STRUCTURAL MACROECONOMETRICS

MASAO OGAKI
The Ohio State University

KYUNGHO JANG
The University of Alabama at Birmingham

HYOUNG-SEOK LIM
The Bank of Korea

First draft: May, 2000

This version: February 29, 2004
This book presents various structural econometric tools used in macroeconomics. The word “structural” has been defined in many ways. In this book, “structural” means that explicit assumptions are made in econometric methods so that estimators or test statistics can be interpreted in terms of an economic model (or models) as explained in Chapter 1.

Many applied macroeconomists link macroeconomic models with econometric methods in this sense of structural econometrics. In principle, recent advances of theoretical time series econometrics make this task easier because they often relax the very restrictive assumptions made in conventional econometrics. There are many textbooks that explain these advanced econometric methods. It is often difficult, however, for applied researchers to exploit these advances because few textbooks in time series econometrics explain how macroeconomic models are mapped into advanced econometric models.\(^1\) To fill this gap, this book presents methods to apply advanced econometric procedures to structural macroeconomic models. The econometric methods covered are mainly those of time series econometrics, and include the generalized method of moments, vector autoregressions, and estimation and testing in the presence of nonstationary variables.

Since this book focuses on applications, proofs are usually omitted with references given for interested readers. When proofs are helpful to understand issues that are important for applied research, they are given in mathematical appendices. Many examples are given to illustrate concepts and methods.

\(^1\)For example, Hamilton (1994) contains exceptional volume of explanations of applications for a time series econometrics textbook, but its main focus is on econometrics, and not on the mapping of economic models into econometric models.
This book is intended for an advanced graduate course in time series econometrics or macroeconomics. The prerequisites for this course would include an introduction to econometrics. This book is also useful to applied macroeconomic researchers interested in learning how recent advances in time-series econometrics can be used to estimate and test structural macroeconomic models.
# Contents

1 **INTRODUCTION**

2 **STOCHASTIC PROCESSES**
2.1 Review of Probability Theory ........................................ 5
2.2 Stochastic Processes .................................................. 7
2.3 Conditional Expectations ............................................. 8
2.4 Stationary Stochastic Processes ................................... 12
2.5 Conditional Heteroskedasticity .................................... 16
2.6 Martingales and Random Walks .................................. 18
2.A A Review of Measure Theory ..................................... 19
2.B Convergence in Probability ........................................ 29
2.B.1 Convergence in Distribution ................................... 30
2.B.2 Propositions 2.2 and 2.3 for Infinite Numbers of R.V.’s (Incomplete) .................................................. 31

3 **FORECASTING**
3.1 Projections ............................................................. 33
3.1.1 Definitions and Properties of Projections .................. 33
3.1.2 Linear Projections and Conditional Expectations ....... 35
3.2 Some Applications of Conditional Expectations and Projections .................... 37
3.2.1 Volatility Tests .................................................. 37
3.2.2 Parameterizing Expectations ................................. 39
3.2.3 Noise Ratio ...................................................... 40
3.A Introduction to Hilbert Space .................................. 42
3.A.1 Vector Spaces ..................................................... 42
3.A.2 Hilbert Space ..................................................... 44

4 **ARMA AND VECTOR AUTOREGRESSION REPRESENTATIONS** 51
4.1 Autocorrelation ....................................................... 51
4.2 The Lag Operator ..................................................... 52
4.3 Moving Average Representation ................................. 53
4.4 Autoregression Representation ................................. 55
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4.1</td>
<td>Autoregression of Order One</td>
<td>55</td>
</tr>
<tr>
<td>4.4.2</td>
<td>The $p$-th Order Autoregression</td>
<td>57</td>
</tr>
<tr>
<td>4.5</td>
<td>ARMA</td>
<td>58</td>
</tr>
<tr>
<td>4.6</td>
<td>The Wold Representation</td>
<td>58</td>
</tr>
<tr>
<td>4.7</td>
<td>Fundamental Innovations</td>
<td>61</td>
</tr>
<tr>
<td>4.8</td>
<td>The Spectral Density</td>
<td>63</td>
</tr>
<tr>
<td>4.A</td>
<td>Difference Equations</td>
<td>65</td>
</tr>
</tbody>
</table>

## 5 STOCHASTIC REGRESSORS IN LINEAR MODELS | 67 |
| 5.1   | The Conditional Gauss Markov Theorem | 68 |
| 5.2   | Unconditional Distributions of Test Statistics | 73 |
| 5.3   | The Law of Large Numbers | 75 |
| 5.4   | Convergence in Distribution and Central Limit Theorem | 76 |
| 5.5   | Consistency and Asymptotic Distributions of OLS Estimators | 80 |
| 5.6   | Consistency and Asymptotic Distributions of IV Estimators | 82 |
| 5.7   | Nonlinear Functions of Estimators | 83 |
| 5.8   | Remarks on Asymptotic Theory | 83 |
| 5.A   | Monte Carlo Studies | 84 |
| 5.A.1 | Random Number Generators | 85 |
| 5.A.2 | Estimators | 88 |
| 5.A.3 | Tests | 89 |
| 5.A.4 | A Pitfall in Monte Carlo Simulations | 90 |
| 5.A.5 | An Example Program | 92 |

## 6 ESTIMATION OF THE LONG-RUN COVARIANCE MATRIX | 101 |
| 6.1   | Serially Uncorrelated Variables | 102 |
| 6.2   | Serially Correlated Variables | 103 |
| 6.2.1 | Unknown Order of Serial Correlation | 103 |
| 6.2.2 | Known Order of Serial Correlation | 108 |

## 7 TESTING LINEAR FORECASTING MODELS | 112 |
| 7.1   | Forward Exchange Rates | 113 |
| 7.2   | The Euler Equation | 116 |
| 7.3   | The Martingale Model of Consumption | 118 |
| 7.4   | The Linearized Euler Equation | 119 |
| 7.5   | Optimal Taxation | 121 |

## 8 VECTOR AUTOREGRESSION TECHNIQUES | 124 |
<p>| 8.1   | OLS Estimation | 125 |
| 8.2   | Granger Causality | 126 |
| 8.3   | The Impulse Response Function | 129 |
| 8.4   | Forecast error decomposition | 132 |</p>
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.5</td>
<td>Structural VAR Models</td>
<td>133</td>
</tr>
<tr>
<td>8.6</td>
<td>Identification</td>
<td>136</td>
</tr>
<tr>
<td>8.6.1</td>
<td>Short-Run Restrictions for Structural VAR</td>
<td>136</td>
</tr>
<tr>
<td>8.6.2</td>
<td>Identification of block recursive systems</td>
<td>138</td>
</tr>
<tr>
<td>8.6.3</td>
<td>Two-step ML estimation</td>
<td>139</td>
</tr>
<tr>
<td>9</td>
<td>GENERALIZED METHOD OF MOMENTS</td>
<td>143</td>
</tr>
<tr>
<td>9.1</td>
<td>Asymptotic Properties of GMM Estimators</td>
<td>143</td>
</tr>
<tr>
<td>9.1.1</td>
<td>Moment Restriction and GMM Estimators</td>
<td>143</td>
</tr>
<tr>
<td>9.1.2</td>
<td>Asymptotic Distributions of GMM Estimators</td>
<td>144</td>
</tr>
<tr>
<td>9.1.3</td>
<td>Optimal Choice of the Distance Matrix</td>
<td>145</td>
</tr>
<tr>
<td>9.1.4</td>
<td>A Chi-Square Test for the Overidentifying Restrictions</td>
<td>146</td>
</tr>
<tr>
<td>9.2</td>
<td>Special Cases</td>
<td>146</td>
</tr>
<tr>
<td>9.2.1</td>
<td>Ordinary Least Squares</td>
<td>146</td>
</tr>
<tr>
<td>9.2.2</td>
<td>Linear Instrumental Variables Regressions</td>
<td>147</td>
</tr>
<tr>
<td>9.2.3</td>
<td>Nonlinear Instrumental Variables Estimation</td>
<td>148</td>
</tr>
<tr>
<td>9.2.4</td>
<td>Linear GMM estimator</td>
<td>148</td>
</tr>
<tr>
<td>9.3</td>
<td>Important Assumptions</td>
<td>149</td>
</tr>
<tr>
<td>9.3.1</td>
<td>Stationarity</td>
<td>150</td>
</tr>
<tr>
<td>9.3.2</td>
<td>Identification</td>
<td>151</td>
</tr>
<tr>
<td>9.4</td>
<td>Extensions</td>
<td>151</td>
</tr>
<tr>
<td>9.4.1</td>
<td>Sequential Estimation</td>
<td>151</td>
</tr>
<tr>
<td>9.4.2</td>
<td>GMM with Deterministic Trends</td>
<td>153</td>
</tr>
<tr>
<td>9.4.3</td>
<td>Minimum Distance Estimation</td>
<td>153</td>
</tr>
<tr>
<td>9.5</td>
<td>Hypothesis Testing and Specification Tests</td>
<td>154</td>
</tr>
<tr>
<td>9.6</td>
<td>Numerical Optimization</td>
<td>156</td>
</tr>
<tr>
<td>9.7</td>
<td>The Optimal Choice of Instrumental Variables</td>
<td>158</td>
</tr>
<tr>
<td>9.8</td>
<td>Small Sample Properties</td>
<td>158</td>
</tr>
<tr>
<td>9.A</td>
<td>Asymptotic Theory for GMM</td>
<td>161</td>
</tr>
<tr>
<td>9.A.1</td>
<td>Asymptotic Properties of Extremum Estimators</td>
<td>162</td>
</tr>
<tr>
<td>9.A.2</td>
<td>Consistency of GMM Estimators</td>
<td>164</td>
</tr>
<tr>
<td>9.A.3</td>
<td>A Sufficient Condition for the Almost Sure Uniform Convergence</td>
<td>165</td>
</tr>
<tr>
<td>9.A.4</td>
<td>Asymptotic Distributions of GMM Estimators</td>
<td>170</td>
</tr>
<tr>
<td>9.B</td>
<td>A Procedure for Hansen’s J Test (GMM.EXP)</td>
<td>174</td>
</tr>
<tr>
<td>10</td>
<td>EMPIRICAL APPLICATIONS OF GMM</td>
<td>181</td>
</tr>
<tr>
<td>10.1</td>
<td>Euler Equation Approach</td>
<td>181</td>
</tr>
<tr>
<td>10.2</td>
<td>Alternative Measures of IMRS</td>
<td>183</td>
</tr>
<tr>
<td>10.3</td>
<td>Habit Formation and Durability</td>
<td>185</td>
</tr>
<tr>
<td>10.4</td>
<td>State-Nonseparable Preferences</td>
<td>187</td>
</tr>
<tr>
<td>10.5</td>
<td>Time Aggregation</td>
<td>188</td>
</tr>
<tr>
<td>10.6</td>
<td>Multiple-Goods Models</td>
<td>189</td>
</tr>
</tbody>
</table>
## CONTENTS

10.7 Seasonality .................................................. 191
10.8 Monetary Models ............................................. 192
10.9 Calculating Standard Errors for Estimates of Standard Deviation, Cor-
relation, and Autocorrelation ................................. 193
10.10 Real Business Cycle Models and GMM Estimation ............... 194
10.11 GMM and an ARCH Process ............................... 199
10.12 Other Empirical Applications ............................. 202

11 UNIT ROOT NONSTATIONARY PROCESSES ................. 210
11.1 Definitions .................................................. 211
11.2 Decompositions .............................................. 212
11.3 Tests for the Null of Difference Stationarity .................. 214
   11.3.1 Dickey-Fuller Tests ................................... 214
   11.3.2 Said-Dickey Test ....................................... 216
   11.3.3 Phillips-Perron Tests ................................. 218
   11.3.4 Park’s J Tests .......................................... 218
11.4 Tests for the Null of Stationarity .......................... 220
11.5 Near Observational Equivalence ............................ 221
11.6 Asymptotics for unit root tests ............................. 222
   11.6.1 DF test with serially uncorrelated disturbances ......... 222
   11.6.2 ADF test with serially correlated disturbances ........ 226
   11.6.3 Phillips-Perron test .................................... 232
11.A Asymptotic Theory ......................................... 239
   11.A.1 Functional Central Limit Theorem .................... 239
11.B Procedures for Unit Root Tests ............................ 239
   11.B.1 Said-Dickey Test (ADF.EXP) .......................... 239
   11.B.2 Park’s J Test (JPQ.EXP) .............................. 240
   11.B.3 Park’s G Test (GPQ.EXP) .............................. 241

12 COINTEGRATING AND SPURIOUS REGRESSIONS .......... 245
12.1 Definitions .................................................. 246
12.2 Exact Finite Sample Properties of Regression Estimators .......... 248
   12.2.1 Spurious Regressions .................................. 249
   12.2.2 Cointegrating Regressions ............................. 253
12.3 Large Sample Properties ................................... 254
   12.3.1 Canonical Cointegrating Regression .................... 255
   12.3.2 Estimation of Long-Run Covariance Parameters .......... 257
12.4 Tests for the Null Hypothesis of No Cointegration ............... 259
12.5 Tests for the Null Hypothesis of Cointegration ................ 260
12.6 Generalized Method of Moments and Unit Roots ................ 261
12.A Procedures for Cointegration Tests ......................... 263
   12.A.1 Park’s CCR and H Test (CCR.EXP) .................... 263
13 ECONOMIC MODELS AND COINTEGRATING REGRESSIONS

13.1 The Permanent Income Hypothesis of Consumption 248
13.2 Present Value Models of Asset Prices 251
13.3 Applications to Money Demand Functions 253
13.4 The Cointegration Approach to Estimating Preference Parameters 253
    13.4.1 The Time Separable Addilog Utility Function 255
    13.4.2 The Time Nonseparable Addilog Utility Function 259
    13.4.3 Engel’s Law and Cointegration 264
13.5 The Cointegration-Euler Equation Approach 267
    13.5.1 The Economy 270
    13.5.2 The 2-Step Estimation Method 274
    13.5.3 Measuring Intertemporal Substitution: The Role of Durable Goods 276
13.6 Purchasing Power Parity 276

14 ESTIMATION AND TESTING OF LINEAR RATIONAL EXPECTATIONS MODELS

14.1 The Nonlinear Restrictions 284
    14.1.1 Stationary $d_t$ 285
    14.1.2 Difference Stationary $d_t$ 287
14.2 Econometric Methods 288
    14.2.1 Stationary $d_t$ 288
    14.2.2 Difference Stationary $d_t$ 289

15 VECTOR AUTOREGRESSIONS WITH UNIT ROOT NONSTATIONARY PROCESSES

15.1 Identification on Structural VAR Models 297
    15.1.1 Long-Run Restrictions for Structural VAR Models 297
    15.1.2 Short-run and Long-Run Restrictions for Structural VAR Models 298
15.2 Vector Error Correction Model 301
    15.2.1 The model and Long-run Restrictions 301
    15.2.2 Identification of Permanent Shocks 303
    15.2.3 Impulse Response Functions 305
    15.2.4 Forecast-Error Variance Decomposition 307
    15.2.5 Summary 308
15.3 Structural Vector Error Correction Models 309
15.4 An Exchange Rate Model with Sticky Prices 311
15.5 The Instrumental Variables Methods 318
15.6 Tests for the Number of Cointegrating Vectors 322
15.7 How Should an Estimation Method be Chosen? 324
CONTENTS

15.7.1 Are Short-Run Dynamics of Interest? .......................... 325
15.7.2 The Number of the Cointegrating Vectors ....................... 325
15.7.3 Small Sample Properties .................................. 326
15.A Estimation of the Model ...................................... 327
15.B Monte Carlo Integration ...................................... 332
15.C Johansen’s Maximum Likelihood Estimation and Cointegration Rank Tests ........................................ 334

16 PANEL AND CROSS-SECTIONAL DATA .................................. 343
16.1 Generalized Method of Moments ................................ 343
16.2 Tests of Risk Sharing ........................................ 345
16.3 Decreasing Relative Risk Aversion and Risk Sharing .......... 347
16.4 Euler Equation Approach ...................................... 349
16.5 Panel Unit Root Tests ......................................... 350
16.6 Cointegration and Panel Data ................................ 352

A INTRODUCTION TO GAUSS ............................................. 357
A.1 Starting and Exiting GAUSS ..................................... 357
    A.1.1 The Windows Version .................................. 357
    A.1.2 The DOS Version ...................................... 357
A.2 Running a Program Stored in a File from the COMMAND Mode ... 358
A.3 Editing a File .................................................. 358
A.4 Rules of Syntax ................................................ 358
    A.4.1 Statements ............................................. 358
    A.4.2 Case .................................................. 358
    A.4.3 Comments ............................................. 359
    A.4.4 Symbol Names ....................................... 359
A.5 Reading and Storing Data ...................................... 359
A.6 Operators ..................................................... 359
    A.6.1 Operators for Matrix Manipulations ..................... 359
    A.6.2 Numeric Operators .................................. 361
A.7 Commands ....................................................... 362
    A.7.1 Functions ............................................. 362
    A.7.2 Printing ............................................... 363
    A.7.3 Preparing an Output File .............................. 364
A.8 Procedure ..................................................... 364
A.9 Examples ...................................................... 364

B COMPLEX VARIABLES, THE SPECTRUM, AND LAG OPERA-
TOR ........................................................................ 365
B.1 Complex Variables ............................................. 366
    B.1.1 Complex Numbers ..................................... 366
## List of Tables

11.1 Critical Values of Park’s $J(p, q)$ Tests for the Null of Difference Stationarity ........................................ 219  
11.2 Probability of smaller values ........................................ 242  
12.1 Critical Values of Park’s $I(p, q)$ Tests for Null of No Cointegration ........................................ 260  
C.1 GMM Results ........................................ 390  
C.2 Data moments and model moments ........................................ 390  
C.3 GPQ tests ........................................ 392  
C.4 ADF tests ........................................ 393  
C.5 CCR estimation and H(p,q) tests ........................................ 393
Chapter 1

INTRODUCTION

The word “structural” has various different meanings in econometrics. In this book, “structural” means that explicit assumptions are made in econometric methods so that estimators or test statistics can be interpreted in terms of an economic model (or models). In some cases, some properties of the estimators and test statistics are known when they are applied to data generated from an economic model. We then use the economic model to interpret empirical results obtained by applying the econometric tools to real data. This is important because an economic model is used to analyze causal relationships between economic variables, and understanding causal relationships is essential for policy evaluations and forecasting.

As a very simple example, consider a model of demand for a good:

\[ Q_t^d = a - bP_t + e_t, \]  

where \( P_t \) is the price and \( Q_t^d \) is the market quantity demanded. In this model \( a \) and \( b \) are constants and \( e_t \) is the demand shock. The model assumes that the observed quantity, \( Q_t \), is equated with \( Q_t^d \), \( P_t \) is nonstochastic, \( e_t \) has mean zero, \( E(e_t^2) = \sigma^2 \), and \( E(e_t e_s) = 0 \) if \( t \neq s \). With these assumptions the Gauss-Markov Theorem can be applied to this model. If the Ordinary Least Squares (OLS) slope coefficient estimator
is applied to data of $Q_t$ and $P_t$ for $t = 1, \cdots, T$ in this model, then the estimator is the Best Linear Unbiased Estimator (BLUE) for the demand slope coefficient, $b$.

One benefit of having this structural model is that we know exactly what the limitations are when we interpret OLS results applied to real data in terms of the model. This knowledge is helpful because we can then study how to improve our econometric methods for better interpretation of data.

For example, consider the assumption made in the model that $P_t$ is nonstochastic. This assumption is sometimes motivated by saying that the price is taken as given by the individual market participants. It is easy to see that this motivation is problematic by considering the supply side of the market. Consider a model of supply of the good:

$$Q_s^t = c + dP_t + u_t, \quad (1.2)$$

where $Q_s^t$ the market quantity supplied and $u_t$ is the supply shock. In equilibrium, the observed quantity, $Q_t$, is equal to $Q_d^t$ and $Q_s^t$. Equating the right hand sides of (1.1) and (1.2), and solving for $P_t$, we obtain

$$P_t = \frac{1}{d + b}(a - c + e_t - u_t). \quad (1.3)$$

Hence $P_t$ is stochastic. Moreover, (1.3) makes it clear that $P_t$ is correlated with $e_t$ and $u_t$. This means that the OLS slope coefficient estimator is not even a consistent estimator for $b$ or $d$ as discussed in Chapter 5. This leads us to consider an improved econometric method, an instrumental variable method, for example.

The structural demand model tells us under what assumptions we can interpret the OLS slope estimator as an unbiased estimator for $b$. By studying the assumptions, we can see what will happen when they are violated. This process leads to better
Another consideration is the trend observed in most aggregate data. The demand model with trends leads to cointegrating regressions as discussed in Chapter 13.

Instead of starting with a demand function, one can start with a utility function as in the Euler Equation Approach discussed in Chapter 10. When data contain trends, cointegrating regressions can be used to estimate preference parameters, and this Cointegration Approach can be combined with the Euler Equation Approach as described in Chapter 13.

We do not claim that structural econometrics as defined here is better than non-structural econometrics. They are tools that serve different purposes. Just as it does not make sense to argue whether a hammer is better than a screwdriver, we cannot compare structural and non-structural econometrics without specifying the purposes. For the purpose of summarizing data properties and finding stylized facts, non-structural econometrics is better. This purpose is obviously very important in economics. Using a structural econometric model that enforces a certain economic interpretation is not good for this purpose. On the other hand, after finding stylized facts with non-structural econometrics, one may wish to understand causal relationships that explain stylized facts and make policy recommendations based on causal relationships. For that purpose, structural econometrics is better than non-structural econometrics.

Similarly, we do not claim that the definition of “structural” in this book is better than other definitions. For example, Hendry (1993) and Ericsson (1995) define a structural model as an econometric model that is invariant over extensions of
the information set in time, interventions or variables. Their definition is useful for their purpose of finding invariant relationships between economic variables in data, but cannot be used for our purpose of interpreting empirical results in terms of an economic model.

References


Chapter 2

STOCHASTIC PROCESSES

In most macroeconomic models, expectations conditional on information sets are used to model the forecasting conducted by economic agents. Economic agents typically observe stochastic processes of random variables (collections of random variables indexed by time) to form their information sets. This chapter defines the concepts of conditional expectations and information sets for the case of a finite number of elements in the probability space.\footnote{For the general probability space, these concepts are defined with measure theory (see Appendix 2.A). For our purpose, it is not necessary for the reader to understand measure theory.}

2.1 Review of Probability Theory

Since the probability statements made in asymptotic theory involve infinitely many random variables instead of just one random variable, it is important to understand basic concepts in probability theory. Thus, we first review those basic concepts.

Imagine that we are interested in making probability statements about a set of the states of the world (or a probability space), which we denote by $S$. For the purpose of understanding concepts, nothing is lost by assuming that there is a finite number of states of the world. Hence we adopt the simplifying assumption that $S$
consists of \(N\) possible states: \(S = \{s_1, \cdots, s_N\}\). We assign a probability \(\pi_i = Pr(s_i)\) to \(s_i\), depending on how likely \(s_i\) is to occur. It is assumed that \(\sum_{i=1}^{N} \pi_i = 1\) and \(0 \leq \pi_i \leq 1\) for all \(i\). Note that we can now assign a probability to all subsets of \(S\). For example, let \(\Lambda\) be \(\{s_1, s_2\}\). Then the probability that the true \(s\) is in \(\Lambda\) is denoted by \(Pr(s \in \Lambda)\), where \(Pr(s \in \Lambda) = \pi_1 + \pi_2\).

Example 2.1 The state of the world consists of \(s_1\): it rains tomorrow, and \(s_2\): it does not rain tomorrow. According to a weather forecast, \(\pi_1 = 0.8\) and \(\pi_2 = 0.2\).

A random variable assigns a real value to each element \(s\) in \(S\) (that is, it is a real valued function on \(S\)). Let \(X(s)\) be a random variable (we will often omit the arguments \(s\)). For a real value \(x\), the distribution function, \(F(x)\), of the random variable is defined by \(F(x) = Pr\{s : X(s) \leq x\}\). A random variable is assigned an expected value or mean value

\[
E(X) = \sum_{i=1}^{N} X(s_i)\pi_i.
\]

Example 2.2 Continuing Example 2.1, let \(X(s)\) be the profit of an umbrella seller in terms of dollars with \(X(s_1) = 100\) and \(X(s_2) = 10\). Then \(E(X) = 100 \times 0.8 + 10 \times 0.2 = 82\). The distribution function \(F(x)\) is given by \(F(x) = 0\) for \(x < 10\), \(F(x) = 0.2\) for \(10 \leq x < 100\), and \(F(x) = 1\) for \(x \geq 100\).

A random vector is a vector of random variables defined on the set of states. For a \(k\)-dimensional random vector \(X(s) = (X_1(s), \cdots, X_k(s))'\), the joint distribution function \(F\) is defined by

\[
F(x_1, \cdots, x_k) = Pr[X_1 \leq x_1, \cdots, X_k \leq x_k].
\]
2.2 Stochastic Processes

A collection of random variables indexed by time is called a *stochastic process* or a *time series*. Let \( X_t(s) \) be a random variable, then a collection \( \{X_t : X_0(s), X_1(s), X_2(s), \cdots \} \) is a univariate stochastic process. It is sometimes more convenient to consider a stochastic process that starts from the infinite past, \( \{\cdots, X_{-2}(s), X_{-1}(s), X_0(s), X_1(s), X_2(s), \cdots \} \). In general, \( \{X_t(s) : t \in A\} \) for any set \( A \) is a stochastic process. If \( A \) is a set of integers, then time is *discrete*. It is also possible to consider a *continuous* time stochastic process for which the time index takes any real value. For example, \( \{X_t(s) : t \) is a nonnegative real number\}. Here, if we take \( X_t \) as a random vector rather than a random variable, then it is a vector stochastic process. When we observe a sample of size \( T \) of a random variable \( X \) or a random vector \( \mathbf{X} : \{X_1, \cdots, X_T\} \), it is considered a particular realization of a part of the stochastic process.

Note that once \( s \) is determined, the complete history of the stochastic process becomes known. For asymptotic theory, it is usually easier to think about the stochastic nature of economic variables this way rather than the alternative, which is to consider a probability space for each period based on independent disturbances.

In a sense, the stochastic process modeled in this manner is deterministic because everything is determined at the beginning of the world when \( s \) is determined. However, this does not mean that there is no uncertainty to economic agents because they do not learn \( s \) until the end of the world. In order to illustrate this, let us consider the following example:

**Example 2.3** Imagine an economy with three periods and six states of the world. The world begins in period 0. We observe two variables, aggregate output \( (Y_t) \) and
the interest rate \((i_t)\), in period 1 and period 2. The world ends in period 2. In each period, \(Y_t\) can take two values, 150 and 300, and \(i_t\) can take two values, 5 and 10. We assume that \(i_2\) is equal to \(i_1\) in all states of the world, and that the \(i_1 = 5\) in all states in which \(Y_1 = 150\). The six states of the world can be described by the triplet, \([Y_1, i_1, Y_2]\).

The six states of the world are, \(s_1 = [300, 10, 300]\), \(s_2 = [300, 10, 150]\), \(s_3 = [300, 5, 300]\), \(s_4 = [300, 5, 150]\), \(s_5 = [150, 5, 300]\), and \(s_6 = [150, 5, 150]\). To illustrate, \(s_1\) means the economy is in a boom (higher output level) with a high interest rate in period 1, and is in a boom in period 2. In period 0, the economic agents assign a probability to each state: \(\pi_1 = 0.20\), \(\pi_2 = 0.10\), \(\pi_3 = 0.15\), \(\pi_4 = 0.05\), \(\pi_5 = 0.15\), and \(\pi_6 = 0.35\). Unconditional expected values are taken with these probabilities.

In this example, let \(X_t(s) = [Y_t(s), i_t(s)]\). Then \([X_1(s), X_2(s)]\) is a stochastic process. The whole history of the process is determined at the beginning of the world when \(s\) is chosen, and the agents learn which state of the world they are in at the end of the world in period 2. In period 1, however, the agents only have partial information as to which state of the world is true. For example, if \(Y_1 = 300\) and \(i_1 = 5\), the agents learn that they are in either \(s_3\) or \(s_4\), but cannot tell which one they are in until they observe \(Y_2\) in period 2.

### 2.3 Conditional Expectations

Economic agents use available information to learn the true state of the world and make forecasts of future economic variables. This forecasting process can be modeled using conditional expectations.

Information can be modeled as a partition of \(S\) into mutually exclusive subsets:
\[ F = \{ \Lambda_1, \cdots, \Lambda_M \} \] where \( \Lambda_1 \cup \cdots \cup \Lambda_M = S \), and \( \Lambda_j \cap \Lambda_k = \emptyset \) if \( j \neq k \). For example, information \( F \) consists of two subsets: \( F = \{ \Lambda_1, \Lambda_2 \} \). Here \( \Lambda_1 = \{ s_1, \cdots, s_M \} \), and \( \Lambda_2 = \{ s_{M+1}, \cdots, s_N \} \). The information represented by \( F \) tells us which \( \Lambda \) contains the true \( s \), but no further information is given by \( F \).

In this situation, once agents obtain the information represented by \( F \), then the agents know which subset contains the true \( s \), and they can assign a probability of zero to all elements in the other subset. There is no reason to change the ratios of probabilities assigned to the elements in the subset containing the true \( s \). Nonetheless, the absolute level of each probability should be increased, so that the probabilities add up to one. The probability conditional on the information that the true \( s \) is in \( \Lambda_j \) is denoted by \( Pr\{s_i|s \in \Lambda_j\} \). The considerations given above lead to the following definition of conditional probability:

\[
Pr\{s_i|s \in \Lambda_j\} = \frac{Pr\{s_i\}}{Pr\{s \in \Lambda_j\}},
\]

when \( s_i \) is in \( \Lambda_j \). Here each probability is scaled by the probability of the subset containing the true \( s \), so that the probabilities add up to one.

We use conditional probability to define the conditional expectation. The expectation of a random variable \( Y \) conditional on the information that the true \( s \) is in \( \Lambda_j \) is

\[
E(Y|s \in \Lambda_j) = \sum_{s \in \Lambda_j} Y(s) \frac{Pr\{s_i\}}{Pr\{s \in \Lambda_j\}},
\]

where the summation is taken over all \( s \) in \( \Lambda_j \).

It is convenient to view the conditional expectation as a random variable. For this purpose, the conditional expectation needs to be defined over all \( s \) in \( S \), not just for \( s \) in a particular \( \Lambda_j \). Given each \( s \), we first find out which \( \Lambda \) contains \( s \).
When \( \Lambda_j \) contains \( s \), the expected value of \( Y \) conditional on \( \mathcal{F} \) for \( s \) is given by
\[
E(Y|\mathcal{F})(s) = E(Y|s \in \Lambda_j).
\]

Instead of a partition, we can use a random variable or a random vector to describe information. Consider information represented by a partition \( \mathcal{F} = \{ \Lambda_1, \cdots, \Lambda_M \} \).

Consider the set \( I \), which consists of all random variables that take the same value for all elements in each \( \Lambda_j \):
\[
I = \{ X(s) : X(s_i) = X(s_k) \text{ if } s_i \in \Lambda_j \text{ and } s_k \in \Lambda_j \text{ for all } i, j, k \}.
\]
Then the information set \( I \) represents the same information as \( \mathcal{F} \) does. A random variable \( X \) is said to be in this information set, when \( X(s_i) = X(s_k) \) if both \( s_i \) and \( s_k \) are in the same \( \Lambda_j \).

A random vector \( X \) is said to be in this information set when each element of \( X \) is in the information set.

If \( X \) is in the information set \( I \), and if \( X \) takes on different values for all different \( \Lambda \) \( (X(s_i) \neq X(s_k) \text{ when } s_i \text{ and } s_k \text{ are not in the same } \Lambda) \), then we say that the random variable \( X \) generates the information set \( I \). If a random vector \( X \) is in \( I \), and if at least one element of \( X \) takes on different values for different \( \Lambda \), then the random vector \( X \) is said to generate the information set \( I \). When a random variable \( X \) or a random vector \( X \) generates the information set \( I \), which represents the same information as a partition \( \mathcal{F} \), we define \( E(Y|I) \) as \( E(Y|\mathcal{F}) \). If \( I \) is generated by \( X \), we define \( E(Y|X) = E(Y|I) \); and if \( I \) is generated by a random vector \( X \), we define \( E(Y|X) = E(Y|I) \). It should be noted that \( E(Y|I) \) is in the information set \( I \).

**Example 2.4** Continuing Example 2.3, let \( I \) be the information set generated by \( X_1 = (Y_1, i_1) \), and let \( \mathcal{F} \) be the partition that represents the same information as \( I \). Then \( \mathcal{F} = \{ \Lambda_1, \Lambda_2, \Lambda_3 \} \), where \( \Lambda_1 = \{ s_1, s_2 \} \), \( \Lambda_2 = \{ s_3, s_4 \} \), and \( \Lambda_3 = \{ s_5, s_6 \} \).

---

\(^2\)In the terminology of probability theory, we consider a set of all possible unions of \( \Lambda \)'s in \( \mathcal{F} \) plus the null set. This set of subsets of \( S \) is called a \( \sigma \)-field, and used to describe information. When a random variable \( X \) is in the information set \( I \), we say that the random variable is measurable with respect to this \( \sigma \)-field.
2.3. CONDITIONAL EXPECTATIONS

Using (2.3), \( Pr(s_1|s \in \Lambda_1) = \frac{0.20}{0.20+0.10} = \frac{2}{3} \) and \( Pr(s_2|s \in \Lambda_1) = \frac{0.10}{0.20+0.10} = \frac{1}{3} \).

Hence \( E(Y_2|s \in \Lambda_1) = 300 \times \frac{2}{3} + 150 \times \frac{1}{3} = 250 \). Similarly, \( Pr(s_3|s \in \Lambda_2) = \frac{3}{4} \), \( Pr(s_4|s \in \Lambda_2) = \frac{1}{4} \), \( Pr(s_5|s \in \Lambda_3) = \frac{3}{10} \), \( Pr(s_6|s \in \Lambda_3) = \frac{7}{10} \), \( E(Y_2|s \in \Lambda_2) = 262.5 \), and \( E(Y_2|s \in \Lambda_3) = 195 \). Hence the random variable \( E(Y_2|I) \) is given by

\[
E(Y_2|I)(s) = \begin{cases} 
250 & \text{if } s \in \Lambda_1 \\
262.5 & \text{if } s \in \Lambda_2 \\
195 & \text{if } s \in \Lambda_3
\end{cases}
\]

Example 2.5 Continuing Example 2.4, consider the information set \( J \) which is generated by \( Y_1 \). Then \( J \) is a smaller information set than \( I \) in the sense that \( J \subset I \).

Similar computations as those in Example 2.4 yield

\[
E(Y_2|J)(s) = \begin{cases} 
255 & \text{if } s \in \{s_1, s_2, s_3, s_4\} \\
195 & \text{if } s \in \{s_5, s_6\}
\end{cases}
\]

Two properties of conditional expectations are very important in macroeconomics.

Proposition 2.1 (Properties of Conditional Expectations)

(a) If a random variable \( Z \) is in the information set \( I \), then

\[
E(ZY|I) = ZE(Y|I) \tag{2.7}
\]

for any random variables \( Y \) with finite \( E(|Y|) \), assuming that \( E(|ZY|) \) is finite.

(b) The Law of Iterated Expectations: If the information set \( J \) is smaller than the information set \( I \) \( (J \subset I) \), then

\[
E(Y|J) = E[E(Y|I)|J] \tag{2.8}
\]

for any random variable \( Y \) with finite \( E(|Y|) \).
CHAPTER 2. STOCHASTIC PROCESSES

Expectation can be viewed as a special case of conditional expectation in which the information set consists of constants. Since a constant is a random variable which takes the same value for all states of the world, any information set includes all constants. Therefore, the Law of Iterated Expectations implies

\[(2.9) \quad E(Y) = E[E(Y|I)].\]

When we wish to emphasize the difference between expectations and conditional expectations, expectations are called unconditional expectations. Relation (2.9) states that an unconditional expected value of a random variable \(Y\) can be computed as an unconditional expected value of the expectation of the random variable conditional on any information set. For a proof of Proposition 2.1 in the general case, see, e.g., Billingsley (1986, Theorem 34.3 and Theorem 34.4).

2.4 Stationary Stochastic Processes

A stochastic process \(\{\cdots, X_{-1}, X_0, X_1, \cdots\}\) is strictly stationary if the joint distribution function of \((X_t, X_{t+1}, \cdots, X_{t+h})\) is the same for all \(t = 0, \pm 1, \pm 2, \cdots\) and all \(h = 0, 1, 2, \cdots\). A stochastic process \(\{\cdots, X_{-1}, X_0, X_1, \cdots\}\) is covariance stationary (or weakly stationary) if \(X_t\) has finite second moments \((E(X_tX_t') < \infty)\) and if \(E(X_t)\) and \(E(X_tX_{t-h}')\) do not depend on the date \(t\) for all \(t = 0, \pm 1, \pm 2, \cdots\) and all \(h = 0, 1, 2, \cdots\).

Because all moments are computed from distribution functions, if \(X_t\) is strictly stationary and has finite second moments, then it is also covariance stationary. If \(X_t\) is covariance stationary, then its mean \(E(X_t)\) and its \(h\)-th autocovariance \(\Phi(h) = E[(X_t - E(X_t))(X_{t-h} - E(X_{t-h}))'] = E(X_tX_{t-h}') - E(X_t)E(X_{t-h}')\) does not depend on date \(t\).
Proposition 2.2 If a \(k\)-dimensional vector stochastic process \(X_t\) is strictly stationary, and if a continuous function \(f(\cdot) : \mathbb{R}^k \rightarrow \mathbb{R}^p\) does not depend on date \(t\), then \(f(X_t)\) is also strictly stationary.\(^3\)

This follows from the fact that the distribution function of \(f(X_t), f(X_{t+1}), \ldots, f(X_{t+h})\) is determined by \(f\) and the joint distributions of \(X_t, X_{t+1}, \ldots, X_{t+h}\) (see Appendix 2.A). Proposition 2.2 will be used frequently to derive the cointegrating properties of economic variables from economic models in Chapter 13.

The next proposition is for covariance stationary processes.

Proposition 2.3 If a \(k\)-dimensional vector stochastic process \(X_t\) is covariance stationary, and if a linear function \(f(\cdot) : \mathbb{R}^k \rightarrow \mathbb{R}^p\) does not depend on date \(t\), then \(f(X_t)\) is also covariance stationary.

This proposition is true because \(f(X_t)\) has finite second moments, and the first and second moments of \(f(X_t)\) do not depend on date \(t\). However, unlike Proposition 2.2 for strictly stationary processes, a nonlinear function of a covariance stationary process may not be covariance stationary. For example, suppose that \(X_t\) is covariance stationary. Imagine that \(X_t\)'s variance is finite but \(E(|X_t|^4) = \infty\). Consider \(Z_t = f(X_t) = (X_t)^2\). Then \(Z_t\)'s variance is not finite, and hence \(Z_t\) is not covariance stationary.

In order to model strictly stationary and covariance stationary processes, it is convenient to consider white noise processes. A univariate stochastic process \(\{e_t : t = \)


\[ \cdots, -1, 0, 1, \cdots \] is \textit{white noise} if \( E(e_t) = 0 \), and

\[
E(e_t e_j) = \begin{cases} 
\sigma^2 & \text{if } t = j \\
0 & \text{if } t \neq j
\end{cases},
\]

where \( \sigma \) is a constant. For a vector white noise, we require

\[
E(e_t e'_j) = \begin{cases} 
\Sigma & \text{if } t = j \\
0 & \text{if } t \neq j
\end{cases},
\]

where \( \Sigma \) is a matrix of constants. A white noise process is covariance stationary.

If a process is independent and identically distributed (i.i.d.), then it is strictly stationary. The simplest example of an i.i.d. process is an i.i.d. white noise. A \textit{Gaussian white noise process} \( \{e_t: -\infty < t < \infty\} \) is an i.i.d. white noise process for which \( e_t \) is normally distributed with zero mean. In these definitions, \( e_t \) can be a vector white noise process.

All linear functions of white noise random variables are covariance stationary because of Proposition 2.3. In addition, by Proposition 2.2, all functions of i.i.d. white noise random variables are strictly stationary. A simple example of this case is:

\textbf{Example 2.6} Let \( X_t = \delta + e_t \), where \( e_t \) is a white noise process, and \( \delta \) is a constant. Then \( E(X_t) = \delta \), and \( X_t \) is covariance stationary. If \( e_t \) is an i.i.d. white noise process, then \( X_t \) is strictly stationary.

If \( X_t \) is strictly stationary with finite second moments, \( X_t \) is covariance stationary. Therefore, \( X_t \)'s first and second moments cannot depend on date \( t \). In empirical work, the easiest case to see that an observed variable is \textit{not} strictly stationary is when a variable’s mean shifts upward or downward over time. A simple example of this case is:
Example 2.7 Let $X_t = \delta + \theta t + e_t$, where $e_t$ is an i.i.d. white noise random variable and $\delta$ and $\theta \neq 0$ are constants. Then $X_t$ is not stationary because $E(X_t) = \delta + \theta t$ depends on time.\footnote{Because $X_t$ is stationary after removing a deterministic trend in this example, we say that $X_t$ is trend stationary as we will discuss in Chapter 11. Trend stationarity is a way to model nonstationarity.}

Strictly stationary and covariance stationary processes can be serially correlated, that is, their $h$-th order autocovariances can be nonzero for $h \neq 0$ as in the next two examples.

Example 2.8 \textit{(The first order Moving Average Process)} Let $X_t = \delta + e_t + Be_{t-1}$, where $e_t$ is a white noise which satisfies (2.10), and $\delta$ and $B$ are constant. This is a moving average process of order 1 (see Chapter 4). Then $X_t$ is covariance stationary for any $B$ because of Proposition 2.3.\footnote{Even though $X_t$ is stationary for any $B$, it is often convenient to impose a restriction $|B| \leq 1$ as explained in Chapter 4.} $E(X_t) = \delta$, and its $h$-th autocovariance is

\begin{equation}
\phi_h = E[(X_t - \delta)(X_{t-h} - \delta)] = \begin{cases} 
\sigma^2 (1 + B^2) & \text{if } h = 0 \\
\sigma^2 & \text{if } |h| = 1 \\
0 & \text{if } |h| > 1
\end{cases}.
\end{equation}

In this example, if $e_t$ is an i.i.d. white noise, then $X_t$ is strictly stationary.

Example 2.9 \textit{(The first order Autoregressive Process)} Consider a process $X_t$ which is generated from an initial random variable $X_0$, where

\begin{equation}
X_t = AX_{t-1} + e_t \quad \text{for } t \geq 1,
\end{equation}

where $e_t$ is a Gaussian white noise random variable, and $A$ is a constant. This is an autoregressive process of order 1 (see Chapter 4). If $|A| < 1$ and $X_0$ is a normally distributed random variable with mean zero and variance of $\frac{\text{Var}(e_t)}{1-A^2}$, then $X_t$ is strictly
stationary (see Exercise 2.3). The methods explained in Chapter 4 can be used to show that \( X_t \) is not strictly stationary when \( X_0 \)’s distribution is different from the one given above.

2.5 Conditional Heteroskedasticity

Using conditional expectations, we can define variance and covariance conditional on an information set just as we use unconditional expectations to define (unconditional) variance and covariance. The variance of \( Y \) conditional on an information set \( I \) is

\[
\text{Var}(Y|I) = E[(Y - E(Y|I))^2|I],
\]

and the covariance of \( X \) and \( Y \) conditional on an information set \( I \) is

\[
\text{Cov}(X, Y|I) = E[(X - E(X|I))(Y - E(Y|I))|I].
\]

Consider a stochastic process \([Y_t : t \geq 1]\). If the unconditional variance of \( Y_t \), \( \text{Var}(Y_t) \), depends on date \( t \), then the \( Y_t \) is said to be heteroskedastic; if not, it is homoskedastic. If \( Y_t \)'s variance conditional on an information set \( I_t \), \( \text{Var}(Y_t|I_t) \), is constant and does not depend on the information set, then \( Y_t \) is said to be conditionally homoskedastic; if not, it is conditionally heteroskedastic.

Example 2.10 Let \( Y_t = \delta + h_t e_t \), where \( e_t \) is an i.i.d. white noise with unit variance \((E(e_t^2) = 1)\), and \( \{h_t : -\infty < t < \infty\} \) is a sequence of real numbers. Then the (unconditional) variance of \( Y_t \) is \( h_t \), and \( Y_t \) is heteroskedastic as long as \( h_t \neq h_j \) for some \( t \) and \( j \).

A heteroskedastic process is not strictly stationary because its variance depends on date \( t \). It should be noted, however, that a strictly stationary random variable can
be conditionally heteroskedastic. This fact is important because many of the financial
time series have been found to be conditionally heteroskedastic. For example, the
growth rates of asset prices and foreign exchange rates can be reasonably modeled
as strictly stationary processes. However, the volatility of such a growth rate at a
point in time tends to be high if it has been high in the recent past. Therefore, such
a growth rate is often modeled as a conditionally heteroskedastic process. A popular
method to model conditional heteroskedasticity, introduced by Engle (1982), is an
autoregressive conditional heteroskedastic (ARCH) process. The following is a simple
example of an ARCH process.

**Example 2.11 (An ARCH Process)** Let $I_t$ be an information set, and $e_t$ be a uni-
variate stochastic process such that $e_t$ is in $I_t$, and $E(e_t|I_{t-1}) = 0$. Assume that

$$e_t^2 = \eta + \alpha e_{t-1}^2 + w_t,$$

where $\eta > 0$, $w_t$ is another white noise process in $I_t$ with $E(w_t|I_{t-1}) = 0$ and

$$E(w_k w_j|I_t) = \begin{cases} 
\lambda^2 & \text{if } k = j \\
0 & \text{if } k \neq j 
\end{cases},$$

where $\lambda$ is a constant. Relation (2.16) implies that $e_t$'s conditional variance depends
on $I_t$:

$$E(e_t^2|I_{t-1}) = \eta + \alpha e_{t-1}^2,$$

and thus $e_t$ is conditionally heteroskedastic.

In order to see whether or not $e_t$'s unconditional variance is constant over time, take
expectations of both sides of (2.18) to obtain

$$E(e_t^2) = \eta + \alpha E(e_{t-1}^2).$$
Hence if the variance of $e_t$ is a constant $\sigma^2$, then $\sigma^2 = \eta + \alpha \sigma^2$, and $\sigma^2 = \frac{\eta}{1 - \alpha}$. Because $\sigma^2$ is positive, this equation implies that $\alpha < 1$. When $\alpha < 1$, an ARCH process can be covariance stationary and strictly stationary.

### 2.6 Martingales and Random Walks

Consider a stochastic process $[Y_t: -\infty < t < \infty]$, and a sequence of information sets $[I_t: -\infty < t < \infty]$ that is increasing ($I_t \subset I_{t+1}$). If $Y_t$ is in $I_t$ and if

$$E(Y_{t+1}|I_t) = Y_t,$$

then $Y_t$ is a martingale adapted to $I_t$. Rational expectations often imply that an economic variable is a martingale (see Section 3.2). If $Y_t$ is a martingale adapted to $I_t$ and if its conditional variance, $E((Y_{t+1} - Y_t)^2|I_t)$, is constant (that is, $Y_t$ is conditionally homoskedastic), then $Y_t$ is a random walk.

As we will discuss later in this book, most of the rational expectations models imply that certain variables are martingales. The models typically do not imply that the variables are conditionally homoskedastic, and hence do not imply that they are random walks. However, if the data for the variable does not show signs of conditional heteroskedasticity, then we may test whether or not a variable is a random walk. It is often easier to test whether or not the variable is a random walk than to test whether or not it is a martingale.

Consider a stochastic process $[e_t: -\infty < t < \infty]$, and a sequence of information sets $[I_t: -\infty < t < \infty]$ which is increasing ($I_t \subset I_{t+1}$). If $e_t$ is in $I_t$ and if

$$E(e_{t+1}|I_t) = 0,$$

then $e_t$ is a martingale difference sequence adapted to $I_t$. If $Y_t$ is a martingale adapted
to $I_t$, then $e_t = Y_t - Y_{t-1}$ is a martingale difference sequence (see Exercise 2.4). A covariance stationary martingale difference sequence is a white noise process (see Exercise 2.5). However, a white noise process may not be a martingale difference sequence for any sequence of information sets. An i.i.d. white noise process is a martingale difference sequence (see Exercise 2.6).

In these definitions, a martingale or a martingale difference sequence can be a vector stochastic process.

Appendix

2.A A Review of Measure Theory

Let $S$ be an arbitrary nonempty set of points $s$. An event is a subset of $S$. A set of subsets is called a class. A set $\mathcal{F}$ of subsets of $S$ is called a field if

(i) $S \in \mathcal{F}$;

(ii) $A \in \mathcal{F}$ implies $A^c \in \mathcal{F}$, where $A^c$ is the complement of $A$;

(iii) $A, B \in \mathcal{F}$ implies $A \cup B \in \mathcal{F}$.

A class $\mathcal{F}$ is a $\sigma$-field if it is a field and if

(iv) $A_1, A_2, \cdots \in \mathcal{F}$ implies $A_1 \cup A_2 \cup \cdots \in \mathcal{F}$.

A set function is a real-valued function defined on some class of subsets of $S$. A set function $Pr$ on a field $\mathcal{F}$ is a probability measure if it satisfies these conditions:

(i) $0 \leq Pr(A) \leq 1$ for $A \in \mathcal{F}$;

(ii) $Pr(0) = 0, Pr(S) = 1$;
(iii) if \( A_1, A_2, \cdots \) is a disjoint sequence of \( \mathcal{F} \)-sets and if \( \bigcup_{k=1}^{\infty} A_k \in \mathcal{F} \), then
\[
Pr(\bigcup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} Pr(A_k).
\]

If \( \mathcal{F} \) is a \( \sigma \)-field in \( S \) and \( Pr \) is a probability measure on \( \mathcal{F} \), the triple \( (S, \mathcal{F}, Pr) \) is called a \textit{probability space}. Given a class \( \mathcal{A} \), consider the class which is the intersection of all \( \sigma \)-fields containing \( \mathcal{A} \). This class is the smallest \( \sigma \)-field which contains \( \mathcal{A} \), and is called the \( \sigma \)-field generated by \( \mathcal{A} \) and is denoted by \( \sigma(\mathcal{A}) \).

**Proposition 2.A.1** A probability measure on a field has a unique extension to the generated \( \sigma \)-field.

In Euclidean \( k \)-space \( \mathbb{R}^k \), consider the class of the bounded rectangles
\[
[x = (x_1, \cdots, x_k) : a_i \leq x \leq b_i, i = 1, \cdots, k].
\]
The \( \sigma \)-field generated from this class is called the \textit{k-dimensional Borel sets}, and denoted by \( \mathcal{R}^k \).

Let \( \mathcal{F} \) be a \( \sigma \)-field of subsets of \( S \) and \( \mathcal{F}' \) be a \( \sigma \)-field of subsets of \( S' \). For a mapping \( T : S \mapsto S' \), consider the inverse images \( T^{-1}(A') = \{ s \in S : T(s) \in A' \} \). The mapping \( T \) is measurable \( \mathcal{F}/\mathcal{F}' \) if \( T^{-1}(A') \in \mathcal{F} \) for each \( A' \in \mathcal{F}' \).

For a real-valued function \( f \), the image space \( S' \) is the line \( \mathbb{R}^1 \), and in this case \( \mathcal{R}^1 \) is always tacitly understood to play the role of \( \mathcal{F}' \). A real-valued function on \( S \) is measurable \( \mathcal{F} \) (or simply measurable when it is clear from the context what \( \mathcal{F} \) is involved) if it is measurable \( \mathcal{F}/\mathcal{R}^1 \). If \( (S, \mathcal{F}, Pr) \) is a probability space, then a real-valued measurable function is called a \textit{random variable}. For a random variable \( X \), we can assign a probability to the event that \( X(s) \) belongs to a Borel set \( \mathcal{B} \) by \( Pr(X^{-1}(\mathcal{B})) \).
For a mapping \( f : S \mapsto \mathbb{R}^k \), \( \mathbb{R}^k \) is always understood to be the \( \sigma \)-field in the image space. If \( (S, \mathcal{F}, P) \) is a probability space, then a measurable mapping \( X : S \mapsto \mathbb{R}^k \) is called a random vector. It is known that \( X \) is a random vector if and only if each component of \( X \) is a random variable.

A mapping \( f : \mathbb{R}^i \mapsto \mathbb{R}^k \) is defined to be measurable if it is measurable \( \mathcal{R}^i/\mathcal{R}^k \). Such functions are called Borel functions.

**Proposition 2.A.2** If \( f : \mathbb{R}^i \mapsto \mathbb{R}^k \) is continuous, then it is measurable.

If \( X \) is a \( j \)-dimensional random vector, and \( g : \mathbb{R}^j \mapsto \mathbb{R}^i \) is measurable, then \( g(X) \) is an \( i \)-dimensional random vector. If the distribution of \( X \) is \( \mu \), the distribution of \( g(X) \) is \( \mu g^{-1} \). Proposition 2.2 can be proven by taking \( X = [Y_1', \cdots, Y_{t+k}]' \).

We now introduce two definitions of conditional expectation. One definition is standard in measure theory. The other definition is given because it is convenient for the purpose of stating a version of the conditional Gauss-Markov theorem used in this book. Intuitively, the conditional Gauss-Markov theorem is obtained by stating all assumptions and results of the Gauss-Markov theorem conditional on the stochastic regressors. Formally, it is necessary to make sure that the conditional expectations of the relevant variables are well defined.

Let \( S \) be a probability space, \( \mathcal{F} \) be a \( \sigma \)-field of \( S \), and \( P \) be a probability measure defined on \( \mathcal{F} \). The random variables we will consider in this section are defined on this probability space. Let \( X = (X_1, X_2, \ldots, X_T)' \) be a \( T \times K \) matrix of random variables, which will be the regressor matrix of the regression to be considered. Let \( y = (y_1, y_2, \ldots, y_T) \) and \( e = (e_1, e_2, \ldots, e_T) \) be \( T \times 1 \) vectors of random variables. We are concerned with a linear model of the form: \( y = Xb_0 + e \), where \( b_0 \) is a \( K \times 1 \) vector of real numbers.
For \( s \) such that \( X(s)'X(s) \) is nonsingular, the OLS estimator is

\[
(2.A.1) \quad b_T = (X'X)^{-1}X'y.
\]

In order to apply a conditional version of the Gauss-Markov Theorem, it is necessary to define the expectation and variance of \( b_T \) conditional on \( X \).

Let \( Z \) be an integrable random variable (namely, \( E(|Z|) < \infty \)), and \( \sigma(X) \) be the smallest \( \sigma \)-field with respect to which the random variables in \( X \) are measurable.

The standard definition of the expectation of \( Z \) given \( X \) is obtained by applying the Radon-Nikodym theorem (see, e.g., Billingsley, 1986). Throughout this paper, we use the notation \( E[Z|\sigma(X)] \) to denote the usual conditional expectation of \( Z \) conditional on \( X \) as defined by Billingsley (1986) for a random variable \( Z \).\(^6\) \( E[Z|\sigma(X)] \) is a random variable, and \( E[Z|\sigma(X)]_s \) denotes the value of the random variable at \( s \) in \( S \). It satisfies the following two properties:

(i) \( E(Z|\sigma(X)) \) is measurable and integrable given \( \sigma(X) \).

(ii) \( E(Z|\sigma(X)) \) satisfies the functional equation:

\[
(2.A.2) \quad \int_G E(Z|\sigma(X))dPr = \int_G ZdPr, \quad G \in \sigma(X).
\]

There will in general be many such random variables which satisfy these two properties; any one of them is called a version of \( E(Z|\sigma(X)) \). Any two versions are equal with probability 1.

It should be noted that this definition is given under the condition that \( Z \) is integrable, namely \( E(|Z|) < \infty \). This condition is too restrictive when we define

\(^6\)If \( z \) is a vector, the conditional expectation is taken for each element in \( z \).
the conditional expectation and variance of the OLS estimator in many applications because the moments of \((X'X)^{-1}\) may not be finite even when \(X\) has many finite moments. For this reason, it is difficult to confirm that \(E(b_T|\sigma(X))\) can be defined in each application even if \(X\) is normally distributed. Thus, Judge et al. (1985) conclude that the Gauss-Markov theorem based on \(E(\cdot|\sigma(X))\) is not very useful.

We avoid this problem by adopting a different definition of conditional expectation based on conditional distribution. For this purpose, we first define conditional probabilities following Billingsley (1986). Given \(A\) in \(\mathcal{F}\), define a finite measure \(v\) on \(\sigma(X)\) by \(v(G) = \Pr(A \cap G)\) for \(G\) in \(\sigma(X)\). Then \(\Pr(G) = 0\) implies that \(v(G) = 0\). The Radon-Nikodym theorem can be applied to the measures \(v\) and \(\Pr\), and there exists a random variable \(f\) that is measurable and integrable with respect to \(\Pr\), such that \(\Pr(A \cap G) = \int_G fd\Pr\) for all \(G\) in \(\sigma(X)\). Denote this random variable by \(\Pr(A|\sigma(G))\). This random variable satisfies these two properties:

(i) \(\Pr(A|\sigma(X))\) is measurable and integrable given \(\sigma(X)\).

(ii) \(\Pr(A|\sigma(X))\) satisfies the functional equation

\[
(2.A.3) \quad \int_G \Pr(A|\sigma(X)) d\Pr = \Pr(A \cap G), \quad G \in \sigma(X).
\]

There will in general be many such random variables, but any two of them are equal with probability 1. A specific such random variable is called a version of the conditional probability.

Given a random variable \(Z\), which may not be integrable, we define a conditional distribution \(\mu(\cdot, s)\) given \(X\) for each \(s\) in \(S\). Let \(\mathcal{R}^1\) be the \(\sigma\)-field of the Borel sets

\[\text{Loeve (1978) slightly relaxes this restriction by defining the conditional expectation for any random variable whose expectation exists (but may not be finite) with an extension of the Radon-Nikodym theorem. This definition can be used for } E(\cdot|\sigma(X)), \text{ but this slight relaxation does not solve our problem.}\]
in $\mathbb{R}^1$. By Theorem 33.3 in Billingsley (1986, p.460), there exists a function $\mu(H, s)$, defined for $H$ in $\mathbb{R}^1$ and $s$ in $S$, with these two properties:

(i) For each $s$ in $S$, $\mu(H, s)$ is, as a function of $H$, a probability measure on $\mathbb{R}^1$.

(ii) For each $H$ in $\mathbb{R}^1$, $\mu(H, s)$ is, as a function of $s$, a version of $Pr(Z \in H|\sigma(X))_s$.

For each $s$ in $S$, we define $E(Z|X)_s$ to be $\int_{\mathbb{R}^1} z \mu(dz, s)$. It should be noted that $E(Z|X)_s$ does not necessarily satisfy the usual properties of conditional expectation such as the law of iterated expectations. In general, $E(Z|X)_s$ may not even exist for some $s$. If $\int_{\mathbb{R}^1} |z| \mu(dz, s)$ is finite, then, $E(Z|X)_s$ is said to exist and be finite.

Given a $T \times K$ matrix of real numbers $x$, $E(Z|X)_s$ is identical for all $s$ in $X^{-1}(x)$. Therefore, we define $E(Z|X = x)$ as $E(Z|X)_s$ for $s$ in $X^{-1}(x)$. This is the definition of the conditional expectation of $Z$ given $X = x$ in this paper.

We are concerned with a linear model of the form:

**Assumption 2.A.1** $y = Xb_0 + e$

where $b_0$ is a $K \times 1$ vector of real numbers. Given a $T \times K$ matrix of real numbers $x$, we assume that the conditional expectation of $e$ given $X = x$ is zero:

**Assumption 2.A.2** $E[e|X = x] = 0$.

Next, we assume that $e$ is homoskedastic and $e_t$ is not serially correlated given $X = x$:

**Assumption 2.A.3** $E[ee'|X = x] = \sigma^2 I_T$.

The OLS estimator can be expressed by (2.A.1) for all $s$ in $X^{-1}(x)$ when the next assumption is satisfied:

**Assumption 2.A.4** $x'x$ is nonsingular.
Under Assumptions 2.A.1–2.A.4, \( E[b_T|X = x] = b_0 \) and \( E[(b_T - b_0)'(b_T - b_0)|X = x] = \sigma^2(x'x)^{-1} \). The conditional version of the Best Linear Unbiased Estimator (BLUE) given \( X = x \) can be defined as follows: An estimator \( b_T \) for \( b_0 \) is BLUE conditional on \( X = x \) if \( 1 \) \( b_T \) is linear conditional on \( X = x \), namely, \( b_T \) can be written as \( b_T = Ay \) for all \( s \) in \( X^{-1}(x) \) where \( A \) is a \( K \times T \) matrix of real numbers; \( 2 \) \( b_T \) is unbiased conditional on \( X = x \), namely, \( E(b_T|X = x) = b \); \( 3 \) for any linear unbiased estimator \( b^* \) conditional on \( X = x \), \( E[(b_T - b_0)(b_T - b_0)'|X = x] \leq E[(b^* - b_0)(b^* - b_0)'|X = x] \), namely, \( E[(b_T - b_0)(b_T - b_0)'|X(s) = x] - E[(b_T - b_0)(b_T - b_0)'|X(s) = x] \) is a positive semidefinite matrix.

With these preparations, the following theorem can be stated:

**Theorem 2.A.1** (The Conditional Gauss-Markov Theorem) Under Assumptions 2.A.1–2.A.4, the OLS estimator is BLUE conditional on \( X = x \).

Applying any of the standard proofs of the (unconditional) Gauss-Markov theorem can prove this theorem by replacing the unconditional expectation with \( E(\cdot|X = x) \).

Modifying some assumptions and adding another yields the textbook version of the conditional Gauss-Markov theorem based on \( E(\cdot|\sigma(X)) \).

**Assumption 2.A.2’** \( E[e|\sigma(X)] = 0 \).

Since \( E[e|\sigma(X)] \) is defined only when each element of \( e \) is integrable, Assumption 2.A.2’ implicitly assumes that \( E(e) \) exists and is finite. It also implies \( E(e) = 0 \) because of the law of iterated expectations. Given \( E(e) = 0 \), a sufficient condition for Assumption 2.A.2’ is that \( X \) is statistically independent of \( e \). Since Assumption 2.A.2’ does not imply that \( X \) is statistically independent of \( e \), Assumption 2.A.2’
is weaker than the assumption of independent stochastic regressors. With the next assumption, we assume that $e$ is conditionally homoskedastic and $e_t$ is not serially correlated:

**Assumption 2.A.3’** $E[ee'|\sigma(X)] = \sigma^2 I_T$.

The next assumption replaces Assumption 2.A.4.

**Assumption 2.A.4’** $XX'$ is nonsingular with probability one.

From Assumption 2.A.1, $b_T = b_0 + (X'X)^{-1}X'e$. Hence we can prove a version of the conditional Gauss-Markov theorem based on $E(\cdot|\sigma(X))$ when the expectations of $(X'X)^{-1}X'e$ and $(X'X)^{-1}X'ee'X(X'X)^{-1}$ exist and are finite. For this purpose, we consider the following assumption:

**Assumption 2.A.5** $E[\text{trace}((X'X)^{-1}X'ee'X(X'X)^{-1})]$ exists and is finite.

The problem with Assumption 2.A.5 is that it is not easy to verify the assumption for many distributions of $X$ and $e$ that are often used in applications and Monte Carlo studies. However, a sufficient condition for Assumption 2.A.5 is that the distributions of $X$ and $e$ have finite supports.

Under Assumptions 2.A.1, 2.A.2’–2.A.4’, and 2.A.5,

$$E(b_T|\sigma(X)) = b_0 + E[(X'X)^{-1}X'e|\sigma(X)] = b_0.$$ 

Moreover, $E[(b_T - b_0)'(b_T - b_0)|\sigma(X)]$ can be defined, and $E[(b_T - b_0)'(b_T - b_0)|\sigma(X)] = E[(X'X)^{-1}X'ee'X(X'X)^{-1}|\sigma(X)] = (X'X)^{-1}X'E[ee'|\sigma(X)]X'(X'X)^{-1} = \sigma^2(X'X)^{-1}$.

We now consider a different definition of the conditional version of the Best Linear Unbiased Estimator (BLUE). The *Best Linear Unbiased Estimator (BLUE)*
conditional on $\sigma(X)$ is defined as follows. An estimator $b_T$ for $b_0$ is BLUE conditional on $\sigma(X)$ in $H$ if (1) $b_T$ is linear conditional on $\sigma(X)$, namely, $b_T$ can be written as $b_T = Ay$ where $A$ is a $K \times T$ matrix, and each element of $A$ is measurable given $\sigma(X)$; (2) $b_T$ is unbiased conditional on $\sigma(X)$ in $G$, equivalently, $E(b_T|\sigma(X)) = b_0$, (3) for any linear unbiased estimator $b^*$ conditional on $\sigma(X)$ for which $E(b^*b^*)$ exists and is finite, $E[(b_T - b_0)(b_T - b_0)'|\sigma(X)] \leq E[(b^* - b_0)(b^* - b_0)'|\sigma(X)]$ with probability 1, namely, $E[(b^* - b_0)(b^* - b_0)'|\sigma(X)] - E[(b_T - b_0)(b_T - b_0)'|\sigma(X)]$ is a positive semidefinite matrix with probability 1.

**Proposition 2.A.3** Under Assumptions 2.A.1, 2.A.2’–2.A.4’, and 2.A.5, the OLS estimator is BLUE conditional on $\sigma(X)$. Moreover, it is unconditionally unbiased and has the minimum unconditional covariance matrix among all linear unbiased estimators conditional on $\sigma(X)$.

**Proof** The proof of this proposition is given in Greene (1997, Section 6.7).

In this proposition, the covariance matrix of $b_T$ is $\sigma^2 E[(X'X)^{-1}]$, which is different from $\sigma^2[E(X'X)]^{-1}$. This property may seem to contradict the standard asymptotic theory, but it does not. Asymptotically, $(1/T)X'X$ converges almost surely to $E[X|X_t]$ if $X_t$ is stationary and ergodic. Hence the limit of the covariance matrix of $\sqrt{T}(b_T - b_0), \sigma^2 E[(1/T)(X'X)]^{-1}$, is equal to the asymptotic covariance matrix, $\sigma^2[E(X|X_t)]^{-1}$.

In order to study the distributions of the $t$ ratios and $F$ test statistics we need an additional assumption:

**Assumption 2.A.6** Conditional on $X$, $e$ follows a multivariate normal distribution.
Given a $1 \times K$ vector of real numbers $R$, consider a random variable

$$N_R = \frac{R(b_T - b_0)}{\sigma[R(X'X)^{-1}R]^{1/2}}$$

and the usual $t$ ratio for $Rb_0$

$$t_R = \frac{R(b_T - b_0)}{\hat{\sigma}[R(X'X)^{-1}R]^{1/2}}.$$  

Here $\hat{\sigma}$ is the positive square root of $\hat{\sigma}^2 = (y - Xb)'(y - Xb)/(T - K)$. With the standard argument, $N_R$ and $t_R$ can be shown to follow the standard normal distribution and Student’s $t$ distribution with $T - K$ degrees of freedom with appropriate conditioning, respectively, under either Assumptions 2.A.1–2.A.6 or Assumptions 2.A.1, 2.A.2', 2.A.3', and 2.A.5–2.A.6. The following proposition is useful in order to derive the unconditional distributions of these statistics.

**Proposition 2.A.4** If the probability density function of a random variable $Z$ conditional on a random vector $Q$ does not depend on the values of $Q$, then the marginal probability density function of $Z$ is equal to the probability density function of $Z$ conditional on $Q$.  

This proposition is obtained by integrating the probability density function conditional on $Q$ over all possible values of the random variables in $Q$. Since $N_R$ and $t_R$ follow a standard normal distribution and a $t$ distribution conditional on $X$, respectively, Proposition 2.A.4 implies the following proposition:

**Proposition 2.A.5** Suppose that Assumptions 2.A.1, 2.A.5, and 2.A.6 are satisfied and that Assumptions 2.A.2 and 2.A.3 are satisfied for all $x$ in a set $H$ such that $Pr(X^{-1}(H)) = 1$. Then $N_R$ is a standard normal random variable and $t_R$ is a $t$ random variable with $T - K$ degrees of freedom.
Alternatively, the assumptions for Proposition 2.A.3 with Assumption 2.A.6 can be used to obtain a similar result:

**Proposition 2.A.5’** Suppose that Assumptions 2.A.1, 2.A.2′–2.A.3′, 2.A.5, and 2.A.6 are satisfied for s and that Assumptions 2.A.2 and 2.A.3 are satisfied for all x in a set H such that \(Pr(X^{-1}(H)) = 1\). Then \(N_R\) is a standard normal random variable and \(t_R\) is a \(t\) random variable with \(T - K\) degrees of freedom.

Similarly, the usual \(F\) test statistics also follow (unconditional) \(F\) distributions. These results are sometimes not well understood by econometricians. For example, a standard textbook, Judge et al. (1985, p.164), states that “our usual test statistics do not hold in finite samples” on the ground that the (unconditional) distribution of \(b_T's\) is not normal. It is true that \(b_T\) is a nonlinear function of \(X\) and \(e\), so it does not follow a normal distribution even if \(X\) and \(e\) are both normally distributed. However, the usual \(t\) and \(F\) test statistics have the usual (unconditional) distributions as a result of Proposition 2.A.4.

## 2.B Convergence in Probability

Let \(c_1, c_2, \cdots, c_T, \cdots\) be a sequence of real numbers and \(c\) be a real number. The sequence is said to converge to \(c\) if for any \(\varepsilon\), there exists an \(N\) such that \(|c_T - c| < \varepsilon\) for all \(T \geq N\). We write \(c_T \rightarrow c\) or \(\lim_{T \rightarrow \infty} c_T = c\). This definition is extended to a sequence of vectors of real numbers \(\{c_1, c_2, \cdots, c_T, \cdots\}\) by interpreting \(|c_T - c|\) as the Euclidean distance \((c_T - c)'(c_T - c)\).

Consider a univariate stochastic process \([X_T : T \geq 1]\), and a random variable \(X\). Fix \(s\), and then \([X_T(s) : T \geq 1]\) is a sequence of real numbers and \(X(s)\) is a real
number. For each \( s \), verify whether or not \( X_T(s) \to X(s) \). Then collect \( s \) such that \( X_T(s) \to X(s) \), and calculate the probability that \( X_T(s) \to X(s) \). If the probability is one, we say the sequence of random variables, \([X_T : T \geq 1]\), converges to \( X \) almost surely or with probability one. We write \( X_T \to X \) almost surely. This definition is extended to a sequence of random vectors by using convergence for a sequence of vectors for each \( s \). In general, if a property holds for all \( s \) except for a set of \( s \) with probability zero, we say that the property holds almost surely or with probability one.

If \( \Omega \) has finite elements, almost sure convergence is the same thing as convergence of \( X_T(s) \) to \( X(s) \) in all states of the world. In general, however, almost sure convergence does not imply convergence in all states.

The sequence of random variables \([X_T : T \geq 1]\) converges in probability to the random variable \( X_T \) if, for all \( \varepsilon > 0 \), \( \lim_{T \to \infty} \text{Prob}(|X_T - X| > \varepsilon) = 0 \). This is expressed by writing \( X_T \overset{p}{\to} c \) or \( \lim_{T \to \infty} X_T = X \). This extension to the vector case is done by using the Euclidean distance. Almost sure convergence implies convergence in probability.

*Slutsky’s Theorem* is important for working with probability limits. It states that, if \( \text{plim} X_T = X \) and if \( f(\cdot) \) is a continuous function, then \( \text{plim}(f(X_T)) = f(\text{plim}(X_T)) \).

### 2.B.1 Convergence in Distribution

Consider a univariate stochastic process \([X_T : T \geq 1]\), and a random variable \( X \) with respective distribution functions \( F_T \) and \( F \). If \( F_T(x) \to F(x) \) for every continuity point \( x \) of \( F \), then \( X_T \) is said to converge in distribution to \( X \); this is expressed by writing \( X_T \overset{D}{\to} X \). The distribution \( F \) is called the asymptotic distribution or the
2.B. CONVERGENCE IN PROBABILITY

limiting distribution of $X_T$.

2.B.2 Propositions 2.2 and 2.3 for Infinite Numbers of R.V.’s (Incomplete)

In Propositions 2.2 and 2.3, we only allow for a finite number of random variables. In many applications, we are often interested in infinite sums of covariance or strictly stationary random variables. We need the convergence concepts explained in Appendix 2.B. A sequence of real numbers $\{a_j\}_{j=0}^{\infty}$ is square summable if $\sum_{j=0}^{\infty} a_j^2$ is finite. A sufficient condition for $\{a_j\}_{j=0}^{\infty}$ is that it is absolutely summable, that is, $\sum_{j=0}^{\infty} |a_j|$ is finite. In the following propositions, the infinite sum $\sum_{j=0}^{\infty} a_j X_{t-j}$ means the convergence in mean square of $\sum_{j=0}^{T} a_j X_{t-j}$ as $T$ goes to infinity.

**Proposition 2.B.1** If $X_t$ is a scalar covariance stationary process, and if $\{a_j\}_{j=0}^{\infty}$ is square summable, then $X = \sum_{j=0}^{\infty} a_j X_{t-j}$ is covariance stationary. □

The vector version of this proposition is:

**Proposition 2.B.2** If $X_t$ is a $k$-dimensional vector covariance stationary process, and if the absolute value of the $i$-th row of a sequence of a $k \times k$ matrix of real numbers $\{A_j\}_{j=0}^{\infty}$ is square summable for $i = 1, \cdots, k$, then $X_t = \sum_{j=0}^{\infty} A_j X_{t-j}$ is covariance stationary. □

**Exercises**

2.1 In Example 2.3, assume that $\pi_1 = 0.15$, $\pi_2 = 0.05$, $\pi_3 = 0.20$, $\pi_4 = 0.30$, $\pi_5 = 0.10$, and $\pi_6 = 0.20$. As in Example 2.4, compute $E(Y_2|I)(s)$ and $E(Y_2|J)(s)$. Then compute $E(E(Y_2|I)|J)(s)$. Verify that $E(Y_2|J)(s) = E(E(Y_2|I)|J)(s)$ for all $s \in S$. 
2.2 In example 2.9, assume that $|A| < 1$. This condition does not ensure that $Y_t$ is strictly stationary. In order to see this, suppose that $Y_0 = 0$. Then compute the expected values of $Y_1$ and $Y_2$ and the variance of $Y_1$ and $Y_2$, and show that $Y_t$ is not strictly stationary if $A \neq 0$.

2.3 In example 2.9, assume that $|A| < 1$ and that $Y_0$ is $N(0, \frac{\sigma^2}{1-A^2})$. Then compute the expected values of $Y_1$ and $Y_2$, the variance of $Y_1$ and $Y_2$, and the $k$-th autocovariance of $Y$. Prove that $Y_t$ is strictly stationary in this case. (Hint: Remember that first and second moments completely determine the joint distribution of jointly normally distributed random variables.)

2.4 Let $Y_t$ be a martingale adapted to $I_t$. Then prove that $e_t = Y_t - Y_{t-1}$ is a martingale difference sequence.

2.5 Prove that a covariance stationary martingale difference sequence is a white noise process.

2.6 Prove that an i.i.d. white noise process is a martingale difference sequence.

References


Chapter 3

FORECASTING

3.1 Projections

In macroeconomics, forecasting is important in many ways. For structural macroeconomic models, we usually need to specify the forecasting rules that economic agents are using and the information set used by them to forecast future economic variables. Taking the conditional expectation is one way to model forecasting. This method generally requires nonlinear forecasting rules which are difficult to estimate. For the purpose of testing the models and parameter estimation, it is sometimes possible for an econometrician to use a simpler forecasting rule and a smaller information set.

In this section, we study projections as a forecasting method. Projections are used to explain the Wold representation, which forms a basis for studying linear and nonlinear stochastic processes.

3.1.1 Definitions and Properties of Projections

In this chapter, we consider random variables with finite second moments unless otherwise noted. We consider the problem of forecasting $y$, using a set $H$ of random variables. Typically, $y$ is a future random variable such as the growth rate of the Gross Domestic Product (GDP) or the growth rate of a stock price, and $H$ contains
current and past economic variables that are observed by economic agents and/or econometricians. Let us denote a forecast of $y$ based on $H$ by $y^f$, so that the forecasting error is $y - y^f$. In most economic applications, we choose the forecast, $y^f$, so that $y^f$ minimizes

$$E[(y - y^f)^2].$$

(3.1)

In other words, $y^f$ is in $H$, and for all $h$ in $H$,

$$E[(y - y^f)^2] \leq E[(y - h)^2].$$

(3.2)

The expression (3.1) is called the mean squared error associated with the forecast, $y^f$.

When two random variables $h_1$ and $h_2$ satisfy

$$E(h_1 h_2) = 0,$$

(3.3)

they are said to be orthogonal to each other. When either $h_1$ or $h_2$ has mean zero, orthogonality means that they are uncorrelated. The concept of orthogonality is closely related to the problem of minimizing the mean squared error. Under certain conditions on $H$, the Classical Projection Theorem (see, e.g., Luenberger, 1969) states that there exists a unique random variable $y^f$ in $H$ that minimizes the mean squared error, and that $y^f$ is the minimizer if and only if the forecasting error is orthogonal to all members of $H$:

$$E((y - y^f) h) = 0$$

(3.4)

for all $h$ in $H$; this is called the orthogonality condition. When such a forecast exists, we call the forecast, $y^f$, a projection of $y$ onto $H$, and denote it by $\hat{E}(y|H)$. When $Y$ is a random vector with finite second moments, we apply the projection to each element of $Y$ and write $\hat{E}(Y|H)$. 
3.1. PROJECTIONS

Some properties of projections are very important:

Proposition 3.1 (Properties of Projections)

(a) Projections are linear: \( \hat{E}(aX + bY | H) = a\hat{E}(X | H) + b\hat{E}(Y | H) \) for any random variables, \( X \) and \( Y \), with finite variance and constants, \( a \) and \( b \).

(b) If a random variable \( Z \) is in the information set \( H \), then

\[
\hat{E}(ZY | H) = Z\hat{E}(Y | H).
\]

(c) The Law of Iterated Projections: If the information set \( H \) is smaller than the information set \( G \) (\( H \subset G \)), then

\[
\hat{E}(Y | H) = \hat{E}[\hat{E}(Y | G) | H]
\]

3.1.2 Linear Projections and Conditional Expectations

The meaning of projection depends on how the information set \( H \) used for the projection is constructed. Let \( X \) be a \( p \times 1 \) vector of random variables with finite second moments. Let \( H = \{ h \text{ is a random variable such that } h = X'b \text{ for some } p\text{-dimensional vector of real numbers } b \} \). Since \( \hat{E}(y | H) \) is also a member of \( H \), there exists \( b_0 \) such that

\[
\hat{E}(y | H) = X'b_0.
\]

(3.5) In this sense, \( \hat{E}(y | H) \) uses a linear forecasting rule. When we use an information set such as \( H \), which only allows for linear forecasting rules, the projection based on such an information set is called a linear projection. We write \( \hat{E}(y | H) = \hat{E}(y | X) \).
Let $H^N = \{ h \text{ is a random variable with a finite variance such that } h = f(X) \}$ for a function $f$.\footnote{As in Proposition 2.2, we require that the function $f$ is measurable.} In this case, there exists a function $f_0(\cdot)$ such that
\begin{equation}
\hat{E}(y|H^N) = f_0(X).
\end{equation}
In this sense, $\hat{E}(y|H^N)$ allows for a nonlinear forecasting rule. It can be shown that
\begin{equation}
\hat{E}(y|H^N) = E(y|X).
\end{equation}
Hence the projection and conditional expectation coincide when we allow for nonlinear forecasting rules. For this reason, the projections we use in this book are linear projections unless otherwise noted.

An important special case is when $y$ and $X$ are jointly normally distributed. In this case, the expectation of $y$ conditional on $X$ is a linear function of $X$. Hence the linear projection of $y$ onto the information set generated by $X$ is equal to the expectation of $y$ conditional on $X$.

When it is necessary to distinguish the information set $I$ generated by $X$ for conditional expectations introduced in Chapter 2 and the information set $H$ generated by $X$ for linear projections, $H$ will be called the linear information set generated by $X$. (????? Unclear! from Billy)

Linear projections are important because it is easy to estimate them in many applications. Note that the orthogonality condition states that
\begin{equation}
E[(y - X'b_0)h] = 0
\end{equation}
for any $h$ in $H$. Since each element of $X$ is in $H$, using the $i$-th element $X_i$ for $h$, we obtain
\begin{equation}
E[(y - X'b_0)X_i] = 0
\end{equation}
for \( i = 1, 2, \cdots, p \), or

\[
E[X(y - X'b_0)] = 0.
\]

Therefore

\[
E(Xy) = E(XX')b_0.
\]

Assuming that \( E(XX') \) is nonsingular, we obtain

\[
b_0 = E(XX')^{-1}E(Xy)
\]

and

\[
\hat{E}(y_t|H_t) = X_t'b_0,
\]

where \( H_t \) is the linear information set generated by \( X_t \). As we will discuss, if \( X_t \) and \( y_t \) are strictly stationary, Ordinary Least Squares (OLS) can be used to estimate \( b_0 \).

### 3.2 Some Applications of Conditional Expectations and Projections

This section presents some applications of conditional expectations and projections in order to illustrate their use in macroeconomics. More explanations of some of these applications and presentations of other applications will be given in later chapters.

In this chapter, all random variables are assumed to have finite second moments.

#### 3.2.1 Volatility Tests

Many rational expectations models imply

\[
X_t = E(Y_t|I_t)
\]
for economic variables $X_t$ and $Y_t$. Here $X_t$ is in the information set $I_t$ which is available to the economic agents at date $t$ while $Y_t$ is not. A testable implication of (3.14) can be obtained by comparing the volatility of $X_t$ with that of $Y_t$. Relation (3.14) implies

\begin{equation}
Y_t = X_t + \epsilon_t
\end{equation}

where $\epsilon_t = Y_t - E(Y_t|I_t)$ is the forecast error. Since $E(\epsilon_t|I_t) = 0$,

\begin{equation}
E(\epsilon_t h_t) = 0
\end{equation}

for any random variable $h_t$ that is in $I_t$. We can interpret (3.16) as an orthogonality condition. The forecast error must be uncorrelated with any variable in the information set. Since $X_t$ is in $I_t$, (3.16) implies $E(\epsilon_t X_t) = 0$. Therefore, from (3.15) we obtain

\begin{equation}
E(Y_t^2) = E(X_t^2) + E(\epsilon_t^2).
\end{equation}

Since (3.14) implies that $E(X_t) = E(Y_t)$, (3.17) implies

\begin{equation}
Var(Y_t) = Var(X_t) + E(\epsilon_t^2).
\end{equation}

Since $E(\epsilon_t^2) \geq 0$, we conclude

\begin{equation}
Var(Y_t) \geq Var(X_t).
\end{equation}

Thus, if $X_t$ forecasts $Y_t$, $X_t$ must be less volatile than $Y_t$. Various volatility tests have been developed to test this implication of (3.14).

LeRoy and Porter (1981) and Shiller (1981) started to apply volatility tests to the present value model of stock prices. Let $p_t$ be the real stock price (after the
dividend is paid) in period \( t \) and \( d_t \) be the real dividend paid to the owner of the stock at the beginning of period \( t \). Then the no-arbitrage condition is

\[
(3.20) \quad p_t = E[b(p_{t+1} + d_{t+1})|I_t],
\]

where \( b \) is the constant real discount rate, and \( I_t \) is the information set available to economic agents in period \( t \). Solving (3.20) forward and imposing the no bubble condition, we obtain the present value formula:

\[
(3.21) \quad p_t = E(\sum_{i=1}^{\infty} b^i d_{t+i} | I_t).
\]

Applying the volatility test, we conclude that the variance of \( \sum_{i=1}^{\infty} b^i d_{t+i} \) is greater than or equal to the variance of \( p_t \). One way to test this is to directly estimate these variances and compare them. However, \( \sum_{i=1}^{\infty} b^i d_{t+i} \) involves infinitely many data points for the dividend. When we have data for the stock price and dividend for \( t = 1, \cdots, T \), we use (3.21) to obtain

\[
(3.22) \quad p_t = E(\sum_{i=1}^{T-t} b^i d_{t+i} + b^{T-t} p_T | I_t).
\]

Let \( Y_t = \sum_{i=1}^{T-t} b^i d_{t+i} + b^{T-t} p_T \). Then we have data on \( Y_t \) from \( t = 1 \) to \( t = T \) when we choose a reasonable number for the discount rate \( b \). We can estimate the variance of \( p_t \) and the variance of \( Y_t \), and compare them to form a test statistic. \( ^2 \)

### 3.2.2 Parameterizing Expectations

As discussed in Section 3.1, conditional expectations allow for nonlinear forecasting rules. For example, consider \( E(Y|I) \) for a random variable \( Y \) and an information set \( I \) generated from a random variable \( X \). Then \( E(Y|I) \) can be written as a function of

\( ^2 \)There are some problems with this procedure. One problem is nonstationarity of \( p_t \) and \( Y_t \). For more detailed explanation of volatility tests, see Campbell, Lo, and MacKinlay (1997).
\( X : E(Y|I) = f(X) \). The function \( f(\cdot) \) can be nonlinear here. In most applications involving nonlinear forecasting rules, the functional form of \( f(\cdot) \) is not known. In order to simulate rational expectations models, it is often necessary to have a method to estimate \( f(\cdot) \).

Marcet’s (1989) parameterizing expectations method (also see den Haan and Marcet, 1990) is based on the fact that the conditional expectation is a projection, and thus minimizes the mean square error. We take a class of functions that approximate any function. For example, take a class of polynomial functions and let \( f_N(X) = a_0 + a_1X + a_2X^2 + \cdots + a_NX^N \). We choose \( a_0, \cdots, a_N \) to minimize the mean square error, \( E[(Y - f_N(X))^2] \). Intuitively, \( f_N(\cdot) \) should approximate \( f(X) \) for a large enough \( N \). This method is used to simulate economic models with rational expectations.

### 3.2.3 Noise Ratio

In econometrics, we often test an economic model with test statistics whose probability distributions are known under the null hypothesis that the model is true. Hansen’s \( J \) test, which will be discussed in Chapter 9, is an example. Given that all economic models are meant to be approximations, however, it seems desirable to measure how good a model is in approximating reality. Durlauf and Hall (1990) and Durlauf and Maccini (1995) propose such a measure called the noise ratio.\(^3\)

Consider an economic model which states

\[
E(g(Y)|I) = 0
\]

for an information set \( I \) and a function \( g(\cdot) \) of a random vector \( Y \). For example, let \( S \) be the spot exchange rate of a currency in the next period, \( F \) be the forward

---

\(^3\)See Komuki (1999) for an application of the noise ratio to foreign exchange rate models.
exchange rate observed today for the currency to be delivered in the next period, \( g(S, F) = S - F \), and \( I \) be the information set available to the economic agents today. Then under the assumption of risk neutral investors, we obtain (3.23).

Let \( \nu = g(Y) - E(g(Y)|I) \). If the model is true, then \( g(Y) = \nu \). Since this model is an approximation, however, \( g(Y) \) deviates from \( \nu \). Let \( N = g(Y) - \nu \), which is called the model noise. A natural measure of how well the model approximates reality is \( \text{Var}(N) \). Durlauf and Hall (1990) propose a method to estimate a lower bound of \( \text{Var}(N) \) using \( \eta = \text{Var}(\hat{E}(g(Y)|H)) \), where \( H \) is an information set generated from some variables in \( I \).

Using the law of iterated projections\(^5\), \( \hat{E}(\nu|H) = 0 \). Thus, \( \hat{E}(g(Y)|H) = \hat{E}(N|H) \), and therefore \( \eta = \text{Var}(\hat{E}(N|H)) \). Because \( N = \hat{E}(N|H) + (N - \hat{E}(N|H)) \), and the forecast error, \( N - \hat{E}(N|H) \), is orthogonal to \( \hat{E}(N|H) \), \( E(N^2) = E[(\hat{E}(N|H))^2] + E[(N - \hat{E}(N|H))^2] \). Since \( E[(N - \hat{E}(N|H))^2] \geq 0 \), \( E(N^2) \geq E[(\hat{E}(N|H))^2] \). Therefore, \( \text{Var}(N) = E(N^2) - (E(N))^2 \geq E[(\hat{E}(N|H))^2] - \{E[\hat{E}(N|H)]\}^2 = \eta \). Thus \( \eta \) is a lower bound of \( \text{Var}(N) \).

In a sense, \( \eta \) is a sharp lower bound. Since we do not know much about the model noise, \( N \), it may or may not be in \( H \). If \( N \) happens to be in \( H \), then \( \hat{E}(N|H) = N \). Therefore, in this case \( \text{Var}(N) = \eta \).

The noise ratio, \( NR \), is defined by \( NR = \eta / \text{Var}(g(Y)) \). Since \( \hat{E}(g(Y)|H) \) is orthogonal to \( g(Y) - \hat{E}(g(Y)|H) \),

\[
(3.24) \quad \text{Var}(g(Y)) = \eta + \text{Var}(g(Y) - \hat{E}(g(Y)|H)).
\]

\(^4\)For example, in the forward exchange rate model mentioned above, some lagged values of \( S - F \) and a constant can be used to generate a linear information set \( H \).

\(^5\)We assume that the second moment exists and is finite. Therefore, the conditional expectation is a projection.

\(^6\)Here, we assumed that the constants are included in \( H \), so that \( E(S) = E[\hat{E}(S|H)] \).
Therefore, the $0 \leq NR \leq 1$.

Appendix

3.A Introduction to Hilbert Space

This Appendix explains Hilbert space techniques used in this book.\footnote{All proofs of the results can be found in Luenberger (1969) or Hansen and Sargent (1991).} Projections explained in this chapter are defined in a Hilbert space. In Appendix B, we will consider another Hilbert space, which provides the foundation for the lag operator methods and the frequency domain analysis which are useful in macroeconomics and time series economics.

A pre-Hilbert space is a vector space on which an inner product is defined. The inner product is used to define a distance. If all Cauchy sequences of a pre-Hilbert space converge, then it is said to be complete. A Hilbert space is a complete pre-Hilbert space. One reason why a Hilbert space is useful is that the notion of orthogonality can be defined with the inner product. Since a Hilbert space is complete, we can prove that the limit of a sequence exists once we prove that the sequence is Cauchy. For example, this technique can be used to prove that a projection can be defined.

Section 3.A.1 reviews definitions regarding vector spaces. Section 3.A.2 gives an introduction to Hilbert space.
3.A.1 Vector Spaces

Given a set of scalars $K$ (either the real line, $R$, or the complex plane, $C$), a vector space (or a linear space) $X$ on $K$ is a set of elements, called vectors, together with two operations (addition and scalar multiplication) which satisfy the following conditions:

For any $x, y, z$ in $X$ and for any $\alpha, \beta$ in $K$, we require

1. $x + y = y + x$ (commutative law) (3.A.1)
2. $(x + y) + z = x + (y + z)$ (associative law) (3.A.2)
3. There is a null vector $0$ in $X$ such that $x + 0 = x$ for all $x$ in $X$. (3.A.3)
4. $\alpha(x + y) = \alpha x + \alpha y$ (distributive laws) (3.A.4)
5. $(\alpha + \beta)x = \alpha x + \beta x$ (associative law) (3.A.5)
6. $0x = 0, 1x = x$. (3.A.6)

Using $\alpha = -1$, we define $x - y = x + (-1)y$. In this Appendix, we give examples of vector spaces on $R$, but state results that are applicable when $K = C$. Examples of vector spaces on $C$ are given in Appendix B.

A nonempty subset $H$ of a vector space $X$ is called a (linear) subspace of $X$ if every vector of the form $\alpha x + \beta y$ is in $H$ whenever $x$ and $y$ are both in $H$ and $\alpha$ and $\beta$ are in $K$. A subspace always contains the null vector $0$, and satisfies conditions (3.A.1)-(3.A.6). Hence a subspace is itself a vector space.

If a subset $H$ of $X$ is not a subspace, it is often convenient to construct the smallest subspace containing $H$. For this purpose, we use linear combinations of vectors in $H$. A linear combination of the vectors $x_1, x_2, \cdots, x_n$ is a sum of the form

---

*In general, an additive group $X$ for which scalar multiplication satisfies (3.A.4)-(3.A.6) for any field $K$ is a vector space on $K$. In this book $K$ is either the real line or the complex plane.*
\( \alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2 + \cdots + \alpha_n \mathbf{x}_n \) where \( \alpha_i \) is a scalar \((i = 1, \cdots, n)\). The set consisting of all vectors in \( X \) which are linear combinations of vectors in \( H \) is called the \( (linear) \) subspace generated by \( H \).

A normed vector space is a vector space \( X \) on which a norm is defined. The norm is a real-valued function that maps each element of \( x \) in \( X \) into a real number \( \|x\| \), which satisfies

\[
\begin{align*}
(3.A.7) & \quad \|x\| \geq 0 \text{ for all } x \text{ in } X \text{ and } \|x\| = 0 \text{ if and only if } x = 0. \\
(3.A.8) & \quad \|x + y\| \leq \|x\| + \|y\| \quad \text{(The triangle inequality)} \\
(3.A.9) & \quad \|\alpha x\| = |\alpha| \|x\| \text{ for all } \alpha \text{ in } K \text{ and } x \text{ in } X.
\end{align*}
\]

A norm can be used to define a metric \( d \) on \( X \) by \( d(x, y) = \|x - y\| \).

A sequence \( \{x_n\}_{n=1}^\infty \) in a normed vector space converges to \( x_0 \) if the sequence \( \{\|x_n - x_0\|\}_{n=1}^\infty \) of real numbers converges to zero, which is denoted by \( x_n \to x_0 \) or \( \lim x_n = x_0 \). A sequence \( \{x_n\}_{n=1}^\infty \) in a normed vector space is a Cauchy sequence if for any \( \epsilon > 0 \), there exists an integer \( N \) such that \( \|x_n - x_m\| < \epsilon \) for all \( n, m > N \). In a normed vector space, every convergent sequence is a Cauchy sequence. A space in which every Cauchy sequence has a limit is said to be complete. A complete normed vector space is called a Banach space.

**Example 3.A.1** The real line, \( \mathbb{R} \), is a vector space on \( K = \mathbb{R} \) with addition and scalar multiplication defined in the usual way. When the norm of a real number is defined as its absolute value, \( \mathbb{R} \) is a Banach space.

**Example 3.A.2** Vectors in the space consist of sequences of \( n \) real numbers, \( \mathbb{R}^n \), which is a vector space on \( \mathbb{R} \) when \( \mathbf{x} + \mathbf{y} \) for \( \mathbf{x} = (x_1, x_2, \cdots, x_n)' \) and \( \mathbf{y} = (y_1, y_2, \cdots, y_n)' \)
is defined by \((x_1+y_1, x_2+y_2, \cdots, x_n+y_n)\) and \(\alpha x\) for \(\alpha\) in \(\mathbb{R}\) is defined by \((\alpha x_1, \alpha x_2, \cdots, \alpha x_n)\).

When we define a norm of \(x\) as \(\|x\| = \sqrt{\sum_{i=1}^{n} x_i^2}\), \(\mathbb{R}^n\) is a Banach space.

### 3.A.2 Hilbert Space

A **pre-Hilbert space** is a vector space \(X\) on \(K\) for which an inner product is defined. The inner product is a scalar-valued function that maps each element of \((x, y)\) in \(X \times X\) into an element \((x|y)\) in \(K\), which satisfies

\[
\begin{align*}
(3.A.10) & \quad (x|y) = (y|x) \\
(3.A.11) & \quad (x + z|y) = (x|y) + (z|y) \\
(3.A.12) & \quad (\alpha x|y) = \alpha(x|y) \\
(3.A.13) & \quad (x|x) \geq 0 \quad \text{and} \quad (x|x) = 0 \quad \text{if and only if} \quad x = 0.
\end{align*}
\]

for any \(x, y, z\) in \(X\) and \(\alpha\) in \(K\). The bar on the right side on (3.A.10) denotes complex conjugation, which can be ignored if \(K\) is \(\mathbb{R}\). By (3.A.10), \((x|x)\) is real for each \(x\) even when \(K\) is \(\mathbb{C}\).

A norm can be defined from an inner product by \(\|x\| = \sqrt{(x|x)}\). Thus a pre-Hilbert space is a normed vector space. A complete pre-Hilbert space is called a **Hilbert space**.

**Example 3.A.3** When we define \((x|y) = \sum_{i=1}^{n} x_i y_i\), \(\mathbb{R}^n\) is a Hilbert space on \(\mathbb{R}\).

The following Hilbert space of random variables with finite second moments is the one we used in Chapter 3.

**Example 3.A.4** Let \((S, \mathcal{F}, \text{Prob})\) be a probability space. Let \(L^2(\text{Prob}) = \{h : h\) is a (real-valued) random variable and \(E(|h|^2) < \infty\}\). Then with an inner product
defined by \((h_1|h_2) = E(h_1h_2)\), \(L^2(Prob)\) is a Hilbert space on \(R\). If two different random variables \(h_1\) and \(h_2\) satisfy \(E[(h_1 - h_2)^2] = 0\), then \(h_1\) and \(h_2\) are the same element in this space. In other words, if \(E[(h_1 - h_2)^2] = 0\), then \(h_1 = h_2\) with probability one. Hence this definition does not cause problems for most purposes. In this space, the distance is defined by the mean square, so the convergence in this space is the convergence in mean square.

One reason why an inner product is useful is that we can define the notion of orthogonality. In a Hilbert space, two vectors \(x\) and \(y\) are said to be orthogonal if \((x|y) = 0\). A vector \(x\) is said to be orthogonal to a set \(H\) if \(x\) is orthogonal to each element \(h\) in \(H\). Some useful results concerning the inner product are:

**Proposition 3.A.1 (The Cauchy-Schwarz Inequality)** For all \(x, y\) in a Hilbert space, \(|(x|y)| \leq ||x||\ ||y||\). Equality holds if and only if \(x = \lambda y\) for some \(\lambda\) in \(K\), or \(y = 0\).

**Proposition 3.A.2 (Continuity of the Inner Product)** Suppose that \(x_n \rightarrow x\) and \(y_n \rightarrow y\) in a Hilbert space. Then \((x_n|y_n) \rightarrow (x|y)\).

**Proposition 3.A.3** If \(x\) is orthogonal to \(y\) in a Hilbert space, then \(\|x + y\|^2 = \|x\|^2 + \|y\|^2\).

**Example 3.A.5** In \(L^2(Prob)\), the Cauchy-Schwarz Inequality becomes \(|E(xy)| \leq \sqrt{E(x^2)}\sqrt{E(y^2)}\) for any random variables with finite second moments. Proposition 3.A.3 states that if \(x\) and \(y\) satisfy \(E(xy) = 0\), then \(E[(x + y)^2] = E(x^2) + E(y^2)\).

Projections can be defined on a Hilbert space due to the following result:

---

These three propositions hold for a pre-Hilbert space. See Luenberger (1969, p.47 and p.49).
Proposition 3.A.4 (The Classical Projection Theorem) Let $X$ be a Hilbert space and $H$ be a closed linear subspace of $X$. Corresponding to any vector $x$ in $X$, there is a unique vector $h_0$ in $H$ such that $\|x - h_0\| \leq \|x - h\|$. Furthermore, a necessary and sufficient condition that $h_0$ in $H$ be the unique minimizing vector is that $x - h_0$ be orthogonal to $H$.

Given a closed linear space $H$, we define a function $\hat{E}(\cdot|H)$ on $X$ by $\hat{E}(x|H) = h_0$ where $h_0$ is an element in $H$ such that $x - h_0$ is orthogonal to $H$. $\hat{E}(x|H)$ is the projection of $x$ onto $H$. The projection defined in Section 3.1 in $L^2(Prob)$ is one example.

If a sequence $\{e_t\}_{t=1}^\infty$ in a Hilbert space satisfies $\|e_t\| = 1$ for all $t$ and $(e_t|e_s) = 0$ for all $t \neq s$, then it is said to be an orthonormal sequence. We are concerned with an infinite series of the form $\sum_{t=1}^\infty \alpha_t e_t$. An infinite series of the form $\sum_{t=1}^\infty x_t$ is said to converge to the element $x$ in a Hilbert space if the sequence of partial sums $s_T = \sum_{t=1}^T x_t$ converges to $x$. In that case we write $x = \sum_{t=1}^\infty x_t$. A necessary and sufficient condition for an infinite series of orthonormal sequence to converge in Hilbert space is known (see Luenberger, 1969, p.59):

Proposition 3.A.5 Let $\{e_j\}_{j=1}^\infty$ be an orthonormal sequence in a Hilbert space $X$. A series of the form $\sum_{j=1}^\infty \alpha_j e_j$ converges to an element $x$ in $X$ if and only if $\sum_{j=1}^\infty |\alpha_j|^2 < \infty$, and in that case we have $\alpha_j = (x|e_j)$.

Example 3.A.6 Applying the above proposition in $L^2(Prob)$, we obtain a necessary and sufficient condition for an MA($\infty$) representation $\sum_{j=0}^\infty b_j v_{t-j}$ to converge for a white noise process $\{v_{t-j}\}_{j=0}^\infty$ with $E(v_t^2) = \sigma_v^2 > 0$. Define $e_t = \frac{v_t}{\sigma_v}$, and $\alpha_j = b_j \sigma_v$, so that $\{e_{t-j}\}_{j=0}^\infty$ is orthonormal because $E(e_t^2) = 1$ and $E(e_t e_s) = 0$ for $t \neq s$. From
the above proposition, $\sum_{j=1}^{\infty} b_j v_j = \sum_{j=1}^{\infty} \alpha_j e_j$ converges in $L^2(\text{Prob})$, if and only if $\sum_{j=1}^{\infty} |\alpha_j|^2 < \infty$. Since $\sum_{j=1}^{\infty} |\alpha_j|^2 < \infty$ if and only if $\sum_{j=1}^{\infty} |b_j|^2 < \infty$, $\sum_{j=1}^{\infty} b_j v_j$ converges in mean square if and only if $\{b_j\}_{j=1}^{\infty}$ is square summable.

Given an orthonormal sequence $\{e_j\}_{j=1}^{\infty}$, we started from a square summable sequence $\{\alpha_j\}$ and constructed $x = \sum_{j=1}^{\infty} \alpha_j e_j$ in $X$ in the above proposition. We now start with a given $x$ in $X$ and consider a series

$$
(3.14) \quad \sum_{j=1}^{\infty} (x|e_j)e_j.
$$

The series is called the *Fourier series* of $x$ relative to $\{e_j\}_{j=1}^{\infty}$, and $(x|e_j)$ is called the *Fourier coefficient* of $x$ with respect to $e_j$.

In general, $x$ is not equal to its Fourier series. Given a subset $H$ of a Hilbert space, the *closed subspace generated by* $H$ is the closure of the linear subspace generated by $H$. Let $M$ be the closed subspace generated by $\{e_j\}_{j=1}^{\infty}$. If $x$ is in $M$, then $x$ is equal to its Fourier series as implied by the next proposition:

**Proposition 3.A.6** Let $x$ be an element in a Hilbert space $X$ and $\{e_j\}_{j=1}^{\infty}$ be an orthonormal sequence in $H$. Then the Fourier series $\sum_{j=1}^{\infty} (x|e_j)e_j$ converges to an element $\hat{x}$ in the closed subspace $M$ generated by $\{e_j\}_{j=1}^{\infty}$. The difference vector $x - \hat{x}$ is orthogonal to $M$.

This proposition shows that the Fourier series of $x$ is the projection of $x$ onto $M$: $\hat{E}(x|M) = \sum_{j=1}^{\infty} (x|e_j)e_j$.\(^{10}\)

**Exercises**

3.1 Let $S_t$ be a spot exchange rate at time $t$ and $F_t$ be a forward exchange rate

\(^{10}\)See Luenberger (1969, p.60).
observed at time $t$ for delivery of one unit of a currency at $t+1$. Assume that $F_t = E(S_{t+1}|I_t)$ where $I_t$ is the information set available for the economic agents at $t$. Prove that $Var(F_t) \leq Var(S_{t+1})$.

### 3.2
Let $i_{n,t}$ be the $n$ year interest rate observed at time $t$. The expectations hypothesis of the term structure of interest rates states that $i_{n,t} = E(A_t|I_t)$ where

$$A_t = \frac{1}{n} \sum_{\tau=0}^{n-1} i_{1,t+\tau},$$

where $I_t$ is the information available at time $t$. Imagine that data on interest rates clearly indicate that $Var(i_{n,t}) \leq Var(A_t)$. Does the data support the expectations theory? Explain your answer.

### 3.3
Let $p_t$ be the real stock price, $d_t$ be the real dividend, and $b$ be the constant ex ante discount rate. Assume that $p_t$ and $d_t$ are stationary with zero mean and finite second moments. Let

$$p_{t}^e = \sum_{\tau=1}^{\infty} b^{\tau} E(d_{t+\tau}|I_t),$$

where $I_t$ is the information set available in period $t$ that includes the present and past values of $p_t$ and $d_t$. Let $E(\cdot|H_t)$ be the linear projection onto an information set $H_t$. Define the model noise $N_t$ by

$$N_t = p_t - p_{t}^e.$$

Let $\eta = Var(E(N_t|H_t))$.

(a) Assume that $H_t$ is generated by $\{d_t\}$. Show that $\eta \leq Var(N_t)$ for any noise $N_t$. 
(b) Assume that $H_t$ is generated by $\{d_t, d_{t-1}, d_{t-2}\}$. Show that $\eta \leq \text{Var}(N_t)$ for any noise $N_t$.

3.4 Derive (3.24) in the text.

References


Chapter 4

ARMA AND VECTOR AUTOREGRESSION REPRESENTATIONS

4.1 Autocorrelation

The Wold representation of a univariate process \( \{X_t : -\infty < t < \infty\} \) provides us with a description of how future values of \( X_t \) depend on its current and past values (in the sense of linear projections). A useful description of this dependence is autocorrelation. The \( j \)-th autocorrelation of a process (denoted by \( \rho_j \)) is defined as the correlation between \( X_t \) and \( X_{t-j} \):

\[
\text{Corr}(X_t, X_{t-j}) = \frac{\text{Cov}(X_t, X_{t-j})}{\sqrt{\text{Var}(X_t)}\sqrt{\text{Var}(X_{t-j})}}.
\]

In general, \( \rho_j \) depends on \( t \). If the process is covariance stationary, \( \rho_j \) does not depend on \( t \), and is equal to its \( j \)-th autocovariance divided by its variance:

\[
\rho_j = \frac{\gamma_j}{\gamma_0}, \tag{4.1}
\]

where \( \gamma_j = \text{Cov}(X_t, X_{t-j}) \) is the \( j \)-th autocovariance, and \( \gamma_0 = \text{Var}(X_t) \). For covariance stationary processes, \( \gamma_j = \gamma_{-j} \), hence \( \rho_j = \rho_{-j} \). When we view \( \rho_j \) as a function of \( j \), it is called the autocorrelation function. Note that \( \rho_0 = 1 \) for any process by
definition. For a white noise process, \( \rho_j = 0 \) for \( j \neq 0 \). The autocorrelation function is a population concept, and can be estimated by its sample counterpart as explained in Chapter 5.

4.2 The Lag Operator

In order to study ARMA representations, it is convenient to use the lag operator, denoted by the symbol \( L \). When the operator is applied to a sequence \( \{X_t : -\infty < t < \infty\} \) of real numbers, it results in a new sequence \( \{Y_t : -\infty < t < \infty\} \), where the value of \( Y \) at date \( t \) is equal to the value \( X \) at date \( t - 1 \):

\[
Y_t = X_{t-1},
\]

and we write

\[
(4.2) \quad LX_t = X_{t-1}.
\]

When we apply the lag operator to a univariate stochastic process \( \{X_t : -\infty < t < \infty\} \), the lag operator is applied to all sequences of real numbers \( \{X_t(\omega) : -\infty < t < \infty\} \) given by fixing the state of the world \( \omega \) to generate a new stochastic process \( \{X_t : -\infty < t < \infty\} \) that satisfies \( X_{t-1}(\omega) = LX_t(\omega) \) for each \( \omega \).

When the lag operator is applied twice to a process \( \{X_t : -\infty < t < \infty\} \), we write \( L^2X_t = X_{t-2} \). In general, for any integer \( k > 0 \), \( L^kX_t = X_{t-k} \). It is convenient to define \( L^0 = 1 \) as the identity operator that gives \( L^0X_t = X_t \), and to define \( L^{-k} \) as the operator that moves the sequence forward: \( L^{-k}X_t = X_{t+k} \) for any integer \( k > 0 \).

We define a \( p \)-th order polynomial in the lag operator \( B(L) = B_0 + B_1L + B_2L^2 + \cdots + B_pL^p \), where \( B_1, \cdots, B_p \) are real numbers, as the operator that yields

\[
B(L)X_t = (B_0 + B_1L + B_2L^2 + \cdots + B_pL^p)X_t = B_0X_t + B_1X_{t-1} + \cdots + B_pX_{t-p}.
\]
When an infinite sum $B_0X_t + B_1X_{t-1} + B_2X_{t-2} + \cdots$ converges in some sense (such as convergence in $L^2$), we write $B(L) = B_0 + B_1L + B_2L^2 + \cdots$, and

$$B(L)X_t = (B_0 + B_1L + B_2L^2 + \cdots)X_t = B_0X_t + B_1X_{t-1} + B_2X_{t-2} + \cdots.$$ 

For a vector stochastic process $\{X_t : -\infty < t < \infty\}$, a polynomial in the lag operator $B_0 + B_1L + B_2L^2 + \cdots + B_pL^p$ for matrices $B_0, \cdots, B_p$ with real numbers is used in the same way, so that

$$(B_0 + B_1L + B_2L^2 + \cdots + B_pL^p)X_t = B_0X_t + B_1X_{t-1} + \cdots + B_pX_{t-p}.$$ 

Using the lag operator, $X_t = \Phi_0e_t + \Phi_1e_{t-1} + \cdots$ can be expressed as

$$X_t = \Phi(L)e_t,$$

where $\Phi(L) = \Phi_0 + \Phi_1L + \Phi_2L^2 + \cdots$.

### 4.3 Moving Average Representation

If $X_t$ is linearly regular and covariance stationary with mean $\mu$, then it has a Moving Average (MA) representation of the form $X_t = \mu + \Phi(L)e_t$ or

$$X_t = \mu + \Phi_0e_t + \Phi_1e_{t-1} + \Phi_2e_{t-2} + \cdots$$

where $\Phi_0 = 1$. If $\Phi(L)$ is a polynomial of infinite order, $X_t$ is a moving average process of infinite order (denoted $\text{MA}(\infty)$). If $\Phi(L)$ is a polynomial of order $q$, $X_t$ is a moving average process of order $q$ (denoted $\text{MA}(q)$). In this section, we study how some properties of $X_t$ depend on $\Phi(L)$. 

Masao needs to check this!
An MA(1) process $X_t$ has a representation $X_t = \mu + e_t + \Phi e_{t-1}$ as in Example 2.8, where $e_t$ is a white noise process that satisfies (2.10), and $\mu$ and $\Phi$ are constants. The mean, variance, and autocovariance of this process are given in Example 2.8, $E(X_t) = \mu$, and its $k$-th autocorrelation is $\rho_k = \frac{\Phi}{1 + \Phi^2}$ if $|k| = 1$, and $\rho_k = 0$ if $|k| > 1$.

An MA($q$) process $X_t$ satisfies

\begin{equation}
X_t = \mu + e_t + \Phi_1 e_{t-1} + \cdots + \Phi_q e_{t-q},
\end{equation}

where $e_t$ is a white noise process that satisfies (2.10), and $\mu$ and $\Phi_1, \cdots, \Phi_q$ are real numbers. A moving average process is covariance stationary for any $(\Phi_1, \cdots, \Phi_q)$.

Using (2.10), we obtain the mean of an MA($q$) process:

\begin{equation}
E(X_t) = \mu,
\end{equation}

its variance:

\begin{equation}
\gamma_0 = E[(X_t - \mu)^2] = \sigma^2(1 + \Phi_1^2 + \cdots + \Phi_q^2),
\end{equation}

and its $j$-th autocovariance:

\begin{equation}
\gamma_j = E[(X_t - \mu)(X_{t-j} - \mu)] = \begin{cases} 
\sigma^2(\Phi_j + \Phi_{j+1} + \cdots + \Phi_q \Phi_{q-j}) & \text{for } |j| \leq q \\
0 & \text{for } |j| > q
\end{cases}.
\end{equation}

Hence the $j$-th autocorrelation of an MA($q$) process is zero when $|j| > q$.

When a vector stochastic process \{\cdots, X_{-2}, X_{-1}, X_0, X_1, \cdots, X_t, \cdots\} can be written as

\begin{equation}
X_t = \mu + \Phi_0 e_t + \Phi_1 e_{t-1} + \cdots + \Phi_q e_{t-q},
\end{equation}

\footnote{We often impose conditions on $(\Phi_1, \cdots, \Phi_q)$ as we will discuss later in this chapter.}
4.4. **THE WOLD REPRESENTATION**

for a white noise process \( e_t \), then \( X_t \) has a \( q \)-th order (one-sided) moving average (MA(\( q \))) representation. For any \( \Phi_0, \cdots, \Phi_q \), a process with MA(\( q \)) representation is covariance stationary. As \( q \) goes to infinity, an MA(\( \infty \)) representation

\[
X_t = \mu + \Phi_0 e_t + \Phi_1 e_{t-1} + \cdots
\]

is well defined and covariance stationary if \( \sum_{j=0}^{\infty} |\Phi^j_i|^2 < \infty \) for the \( i \)-th row of \( \Phi_j \), \( \Phi^j_i \). In this case, \( X_t \) has a moving average representation of infinite order.

### 4.4 The Wold Representation

Let \( \{\cdots, X_{-2}, X_{-1}, X_0, X_1, \cdots, X_t, \cdots\} \) be a covariance stationary \( n \)-dimensional vector process with mean zero. Let \( H_t \) be the linear information set generated by the current and past values of \( X_t \).\(^2\) We use the notation, \( \hat{E}(y|X_t, X_{t-1}, X_{t-2}, \cdots) \) for \( \hat{E}(y|H_t) \). Note that the information set grows larger over time and the sequence \( \{H_t: -\infty < t < \infty\} \) is increasing in the sense that \( H_t \subset H_{t+1} \) for all \( t \). Let \( H_{-\infty} \) be the set of random variables that are in \( H_t \) for all \( t \): \( H_{-\infty} = \bigcap_{n=1}^{\infty} H_{t-n} \). Then \( 0'X_t \) is a member of \( H_t \). Therefore, the constant zero is always a member of \( H_{-\infty} \). The stochastic process \( X_t \) is linearly regular if \( H_{-\infty} \) contains only the constant zero when \( H_{-\infty} = \bigcap_{n=1}^{\infty} H_{t-n} \), in which \( H_t \) is generated by the current and past values of \( X_t \).

The stochastic process \( X_t \) is linearly deterministic if \( H_t = H_{-\infty} \) for all \( t \). For example, if \( X_t \) is an \( n \)-dimensional vector of constants, then \( X_t \) is linearly deterministic.

We can now state the Wold decomposition theorem, which states that any covariance stationary process can be decomposed into linearly regular and linearly deterministic components:

\(^2\)We only define the linear information set for a finite number of random variables. See Appendix 3.A for further explanation.
**Proposition 4.1 (The Wold Decomposition Theorem)** Let \( \cdots, X_{-1}, X_0, X_1, \cdots, X_t, \cdots \) be a covariance stationary vector process with mean zero. Then it can be written as

\[
X_t = \sum_{j=0}^{\infty} \Phi_j e_{t-j} + g_t,
\]

where \( \Phi_0 = I_n, \sum_{j=0}^{\infty} |\Phi_j|^2 < \infty \) for the \( i \)-th row of \( \Phi_j, \Phi_j^i \), and

\[
e_t = X_t - \hat{E}(X_t|X_{t-1}, X_{t-2}, X_{t-3}, \cdots)
\]

and

\[
g_t = \hat{E}(X_t|H_{-\infty}).
\]

It can be shown that \( \sum_{j=0}^{\infty} \Phi_j e_{t-j} \) is a linearly regular covariance stationary process and \( g_t \) is linearly deterministic. Hence if \( X_t \) is not linearly regular, it is possible to remove \( g_t \) and work with a linearly regular process as long as \( g_t \) can be estimated.

**Proposition 4.2 (The Wold Representation)** Let \( \cdots, X_{-1}, X_0, X_1, \cdots, X_t, \cdots \) be a linearly regular covariance stationary vector process with mean zero. Then it can be written as

\[
X_t = \sum_{j=0}^{\infty} \Phi_j e_{t-j},
\]

where \( \Phi_0 = I_n, \sum_{j=0}^{\infty} |\Phi_j|^2 < \infty \) for the \( i \)-th row of \( \Phi_j, \Phi_j^i \), and \( e_t \) is defined by (4.12).

The Wold representation gives a unique MA representation when the MA innovation \( e_t \) is restricted to the form given by Equation (4.12). There may exist infinitely
many other MA representations when the MA innovation is not restricted to be given by (4.12) as we will discuss below.

In many macroeconomic models, stochastic processes that we observe (real GDP, interest rates, stock prices, etc.) are considered to be generated from the nonlinear function of underlying shocks. In this sense, the processes in these models are nonlinear, but Proposition 4.1 states that even a nonlinear stochastic process has a linear moving average representation as long as it is linearly regular and covariance stationary.

**Example 4.1** Suppose that \( u_t \) is a Gaussian white noise. Let \( X_t = u_t^2 - 1 \). Then the Wold representation of \( X_t \) is \( X_t = e_t \), where \( e_t = u_t^2 - 1 \).

In this example, \( X_t \) is a nonlinear transformation of a Gaussian white noise. The shock that generates \( X_t \), \( u_t \), is normally distributed. However, the innovation in its Wold representation, \( e_t \), is not normally distributed. Thus, the innovation in the Wold representation of a process can have a different distribution from the underlying shock that generates the process.

Even when the underlying shocks that generate processes are i.i.d., the innovations in the Wold representation may not be i.i.d. as in the next example.

**Example 4.2** Suppose that \( u_t \) is an i.i.d Gaussian white noise, so that \( E(u_t^3) = 0 \). Let \( X_t \) be generated by \( X_t = u_t + \Phi(u_{t-1}^2 - 1) \). Then \( E(X_t X_{t-1}) = E[u_t u_{t-1} + \Phi u_{t-1}^3 - \Phi u_{t-1} + \Phi u_t u_{t-2}^2 - \Phi u_t + \Phi^2(u_{t-1}^2 - 1)(u_{t-2}^2 - 1)] = 0 \). Hence the Wold representation of \( X_t \) is \( X_t = e_t \), where \( e_t = u_t + \Phi(u_{t-1}^2 - 1) \).

Note that the Wold representation innovation \( e_t \) in this example is serially uncorrelated, but not i.i.d. because \( e_t(= u_t + \Phi u_{t-1}^2) \) and \( e_{t-1}(= u_{t-1} + \Phi u_{t-2}^2) \) are related.
nonlinearly through the $\Phi u_{t-1}^2$ and $u_{t-1}$ terms.

The Wold representation states that any linearly regular covariance stationary process has an MA representation. Therefore, it is useful to estimate an MA representation in order to study how linear projections of future variables depend on their current and past values. Higher order MA representations and vector MA representations are hard to estimate, however, and it is often convenient to consider AR representations and ARMA representations, which are closely related to MA representations.

### 4.5 Autoregression Representation

A process $X_t$, which satisfies $B(L)X_t = \delta + e_t$ with $B_0 = 1$ or

$$X_t + B_1X_{t-1} + B_2X_{t-2} + \cdots = \delta + e_t$$

for a white noise process $e_t$, is an autoregression. If $B(L)$ is a polynomial of infinite order, $X_t$ is an autoregression of infinite order (denoted AR($\infty$)). If $B(L)$ is a polynomial of order $p$, $X_t$ is an autoregression of order $p$ (denoted AR($p$)). In this section, we study how some properties of $X_t$ depend on $B(L)$.

#### 4.5.1 Autoregression of Order One

Consider a process $X_t$ that satisfies

$$(4.15) \quad X_t = \delta + BX_{t-1} + e_t \quad \text{for } t \geq 1,$$

where $e_t$ is a white noise process with variance $\sigma^2$ and $X_0$ is a random variable that gives an initial condition for (4.15). Such a process is called an autoregression of order
1, denoted by AR(1). It is often convenient to consider (4.15) in a deviation-from-the-mean form:

\[(4.16) \quad X_t - \mu = B(X_{t-1} - \mu) + e_t \quad \text{for} \quad t \geq 1,\]

where \(\mu = \frac{\delta}{1 - B}\). Substituting (4.16) recursively, we obtain

\[X_1 - \mu = B(X_0 - \mu) + e_1\]
\[X_2 - \mu = B(X_1 - \mu) + e_2 = B^2(X_0 - \mu) + Be_1 + e_2, \text{ so that} \]

\[(4.17) \quad X_t - \mu = B^t(X_0 - \mu) + B^{t-1}e_1 + B^{t-2}e_2 + \cdots + Be_{t-1} + e_t \quad \text{for} \quad t \geq 1.\]

In this way, \(X_t\) is defined for any real number \(B\).

Suppose that \(X_0\) is uncorrelated with \(e_1, e_2, \cdots\). When the absolute value of \(B\) is greater than or equal to one, then the variance of \(X_t\) increases over time. Hence \(X_t\) cannot be covariance stationary. In macroeconomics, the case in which \(B = 1\) is of importance, and will be discussed in detail in Chapter 11.

Consider the case where the absolute value of \(B\) is less than one. In this case, \(B^tX_0(\omega)\) becomes negligible as \(t\) goes to infinity for a fixed \(\omega\). As seen in Example 2.9, however, the process \(X_t\) is not covariance stationary in general. Whether or not \(X_t\) is stationary depends upon the initial condition \(X_0\). In order to choose \(X_0\), consider an MA process

\[(4.18) \quad X_t = \mu + e_t + Be_{t-1} + B^2e_{t-2} + \cdots,\]

and choose the initial condition for the process \(X_t\) in (4.15) by

\[(4.19) \quad X_0 = \mu + e_0 + Be_{-1} + B^2e_{-2} + \cdots.\]

When this particular initial condition is chosen, \(X_t\) is covariance stationary.

With the lag operator, (4.16) can be written as

\[(4.20) \quad (1 - BL)(X_t - \mu) = e_t.\]
We define the inverse of \((1 - BL)\) as

\[
(1 - BL)^{-1} = 1 + BL + B^2L^2 + B^3L^3 + \cdots,
\]

when the absolute value of \(B\) is less than one. When a process \(X_t\) has an MA representation of the form (4.18), we write

\[
X_t = \mu + (1 - BL)^{-1}e_t,
\]

which is the MA(\(\infty\)) representation of an AR(1) process.

### 4.5.2 The \(p\)-th Order Autoregression

A \(p\)-th order autoregression satisfies

\[
X_t = \delta + B_1X_{t-1} + B_2X_{t-2} + \cdots + B_pX_{t-p} + e_t \quad \text{for} \quad t \geq 1.
\]

The stability condition is that all the roots of

\[
1 - B_1z - B_2z^2 - \cdots - B_pz^p = 0
\]

are larger than one in absolute value, or equivalently, all the roots of

\[
z^p - B_1z^{p-1} - B_2z^{p-2} - \cdots - B_p = 0
\]

are smaller than one in absolute value.

Consider, for instance, the special case of a AR(1) process with \(B_1 = 1\) and \(X_0 = 0\):

\[
X_t = X_{t-1} + e_t
\]

\[
= e_1 + e_2 + \cdots + e_{t-1} + e_t \quad \text{for} \quad t \geq 1,
\]
where $E(X_t) = 0$ and $E(X_{t-i}X_{t-j}) = \sigma^2$ for $i = j$. Note that $Var(X_1) = \sigma^2$, 
$Var(X_2) = 2\sigma^2$, \ldots, $Var(X_t) = t\sigma^2$. Since the variance of $X_t$ varies over time, $X_t$ is nonstationary. Note also that its first difference is stationary since $e_t(= X_t - X_{t-1})$ is stationary. Such a process is called difference stationary. In this case,

When a (possibly infinite order) polynomial in the lag operator $\Phi(L) = \Phi_0 + \Phi_1L + \Phi_2L^2 + \cdots$ is given, we consider a complex valued function $\Phi(z^{-1}) = \Phi_0 + \Phi_1z^{-1} + \Phi_2z^{-2} + \cdots$ by replacing the lag operator $L$ by a complex number $z$. Consider a condition

(4.28) \hspace{1cm} \Phi(z) = \Phi_0 + \Phi_1z + \Phi_2z^2 + \cdots = 0.

If a complex number $z_i$ satisfies the condition (4.28), then $z_i$ is a zero of $\Phi(z)$. We also say that $z_i$ is a root of the equation $\Phi(z) = 0$.

4.6 \hspace{0.5cm} ARMA

An ARMA($p, q$) process satisfies

(4.29) \hspace{1cm} X_t = \delta + B_1X_{t-1} + B_2X_{t-2} + \cdots + B_pX_{t-p} + e_t + \theta_1e_{t-1} + \theta_2e_{t-2} + \cdots

If $B(1) = 1 - B_1 - \cdots - B_p \neq 1$, we have the deviation-from-the-mean form

(4.30) \hspace{1cm} B(L)(X_t - \mu) = \theta(L)e_t,

where $\mu = \frac{\delta}{B(1)}$. Provided that the $p$-th order polynomial $B(z)$ satisfies stability conditions, the ARMA($p, q$) process yields the MA($\infty$) representation

(4.31) \hspace{1cm} X_t = \mu + \Phi(L)e_t,

where $\Phi(L) = B(L)^{-1}\theta(L) = \Phi_0 + \Phi_1L + \theta_2L^2 + \cdots$ and $\sum_{j=0}^{\infty} |\theta_j|^2 \leq \infty$. 


On the other hand, if \( \theta(z) \) satisfies stability conditions that all roots of \( \theta(z) = 0 \) lie outside the unit circle, then \( \theta(L) \) is invertible and the ARMA\((p,q)\) process yields the AR\((\infty)\) representation\(^3\)

\[
(4.32) \quad \theta(L)^{-1}B(L)X_t = \delta^* + e_t,
\]

where \( \delta^* = \frac{\delta}{\theta(1)} \). Therefore, if both \( B(z) \) and \( \theta(z) \) satisfy stability conditions, then the ARMA\((p,q)\) process has both the MA\((\infty)\) and AR\((\infty)\) representations.

### 4.7 Fundamental Innovations

Let \( X_t \) be a covariance stationary vector process with mean zero that is linearly regular. Then the Wold representation in (4.14) gives an MA representation. There are infinitely many other MA representations.

**Example 4.3** let \( u_t \) be a white noise, and \( X_t = u_t \). Then \( X_t = u_t \) is an MA representation. Let \( u_t^* = u_{t+1} \). Then \( X_t = u_{t-1}^* \) is another MA representation.

In this example, another MA representation is obtained by adopting a different dating procedure for the innovation.

It is often convenient to restrict our attention to the MA representations for which the information content of the current and past values of the innovations is the same as that of the current and past values of \( X_t \). Let

\[
(4.33) \quad X_t = \sum_{j=0}^{\infty} \Phi_j u_{t-j} = \Phi(L)u_t
\]

\(^3\)Without any loss of generality, we assume that there are no common roots of \( B(z) = 0 \) and \( \theta(z) = 0 \). In such a case, we can write the ARMA\((p,q)\) process by the ARMA\((p-m,q-m)\) process that has no common roots, where \( m \) is the number of common roots. See Hayashi (2000, p. 382) for further discussion.
be an MA representation for $X_t$. Let $H_t$ be the linear information set generated by the current and past values of $X_t$, and $H_t^u$ be the linear information set generated by the current and past values of $u_t$. Then $H_t \subset H_t^u$ because of (4.33). The innovation process $u_t$ is said to be \textit{fundamental} if $H_t = H_t^u$. The innovation in the Wold representation is fundamental.

In Example 4.3, $X_t = u_t$ is a fundamental MA representation while $X_t = u_{t-1}^*$ is not. As a result of the dating procedure used for $X_t = u_{t-1}^*$, the information set generated by the current and past values of $u_{t}^*$ : $\{u_{t}^*, u_{t-1}^*, \ldots\}$ is equal to $H_{t+1}$, and is strictly larger than the information set generated by $H_t$.

The concept of fundamental innovations is closely related to the concept of invertibility. If the MA representation (4.33) is invertible, then $u_t = \Phi(L)^{-1}X_t$. Therefore, $H_t^u \subset H_t$. Since (4.33) implies $H_t \subset H_t^u$, $H_t = H_t^u$. Thus if the MA representation (4.33) is invertible, then $u_t$ is fundamental.

If all the roots of $\text{det}[\Phi(z)] = 0$ lie outside the unit circle, then $\Phi(L)$ is invertible, and $u_t$ is fundamental. If all the roots of $\text{det}[\Phi(z)] = 0$ lie on or outside the unit circle, then $\Phi(L)$ may not be invertible, but $u_t$ is fundamental. Thus for fundamentalness, we can allow some roots of $\text{det}[\Phi(z)] = 0$ to lie on the unit circle.

In the univariate case, if $X_t = \Phi(L)u_t$ and all the roots of $\Phi(z) = 0$ lie on or outside the unit circle, then $u_t$ is fundamental. For example, let $X_t = u_t + \Phi u_{t-1}$. If $|\Phi| < 1$, then this MA representation is invertible, and $u_t$ is fundamental. If $\Phi = 1$ or if $\Phi = -1$, then this MA representation is not invertible, but $u_t$ is fundamental. If $|\Phi| > 1$, then $u_t$ is not fundamental.

The MA representations with fundamental innovations are useful; it is easier to express projections of variables onto $H_t$ with them than if they had non-fundamental
innovations. For example, let $X_t$ be a univariate process with an MA(1) representation: $X_t = u_t + \Phi u_{t-1}$. It is natural to assume that economic agents observe $X_t$, but not $u_t$. Therefore, the economic agents’ forecast for $X_{t+1}$ can be modeled as $\hat{E}(X_{t+1}|H_t)$ rather than $\hat{E}(X_{t+1}|H_u)$. If $|\Phi| \leq 1$, $u_t$ is fundamental, and $\hat{E}(X_{t+1}|H_t) = \hat{E}(X_{t+1}|H_u) = \Phi u_t$. On the other hand, if $|\Phi| > 1$, $u_t$ is not fundamental, and $\hat{E}(X_{t+1}|H_t) \neq \hat{E}(X_{t+1}|H_u) = \Phi u_t$, and there is no easy way to express $\hat{E}(X_{t+1}|H_u)$.

4.8 The Spectral Density

Consider a covariance stationary process $Y_t$ such that $Y_t - E(Y_t)$ is linearly regular. Then $Y_t - E(Y_t) = b(L)e_t = \sum_{j=0}^{\infty} b_j e_{t-j}$ for a square summable $\{b_j\}$ and a white noise process $e_t$ such that $E(e_t^2) = 1$ and $E(e_t e_s) = 0$ for $t \neq s$. Its $k$-th autocovariance $\Phi(k) = E[(Y_t - E(Y_t))(Y_{t-k} - E(Y_{t-k}))']$ does not depend on date $t$. For a real number $r$, define

\begin{equation}
\exp(ir) = \cos(r) + i \sin(r),
\end{equation}

where $i = \sqrt{-1}$. The spectral density of $Y_t$, $f(\lambda)$ is defined by

\begin{equation}
f(\lambda) = \sum_{j=0}^{\infty} b_j \exp(-i\lambda j)(\sum_{j=0}^{\infty} b_j \exp(i\lambda j)).
\end{equation}

Then

\begin{equation}
f(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \Phi(k) \exp(i\lambda k)
\end{equation}

for a real number $\lambda (-\pi < \lambda < \pi)$ when the autocovariances are absolutely summable. The spectral density is a function of $\lambda$, which is called the frequency. Using the
properties of the \( \cos \) and \( \sin \) functions and the fact that \( \Phi(k) = \Phi(-k) \), it can be shown that

\[
f(\lambda) = \frac{1}{2\pi} \Phi(0) + 2 \sum_{k=1}^{\infty} \Phi(k) \cos(i\lambda k),
\]

(4.37)

where \( f(\lambda) = f(-\lambda) \) and \( f(\lambda) \) is nonnegative for all \( \lambda \).

Equation (4.36) gives the spectral density from the autocovariances. When the spectral density is given, the autocovariances can be calculated form the following formula:

\[
\int_{-\pi}^{\pi} f(\lambda) \exp(i\lambda k) d\lambda = \Phi(k).
\]

(4.38)

Thus the spectral density and the autocovariances contain the same information about the process. In some applications, it is more convenient to examine the spectral density than the autocovariances. For example, it requires infinite space to plot the autocovariance for \( k = 0, 1, 2, \cdots \), whereas the spectral density can be concisely plotted.

An interpretation of the spectral density is given by the special case of (4.38) in which \( k = 0 \):

\[
\int_{-\pi}^{\pi} f(\lambda) d\lambda = \Phi(0).
\]

(4.39)

This relationship suggests an intuitive interpretation that \( f(\lambda) \) is the contribution of the frequency \( \lambda \) to the variance of \( Y_t \).

This intuition can be formalized by the \textit{spectral representation theorem} which states that any covariance stationary process \( Y_t \) with absolutely summable autocovariances can be expressed in the form

\[
Y_t = \mu + \int_{0}^{\pi} [\alpha(\lambda) \cos(\lambda t) + \delta(\lambda) \sin(\lambda t)] d\lambda,
\]

(4.40)
where $\alpha(\lambda)$ and $\delta(\lambda)$ are random variables with mean zero for any $\lambda$ in $[0, \pi]$. These variables have the further properties that for any frequencies $0 < \lambda_1 < \lambda_2 < \lambda_3 < \lambda_4 < \pi$, the variable $\int_{\lambda_1}^{\lambda_2} \alpha(\lambda)$ is uncorrelated with $\int_{\lambda_1}^{\lambda_3} \alpha(\lambda)$, and the variable $\int_{\lambda_1}^{\lambda_2} \delta(\lambda)$ is uncorrelated with $\int_{\lambda_3}^{\lambda_4} \delta(\lambda)$. For any $0 < \lambda_1 < \lambda_2 < \pi$ and $0 < \lambda_3 < \lambda_4 < \pi$, the variable $\int_{\lambda_1}^{\lambda_2} \alpha(\lambda)$ is uncorrelated with $\int_{\lambda_3}^{\lambda_4} \delta(\lambda)$. For such a process, the portion of the variance due to cycles with frequency less than or equal to $\lambda_1$ is given by

$$2 \int_{0}^{\lambda_1} f(\lambda) d\lambda. \tag{4.41}$$

**Exercises**

4.1 Let $u_t$ be a white noise, and $x_t = u_t + 0.8u_{t-1}$. Is $x_t$ covariance stationary? Is $u_t$ fundamental for $x_t$? Give an expression for $\hat{E}(x_t | u_{t-1}, u_{t-2}, \cdots)$ in terms of past $u_t$'s. Is it possible to give an expression for $\hat{E}(x_t | x_{t-1}, x_{t-2}, \cdots)$ in terms of past $u_t$'s? If so, give an expression. Explain your answers.

4.2 Let $u_t$ be a white noise, and $x_t = u_t + 1.2u_{t-1}$. Is $x_t$ covariance stationary? Is $u_t$ fundamental for $x_t$? Give an expression for $\hat{E}(x_t | u_{t-1}, u_{t-2}, \cdots)$ in terms of past $u_t$'s. Is it possible to give an expression for $\hat{E}(x_t | x_{t-1}, x_{t-2}, \cdots)$ in terms of past $u_t$'s? If so, give an expression. Explain your answers.

4.3 Let $u_t$ be a white noise, and $x_t = u_t + u_{t-1}$. Is $x_t$ covariance stationary? Is $u_t$ fundamental for $x_t$? Give an expression for $\hat{E}(x_t | u_{t-1}, u_{t-2}, \cdots)$ in terms of past $u_t$'s. Is it possible to give an expression for $\hat{E}(x_t | x_{t-1}, x_{t-2}, \cdots)$ in terms of past $u_t$'s? If so, give an expression. Explain your answers.

**References**

Chapter 5

STOCHASTIC REGRESSORS IN LINEAR MODELS

This chapter explains asymptotic theory for linear models in the form that is convenient for most applications of structural econometrics for linear time series models. In many applications of rational expectations models, stringent distributional assumptions, such as an assumption that the disturbances are normally distributed, are unattractive. Without such assumptions, however, it is not possible to obtain the exact distributions of estimators in finite samples. For this reason, asymptotic theory describes the properties of estimators as the sample size goes to infinity.

Many researchers use asymptotic theory at initial stages of an empirical research project. Given the difficulties of obtaining the exact small sample distributions of estimators in many applications, this utilization seems to be a sound strategy. If the sample size is “large”, then asymptotic theory must be a good approximation of the true properties of estimators. The problem is that no one knows how large the sample size should be, because the answer depends on the nature of each application. After the importance of a research project is established, small sample properties of the estimators used in the project are often studied. For this purpose, Monte Carlo
experiments can be used as described later in this book.

5.1 The Conditional Gauss Markov Theorem

In regressions (5.4?????????) and (5.7????????), $X_t$ is strictly exogenous in the time series sense if $E(e_t|X_{t+2}, X_{t+1}, X_t, X_{t-1}, X_{t-2}, \ldots) = 0$. This is a very restrictive assumption that does not hold in all applications of cointegration discussed in Chapter 13. For example, $E(e_t|X_t, X_{t-1}, X_{t-2}, \ldots) = 0$ in some applications because $e_t$ is a forecast error. However, the forecast error is usually correlated with future values of $X_t$. Hence the strict exogeneity assumption is violated. Nevertheless, as Choi and Ogaki (1999) argue, it is useful to observe that the Gauss Markov theorem applies to cointegrating regressions in order to understand small sample properties of various estimators for cointegrating vectors. Moreover, this observation leads to a Generalized Least Squares (GLS) correction to spurious regressions.

Let $\sigma(X)$ be the smallest $\sigma$-field with respect to which the random variables in $X$ are measurable. We use the notation $E[Z|\sigma(X)]$ to denote the usual conditional expectation of $Z$ conditional on $X$ as defined by Billingsley (1986) for a random variable $Z$. $E[Z|\sigma(X)]$ is a random variable, and $E[Z|\sigma(X)](s)$ denotes the value of the random variable at $s$ in $S$ (????? what is s?). It should be noted that the definition is given under the condition that $Z$ is integrable, namely $E(|Z|) < \infty$. This condition can be too restrictive when we define the conditional expectation of the OLS estimator in some applications as we discuss later. "

For this reason, we will also use a different concept of conditional expectation

\[1\text{Loeve (1978) slightly relaxes this restriction by defining the conditional expectation for any random variable whose expectation exists (but may not be finite) with an extension of the Radon-Nikodym theorem. This definition can be used for $E(|\sigma(X))$, but this slight relaxation does not solve our problem which we describe later.} \]
conditional on $X$ that can be used when $Z$ and $\text{vec}(X)$ have probability density functions $f_Z(z)$ and $f_X(\text{vec}(x))$, respectively. In this case, if $f_X(\text{vec}(x))$ is positive, we define the expectation of $Z$ conditional on $X(s) = x$ as

$$E[Z|X(s) = x] = \int_{-\infty}^{\infty} \frac{f_Z(z)}{f_X(\text{vec}(x))} dz.$$  

(5.1)

For this definition, we use the notation $E[Z|X(s) = x]$. This definition can only be used when the probability density functions exist and $f_X(\text{vec}(x))$ is positive, but the advantage of this definition for our purpose is that the conditional expectation can be defined even when $E(Z)$ does not exist. For example let $Z = \frac{Y}{X}$ where $Y$ and $X$ are independent random variables with a standard normal distribution. Then $Z$ has the Cauchy distribution, and $E(Z)$ does not exist. Thus, $E[Z|\sigma(X)]$ cannot be defined.\(^2\)

However, we can define $E[Z|X(s) = x]$ for all $s$ in the probability space because the density function of $X$ is always positive.

In the special case in which both types of conditional expectations can be defined, they coincide. More precisely, suppose that $Z$ and $\text{vec}(X)$ have the probability density functions, that the probability density function of $\text{vec}(X)$ is always positive, and that $Z$ is integrable. Then $E[Z|\sigma(X)](s) = E[Z|X(s)]$ with probability one.

Let $y = (y_1, y_2, \ldots, y_T)'$ be a $T \times 1$ vector of random variables, and $e = (e_1, e_2, \ldots, e_T)'$ be a $T \times 1$ vector of random variables. We are concerned with a linear model of the form:

Assumption 5.1 $y = Xb_0 + e$,

where $b_0$ is a $K \times 1$ vector of real numbers. We assume that the expectation of $e$ conditional on $X$ is zero:

\(^2\)It should be noted that we cannot argue that $E(Z) = E(E(Y|X)|\sigma(X))$ = $E(E(Y|X)|\sigma(X))$ = $E(Y|X)|\sigma(X))$ = 0 even though $Y$ is measurable in $\sigma(X)$ because $E(Y|X)|\sigma(X))$ is not defined.
CHAPTER 5. STOCHASTIC REGRESSORS

Assumption 5.2 \( E[e|\sigma(X)] = 0 \).

Since \( E[e|\sigma(X)] \) is only defined when each element of \( e \) is integrable, Assumption 5.2 implicitly assumes that \( E(e) \) exists and is finite. It also implies \( E(e) = 0 \) because of the law of iterated expectations. Given \( E(e) = 0 \), a sufficient condition for Assumption 5.2 is that \( X \) is statistically independent of \( e \). Since Assumption 5.2 does not imply that \( X \) is statistically independent of \( e \), Assumption 5.2 is weaker than the assumption of the independent stochastic regressors. With the next assumption, we assume that \( e \) is conditionally homoskedastic and \( e_t \) is not serially correlated:

Assumption 5.3 \( E[ee'|\sigma(X)] = \sigma^2 I_T \).

Let \( G = \{ s \in S : X(s)'X(s) \) is nonsingular\}. Since the determinant of a matrix is a continuous function of the elements of a matrix, \( G \) is a member of the \( \sigma \)-field \( \mathcal{F} \).

For any \( s \) in \( G \), the OLS estimator is

\[
(5.2) \quad b_T = (X'X)^{-1}X'y.
\]

From Assumption 5.1, \( b_T = b_0 + (X'X)^{-1}X'e \). Hence the conditional Gauss-Markov theorem can be proved when the expectation of \( (X'X)^{-1}X'e \) and \( (X'X)^{-1}X'ee'X(X'X)^{-1} \) can be defined. For this purpose, we consider the following two alternative assumptions:

Assumption 5.4 \( E[(X'X)^{-1}X'ee'X(X'X)^{-1}] \) exists and is finite.

Assumption 5.4’ \( e \) and \( vec(X) \) have probability density functions, and the probability density functions of \( vec(X) \) are positive for all \( s \) in \( G \).
A sufficient condition for Assumption 5.4 is that the distributions of $X$ and $e$ have finite supports. Under Assumption 5.4, $E[(X'X)^{-1}X'e]$ also exists and is finite. Hence $E(b_T|\sigma(X))$ can be defined. From Assumptions 5.1-5.3, $E(b_T|\sigma(X)) = b_0 + E[(X'X)^{-1}X'e|\sigma(X)] = b_0$ for $s \in G$ with probability $Pr(G)$. Under Assumptions 5.1-5.4, $E[(b_T-b_0)'(b_T-b_0)|\sigma(X)]$ can be defined, and $E[(b_T-b_0)'(b_T-b_0)|\sigma(X)] = E[(X'X)^{-1}X'ee'X(X'X)^{-1}|\sigma(X)] = (X'X)^{-1}X'E[ee'|\sigma(X)]X(X'X)^{-1} = \sigma^2(X'X)^{-1}$ for $s \in G$ with probability $Pr(G)$. The problem with Assumption 5.4 is that it is not easy to verify Assumption 5.4 for many distributions of $X$ and $e_t$ which are often used in applications and Monte Carlo studies.

Under Assumptions 5.1-5.3 and 5.4', $E[b_T|X(s)] = b_0$ and $E[(b_T-b_0)'(b_T-b_0)|X(s)] = \sigma^2(X(s)'X(s))^{-1}$ for any $s \in G$.

Corresponding with Assumption 5.4 and 5.4', we consider two definitions of the conditional version of the Best Linear Unbiased Estimator (BLUE). Given a set $H$ in the \( \mathcal{F} \), the \emph{Best Linear Unbiased Estimator (BLUE) conditional on} \( \sigma(X) \) in $H$ is defined as follows. An estimator $b_T$ for $b_0$ is the BLUE conditional on $\sigma(X)$ in $H$ if (1) $b_T$ is linear conditional on $\sigma(X)$, namely, $b_T$ can be written as $b_T = Ay$ where $A$ is a $K \times T$ matrix, and each element of $A$ is measurable $\sigma(X)$; (2) $b_T$ is unbiased conditional on $\sigma(X)$ in $G$, namely, $E(b_T|\sigma(X)) = b_0$ for $s \in H$ with probability $Pr(H)$; (3) for any linear unbiased estimator $b^*$ conditional on $X(s) = x$ for which $E(b^*b'^*)$ exists and is finite, $E[(b_T-b_0)'(b_T-b_0)|X(s) = x] \leq E[(b^*-b_0)'(b^*-b_0)|X(s) = x] - E[(b_T-b_0)'(b_T-b_0)|X(s) = x]$ is a positive semidefinite matrix with probability one for $s \in H$ with probability $Pr(H)$.

An estimator $b_T$ for $b_0$ is the BLUE conditional on $X(s) = x$ in $H$ if (1) $b_T$
is linear conditional on $X(s)$ in $H$, namely, $b_T$ can be written as $b_T = Ay$ where $A$ is a $K \times T$ matrix, and each element of $A$ is measurable $\sigma(X)$; (2) $b_T$ is unbiased conditional on $X(s) = x$ in $H$, namely, $E(b_T|X(s) = x) = b_0$ for any $s$ in $H$; (3) for any linear unbiased estimator $b^*$ conditional on $X(s) = x$ for which $E((b^* - b_0)(b^* - b_0)'|X(s) = x)$ exists and is finite, $E((b_T - b_0)(b_T - b_0)'|X(s) = x) - E((b_T - b_0)(b_T - b_0)'|X(s) = x)$ is a positive semidefinite matrix for any $s$ in $H$.

With these preparations, the following theorem can be stated:

**Theorem 5.1 (The Conditional Gauss-Markov Theorem)** Under Assumptions 5.1-5.4, the OLS estimator is the BLUE conditional on $\sigma(X)$ in $G$. Under Assumptions 5.1-5.3 and 5.4, the OLS estimator is the BLUE conditional on $X(s) = x$ in $G$.

The theorem can be proved by applying any of the standard proofs of the (unconditional) Gauss-Markov theorem by replacing the unconditional expectation with the appropriate conditional expectation.

Under Assumptions 5.1-5.4, the unconditional expectation and the unconditional covariance matrix of $b_T$ can be defined. With an additional assumption that $Pr(G) = 1$ or $Assumption 5.5 X'X$ is nonsingular with probability one,

we obtain the following corollary of the theorem:

**Proposition 5.1** Under Assumptions 5.1-5.5, the OLS estimator is unconditionally unbiased and has the minimum unconditional covariance matrix among all linear unbiased estimators conditional on $\sigma(X)$. 
Proof Using the law of iterated expectations, 
\[ E(b_T) = E\{E[b_T|\sigma(X)]\} = E(b_0) = b_0, \]
and 
\[ E[(b_T - b_0)(b_T - b_0)^\prime] = E[E[(b_T - b_0)(b_T - b_0)^\prime|\sigma(X)]] = \sigma^2 E[(X^\prime X)^{-1}]. \]
For the minimum covariance matrix part, let \( b^\ast \) be another linear unbiased estimator conditional on \( \sigma(X) \). Then
\[ E[(b^\ast - b_0)(b^\ast - b_0)^\prime]\frac{1}{\sigma(X)} = E[(b_T - b_0)(b_T - b_0)^\prime|\sigma(X)] + \Delta, \]
where \( \Delta \) is a positive semidefinite matrix with probability one. Then 
\[ E[(b^\ast - b_0)(b^\ast - b_0)^\prime] - E[(b_T - b_0)(b_T - b_0)^\prime] = E[E[b^\ast b^\prime|\sigma(X)] - E[E(b_T b_T^\prime|\sigma(X)]] = E(\Delta) \text{ is a positive semidefinite matrix}. \]

A few remarks for this proposition are in order:

**Remark 5.1** Assumption 5.4 cannot be replaced by Assumption 5.4' for this proposition. Under Assumption 5.4', \( E(b_T) \) and \( E[(b_T - b_0)(b_T - b_0)^\prime] \) may not exist.

**Remark 5.2** In this proposition, the covariance matrix of \( b_T \) is \( \sigma^2 E[(X^\prime X)^{-1}] \), which is different from \( \sigma^2 E[X^\prime X]^{-1} \). This result may seem to contradict the standard asymptotic theory, but it does not. Asymptotically, \( \frac{1}{T}X^\prime X \) converges almost surely to \( E[X_t X_t^\prime] \) if \( X_t \) is stationary and ergodic. Hence the limit of the covariance matrix of \( \sqrt{T}(b_T - b_0), \sigma^2 E[(X^\prime X)^{-1}] \), is equal to the asymptotic covariance matrix, \( \sigma^2 [E(X_t X_t^\prime)]^{-1} \).

### 5.2 Unconditional Distributions of Test Statistics

In order to study distributions of the \( t \) ratios and \( F \) test statistics, we need an additional assumption:

**Assumption 5.6** Conditional on \( X \), \( e \) follows a multivariate normal distribution.

Given a \( 1 \times K \) vector of real numbers \( R \), consider a random variable
\[ N_R = \frac{R(b_T - b_0)}{\sigma[R(X^\prime X)^{-1}R]^{\frac{1}{2}}} \]
and the usual $t$ ratio for $Rb_0$

$$t_R = \frac{R(b_T - b_0)}{\hat{\sigma} \sqrt{R(X'X)^{-1}R'}}.$$  

Here $\hat{\sigma}$ is the positive square root of $\hat{\sigma}^2 = \frac{1}{T-K}(y - Xb_T)'(y - Xb_T)$. With the standard argument, $N_R$ and $t_R$ can be shown to follow the standard normal distribution and Student’s $t$ distribution with $T - K$ degrees of freedom conditional on $X$, respectively, under either Assumptions 5.1-5.6 or Assumptions 5.1-5.3, 5.4, and 5.5-5.6. The following proposition is useful in order to derive unconditional distributions of these statistics.

**Proposition 5.2** If the probability density function of a random variable $Z$ conditional on a random vector $Q$ does not depend on the values of $Q$, then the marginal probability density function of $Z$ is equal to the probability density function of $Z$ conditional on $Q$.

This proposition is obtained by integrating the probability density function conditional on $Q$ over all possible values of the random variables in $Q$. Since $N_R$ and $t_R$ follow the standard normal and the Student’s $t$ distribution conditional on $X$, respectively, Proposition 5.2 implies the following proposition:

**Proposition 5.3** Under the Assumptions 5.1-5.6, or under the Assumptions 5.1-5.3, 5.4, and 5.5-5.6, $N_R$ is the standard normal random variable and $t_R$ is the Student’s $t$ random variable with $T - K$ degrees of freedom.

Similarly, the usual $F$ test statistics also follow (unconditional) $F$ distributions. These results are sometimes not well understood by econometricians. For example, a standard textbook, Judge et al. (1985, p.164), states that “our usual test statistics
do not hold in finite samples on the grounds that $b_T$’s (unconditional) distribution is not normal. It is true that $b_T$ is a nonlinear function of $X$ and $e$, so it does not follow a normal distribution even if $X$ and $e$ are both normally distributed. However, the usual $t$ and $F$ test statistics have usual (unconditional) distributions as a result of Proposition 5.2.

5.3 The Law of Large Numbers

If an estimator $b_T$ converges almost surely to a vector of parameters $b_0$, then $b_T$ is **strongly consistent** for $b_0$. If an estimator $b_T$ converges in probability to a vector of parameters $b_0$, then $b_T$ is **weakly consistent** for $b_0$.

Consider a univariate stationary stochastic process $\{X_t\}$. When $X_t$ is stationary, $E(X_t)$ does not depend on date $t$. Therefore, we often write $E(X)$ instead of $E(X_t)$. Assume that $E(|X|)$ is finite, and consider a sequence of random variables $[Y_T : T \geq 1]$, where $Y_T = \frac{1}{T} \sum_{t=1}^{T} X_t$ is the sample mean of $X$ computed from a sample of size $T$. In general, the sample mean does not converge to its unconditional expected value, but converges almost surely to an expectation of $X$ conditional on an information set. For the sample mean to converge almost surely to its unconditional mean, we require the series to be ergodic. A stationary process $\{X_t\}$ is said to be **ergodic** if, for any bounded functions $f : \mathbb{R}^{i+1} \rightarrow \mathbb{R}$ and $g : \mathbb{R}^{j+1} \rightarrow \mathbb{R}$,

\[
\lim_{T \to \infty} |E[f(X_t, \cdots, X_{t+i})g(X_{t+T}, \cdots, X_{t+T+j})]| = |E[f(X_t, \cdots, X_{t+i})]| |E[g(X_t, \cdots, X_{t+j})]|.
\]

(5.6)

Heuristically, a stationary process is ergodic if it is asymptotically independent: that is, if $(X_t, \cdots, X_{t+i})$ and $(X_{t+T}, \cdots, X_{t+T+j})$ are approximately independent for large enough $T$. 

Proposition 5.4 (The strong law of large numbers) If a stochastic process \([X_t : t \geq 1]\) is stationary and ergodic, and if \(E(|X|)\) is finite, then \(\frac{1}{T} \sum_{t=1}^{T} X_t \rightarrow E(X)\) almost surely.

5.4 Convergence in Distribution and Central Limit Theorem

This section explains a definition of convergence in distribution and presents some central limit theorems. These central limit theorems are based on martingale difference sequences, and are useful in many applications of rational expectations models.

Central limit theorems establish that the sample mean scaled by \(T\) converges in distribution to a normal distribution\(^3\) under various regularity conditions. The following central limit theorem by Billingsley (1961) is useful for many applications because we can apply it when economic models imply a variable is a martingale difference sequence.

Proposition 5.5 (Billingsley’s Central Limit Theorem) Suppose that \(e_t\) is a stationary and ergodic martingale difference sequence adapted to \(I_t\), and that \(E(|e|^2) < \infty\). Assume that \(I_{t-1} \subset I_t\) for all \(t\). Then

\[
\frac{1}{\sqrt{T}} \sum_{t=1}^{T} e_t \overset{D}{\rightarrow} N(0, E(e^2)).
\]

If \(e_t\) is an i.i.d. white noise, then it is a stationary and ergodic martingale difference sequence adapted to \(I_t\) which is generated from \(\{e_t, e_{t-1}, \cdots\}\). Hence the Billingsley’s Central Limit Theorem is more general than the central limit theorems

\(^3\)In some central limit theorems, the limiting distribution is not normal.
for i.i.d. processes such as the Lindeberg-Levy theorem, which is usually explained in econometric text books. However, Billingsley’s Central Limit Theorem cannot be applied to any serially correlated series.

A generalization of the theorem to serially correlated series is due to Gordin (1969):

**Proposition 5.6 (Gordin’s Central Limit Theorem)** Suppose that $e_t$ is a univariate stationary and ergodic process with mean zero and $E(|e|^2) < \infty$, that $E(e_t|e_{t-j}, e_{t-j-1}, \cdots)$ converges in mean square to 0 as $j \to \infty$, and that

$$
(5.7) \quad \sum_{j=0}^{\infty} [E(r_{tj}^2)]^{1/2} < \infty,
$$

where

$$
(5.8) \quad r_{tj} = E(e_t|I_{t-j}) - E(e_t|I_{t-j-1}),
$$

where $I_t$ is the information set generated from $\{e_t, e_{t-1}, \cdots\}$. Then $e_t$’s autocovariances are absolutely summable, and

$$
(5.9) \quad \frac{1}{\sqrt{T}} \sum_{t=1}^{T} e_t \xrightarrow{D} N(0, \Omega),
$$

where

$$
(5.10) \quad \Omega = \lim_{T \to \infty} \sum_{j=-T+1}^{T-1} E(e_t e_{t-j}).
$$

When $e_t$ is serially correlated, the sample mean scaled by $T$ still converges to a normal distribution, but the variance of the limiting normal distribution is affected by serial correlation as in (5.10).
In (5.10), \( \Omega \) is called a \textit{long-run variance} of \( e_t \). Intuition behind the long-run variance can be obtained by observing

\[
E[\frac{1}{\sqrt{T}} \sum_{t=1}^{T} e_t^2] = \sum_{j=-T+1}^{T-1} \frac{T - |j|}{T} E(e_t e_{t-j})
\]

and that the right hand side (5.11) is the Cesaro sum of \( \sum_{j=-T+1}^{T-1} E(e_t e_{t-j}) \). Thus when \( \sum_{j=-T+1}^{T-1} E(e_t e_{t-j}) \) converges, its limit is equal to the limit of the right hand side of (5.11) (Apostol, 1974).

Another expression for the long-run variance can be obtained from an MA representation of \( e_t \). Let \( e_t = \Psi(L)u_t = \Psi_0 u_t + \Psi_1 u_{t-1} + \cdots \) be an MA representation. Then \( E(e_t e_{t-j}) = (\Psi_j \Psi_0 + \Psi_{j+1} \Psi_1 + \Psi_{j+2} \Psi_2 + \cdots) E(u_t^2) \), and \( \Omega = \{ (\Psi_0^2 + \Psi_1^2 + \Psi_2^2 + \cdots) + 2(\Psi_1 \Psi_0 + \Psi_2 \Psi_1 + \Psi_3 \Psi_2 + \cdots) + 2(\Psi_2 \Psi_0 + \Psi_3 \Psi_1 + \Psi_4 \Psi_2 + \cdots) + \cdots \} E(u_t^2) = (\Psi_0 + \Psi_1 + \Psi_2 + \cdots)^2 E(u_t^2) \). Hence

\[
\Omega = \Psi(1)^2 E(u_t^2).
\]

In the next example, we consider a multi-period forecasting model. For this model, it is easy to show that Gordin’s Theorem is applicable to the serially correlated forecast error.

\textbf{Example 5.1 (The Multi-Period Forecasting Model)} Suppose that \( I_t \) is an information set generated by \( \{ Y_t, Y_{t-1}, Y_{t-2}, \cdots \} \), where \( Y_t \) is a stationary and ergodic vector stochastic process. In typical applications, economic agents are assumed to use the current and past values of \( Y_t \) to generate their information set. Let \( X_t \) be a stationary and ergodic random variable in the information set \( I_t \) with \( E(|X_t|^2) < \infty \).

We consider an \( s \)-period ahead forecast of \( X_t \), \( E(X_{t+s}|I_t) \), and the forecast error, \( e_t = X_{t+s} - E(X_{t+s}|I_t) \).
It is easy to verify that all the conditions for Gordin’s Theorem are satisfied for $e_t$. Moreover, because $E(e_t|I_t) = 0$ and $e_t$ is in the information set $I_{t+s}$, $E(e_t e_{t-j}) = E(E(e_t e_{t-j}|I_t)) = E(e_{t-j}E(e_t|I_t)) = 0$ for $j \geq s$. Hence $\Omega = \lim_{j \to \infty} \sum_{j=-s}^{s} E(e_t e_{t-j})$.

Hansen (1985) generalized Gordin’s Central Limit Theorem to vector processes. In this book, we call the generalized theorem Gordin and Hansen’s Central Limit Theorem.

**Proposition 5.7** (*Gordin and Hansen’s Central Limit Theorem*) Suppose that $e_t$ is a vector stationary and ergodic process with mean zero and finite second moments, that $E(e_t|e_{t-j}, e_{t-j-1}, \cdots)$ converges in mean square to 0 as $j \to \infty$, and that

\[
\sum_{j=0}^{\infty} [E(r'_t r_{tj})]^2 < \infty,
\]

where

\[
r_{tj} = E(e_t|I_{t-j}) - E(e_t|I_{t-j-1}),
\]

where $I_t$ is the information set generated from $\{e_t, e_{t-1}, \cdots\}$. Then $e_t$’s autocovariances are absolutely summable, and

\[
\frac{1}{\sqrt{T}} \sum_{t=1}^{T} e_t \xrightarrow{D} N(0, \Omega)
\]

where

\[
\Omega = \lim_{T \to \infty} \sum_{j=-T}^{T-1} E(e_t e'_{t-j}).
\]
The matrix $\Omega$ in Equation (5.15) is called the long-run covariance matrix of $e_t$.

As in the univariate case, another expression for the long-run covariance can be obtained from an MA representation of $e_t$. Let $e_t = \Psi(L)u_t = \Psi_0 u_t + \Psi_1 u_{t-1} + \cdots$ be an MA representation. Then $E(e_t e_{t-j}) = (\Psi_j + \Psi_{j+1} + \Psi_{j+2} + \cdots) E(u_t u_t') (\Psi_0 + \Psi_1 + \Psi_2 + \cdots)'$, and $\Omega = (\Psi_0 + \Psi_1 + \Psi_2 + \cdots) E(u_t u_t') (\Psi_0 + \Psi_1 + \Psi_2 + \cdots)'$. Hence

$$\Omega = \Psi(1) E(u_t u_t') \Psi(1)'$$

(5.16)

In the next example, Gordin and Hansen’s Central Limit Theorem is applied to a serially correlated vector process:

**Example 5.2** Continuing Example 5.1, let $Z_t$ be a random vector with finite second moments in the information set $I_t$. Define $f_t = Z_t e_t$. Then $E(f_t | I_t) = E(Z_t e_t | I_t) = E(Z_t E(e_t | I_t)) = 0$. In empirical work, it is often necessary to apply a central limit theorem to a random vector such as $f_t$. It is easy to verify that all conditions for Gordin and Hansen’s Theorem are satisfied for $f_t$. Moreover, $E(f_t | I_t) = 0$ and $f_t$ is in the information set $I_{t+s}$, thus $E(f_t f_{t-j}') = E(E(f_t f_{t-j}' | I_t)) = E(E(f_t | I_t) f_{t-j}') = 0$ for $j \geq s$. Hence $\Omega = \lim_{j \to \infty} \sum_{j-s}^{j-1} E(f_t f_{t-j}') = \sum_{j=-s+1}^{s-1} E(f_t f_{t-j}')$.

5.5 Consistency and Asymptotic Distributions of OLS Estimators

Consider a linear model,

$$y_t = x_t' b_0 + e_t,$$

(5.17)

where $y_t$ and $e_t$ are stationary and ergodic random variables, and $x_t$ is a $p$-dimensional stationary and ergodic random vector. We assume that the orthogonality conditions

$$E(x_t e_t) = 0$$

(5.18)
are satisfied, and that $E(x_t x'_t)$ is nonsingular. Imagine that we observe a sample of $(y_t, x'_t)$ of size $T$. Proposition 5.4 shows that $\frac{1}{T} \sum_{t=1}^{T} x_t x'_t$ converges to $E(x_t x'_t)$ almost surely. Hence with probability one, $\sum_{t=1}^{T} x_t x'_t(s)$ is nonsingular for large enough $T$, and the Ordinary Least Squares (OLS) estimator for (5.17) can be written as

$$\mathbf{b}_T = \left( \sum_{t=1}^{T} x_t x'_t \right)^{-1} \left( \sum_{t=1}^{T} x_t y_t \right). \quad (5.19)$$

In order to apply the Law of Large Numbers to show that the OLS estimator is strongly consistent, rewrite (5.19) from (5.17) after scaling each element of the right side by $T$:

$$\mathbf{b}_T - \mathbf{b}_0 = \left( \frac{1}{T} \sum_{t=1}^{T} x_t x'_t \right)^{-1} \left( \frac{1}{T} \sum_{t=1}^{T} (x_t e_t) \right). \quad (5.20)$$

Applying Proposition 5.4, we obtain

$$\mathbf{b}_T - \mathbf{b}_0 \rightarrow [E(x_t x'_t)]^{-1}(E(x_t e_t)) = 0 \quad \text{almost surely.} \quad (5.21)$$

Hence the OLS estimator, $\mathbf{b}_T$, is a strongly consistent estimator. In order to obtain the asymptotic distribution of the OLS estimator, we make an additional assumption that a central limit theorem applies to $x_t e_t$. In particular, assuming that the Gordin and Hansen’s Martingale Approximation Central Limit Theorem is applicable, we multiply both sides of (5.20) by the square root of $T$:

$$\sqrt{T} (\mathbf{b}_T - \mathbf{b}_0) = \left( \frac{1}{T} \sum_{t=1}^{T} x_t x'_t \right)^{-1} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^{T} (x_t e_t) \right). \quad (5.22)$$

Therefore,

$$\sqrt{T} (\mathbf{b}_T - \mathbf{b}_0) \overset{D}{\rightarrow} N(0, [E(x_t x'_t)]^{-1} \Omega [E(x_t x'_t)]^{-1}) \quad (5.23)$$

$^4$Appendix 3.A explains why these types of conditions are called orthogonality conditions.
where $\Omega$ is the long-run covariance matrix of $x_t e_t$:

\[
\Omega = \sum_{j=-\infty}^{\infty} E(e_t e_{t-j} x_t' x_{t-j}).
\]

### 5.6 Consistency and Asymptotic Distributions of IV Estimators

Consider the linear model (5.17) for which the orthogonality conditions (5.18) are not satisfied. In this case, we try to find a $p$-dimensional stationary and ergodic random vector $z_t$, which satisfies two types of conditions: the orthogonality condition

\[
E(z_t e_t) = 0,
\]

and the relevance condition that $E(z_t x_t')$ is nonsingular. We define the Linear Instrumental Variable (IV) estimator as

\[
b_T = \left( \sum_{t=1}^{T} z_t x_t' \right)^{-1} \sum_{t=1}^{T} z_t y_t.
\]

Then

\[
b_T - b_0 = \left( \frac{1}{T} \sum_{t=1}^{T} z_t x_t' \right)^{-1} \left( \frac{1}{T} \sum_{t=1}^{T} z_t e_t \right).
\]

Applying Proposition 5.4, we obtain

\[
b_T - b_0 \rightarrow [E(z_t x_t')]^{-1}(E(z_t e_t)) = 0 \quad \text{almost surely.}
\]

Hence the linear IV estimator, $b_T$, is a strongly consistent estimator. Assuming that the Vector Martingale Approximation Central Limit Theorem is applicable to $z_t e_t$,

\[
\sqrt{T}(b_T - b_0) \xrightarrow{D} N(0, [E(z_t x_t')]^{-1} \Omega [E(z_t x_t')]^{-1})
\]

where $\Omega$ is the long-run covariance matrix of $z_t e_t$:

\[
\Omega = \sum_{j=-\infty}^{\infty} E(e_t e_{t-j} z_t' z_{t-j}).
\]
5.7 Nonlinear Functions of Estimators

In many applications of linear models, we are interested in nonlinear functions of \( b_0 \), say \( a(b_0) \). This section explains the delta method, which is a convenient method to derive asymptotic properties of \( a(b_T) \) as an estimator for \( b_0 \) where \( b_T \) is a weakly consistent estimator for \( b_0 \). In many applications, \( b_T \) is an OLS estimator or a linear IV estimator. Later in this book we will use the proof of the delta method to prove the asymptotic normality of the GMM estimator. (????? a not bold, f is better?)

**Proposition 5.8** Suppose that \( \{b_T\} \) is a sequence of \( p \)-dimensional random vectors such that \( \sqrt{T}(b_T - b_0) \xrightarrow{D} z \) for a random vector \( z \). If \( a(\cdot) : \mathbb{R}^p \rightarrow \mathbb{R}^r \) is continuously differentiable at \( b \), then

\[
\sqrt{T}[a(b_T) - a(b_0)] \xrightarrow{D} d(b_0)z,
\]

where \( d(b_0) = \left. \frac{\partial a(b)}{\partial b} \right|_{b=b_0} \) denotes the \( r \times p \) matrix of first derivatives evaluated at \( b_0 \). In particular, if \( z \sim \mathcal{N}(0, \Sigma) \), then

\[
\sqrt{T}[a(b_T) - a(b_0)] \xrightarrow{D} \mathcal{N}(0, d(b_0)\Sigma d(b_0)') .
\]

**Proof**

5.8 Remarks on Asymptotic Theory

When we use asymptotic theory, we do not have to make restrictive assumptions that the disturbances are normally distributed. Serial correlation and conditional heteroskedasticity can be easily taken into account as long as we can estimate the long-run covariance matrix (which is the topic of the next chapter).
It is a common mistake to think that the linearity of the formula for the long-run covariance matrix means a linearity assumption for the process of $x_t e_t$ (for the OLS estimator) or $z_t e_t$ (for the IV estimator). It should be noted that we did not assume that $x_t e_t$ or $z_t e_t$ was generated by linear functions (i.e., a moving average process in the terminology of Chapter 4) of independent white noise processes. Even when $x_t e_t$ or $z_t e_t$ is generated from nonlinear functions of independent white noise processes, the distributions based on the long-run covariance matrices give the correct limiting distributions. This point is related to the Wold representation for nonlinear processes discussed in Chapter 4. Even when $z_t e_t$ is generated as a nonlinear process, as long as it is a linearly regular and covariance stationary process, it has the Wold representation: $z_t e_t = \Psi(L)u_t$, and its long-run covariance matrix is given by (5.30).

Appendix

5.A Monte Carlo Studies

This appendix explains Monte Carlo methods. Example programs in GAUSS explained in Appendix A are given. Asymptotic theory is used to obtain approximations of the exact finite sample properties of estimators and test statistics. In many time series applications, the exact finite sample properties cannot be obtained. For example, in a regression with lagged dependent variables, we can assume neither that the regressor is nonrandom nor that the error term is strictly exogenous in the time series sense. In many applications with financial variables, the assumption that the error term in a regression is normal is inappropriate because many authors have found evidence against normality for several financial variables. Asymptotic theory gives accurate approximations when the sample size is “large,” but exactly how “large” is
enough depends on each application. One method to study the quality of asymptotic approximations is the Monte Carlo simulations.

5.A.1 Random Number Generators

In Monte Carlo studies, data are generated with computer programs called pseudo-random number generators. These programs generate sequences of values that appear to be draws from a specified probability distribution. Modern pseudo-random generators are accurate enough that we can ignore the fact that numbers generated are not exactly independent draws from a specified probability distribution for most purposes.\(^5\) Hence in the rest of this appendix, phrases such as “values that appear to be” are often suppressed.

Recall that when a probability space \(\Omega\) is given, the whole history of a stochastic process \(\{e_t(s)\}_{t=1}^N\) is determined when a point in the probability space \(s\) is given. For a random number generator, we use a number called the starting seed to determine \(s\). Then the random number generator automatically updates the seed each time a number is generated. It should be noted that the same sequence of numbers is generated whenever the same starting seed is given to a random number generator.

Generated random numbers are used to generate samples. From actual data, we obtain only one sample, but in Monte Carlo studies, we can obtain many samples from generated random numbers. Each time a sample is generated, we compute estimators or test statistics of interest. After replicating many samples, we can estimate small sample properties of the estimators or test statistics by studying the generated

\(^5\)One exception is that a pseudo-random number generator ultimately cycles back to the initial value generated and repeats the sequence when too many numbers are generated. Most modern pseudo-random number generators cycle back after millions of values are drawn, and this tendency is not a problem for most Monte Carlo studies. However, in some studies in which millions or billions of values are needed, there can be a serious problem.
distributions of these variables and compare them with predictions of asymptotic theory.

Most programs offer random number generators for the uniform distribution and the standard normal distribution. For example,

\[ y = \text{RNDN}(r, c); \]

in GAUSS generates \( r \times c \) values that appear to be a realization of independent standard normal random variables that will be stored in an \( r \times c \) matrix. The starting seed for RNDN can be given by a statement

\[ \text{RNDSEED } n; \]

where the value of the seed \( n \) must be in the range \( 0 < n < 2^{31} - 1 \).

One can produce random numbers with other distributions by transforming generated random numbers. The following examples are some of the transformations which are often used.

**Example 5.A.1** A \( \chi^2 \) random variable with \( d \) degrees of freedom can be created from \( d \) independent random variables with the standard normal distribution. If \( e_i \sim N(0, 1) \), and if \( e_i \) is independent from \( e_j \) for \( j \neq i \), then \( \sum_{i=1}^{d} e_i^2 \) follows the \( \chi^2 \) distribution with \( d \) degrees of freedom.

For example, in GAUSS one can generate a \( T \times 1 \) vector with values that appear to be a realization of an i.i.d. \( \{x\}_{t=1}^{T} \) of random variables with the \( \chi^2 \) distribution with \( d \) degrees of freedom by the following program:

\[ e = \text{RNDN}(T, d); \]
\[ x = \text{sumc}((e^2)^{\prime}); \]
Example 5.A.2 A random variable which follows the Student’s $t$ distribution with $d$ degrees of freedom can be generated from $d + 1$ independent random variables with the standard normal distribution. If $e_i \sim N(0, 1)$, and if $e_i$ is independent from $e_j$ for $j \neq i$, then $x = e_1 / \sqrt{\sum_{i=2}^{d+1} e_i^2 / d}$ follows the $t$ distribution with $d$ degrees of freedom.

For example, in GAUSS one can generate a $T \times 1$ vector with values that appear to be a realization of an i.i.d. $\{x\}_{i=1}^{T}$ of random variables with the $t$ distribution with $d$ degrees of freedom by the following program:

\[
e = \text{RNDN}(T, d+1);
\]
\[
c = \text{sumc}((e[.,2:d+1]^2)');
\]
\[
x = e[.,1]./\text{sqrt}(c/d);
\]

Example 5.A.3 A $K$-dimensional random vector which follows $N(0, \Psi)$ for any positive definite covariance matrix $\Psi$ can be generated from $K$ independent random variables with the standard normal distribution. Let $\Psi = PP'$ be the Cholesky decomposition of $\Psi$, in which $P$ is a lower triangular matrix. If $e_i \sim N(0, 1)$, and if $e_i$ is independent from $e_j$ for $j \neq i$, then $X = Pe \sim N(0, \Psi)$ where $e = (e_1, e_2, \cdots, e_K)'$.

For example, in GAUSS one can generate a $T \times K$ matrix with values that appear to be a realization of an i.i.d. $\{X_i\}_{i=1}^{T}$ of $K$-dimensional random vectors with the $N(0, C)$ distribution with the following program provided that the matrix $C$ is already defined.

\[
e = \text{RNDN}(T, K);
\]
\[
P = \text{chol}(C)';
\]
\[
x = eP;
\]

Note that the Cholesky decomposition in GAUSS gives an upper triangular matrix. Thus, the above program transposes the matrix to be lower triangular.
5.A.2 Estimators

When a researcher applies an estimator to actual data without the normality assumption, asymptotic theory is used as a guide of small sample properties of the estimator. In some cases, asymptotic theory does not give a good approximation of the exact finite sample properties. A Monte Carlo study can be used to estimate the true finite sample properties. For example, the mean, median, and standard deviation of the realized values of the estimator over generated samples can be computed and reported as estimates of the true values of these statistics. For example, \( N \) independent samples are created and an estimate \( b_i \) \((i \geq 1)\) for a parameter \( b_0 \) is calculated for the \( i \)-th sample. Then the expected value of the estimator \( E(b_i) \) can be estimated by its mean over the samples: \( \frac{1}{N} \sum_{i=1}^{N} b_i \). By the strong law of large numbers, the mean converges almost surely to the expected value.

Other properties can also be reported, depending on the purpose of the study. For example, Nelson and Startz (1990) report estimated 1%, 5%, 10%, 50%, 90%, and 99% fractiles for an IV estimator and compared them with fractiles implied by the asymptotic distribution. This influential paper uses Monte Carlo simulations to study the small sample properties of IV estimator and its \( t \)-ratio when instruments are poor in the sense that the relevance condition is barely satisfied.

When the deviation from the normal distribution is of interest, the skewness and kurtosis are often estimated and reported. The skewness of a variable \( Y \) with mean \( \mu \) is

\[
(5.A.1) \quad \frac{E(Y - \mu)^3}{[Var(Y)]^{\frac{3}{2}}}.
\]

A variable with negative skewness is more likely to be far below the mean than it is
to be far above, and conversely a variable with positive skewness is more likely to be
far above the mean than is to be below. If $Y$ has a symmetric distribution such as
a normal distribution, then the skewness is zero. The kurtosis of $Y$ is

\[
E(Y - \mu)^4 \quad \quad [V\text{ar}(Y)]^2.
\]

If $Y$ is normally distributed, the kurtosis is 3. If the kurtosis of $Y$ exceeds 3, then its
distribution has more mass in the tails than the normal distribution with the same
variance.

5.A.3 Tests

When a researcher applies a test to actual data without the normality assumption,
asymptotic theory is typically used. For example, the critical value of 1.96 is used
for a test statistic with the asymptotic normal distribution for the significance level
of 5%. The significance level and critical value based on the asymptotic distribution
are called the nominal significance level and the nominal critical value, respectively.
The probability of rejecting the null hypothesis when it is true is called the size of the
test. Since the asymptotic distribution is not exactly equal to the exact distribution
of the test statistic, the true size of the test based on the nominal critical value is
usually either larger or smaller than the nominal significance level. This property is
called the size distortion. If the true size is larger than the nominal significance level,
the test overrejects the null hypothesis and is said to be liberal. If the true size is
smaller than the nominal significance level, the test underrejects the null hypothesis
and is said to be conservative. Using the distribution of the test statistic produced
by a Monte Carlo simulation, one can estimate the true critical value.

The power of the test is the probability of rejecting the null hypothesis when
the alternative hypothesis is true. In Monte Carlo studies, two versions of the power can be reported for each point of the alternative hypothesis: the power based on the nominal critical value and the power based on the estimated true critical value. The latter is called the size corrected power. The power based on the nominal critical value is also of interest because it is the probability of rejecting the null hypothesis in practice if asymptotic theory is used. On the other hand, the size corrected power is more appropriate for the purpose of comparing tests. For example, a liberal test tends to have a higher power based on the nominal critical value than a conservative test. However, we cannot conclude the liberal test is better from this observation because the probability of Type I error is not equal for the two tests.

5.A.4 A Pitfall in Monte Carlo Simulations

Common mistakes are made by many graduate students when they first use Monte Carlo simulations. These mistakes happen when they repeatedly use a random number generator to conduct simulations. These mistakes are caused by updating seeds arbitrarily in the middle of a simulation. Recall that once the starting seed is given, a random number generator automatically updates the seed whenever it creates a number. The way the seed is updated depends on the program.

The following example illustrates common mistakes in a simple form:

Example 5.A.4 A Monte Carlo Program with a Common Mistake (I)

```plaintext
ss=3937841;
i=1;
vecm=zeros(100,1);
do until i>100;
    RNDSEED ss;
y=RNDN(50,1);
m=meanc(y);
```
In this example, the programmer is trying to create 100 samples of the sample mean of a standard normal random variable $y$ when the sample size is 50. However, exactly the same data are generated 100 times because the same starting seed is given for each replication inside the do-loop. This mistake is innocuous because it is easy to detect. The following program contains a mistake which is harder to detect:

Example 5.A.5 A Monte Carlo Program with a Common Mistake (II)

```plaintext
ss=3937841;
i=1;
vecm=zeros(100,1);
do until i>100;
    RNDSEED ss+i;
    y=RNDN(50,1);
    m=meanc(y);
    vecm[i]=m;
i=i+1;
endo;
```

The problem is that the seed is updated in an arbitrary way in each sample by giving a different starting seed. There is no guarantee that one sample is independent from the others. A correct program would put the RNDSEED statement before the do loop. For example, the RNDSEED statement inside the do loop should be removed and the statement

```plaintext
RNDSEED ss;
```

can be added after the first line.
In Monte Carlo simulations, it is also important to control the starting seed so that the simulation results can be replicated. When you publish Monte Carlo results, it is advisable to put enough information in the publication so that others can exactly replicate the results.\(^6\) At the very least, a record of the information should be kept. If no RNDSEED statement is given before the RNDN command is used, GAUSS will take the starting seed from the computer clock. Then there is no way to exactly replicate these Monte Carlo results.

### 5.A.5 An Example Program

This section describes basic Monte Carlo methods with an example program. In the following example, the sample mean is calculated as an estimator for the expected value of \(X_t, E(X_t)\), where \(X_t = \mu + e_t\) and \(e_t\) is drawn from the \(t\) distribution with 3 degrees of freedom. The \(t\) distribution with 3 degrees of freedom has thick tails and large**, outlying values have high probability. Hence the \(t\) distribution is often considered a better distribution to describe some financial variables. Because \(X_t\) is not normally distributed, the standard theory for the exact finite sample properties cannot be applied. The example is concerned with the \(t\) test of the null hypothesis that \(\mu = 0\). Because a random variable with the \(t\) distribution with 3 degrees of freedom has zero mean and a finite second moment, asymptotic theory predicts that the \(t\) test statistic of the sample mean divided by the estimated standard error approximately follows the standard normal distribution.

**Example 5.A.6 The program.**

```plaintext
@MCMEAN.PRG @ @Monte Carlo Program for the sample mean@
```

\(^6\)This information is also relevant because different computer specifications and different versions of the program (such as GAUSS) can produce different results.
This example program is a GAUSS program to calculate the empirical size and power of the t-test for \( H_0 : E(X) = 0 \), where \( X \) follows \( t \)-distribution with 3 degrees of freedom. The power is calculated for the case when \( E(X) = 0.2 \).

```gauss
RNDSEED 382974;
output file=mc.out reset;
tend=25; // the sample size
nor=1000; // the number of replications
df=3; // d.f. for the \( t \)-distribution of \( X \)
i=1;
 tn=zeros(nor,1); // used to store \( t \)-values under \( H_0 \)
ta=zeros(nor,1); // used to store \( t \)-values under \( H_1 \)
do until i>nor;
 nrv=RNDN(tend,df+1); // normal r.v.'s
 crv=nrv[,2:df+1]^2; // chi square r.v.'s
 x0=nrv[,1]./sqrt(sumc(crv')/df); // \( t \)-distribution: used under \( H_0 \)
x1=x0+0.2; // used for \( H_1 \)
 mx0=meanc(x0);
 mx1=meanc(x1);
 sighat0=sqrt((x0-mx0)'(x0-mx0)/(tend-1)); // \( \sigma_{\text{hat}} \) under \( H_0 \)
 sighat1=sqrt((x1-mx1)'(x1-mx1)/(tend-1)); // \( \sigma_{\text{hat}} \) under \( H_1 \)
 tn[i]=meanc(x0)*sqrt(tend)/sighat0; // \( t \)-value under \( H_0 \)
 ta[i]=meanc(x1)*sqrt(tend)/sighat1; // \( t \)-value under \( H_1 \)
i=i+1;
endo;

? "***** When \( H_0 \) is true *****";
? "The estimated size with the nominal critical value";
? meanc(abs(tn).>1.96);
? "The estimated true 5-percent critical value"
? sorttn=sortc(abs(tn),1);
? etcv=sorttn[int(nor*0.95)];
? etcv;
?
? "***** When \( H_1 \) is true *****";
? "The estimated power with the nominal critical value";
? meanc(abs(ta).>1.96);
? "The estimated size corrected power"
? meanc(abs(ta).>etcv);
output off;
```
Some features of the example are important. Before the do-loop of the replications, the program set up an output file by

```
output file=mc.out;
```

Then to avoid the common mistake explained in 5.A.4, it makes the RNDNSEED statement before the do-loop.

It is a good idea to minimize the content inside the do-loop to speed up replications. Everything that can be done outside the do-loop should be done here. For example, the program defines variables to store the test results:

```
  tn=zeros(nor,1);
  ta=zeros(nor,1);
```

In GAUSS, the do-loop can be set up as follows:

```
i=1;
do until i>250;
  ... (Program for each replication)
i=i+1;
endo;
```

After the do-loop, the program calculates characteristics of the generated distributions of test statistics under the null hypothesis and the alternative hypothesis such as the frequency of rejecting the null with the nominal critical value.

**Exercises**

5.1 Show that all conditions of Gordin’s Central Limit Theorem are satisfied for $e_t$ in Example 5.1.

5.2 Show that all conditions of Gordin and Hansen’s Central Limit Theorem are satisfied for $f_t$ in Example 5.2.
5.3 Let $e_t = \Psi(L)u_t = \Psi_0 u_t + \Psi_1 u_{t-1} + \cdots$ be an MA representation. What is the long-run variance of $f_t = (1 - L)e_t$?

5.4 Explain what it means to say that “a test under-rejects in small samples” (or “a test is conservative”). When a test is conservative, which is greater, the true critical value or the nominal critical value?

5.5 Consider the linear model

$$y_t = x_t' \beta + e_t,$$

where $x_t$ is a $k$-dimensional vector.

Let $z_t$ be a $k \times 1$ vector of instrumental variables. We will adopt the following set of assumptions:

(A1) $(e_t, x_t, z_t)_{t=1}^{\infty}$ is a stationary and ergodic stochastic process.

(A2) $z_t e_t$ have finite second moments.

(A3) $E(e_t^2 | z_t) = \sigma^2$, where $\sigma$ is a constant.

(A4) $E(e_t | I_t) = 0$ for a sequence of information sets $(I_t)_{t=1}^{\infty}$ which is increasing (i.e., $I_t \subset I_{t+1}$), $z_t$ and $x_t$ are in $I_t$, and $y_t$ is in $I_{t+1}$.

(A5) $E(z_t x_t')$ is nonsingular.

Note that $E(e_t) = 0$ is implied by (A4) if $z_t$ includes a constant.

Note that many rational expectations models imply (A4). For the following problems, prove the asymptotic properties of the instrumental variable (IV) estimator, $b_{IV}$, for $\beta$ under (A1)-(A5). Use a central limit theorem and a strong law of large
numbers given in this chapter, and indicate which ones you are using and where you are using them in your proof.

(a) Express the IV estimator \( \mathbf{b}_{IV} \) in terms of \( \mathbf{z}_t, \mathbf{x}_t, \) and \( y_t(t = 1, \ldots, T) \) when \( \sum_{t=1}^{T} \mathbf{z}_t \mathbf{x}_t' \) is nonsingular.

(b) Let \( \mathbf{g}_t = \mathbf{z}_t \epsilon_t \). Prove that \( \mathbf{g}_t \) is a martingale difference sequence.

(c) Prove that the IV estimator is consistent under (A1)-(A5).

(d) Prove that the IV estimator is asymptotically normally distributed. Derive the formula of the covariance matrix of the asymptotic distribution.

(e) Explain what happens if \( y_t \) is in \( I_{t+2} \) in (A4).

5.6 Consider the linear model

\[ y_t = \mathbf{x}_t' \beta + \epsilon_t, \]

where \( \mathbf{x}_t \) is a \( k \)-dimensional vector. Following Hayashi (2000), suppose that this model satisfies the classical linear regression model assumptions for any sample size \( n \) as follows:

(A1) Linearity: \( y_t = \mathbf{x}_t' \beta + \epsilon_t \).

(A2) Ergodic stationarity: \( \{y_t, \mathbf{x}_t\} \) is jointly stationary and ergodic.

(A3) Predetermined regressors: \( E(\epsilon_t \mathbf{x}_t) = \mathbf{0} \).

(A4) Rank condition: \( E(\mathbf{x}_t \mathbf{x}_t') \) is nonsingular (and hence finite).

(A5) \( \mathbf{x}_t \epsilon_t \) is a martingale difference sequence with finite second moments.
(A6) Finite fourth moments for regressors: \( E[(x_{it}x_{jt})^2] \) exists and finite for all \( i, j (= 1, 2, \ldots, k) \).

(A7) Conditional homoskedasticity: \( E(e_t^2|x_t) = \sigma^2 > 0 \).

Further, assume that \( e_t \) is normally distributed conditional on \( X \), where \( X \) is an \( n \times k \) matrix with \( x'_t \) in its \( t \)-th row. Let
\[
t_k = \frac{b_k - \bar{\beta}_k}{SE(b_k)} = \frac{b_k - \bar{\beta}_k}{\sqrt{s^2[(X'X)^{-1}]}_{kk}}
\]
be the \( t \) statistic for the null hypothesis \( \beta_k = \bar{\beta}_k \).

(a) Prove that \( t_k \) converges in distribution to the standard normal distribution as the sample size goes to infinity. You do not have to prove that \( s^2 \) is consistent \( \sigma^2 \) for this question. You can assume that \( s^2 \) is consistent.

(b) Based on the asymptotic result in (a), suppose that you set the nominal size to be 5 percent and reject the null hypothesis when \(|t_k|\) is greater than 1.96. Does this test overreject or underreject. How do you know? Suppose that \( k = 3 \). Is the actual size larger than 10 percent when \( n = 4 \). What if \( n = 8, 9, 10, 11 \)? Explain.

5.7 Consider the linear model

(5.E.1) \[
y = X\beta + e
\]

Let \( k \times 1 \) matrix \( x'_t \) be the \( t \)-th row of the regressor matrix \( X \). The model (5.E.1) can be written as

(5.E.2) \[
y_t = x'_t\beta + e_t
\]
We will adopt the following set of assumptions:

(A1) \((e_t, x_t)_{t=1}^\infty\) are independent and identically distributed (i.i.d.) random vectors.

(A2) \(x_t\) and \(e_t\) have finite second moments.

(A3) \(E(e_t^2|x_t) = \sigma^2\) which is a constant.

(A4) \(E(x_t e_t) = 0\) for all \(t \geq 1\)

(A5) \(E(x_t x'_t)\) is nonsingular.

Note that \(E(e_t) = 0\) is implied by (A) if \(x_t\) includes a constant.

Consider the model (5.E.1) with \(k = 1\). Assume that \(x_t\) follows \(N(0,1)\). Assume that \(x_t\) and \(e_t\) are independent. Under the null hypothesis \(H_0\), the true value of \(\beta\) is 0, so that \(y_t = e_t\).

Consider the standard \(t\) statistic,

\[
(5.E.3) \quad t_1 = \frac{b - \beta}{\hat{\sigma}_1 \sqrt{X'X}}^{-1}
\]

where \(\hat{\sigma}_1^2 = (Y - Xb)'(Y - Xb)/(n - k)\). Consider another version of the \(t\) statistic

\[
(5.E.4) \quad t_2 = \frac{b - \beta}{\hat{\sigma}_2 \sqrt{X'X}}^{-1}
\]

where \(\hat{\sigma}_2^2 = (Y - Xb)'(Y - Xb)/n\). Note that both \(t_1\) and \(t_2\) converge in distribution to a random variable with the standard normal distribution.

Consider two alternative assumptions for \(e_t\).

(A6) \(e_t\) follows the \(t\) distribution with 4 degrees of freedom.

(A6') \(e_t\) follows the standard normal distribution.

Note that Assumptions 1.1 - 1.5 are satisfied with (A6'), so that \(t_1\) has the exact \(t\) distribution with \(n - k\) degrees of freedom.

Using GAUSS, conduct a Monte Carlo simulation with the sample size of 26 and 500 replications under Assumption (A6).
(a) Use the $t_1$ in (5.E.3) to estimate

(i) the true size of the $t$ test for $H_0: \beta = 0$ based on the nominal significance level of 5% and the nominal critical value based on the standard normal distribution are used.

(ii) the true size of the $t$ test for $H_0: \beta = 0$ based on the nominal significance level of 5% and the nominal critical value based on the $t$ distribution with 25 degrees of freedom.

(iii) the true critical value of the $t$ test for the 5% significance level,

(iv) the power of the test at $\beta = 0.15$ based on the nominal critical value,

(v) the size corrected power of the test.

(b) Use the $t_2$ in (5.E.4) and repeat the exercises (a) – (e).

For the starting seed, use 3648xxxx, where xxxx is your birth date, such as 0912 for September 12. Submit your program and output. For each $t$ ratio, discuss whether it is better to use the standard distribution or the $t$ distribution critical values for this application. Also discuss whether $t_1$ or $t_2$ is better for this application.

References


