf{Z}'\mathbf{Z}\mathbf{A}(\hat{\xi}) \quad (25.77)$$

(see exercise 2) where

$$\mathbf{A} \equiv \begin{pmatrix} \Gamma \\ \Delta \end{pmatrix} \quad \text{and} \quad \mathbf{Z} \equiv (\mathbf{Y}, \mathbf{X})$$

(see Hendry and Richard (1983)). Substituting (76) and (77) into the log likelihood function of the multivariate linear regression model  $\log L(\theta; \mathbf{Y})$  we get the ‘concentrated’ likelihood function  $\log L(\xi; \mathbf{Y})$  defined by

$$\begin{aligned} \log L(\xi; \mathbf{Y}) &= \text{const} - \frac{T}{2} \log(\det \hat{\Omega}) \propto \\ &\quad -\frac{T}{2} \log(\det[(\Gamma'\mathbf{Y}'\mathbf{M}_x\mathbf{Y}\Gamma')^{-1}(\mathbf{A}'\mathbf{Z}'\mathbf{Z}\mathbf{A})]). \end{aligned} \quad (25.78)$$

This function can be viewed as the objective function for estimating  $\xi$  using  $\hat{\theta} \equiv (\hat{\mathbf{B}}, \hat{\Omega})$  as opposed to the direct estimation of  $\xi$  by solving (76) and (77) for  $\xi$ . The log likelihood function (78) has the advantage that it provides us with

a natural objective function to ‘solve’ (76) and (77). In the case of general restrictions  $\xi = \mathbf{H}(\theta)$  the ‘solution’ is rather prohibitive and thus we consider the case where the restrictions are *exclusion restrictions*.

When the identification of  $\xi$  is achieved by exclusion restrictions the constrained structural parameter matrices  $\Gamma(\xi)$  and  $\Delta(\xi)$  are *linear in  $\xi$* . This implies that the first-order conditions can be derived explicitly. The conditions

$$\frac{\partial \log L(\xi; \mathbf{Y})}{\partial \xi_i} = 0, \quad i = 1, 2, \dots, p, \quad (25.79)$$

give rise to the following system of equations:

$$\text{tr} \left[ (\tilde{\Gamma}' \hat{\Omega} \tilde{\Gamma})^{-1} \tilde{\Gamma}' \hat{\Omega} \frac{\partial \Gamma}{\partial \xi_i} - (\tilde{\mathbf{A}}' \mathbf{Z}' \mathbf{Z} \tilde{\mathbf{A}})^{-1} \tilde{\mathbf{A}}' \mathbf{Z}' \mathbf{Z} \frac{\partial \mathbf{A}}{\partial \xi_i} \right] = 0, \quad i = 1, 2, \dots, p. \quad (25.80)$$

Using the relation

$$\frac{\tilde{\mathbf{A}}' \mathbf{Z}' \mathbf{Z} \tilde{\mathbf{A}}}{T} = \tilde{\Gamma}' \hat{\Omega} \tilde{\Gamma}, \quad (25.81)$$

(80) takes the form

$$\text{tr} \left[ (\tilde{\Gamma}' \hat{\Omega} \tilde{\Gamma})^{-1} \tilde{\Gamma}' \hat{\Omega} \frac{\partial \Gamma}{\partial \xi_i} - (\tilde{\Gamma}' \hat{\Omega} \tilde{\Gamma})^{-1} \tilde{\mathbf{A}}' \left( \frac{\mathbf{Z}' \mathbf{Z}}{T} \right) \frac{\partial \mathbf{A}}{\partial \xi_i} \right] = 0. \quad (25.82)$$

The system of equations (82) could be used to derive the MLE’s of  $\xi$  and hence  $\tilde{\Gamma}$ ,  $\tilde{\mathbf{A}}$ ,  $\tilde{\mathbf{B}}$  and  $\hat{\Omega}$ . An asymptotically equivalent form can be derived using  $\hat{\Omega}$  for  $\Omega$  in view of the fact that  $(\tilde{\Omega} - \hat{\Omega}) \xrightarrow{P} \mathbf{0}$  and

$$\frac{1}{T} \left( \tilde{\Gamma}' \hat{\Omega} \frac{\partial \Gamma}{\partial \xi_i} - \tilde{\mathbf{A}}' \mathbf{Z}' \mathbf{Z} \frac{\partial \mathbf{A}}{\partial \xi_i} \right) = - \tilde{\mathbf{A}}' \mathbf{Z}' \mathbf{X} \hat{\Pi} \frac{\partial \mathbf{A}}{\partial \xi_i}, \quad \hat{\Pi} \equiv (\hat{\mathbf{B}} : \mathbf{I}). \quad (25.83)$$

Using these, (82) can be expressed in the asymptotically equivalent form

$$\text{tr} \left[ (\tilde{\Gamma}' \hat{\Omega} \tilde{\Gamma})^{-1} \tilde{\mathbf{A}}' \mathbf{Z}' \mathbf{X} \hat{\Pi} \frac{\partial \mathbf{A}}{\partial \xi_i} \right] = 0, \quad i = 1, 2, \dots, p. \quad (25.84)$$

For the details of the derivation see Hendry (1976), Hendry and Richard (1983). The reformulated system (84) is particularly intuitive because it can be interpreted as providing an estimator of  $\xi$  in terms of the sufficient statistics  $\hat{\mathbf{B}}$  and  $\hat{\Omega}$  of the form

$$\tilde{\xi} = \mathbf{H}(\hat{\mathbf{B}}, \hat{\Omega}). \quad (25.85)$$

$\hat{\mathbf{B}}$  and  $\hat{\Omega}$  are sufficient statistics for the MLR statistical model because, as

argued in Chapter 24, they are functions of the minimal sufficient statistics

$$\tau(\mathbf{Y}) \equiv (\mathbf{Y}'\mathbf{Y}, \mathbf{Y}'\mathbf{X}). \quad (25.86)$$

Moreover, the orthogonality between the systematic and non-systematic components, preserved with their estimated counterparts by the MLE's (discussed in Chapters 19 and 23) holds asymptotically for (84). In order to see this consider the systematic and non-systematic components related to the system of equations (73) defined by

$$E(\mathbf{Z}_t/\mathbf{X}_t = \mathbf{x}_t) = \boldsymbol{\Pi}'\mathbf{x}_t = \begin{pmatrix} \mathbf{B}' \\ I_k \end{pmatrix} \mathbf{x}_t, \quad (25.87)$$

and  $\boldsymbol{\varepsilon}_t$ , where  $\boldsymbol{\varepsilon}_t = -\mathbf{A}'\mathbf{Z}_t$ ,  $\mathbf{Z}_t \equiv (\mathbf{y}'_t, \mathbf{X}'_t)$ . (25.88)

It can be shown (see Hendry (1976)) that

$$\frac{1}{T} \mathbf{A}'\mathbf{Z}'\mathbf{X}\hat{\boldsymbol{\Pi}} \xrightarrow{P} \mathbf{0}. \quad (25.89)$$

a condition which suggests that the MLE can be interpreted as an IV estimator (see Section 25.7).

Unfortunately, explicit solutions for  $\boldsymbol{\xi}$  of the form given in (85) are not possible because (84) is non-linear in  $\boldsymbol{\xi}$  and it has to be solved by some numerical optimisation procedure (see Hendry (1976)). As emphasised by Hendry the numerical optimisation rule for deriving the MLE's  $\tilde{\boldsymbol{\xi}}$  of  $\boldsymbol{\xi}$  from (84) must be distinguished from the alternative *approximate solutions* of the system

$$\text{tr}\left(\mathbf{V}^{-1}\mathbf{A}'\mathbf{Z}'\mathbf{X}\boldsymbol{\Pi}\frac{\partial \mathbf{A}}{\partial \boldsymbol{\xi}_i}\right) = 0, \quad i = 1, 2, \dots, p, \quad (25.90)$$

where

$$\mathbf{V} = \frac{1}{T} \mathbf{A}'\mathbf{Z}'\mathbf{Z}\mathbf{A}, \quad \boldsymbol{\Pi} = (\mathbf{B}; \mathbf{I}), \quad (25.91)$$

$$\mathbf{B} = \hat{\mathbf{B}} - (\hat{\mathbf{B}}\boldsymbol{\Gamma} + \Delta)'(\boldsymbol{\Gamma}'\hat{\boldsymbol{\Omega}}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}'\hat{\boldsymbol{\Omega}}. \quad (25.92)$$

When  $\mathbf{V}$  and  $\boldsymbol{\Pi}$  are given, (90) is linear in  $\mathbf{A}$  and an explicit solution can be derived. On the other hand, when  $\mathbf{A}$  is given,  $\mathbf{V}$  and  $\boldsymbol{\Pi}$  can be derived easily. This suggests that (90)–(92) can be used to generate estimators of  $\mathbf{A}$  for different estimators of  $\mathbf{V}$  and  $\boldsymbol{\Pi}$ . Indeed, Hendry (1976) showed that most of the conventional estimators in the econometric literature such as three-stage least-squares (3SLS), two-stage least-squares (2SLS), can be easily generated by (90)–(92). Hendry referred to (90) as the *estimator generating*

equation (EGE). The usefulness of the EGE is that it unifies and summarises a huge literature in econometrics by showing that:

- (a) Every solution is an approximation to the maximum likelihood estimator obtained by variations in the ‘initial values’ selected for  $\mathbf{V}$  and  $\mathbf{\Pi}$  and in the number of steps taken through the cycle (90)–(92), including not iterating.
  - (b) All known econometric estimators for linear dynamic systems can be obtained in this way, which provides a systematic approach to estimation theory in an area where a very large number of methods have been proposed.
  - (c) Equation (90) classifies methods immediately into distinct groups of varying asymptotic efficiency as follows:
    - (i) as efficient asymptotically as the maximum likelihood  $\tilde{\mathbf{A}}$  if  $(\mathbf{V}, \mathbf{\Pi})$  are estimated consistently;
    - (ii) consistent for  $\mathbf{A}$  if any convergent estimator is used for  $(\mathbf{V}, \mathbf{\Pi})$ ;
    - (iii) asymptotically as efficient as can be obtained for a single equation out of a system if  $\mathbf{V} = \mathbf{I}$  but  $\mathbf{\Pi}$  is consistently estimated
- ...

(see Hendry and Richard (1983)).

#### *Definition 4*

*In the case where  $\Gamma$  is  $m \times m$  and non-singular the MLE of  $\xi$ , called the **full information maximum likelihood (FIML) estimator**, is the solution of the system (84) for  $\xi$ .*

It is interesting to note that in this case the system of equations (76) determining  $\xi$  takes the simple form

$$\tilde{\mathbf{B}} = -\Delta(\xi)\Gamma(\xi)^{-1}, \quad (25.93)$$

when the theoretical parameters of interest  $\xi$  are just identified then the statistical parameters are not constrained, i.e.  $\hat{\mathbf{B}} = \tilde{\mathbf{B}}$  and  $\hat{\Omega} = \tilde{\Omega}$  and thus (93) reduces to the *indirect maximum likelihood estimator*

$$\hat{\mathbf{B}} = -\Delta(\xi)\Gamma(\xi)^{-1}. \quad (25.94)$$

The score function for the general case is defined as

$$\mathbf{q}(\xi) = \frac{\partial \log L}{\partial \xi} \quad (25.95)$$

with  $\log L$  as defined in (78). The information matrix can be defined by

$$\mathbf{I}_T(\xi) = E[\mathbf{q}(\xi)\mathbf{q}(\xi)'] \quad (25.96)$$

(see Chapter 13), where  $E(\cdot)$  is with respect to the underlying probability model distribution  $D(\mathbf{y}_t; \mathbf{X}_t; \boldsymbol{\theta})$ . Using the asymptotic information matrix

defined by

$$\mathbf{I}_x(\xi) = \lim_{T \rightarrow \infty} \left( \frac{1}{T} \mathbf{I}_T(\xi) \right) \quad (25.97)$$

we can deduce that

$$\sqrt{T}(\tilde{\xi} - \xi)_{\text{FIML}} \underset{x}{\sim} N(\mathbf{0}, \mathbf{I}_x(\xi)^{-1}). \quad (25.98)$$

A more explicit form of  $\mathbf{I}_x(\xi)$  will be given in the next section in the case where only exclusion restrictions exist, for the 3SLS (three-stage least-squares) estimator.

One important feature of the above derivation of (84) and (90)–(92) is that it does not depend on  $\Gamma$  being  $m \times m$  and non-singular. These results hold true with  $\Gamma$  being  $m \times q$  ( $q \leq m$ ). In this case the structural formulation is said to be incomplete (see Richard (1984)). For  $\Gamma$   $m \times m_1$  ( $m_1 < m$ ), ‘solving’ (84) gives rise to a direct sub-system generalisation of the limited information maximum likelihood (LIML) estimator (with  $m_1 = 1$ ) (see Richard (1979)).

## 25.6 Least-squares estimation

As argued in the previous section most of the conventional estimators of the structural parameters  $\xi$  can be profitably viewed as particular approximations of the FIML estimator. Instead of deriving such estimators using the estimator generating equation (EGE) discussed above (see Hendry (1976) for a comprehensive discussion) we will consider the derivation of two of the most widely used (and discussed) least-squares estimators, the two-stage (2SLS) and three-stage least-squares (3SLS) estimators, in order to illustrate some of the issues raised in the previous sections. In particular, the role of weak exogeneity and a priori restrictions in the reparametrisation.

### (1) Two-stage least-squares (2SLS)

2SLS is by far the most widely used method of estimation for the structural parameters of interest in a particular equation. For expositional purposes let us consider the first structural equation of the system

$$\Gamma' \mathbf{y}_t + \Delta' \mathbf{x}_t + \varepsilon_t = 0, \quad (25.99)$$

$$y_{1t} = \Gamma_1^{0'} \mathbf{y}_t^{(1)} + \Delta_1' \mathbf{x}_t + \varepsilon_{1t}. \quad (25.100)$$

As argued in Section 25.2, equation (100) constitutes a proper statistical

GM based on the following decomposition of the probability model:

$$D(\mathbf{y}_t/\mathbf{X}_t; \boldsymbol{\theta}) = D(y_{1t}/\mathbf{y}_t^{(1)}, \mathbf{X}_t; \boldsymbol{\eta}_1) \cdot D(\mathbf{y}_t^{(1)}/\mathbf{X}_t; \boldsymbol{\eta}_{(1)}) \quad (25.101)$$

with systematic component

$$\mu_{1t}^s = E(y_t/\mathcal{F}_t^{(1)}) \quad (25.102)$$

and non-systematic component

$$\varepsilon_{1t} = y_t - E(y_t/\mathcal{F}_t^{(1)}), \quad \mathcal{F}_t^{(1)} = (\sigma(\mathbf{y}_t^{(1)}), \mathbf{X}_t = \mathbf{x}_t). \quad (25.103)$$

This suggests that, because of the normality of  $D(y_t/\mathbf{X}_t; \boldsymbol{\theta})$ ,  $\boldsymbol{\eta}_1$  and  $\boldsymbol{\eta}_{(1)}$  are variation free and thus  $\mathbf{y}_t^{(1)}$  is *weakly exogenous* with respect to  $\boldsymbol{\eta}_1$  (see Chapter 20). If no restrictions are placed on

$$\boldsymbol{\alpha}_1 \equiv \begin{pmatrix} \Gamma_1^0 \\ \Delta_1 \end{pmatrix}$$

its MLE is

$$\hat{\boldsymbol{\alpha}}_1 = (\mathbf{Z}'_{(1)} \mathbf{Z}_{(1)})^{-1} \mathbf{Z}'_{(1)} \mathbf{y}_1, \quad (25.104)$$

where  $\mathbf{Z}_{(1)} \equiv (\mathbf{Y}^{(1)}, \mathbf{X})$ . This estimator has the same properties in the estimator of  $\boldsymbol{\beta}^*$  in the stochastic linear regression model given that (100) is a hybrid of the linear and stochastic linear regression models. Moreover, the variance  $v_{11}$  can be estimated by

$$\hat{v}_{11} = \frac{1}{T} \hat{\varepsilon}_1' \hat{\varepsilon}_1, \quad \hat{\varepsilon}_1 = \mathbf{y}_1 - \mathbf{Z}_{(1)} \hat{\boldsymbol{\alpha}}_1. \quad (25.105)$$

The optimality of the estimators (104) and (105) arises because of the orthogonality between the systematic and non-systematic components

$$E(\mu_{1t}^s \varepsilon_{1t}) = E\{E[\mu_{1t}^s \varepsilon_{1t} / \sigma(\mathbf{y}_t^{(1)})]\} = E\{\mu_{1t}^s E(\varepsilon_{1t} / \sigma(\mathbf{y}_t^{(1)}))\} = 0 \quad (25.106)$$

since  $E[\varepsilon_{1t} / \sigma(\mathbf{y}_t^{(1)})] = 0$ . Note that the expectation operator in  $E(\mu_{1t}^s \varepsilon_{1t})$  is defined relative to  $D(\mathbf{y}_t/\mathbf{X}_t; \boldsymbol{\theta})$ , the distribution underlying the probability model.

The above scenario, however, changes drastically by the imposition of a priori restrictions on  $\boldsymbol{\eta}_1$ . In order to see this let us consider the simple case of *exclusion restrictions*. Let us rearrange the vectors  $\mathbf{y}_t$  and  $\mathbf{x}_t$  so as to get the  $m_1$  and  $k_1$  included endogenous ( $\mathbf{y}_{1t}$ ) and exogenous ( $\mathbf{x}_{1t}$ ) variables first, i.e.  $\mathbf{y}_t \equiv (y_{1t}, y'_{1t}, y'_{(1)t})'$ ,  $\mathbf{x}_t \equiv (x'_{1t}, x'_{(1)t})'$ . In terms of the decomposition in (101) this rearrangement suggests that

$$\begin{aligned} D(\mathbf{y}_t/\mathbf{X}_t; \boldsymbol{\theta}) &= D(y_{1t}/\mathbf{y}_{1t}, \mathbf{y}_{(1)t}, \mathbf{X}_t; \boldsymbol{\eta}_1) \\ &\quad \cdot D(\mathbf{y}_{1t}/\mathbf{y}_{(1)t}, \mathbf{X}_t; \boldsymbol{\eta}_1^*) \cdot D(\mathbf{y}_{(1)t}/\mathbf{X}_t; \boldsymbol{\eta}_{(1)}^*). \end{aligned} \quad (25.107)$$

In terms of this decomposition the statistical GM (100) becomes

$$y_{1t} = \gamma'_1 y_{1t} + \gamma'_{(1)} y_{(1)t} + \delta'_1 x_{1t} + \delta'_{(1)} x_{(1)t} + \varepsilon_{1t}, \quad (25.108)$$

where

$$\gamma_1: (m_1 - 1) \times 1, \quad \gamma_{(1)}: (m - m_1 - 1) \times 1,$$

$$\delta_1: k_1 \times 1, \quad \delta_{(1)}: (k - k_1) \times 1.$$

If we impose the exclusion restrictions

$$\gamma_{(1)} = \mathbf{0} \quad \text{and} \quad \delta_{(1)} = \mathbf{0} \quad (25.109)$$

(108) takes the *restricted structural form*

$$y_{1t} = \gamma'_1 y_{1t} + \delta'_1 x_{1t} + \varepsilon_{1t}^*. \quad (25.110)$$

Let us consider the implications of these restrictions for the rest of the system via

$$\mathbf{B}\Gamma_1^* + \Delta_1^* = \mathbf{0}. \quad (25.111)$$

Given that the restricted coefficient vectors take the form

$$\Gamma_1^* = (-1, \gamma'_1, \mathbf{0}')', \quad \Delta_1^* = (\delta'_1, \mathbf{0})', \quad (25.112)$$

(111) is

$$\begin{pmatrix} \beta_{11} & \mathbf{B}_{12} & \mathbf{B}_{13} \\ \beta_{21} & \mathbf{B}_{22} & \mathbf{B}_{23} \end{pmatrix} \begin{pmatrix} -1 \\ \gamma_1 \\ 0 \end{pmatrix} + \begin{pmatrix} \delta_1 \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix} \quad (25.113)$$

or

$$\mathbf{B}_{21}\gamma_1 + \delta_1 = \beta_{11}, \quad (25.114)$$

$$\mathbf{B}_{22}\gamma_1 = \beta_{21}. \quad (25.115)$$

These equations relate  $\eta_1$  and  $\eta_{(1)}$  via

$$\beta_1 = \begin{pmatrix} \beta_{11} \\ \beta_{21} \end{pmatrix} \quad \text{and} \quad \mathbf{B}_{(1)} = \begin{pmatrix} \mathbf{B}_{21} & \mathbf{B}_{31} \\ \mathbf{B}_{22} & \mathbf{B}_{32} \end{pmatrix}.$$

In the case where the equation (110) is *just identified* ( $k - k_1 = m_1 - 1$ )  $\mathbf{B}_{22}$  is  $(m_1 - 1) \times (m_1 - 1)$  square matrix with rank  $m_1 - 1$ . Hence, (114) and (115) can be solved uniquely for  $\gamma_1$  and  $\delta_1$ :

$$\gamma_1 = \mathbf{B}_{22}^{-1}\beta_{21}, \quad \delta_1 = \beta_{11} - \mathbf{B}_{21}\mathbf{B}_{22}^{-1}\beta_{21}. \quad (25.116)$$

Looking at (116) we can see that in this case  $\eta_1$  and  $\eta_{(1)}$  are still variation free. Moreover, no structural parameter estimator is needed because these can be estimated via the indirect MLE's:

$$\hat{\gamma}_1 = \hat{\mathbf{B}}_{22}^{-1}\hat{\beta}_{21}, \quad \hat{\delta}_1 = \hat{\beta}_{11} - \hat{\mathbf{B}}_{21}\hat{\mathbf{B}}_{22}^{-1}\hat{\beta}_{21}. \quad (25.117)$$

However, when  $(k - k_1) > m_1 - 1$ , the first equation is *overidentified*, the equations in (114), (115), impose restrictions on  $\mathbf{B}_{(1)}$  and thus the *variation free* condition between  $\boldsymbol{\eta}_1$  and  $\boldsymbol{\eta}_{(1)}$  is *invalidated*. In an attempt to enhance our understanding of this condition let us return to the decomposition in (107) and define the structural parameters involved.

$$\begin{aligned} \begin{pmatrix} y_{1t} \\ y_{1t} \\ y_{(1)t} \end{pmatrix} / \mathbf{X}_t &= \mathbf{x}_t \\ &\sim N \left( \begin{pmatrix} \beta'_{11} \mathbf{x}_{1t} + \beta'_{21} \mathbf{x}_{(1)t} \\ \mathbf{B}'_{12} \mathbf{x}_{1t} + \mathbf{B}'_{22} \mathbf{x}_{(1)t} \\ \mathbf{B}'_{13} \mathbf{x}_{1t} + \mathbf{B}'_{23} \mathbf{x}_{(1)t} \end{pmatrix} \begin{pmatrix} \omega_{11} & \omega_{12} & \omega_{13} \\ \omega_{21} & \Omega_{22} & \Omega_{23} \\ \omega_{31} & \Omega_{32} & \Omega_{33} \end{pmatrix} \right), \quad (25.118) \end{aligned}$$

$$\begin{aligned} \gamma_1 &= \Omega^{22} \omega_{21} + \Omega^{32} \omega_{31}, \quad \gamma_{(1)} = \Omega^{23} \omega_{21} + \Omega^{33} \omega_{31}. \\ \delta_{(1)} &= \beta_{11} - \mathbf{B}_{12} \gamma_1 - \mathbf{B}_{13} \gamma_{(1)}. \end{aligned} \quad (25.119)$$

$$\delta_{(1)} = \beta_{21} - \mathbf{B}_{22} \gamma_1 - \mathbf{B}_{23} \gamma_{(1)}, \quad v_{11} = \omega_{11} - \omega_{12} \gamma_1 - \omega_{13} \gamma_{(1)},$$

$$\begin{pmatrix} \Omega_{22} & \Omega_{23} \\ \Omega_{32} & \Omega_{33} \end{pmatrix}^{-1} = \begin{pmatrix} \Omega^{22} & \Omega^{23} \\ \Omega^{32} & \Omega^{33} \end{pmatrix}. \quad (25.120)$$

Under

$$\gamma_{(1)} = \mathbf{0}, \quad \delta_{(1)} = \mathbf{0} \Rightarrow \gamma_1 = \Omega_{22}^{-1} \omega_{21}, \quad (25.121)$$

$$v_{11} = \omega_{11} - \omega_{12} \Omega_{22}^{-1} \omega_{21}, \quad (25.122)$$

as well as (114), (115).

In the just identified case

$$\Omega_{22}^{-1} \omega_{21} = \mathbf{B}_{22}^{-1} \beta_{21} \quad (25.123)$$

and  $\gamma_1$  as defined by (121) and (115) coincide. On the other hand, in the overidentified case (121) and (115) define  $\gamma_1$  differently with (115) invalidating the variation free condition.

An alternative way to view the above argument is in terms of the systematic and non-systematic components. If we define  $\mu_{1t}^*$  in (110) by

$$\mu_{1t}^* = E(y_{1t} | \sigma(\mathbf{y}_{1t}), \mathbf{X}_{1t} = \mathbf{x}_{1t}), \quad (25.124)$$

then

$$E(\mu_{1t}^* e_{1t}) = E\{E(\mu_{1t}^* e_{1t} | \sigma(\mathbf{y}_t^{(1)}))\} \neq 0, \quad (25.125)$$

where  $E(\cdot)$  is defined in terms of  $D(\mathbf{y}_t | \mathbf{X}_t; \theta)$ . However, it can be verified directly that *under the restrictions* (114) and (115)

$$E(\mu_{1t}^* e_{1t}^*) = E(\mu_{1t}^* e_{1t}) = E\{E(\mu_{1t}^* e_{1t} | \sigma(\mathbf{y}_t^{(1)}))\} = 0, \quad (25.126)$$

where

$$\varepsilon_{1t}^* = y_{1t} - E(y_{1t}/\sigma(\mathbf{y}_{1t}), \mathbf{X}_{1t} = \mathbf{x}_{1t})$$

(see Spanos (1985d)). This suggests that the natural way to proceed in order to estimate  $(\gamma_1, \delta_1, v_{11})$  is to find a way to impose the restrictions (114)–(115) and then construct an estimator which preserves the orthogonality between  $\mu_{1t}^*$  and  $\varepsilon_{1t}^*$ . The LIML estimator briefly considered in the previous section is indeed such an estimator (see Theil (1971)). Another estimator based on the same argument is the so-called *two-stage least-squares* (2SLS) estimator. Let us consider its derivation in some detail.

The orthogonality in (126) suggests that

$$y_{1t} = \gamma'_1 \mathbf{y}_{1t} + \delta'_1 \mathbf{x}_{1t} + \varepsilon_{1t}^* \quad (25.127)$$

is equivalent to

$$y_{1t} = \beta'_{12} \mathbf{x}_{1t} + \beta'_{21} \mathbf{x}_{(1)t} + u_t, \quad (25.128)$$

subject to (114) and (115). In order to see this let us substitute

$$\mathbf{y}_{1t} = \mathbf{B}'_{12} \mathbf{x}_{1t} + \mathbf{B}'_{22} \mathbf{x}_{(1)t} + \mathbf{u}_{1t} \quad (25.129)$$

into (127):

$$y_{1t} = \gamma'_1 (\mathbf{B}'_{12} \mathbf{x}_{1t} + \mathbf{B}'_{22} \mathbf{x}_{(1)t} + \mathbf{u}_{1t}) + \delta'_1 \mathbf{x}_{1t} + \varepsilon_{1t}^* \quad (25.130)$$

$$= (\gamma'_1 \mathbf{B}'_{12} + \delta'_1) \mathbf{x}_{1t} + \delta'_1 \mathbf{B}'_{22} \mathbf{x}_{(1)t} + \gamma'_1 \mathbf{u}_{1t} + \varepsilon_{1t}^*. \quad (25.131)$$

Given  $\varepsilon_{1t}^* = -\Gamma_1^* \mathbf{u}_t = u_{1t} - \gamma'_1 \mathbf{u}_{1t}$ , (130) becomes

$$y_{1t} = (\gamma'_1 \mathbf{B}'_{12} + \delta'_1) \mathbf{x}_{1t} + \gamma'_1 \mathbf{B}'_{22} \mathbf{x}_{(1)t} + u_{1t}. \quad (25.132)$$

The LIML estimator can be viewed as a constrained MLE of  $\beta_{11}$  and  $\beta_{21}$  in (128) subject to the restrictions (114)–(115) as shown in (132) (see Theil (1971)). Similarly, the 2SLS estimator of  $\alpha_1^* \equiv (\gamma', \delta')'$  can be interpreted as achieving the same effect by a two-stage procedure. The method is based on a re-arranged formulation of (130) for  $t = 1, 2, \dots, T$ :

$$\mathbf{y}_1 = (\mathbf{X}_1 \mathbf{B}_{12} + \mathbf{X}_{(1)} \mathbf{B}_{22}) \gamma_1 + \mathbf{X}_1 \delta_1 + u_1, \quad (25.133)$$

in an attempt to impose (114) and (115) in stage one by estimating  $(\mathbf{X}_1 \mathbf{B}_{12} + \mathbf{X}_{(1)} \mathbf{B}_{22})$  separately using (129). Once the restrictions are imposed the next step is to construct an estimator of  $\alpha_1^*$  which preserves the orthogonality between the systematic and non-systematic component of (133). More explicitly:

*Stage one:* using the formulation provided by (133) estimate

$$\mathbf{X} \mathbf{B}_{12} = (\mathbf{X}_1 \mathbf{B}_{12} + \mathbf{X}_{(1)} \mathbf{B}_{22}) \quad (25.134)$$

in the context of (129) or

$$\mathbf{Y}_1 = \mathbf{X}\hat{\mathbf{B}}_{,2} + \mathbf{U}_1 \quad \text{using } \hat{\mathbf{B}}_{,2} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_1. \quad (25.135)$$

*Stage two:* using  $\hat{\mathbf{Y}}_1 = \mathbf{X}\hat{\mathbf{B}}_{,2}$  and the equality  $\mathbf{Y}_1 = \hat{\mathbf{Y}}_1 + \hat{\mathbf{U}}_1$ , where

$$\hat{\mathbf{U}}_1 = \mathbf{Y}_1 - \mathbf{X}\hat{\mathbf{B}}_{,2}, \quad (25.136)$$

(130) can be transformed into

$$\mathbf{y}_1 = \hat{\mathbf{Y}}_1\gamma_1 + \mathbf{X}_1\delta_1 + \mathbf{u}_1^* \quad (25.137)$$

or

$$\mathbf{y}_1 = \hat{\mathbf{Z}}_1\alpha_1^* + \mathbf{u}_1^*, \quad \mathbf{u}_1^* = \varepsilon_1^* + \hat{\mathbf{U}}'\gamma_1, \quad \hat{\mathbf{Z}}_1 = (\hat{\mathbf{Y}}_1; \mathbf{X}_1). \quad (25.138)$$

Estimate  $\alpha_1^*$  by

$$\hat{\alpha}_1^* = (\hat{\mathbf{Z}}_1'\hat{\mathbf{Z}}_1)^{-1}\hat{\mathbf{Z}}_1'\mathbf{y}_1 \quad (25.139)$$

or

$$\begin{pmatrix} \hat{\gamma}_1 \\ \hat{\delta}_1 \end{pmatrix}_{2SLS} = \begin{pmatrix} \hat{\mathbf{Y}}_1'\hat{\mathbf{Y}}_1 & \hat{\mathbf{Y}}_1'\mathbf{X}_1 \\ \mathbf{X}_1'\hat{\mathbf{Y}}_1 & \mathbf{X}_1'\mathbf{X}_1 \end{pmatrix}^{-1} \begin{pmatrix} \hat{\mathbf{Y}}_1'\mathbf{y}_1 \\ \mathbf{X}_1'\mathbf{y}_1 \end{pmatrix}. \quad (25.140)$$

In the context of the model (1)–(2) estimating the parameters of (1) by 2SLS amounts to estimating the statistical form in equation (4):

$$i_t = \beta_{21} + \beta_{22}p_t + \beta_{23}y_t + \beta_{24}g_t + u_{2t},$$

by OLS and substitute the fitted values  $\hat{i}_t = \hat{\beta}_{21} + \hat{\beta}_{22}p_t + \hat{\beta}_{23}y_t + \hat{\beta}_{24}g_t$  into the structural form of (1) to get:

$$m_t = \alpha_{11} + \alpha_{21}\hat{i}_t + \alpha_{31}p_t + \alpha_{41}y_t + u_t^*.$$

Applying OLS to this equation yields the 2SLS estimator of  $\alpha_{11}$ ,  $\alpha_{21}$ ,  $\alpha_{31}$  and  $\alpha_{41}$ .

Given that (133) preserves the orthogonality between the systematic and non-systematic component by redefining the latter, the 2SLS method uses the equivalence between (127) and (130) and via an operational form of the latter estimates  $\alpha_1^*$  consistently. Consistency ( $\hat{\alpha}_1^* \rightarrow \alpha_1^*$ ) stems from the asymptotic orthogonality

$$E(\hat{\mathbf{Z}}_1'\mathbf{u}_1^*) \rightarrow \mathbf{0} \quad \text{as } T \rightarrow \infty. \quad (25.141)$$

For a given  $T$ , however,  $E(\hat{\mathbf{Z}}_1'\mathbf{u}_1^*) \neq \mathbf{0}$  and thus  $\hat{\alpha}_1^*$  is a biased estimator, i.e.

$$E(\hat{\alpha}_1^*) \neq \alpha_1^*. \quad (25.142)$$

The 2SLS estimator can be compared directly with the LIML estimator

$$\begin{pmatrix} \hat{\gamma}_1 \\ \hat{\delta}_1 \end{pmatrix}_{LIML} = \begin{pmatrix} \mathbf{Y}_1'\mathbf{Y}_1 - \hat{l}^*\hat{\mathbf{U}}_1'\hat{\mathbf{U}}_1 & \mathbf{Y}_1'\mathbf{X}_1 \\ \mathbf{X}_1'\mathbf{Y}_1 & \mathbf{X}_1'\mathbf{X}_1 \end{pmatrix}^{-1} \begin{pmatrix} (\mathbf{Y}_1 - \hat{l}^*\hat{\mathbf{U}}_1)' \mathbf{y}_1 \\ \mathbf{X}_1' \mathbf{y}_1 \end{pmatrix}, \quad (25.143)$$

which, in view of the fact that

$$\hat{\mathbf{Y}}_1' \mathbf{X}_1 = \mathbf{Y}_1' \mathbf{X}_1 \quad \text{and} \quad \hat{\mathbf{Y}}_1' \hat{\mathbf{Y}}_1 = \mathbf{Y}_1' \mathbf{Y}_1 - \hat{\mathbf{U}}_1' \hat{\mathbf{U}}_1, \quad (25.144)$$

it differs from the 2SLS estimator in so far as  $\hat{l}^*$  is not given the value one but

$$\hat{l}^* = \frac{\hat{\gamma}_1^0 \mathbf{Y}_1^0' \mathbf{M}_1 \mathbf{Y}_1^0 \hat{\gamma}_1^0}{\hat{\gamma}_1^0 \mathbf{Y}_1^0' \mathbf{M}_x \mathbf{Y}_1^0 \hat{\gamma}_1^0}, \quad (25.145)$$

where

$$\hat{\gamma}_1^0 = \begin{pmatrix} -1 \\ \gamma_1 \end{pmatrix}, \quad \mathbf{Y}_1^0 = (\mathbf{y}_1, \mathbf{Y}_1), \quad \mathbf{M}_1 = \mathbf{I} - \mathbf{X}_1(\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1'$$

(see Schmidt (1976)). If we substitute  $k$  for  $\hat{l}^*$  in (143) where  $k$  is a scalar (stochastic or non-stochastic) (143) defines the so-called *k-class estimator* of  $\alpha_1^*$  which includes both 2SLS and LIML estimators. For the 2SLS estimator,  $k = 1$ .

Looking at (140) we can see that the 2SLS estimator exists if the rank of

$$\begin{pmatrix} \hat{\mathbf{Y}}_1' \hat{\mathbf{Y}}_1 & \hat{\mathbf{Y}}_1' \mathbf{X}_1 \\ \mathbf{X}_1' \hat{\mathbf{Y}}_1 & \mathbf{X}_1' \mathbf{X}_1 \end{pmatrix}$$

is  $m_1 + k_1 - 1$ . Given that  $\text{rank}(\mathbf{X}_1' \mathbf{X}_1) = k_1$  by assumption [5], Section 25.3, this rank condition is satisfied if  $\text{rank}(\hat{\mathbf{Y}}_1' \hat{\mathbf{Y}}_1) = m_1 - 1$ . The latter condition holds when  $\text{rank}(\mathbf{B}_{22}) \geq m_1 - 1$ , i.e. the order condition for the first equation is satisfied. It must be noted that in the case where this equation is exactly identified, i.e.

$$\text{rank}(\Phi_1) = \text{rank}(\Phi_1 \mathbf{A}^*) = m - 1, \quad (25.146)$$

the 2SLS estimator is equivalent to solving

$$\hat{\mathbf{B}} \Delta_1^* + \Gamma_1^* = \mathbf{0}, \quad (25.147)$$

where

$$\hat{\mathbf{B}} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{Y}, \quad \Gamma_1^* = \begin{pmatrix} -1 \\ \gamma_1 \\ \mathbf{0} \end{pmatrix}, \quad \Delta_1^* = \begin{pmatrix} \delta_1 \\ \mathbf{0} \end{pmatrix} \quad \text{for } \alpha_1^* \quad (25.148)$$

The solution of (147) for  $\alpha_1^*$  is the indirect MLE.

In order to discuss the properties of the 2SLS estimator we need to derive its distribution. From (140) we can deduce that

$$\hat{\gamma}_{2SLS} = (\mathbf{Y}_1' (\mathbf{P}_x - \mathbf{P}_{x_1}) \mathbf{Y}_1)^{-1} \mathbf{Y}_1' (\mathbf{P}_x - \mathbf{P}_{x_1}) \mathbf{y}_1, \quad (25.149)$$

$$\hat{\delta}_{2SLS} = (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' (\mathbf{y}_1 - \mathbf{Y}_1 \hat{\gamma}_{2SLS}), \quad (25.150)$$

where  $\mathbf{P}_x = \mathbf{X}(\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'$ ,  $\mathbf{P}_{x_1} = \mathbf{X}_1(\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1'$ . These results show that the distribution of  $\hat{\delta}_{2SLS}$  can be derived directly from that of  $\hat{\gamma}_{2SLS}$ . Concentrating

on the latter estimator we can express it in the form

$$\hat{\gamma}_{2SLS} = \mathbf{W}_{22}^{-1} \mathbf{W}_{21}, \quad (25.151)$$

where

$$\mathbf{W}_1^0 \equiv \begin{pmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{22} \end{pmatrix} = \mathbf{Y}_1^{0\prime} (\mathbf{P}_x - \mathbf{P}_{x_1}) \mathbf{Y}_1^0, \quad \mathbf{Y}_1^0 = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{Y}_1 \end{pmatrix} \quad (25.152)$$

(see Mariano (1982)). In view of the fact that

$$(\mathbf{y}_{1t}^0 / \mathbf{X}_t = \mathbf{x}_t) \sim N(\mathbf{B}_1^0 \mathbf{x}_t, \boldsymbol{\Omega}_1^0), \quad (25.153)$$

where  $\mathbf{B}_1^0 \equiv (\mathbf{B}'_{1,1}, \mathbf{B}'_{1,2})'$ , we can see that the distribution of  $\mathbf{Y}_1^0 \mathbf{Q} \mathbf{Y}_1^0$  where  $\mathbf{Q} = \mathbf{P}_x - \mathbf{P}_{x_1}$  is an idempotent matrix must be a matrix extension to the non-central chi-square (see Appendix 24.1). That is,

$$\mathbf{W}_1^0 \sim W_{m_1}(\boldsymbol{\Omega}_1^0, T; \mathbf{M}_1^0), \quad (25.154)$$

where  $W_{m_1}(\cdot)$  stands for the non-central Wishart distribution with  $T$  degrees of freedom, scale matrix  $\boldsymbol{\Omega}_1^0$  and non-centrality (or means-sigma) matrix

$$\mathbf{M}_1^0 = \boldsymbol{\Omega}_1^{0-1} E(\mathbf{Y}_1^0)' \mathbf{Q} E(\mathbf{Y}_1^0). \quad (25.155)$$

The 2SLS estimator of  $\delta_1$  as given in (150) can be viewed as a regressor function of  $\mathbf{W}_1^0$ . The  $k$ -class and LIML estimators of  $\gamma_1$  can be expressed similarly as

$$\hat{\gamma}_{(k)} = (\mathbf{W}_{22} + \bar{k} \mathbf{S}_{22})^{-1} (\mathbf{W}_{21} + \bar{k} \mathbf{S}_{21}), \quad (25.156)$$

$$\hat{\gamma}_{LIML} = (\mathbf{W}_{22} - \hat{l}^* \mathbf{S}_{22})^{-1} (\mathbf{W}_{21} - \hat{l}^* \mathbf{S}_{21}), \quad (25.157)$$

where

$$\bar{k} = 1 - k, \quad \mathbf{S}_1^0 = \frac{1}{T} (\mathbf{Y}_1^0 - \mathbf{X} \mathbf{B}_1^0)' (\mathbf{Y}_1^0 - \mathbf{X} \mathbf{B}_1^0) \equiv \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix} \quad (25.158)$$

(see Mariano (1982)). These results suggest that the same methods for the derivation of the distribution of  $\hat{\gamma}_{2SLS}$  can be used to derive those of  $\hat{\gamma}_{(k)}$  and  $\hat{\gamma}_{LIML}$ . In Chapter 6, several ways to derive the distribution of Borel functions of random vectors were discussed and the general impression was that such derivations can be very difficult exercises in multivariate calculus. In the case of  $\hat{\gamma}_1$  above, the derivations are further complicated by the fact that the distribution function of the non-central Wishart is a highly complicated function based on infinite series (see Muirhead (1982)). The complexity of the distribution increases rapidly with the rank( $\mathbf{M}_1^0$ ). This is basically the reason why the econometric literature on the distribution of these estimators concentrated mainly on the case where  $m_1 = 1$  (see Basmann (1974) for an excellent survey). In the general case of  $m_1 > 1$  (see Phillips (1980)) the distribution of  $\psi$  is so complicated so as to be almost

non-operational. This encouraged the development of asymptotic results and related expansions such as the Edgeworth expansions (see Chapter 10). Before we consider these results it is important to summarise some of the most interesting aspects of the finite sample research related to the above estimators; in particular to the existence of moments of the various estimators discussed above.

- (i)  $\hat{\gamma}_{2SLS}$  has moments up to order  $(k_2 - m_1 + 1)$ , i.e. the 2SLS estimator has moments up to the degree of overidentification. Thus, for a just-identified equation where  $k_2 = m_1 - 1$  no moments exist; see Kinal (1980).
- (ii)  $\hat{\gamma}_{(k)}$  has moments up to order  $T - (k_1 + m_1 - 1)$  if  $0 \leq k < 1$  and  $k$  is non-stochastic; see Kinal (1980).
- (iii)  $\hat{\gamma}_{LIML}$  has no moments (see Sargan (1970)).

See Mariano (1982) for an excellent survey of these results. It must be emphasised that the non-existence of moments should not be interpreted as making the relevant estimators inferior to ones whose moments exist. It implies, however, that comparisons of estimators based on criteria such as mean square error cannot be made since they presuppose the existence of some moments. Moreover, in Monte Carlo studies (see Hendry (1984)) the non-existence of moments provides useful information.

The asymptotic distribution of the 2SLS estimator takes the form

$$\sqrt{T}(\hat{\alpha}_1^* - \alpha_1^*)_{2SLS} \underset{a}{\sim} N(\mathbf{0}, v_{11}^* \mathbf{D}_{11}^{-1}), \quad (25.159)$$

where

$$\mathbf{D}_{11} = \begin{pmatrix} \mathbf{B}'_{.2} \mathbf{Q}_x \mathbf{B}_{.2} & \mathbf{B}'_{.2} \mathbf{Q}_1 \\ \mathbf{Q}_1 \mathbf{B}'_{.2} & \mathbf{Q}_{11} \end{pmatrix}, \quad (25.160)$$

$$\mathbf{Q}_x = \lim_{T \rightarrow \infty} \left( \frac{\mathbf{X}' \mathbf{X}}{T} \right), \quad Q_{11} = \lim_{T \rightarrow \infty} \left( \frac{\mathbf{X}'_1 \mathbf{X}_1}{T} \right),$$

$$\mathbf{Q}_1 = \lim_{T \rightarrow \infty} \left( \frac{\mathbf{X}' \mathbf{X}_1}{T} \right), \quad v_{11}^* \equiv E(\epsilon_{1t}^* \epsilon_{1t}^*).$$

The asymptotic covariance of  $\hat{\alpha}_1$  can be estimated by

$$\text{Cov}(\hat{\alpha}_1^*) = \hat{v}_{11}^* \begin{pmatrix} \hat{\mathbf{Y}}'_1 \hat{\mathbf{Y}}_1 & \hat{\mathbf{Y}}'_1 \mathbf{X}_1 \\ \mathbf{X}'_1 \hat{\mathbf{Y}}_1 & \mathbf{X}'_1 \mathbf{X}_1 \end{pmatrix}^{-1}, \quad (25.161)$$

where

$$\hat{v}_{11}^* = \frac{1}{T} \hat{\epsilon}_1^* \hat{\epsilon}_1^*, \quad \hat{\epsilon}_1^* = \mathbf{y}_1 - \mathbf{Y}_1 \hat{\gamma}_1 - \mathbf{X}_1 \hat{\delta}_1$$

(see Schmidt (1976)). It is important to note that  $\mathbf{D}_{11}$  is non-singular if  $\text{rank}(\mathbf{B}_{22}) = m_1 - 1$ , i.e. the first equation is *identified*.

Under the condition that

$$\sqrt{T}(k-1) \xrightarrow{P} 0, \quad (25.162)$$

the  $k$ -class estimators are asymptotically equivalent to the 2SLS. In particular the LIML and 2SLS are asymptotically equivalent.

A brief introduction to Edgeworth expansions was considered in Chapter 10. Sargan (1976) considered the question of applicability of such expansions to econometric estimators. He proposed an Edgeworth expansion for the difference

$$(\hat{\alpha} - \alpha) = e_T(\mathbf{p}, \mathbf{w}), \quad (25.163)$$

where  $\mathbf{p}$  is a normally distributed random vector, and  $\mathbf{w}$  a random vector, independent of  $\mathbf{p}$ . The conditions placed on  $e_T(\mathbf{p}, \mathbf{w})$  are general enough to be applicable to many of the conventional estimators mentioned above; see also Phillips (1977). These conditions are related to the smoothness and invertibility of  $e_T(\cdot)$  and the boundedness of the moments of  $\sqrt{T}\mathbf{w}$  of all orders. In the case of the 2SLS it can be shown that

$$\hat{\gamma}_1 = h(\mathbf{p}, \boldsymbol{\beta}_{.1}, \mathbf{B}_{.2}), \quad (25.164)$$

where

$$\mathbf{p} \equiv \left( \frac{\mathbf{X}'\mathbf{u}_1}{T}, \frac{\mathbf{X}'\mathbf{U}_1}{T} \right), \quad \mathbf{w} = \mathbf{0} \quad (25.165)$$

and the Sargan conditions apply (see Phillips (1980a)). The asymptotic expansion of order  $O_p(T^{-\frac{1}{2}})$  takes the general form

$$(\hat{\gamma}_1 - \gamma_1) = F_0 + F_{-\frac{1}{2}} + F_{-1} + O_p(T^{-\frac{3}{2}}), \quad (25.166)$$

where the components of  $F_0, F_{-\frac{1}{2}}, F_{-1}$  include the terms of order  $O_p(1), O_p(T^{-\frac{1}{2}}), O_p(T^{-1})$  respectively.

## (2) Three-stage least squares (3SLS)

The 3SLS estimator is a system GLS estimator based on the SURE formulation considered in Chapter 24. As argued there, this formulation enables us to accommodate exclusion restrictions directly into the formulation itself.

The results related to the estimation of the MLR model when a priori information is available (see Section 24.4) suggest that in estimating (110) separated from the system there must be some loss of efficiency. In the case of exclusion restrictions the reformulation of the MLR model into the SURE form enabled us to incorporate such a priori information. Intuition

suggests that the same formulation should lead us to a more efficient system estimator than the 2SLS estimator.

Expressing all  $m$  equations of (99) in the same form as equation one in (110) the SURE formulation in the present context gives rise to the system

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_m \end{pmatrix} = \begin{pmatrix} \mathbf{Z}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & & \\ \vdots & & & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{Z}_m \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_m \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1^* \\ \boldsymbol{\varepsilon}_2^* \\ \vdots \\ \boldsymbol{\varepsilon}_m^* \end{pmatrix}, \quad (25.167)$$

where

$$\mathbf{Z}_i \equiv (\mathbf{Y}_i; \mathbf{X}_i), \quad \boldsymbol{\alpha}_i = \begin{pmatrix} \gamma_i \\ \delta_i \end{pmatrix}, \quad i = 1, 2, \dots, m,$$

or, more compactly,

$$\mathbf{y}_* = \mathbf{Z}_* \boldsymbol{\alpha}_* + \boldsymbol{\varepsilon}_*^* \quad (25.168)$$

is an obvious notation. The 2SLS estimator amounts to applying OLS to the system

$$\mathbf{y}_* = \hat{\mathbf{Z}}_* \boldsymbol{\alpha}_* + \boldsymbol{\varepsilon}_*^*, \quad (25.169)$$

where  $\hat{\mathbf{Z}}_*$  is  $\mathbf{Z}_*$  with  $\hat{\mathbf{Y}}_i$  instead of  $\mathbf{Y}_i$ , i.e.  $\hat{\mathbf{Z}}_i \equiv (\hat{\mathbf{Y}}_i; \mathbf{X}_i)$ . That is, the 2SLS estimator for the system as a whole is

$$\hat{\boldsymbol{\alpha}}_{*2SLS} = (\hat{\mathbf{Z}}'_* \hat{\mathbf{Z}}_*)^{-1} \hat{\mathbf{Z}}'_* \mathbf{y}_*. \quad (25.170)$$

In view of the fact that

$$(\boldsymbol{\varepsilon}_*^* / \mathbf{X}) \sim N(\mathbf{0}, \mathbf{V} \otimes \mathbf{I}_T), \quad (25.171)$$

intuition suggests that the generalised-least-squares (GLS) estimator used in the context of the MLR model should be more appropriate. This estimator takes the form (see Section 24.4):

$$\hat{\boldsymbol{\alpha}}_{*3SLS} = (\hat{\mathbf{Z}}'_* (\hat{\mathbf{V}}^{-1} \otimes \mathbf{I}_T) \hat{\mathbf{Z}}_*)^{-1} \hat{\mathbf{Z}}'_* (\hat{\mathbf{V}}^{-1} \otimes \mathbf{I}_T) \mathbf{y}_*, \quad (25.172)$$

where  $\hat{\mathbf{V}}$  is estimated from the 2SLS residuals applied to all the equations of the system via

$$\hat{v}_{ij} = \frac{1}{T} \hat{\boldsymbol{\varepsilon}}_{i*}' \hat{\boldsymbol{\varepsilon}}_j^*, \quad i, j = 1, 2, \dots, m. \quad (25.173)$$

For obvious reasons this estimator of  $\boldsymbol{\alpha}_*$  is called the three-stage least-squares (3SLS) estimator, first suggested by Zellner and Theil (1962).

It is important to note that for  $(\hat{\mathbf{Z}}'_* (\hat{\mathbf{V}}^{-1} \otimes \mathbf{I}_T) \hat{\mathbf{Z}}_*)$  in (189) to be invertible  $\hat{\mathbf{Z}}_*$  must be of full column rank which requires that each  $\hat{\mathbf{Z}}_i = (\hat{\mathbf{Y}}_i, \mathbf{X}_i)$ ,  $i = 1, 2, \dots, m$ .

$2, \dots, m$ , i.e. all equations comprising the system must be *identified*. Moreover, when the system is identified

$$\frac{1}{T} (\hat{\mathbf{Z}}'_* (\hat{\mathbf{V}}^{-1} \otimes \mathbf{I}_T) \hat{\mathbf{Z}}_*) \xrightarrow{P} \mathbf{D} \quad (25.174)$$

and  $\mathbf{D}$  is non-singular. This enables us to deduce that

$$\sqrt{T}(\hat{\boldsymbol{\alpha}}_{*3SLS} - \boldsymbol{\alpha}_*) \underset{\alpha}{\sim} N(\mathbf{0}, \mathbf{D}^{-1}) \quad (25.175)$$

(see Schmidt (1976)). Sargan (1964a) showed that when only exclusion restrictions are present and the system is identified, 3SLS and FIML are asymptotically equivalent, i.e.  $\mathbf{D} = \mathbf{I}_\infty(\xi)$ ; see Section 25.3.

In relation to the finite sample distribution of  $\hat{\boldsymbol{\alpha}}_{*3SLS}$  the result related to the moments of the 2SLS applies to this case as well. That is, moments exist up to the order of overidentification (see Sargan (1978)). Moreover, the non-existence of moments for the LIML estimator extends to the FIML estimator (see Sargan (1970)). As argued above, however, existence or non-existence of moments should not be considered as a criterion for choosing between estimators in general.

## 25.7 Instrumental variables

The method of instrumental variables (IV) was initially proposed by Reiersol (1941) in the context of the errors-in-variables model (see Judge *et al.* (1982)) as an alternative estimation method which ‘solved’ the inconsistency problem associated with the OLS estimator. This was then systematised and extended by Durbin (1954) to a more general linear statistical model where the explanatory variables are correlated with the error term. He also conjectured the direct relationship between the IV estimator and the LIML estimator in the context of the simultaneous equations model. The current approach to the IV method was formalised by Sargan (1958), who considered the asymptotic efficiency as well as the consistency of IV estimators. Let us summarise the current ‘textbook’ approach in order to motivate the formalisation proposed in the sequel.

Consider the statistical formulation

$$y_t = \boldsymbol{\alpha}' \mathbf{X}_t + \varepsilon_t, \quad t \in \mathbb{T} \quad (25.176)$$

where  $\mathbf{X}_t$ :  $p \times 1$  vector of (possibly stochastic) explanatory variables such that

$$E(\mathbf{X}_t \varepsilon_t) \neq 0, \quad t \in \mathbb{T}. \quad (25.177)$$

If we estimate  $\alpha$  using the usual orthogonal projection (OLS) estimator

$$\hat{\alpha} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

we can see that this estimator is in general *biased* and *inconsistent* since

$$\hat{\alpha} = \alpha + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \quad \text{and} \quad E\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon\} \neq 0$$

and the bias does not go to zero as  $T \rightarrow \infty$ .

The IV method ‘solves’ the bias (and inconsistency) problem by introducing a new vector of instruments  $\mathbf{Z}_t: m \times 1, t = 1, 2, \dots, T$  such that:

$$(i) \quad E(\mathbf{Z}_t'\varepsilon_t) = \mathbf{0} \quad \text{or} \quad ((\mathbf{Z}'\varepsilon/T) \xrightarrow{P} \mathbf{0})$$

$$(ii) \quad \text{Cov}(\mathbf{Z}_t) = \Sigma_{33}^P \quad \text{or} \quad ((\mathbf{Z}'\mathbf{Z}/T) \xrightarrow{P} \Sigma_{33})$$

$$(iii) \quad \text{Cov}(\mathbf{Z}_t, \mathbf{X}_t) = \Sigma_{32}^P \quad \text{or} \quad ((\mathbf{Z}'\mathbf{X}/T) \xrightarrow{P} \Sigma_{32})$$

where  $\Sigma_{33} > 0$ . In the case where the number of instruments equals the number of unknown parameters in  $\alpha$ , i.e.  $m = p$ , the IV estimator takes the form:

$$\hat{\alpha}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}. \quad (25.178)$$

Assuming that  $\Sigma_{32}$  is bounded and non-singular we can deduce that

$$E(\hat{\alpha}_{IV}) = \alpha \quad \text{and} \quad \hat{\alpha}_{IV} \xrightarrow{P} \alpha$$

Moreover, if in addition to (i)–(iii) we also assume that

$$(iv) \quad (\mathbf{Z}'\varepsilon/\sqrt{T}) \sim_{\alpha} N(\mathbf{0}, \sigma^2 \Sigma_{33})$$

then

$$\sqrt{T}(\hat{\alpha}_{IV} - \alpha) \sim_{\alpha} N(\mathbf{0}, \sigma^2 \Sigma_{23}^{-1} \Sigma_{33} \Sigma_{32}^{-1})$$

In the case where  $m > p$  (we have more instruments than  $\alpha$ s) Sargan (1958) proposed the so-called generalised instrumental variable estimator (GIVE)

$$\hat{\alpha}_{IV}^* = (\mathbf{X}'\mathbf{P}_z\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_z\mathbf{y} \quad (25.179)$$

where  $\mathbf{P}_z = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ . The GIVE estimator was derived by Sargan as an extension of  $\alpha_{IV}$  where the ‘optimal instruments’ are chosen as linear functions of  $\mathbf{Z}$ , i.e.  $\mathbf{Z}^* = \mathbf{ZH}$ ,  $\mathbf{H}$  being a  $m \times p$  matrix and  $\mathbf{H}$  is chosen so as to minimise the asymptotic covariance matrix

$$\text{Cov}(\hat{\alpha}_{IV}^*) = \sigma^2 (\mathbf{H}\Sigma_{23})^{-1} (\mathbf{H}'\Sigma_{33}\mathbf{H})(\Sigma_{23}'\mathbf{H}')^{-1}$$

of  $\hat{\alpha}_{IV}^* = (\mathbf{Z}^{*'} \mathbf{X})^{-1} \mathbf{Z}^{*'} \mathbf{y}$ .

Given that for  $l(\mathbf{H}) = (\mathbf{H} \boldsymbol{\Sigma}_{23})^{-1} (\mathbf{H}' \boldsymbol{\Sigma}_{33} \mathbf{H}) (\boldsymbol{\Sigma}_{23}' \mathbf{H}')^{-1} l(\mathbf{H}) = l(\mathbf{A}\mathbf{H})$  for any  $m \times m$  non-singular matrix we can choose  $\mathbf{H}$  by minimising

$$(\mathbf{H}' \boldsymbol{\Sigma}_{33} \mathbf{H}) \text{ subject to } \mathbf{H} \boldsymbol{\Sigma}_{23} = \mathbf{I}.$$

The optimal  $\mathbf{H}$  takes the form:

$$\mathbf{H} = \boldsymbol{\Sigma}_{33}^{-1} \boldsymbol{\Sigma}_{32} (\boldsymbol{\Sigma}_{23} \boldsymbol{\Sigma}_{33}^{-1} \boldsymbol{\Sigma}_{23})^{-1}.$$

Using its sample equivalent the ‘optimal’ set of instruments is

$$\mathbf{Z}^* = \mathbf{P}_z \mathbf{X} (\mathbf{X}' \mathbf{P}_z \mathbf{X})^{-1} \text{ and thus } \hat{\alpha}_{IV}^* = \hat{\alpha}_{IV}^*.$$

$$\sqrt{T}(\hat{\alpha}_{IV}^* - \alpha) \sim_{\mathcal{D}} N(\mathbf{0}, \sigma^2 (\boldsymbol{\Sigma}_{23} \boldsymbol{\Sigma}_{33}^{-1} \boldsymbol{\Sigma}_{32})^{-1}).$$

The GIVE estimator as derived above is not just consistent but also asymptotically efficient. Because of this it should come as no surprise to learn that a number of well-known estimators in econometrics including OLS, GLS, 2SLS, 3SLS, LIML and FIML can be viewed as GIVE estimators (see Bowden and Turkington (1984), *inter alia*).

Note that in the case where  $m = p$ ,  $\hat{\alpha}_{IV}^* = \hat{\alpha}_{IV}$  as defined in (178).

The main problem in implementing the IV method in practice is the choice of the instruments  $\mathbf{Z}_t$ . The above argument begins with the presupposition that such a set of instruments exists. As for the conditions (i)–(iv), they are of rather limited value in practice because the basic orthogonality condition (i), as it stands, is non-operational. In order to make it operational we need to specify explicitly the distribution in terms of which  $E(\cdot)$  in (176) and (i)–(iv) are defined. If we look at the argument leading to the GIVE estimator (179) we can see that the whole argument ‘revolves’ around an (implicit) distribution which ‘somehow’ involves  $y_t$ ,  $\mathbf{X}_t$  and  $\mathbf{Z}_t$ . In order to see this let us assume that the underlying distribution is  $D(y_t, \mathbf{X}_t; \psi)$  (assumed to be normal for simplicity). If we interpret the systematic and non-systematic components of (176) as in Chapters 17 and 20, i.e.

$$\mu_t = E(y_t | \sigma(\mathbf{X}_t)) = \alpha' \mathbf{X}_t \quad \text{and} \quad \varepsilon_t = y_t - E(y_t | \sigma(\mathbf{X}_t)) \quad (25.180)$$

respectively, then

$$(a) \quad \alpha = \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\sigma}_{21} \quad (\boldsymbol{\Sigma}_{22} = \text{Cov}(\mathbf{X}_t), \quad \boldsymbol{\sigma}_{21} = \text{Cov}(\mathbf{X}_t, y_t)),$$

$$(b) \quad E(\mathbf{X}'_t \varepsilon_t) = \mathbf{0}, \quad \text{by construction}$$

and

$$(c) \quad \hat{\alpha} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y} \text{ is the MLE of } \alpha \text{ and it is unbiased, consistent and fully efficient.}$$

This suggests immediately that  $D(y_t, \mathbf{X}_t; \psi)$  is *not* the distribution underlying the above IV argument and thus (180) is not the (implicit) interpretation of  $\alpha' \mathbf{X}_t$  and  $\varepsilon_t$ .

Another question which raises the issue of the underlying distribution is related to the possible conflict between condition (i) and

$$(v) \quad \text{Cov}(\mathbf{Z}_t, y_t) = \sigma_{31} \neq 0.$$

A glance at the IV estimators (178) and (179) confirm that the latter condition is needed in order to ensure the existence of these estimators, in addition to (i)–(iii). This, however, requires  $\mathbf{Z}_t$  to be correlated with  $y_t$  but uncorrelated with a random variable  $\varepsilon_t$ , directly related to  $y_t$ . ‘How do we resolve this apparent paradox?’

In addition to the above raised questions the discerning reader might be wondering how the GIVE estimator (179) can be a ‘good’ estimator of  $\alpha$  in (176) when apparently the latter parameter has ‘nothing to do’ with  $\mathbf{Z}_t$ . As emphasised throughout Chapters 19–24 there is nothing coincidental about the form of the parameter to be estimated and its ‘best’ estimator. For example the fact that  $\hat{\alpha} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$  is a ‘good’ estimator of  $\alpha = \Sigma_{22}^{-1}\sigma_{21}$  is no accident, given that the natural sample analogues of  $\Sigma_{22}$  and  $\sigma_{21}$  are  $\mathbf{X}'\mathbf{X}$  and  $\mathbf{X}'\mathbf{y}$  respectively. Using the same analogy in the case of  $\hat{\alpha}_{IV}^*$  we can see that the parameter this is a ‘good’ estimator of must be:

$$\alpha^* = (\Sigma_{23}\Sigma_{33}^{-1}\Sigma_{32})^{-1}\Sigma_{23}\Sigma_{33}^{-1}\sigma_{31}. \quad (25.181)$$

How is this parameter implicitly assumed to be the parameter of interest in (176)?

The above raised questions indicate that the issues of the (implicit) distribution underlying (176), the non-orthogonality condition (177) and the (implicit) parametrisation are inextricably bound up. The fact that (181) involves the moments of all random variables ( $y_t, \mathbf{X}_t, \mathbf{Z}_t$ ) suggests that the (implicit) distribution is related to their joint distribution. Moreover, the apparent conflict between conditions (i) and (v) can only be resolved if  $\mathbf{Z}_t$  refer to conditioning variables. Putting these ‘clues’ together they suggest that the underlying distribution should be:

$$D(y_t, \mathbf{X}_t, \mathbf{Z}_t; \theta). \quad (25.182)$$

Let us consider how this choice of the implicit distribution could help us resolve the above raised issues.

Firstly, assuming that the parametrisation of interest in (176) is  $\alpha^*$ , as defined in (181), we can see how the non-orthogonality (177) arises. If we ignore (182) and instead treat  $E(\cdot)$  in (177) as defined in terms of  $D(y_t, \mathbf{X}_t; \psi)$  then

$$E(\mathbf{X}'_t \varepsilon_t) = E(\mathbf{X}'_t (y_t - \alpha' \mathbf{X}_t)) = (\sigma_{21} - \Sigma_{22} \alpha^*) \neq 0$$

unless  $\alpha^* = \Sigma_{22}^{-1}\sigma_{21}$ . This leaves us with having to explain how the parametrisation of  $\alpha^*$  in (181) arises. Using (182) as the underlying distribution we can deduce that

$$\begin{aligned} E(\mathbf{X}'_t \varepsilon_t / \sigma(\mathbf{Z}_t)) &= E\{E[\mathbf{X}'_t \varepsilon_t / \sigma(\mathbf{Z}_t)] / \sigma(\mathbf{X}_t)\} = E\{\mathbf{X}'_t E[\varepsilon_t / \sigma(\mathbf{X}_t)] / \sigma(\mathbf{Z}_t)\} \\ &= E\{\mathbf{X}'_t E[(y_t - \alpha^* \mathbf{X}_t) / \sigma(\mathbf{Z}_t)] / \sigma(\mathbf{X}_t)\} \\ &= E\{\mathbf{X}'_t [\mathbf{Z}_t \Sigma_{33}^{-1} \sigma_{31} - \mathbf{Z}_t \Sigma_{33}^{-1} \Sigma_{32} \alpha^*] / \sigma(\mathbf{X}_t)\} \\ &= \Sigma_{23} \Sigma_{33}^{-1} \sigma_{31} - \Sigma_{23} \Sigma_{33}^{-1} \Sigma_{32} \alpha^* \end{aligned}$$

(see Section 7.2).

This shows clearly that if (182) is the underlying distribution, then

$$E(\mathbf{X}'_t \varepsilon_t / \sigma(\mathbf{Z}_t)) = \mathbf{0} \quad (25.183)$$

for  $\alpha^*$  as defined in (181). In other words, the non-orthogonality in (177) arose because  $\alpha^*$  was defined in terms of (182) but the expectation in terms of  $D(y_t, \mathbf{X}_t; \psi)$ .

The above discussion suggests that the instruments  $\mathbf{Z}_t$  are random variables jointly distributed with the  $y_t$  and  $\mathbf{X}_t$ , but treated (implicitly) as conditioning variables. In defining the statistical specification (176), however, we did not include  $\mathbf{Z}_t$  explicitly. That is, the natural statistical GM:

$$y_t = \alpha'_0 \mathbf{X}_t + \gamma'_0 \mathbf{Z}_t + u_t \quad (25.184)$$

where  $\alpha_0 = \Sigma_{2,3}^{-1}(\sigma_{12} - \sigma_{13} \Sigma_{33}^{-1} \Sigma_{32})'$  and  $\gamma_0 = \Sigma_{33}^{-1} \sigma_{31} - \Sigma_{33}^{-1} \Sigma_{32} \alpha_0$  where  $\Sigma_{2,3} = (\Sigma_{22} - \Sigma_{23} \Sigma_{33} \Sigma_{32})$  is not the parametrisation of interest. Instead  $\mathbf{Z}_t$  is marginalised out from the systematic component  $E(y_t / \sigma(\mathbf{X}_t, \mathbf{Z}_t))$ . In the present case this amounts to imposing the restriction  $\gamma_0 = \mathbf{0}$  or equivalently

$$\Sigma_{23} \gamma_0 = \mathbf{0} \quad (25.185)$$

since  $m \geq p$ . Given the relationship between  $\gamma_0$  and  $\alpha_0$  in (184), however, (185) implies that

$$\alpha_0 = (\Sigma_{23} \Sigma_{33}^{-1} \Sigma_{32})^{-1} \Sigma_{23} \Sigma_{33}^{-1} \sigma_{31} = \alpha^* \quad (25.186)$$

and (184) reduces to (176) with  $\alpha$  as defined in (181). This relationship between (184) and (176) enhances our understanding of the IV argument considerably. It shows that IV are indeed ‘omitted’ conditioning variables whose separate effect is of no interest in a particular formulation. If we were to estimate  $\alpha_0$  and  $\gamma_0$  in (184) the MLE’s would be the usual orthogonal projection estimator:

$$\hat{\alpha}_0 = (\mathbf{X}' \mathbf{M}_z \mathbf{X})^{-1} \mathbf{X}' \mathbf{M}_z \mathbf{y} \quad (25.187)$$

and

$$\hat{\gamma}_0 = (\mathbf{Z}' \mathbf{M}_x \mathbf{Z})^{-1} \mathbf{X}' \mathbf{M}_x \mathbf{y} = (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{y} - (\mathbf{Z}' \mathbf{Z})^{-1} (\mathbf{Z}' \mathbf{X}) \hat{\alpha}_0$$

$\mathbf{P}_z = \mathbf{Z}(\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}'$ ,  $\mathbf{M}_z = \mathbf{I} - \mathbf{P}_z$ . The estimator  $\hat{\alpha}_0$  of  $\alpha_0$ , however, is not an estimator of the parameters of interest  $\alpha^*$ . On the other hand, when  $\mathbf{Z}_t$  in (184) is dropped without taking into account its indirect effect (through the parametrisation) the estimator  $\hat{\alpha} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}$  is inappropriate. This is because  $\hat{\alpha}$  introduces a non-existent orthogonality between the estimated ‘systematic’ and ‘non-systematic’ components. That is, for  $\hat{\mu} = \mathbf{X} \hat{\alpha} = \mathbf{P}_x \mathbf{y}$  and  $\hat{\varepsilon} = \mathbf{y} - \mathbf{X} \hat{\alpha} = \mathbf{M}_x \mathbf{y}$  we can see that the sample equivalent to (176) induced by  $\hat{\alpha}$  is

$$\mathbf{y} = \mathbf{P}_x \mathbf{y} + \mathbf{M}_x \mathbf{y} \quad (25.188)$$

where  $\hat{\mu} \perp \hat{\varepsilon}$  given that  $\mathbf{P}_x \mathbf{M}_x = \mathbf{0}$ ,  $\mathbf{P}_x = \mathbf{X}(\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'$ ,  $\mathbf{M}_x = \mathbf{I} - \mathbf{P}_x$ . In view of (177), however, no such orthogonality between  $\mu_t$  and  $\varepsilon_t$  exists within the framework defined by  $D(y_t, \mathbf{X}_t; \psi)$  and thus  $\alpha$  is inappropriate. From (183) we can see that in order to achieve an orthogonality between  $\mu_t$  and  $\varepsilon_t$  we need to take account of the conditioning variables  $\mathbf{Z}_t$ . The way this was achieved in (184) was inappropriate not because we introduced a non-existent orthogonality (the orthogonality was valid) but because we achieved the orthogonality at the expense of the parametrisation of interest. The question which naturally arises at this stage is how we can introduce an orthogonality between  $\mu_t$  and  $\varepsilon_t$  without losing the parametrisation of interest.

Intuition suggests that when  $\mu_t = \alpha' \mathbf{X}_t$  and  $\varepsilon_t = y_t - \alpha' \mathbf{X}_t$  are not orthogonal it must be the case that  $\varepsilon_t$  includes information which ‘rightfully’ belongs to  $\mu_t$ . In the present context what is systematic and non-systematic information is determined by (182), the underlying distribution. This being the case the obvious way to proceed is to redefine both components so as to ensure that they are indeed systematic and non-systematic relative to  $D(y_t, \mathbf{X}_t / \mathbf{Z}_t; \theta)$ . This can be achieved by the new components:

$$\mu_t^* = E(\mu_t / \sigma(\mathbf{Z}_t)) \quad \text{and} \quad \varepsilon_t^* = \varepsilon_t - (\mu_t^* - \mu_t) \quad (25.189)$$

Note how the systematic information  $(\mu_t^* - \mu_t)$  is subtracted from  $\varepsilon_t$  to define the new error term  $\varepsilon_t^*$ . The redesigned form of (176) is

$$y_t = E(\mu_t / \sigma(\mathbf{Z}_t)) + [\mu_t - E(\mu_t / \sigma(\mathbf{Z}_t))] + \varepsilon_t^* \quad (25.190)$$

with  $E(\mu_t^* \varepsilon_t^*) = 0$ .

(190) is different from (184) in so far as the original parametrisation has been retained.

It is interesting to note that the above ‘re-design’ can be motivated

directly in terms of condition (i). Given that

$$E(\mathbf{Z}'_t \varepsilon_t) = E\{E[\mathbf{Z}'_t \varepsilon_t / \sigma(\mathbf{Z}_t)]\} = E\{\mathbf{Z}_t E(\varepsilon_t / \sigma(\mathbf{Z}_t))\}$$

(see Chapter 7) we can deduce that condition (i) holds if

$$E(\varepsilon_t / \sigma(\mathbf{Z}_t)) = 0 \quad (25.191)$$

which in turn is valid for

$$E(\mu_t / \sigma(\mathbf{Z}_t)) = E(y_t / \sigma(\mathbf{Z}_t)) \quad (25.192)$$

The conditions (191) and (192) are equivalent to (189) and thus the ‘choice’ of  $\mathbf{Z}_t$  so as to ensure condition (i) is equivalent to the above ‘re-design’ of (176).

In order to understand how the IV estimator (179) is indeed the most appropriate estimator of  $\alpha^*$  in the context of (176) let us derive the sample equivalent of (190). Using the well-known analogy between conditional expectations and orthogonal projections we can deduce the sample equivalents to (189) and (190) for  $t = 1, 2, \dots, T$  ( $T > m + p$ ):

$$\mu^* = \mathbf{P}_z \mathbf{X} \alpha^*, \quad \varepsilon^* = \varepsilon + \mathbf{M}_z \mathbf{X} \alpha_0$$

and

$$\mathbf{y} = \mathbf{P}_z \mathbf{X} \alpha^* + \mathbf{M}_z \mathbf{X} \alpha_0 + \varepsilon. \quad (25.193)$$

$\mathbf{P}_z$  and  $\mathbf{M}_z = \mathbf{I} - \mathbf{P}_z$  are the orthogonal projections onto the space spanned by the columns of  $\mathbf{Z}$  and its orthogonal complement, respectively. In view of the orthogonality between  $\mathbf{P}_z$  and  $\mathbf{M}_z$  ( $\mathbf{P}_z \mathbf{M}_z = \mathbf{0}$ ) the usual orthogonal (OLS) estimators of  $\alpha^*$  and  $\alpha_0$  coincide with (179) and (187), respectively.

This derivation of the IV estimator provides us with additional insight as to how the method ‘solves’ the original estimation problem. In a sense the IV method is a *two-stage* method (see Pagan (1986)) where the *first stage* ‘re-designs’ the original formulation so as to achieve orthogonality between the systematic and non-systematic components, without changing the parametrisation, and the *second stage* the parameters of interest are estimated using the usual orthogonal projection (OLS) estimator.

The re-designed formulation is particularly illuminating in so far as the relationship between (176) and (184) is concerned. If we rearrange (193) in the form:

$$\mathbf{y} = \mathbf{X} \alpha^* + \mathbf{M}_z \mathbf{X} (\alpha_0 - \alpha^*) + \varepsilon \quad (25.194)$$

we can see that for  $\delta = (\alpha_0 - \alpha^*)$  the hypothesis

$$H_0: \delta = \mathbf{0} \quad \text{against} \quad H_1: \delta \neq \mathbf{0}$$

can be easily tested in the context of (194) when  $T > m + p$ .  $H_0$  is often viewed as referring to the *admissibility* of the instruments in  $\mathbf{Z}_t$ . From (184)–

(186) we can see that  $H_0$  is an indirect test of the hypothesis that the coefficient of the conditioning variables  $Z_t$  is zero, i.e.

$$\text{cov}(y_t, Z_t/X_t) = 0$$

This brings out the connection between ‘omitted’ conditioning variables and instrumental variables (see Section 25.8 below).

In the context of the simultaneous equations model all the estimators discussed (2SLS, LIML, 3SLS and FIML) can be viewed as IV estimators; see the recent monograph by Bowden and Turkington (1984). In order to see this let us consider the case of the 2SLS (see Section 25.6 above).

The first restricted structural equation for the sample period  $t = 1, 2, \dots, T$  takes the form:

$$y_1 = Z_1 \alpha_1 + \varepsilon_1^* \quad (25.195)$$

where  $Z_1 \equiv (Y_1, X_1)$  and  $\alpha_1 = (\gamma'_1, \delta'_1)'$ . Given that the underlying distribution is  $D(y_t/X_t; \theta)$  where  $y_t \equiv (y_{1t}, y'_{1t}, y'_{(1)t})'$  and  $X_t \equiv (X'_{1t}, X'_{(1)t})'$  we can see that  $X_t$  is the set of instruments and (195) includes some of those instruments as genuine conditioning variables. Expressing (195) in the form (193) yields:

$$y_1 = P_x Y_1 \gamma_1 + X_1 \delta_1 + M_x Y_1 \gamma_1^0 + \varepsilon_1^* \quad (25.196)$$

since  $P_x X_1 = X_1$ . From this we can deduce that the IV estimator of  $\alpha_1$  is:

$$\begin{pmatrix} \hat{\gamma}_1 \\ \hat{\delta}_1 \end{pmatrix}_{IV} = \begin{pmatrix} Y'_1 P_x Y_1 & Y'_1 P_x X_1 \\ X'_1 P_x Y_1 & X'_1 X_1 \end{pmatrix}^{-1} \begin{pmatrix} Y'_1 P_x y_1 \\ X'_1 y_1 \end{pmatrix} \quad (25.197)$$

which coincides with the 2SLS estimator (140).

## 25.8 Misspecification testing

For theoretical as well as practical reasons misspecification testing in the context of the simultaneous equations model will be considered at three different but interrelated levels:

(i) the statistical system level, in terms of the statistical GM,

$$y_t = B' x_t + u_t, \quad t \in \mathbb{T}; \quad (25.198)$$

(ii) the unrestricted structural equation, in terms of the statistical GM,

$$y_{1t} = \gamma'_1 y_{1t} + \gamma'_{(1)} y_{(1)t} + \delta'_1 x_{1t} + \delta'_{(1)} x_{(1)t} + \varepsilon_{1t}, \quad t \in \mathbb{T}; \quad (25.199)$$

- (iii) the restricted structural equation, in terms of the theoretical parameters of interest,

$$Y_{1t} = \gamma'_1 Y_{1t} + \delta'_1 X_{1t} + \varepsilon_{1t}^*. \quad (25.200)$$

**(1) The statistical system level**

If we look at the specification of the simultaneous equations model given in Section 25.4 we can see that a sufficient condition for the theoretical parameters of interest  $\xi$  to be well defined is that the statistical parameters  $\theta$  are well defined. This suggests that the natural way to proceed with misspecification testing is to test assumptions [6] to [8] in the context of the multivariate linear regression model (see Chapter 24). In the case where the system defined by the structural form with the restrictions imposed is just identified, the mapping  $H(\cdot): \Theta \rightarrow \Xi$ ,  $\xi = H(\theta)$ , is one-to-one and onto, implying that the inverse image of  $H(\cdot)$  is  $\Theta$  itself. Hence,  $\xi$  is a simple reparametrisation and as well defined as  $\theta$ . In the case where the system is overidentified then the mapping  $H(\cdot)$  is many-to-one whose inverse image  $\Theta_0$  is a subset of  $\Theta$ . That is, in the case of overidentification  $H(\cdot)$  imposes restrictions on  $\Theta$  which are, however, testable in the context of (198) (see Section 25.9).

The above argument suggests that before any questions related to the theoretical parameters of interest  $\xi$  are considered we need to ensure that the statistical model in terms of the statistical parameters  $\theta$  is well defined. This is achieved by testing for departures from the underlying assumptions *using the procedures discussed in Chapter 24*. If this misspecification testing reveals that the statistical assumptions underlying (198) are valid we can proceed to consider the identification, estimation as well as specification testing in the context of the structural form reparametrisation. Otherwise we need to respecify the underlying statistical model. For example, the assumption of independence is likely to prove inappropriate in econometric modelling with time-series data. A respecification of the underlying statistical model will give rise to the multivariate dynamic linear regression model whose statistical GM is

$$y_t = B'_0 x_t + \sum_{i=1}^l A'_i y_{t-i} + \sum_{i=1}^l B'_i x_{t-i} + u_t, \quad t > l \quad (25.201)$$

(see Chapter 24). This suggests that if the identification problem was ‘solved’ in terms of  $\theta$  in (198) we need to reconsider it in terms of  $\theta^*$  in (199). Given that (198) and (199) coincide under

$$H_0: B_i = 0, \quad A_i = 0, \quad i = 1, 2, \dots, l, \quad (25.202)$$

we need to account for these implicitly imposed ‘testable’ restrictions. In the context of (198) the restrictions in (202) are viewed as ‘phoney’ restrictions which fail the rank condition (see Section 25.3) because all equations satisfy these restrictions. When  $H_0$  is tested and rejected, however, we need to reconsider our estimable model (see Chapters 1 and 26) in view of the fact that the original model allowed for no lags in the variables involved. When a situation like this arises in practice the modelling is commonly done equation by equation and not in terms of the system as a whole because of the relative uncertainty about which variables enter which equation. For this reason it is important to consider misspecification testing in terms of individual equations as well.

### **(2)      *The unrestricted single structural equation level***

From the theoretical viewpoint (194) is of no interest because no structural parameter is identified in its context. From the statistical viewpoint, however, (194) is of some interest because it can be viewed as a ‘bridge’ between (193) and (195) which can be used for misspecification testing purposes. Treating it as a hybrid of the linear and stochastic linear regression models (see Chapters 19–20) (194) presents us with no new problems as far as misspecification testing is concerned. The testing procedures proposed in Chapters 21 and 22 can be easily adapted for the present context.

### **(3)      *The restricted single structural equation level***

Having tested for misspecification in the context of (198) and accepted the underlying assumptions as valid we have a well-defined estimated statistical GM. From the theoretical viewpoint, however, (198) makes very little sense (if any). For that we need to reparametrise it (by imposing a priori restrictions) in terms of the theoretical parameters of interest. The question of how we achieve such a reparametrisation will be discussed in Section 25.9. At this stage we need to assume that this has been achieved by imposing the exclusion restrictions  $\gamma_{(1)} = \mathbf{0}$ ,  $\delta_{(1)} = 0$  (enough in number to identify the first equation) in order to get (200).

From the misspecification viewpoint the reparametrisation expressed in (200) is of interest in so far as this has not been achieved *at the expense* of the statistical properties of the estimated GM (198) we started with. That is, we need to check that (200) is not just a theoretically meaningful but also statistically well-defined relationship. The discerning reader would have noticed that the same problem arose in the context of the linear and dynamic linear regression models (see Chapter 23).

In Sections 25.6 and 25.7 it was argued that (200) can be viewed in various alternative ways which are useful for different purposes. The two interpretations we are interested in here are:

$$y_{1t} = E(y_{1t}/\sigma(y_{1t}), \mathbf{X}_{1t} = \mathbf{x}_{1t}) + \varepsilon_{1t}^* \quad (25.203)$$

and

$$y_{1t} = \gamma'_1 E(\mathbf{y}_{1t}/\mathbf{X}_t = \mathbf{x}_t) + \delta'_1 \mathbf{x}_{1t} + u_{1t}. \quad (25.204)$$

Given that

$$E(y_{1t}/\sigma(y_{1t}), \mathbf{X}_{1t} = \mathbf{x}_{1t}) = \gamma'_1 \mathbf{y}_{1t} + \delta'_1 \mathbf{x}_{1t}, \quad \varepsilon_{1t}^* = -\gamma'_1 \mathbf{u}_{1t} + u_{1t}$$

and

$$E(\mathbf{y}_{1t}/\mathbf{X}_t = \mathbf{x}_t) = \mathbf{B}'_{12} \mathbf{x}_{1t} + \mathbf{B}'_{22} \mathbf{x}_{(1)t},$$

we can see that we can go from (204) to (203) by subtracting  $\gamma'_1 \mathbf{y}_{1t}$  from both sides of (204). For estimation purposes (204) is preferable because of the orthogonality between the systematic and non-systematic components (see Section 25.6). For misspecification testing, however, it is more convenient to use (203).

The normality of  $D(y_{1t}/y_{1t}, \mathbf{X}_{1t}; \xi_1)$  can be tested using a direct extension of the skewness–kurtosis test discussed in Chapter 21 based on the residuals

$$\hat{\varepsilon}_{1t} = y_{1t} - \hat{\gamma}_{IV} \mathbf{y}_{1t} - \hat{\delta}_{IV} \mathbf{x}_{1t}, \quad t = 1, 2, \dots, T, \quad (25.205)$$

where  $\hat{\gamma}_{IV}, \hat{\delta}_{IV}$  refer to some IV estimator of  $\gamma_1$  and  $\delta_1$  respectively such as the 2SLS or LIML estimators.

The linearity of the conditional expectation

$$E(y_{1t}/\sigma(y_{1t}), \mathbf{X}_{1t} = \mathbf{x}_{1t}) = \alpha_1^{*'} \mathbf{Z}_{1t}, \quad (25.206)$$

where  $\mathbf{Z}_{1t} \equiv (\mathbf{y}'_{1t}, \mathbf{x}'_{1t})'$  and  $\alpha_1^{*'} \equiv (\gamma'_1, \delta'_1)$  can be tested using the *F*-type test discussed in the context of the linear regression model (see Chapter 21). That is, generate the higher-order terms  $\psi_t$ , using  $\mathbf{Z}_{1t}$  or powers of  $\hat{\mu}_t^*$ , and test the null hypothesis

$$H_0: \mathbf{d} = \mathbf{0} \quad \text{against} \quad H_1: \mathbf{d} \neq \mathbf{0}$$

in the auxiliary regression

$$\mathbf{y}_1 = \mathbf{Z}_1 \mathbf{z}_1^0 + \Psi \mathbf{d} + \varepsilon_1^0 \quad (25.207)$$

or

$$\hat{\varepsilon}_1^* = \mathbf{Z}_1 (\mathbf{z}_1^0 - \hat{\mathbf{z}}_{IV}) + \Psi \mathbf{d} + \mathbf{v} \quad (25.208)$$

(see Chapter 21). The augmented equation (207) or (208) should be estimated using an IV estimator, say 2SLS. Using the asymptotic normality of  $\hat{\mathbf{z}}_{IV}^*$  the *F*-type test procedure discussed in Chapter 21 can be justified asymptotically in the present context.

The homoskedasticity of the conditional variance

$$\text{Var}(y_{1t}/\sigma(y_{1t}), \mathbf{X}_{1t} = \mathbf{x}_{1t}) = v_{11} \quad (25.209)$$

can be tested by extending the White and related tests discussed in Chapter 21 to the present context. If we *assume* that no heteroskedasticity is present in the simultaneous equation model as a whole we can test (209) against

$$\text{Var}(y_{1t}/\sigma(y_{1t}), \mathbf{X}_{1t} = \mathbf{x}_{1t}) = h(\mathbf{Z}_{1t}), \quad (25.210)$$

using  $H_0: \mathbf{c}_1 = \mathbf{0}$  against  $H_1: \mathbf{c}_1 \neq \mathbf{0}$  in the auxiliary regression

$$\hat{\epsilon}_{1t}^{*2} = c_0 + \mathbf{c}'_1 \psi_t + v_t \quad (25.211)$$

(see Kelejian (1982)). This can be tested using the *F*-type test or its asymptotic variants (e.g.  $TR^2$ ) discussed in Chapter 21. In the case where the assumption related to the presence of heteroskedasticity in the rest of the system is not appropriate, we need to modify the *F*-type tests because the asymptotic distribution of  $\hat{\mathbf{c}} \equiv (\hat{c}_0, \hat{\mathbf{c}}_1)$  is not

$$\sqrt{x} T(\hat{\mathbf{c}} - \mathbf{c}) \sim N(0, \sigma^2 \mathbf{Q}_\psi^{-1}) \quad (25.212)$$

but

$$\sqrt{x} T(\hat{\mathbf{c}} - \mathbf{c}) \sim N(\mathbf{0}, 2\sigma^4 (\mathbf{Q}_\psi^{-1} + 2\mathbf{L}\mathbf{Q}_{z_1}^{-1}\mathbf{L}')), \quad (25.213)$$

where  $\mathbf{Q}_\psi$  and  $\mathbf{Q}_{z_1}$  refer to the probability limits of  $[\sum_{t=1}^T \psi_t^* \psi_t^{*'}]$ ,  $\psi_t^* = (1, \psi_t')'$ , and  $(\sum_{t=1}^T \hat{\mathbf{Z}}_{1t} \hat{\mathbf{Z}}_{1t}')$ ,  $\hat{\mathbf{Z}}_{1t} = (\hat{y}'_{1t}, \mathbf{x}'_{1t})$  respectively and  $\mathbf{L}$  is the matrix of coefficients of the auxiliary regression:

$$E(\mathbf{Z}_{1t} \epsilon_{1t}^*) = \mathbf{L}' \psi_t^* + v_t^* \quad (25.214)$$

(see Pagan and Hall (1983), White (1982a)).

An important assumption which should be tested in the present context is whether by imposing the exclusive restrictions on (198) the *parameter time invariance* no longer holds. The diagnostic graphs related to the recursive estimates of the coefficients as well as the  $\tau$ -observation windows discussed in Chapter 21 apply to the present case in terms of some IV estimator of  $\alpha_1^*$ . Moreover, the *F*-type tests related to time-dependence and structural change can be applied to equation (203) using some IV estimator for  $\alpha_1^*$ . Time-dependence of the parameters is particularly important in the present context because this formulation is commonly used for prediction and policy evaluation purposes. In cases where no misspecification tests have been applied to the statistical GM (198), testing for parameter time dependence is of paramount importance because heteroskedasticity in its context might appear as time dependence in the context of (204). This is because the parameters in (204) are defined in terms of  $\mathbf{B}$  and  $\Omega$  and if

$$\text{Cov}(\mathbf{y}_t / \mathbf{X}_t = \mathbf{x}_t) = \boldsymbol{\Omega}(\mathbf{x}_t), \quad (25.215)$$

the vector of coefficients  $\boldsymbol{\alpha}_1^*$  will be a function of  $\mathbf{x}_t$  via  $\boldsymbol{\Omega}(\mathbf{x}_t)$ .

The sampling model assumption of *independence* can be tested by applying the *F*-type test to test the significance of the coefficients of the lagged  $\mathbf{Z}_{1t}$  in

$$y_{1t} = \boldsymbol{\alpha}_1^0' \mathbf{Z}_{1t} + \sum_{i=1}^l \boldsymbol{\alpha}_i' \mathbf{Z}_{1t-i} + v_t \quad (25.216)$$

or

$$\hat{e}_{1t}^* = (\boldsymbol{\alpha}_1^0 - \hat{\boldsymbol{\alpha}}_{1v}^*)' \mathbf{Z}_{1t} + \sum_{i=1}^l \boldsymbol{\alpha}_i' \mathbf{Z}_{1t-i} + v_t, \quad l \geq 1. \quad (25.217)$$

The only modification of the *F*-type test discussed in Chapter 22 needed is to estimate the parameters of (216) or (217) by some IV estimator because of the presence of simultaneity. The *F*-type autocorrelation test is based on the auxiliary equation

$$\hat{e}_{1t}^* = (\boldsymbol{\alpha}_1^0 - \hat{\boldsymbol{\alpha}}_{1V}^*)' \mathbf{Z}_{1t} + \sum_{i=1}^l c_i' \hat{e}_{1t-i}^* + v_t, \quad l \geq 1 \quad (25.218)$$

(see Godfrey (1976), Breusch and Godfrey (1981)).

## 25.9 Specification testing

In practice, specification testing in the simultaneous equations model is almost exclusively considered in the context of the single identified structural equation

$$y_{1t} = \gamma_1' \mathbf{y}_{1t} + \delta_1' \mathbf{x}_{1t} + \varepsilon_{1t}, \quad t \in \mathbb{T}, \quad (25.219)$$

where  $\mathbf{y}_{1t}: (m_1 - 1) \times 1$ ,  $\mathbf{x}_{1t}: k_1 \times 1$ ,  $(k - k_1) \geq m_1 - 1$  (for identification).

### (1) Testing $H_0: \gamma_1 = \bar{\gamma}_1$ against $H_1: \gamma_1 \neq \bar{\gamma}_1$

(Anderson and Rubin (1949)). Under  $H_0$  the GM (219) takes the form

$$y_{1t}^* = \delta_1' \mathbf{x}_{1t} + \varepsilon_{1t}, \quad t \in \mathbb{T}, \quad (25.220)$$

where  $y_{1t}^* = y_{1t} - \bar{\gamma}_1' \mathbf{y}_{1t}$ . Under  $H_1$ , however,  $y_{1t}^* = y_{1t} - \gamma_1' \mathbf{y}_{1t}$  and a function of both  $\mathbf{x}_{1t}$  and  $\mathbf{x}_{(1)t}$  given that

$$y_{1t} = \mathbf{B}'_{12} \mathbf{x}_{1t} + \mathbf{B}'_{22} \mathbf{x}_{(1)t} + u_{1t}. \quad (25.221)$$

Hence,  $H_0$  can be tested indirectly by testing that the coefficient of  $\mathbf{X}_{(1)}$  is zero in the regression equation

$$\mathbf{y}_1^* = \mathbf{X}_1 \boldsymbol{\beta}_1^* + \mathbf{X}_{(1)} \boldsymbol{\beta}_{(1)}^* + v_1. \quad (25.222)$$

This can be tested using the  $F$ -type test statistic

$$FT(\mathbf{y}_1^*) = \left( \frac{T-k}{k_2} \frac{\mathbf{y}_1^{*'} (\mathbf{M}_{X_1} - \mathbf{M}_X) \mathbf{y}_1^*}{\mathbf{y}_1^{*'} \mathbf{M}_X \mathbf{y}_1^*} \right)^{H_0} \sim F(k_2, T-k), \quad (25.223)$$

where  $\mathbf{M}_X = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  and  $\mathbf{M}_{X_1} = \mathbf{I} - \mathbf{X}_1(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1$ , the rejection region being

$$C_1 = \{ \mathbf{y}_1^* : FT(\mathbf{y}_1^*) > c_\alpha \}, \quad \alpha = \int_{c_\alpha}^{\infty} dF(k_2, T-k). \quad (25.224)$$

A special case of  $H_0$  above of particular interest is  $\gamma_1 = \mathbf{0}$ , i.e. the endogenous variables are insignificant in (219).

This test can be readily extended to include some of the coefficients in  $\delta_1$  as long as all the coefficients in  $\gamma_1$  are specified under  $H_0$ . The difficulty with including only a subset of  $\gamma_1$  in  $H_0$  is that we cannot apply the usual linear regression estimator under  $H_0$ , we need to apply an IV estimator such as 2SLS or LIML. The most convenient way to extend  $H_0$  to a more general set of linear restrictions on  $\alpha^* \equiv (\gamma'_1, \delta'_1)'$  is to use the asymptotic distribution of a consistent and asymptotically normal estimator.

## (2) Testing $H_0: \mathbf{R}\alpha^* = \mathbf{r}$ against $H_1: \mathbf{R}\alpha^* \neq \mathbf{r}$

where  $\mathbf{R}$  and  $\mathbf{r}$  are  $p \times (m_1 + k_1 - 1)$  and  $p \times 1$  known matrices with  $\text{rank}(\mathbf{R}) = p$ . Using the asymptotic normality of  $\hat{\alpha}_{2SLS}^*$  (or  $\hat{\alpha}_{LIML}^*$ ),

$$\sqrt{\frac{T}{\alpha}} (\hat{\alpha}_{2SLS}^* - \alpha^*) \sim N(0, v_{11} \mathbf{D}_{11}^{-1}) \quad (25.225)$$

(see Section 25.6), we can use the same intuitive argument as in the case of the  $F$ -test (see Chapter 20) to suggest that a natural choice for a test statistic in the present context is

$$FT(\mathbf{y}_1) = \frac{(\mathbf{R}\hat{\alpha}_{2SLS}^* - \mathbf{r})' [\mathbf{R}(\hat{\mathbf{Z}}'_1 \hat{\mathbf{Z}}_1)^{-1} \mathbf{R}']^{-1} (\mathbf{R}\hat{\alpha}_{2SLS}^* - \mathbf{r})}{\hat{v}_{11}^*} \left( \frac{1}{p} \right), \quad (25.226)$$

where  $\hat{\mathbf{Z}}_1 \equiv (\hat{\mathbf{Y}}_1 : \hat{\mathbf{X}}_1)$ . Asymptotically we can argue that

$$pFT(\mathbf{y}_1) \underset{\alpha}{\overset{H_0}{\sim}} \chi^2(p). \quad (25.227)$$

In practice it is preferable to use the  $F$ -type form (226) based on the approximation

$$FT(\mathbf{y}_1) \underset{\text{appx}}{\overset{H_0}{\sim}} F(p, T-k), \quad (25.228)$$

with a rejection region similar to (224) above.

The above results suggest that asymptotically the specification testing results discussed in Chapter 19 in the context of the linear regression model can be extended to the present case. In particular, individual ( $t$ -tests) and joint tests of significance are asymptotically justified. Two special cases of linear homogeneous restrictions which can be tested using the above  $F$ -type test are:

- (i) the overidentifying restrictions; and
- (ii) the exogeneity restrictions;

which we consider next.

### (3) Testing the overidentifying restrictions

As argued in Section 25.3 the identification restrictions in the case of exclusion restrictions come in the form of the system (51)–(52). For identification we need  $(k - k_1) \geq (m_1 - 1)$ . In the case of overidentification the overidentifying restrictions  $q = k - k_1 - m_1 + 1$  can be tested via (51)–(52). Given that (51) is not directly involved in the overidentification the latter conditions can be tested using (52), i.e.

$$H_0: \mathbf{B}_{22}\gamma_1 - \boldsymbol{\beta}_{21} = \mathbf{0}, \quad H_1: \mathbf{B}_{22}\gamma_1 - \boldsymbol{\beta}_{21} \neq \mathbf{0}.$$

A convenient formulation which enables us to test  $H_0$  is the statistical sub-model:

$$y_{1t} = \boldsymbol{\beta}'_{11}\mathbf{x}_{1t} + \boldsymbol{\beta}'_{21}\mathbf{x}_{(1)t} + u_{1t}$$

$$\mathbf{y}_{1t} = \mathbf{B}'_{12}\mathbf{x}_{1t} + \mathbf{B}'_{22}\mathbf{x}_{(1)t} + \mathbf{u}_{1t}$$

Multiplying the second equation by  $\gamma'_2$  and subtracting from the first we can define the *unrestricted structural form* of  $y_{1t}$  as:

$$y_{1t} = \gamma'_1 y_{1t} + (\boldsymbol{\beta}_{11} - \mathbf{B}_{12}\gamma_1)' \mathbf{x}_{1t} + (\boldsymbol{\beta}_{21} - \mathbf{B}_{22}\gamma_1)' \mathbf{x}_{(1)t} + \varepsilon_{1t}. \quad (25.229)$$

This form presents itself as the natural formulation in the context of which  $H_0$  can be tested. If we compare (229) with *restricted structural form* (219) we can derive the auxiliary regression:

$$\varepsilon_{1t}^* = (\boldsymbol{\beta}_{11} - \mathbf{B}_{12}\gamma_1 - \boldsymbol{\delta}_1)' \mathbf{x}_{1t} + (\boldsymbol{\beta}_{21} - \mathbf{B}_{22}\gamma_1)' \mathbf{x}_{(1)t} + \varepsilon_{1t}$$

or its operational form:

$$\hat{\varepsilon}_{1t}^* = \boldsymbol{\delta}_1^* \mathbf{x}_{1t} + \boldsymbol{\delta}_{(1)}^* \mathbf{x}_{(1)t} + v_t. \quad (25.230)$$

The obvious way to test  $H_0$  is to use the  $R^2$  from (230) to define the LM test statistic

$$LM = TR^2 \stackrel{H_0}{\sim} \chi^2(q). \quad (25.231)$$

An asymptotically equivalent form suggested by Bassman (1960) as better for small  $T$  is:

$$FT = \frac{RRSS - URSS}{URSS} \left( \frac{T-k}{q} \right) \stackrel{H_0}{\underset{\alpha}{\sim}} F(q, T-k) \quad (25.232)$$

where

$$RRSS = (\mathbf{y}_1 - \mathbf{y}_1 \hat{\gamma}_{IV} - \mathbf{X}_1 \boldsymbol{\delta}_{IV})' (\mathbf{y}_1 - \mathbf{Y}_1 \hat{\gamma}_{IV} - \mathbf{X}_1 \hat{\boldsymbol{\delta}}_{IV})$$

$$URSS = (\mathbf{y}_1 - \mathbf{Y}_1 \hat{\gamma}_{IV})' (\mathbf{M}_x - \mathbf{M}_{x_1}) (\mathbf{y}_1 - \mathbf{Y}_1 \hat{\gamma}_{IV}).$$

The main problems with (231) and (232) are:

(i) when  $H_0$  is rejected the tests provide no information as to the likely source of the problem, and

(ii) for a large  $k$  their power is likely to be low (see Section 19.5).

In an attempt to alleviate the problem of low power and get some idea about the likely source of any rejection, the modeller could choose  $q$  instruments at a time. That is, for  $\mathbf{X}_{(1)} \equiv (\mathbf{X}_2; \mathbf{X}_3)$  where  $\mathbf{X}_2: T \times q$  we could consider the following augmented form of (219) for  $t=1, 2, \dots, T$ :

$$\mathbf{y}_1 = \mathbf{Y}_1 \gamma_1^* + \mathbf{X}_1 \boldsymbol{\delta}_1^* + \mathbf{X}_2 \boldsymbol{\delta}_2^* + \varepsilon_1, \quad (25.233)$$

where  $\mathbf{X}_1: T \times k_1$  and  $\mathbf{X}_2: T \times q$ . Using (233) the test for the overidentifying exclusion restrictions takes the form

$$H_0: \boldsymbol{\delta}_2^* = \mathbf{0} \quad \text{against} \quad H_1: \boldsymbol{\delta}_2^* \neq \mathbf{0}.$$

The  $F$ -type test statistic for this hypothesis is

$$FT_{IV}(\mathbf{y}_1) = \left( \frac{RRSS_{IV} - URSS_{IV}}{URSS_{IV}} \right) \left( \frac{T-k}{q} \right), \quad (25.234)$$

where

$$RRSS_{IV} = (\mathbf{y}_1 - \mathbf{Y}_1 \hat{\gamma}_{1IV} - \mathbf{X}_1 \hat{\boldsymbol{\delta}}_{1IV})' (\mathbf{y}_1 - \mathbf{Y}_1 \hat{\gamma}_{1IV} - \mathbf{X}_1 \hat{\boldsymbol{\delta}}_{1IV})$$

and

$$URSS_{IV} = (\mathbf{y}_1 - \mathbf{Y}_1 \hat{\gamma}_{1IV}^* - \mathbf{X}_1 \hat{\boldsymbol{\delta}}_{1IV}^* - \mathbf{X}_2 \hat{\boldsymbol{\delta}}_{2IV}^*)'$$

$$\times (\mathbf{y}_1 - \mathbf{Y}_1 \hat{\gamma}_{1IV}^* - \mathbf{X}_1 \hat{\boldsymbol{\delta}}_{1IV}^* - \mathbf{X}_2 \hat{\boldsymbol{\delta}}_{2IV}^*),$$

and IV refers to some instrumental variables estimator such as 2SLS or LIML of the parameter in question. Asymptotically, we can argue by analogy to the linear regression case that

$$qFT_{IV}(\mathbf{y}_1) \stackrel{H_0}{\underset{\alpha}{\sim}} \chi^2(q).$$

In practice, however, it might be preferable to use the  $F$ -type approximation:

$$FT_{IV}(\mathbf{y}_1) \stackrel{H_0}{\underset{\text{aprx}}{\sim}} F(q, T-k), \quad (25.235)$$

with the usual rejection region. For a readable survey of the finite sample results related to the test statistics (226) and (234) see Phillips (1983).

The above tests can be generalised directly to the system as a whole using either the 3SLS formulation in conjunction with the *F*-type tests or the FIML formulation using the likelihood ratio test procedure.

#### (4) Testing for exogeneity

In the specification of (219) above,  $\mathbf{y}_{1t}$  was treated as a random vector of endogenous variables by defining the systematic component as the conditional expectation of  $y_{1t}$  on the  $\sigma$ -field generated by  $\mathbf{y}_{1t}$ , ( $\sigma(\mathbf{y}_{1t})$ ) and the observed value of  $\mathbf{X}_{1t}$ , ( $\mathbf{X}_{1t} = \mathbf{x}_{1t}$ ). This asymmetric treatment of  $\mathbf{y}_{1t}$  and  $\mathbf{X}_{1t}$  was used as the main implication of the endogenous–exogenous distinction. Testing for exogeneity in the present context refers to the possibility of treating  $\mathbf{y}_{1t}$ , or some subset of it in the same way as  $\mathbf{X}_{1t}$ . This amounts to testing whether the stochastic component of  $\mathbf{y}_{1t}$  contributes significantly to the equation (219).

A convenient formulation for testing the appropriateness of this asymmetric treatment of  $\mathbf{y}_{1t}$  is the equation (194) suggested in the context of IV estimation. In the present context (194) takes the form:

$$\mathbf{y}_1 = \mathbf{Y}_1 \gamma_1 + \mathbf{X}_1 \delta_1 + \mathbf{M}_x \mathbf{Y}_1 (\gamma_1^* - \gamma_1) + \varepsilon_1^*. \quad (25.236)$$

In this formulation we can see that  $\mathbf{y}_{1t}$  and  $\mathbf{x}_{1t}$  in (219) are treated similarly when  $\gamma_1^* - \gamma_1 = 0$ . Hence, an obvious way to parametrise the ‘endogeneity’ of  $\mathbf{y}_{1t}$  is in the form of:

$$H_0: (\gamma_1^* - \gamma_1) = 0, \quad H_1: (\gamma_1^* - \gamma_1) \neq 0. \quad (25.237)$$

This can be tested using the *F*-type test statistic:

$$FT = \frac{RRSS - URSS}{URSS} \left( \frac{T - 2(m_1 - 1) - k_1}{m_1 - 1} \right)^{H_0} \sim F(m_1 - 1, T - 2(m_1 - 1) + k_1) \quad (25.238)$$

where *RRSS* and *URSS* refer to the *RSS* from (219) and (236) respectively, both estimated by OLS (see Wu (1973), Hausman (1978), *inter alia*).

The above test can be easily extended to the case where only the exogeneity of a subset of the variables in  $\mathbf{y}_{1t}$  is tested by re-arranging (236)

$$\mathbf{y}_1 = \mathbf{P}_x \mathbf{Y}_1^* \gamma_1^* + \mathbf{Y}_2^* \gamma_2^* + \mathbf{X}_1 \delta_1 + \mathbf{M}_x \mathbf{Y}_2^* (\gamma_2^0 - \gamma_2^*) + \mathbf{M}_x \mathbf{Y}_1^* \gamma_1^0 + \varepsilon_1^* \quad (25.239)$$

(see Hausman and Taylor (1981) for a similar test).

As a conclusion to this section it is important to emphasise that all the above specification tests will be very sensitive to any departures from assumptions [1]–[8] (see Section 25.4). If any of these assumptions are invalid the conclusions of the above tests will be at best misleading. In particular the above ‘exogeneity’ test is likely to be inappropriate in cases where the independent sample assumption is invalid; see Section 25.7.

### 25.10 Prediction

If we assume that the observed value  $\mathbf{x}_{T+l}$  of the exogenous random vector  $\mathbf{X}_{T+l}$  is available, the obvious way to predict  $\mathbf{y}_{T+l}$  is to use the statistical GM,

$$\mathbf{y}_t = \mathbf{B}'\mathbf{x}_t + \mathbf{u}_t, \quad t \in \mathbb{T}, \quad (25.240)$$

in conjunction with a ‘good’ estimator of the unknown parameters  $\boldsymbol{\theta} \equiv (\mathbf{B}, \boldsymbol{\Omega})$ . The problem, however, is that (240) can also be expressed (parametrised) in terms of the theoretical parameters of interest  $\boldsymbol{\xi}$ , i.e.

$$\mathbf{y}_t = \mathbf{B}(\boldsymbol{\xi})'\mathbf{x}_t + \mathbf{u}_t, \quad t \in \mathbb{T}, \quad (25.241)$$

and the MLE of  $\boldsymbol{\theta}$  and  $\boldsymbol{\xi}$  coincide only in the case where the system of structural equations (with the identifying restrictions imposed),

$$\Gamma^*'\mathbf{y}_t + \Delta^*'\mathbf{x}_t + \varepsilon_t^* = \mathbf{0}, \quad (25.242)$$

*is just identified.* Otherwise, if the system is overidentified the FIML (or 3SLS) estimator of  $\boldsymbol{\xi}$  has a smaller variance than  $\hat{\boldsymbol{\theta}}$ .

This discussion suggests that in the case where the system is just identified we apply the prediction methods discussed in the context of the multivariate linear regression model (see Chapter 24). In the case of overidentification, however, we derive the restricted estimator of the statistical parameters  $\mathbf{B}$  via

$$\mathbf{B}(\hat{\boldsymbol{\xi}}) = -\Delta(\hat{\boldsymbol{\xi}})\Gamma(\hat{\boldsymbol{\xi}})^{-1}, \quad (25.243)$$

where  $\Delta(\hat{\boldsymbol{\xi}})$  and  $\Gamma(\hat{\boldsymbol{\xi}})$  refer to some asymptotically efficient estimator  $\hat{\boldsymbol{\xi}}$  of the structural parameters  $\boldsymbol{\xi}$ , such as FIML or 3SLS. It can be shown that in the overidentified case:

$$\sqrt{T}(\mathbf{B}(\hat{\boldsymbol{\xi}}) - \mathbf{B}(\boldsymbol{\xi})) \underset{\alpha}{\sim} N(\mathbf{0}, \mathbf{F}_3), \quad (25.244)$$

$$\sqrt{T}(\hat{\mathbf{B}} - \mathbf{B}) \underset{\alpha}{\sim} N(\mathbf{0}, \mathbf{F}_0), \quad (25.245)$$

with  $(\mathbf{F}_0 - \mathbf{F}_3)$  being positive semi-definite.

In the case where  $\xi$  is estimated by a single equation estimation method such as LIML or 2SLS then

$$\sqrt{T}(\mathbf{B}(\hat{\xi}_{IV}) - \mathbf{B}(\xi)) \underset{x}{\sim} N(\mathbf{0}, \mathbf{F}_2), \quad (25.246)$$

and both  $(\mathbf{F}_0 - \mathbf{F}_2)$  and  $(\mathbf{F}_2 - \mathbf{F}_3)$  are positive semi-definite (see Dhrymes (1978)). It is not very difficult to see that this asymptotic efficiency in estimation carries over to prediction as well given that the prediction error defined by

$$(\hat{\mathbf{y}}_{T+l} - \mathbf{y}_{T+l}) = (\hat{\mathbf{B}} - \mathbf{B})' \mathbf{x}_{T+l} + \mathbf{u}_{T+l} \quad (25.247)$$

is a function of the difference  $(\hat{\mathbf{B}} - \mathbf{B})$  and  $\Omega$ .

### ***Important concepts***

Reparametrisation, theoretical (structural) parameters of interest, simultaneity, overparametrisation, recursive system of equations, identification, linear homogeneous restrictions, exclusion restrictions, order and rank conditions for identification, underidentification, just and overidentification, indirect MLE, estimator generating equation, full information maximum likelihood (FIML), limited information maximum likelihood (LIML), two-stage least squares (2SLS),  $k$ -class estimator, non-central Wishart distribution, instrumental variables, three-stage least-squares (3SLS).

### ***Questions***

1. Compare the multivariate linear regression and the simultaneous equations models.
2. Discuss the relationship between statistical and structural parameters as well as structural and theoretical parameters of interest.
3. Show how the identification problem arises because of the overparametrisation induced by the structural formulation.
4. ‘In the context of the statistical GM

$$y_{1t} = \Gamma_1^0 y_t^{(1)} + \Delta_1' \mathbf{x}_t + \varepsilon_{1t}$$

of Section 25.2 no endogeneity problem arises and the usual orthogonal estimator (OLS) is the appropriate estimator to use.’ Explain.

5. ‘Endogeneity arises because of the violation of the variation free condition of exogeneity induced by the identifying restrictions.’ Discuss.

656      **The simultaneous equations model**

6. Explain what we mean by a recursive system of simultaneous equations and discuss the estimation of the parameters of interest.
7. ‘Identification refers to the uniqueness of the reparametrisation from the statistical to the theoretical parameters of interest.’ Discuss.
8. Explain the order and rank conditions of identification in the case of linear homogeneous restrictions.
9. Discuss the two cases where the order condition might be satisfied but the rank condition fails.
10. ‘Before we could even discuss the identification problem we need to have a well-defined statistical GM.’ Discuss.
11. Explain the concept of an indirect MLE of the theoretical parameters of interest. Under what circumstances is the IMLE fully efficient?
12. Explain the intuition underlying the concept of an estimator generating equations (EGE).
13. How do restrictions of the form  $\mathbf{B}\boldsymbol{\Gamma} + \boldsymbol{\Delta} = \mathbf{0}$  differ in the context of the simultaneous equations and multivariate linear regression model?
14. ‘The structural parameters estimators such as 2SLS, IV, LIML,  $k$ -class, 3SLS are numerical approximations of the FIML estimator.’ Discuss (see Hendry (1976)).
15. Explain the derivation of the 2SLS estimator of  $\boldsymbol{\alpha}_1^*$  in

$$\mathbf{y}_1 = \mathbf{Z}_1 \boldsymbol{\alpha}_1^* + \boldsymbol{\varepsilon}_1^*,$$

both as constrained least-squares as well as an instrumental variables estimator.

16. ‘The 2SLS estimator can be viewed as being a numerical approximation to the LIML estimator.’ Discuss.
17. ‘If the rank condition for identification is not satisfied then 2SLS does not exist.’ Discuss.
18. ‘The 3SLS estimator is derived by using the Zellner formulation in the context of the simultaneous equation model.’ Explain.
19. ‘Misspecification testing in the context of the simultaneous equation model coincides with that in the context of the multivariate linear regression model.’ Discuss.
20. Consider the question of misspecification testing in the context of a single structural equation before and after the identifying restrictions have been imposed.
21. Explain how you could construct a test for the overidentifying restrictions in the context of a single structural equation.
22. Discuss the question of testing for the exogeneity of a subset of the endogenous variables included in the first equation.

23. 'Using the derived statistical parameter estimators defined via

$$\mathbf{B}(\hat{\xi}) = -\Delta(\hat{\xi})\Gamma(\hat{\xi})^{-1}$$

will provide more efficient predictors for  $\mathbf{y}_{T+l}$ .' Discuss.

### Exercises

1. Consider the two equation estimable model:

$$m_t = \alpha_{11} + \alpha_{12}i_t + \alpha_{13}p_t + \alpha_{14}y_t; \quad (1)$$

$$i_t = \alpha_{21} + \alpha_{22}m_t + \alpha_{23}p_t + \alpha_{24}g_t. \quad (2)$$

- (i) Express equations (1) and (2) in the forms

$$\mathbf{y}_t = \mathbf{B}'\mathbf{x}_t + \mathbf{u}_t;$$

$$\Gamma^*\mathbf{y}_t + \Delta^*\mathbf{x}_t + \varepsilon_t^* = 0.$$

- (ii) Derive the parameters  $\Gamma^*$  and  $\Delta^*$  in terms of  $(\mathbf{B}, \Omega)$  and show that the theoretical parameters of interest are uniquely defined.

2. Verify that in the case of a recursive simultaneous equations model

$$y_{it} = \gamma_i^0 \mathbf{y}_{i-1t}^0 + \delta_i' \mathbf{x}_t + \varepsilon_{it}, \quad i = 1, 2, \dots, m, \quad t \in \mathbb{T},$$

the orthogonal estimators (where  $\boldsymbol{\alpha}_i^0 \equiv (\gamma_i^0, \delta_i')'$ ,  $\mathbf{Z}_i \equiv (\mathbf{y}_{i-1}, \mathbf{X})$ ):

$$\hat{\boldsymbol{\alpha}}_i^0 = (\mathbf{Z}_i' \mathbf{Z}_i)^{-1} \mathbf{Z}_i' \mathbf{y}_i$$

and

$$\hat{v}_{ii} = \frac{1}{T} \hat{\mathbf{e}}_i' \hat{\mathbf{e}}_i, \quad \hat{\mathbf{e}}_i = \mathbf{y}_i - \mathbf{Z}_i \hat{\boldsymbol{\alpha}}_i^0, \quad i = 1, 2, \dots, m$$

are indeed the MLE's of  $\boldsymbol{\alpha}_i^0$  and  $v_{ii}$ .

3. Consider the following structural equations with the exclusion restrictions imposed:

$$\begin{aligned} -y_{1t} + \gamma_{21}y_{2t} &+ \gamma_{41}y_{4t} + \delta_{11}x_{1t} &+ \delta_{31}x_{3t} &= \varepsilon_{1t}; \\ -y_{2t} &+ \gamma_{32}y_{3t} &+ \delta_{12}x_{1t} + \delta_{22}x_{2t} &= \varepsilon_{2t}; \\ \gamma_{13}y_{1t} &+ y_{3t} &+ \gamma_{43}y_{4t} + \delta_{13}x_{1t} &+ \delta_{33}x_{3t} = \varepsilon_{3t}; \\ \gamma_{14}y_{1t} + \gamma_{24}y_{2t} &- y_{4t} &+ \delta_{14}x_{1t} + \delta_{24}x_{2t} &= \varepsilon_{4t}. \end{aligned}$$

- (i) Discuss the identifiability of the above equations under the following conditions:

- (a) No additional restrictions;
- (b)  $\delta_{24} = 0, \delta_{22} = 0$ ;
- (c)  $\gamma_{32} = 0, \gamma_{41} = 0$ .

- (ii) Explain how you would estimate the first equation by 2SLS.

658      **The simultaneous equations model**

4. Show that in the case of a just identified equation the 2SLS and IMLE estimators coincide.
5. Show that  $\Gamma(\tilde{\xi})' \tilde{\Omega} \Gamma(\tilde{\xi}) = (1/T) A(\tilde{\xi})' Z' Z A(\tilde{\xi})$  in Section 25.5.
6. Derive the 2SLS estimator of  $\alpha_1^*$  in  $y_1 = Z_1 \alpha_1^* + \varepsilon_1^*$  and explain why  $\hat{v}_{11}^* = (1/T) \tilde{\varepsilon}_1^* \tilde{\varepsilon}_1^*$  where  $\tilde{\varepsilon}^* = (y - \hat{Y}_1 \hat{\gamma}_{2SLS} - X_1 \hat{\delta}_{2SLS})$  is an inconsistent estimator of  $v_{11}^*$ .
7. Compare your answer in 6 with the derivation of the 2SLS estimator as an instrumental variables estimator.
8. ‘The GM  $y_1 = P_x Z_1 \alpha_1^* + \varepsilon_1^{**}$  (see Section 25.6) can be viewed as a sample equivalent to  $y_{1t} = y_1' E(y_{1t}/\sigma(X_t)) + \delta_1' x_{1t} + \varepsilon_{1t}^{**}$  for the sample period  $t = 1, 2, \dots, T$ .’ Explain.
9. Derive a test for the exogeneity of a subset  $y_{2t}$  of  $y_{1t}$  in the context of the single equation

$$y_{1t} = y_1' y_{1t} + \delta_1' x_{1t} + \varepsilon_{1t}^*.$$

**Additional references**

Anderson (1982); Hausman (1983); Judge *et al.* (1982).

## CHAPTER 26

---

### Epilogue: towards a methodology of econometric modelling

---

#### 26.1 A methodologist's critical eye

The purpose of this chapter is to formalise the methodology sketched in Chapter 1 using the concepts and procedures developed and discussed in the intervening chapters. The task of writing this chapter has become considerably easier since the publication of Caldwell (1982) which provides a lucid introduction to the philosophy of science for economists and establishes the required terminology. Indeed, the chapter can be seen as a response to Caldwell's challenge in his discussion of possible alternative approaches to economic methodology:

... One approach which to my knowledge has been completely ignored is the integration of economic methodology and philosophy with econometrics. Methodologists have generally skirted the issue of methodological foundations of econometric theory, and the few econometricians who have addressed philosophical issues have seldom gone beyond gratuitous references to such figures as Feigl or Carnap....

(See *ibid*, p. 216.)

In order to avoid long digressions into the philosophy of science the discussion which follows assumes that the reader has some basic knowledge of philosophy of science at the level covered in the first five chapters of Caldwell (1982) or Chalmers (1982).

Let us begin the discussion by considering the textbook econometric methodology criticised in Chapter 1 from a philosophy of science perspective. Any attempt to justify the procedure given in Fig. 1.1 reveals a deeply rooted influence from the logical positivist tradition of the late 1920s early 30s. Firstly, the preoccupation of logical positivism with criteria of cognitive significance, and in particular their verifiability criterion, is clearly

discernible in defining the intended scope of econometrics as 'the measurement of theoretical relationships'. A theory or a theoretical concept was considered meaningful in so far as it can be verified by observational evidence. Secondly, the treatment of the observed data as not directly related to the specification of the statistical model was based on the logical positivists' view that observed data represent 'objective facts' and any theory which does not 'comply' with the facts was rendered meaningless. Unless the theoretical model, now re-interpreted in terms of the observational language, differs from the statistical model only by some non-systematic effects (white-noise error) the theory had no meaning. Moreover, there was no need to bring the actual DGP into the picture because the theory could only have cognitive significance if it constitutes a description of the former. Thirdly, emphasis in the textbook methodology tradition is placed on testing theories (testability criterion of cognitive significance) and choosing between theories on empirical grounds. Despite such an emphasis, however, to my knowledge no economic theory was ever abandoned because it was rejected by some empirical econometric test, nor was a clear-cut decision between competing theories made in lieu of the evidence of such a test.

From the philosophy of science viewpoint the textbook econometric methodology largely ignored the later reformulations of logical positivism in the shape of logical empiricism in the late 1950s early 60s (see Caldwell (1982)). For example, the distinction between theoretical and observational concepts was never an issue in econometric modelling, adhering to the logical positivist view that the two should coincide for the former to have any meaning. Moreover, the later developments in philosophy of science challenging every aspect of logical positivism and in particular the verification principle and objectivity, as well as the incorrigibility of observed data (see Chalmers (1982)), are yet to reach the textbook econometric methodology. The structure of theories in economics has been influenced by the axiomatic hypothetico-deductive formulation of logical empiricism but this served to complicate the implementation of the textbook econometric methodology even further. Research workers found themselves having to use not only 'illegitimate' statistical procedures but also 'illegitimate' theory construction procedures. That is, they would begin with a very 'vague' but 'estimable' theoretical model and after estimating a variety of similar equations and a series of 'illegitimate' statistical techniques end up with the 'best' (in some sense) and return to revise the theoretical model in order to rationalise the 'best' empirical equation. Commonly, research workers find themselves having to include lagged variables in their estimated equations (because of the nature of the observed data used), even though the original theoretical model could not account

for them. Often, research workers felt compelled, and referees encouraged them, not to report their modelling strategy but to devise a new theoretical hypothetico-deductive model using some form of dynamic optimisation method and pretend that that was their theoretical model all along (see Ward (1972)). As argued in Chapter 1, research workers are driven to use ‘illegitimate’ procedures because of the ‘straightjacket’ the textbook methodology forces them to wear. Moreover, most of these ‘illegitimate’ procedures become the natural way to proceed in the context of the alternative methodology sketched in Chapter 1. In order to see this let us consider a brief formalisation of this methodology in the context of the philosophy of science.

## 26.2 Econometric modelling, formalising a methodology

Having criticised the textbook econometric methodology for being deeply rooted in an outdated philosophy of science the task of formalising an alternative methodology would have been easier if the new methodology could be founded on the most recent accepted view in philosophy of science. Unfortunately (or fortunately) no such generally accepted view has emerged since the dethronement of the positivist (logical positivism as well as logical empiricism) philosophy (see Caldwell (1982), Chalmers (1982), Suppe (1977), *inter alia*). The discussion which follows, purporting to bring together the various threads of methodological arguments in the book, cannot be related to any one school of thought in the current philosophy of science discussions aspiring to become the new orthodoxy. At best it could be seen as the product of a cross-fertilisation between some of the current views on the structure, status and function of theories and the particular features of economic theory and the associated observed data. For expositional purposes the discussion will be related to Fig. 1.2 briefly discussed in Chapter 1.

The concept of the *actual data generation process* (DGP) is used to designate the phenomenon of interest which a theory purports to explain. The concept is used in order to emphasise the intended scope of the theory as well as the source of the observable data. Defined this way, the concept of the actual DGP might be a real observable phenomenon or an experimental-like situation depending on the intended scope of the theory. For example, in the case of the demand schedule discussed in Chapter 1, if the intended scope of the theory is to determine a relationship between a hypothetical range of prices and the corresponding intentions to buy by a group of economic agents, at some particular point in time, the actual DGP might be used to designate the experimental-like situation where such data could be generated. On the other hand, if the intended scope of the theory is

to explain observed quantity or/and price changes over time, then the actual DGP should refer to the actual market process giving rise to the observed data. It should be noted that the intended scope of the theory is used to determine the choice of the observable data to be used. This will be taken up in the discussion of the theory-dependence of observation.

*A theory* is defined as a conceptual construct which purports to provide an idealised description of the phenomena within its intended scope (the actual DGP). A theory is not intended as a carbon copy of 'reality' providing an exact description of the observable phenomena in its intended scope. Economic 'reality' is much too complicated for such an exact copy to be comprehensible and thus useful in explaining the phenomena in question. A theory provides only an idealised projected image of reality in terms of certain abstracted features of the phenomena in its intended scope. These abstracted features, referred to as concepts, provide the means by which such generalised (idealised) descriptions are possible. To some extent the theory assumes that the phenomena within its intended scope can be 'adequately explained' in terms of the proposed idealised replicas viewed as isolated systems built in terms of the devised concepts.

In the context of the centuries-old dichotomy of *instrumentalism* versus *realism* the above view of theory borrows elements from both. It is instrumentalist in so far as it assumes that a theory does not purport to provide an exact picture of reality. Moreover, concepts are not viewed as describing entities of the real world which exist independently of any theory. It is also realist in two senses. Firstly, it is realist in so far as under the circumstances assumed by the theory (as an isolated system) its 'validity' can be ascertained. There is something realistic about a demand schedule in so far as it can be established or not under the circumstances assumed by the theory of demand. Secondly, it is realist because its main aim is to provide an 'adequate' explanation of the phenomena in its intended scope. As such, the adequacy of a theory can be evaluated by how successful it is in coming to grips with the reality it purports to explain. Theories are viewed as providing 'approximations' to reality and they are judged by the extent to which such approximations enhance our understanding of the phenomena in question. We cannot, however, appraise theories by the extent to which they provide exact pictures of reality

... simply because we have no access to the world independently of our theories in a way that would enable us to assess the adequacy of those descriptions . . .

(See Chalmers (1982), p. 163.)

This stems from the view that observation itself presupposes some theory providing the terms of reference. Hence, in the context of the adopted

methodology of econometric modelling, theories are treated as providing approximations to the observable phenomena without being exact copies of reality; a realist position without the logical empiricists' correspondence theory of truth (see Suppe (1977)).

In view of the realist elements of a theory propounded above the logical positivists' contention that the only function of theories is description is rejected. Theories are constructed so as to enable us to consider a variety of questions related to the phenomena within their intended scope, that includes *description, explanation and prediction*.

The distinction between theoretical and observational concepts associated with logical empiricism as well as naive instrumentalism (see Caldwell (1982)) is not adhered to in the context of the proposed methodology. This is because observation itself is theory-laden. We do not just observe, we observe within a reference framework, however rudimentary. In constructing a theory we devise the concepts and choose the assumptions in terms of which an idealised picture of the phenomena within its intended scope is projected. This amounts to devising a language and accepting a system of picturing and conceiving the structure of the phenomena in question. Such a system specifies the 'important' features of the phenomena to be observed. The main problem in econometric modelling, however, is that what a theory suggests as important features to be observed and the available observed data can differ substantially. Commonly, what is observed (data collected) was determined by some outdated theory and what was deemed possible at the time given the external constraints on data collection. Although there is an on-going revision process on what aspects of the observable phenomena data are collected, there is always a sizeable time-lag between current theories and the available observed data. For this reason in the context of the proposed methodology we distinguish between the *concepts of the theory* and the *observed data* available. It should be stressed that this is not the theoretical–observational concepts distinction of logical positivism in some disguise. It is forced upon the econometric modelling because of the gap between what is needed to observe (as suggested by the theory in question) and what is available in terms of observed data series. This distinction becomes even more important when the theory is contrasted with the actual DGP giving rise to the available data.

*Observed data* in econometric modelling are rarely the result of the experiments on some isolated system as projected by a theory. They constitute a sample taken from an on-going real DGP with all its variability and 'irrelevant' features (as far as the theory in question is concerned). These, together with the sampling impurities and observational errors, suggest that published data are far from being objective facts against which

theories are to be appraised, striking at the very foundation of logical positivism. Clearly the econometrician can do very little to improve the quality of the published data in the short-run apart from suggesting better ways of collecting and processing data. On the other hand, bridging the gap between the isolated system projected by a theory and the actual DGP giving rise to the observed data chosen is the econometrician's responsibility. Hence, in view of this and the multitude of observed data series which can be chosen to correspond to the concepts of theory, a distinction is suggested between a theoretical and an estimable model. A *theoretical model* is simply a mathematical formulation of a theory. This differs from the underlying theory in two important respects. Firstly, there might be more than one possible mathematical formulation of a theory. Secondly, in the context of a theory the initial conditions and the simplifying assumptions are explicitly recognised; in the context of the model these simplifying assumptions govern its characterisation of the phenomena of interest. This is to be contrasted with an *estimable model* whose form depends crucially on the nature of the observed data series chosen (see Chapter 1). To some extent the estimable model refers to a mathematical formulation of the observational implications of a theory, in view of the observed data and the underlying actual DGP. In order to determine the form of the estimable model the econometrician might be required to use auxiliary hypotheses in an attempt to bridge the gap between the theory and the actual DGP. It is, however, important to emphasise that an estimable model is defined in terms of the concepts of the theory and not the observed data chosen. In practice, the form of estimable models might not be readily available. This, however, is no reason to abandon the distinction and return to the textbook methodology assumption that the theory coincides with the actual DGP apart from a white-noise error. The estimable form of a theory depends crucially on the nature of the available observed data and the gap between the theory and the actual DGP, given that its aim is to bridge this gap. The estimable model plays an important role in the proposed methodology because it provides the link between a theory and a statistical model.

Having rejected the presupposition that the actual DGP, the theory and the statistical model coincide apart from a white-noise error, the task of specifying a statistical model, for the problem in hand, is no longer a trivial matter. The statistical model can no longer be specified by attaching a white-noise process, assumed to follow a certain distribution, to the theoretical model. The nature of the observed data has a crucial role to play because of their relationship with the estimable form of the theory. Because of this it is important to specify the statistical model, in the context of which the estimable model will be analysed, taking account of the nature of the

observed data in question.

The procedure from the observed data to the statistical model has been discussed in various places throughout the book but it has not been formalised in any systematic way. Because of its central role in the proposed methodology, however, it is imperative to provide such a formalisation. The first step *from the observed data to the statistical model* is made by assuming that:

*the observed data constitute a realisation  $\mathbf{Z} \equiv (\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_T)'$  of a sequence of random vectors (stochastic processes)  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$ .*

This assumption provides the necessary link between the actual DGP and probability theory. It enables us to postulate a probabilistic structure for  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  in the form of its joint distribution function  $D(\mathbf{Z}; \psi)$ . This will provide the basis for the statistical model specification.

The specification of a statistical model is based on three sources of information:

- (i) theory information;
- (ii) measurement information; and
- (iii) sample information.

*The theory information* comes in the form of the estimable model. Its role in the initial specification of the model is related to the choice of the observable variables underlying the model as well as the general form of the statistical GM.

*The measurement information* is related to the quantification and the measurement system properties of the variables involved. These include the units of measurement, the measurement system (nominal, ordinal, interval, ratio; see Appendix 19.1), as well as exact relationships among the observed data series such as accounting identities. Such information is useful for the specification of the statistical model because it enables us to postulate a sensible statistical GM for the problem in question. For example, if the statistical GM  $y_t = \beta' \mathbf{x}_t + u_t$  allows  $y_t$  to take values outside its range we need to reconsider it. Also, if an accounting identity holds either among some of the  $x_{it}$ s or among  $y_t$  and some  $x_{it}$ s the statistical GM is useless. For further discussion on accounting identities and their role in econometric modelling see Spanos (1982a).

*Sample information* comes in the form of the observable random variables involved and their structure. It is helpful to divide this information into three mutually exclusive sets related to  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$ :

- (a) past  $\{\mathbf{Z}_{t-i}, i = 1, 2, \dots\}$ ,
- (b) present  $\{\mathbf{Z}_t\}$ ,
- (c) future  $\{\mathbf{Z}_{t+i}, i = 1, 2, \dots\}$ .

Such information is useful in relation to important concepts underlying the specification of a statistical model such as exogeneity, Granger-causality,

structural invariance (see Hendry and Richard (1983)).

The specification of a statistical model is indirectly related to the distribution of the stochastic process  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  because in practice it is easier to evaluate the appropriateness of probabilistic assumptions about the marginal distributions of the  $Z_{it}$ s instead of any conditional distributions directly related to the postulated statistical GM. According to Fisher the problem of statistical model specification

... is not arbitrary, but requires an understanding of the way in which the data are supposed to, or did in fact originate ...

(See Fisher (1958), p. 8.)

For this reason the statistical model is directly related to the actual DGP being defined in terms of the observed data and not some arbitrary white-noise error term. From the modelling viewpoint it is convenient to specify a statistical model in terms of three interrelated components:

- (i) statistical GM;
- (ii) probability model; and
- (iii) sampling model.

The statistical GM defines a probabilistic mechanism purporting to provide a generalised approximation to the actual DGP. It is a generalised approximation in the sense that it incorporates no theoretical information apart from the choice of the observed data and the general form of the estimable model. That is, it is 'designed' to provide the framework in the context of which any possibly testable theoretical information of interest could be tested. In particular, any theoretical information not needed in determining the distribution of  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  for the sample period  $t = 1, 2, \dots, T$  might be testable in the context of the postulated statistical model. On the other hand, the other two sets of information, the measurement and sample information, play a vital role in determining  $D(\mathbf{Z}; \psi)$ ,  $\mathbf{Z} \equiv (\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_T)$ , and should be taken into consideration at the outset. The theoretical information which is taken into account in designing a statistical GM is the form of the estimable model. This is because the estimable model constitutes a reformulation of the theoretical model so as to provide an idealised approximation to the actual DGP in view of the available data and the statistical GM postulates a probabilistic mechanism which could conceivably have given rise to the observed data. The latter, however, does not necessarily coincide with the former apart from some white-noise error. The statistical GM is a stochastic mechanism whose particular form depends on the nature of  $D(\mathbf{Z}; \psi)$ . In particular the statistical GM is defined in terms of a distribution derivable from  $D(\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_T; \psi)$  by marginalisation and conditioning (see Chapters 5 and 6) in such a way so as to enable us to analyse (statistically) the estimable model in its context. That

is, the estimable model ‘translated’ in terms of the observable variables involved should be ‘nestable’ (a special case) within the postulated statistical GM.

The probability distribution underlying a statistical GM defines the *probability model* component of the statistical model in question. For the completeness of the argument and the additional insights available at the misspecification stage the distribution defining the probability model is related to  $D(\mathbf{Z}; \psi)$ . The apparent generality we lose by ignoring  $D(\mathbf{Z}; \psi)$  and going directly to the distribution of the probability model is illusory. The probability model in the context of statistical inference plays the role of the ‘anchor distribution’ in terms of which every other distribution or distribution related quantity is definable. In Part IV we discussed many instances where if the modeller does not adhere to the same probability model the argument can be led astray. Such instances include the omitted variables argument, the stochastic linear regression model as well as the problem of simultaneity. In the last case the argument about simultaneity bias is based totally on defining the expectation operator  $E(\cdot)$  in terms of  $D(y_t/\mathbf{X}_t; \theta)$ . This bias can be made to disappear if the underlying distribution can be changed (see Chapter 25).

As argued in Chapter 17, the sampling model, which constitutes the third component of a statistical model, provides the link between the probability model and the observed data in hand. Its role might not seem so vital at the specification stage because of its close relationship with the other two components. At the misspecification stage, however, it can play a crucial role and its close relationship with the other two components can be utilised to determine the way to proceed. Given that the statistical model is defined directly in terms of the observable variables giving rise to the observed data in hand and not some unobservable error term, the sampling model has a very crucial role to play in econometric modelling. For example, in the case of the linear regression model very few eyebrows will be raised if a modeller assumes that the error is independent. On the other hand, if time-series data are used and the modeller postulates that  $y_t$  (conditional on  $\mathbf{X}_t = \mathbf{x}_t$ ) is independent for  $t = 1, 2, \dots, T$ , a sizeable number of econometricians will advise caution; even though the two assumptions are equivalent in this context. Making probabilistic assumptions about observable random variables which gave rise to the data in hand seems to provide a better perspective for judging the appropriateness of such assumptions.

Once a statistical model is postulated we can proceed to estimate the parameters in terms of which the statistical GM is defined, what we called the *statistical parameters of interest*. These parameters do not necessarily coincide with the *theoretical parameters of interest* in terms of which the estimable model is defined. Before we are in a position to relate the two sets

of parameters we need to ensure that the estimated *statistical model is well defined*. Given that statistical arguments will be used to 'define' as well as test any hypotheses related to the theoretical parameters of interest, we need to ensure that their foundation (the estimated statistical model) is well defined in the sense that the underlying assumptions are valid. The estimated statistical model might be viewed as a convenient summarisation of the sample information in the Fisher paradigm analogous to the histogram in the Pearson paradigm. As such, the estimated statistical model is well defined if the underlying assumptions (defining it) are valid. Otherwise, any statistical arguments based on invalid assumptions will be at best misleading. *Misspecification testing* refers to the testing of the (testable) assumptions underlying the statistical model in question. The estimated statistical parameters of interest acquire a meaning only after the misspecification testing has been completed without any assumptions being rejected. This is to ensure that the estimated parameters are indeed 'good' estimates of the intended statistical parameters of interest. It will be very difficult to overestimate the importance of misspecification testing in the context of econometric modelling. A misspecified estimated statistical model is essentially useless in this context.

The concept of an *estimated well-defined statistical model* plays a vital role in the context of the proposed methodology of econometric modelling. No valid statistical inference argument can be made unless it is based on such an estimated statistical model. For this reason the statistical GM is not 'constrained' to coincide with the estimable model. For example, there is no point in adhering to one lag in the variables involved in view of the estimable model if the temporal structure of these variables requires more than one in order to yield a well-defined statistical model. The modeller should not feel constrained to allow for features necessitated by the nature of the observed data chosen simply because the theory (estimable model) does not account for. More often than not the theoretical information relating to the estimable model is vague because of the isolated system characteristics of economic theories. An important implication of this is the apparent non-rejection of theories on the basis of statistical tests. Such theories can be accepted or rejected if the 'isolated system' conditions assumed by the theory are made to hold under an experimental-like situation where the theoretical model and the actual DGP can be ensured to coincide apart from a white-noise error term. When this is not the case rejecting a theory whose estimable model can take a number of different forms will be very difficult.

In the case where misspecification testing leads to the rejection of one or more of the assumptions underlying the statistical model we proceed by respecifying the model so as to take account of the apparent invalid

assumption. What we do not do is to ‘engage in local surgery’ by ‘drafting’ the alternative hypothesis of a misspecification test into an otherwise unchanged statistical model such as postulating an AR(1) error because the Durbin–Watson test rejects the independence assumption, in the context of the linear regression model; see Hendry (1983) for a similar viewpoint. Once a well-defined estimated statistical model is reached we can proceed to determine (construct) the empirical econometric model.

Starting from a well-defined statistical model we can proceed to test any theoretical restrictions which can be related to the statistical formulation. The specification of the empirical econometric model can be viewed as a reparametrisation/restriction of the estimated statistical GM in view of the estimable model, so as to be expressed in terms of the *theoretical parameters of interest*. Any reparametrisation which imposes restrictions on the statistical parameters can be tested for on a formal hypothesis-testing framework and accepted or rejected. The aim is to construct an approximation to the actual DGP in terms of the theoretical parameters of interest as suggested by the estimable form of the theory, *an empirical econometric model*. This should be done not at the expense of the statistical properties enjoyed by the estimated statistical GM because the empirical econometric model needs to be itself a well-defined statistical model for prediction and policy evaluation purposes. Hence, when the empirical econometric model is constructed by reparametrising/restricting a well-defined statistical GM we need to ‘check’ that it satisfies the underlying statistical assumptions we started with. This does not constitute proper statistical testing but informal diagnostic ‘checking’. It is important, however, in order to be able to use formal statistical arguments in relation to prediction or/and policy evaluation in the context of the empirical econometric model.

Although we need to ensure that the theoretical parameters of interest  $\xi$  are uniquely defined in terms of the statistical parameters of interest  $\theta$ , there is nothing unique about  $\xi$ . Numerous theoretical parametrisations are possible for any well-defined set of statistical parameters. In practice, we need to choose one of such possible reparametrisations/restrictions of the estimated statistical GM. Several criteria for *model selection* (see Fig. 1.2) have been proposed in the econometric literature depending on the potential uses of the empirical econometric model in question. The most important of such criteria are:

- (i) theory consistency;
- (ii) goodness of fit;
- (iii) predictive ability;
- (iv) robustness (including nearly orthogonal explanatory variables);
- (v) encompassing;

(vi)      parsimony

(see Hendry and Richard (1982), (1983), Hendry (1983), Hendry and Wallis (1984)).

In cases where the observed data can be relied upon as accurate measurements of the underlying variables, there is something realistic about an empirical econometric model in so far as it can be a ‘good’ or a ‘bad’ approximation to the actual DGP. In such cases the instrumentalists’ interpretations of such an empirical econometric model is not very useful because it will stop the modeller seeking to improve the approximation. The realistic interpretation, however, should not be viewed as implying the existence of the ultimate ‘truth’ econometric modelling is aiming at; we have no criteria outside our theoretical perspective to establish ultimate ‘truth’. This should not stop us from seeking better empirical econometric models which provide us with additional insight in our attempt to explain observable economic phenomena of interest. In particular, an empirical econometric model which explains why and how other empirical studies have reached the conclusions they did is invaluable. In such a case we say that the empirical econometric model *encompasses* others purporting to explain the same observable phenomenon; on the subject of encompassing see Hendry and Richard (1982), (1983), and Mizon (1984). Hendry and Richard (1982), in their attempt to formalise the concept of a well-defined empirical model, include the encompassing of all rival models as one of the important conditions for what they call ‘a tentatively adequate conditional data characterisation’ (TACD); for further discussion on the similarities and differences between the two approaches see Spanos (1985).

The empirical econometric model is to some extent as close as we can get to an actual DGP within the framework of an underlying theory and the available observed data chosen. As argued above, its form takes account of all three main sources of information – theory, measurement and sample. As such, the above-discussed methodology differs from both extreme approaches to econometric modelling where only theory or data are used for the specification of empirical models. The first extreme approach requires the statistical model to coincide with the theoretical model apart from a white-noise error term. The second approach ignores the theoretical model altogether and uses only the structure of the observed data chosen as the only source of information for the determination of the empirical model; see Bos and Jenkins (1976); see Spanos (1985).

This concludes the formalisation of the alternative methodology sketched in Chapter 1. A main feature of the new methodology is the broadening of the intended scope of econometrics. Econometric modelling is viewed not as the estimation of theoretical relationships nor as a procedure in establishing the ‘trueness’ of economic theories, but as an

endeavour to understand observable economic phenomena of interest using observed data in conjunction with some underlying theory in the context of a statistical framework.

### 26.3 Conclusion

The methodology formulated in Section 26.2 can be viewed as an attempt towards a coherent approach to econometric modelling where economic theory as well as the structure of the observed data have a role to play. In a certain sense the adopted methodology can be seen as an attempt to formalise certain procedures which are currently practised by an increasing number of applied econometricians. The most important features of the adopted methodology are:

- (i) the role of the actual DGP;
- (ii) the distinction between a theoretical and an estimable model;
- (iii) the role of the observed data in the statistical model specification;
- (iv) the notion of a well-defined estimated statistical model; and
- (v) the distinction between an estimated statistical GM and an empirical econometric model.

As mentioned in the preface the methodology formalised in the present book can be viewed as having evolved out of the LSE tradition in econometric modelling (see Gilbert (1985)) that owes a lot to the work of Denis Sargan and David Hendry. The main features of the LSE tradition in time series econometric modelling, including its emphasis on specification and misspecification testing, maximum likelihood and instrumental variables estimation, asymptotic approximations, dynamic specifications, common factor restrictions and error-correction parametrisations, are integrated within a (hopefully) coherent framework in order to formalise an approach to econometric modelling.

Another related methodology at odds with the textbook methodology was proposed by Sims (1980). The Sims' methodology, in terms of Fig. 1.2, essentially ignores the left-hand side of the diagram and concentrates almost exclusively on modelling the observed data. For a discussion of the relationships between the methodology adopted in the present book and these alternative methodologies see Spanos (1985).

Finally, it is important to warn the reader about the nature of any methodological discussion. On this I can do no better than quote from Caldwell (1982):

...methodology is a frustrating and rewarding area in which to work. Just as there is no best way to listen to a Tchaikovsky symphony, or to write a book, or to raise a child, there is no best way to investigate social reality.

Yet methodology has a role to play in all of this. By showing that science is not objective, rigorous, intellectual endeavour it was once thought to be, and by demonstrating that this need not lead to anarchy, that critical discourse still has a place, the hope is held out that a true picture of the strengths and limitations of scientific practice will emerge. And with luck, this insight may lead to a better and certainly more honest, science.

(See *ibid*, p. 252.)

**Additional references**

Blaug (1980); Boland (1982); Caldwell (1984); Hendry and Wallis (1984).

# Index

- acceptance region, 286–7
  - and confidence region, 304–5
- actual DGP, 20, 661
- adjusted  $R^2$ , 382
- admissible estimator, 236
- almost sure convergence, 188
- alternative hypothesis, 287
- approximate MLE's, 533–6
- a priori restrictions, 377
  - exclusion (zero-one), 616–19
  - linear, 396–401, 422–7
  - linear homogeneous, 615–19
  - non-linear, 427–32
- ARIMA ( $p, d, q$ ) process, 156
  - stability conditions, 161
- ARMA ( $p, q$ ) process, 159–61
- ARCH test, 550
- AR(1) process, 150–5
  - error, 506–7
  - estimation, 279–81
  - stability condition, 152–4
- AR( $m$ ) process, 155–8, 506–7
- asymptotic expansions, 203–8
- asymptotic independence, 140–1, 501
- asymptotic moments, 192
- asymptotic power function, 327
- asymptotic properties of estimators, 244–7
  - consistency, 244
  - efficiency, 247
  - normality, 246
  - unbiasedness, 247
- asymptotic properties of tests, 326–8
  - consistency, 327
  - locally UMP, 328
  - UMP, 327–8
  - unbiasedness, 327
- asymptotic stationarity, 153
- asymptotic test procedures, 328–35
- asymptotic uncorrelatedness, 141
- autocorrelation, 134
- autocorrelation, errors, 501–3, 505–11
  - tests for, 513–21
- autocovariance, 134
- autoproduct moment, 134
- auxiliary regressions, 446–7, 460–1, 467, 470
- Bassman test, 652
- Bayes' formula, 121
- Bayesian approach, 220
- Bernoulli distribution, 62–3, 166
- Bernoulli's theorem, 165
- best linear unbiased estimator (BLUE), 239, 255–6, 450–1
- best linear unbiased scalar (BLUS)
  - residuals, 407
- beta distribution, 401, 479
- bias, 235
- binomial distribution, 63–4, 166
- bivariate distributions, 79–93
  - binomial, 84
  - exponential, 92, 124
  - logistic, 91, 125
  - normal, 83, 88, 93, 120, 122
  - Pareto, 84, 124
- Borel field, 41, 52
- Borel function, 95
- Box–Cox transformation, 455–7
- Box–Jenkins approach (*see* ARMA, ARIMA)
- Breusch–Pagan test, 469–70
- Brownian motion process, 149–50

- CAN estimators, 271  
 canonical correlations, 314  
 Carleman's condition, 74  
 Cauchy distribution, 70–1, 105  
 causality (*see* Granger non-causality)  
 central limit theorem  
     De Moivre–Laplace, 64, 165  
     Liapounov, 174  
     Lindeberg–Feller, 174  
     Lindeberg–Levy, 173  
 characteristic function, 73–4  
 Chebyshev's inequality, 73  
 chi-square distribution, 98–9, 108, 111  
     non-central, 108, 111  
 Chow test, 487–8  
 collinearity, exact, 432–4  
     ‘near’, 434–40  
 common factor restrictions, 507–11  
 condition numbers, 436  
 conditional distributions, 89–94  
     exponential, 92  
     logistic, 91  
     normal, 93  
     Pareto, 92  
 conditional expectation, 121–7  
     wrt a  $\sigma$ -field, 125–7  
     wrt an observed value, 121–5  
     properties, 122, 125, 126–7  
 conditional moments, 122–5  
     mean, 122–3  
     variance, 122–3  
 conditional probability, 43–4  
 confidence region, 303–6  
 confluence analysis, 12  
 consistency, weak, 244–6  
     strong, 246  
 constant, in linear regression, 370–1, 410  
 constrained MLE's, 423–4  
 continuous rv's, 56  
 convergence, mathematical, 185–8  
     of a function, 185–6  
     of a sequence, 185  
     pointwise, 187  
     uniform, 187  
 convergence, modes of, 188–92  
     almost sure, 188, 167  
     in distribution, 189, 167  
     in probability, 189, 166  
     in  $r$ th mean, 188  
 convergence of moments, 192–4  
 correlation coefficient, 119  
 covariance, 119  
     matrix, 312–3  
 Cramer–Rao, lower bound, 237  
     regularity conditions, 237  
 Cramer–Wold lemma, 191  
     cross-correlation, 135  
     cross-covariance, 135  
     cross-section data, 342–3  
     cumulants, 74  
     cumulative distribution function (*see*  
         distribution function)  
     cumulative frequency, 25  
     CUSUM test, 477  
     CUSUMSQ test, 477  
 data, economic, 342–6  
     and the probability model, 346–9  
 degrees of freedom, 108, 111–13  
 $\delta$ -method, 201  
 demand function, 10–11  
 de Moivre–Laplace CLT, 64, 165  
 density function, definition, 57  
     conditional, 90  
     joint, 82  
     marginal, 86  
     properties, 59  
 diagnostic checking, 557  
 difference equation, 155–6, 543–5  
 differencing, 161, 479–81, 528  
 differentiation of vectors and matrices,  
     603–4  
 distribution function, 55–60  
     conditional, 89–92  
     joint, 78–85  
     marginal, 85–7  
 distribution of the sample, 216  
 disturbance (*see* error term, non-systematic  
     component)  
 dummy variables, 369, 536–7  
 Durbin's  $h$  test, 541–2  
 Durbin–Watson test, 515–18  
 dynamic linear regression model, 526–70  
 dynamic multipliers, 601–2  
 econometrics, definition of, 3, 676  
 Edgeworth expansion, 206–7  
 efficiency, relative, 234–5  
     full, 237–8  
 efficient score test (*see* Lagrange multiplier  
     test)  
 eigenvalue, 433, 436  
 eigenvector, 433, 436  
 elliptical family of distributions, 458  
 empirical distribution function, 228–9  
 empirical econometric model, 21, 23, 670  
 encompassing, 568, 670  
 endogeneity (non-exogeneity), 629  
 endogenous variables, 608  
 Engel's law, 6  
 equilibrium, long-run, 558–9  
 ergodicity, 143, 500

- error, autocorrelation, 505–7  
error bounds, Berry–Esseen, 202–3  
error-correction model, 554  
errors-in-variables, 12  
error term, 349–50, 374–5 (*see also* non-systematic component)  
estimable model, 23, 668  
estimate, 231  
estimation, methods, 252–84  
estimation, properties of estimators, 231–49  
estimators,  
CAN, 271  
FIML, 625  
GLS, 463, 503, 587–8  
IV, 637–44  
*k*-class, 632  
LIML, 629, 631, 633  
OLS, 449–52  
3SLS, 639–40  
2SLS, 635–7  
estimator generation equation, 624–6  
events, 38  
elementary, 38  
impossible, 39  
mutually exclusive, 44  
sure, 39  
exclusion restrictions (*see a priori* restrictions)  
exogeneity, weak, 273, 376–7, 421–2  
strong, 505, 629–30  
tests for, 653  
expectation, 68–9  
conditional, 121–7  
properties of, 70–1, 116–20  
experimental design, 366–7  
exponential distribution, 76, 92, 124  
exponential family of distributions, 68, 299
- F* distribution, central, 104, 108, 113, 319  
non-central, 113, 319, 320, 324  
*F* test, 398–402, 425–6, 553  
power of, 401  
*F*-type misspecification test, 446  
homoskedasticity, 467, 547, 555  
linearity, 460–1, 547–8, 555  
parameter time-invariance, 477  
sample independence, 511, 541  
structural invariance, 482–6, 556
- FIML, 625  
final form, 601  
finite sample properties, 232–44  
efficiency, 234–8  
linearity, 238  
sufficiency, 242–4  
unbiasedness, 232–4
- Fisher–Neyman factorisation, 242  
Fisher's *F* distribution (*see* *F* distribution)  
Fisher paradigm, 7–9  
forecasting (*see* prediction)  
frequency curves, 27  
frequency polygon, 24  
functions of r.v.'s, distribution of, 96–107  
addition, 100–2  
min, 105–6  
quotient, 102–5
- Gamma function, 99  
Gaussian distribution (*see* normal distribution)  
Gauss linear model, 6–8, 348, 353, 357–68  
Gauss–Markov theorem, 239, 449  
generalised *F*-test, 590–3  
GIVE PC (computer package), xviii  
GIVE (estimator), 638  
GLS, 463–6, 503, 587  
goodness of fit (*see*  $R^2$ )  
Gram–Charlier series A, 205  
Granger non-causality, 505, 509, 529  
test for, 544
- hermite polynomials, 204  
heteroskedasticity, 463–71  
versus parameter time-dependence, 473  
histogram, 23  
homogeneous non-stationary process, 161, 479–81, 527–8  
homoskedasticity, 126, 378, 463–71  
misspecification tests for, 464–71, 547, 648–9
- hypothesis testing, 285–303  
alternative, 287  
composite, 287  
null, 286–7  
simple, 287
- idempotent matrix, 319, 381, 411  
identically distributed r.v.'s, 94, 216–17  
identification, 614–9  
exact, 618  
order condition, 615  
overidentification, 618  
rank condition, 617–8  
underidentification, 618  
impact multipliers, 601–2  
incidental parameters, 136, 346–7, 499  
independence, 44, 87, 93  
independent r.v.'s, 87  
linear, 118  
necessary and sufficient condition, 117–18  
versus orthogonality, 118

- independent r.v.'s (*continued*)  
   versus uncorrelatedness, 118  
 independent sample, 217, 378  
   misspecification tests for, 511–21  
 indirect MLE, 621–2, 627  
 information matrix, 239  
   asymptotic, 247  
   sample, 239  
   single observation, 247  
 information matrix test procedure, 467  
 initial conditions, 151–3, 156, 528, 531  
 innovation process, 147  
 instrumental variables, 637–44  
   estimator, 638  
 instrumentalism, 665  
 integral, Riemann–Stieltjes, 69  
 integrability, 193  
   square, 203  
   uniform, 193  
 interim multipliers, 601  
 intercept (*see* constant)  
 interval estimation (*see* confidence region)  
 invariance, linear transformations, 438–9  
 invariance of MLE's, 266–7  
 inverse matrix, partitioned, 442  
 invertibility conditions, 161  
 iterative estimation procedure, 588
- Jacobian transformation, 106, 257  
 joint central moments, 119  
 joint density function, 81–2  
   continuous, 82  
   discrete, 82
- k*-class estimator, 632  
 Khinchin's WLLN, 169  
 King's law, 5  
 Kolmogorov–Gabor polynomial, 446  
 Kolmogorov's axiomatic approach, 37–43  
 Kolmogorov's inequality, 171  
 Kolmogorov's stochastic process  
   conditions, 133  
 Kolmogorov's SLLN, 170–1  
 Kolmogorov's WLLN, 169  
 Kolmogorov–Smirnov test, 229, 453  
 Kronecker product, 573, 603–4  
 kurtosis, 73, 452
- lag operator, 155, 161, 509  
 Lagrange multiplier test procedure, 330,  
   333–4, 430–2  
   in misspecification testing, 446, 453, 460,  
   466, 468–9, 519–21  
 Laguerre polynomials, 208  
 law of large numbers (*see* WLLN, SLLN)  
 leading indicator model, 554
- least squares method, 6–7, 233–6, 448–50  
 least squares estimators, 448  
   GLS, 463, 503, 587–8  
   OLS, 448–9, 638  
 Lehmann–Scheffé theorem, 242–3, 387, 577  
 level of significance of a test (*see* size of a  
   test)  
 Liapunov's CLT, 174  
 likelihood function, 258–60  
 likelihood ratio test procedure, 299–303,  
   328–9, 335, 425–6, 432  
 limit of a function, 186  
 limit of a sequence, 185  
 limit of moments, 192  
 limit, probability (*see* convergence in  
   probability)  
 LIML estimator, 629, 631, 633  
 Lindeberg condition, 174–5  
 Lindeberg–Feller CLT, 174, 177  
 Lindeberg–Levy CLT, 173–4  
 linear regression model, 369–410  
 linear restrictions (*see* a priori restrictions)  
 linearity, 370, 378, 457–63  
   and normality, 316  
   inducing transformations, 462–3  
   misspecification tests for, 459–61, 597,  
   648
- locally UMP test, 335  
 logical empiricism, 662–3  
 logical positivism, 3, 662–3  
 logistic distribution, 91, 125  
 log-likelihood, 258–60  
 log-normal distribution, 283, 457  
 long-run equilibrium solution, 558–9  
 long-run multipliers, 602  
 lower bound (*see* Cramer–Rao lower  
   bound)
- MA( $p$ ) stochastic process, 158  
 Malinvaud formulation, 588  
 Mann–Wald theorem, 197  
 marginal distributions, 85–9  
   normal distribution, 88, 317  
 Markov inequality, 71  
 Markov process, 148–9  
 Martingale difference process, 147, 273–5  
 Martingale orthogonality, 118, 171  
 Martingale process, 145  
   CLT for, 178  
   SLLN for, 172  
   WLLN for, 172
- maximum likelihood, method, 257–81  
 mean, 25, 70–1  
 mean square error (MSE), 234–6, 249  
 measurement equations, 12  
 measurement information, 352, 665

- measurement systems, 409–11  
*m*-dependent process, 141  
median, 25, 71  
methodology, 15–21, 659–72  
misspecification testing, 21, 221, 392  
mixing, strong, 142, 179  
  uniform, 143, 179  
MLE, 260  
  constrained, 423  
  properties, 266, 82  
mode, 25, 71  
model selection, 523, 669–70  
moments, approximate, 194  
  asymptotic, 192  
  central, 73, 119  
  limit of, 192  
  raw, 73, 118  
moments, method of, 256–7  
Monte Carlo, 435  
*m*th-order Markov process, 149  
multicollinearity (*see* collinearity)  
multiplication rule of probability, 44  
multiple correlation coefficient, 313–14,  
  318, 322–3, 382, 439  
multivariate linear regression model,  
  571–607  
  in relation to the SEM, 610–14  
multivariate normal distribution, 312–24  
multivariate *t* distribution, 471
- Neyman–Pearson theorem, 296  
non-centrality parameter, 108, 111–13  
  in the *F*-test, 399, 401  
nonlinear model, 461–3  
non-linear restrictions (*see* a priori  
  restrictions)  
non-parametric inference, 218  
non-parametric processes, 146–52  
non-parametric tests, 453  
non-random sample, 218, 343, 494–7  
non-stochastic variables, 357  
non-systematic component, 350, 370, 374,  
  376  
normal distribution, 64–6  
  bivariate, 83–4, 88  
  mean of, 70  
  multivariate, 315–24  
  standard, 68  
  variance, 71  
normality, 447–57  
  misspecification tests for, 451–5  
normalising transformations, 455–7  
normal (Gaussian) stochastic process,  
  135–6  
  time homogeneity restrictions on, 139
- nuisance parameters, 414  
null hypothesis, 286
- observation space, 217  
ogive, 27
- omitted variables bias argument, 419–21  
  and auxiliary regressions, 446, 458–61,  
    468, 471, 502, 515, 523  
  reformulated, 445–7
- OLS, 448–51
- $O, o$  notation, 195–6
- $O_p, o_p$  notation, 196–9
- order condition (*see* identification)  
order of magnitude, 171, 174, 179, 194–8  
orthogonal projection, 381, 411, 642  
orthogonality, 118  
  between systematic and non-systematic  
    components, 350, 358, 371, 381  
overdifferencing, 479  
overidentifying restrictions, 618  
  test for, 651–3  
overparametrisation, 612–13
- panel data, 42
- parameter space, 60
- parameter structural change, 481–7
- parameter time-invariance, 378, 472–81
- parametric family of densities (*see*  
  probability model)  
parametric processes, 146
- Pareto distribution, 61, 339–41
- partial adjustment model, 552
- partial correlation coefficient, 314, 318,  
  323, 439–40
- Pearson family of densities, 28, 452–3
- Pearson paradigm, 7–8
- Pillai's trace test, 593
- pivots, 295
- Political Arithmetik*, 4–5
- power function, 291
- power of a test, 290
- power set, 39
- predetermined variables, 610
- prediction, 221, 247–9, 306–9  
  in the linear regression model, 402–5  
  in the multivariate linear regression  
    model, 599, 601–2, 654–5
- principal components, 434
- probability, definition,  
  axiomatic, 43  
  classical, 34  
  frequency, 34  
  subjective, 35
- probability limit (*see* convergence)
- probability model, 60–1, 214

- probability set function, 42–3  
 probability space, 42
- quadratic forms related to the normal distribution, 319–20  
 quotient of two r.v.'s, 102–5
- $R^2$  (*see* multiple correlation coefficient)  
 random experiment, 37  
 random matrix, 135  
 random sample, 216–17  
 random variable, 48–76
  - continuous, 56
  - definition, 50
  - discrete, 56
  - functions of, 97, 99–110
  - minimal  $\sigma$ -field generated by, 50
 random vector, 78–93
- rank condition (*see* identification)  
 Rao–Blackwell lemma, 243  
 realism, 663  
 recursive estimator, 407, 474–78  
 recursive system, 612–14  
 regression curve, 122–4  
 regressors, order of magnitude, 391–2  
 rejection region, 286  
 reparametrisation/restriction, 21, 352  
 RESET type tests, 446, 460–1, 555, 597  
 residuals, 405–8
  - BLUS, 407
  - recursive, 407, 474–8
 residual sum of squares (RSS), 428  
 respecification approach, 498–502, 505–9  
 restrictions (*see* *a priori* restrictions)  
 Russian school, 36, 64
- sample information, 352, 667  
 sample moments, 227  
 sample space, 38  
 sampling model, 215–19  
 score function, 260  
 second order stochastic process, 138–9  
 selection matrices, 619  
 sequential conditioning, 273–4, 495  
 serial correlation (*see* autocorrelation)  
 Shapiro–Wilk test, 452  
 $\sigma$ -field, definition, 40
  - generated by a r.v., 50–1
  - generated by a set, 40
  - increasing sequence of, 51
 simple random sampling, 343  
 simultaneous equation model, 608–58  
 singular normal distribution, 406  
 size of a test, 291  
 skedasticity, 123–4  
 skewness, 26, 73
- skewness-kurtosis test, 452–5  
 SLLN, 170–2  
 small sample properties (*see* finite sample properties)  
 specification testing, 392  
 spectral decomposition (*see* eigenvalues)  
 standard deviation, 72  
 stationarity, strict, 137–8
  - $l$ th order, 138–9
 statistic, 224  
 statistical GM, 344, 349–52  
 statistical model, 218, 339
  - well-defined estimated, 352, 409, 522, 668
 statistical parameters of interest, 351, 371, 376, 419–22, 575, 666–7
  - versus theoretical parameters of interest, 351, 376, 667–9
 stochastic linear regression model, 413–18  
 stochastic process, 131–7
  - realisation of, 131
 stratified sampling, 343  
 structural form, unrestricted, 614
  - restricted, 616
 structural parameters (*see* theoretical parameters of interest)  
 Student's  $t$  distribution, 104, 108
  - multivariate, 471
 sufficiency, 242
  - minimal, 243
 supermartingale, 145  
 SURE formulation, 585–8, 636  
 systematic component, 349–51, 370–1, 375–6
- $t$  distribution (*see* Student's  $t$  distribution)  
 $t$  test, 364, 396–7, 392  
 test, definition, 294  
 test statistic, 289  
 testing, 221, 285–303  
 Theil's inequality coefficient, 405  
 theoretical parameters of interest, 349, 351, 553, 569, 613–14, 620, 669  
 theoretical model, 21, 667  
 theory, 20, 662–6  
 3SLS estimator, 635–7  
 time-homogeneity restrictions, 137–9  
 time series data, 130, 342  
 Toeplitz matrix, 495, 505  
 total sum of squares, 382  
 2SLS estimator, 626–35  
 type I error, 286  
 type II error, 286
- UMA region, 305  
 UMP test, 291

- unbiased confidence region, 306  
unbiased estimator, 232  
unbiased test, 293  
uncorrelated r.v.'s, 118  
underidentification, 621  
uniform convergence (*see* convergence)  
uniform distribution, 57–8  
unimodal distribution, 71  
univariate distributions, 62–8  
  
variance, 72, 119  
    properties, 73  
variance–covariance matrix, 312  
variance ratio tests, 395–6  
variance stabilising transformations, 487–8  
variation free condition, 377, 422  
    violation of, 629  
vectoring, 573, 604  
vector stochastic process, 135  
  
Wald test procedure, 329, 332–3, 429–30  
weak convergence (*see* convergence in probability)  
weakly stationary process (*see* second order stationarity)  
Weibull distribution, 105  
white-noise process, 150, 151  
White test for homoskedasticity, 465–7  
Wilk's ratio test, 593  
Window,  $t$ -period, 40, 562  
Wishart distribution, 321, 577, 602–3  
WLLN, 168–9  
Wold decomposition, 159  
  
Yule–Walker equations, 157  
  
Zellner formulation (*see* SURE)  
Zero-one restrictions (*see* exclusion restrictions)

# Symbols and abbreviations

## (1) Symbols

- $N(\mu, \sigma^2)$  – normal distribution with mean  $\mu$  and variance  $\sigma^2$   
 $\mathcal{E}$  – random experiment  
 $\mathcal{F}$  –  $\sigma$ -field  
 $\mathcal{B}$  – Borel field  
 $\cup$  – union  
 $\cap$  – intersection  
 $\bar{\cdot}$  – complementation  
 $\subseteq$  – subset of  
 $\in$  – belongs to  
 $\notin$  – does not belong to  
 $\emptyset$  – empty set  
 $\sigma(A)$  – minimal  $\sigma$ -field generated by  $A$   
 $\forall$  – for all  
 $U(\alpha, \beta)$  – uniform distribution between  $\alpha$  and  $\beta$   
 $B(n, p)$  – binomial distribution with parameters  $n$  and  $p$   
 $\Phi = \{ \}$  – probability model  
 $\chi^2(n)$  – chi-square distribution with  $n$  degrees of freedom  
 $F_{\substack{n_1, n_2}}(P)$  –  $F$  distribution with  $n_1$  and  $n_2$  degrees of freedom  
 $\rightarrow_p$  – convergence in probability  
 $\xrightarrow{D}$  – convergence in distribution  
 $\xrightarrow{a.s.}$  – almost sure convergence  
 $\xrightarrow{r}$  – convergence in  $r$ th mean  
 $\mathbb{R}$  – the real line  $(-\infty, \infty)$

$\mathbb{R}^k$	– Euclidean $k$ -dimensional space
$\mathbb{R}_+$	– positive real line $[0, \infty)$
$I_T(\theta)$	– sample information matrix
$I(\theta)$	– single observation information matrix
$I_\infty(\theta)$	– asymptotic information matrix
$\sim$	– distributed as
$\underset{x}{\sim}$	– asymptotically distributed as
$\overset{H}{\sim}$	– distributed under $H$
$E(\cdot)$	– expected value
$E_x(\cdot)$	– asymptotic expected value
$\text{Var}(\cdot)$	– variance
$\text{Var}_x(\cdot)$	– asymptotic variance
$\mathbb{T}$	– index set (usually $\mathbb{T} = \{1, 2, 3, \dots\}$ )
$O_p$	– at most of order in probability
$o_p$	– of smaller order in probability
$\simeq$	– approximately equal to
$\ln$	– $\log_e$
$\otimes$	– Krönecker product
$\text{vec}$	– vectoring
$\text{tr}$	– trace
$\det$	– determinant

## (2) Abbreviations

DGP	– data generating process
CLT	– central limit theorem
DF	– distribution function
pdf	– probability density function
WLLN	– weak law of large numbers
SLLN	– strong law of large numbers
URSS	– unrestricted residual sum of squares
RRSS	– restricted residual sum of squares
DLR	– dynamic linear regression
MLR	– multivariate linear regression
MDLR	– multivariate dynamic linear regression
MSE	– mean square error
UMP	– uniformly most powerful
IID	– independent and identically distributed
MLE	– maximum likelihood estimator

- OLS – ordinary least-squares  
GLS – generalised least-squares  
AR – autoregressive  
MA – moving average  
ARMA – autoregressive, moving average  
ARIMA – autoregressive, integrated, moving average  
ARCH – autoregressive, conditional heteroskedasticity  
LR – likelihood ratio test  
LM – Lagrange multiplier  
GM – generating mechanism  
rv – random variable  
IV – instrumental variable  
GIVE – generalised instrumental variables estimator  
2SLS – two stage least-squares  
LIML – limited information maximum likelihood  
FIML – full information maximum likelihood  
3SLS – three stage least-squares  
wrt – with respect to  
IMLE – indirect maximum likelihood estimator  
CAN – consistent, asymptotically normal  
NIID – normal IID  
UMA – uniformly most accurate  
MAE – mean absolute error  
RESET – regression specification error test  
BLUE – best linear unbiased estimator  
BLUS – best linear unbiased scalar (residuals)