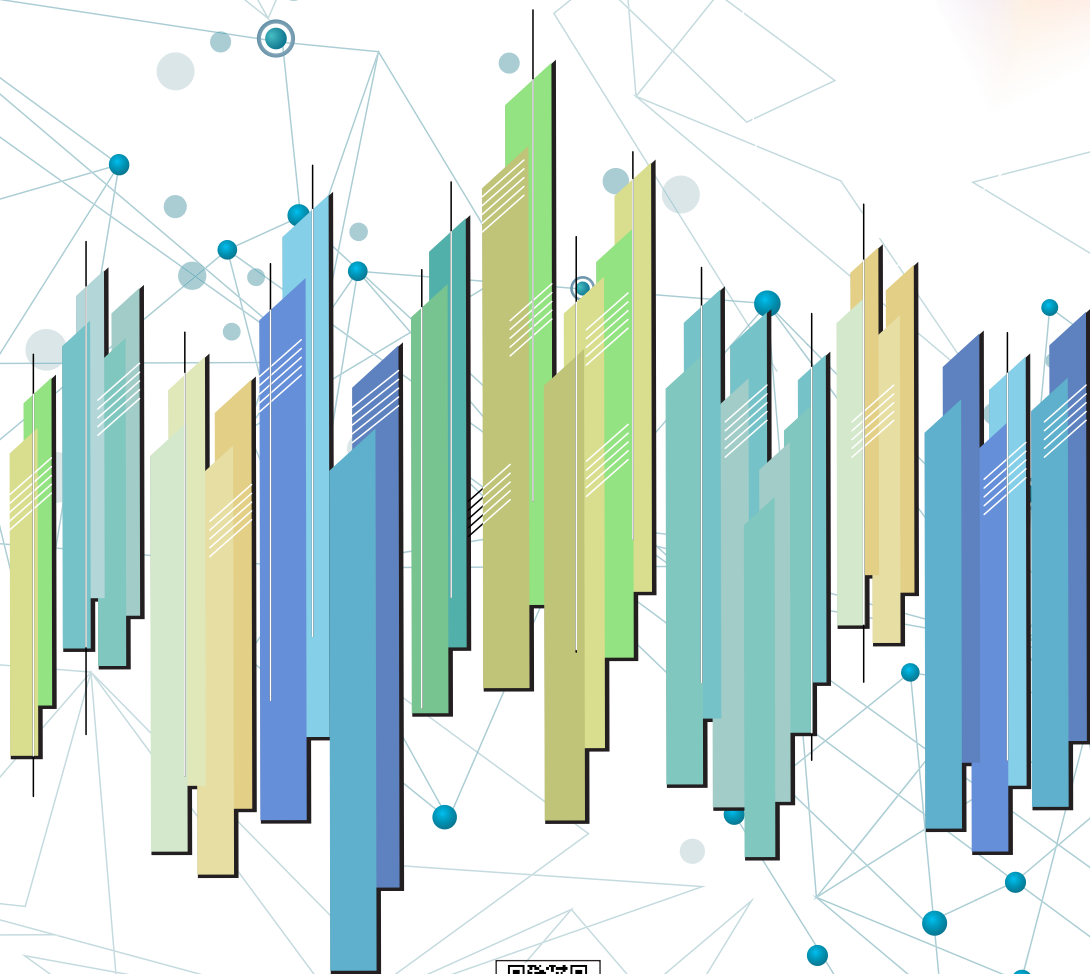


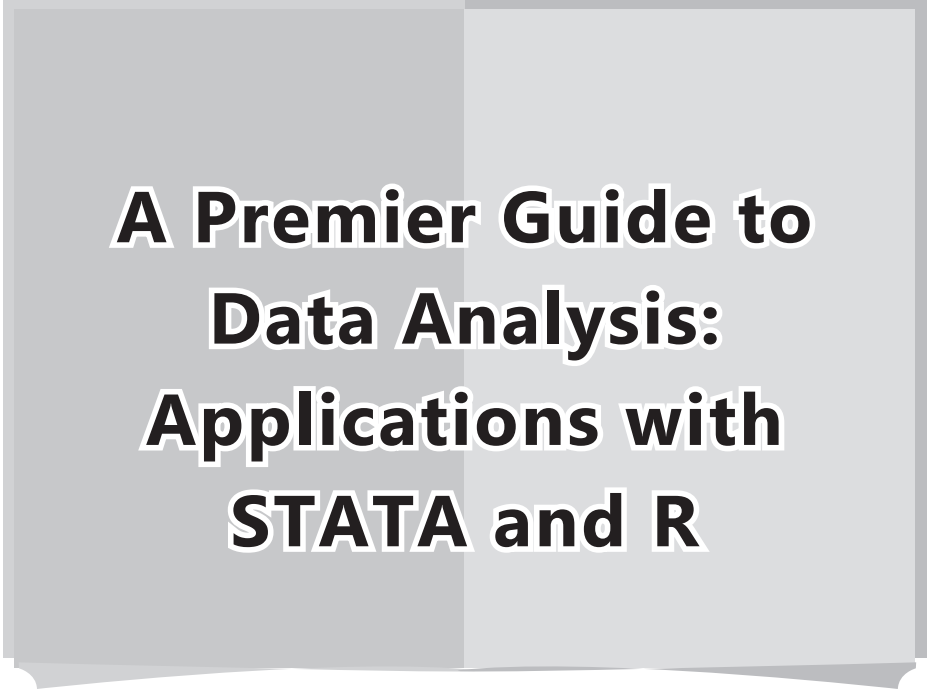
Sajal Jana | Jhumur Sengupta

# A PREMIER GUIDE TO DATA ANALYSIS

Applications with STATA and R



**Sultan Chand & Sons**



**A Premier Guide to  
Data Analysis:  
Applications with  
STATA and R**



# **A Premier Guide to Data Analysis: Applications with STATA and R**

● Dr. Sajal Jana

● Dr. Jhumur Sengupta



**SULTAN CHAND & SONS<sup>®</sup>**  
*Educational Publishers*  
New Delhi

## SULTAN CHAND & SONS®

23, Daryaganj, New Delhi-110002

Phone : 011-23281876, 23266105, 23277843 (*Showroom & Shop*)

011-40234454, 23247051 (*office*)

E-mail : sultanchand74@yahoo.com; info@sultanchandandsons.com

Fax : 011-23266357; Website: www.sultanchandandsons.com

ISBN : 978-93-91820-89-3 (TC-1281)

Price : ₹ 495.00

First Edition: 2024

### EVERY GENUINE COPY OF THIS BOOK HAS A HOLOGRAM



In our endeavour to protect you against counterfeit/fake books, we have pasted a copper hologram over the cover of this book. The hologram displays the full visual image, unique 3D multi-level, multi-colour effects of our logo from different angles when tilted or properly illuminated under a single light source, such as 3D depth effect, kinetic effect, pearl effect, gradient effect, trailing effect, emboss effect, glitter effect, randomly sparking tiny dots, micro text, laser numbering, etc.

*A fake hologram does not display all these effects.*

Always ask the bookseller to put his stamp on the first page of this book.

**All Rights Reserved:** No part of this book, including its style and presentation, can be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopying, recording or otherwise without the prior written consent of the publishers. Exclusive publication, promotion and distribution rights reserved with the Publishers.

**Warning:** An unauthorised act done in relation to a copyright work may result in both civil claim for damages and criminal prosecution.

**Special Note:** Photocopy or Xeroxing of educational books without the written permission of publishers is illegal and against Copyright Act. Buying and Selling of pirated books is a criminal offence. Publication of a key to this book is strictly prohibited.

**General:** While every effort has been made to present authentic information and avoid errors, the author and the publishers are not responsible for the consequences of any action taken on the basis of this book.

**Limits of Liability/Disclaimer of Warranty:** The publisher and the author make no representation or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation warranties of fitness for a particular purpose. No warranty may be created or extended by sales or promotional materials. The advice and strategies contained therein may not be suitable for every situation. This work is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional services. If professional assistance is required, the services of a competent professional person should be sought. Neither the publishers nor the author shall be liable for damages arising herefrom.

**Disclaimer:** The publishers have taken all care to ensure highest standard of quality as regards typesetting, proofreading, accuracy of textual material, printing and binding. However, they accept no responsibility for any loss occasioned as a result of any misprint or mistake found in this publication.

**Author's Acknowledgement :** The writing of a Textbook always involves creation of a huge debt towards innumerable authors and publications. We owe our gratitude to all of them. We acknowledge our indebtedness in extensive footnotes throughout the book. If, for any reason, any acknowledgement has been left out we beg to be excused. We assure to carry out correction in the subsequent edition, as and when it is known.

Printed at: Himani Print Solution, Badarpur, New Delhi-110044.

# Preface



This book is designed for Economics Honours students taking the course on Data Analysis. However, this book can be used by both students and professionals in the industry to develop skills in data analysis using STATA and R. The objective of this book is to discuss data analysis techniques based on statistical and econometric models using STATA and R software packages. This book will assist those who wish to pursue a career as a data analyst. Because data analysis using software packages is important in today's world, every student should be familiar with the techniques for advancing their careers. Companies nowadays require data analysts to create products and solutions for them.

This book is an attempt to provide hands-on STATA and R training. The book is divided into fifteen chapters. The first two chapters cover topics such as data types, different types of variables found in data, and exploratory data analysis techniques. Chapters 3–5 cover the fundamentals of MS Excel, STATA, and R. The chapters go over package installation techniques as well as the use of basic statistical analysis commands. The basic theories of statistics, such as distributions, index numbers, and hypothesis testing for statistical inferences required for data analysis, are covered in chapters 6, 7, 8, and 9. Correlation and regression analysis are covered in Chapters 10 and 11. Time series models, panel data, and non-linear regression models are all covered in Chapter 11. Chapters 12 to 15 analyze Indian data sets from sources such as National Accounts Statistics, National Sample Survey Organization, Reserve Bank of India Data set, and data provided by the Sample Registration System of India.

One of the book's strengths is that it presents difficult techniques in a straightforward, yet rigorous manner. We have included numerous illustrations and practice problems in each chapter to aid comprehension in a simple manner.

We would like to thank our teachers, friends and families whose help and support were essential for completing this book. A special thanks to Sultan Chand & Sons, New Delhi for their assistance throughout the publishing process.

**Dr. Sajal Jana**  
**Dr. Jhumur Sengupta**



# Snapshot of the Book

	<i>Chapter</i>	<i>Exercise</i>	<i>Figures</i>	<i>Tables</i>	<i>Illustrations</i>
1.	Data Management and Data Source	8	18	8	—
2.	Basic Data Analysis	8	15	28	11
3.	Getting Started with MS-Excel	5	30	2	—
4.	An Introduction to STATA	5	9	22	—
5.	An Introduction to R	5	26	12	—
6.	Distribution Functions	7	13	1	7
7.	Sampling Techniques and Survey Design	8	4	17	6
8.	Index Number	5	—	24	24
9.	Hypothesis Testing and Statistical Inference	6	17	8	17
10.	Linear Correlation and Regression	6	5	23	10
11.	Time Series, Panel and Non-Linear Regression Model	4	8	25	4
12.	Analysis of National Accounts Statistics in RStudio	4	9	10	—
13.	Analysis of National Sample Survey Data Using STATA	5	13	13	—
14.	Reserve Bank of India Data Analysis in STATA	7	5	25	—
15.	Analysis of Census Data Using RStudio	5	18	19	—
	<b>Total</b>	<b>88</b>	<b>190</b>	<b>237</b>	<b>78</b>

# Contents



## 01. Data Management and Data Source

01-19

1.1. What is Data?	1
1.2. Definition and Types of Variables	2
1.2.1. Definition	2
1.2.2. Types of Variables	2
1.3. Measurement of Variables	3
1.3.1. Nominal Scale	4
1.3.2. Ordinal Scale	4
1.3.3. Interval Scale	4
1.3.4. Ratio Scale	4
1.4. Errors in Measurement	5
1.5. Primary Data and Secondary Data: Advantages and Limitations	6
1.5.1. Advantages of Primary Data	6
1.5.2. Limitations of Primary Data	6
1.5.3. Advantages of Secondary Data	6
1.5.4. Limitations of Secondary Data	6
1.6. Types of Primary Data	6
1.7. Collection of Primary Data with Questionnaire	7
1.7.1. Dividing the Questionnaire into Different Sections	7
1.7.2. Order of Questions	7
1.7.3. Format of Questions	7
1.7.4. Length of the Questionnaire	7
1.8. A Sample Questionnaire	7
1.9. Scaling and Measurement Techniques for Collection of Primary Data	8
1.9.1. Comparative Scaling Techniques	8
1.9.2. Non-Comparative Scaling Techniques	10
1.10. Secondary Data: Various Sources	11
1.10.1. Publications by Government & Autonomous Organizations	11
1.10.2. Publications by Private Organizations	11





1.10.3. Data from the Internet	11
1.11. Collection of Primary Data	11
1.11.1. Structured Survey	11
1.11.2. Unstructured Survey	12
1.12. Coding and Editing of Data & Construction of Data File	13
1.12.1. Coding and Editing	13
1.12.2. Construction of Data File	14
1.13. Types of Data Analysis	15
1.14. Big Data	18
<i>Exercises</i>	19
<i>Suggested Readings</i>	19

## 02. Basic Data Analysis

21-44

2.1. Understanding the Data	21
2.1.1. Identification of Variables	21
2.1.2. Understanding Data Structure	22
2.2. Coding	23
2.3. Detection of Outliers	23
2.4. Missing Data Management	24
2.5. Presenting Data in Tabular Forms	24
2.6. Frequency Distribution	25
2.7. Graphical Presentation of Data	26
2.7.1. Bar Chart	26
2.7.2. Subdivided Bar Chart	27
2.7.3. Multiple Bar Chart	28
2.7.4. Pie Chart	28
2.7.5. Histogram	29
2.7.6. Pareto Chart	30
2.7.7. Frequency Polygon	31
2.7.8. Ogive	31
2.7.9. Lorenz Curve	33
2.7.10. Scatter Diagram	34
2.7.11. Box Plot	34
2.8. Exploratory Data Analysis	35
2.8.1. Measures of Central Tendency	35
2.8.2. Measures of Dispersions	39
2.8.3. Skewness	40
2.8.4. Kurtosis	42
2.9. Cross Tabulation	42
<i>Exercises</i>	43
<i>Suggested Readings</i>	44

## 03. Getting Started with MS-Excel

45-72

3.1. Excel Spreadsheet	45
------------------------	----

3.2.	Basic Facts about Spreadsheet	45
3.3.	Fundamentals of File Management in Excel	45
3.3.1.	Creating a New Workbook	45
3.3.2.	Opening an Existing Workbook	46
3.3.3.	Saving a Workbook	46
3.4.	Techniques of Data Entry	46
3.5.	Importing Text File	46
3.6.	Export of Data to Text File	48
3.7.	Understanding Functions and Formulas	48
3.7.1.	Structure of Functions	48
3.7.2.	Using the Insert Function Option	49
3.8.	Working with Mathematical Functions	49
3.8.1.	SUMPRODUCT Function	49
3.8.2.	PRODUCT Function	50
3.8.3.	SUM Function	50
3.8.4.	POWER Function	51
3.8.5.	COUNT Function	52
3.8.6.	COUNTA Function	52
3.9.	Creating Basic Formulas	53
3.9.1.	Entering and Editing of Formulas	53
3.9.2.	Arithmetic Formulas	53
3.10.	Statistical Function	54
3.10.1.	LARGE Function	54
3.10.2.	SMALL Function	55
3.10.3.	MEDIAN Function	55
3.10.4.	QUARTILE Function	56
3.10.5.	STDEV Function	57
3.10.6.	MODE Function	57
3.10.7.	RANK Function	58
3.11.	The Analysis Toolpak	58
3.12.	Descriptive Statistics	59
3.13.	Determining Correlation	60
3.14.	Regression Analysis	61
3.15.	Excel Charts	61
3.15.1.	Embedded Chart	61
3.15.2.	Move and Resize a Chart	62
3.15.3.	Change the Chart Type	62
3.15.4.	Basic Steps for Creating a Chart	62
3.15.5.	Insert Axis Title	63
3.15.6.	Plotting a Best-Fit Trend Line	63
3.16.	Forecasting with Trend () Function	64
3.17.	Matrix Operation in Excel	64
3.17.1.	Solving Matrices with Excel	65
3.17.2.	Steps to Calculate $A^{-1}$	65
3.17.3.	Finding the Matrix Product AB	66
3.17.4.	Solving the System of Equations	66
3.18.	Solver in Excel	67
3.18.1.	Steps to Install the Solver Add-Ins Programme	68
3.18.2.	Solving an LP Problem Using Solver	68



<i>Exercises</i>	71
<i>Suggested Readings</i>	72

## 04. An Introduction to STATA

73-90

4.1. About STATA	73
4.2. STATA Interface	73
4.3. Log Files and Do Files in STATA	74
4.4. Accessing Data in STATA	75
4.5. Data Visualization and Statistical Analysis in STATA	77
4.6. Frequency Distribution	81
4.7. Correlation and Regression Analysis in STATA	82
4.8. Graphics in STATA	84
4.9. Time Series Data Analysis in STATA	86
4.9.1. Generation of Differenced and Lagged Time Series	87
4.9.2. Stationarity Test in Time Series Analysis	87
<i>Exercises</i>	90
<i>Suggested Readings</i>	90

## 05. An Introduction to R

91-112

5.1. About R	91
5.2. Installation and Use of R Software	91
5.3. RStudio Interface	95
5.4. Library	95
5.5. Data Management	96
5.5.1. Import of Data in RStudio	96
5.5.2. Management of Missing Data	99
5.5.3. Saving the Program and Working with Data in RStudio	100
5.6. Functions and Assignments	100
5.7. Calculation of Descriptive Statistics	101
5.7.1. Measures of Central Tendency	102
5.7.2. Measures of Skewness and Kurtosis	103
5.7.3. Regression and Correlation	104
5.7.4. Absolute and Relative Frequency	107
5.7.5. Graphics in RStudio	107
<i>Exercises</i>	112
<i>Suggested Readings</i>	112

## 06. Distribution Functions

113-136

6.1. Introduction	113
6.2. Normal Distribution	113
6.2.1. Calculation of Cumulative Distribution Function of Normally Distributed Random Variable	114

6.2.2.	Characteristics of Normal Distribution and Normal Probability Curve	115
6.2.3.	Calculation of Mean and Variance of a Normally Distributed Random Variable	119
6.2.4.	Median of Normal Distribution	121
6.2.5.	Mode of Normal Distribution	122
6.2.6.	Moment Generating Function of Normal Distribution	123
6.2.7.	Moments of Normal Distribution	123
6.2.8.	Additive Property of Normal Distribution	125
6.2.9.	Normal Distribution and Central Limit Theorem	126
6.2.10.	Mean and Variance of Sample Mean	126
6.3.	The Chi-Square Distribution	127
6.3.1.	Characteristics of Chi-Square Distribution	127
6.3.2.	Moment Generating Function (M.G.F.) of $\chi^2$ Distribution	128
6.3.3.	Mode of $\chi^2$ Distribution	129
6.3.4.	Skewness of $\chi^2$ Distribution	129
6.3.5.	$\chi^2$ Distribution of the Variance of a Sample from Normal Distribution	130
6.3.6.	Additive Property of $\chi^2$ Variate	130
6.4.	The Student <i>t</i> -Distribution	130
6.5.	<i>F</i> -Distribution	131
6.6.	Relation between <i>t</i> and <i>F</i> -Distribution	132
6.7.	Working with Microsoft Excel	132
	<i>Exercises</i>	135
	<i>Suggested Readings</i>	136

## 07. Sampling Techniques and Survey Design

137-168

7.1.	Need for Statistical Information	137
7.2.	Complete Enumeration Survey	137
7.3.	Sampling	138
7.4.	Sampling and Non-Sampling Errors	139
7.5.	Cost Aspects of Complete Enumeration and Survey	139
7.6.	Choice between Sampling and Complete Enumeration	141
7.7.	Simple Random Sampling	142
7.7.1.	The Procedure of Selection of Random Sample	142
7.7.2.	Probability of Drawing a Sample Under SRSWOR and SRSWR	144
7.7.3.	Expectation and Variance of Sample Mean Under SRSWOR and SRSWR	145
7.7.4.	Efficiency of Sample Mean Under SRSWOR and SRSWR	147
7.7.5.	Sample Variance Under SRSWOR and SRSWR	150
7.8.	Stratified Sampling	152
7.8.1.	Procedure of Stratified Sampling	153
7.8.2.	Sample Mean Under Stratified Sampling	154
7.8.3.	Advantages of Stratified Sampling	156
7.8.4.	Allocation Procedure and Choice of Sample Size in Different Strata	156
7.9.	Systematic Sampling	159
7.9.1.	Advantages of Systematic Sampling	160
7.9.2.	Sample Mean Under Systematic Sampling	160



7.10. Cluster Sampling	160
7.10.1. Conditions for Cluster Sampling	161
7.10.2. Open Segment and Closed Segment	161
7.10.3. Construction of Clusters	161
7.10.4. Case of Equal Clusters	161
7.11. Sampling in STATA	162
<i>Exercises</i>	166
<i>Suggested Readings</i>	167

## 08. Index Number

169-190

8.1. Meaning of Index Number	169
8.2. Characteristics of Index Numbers	169
8.3. Use of Index Numbers	170
8.4. Methods of Construction of Index Number	170
8.4.1. Aggregative Methods	170
8.4.2. Relative Methods	172
8.5. Methods of Construction of Quantity Index Number	173
8.6. Tests of Index Numbers	174
8.6.1. Time Reversal Test	174
8.6.2. Factor Reversal Test	175
8.6.3. Circular Test	177
8.7. Chain Index Number	177
8.8. Cost of Living Index Number (CLI)	178
8.9. Base Shifting, Splicing and Deflating	184
8.9.1. Base Shifting	184
8.9.2. Splicing Method	185
8.9.3. Deflating Method	187
<i>Exercises</i>	188
<i>Suggested Readings</i>	190

## 09. Hypothesis Testing and Statistical Inference

191-222

9.1. Introduction	191
9.2. Use of Test Statistics for Estimation of Population Parameters	191
9.2.1. Standard Normal Variable ( $z$ Statistic) for Estimating the Population Mean $\mu$ When the Population Variance $\sigma$ is Known	191
9.2.2. $t$ Statistic for Estimating the Population Mean $\mu$ When $\sigma$ Unknown	191
9.2.3. $t$ Statistic for Estimating the Difference between Two Population Means $\mu_1$ and $\mu_2$ When a Common Variance $\sigma^2$ Unknown	192
9.2.4. $\chi^2$ Statistic for Estimating the Population Variance $\sigma^2$ when $\mu$ is Known	192
9.2.5. $\chi^2$ Statistic for Estimating the Population Variance $\sigma^2$ when $\mu$ is Unknown	192

9.2.6.	$\chi^2$ Statistic for Testing Goodness of Fit	193
9.2.7.	$F$ Statistic for Estimating the Differences between Two Population Variances $\sigma_1^2$ and $\sigma_2^2$ when $\mu_1$ and $\mu_2$ are Unknown	193
9.3.	Computation of Confidence Interval	193
9.3.1.	Confidence Interval of $\mu$ when $\sigma^2$ is Known	193
9.3.2.	Confidence Interval of Population Variance $\sigma^2$ when $\mu$ is Known	194
9.3.3.	Confidence Interval of Population Variance $\sigma^2$ when $\mu$ is Unknown	194
9.4.	Hypothesis Testing	196
9.4.1.	Null and Alternative Hypothesis	196
9.4.2.	Steps of Hypothesis Testing	197
9.5.	$\alpha$ -Value Approach to Hypothesis Testing	199
9.6.	Tests of Population Mean Using $t$ Test	200
9.7.	Tests of the Population Proportion (Large Samples)	201
9.8.	Hypothesis Testing of Regression Coefficient	205
9.8.1.	Formation of $H_0$ & $H_A$	205
9.8.2.	Use of Standard Normal Test Statistic in Testing the Statistical Significance of Estimated Values of Regression Parameters (When $\sigma^2$ Is Known)	205
9.8.3.	Use of $t$ -Test Statistic in Testing the Statistical Significance of Estimated Values of Regression Parameters (When $\sigma^2$ Is Unknown)	205
9.8.4.	Use of Chi-Square ( $\chi^2$ ) Test Statistic in Testing Significance of $\sigma^2$	205
9.8.5.	One Tailed $t$ -Test in Regression to Test the Statistical Significance of $\hat{\beta}$	206
9.8.6.	Two Tailed $t$ -Test in Regression to Test the Statistical Significance of $\hat{\beta}$	206
9.9.	Type I and Type II Errors	207
9.10.	Power of a Test	207
9.11.	Hypothesis Testing with Excel	209
9.12.	Nonparametric Test	217
9.12.1.	Test for Randomness: The Run Test	217
9.12.2.	Signed Rank Test	217
9.12.3.	Goodness of Fit Test: Kolmogorov-Smirnov Test	218
9.12.4.	Equality of Two Means: Mann-Whitney $U$ Test	219
	<i>Exercises</i>	220
	<i>Suggested Readings</i>	221

## 10. Linear Correlation and Regression

223-251

10.1.	Correlation	223
10.1.1.	Interpretation of Correlation Coefficient	225
10.1.2.	Rank Correlation	226
10.1.3.	Partial Correlation	227
10.2.	Regression Analysis	227
10.2.1.	Least Squares Method of Estimation	228



10.2.2.	Assumptions of Regression Analysis	228
10.2.3.	Relation between Regression Parameters and Correlation Coefficient	230
10.2.4.	Steps in Bivariate Regression Analysis	231
10.2.5.	Scatter Plot and Best Fit Line	233
10.2.6.	Coefficient of Determination for Measuring the Strength of Association	233
10.3.	Multiple Regression	235
10.3.1.	Estimation of Regression Parameters	235
10.3.2.	Statistical Measures Related to Multiple Regression	236
10.4.	Data Problems in Regression Analysis	238
10.4.1.	Multicollinearity Problem	238
10.4.2.	Heteroscedasticity	240
10.4.3.	Autocorrelation	245
10.5.	Dummy Variable Regression	248
	<i>Exercises</i>	250
	<i>Suggested Readings</i>	251

## 11. Time Series, Panel and Non-Linear Regression Model

253-282

11.1.	Time Series Data and Components	253
11.2.	Stationarity of Time Series Data	253
11.2.1.	Conditions of Stationarity	253
11.2.2.	Stochastic and Deterministic Trend	255
11.2.3.	Conversion of a Non-stationary Series to a Stationary Series	256
11.2.4.	Dicky Fuller Test for Checking Stationarity of Time Series	256
11.2.5.	Augmented Dicky Fuller Test	257
11.3.	Selection of Lag Lengths	257
11.4.	Time Series Process	259
11.4.1.	White Noise Time Series	259
11.4.2.	Auto Regressive Process (AR)	260
11.4.3.	Moving Average Process (MA)	260
11.4.4.	Auto Regressive Moving Average Process (ARMA)	260
11.4.5.	Auto Regressive Integrated Moving Average Process (ARIMA)	260
11.5.	Box-Jenkins Methodology and Forecasting	260
11.5.1.	Identification of the ARIMA ( $p, d, q$ ) Model	261
11.5.2.	ACF and PACF to Choose the Appropriate ARIMA Model	261
11.6.	Vector Autoregressive Model	264
11.6.1.	VAR Model Specification	264
11.6.2.	Granger Causality in VAR	265
11.7.	Volatility in Time Series	267
11.8.	Panel Data Analysis	269
11.8.1.	Pooled Regression Model	271
11.8.2.	Fixed Effect Model	271
11.8.3.	Random Effect Model	271
11.8.4.	Tests for Choosing the Appropriate Panel Data Model	271
11.9.	Non-Linear Regression	274
11.9.1.	Linear Probability Model (LPM)	274
11.9.2.	Logit Model	276

11.9.3. Probit Model	278
11.9.4. Tobit Regression	279
<i>Exercises</i>	281
<i>Suggested Readings</i>	282

## 12. Analysis of National Accounts Statistics in RStudio 283-301

12.1. Basic Structure of SNA, 2008	283
12.2. Components of SNA 2008	283
12.3. Measurement of National Income	283
12.3.1. Gross Value-Added Method	283
12.3.2. Income Method	284
12.3.3. Expenditure Method	284
12.4. National Income Identities	285
12.5. Regional Accounts	285
12.6. Supraregional Sectors	286
12.7. State-Wise Estimates and Other Estimates of Economic Activities by Sectors	286
12.7.1. Agriculture	286
12.7.2. Fishing	286
12.7.3. Mining and Quarrying	286
12.7.4. Registered Manufacturing	286
12.7.5. Unregistered Manufacturing	286
12.7.6. Electricity, Gas, and Water Supply	287
12.7.7. Construction	287
12.7.8. Trade, Hotels, and Restaurants	288
12.7.9. Transport, Storage, Communication	288
12.8. New Series of National Accounts Statistics (Base Year: 2011-12)	288
12.8.1. Improvement in the Coverage of the New Series	288
12.8.2. Classification of Enterprises in the New Series	289
12.9. Estimation of GVA for Unorganized Sector	289
12.10. National Accounts Data Analysis in RStudio	290
12.10.1. Import of GDP Data	290
12.10.2. Decomposition	292
12.11. Forecasting	294
12.11.1. ADF Test	295
12.11.2. Choosing the Appropriate Model of Forecasting	298
12.11.3. Forecasting of GDP	299
<i>Exercises</i>	301
<i>Suggested Readings</i>	301

## 13. Analysis of National Sample Survey Data Using STATA 303-320

13.1. Sample Design and Estimation Procedure of Household Consumption Expenditure Survey and Employment and Unemployment Survey 68th Round	303
13.2. NSSO 68th Round Sample Design	303





13.2.1. Schedules of Enquiry	303
13.2.2. Sample Design of Stage One	304
13.2.3. Sample Design of Stage Two	305
13.2.4. Selection of Households from Each SSS	305
13.3. The Procedure of Extraction of NSSO Data in STATA	306
13.4. Analysis of NSSO Data in STATA	310
13.4.1. Data Visualization and Statistical Analysis	310
13.4.2. Missing Data Management	313
13.4.3. Creation of New Variable	314
13.4.4. Conversion of String Variables	314
13.4.5. Recoding of Variables	314
13.4.6. Bivariate and Multivariate Analysis with NSS Data	315
13.4.7. Regression Model Estimation	316
13.4.8. Regression Model Checking	318
13.4.9. Merge and Append between Two Unit-Level Data Sets	319
<i>Exercises</i>	320
<i>Suggested Readings</i>	320

## 14. Reserve Bank of India Data Analysis in STATA

321-338

14.1. Introduction	321
14.2. Functions of RBI	321
14.3. Hand Book of Statistics on Indian Economy Database	322
14.4. Analysis of Exchange Rate: Exchange Rate Line Plots and Unit Root Tests	323
14.4.1. Lag Selection for Log of US Dollar Series	323
14.4.2. ADF Test of Log of US Dollar Series	324
14.4.3. Lag Selection of First Differenced Log of US Dollar Series	325
14.4.4. ADF Test of First Differenced Series of Log US Dollar	325
14.5. Savings-Investment Analysis: A VAR Model	326
14.5.1. Basic Formulations of Vector Autoregression Model	326
14.5.2. Selection of Lags of VAR Model	326
14.5.3. Vector Autoregressive Model	327
14.5.4. Granger Causality Test	328
14.5.5. Checking for Autocorrelation	329
14.5.6. Checking for Stability of the VAR Model	329
14.5.7. Testing the Cointegration of the Savings Investment Series	330
14.6. Cointegration Analysis on Money Supply and the Price	333
14.7. ARCH GARCH Models for Volatility and Forecasting	335
<i>Exercises</i>	338
<i>Suggested Readings</i>	338

## 15. Analysis of Census Data Using RStudio

339-367

15.1. Sample Registration System	339
15.2. Structure of Sample Registration System	340
15.3. Sample Design in SRS	340

15.3.1. Rural Areas	340
15.3.2. Urban Areas	340
15.3.3. SRS Forms for Data Collection	341
15.3.4. Use of Automation in SRS	341
15.4. Census Data Analysis in RStudio	341
15.4.1. Required Packages	341
15.4.2. Subset Selection of Rows and Columns of Census Data	342
15.4.3. Understanding the Census Data	342
15.4.4. Use of Sample Function for Drawing of Sample from the Population	343
15.4.5. Tidy and Manipulation of Census Data	343
15.4.6. Removing Whitespaces of Census Data	345
15.4.7. Scanning of Census Data	346
15.4.8. Creation and Transformation of Variables	349
15.5. Plots Using Census Data	350
15.5.1. Pie Chart	350
15.5.2. Scatter Plot	351
15.5.3. Histogram and Checking the Normality of Population	351
15.6. Spatial Analysis of Census Data	353
15.6.1. Required Packages	353
15.6.2. Download of Spatial Files	353
15.6.3. Drawing of India Map	354
15.6.4. Use of Summarise Function to Obtain State-Level Information from District-Level Data	356
15.6.5. Mapping Population Figures on the State-Level India Map	357
15.6.6. Finding the Most Populated States	359
15.6.7. Mapping the State-Wise Sex Ratio	360
15.7. Histogram Plots of Male and Female Literacy Rates Based on India District Census Data 2001	363
15.8. Histogram Plots of State Wise Male and Female Literacy Rates Based on India District Census 2001	363
15.9. Scatter Plot of Literacy Rates and Sex Ratio Based on India District Census 2001	365
15.10. Relation Between State-Wise Literacy Rates and Sex Ratio Based on India District Census 2001	366
<i>Exercises</i>	367
<i>Suggested Readings</i>	367

# List of Figures

LF

## 1. Data Management and Data Source

1.1. Data on Car Make and Price	1
1.2. Image Data	1
1.3. Classifications of Variables	2
1.4. Discrete and Continuous Variables	3
1.5. Classifications of Data	3
1.6. Variable Measured on Nominal Scale	4
1.7. Ranking on Ordinal Scale	4
1.8. Temperature on Interval Scale	4
1.9. Height Measurement on Ratio Scale	4
1.10. Different Scaling Techniques	9
1.11. Personal Interview	12
1.12. Telephonic Survey	12
1.13. Mail Survey Through Internet	12
1.14. Observation of Footfalls of Customers	12
1.15. Focus Group Interview	13
1.16. Pie Diagram Showing Preference for Shopping	16
1.17. Bar Plot Showing Frequency Distribution of Income	16
1.18. Bar Plots Showing Frequencies and Preferences for Online Shopping Across Income Groups	17

## 2. Basic Data Analysis

2.1. Bar Chart	27
2.2. Subdivided Bar Chart	27
2.3. Multiple Bar Chart	28
2.4. Pie Chart	29
2.5. Histogram Plot	30
2.6. Pareto Chart	30
2.7. Frequency Polygon	31
2.8. Ogive (Less than Type)	31
2.9. Ogive (More Than Type)	32
2.10. Ogive	33
2.11. Lorenz Curve	33

2.12. Scatter Diagrams	34
2.13. Box Plot	35
2.14. (a) Zero Skewness (b) Positive Skewness (c) Negative Skewness	41
2.15. Kurtosis – (a) Platykurtic, (b) Mesokurtic, (c) Leptokurtic	42

### 3. Getting Started with MS-Excel

3.1. Opening Text Import Wizard	47
3.2. Choosing the Delimiter Comma	47
3.3. Appearance of Import Data Dialog Box	48
3.4. Calculation Using SUMPRODUCT Function	49
3.5. Calculation Using PRODUCT Function	50
3.6. Calculation Using SUM Function	51
3.7. Calculation Using POWER Function	51
3.8. Calculation Using COUNT Function	52
3.9. Calculation Using COUNTA Function	53
3.10. Calculation Using LARGE Function	54
3.11. Calculation Using SMALL Function	55
3.12. Calculation of Median	56
3.13. Calculation of Quartiles	56
3.14. Calculation of Standard Deviation	57
3.15. Calculation of Mode	57
3.16. Calculation Using RANK Function	58
3.17. Computations of Descriptive Statistics	59
3.18. Calculation of Correlation Coefficient	60
3.19. Regression Analysis in Excel	61
3.20. Multiple Bar Chart in Excel	62
3.21. Dialog Box for Drawing a Trend Line in Excel	63
3.22. Trend Line and Forecasting in Excel	64
3.23. Calculation of Inverse Matrix Using MINVERSE Function	65
3.24. Matrix Multiplication Using MMULT Function	66
3.25. Use of MMULT(MINVERSE(),) in Matrix Operations	67
3.26. Solution of Matrix Operations on Systems of Equations	67
3.27. Specifications of Objective Function and Constraints	68
3.28. Use of Simplex Method in Excel	69
3.29. Solver Results Dialog Box for Obtaining the Solution of the LP Problem	70
3.30. Results of LP Problem Using Solver	70

### 4. An Introduction to STATA

4.1. STATA Interface	74
4.2. Creation of Do-Files	74
4.3. Browse Option in Data Import	75
4.4. Data Import in STATA	75
4.5. Scatter Diagram of Income and Education	83
4.6. Histogram of Monthly Income	84
4.7. Histogram for a Range of Income	85
4.8. Pie Diagram Showing Percentage of Males and Females	85
4.9. TS Line Plot of GNIPC	86

### 5. An Introduction to R

5.1. Getting Started with R Installation	91
5.2. CRAN Mirrors	92
5.3. R Download for Windows	92



5.4.	Starting Installation Process	93
5.5.	Download R 4.3.1 for Windows	93
5.6.	RStudio Download	94
5.7.	Install RStudio	94
5.8.	RStudio Interface	95
5.9.	Data from Excel by read_excel	96
5.10.	Import of Data in RStudio by Using Syntax	97
5.11.	Import Data in RStudio by Browser Step 1	98
5.12.	Import Data in RStudio by Browser Step 2	98
5.13.	Import Data in RStudio by Browser Step 3	99
5.14.	Import Data in RStudio by Browser Step 4	99
5.15.	Summary Statistics of BOP Data	101
5.16.	Summary Statistics of Variables in BOP Data	102
5.17.	Exploring BOP Data	102
5.18.	Measures of Central Tendency of the Variable CA in BOP Data	103
5.19.	Measures of Dispersion of Variable CA in BOP Data	103
5.20.	Measures of Quantiles	104
5.21.	Measures of Absolute and Relative Frequencies	107
5.22.	Bar Plot with Absolute Frequency	108
5.23.	Bar Plot with Relative Frequency	108
5.24.	Histogram Plot	109
5.25.	Lorenz Curve Plot	109
5.26.	Correlation Plot	110

## 6. Distribution Functions

6.1.	p.d.f. of Normal Distribution	114
6.2.	p.d.f. of $N(0, 2)$ , $N(0, 1)$ and $N(0, 0.5)$ Distribution	114
6.3.	Area under the Curve for Standard Normal Distribution	117
6.4.	Area under the Standard Normal Curve for $X = 26$ and $X = 40$	118
6.5.	Probability Density of $\chi^2$ with Different Degrees of Freedom	127
6.6.	Calculation of Cumulative Probability of Normal Distribution	133
6.7.	Output of Z-value Calculation	133
6.8.	Calculation of Cumulative Probability of $t$ Distribution	133
6.9.	Output of Calculation for $t$ Value	134
6.10.	Calculation of Cumulative Probability of $F$ Distribution	134
6.11.	Output of Calculation for $F$ Value	134
6.12.	Calculation of Cumulative Probability of Chi Square Distribution	135
6.13.	Output of Cumulative Probability Calculation for Chi Square Distribution	135

## 7. Sampling Techniques and Survey Design

7.1.	Behavior of Sampling Error with Sample Size	139
7.2.	Survey Cost Curve	140
7.3.	Diagrammatic Representation of Stratified Sampling	153
7.4.	Diagrammatic Representation of Cluster Sampling	162

## 9. Hypothesis Testing and Statistical Inference

9.1.	Area Under Standard Normal Curve and Confidence Interval	194
9.2.	Area Under $\chi^2$ Distribution Curve and Confidence Interval	195
9.3.	Region of Acceptance of $H_0$ and Critical Regions	198
9.4.	Hypothesis Testing of Population Mean when Population Variance is Known	209

9.5.	Results of Hypothesis Testing of Population Mean when Population Variance is Known	210
9.6.	Hypothesis Testing of Population Mean when Population Variance is Unknown	211
9.7.	Results of Hypothesis Testing of Population Mean when Population Variance is Unknown	211
9.8.	Data Analysis Dialog Box in Excel in Hypothesis Testing of Equality between Population Means when Population Variance is Known	212
9.9.	Dialog Box in Excel for Hypothesis Testing of the Equality between Two Population Means when Population Variance is Known	212
9.10.	Results of Hypothesis Testing of the Equality between Two Population Means when Population Variance is Known	213
9.11.	Data Analysis Dialog Box in Excel in Hypothesis Testing of Equality between Population Means when Population Variance is Unknown	213
9.12.	Dialog Box in Excel for Hypothesis Testing of the Equality between Two Population Means when Population Variance is Unknown	214
9.13.	Results of Hypothesis Testing of the Equality between Two Population Means when Population Variance is Unknown	214
9.14.	Dialog Box in Excel for Hypothesis Testing Using $F$ Test of Equality between Population Variances When Population Variance is Unknown	215
9.15.	Dialog Box in Excel for Hypothesis Testing of the Equality between Two Variances Using $F$ Test	215
9.16.	Results of Hypothesis Testing of the Equality between Two Population Variances using $F$ Test	216
9.17.	Hypothesis Testing of Goodness of fit using Chi-Square Test	216

## 10. Linear Correlation and Regression

10.1.	Diagrams for Different Nature of Correlation between $X$ and $Y$	223
10.2.	Scatter Plots Showing Relations between $X$ and $Y$	231
10.3.	Choosing the Best Fit Regression Line	233
10.4.	Relations between $X_i$ and $\hat{U}_i^2$	241
10.5.	Wage Differential of Males and Females	248

## 11. Time Series, Panel and Non-Linear Regression Model

11.1.	A Non-Stationary Time Series on Gross National Income Per Capita	254
11.2.	A Stationary Series of Nifty50 Index Closing Price	254
11.3.	Stochastic and Deterministic Trends	255
11.4.	TS Line Plot of Log of Price Series	261
11.5.	ACF of $\ln\_price$ Series	262
11.6.	PACF of $\ln\_price$ Series	263
11.7.	TS Line Plot of First Differenced Series of Exchange Rate	268
11.8.	Logistic Distribution	276

## 12. Analysis of National Accounts Statistics in RStudio

12.1.	Classification of New Series	289
12.2.	TS Line Plot of India GDP Agriculture Data	291
12.3.	India GDP Agriculture Series Decomposition	294
12.4.	GDP of India Data at Constant Prices	295
12.5.	First Differenced Series of GDP Constant Prices	296
12.6.	Second Differenced Series of GDP Constant Prices	297
12.7.	ACF of Second Differenced Series of GDP	298
12.8.	PACF of Second Differenced Series of GDP	299
12.9.	Forecast Plot of GDP in RStudio	300



### 13. Analysis of National Sample Survey Data Using STATA

13.1. Pictorial Representation of Sample Design in 68th Round NSSO Survey	306
13.2. 68th Round NSSO Winzip File	307
13.3. Webpage of 68th Round NSS Survey	307
13.4. Nesstar Explorer Page	308
13.5. STATA 68th Round Data File	308
13.6. Nesstar Explorer Page for Export of Data	309
13.7. Exporting 68th Round Data File to STATA	309
13.8. Data Browser Window in STATA	310
13.9. Data Browser Showing Missing Values Replaced by -1 in STATA	314
13.10. Scatter Diagram Showing Relationship between Production and Consumption	315
13.11. Graphs Showing Distribution Plots of Various Transformation of Consumption	316
13.12. Merging of Two Data Sets	319
13.13. Append of Two Data Sets	320

### 14. Reserve Bank of India Data Analysis in STATA

14.1. TS line Plot of Log of Exchange Rate	323
14.2. TS Line Plot of First Difference of Log US Dollar	325
14.3. Companion Matrix Plot	330
14.4. TS Line Plot Price Index	333
14.5. TS Line Plot Broad Money	333
14.6. TS Line Plot of First Difference of US Dollar Exchange Rate	335

### 15. Analysis of Census Data Using RStudio

15.1. Box Plot of Primary Education with Outlier	347
15.2. Box Plot of Primary Education Without Outlier	349
15.3. Pie Diagram of Literacy	350
15.4. Scatter Plot Internet Use Versus Graduate Education	351
15.5. Population Histogram	352
15.6. Population Histogram with Density Plot	352
15.7. Population Density Plot	353
15.8. Spatial Data Download Web Page	354
15.9. India Country Plot	355
15.10. State Wise India Map Plot	355
15.11. District Wise India Map Plot	356
15.12. State Wise Population India Map Plot	358
15.13. Most Populated States of India Map	359
15.14. State-Wise Sex Ratio India Map Plot	362
15.15. Histogram of Female Literacy Rate (District Level)	363
15.16. Histogram of Female Literacy Rate (State Wise)	364
15.17. Histogram of Male Literacy Rate (State Wise)	365
15.18. Scatter Plot Literacy Rate Versus Sex Ratio	365



# List of Tables

## 1. Data Management and Data Source

1.1. Distinction between Quantitative & Qualitative Variables	2
1.2. Variables Measured on Different Types of Scales	5
1.3. Descriptive Data Analysis Based on Variable Types	5
1.4. Data on Online Shopping	14
1.5. Codebook with Variables & Coding Instructions	15
1.6. Frequency Distribution of Online Shopping	16
1.7. Data on Hours of Internet Use & Other Factors	17
1.8. Effect of Income, Age & Gender on Hours of Internet Use	17

## 2. Basic Data Analysis

2.1. Types of Variables in Data Set	21
2.2. Selected Characteristics by Industry Group	22
2.3. CO <sub>2</sub> Emissions Per Capita in India	22
2.4. Panel Data in Long Format	23
2.5. Panel Data in Wide Format	23
2.6. Banking Preference by Customers in Saltlake	24
2.7. Marks Obtained by the Students in Ordered Array Format	24
2.8. Stem and Leaf Display of Marks Obtained in Maths by the Students	25
2.9. Frequency Table Showing Daily Temperature in Delhi	26
2.10. Refrigerator Sales in May 2020 in Kolkata	26
2.11. Educational Qualification as Per Income Group	27
2.12. Sales of Refrigerators in 2018 and 2020	28
2.13. Percentage of Workers' Education in Gems & Jewellery Sector	28
2.14. Percentage of Workers' Education & the Segments	29
2.15. Frequency Distribution of Age Groups	29
2.16. Frequency Distribution of Income Class	30
2.17. Cumulative Frequency Distribution (Less Than Type) of Age	31
2.18. Cumulative Frequency Distribution (More Than Type) of Age	32
2.19. Cumulative Frequency Distribution of Age	32
2.20. Prices of Food Items	34
2.21. Frequency Distribution of Income Class	36
2.22. Monthly Consumption of Food Items & Prices	37
2.23. Frequency Distribution of Age Groups	38





2.24.	Income Class and Frequencies	38
2.25.	Calculation of Standard Deviation	39
2.26.	Age Distribution of Workers	40
2.27.	Gender-Wise Use of Computers	42
2.28.	Gender-Wise Computer Use in Different Age Groups	43

### 3. Getting Started with MS-Excel

3.1.	Mathematical Operators in Excel	54
3.2.	Output of Descriptive Statistics in Excel	59

### 4. An Introduction to STATA

4.1.	Household-Level Data	76
4.2.	STATA Output Describing the Data	77
4.3.	STATA Output Showing Duplicate Variables	77
4.4.	STATA Output Summary Statistics	78
4.5.	STATA Output Detailed Summary Statistics	78
4.6.	STATA Output of Detailed Summary Statistics of Variable education	80
4.7.	STATA Output Mean Calculation	80
4.8.	STATA Output Median Calculation	80
4.9.	STATA Output Standard Deviation Calculation	80
4.10.	STATA Output List of Variables	81
4.11.	STATA Output List of Variable age for a First Few Observations	81
4.12.	STATA Output List of Variable age for a Last Few Observations	81
4.13.	STATA Output One Way Frequency Distribution of Marital Status	82
4.14.	STATA Output Two-Way Frequency Distribution of Sex-Wise Marital Status	82
4.15.	STATA Output Two-Way Frequency Distribution for Sex = 1	82
4.16.	STATA Output Two-way Frequency Distribution for Sex = 2	82
4.17.	STATA Output Correlation Coefficient	83
4.18.	STATA Output Regression Analysis	83
4.19.	Time Series Data on Gross National Income Per Capita	86
4.20.	STATA Output Dicky-Fuller Test on GNIPC	87
4.21.	STATA Output Dicky-Fuller Test on First Difference of GNIPC	88
4.22.	Summary of Commands	88

### 5. An Introduction to R

5.1.	Data on International Trade of India	97
5.2.	Basic Mathematical Operators Used in RStudio	100
5.3.	Logical Operators in RStudio	100
5.4.	MLRM Data on GDP, and Health Expenditure	104
5.5.	RStudio Output of Regression of GDP on Health Expenditure	105
5.6.	RStudio Output Fitted Values of Health Expenditure	105
5.7.	RStudio Output Regression Residuals	105
5.8.	RStudio Output Summary of Regression Results	105
5.9.	RStudio Output Confidence Intervals of Estimators	106
5.10.	RStudio Output ANOVA	106
5.11.	Syntax for Types of Bivariate Plots	110
5.12.	Summary of the Commands	111

## 6. Distribution Functions

6.1. Range and Probabilities of Normal Distribution	117
---	-----

## 7. Sampling Techniques and Survey Design

7.1. Scores of 40 Students in Physics	143
7.2. A Part of Random Number Table	143
7.3. Physics Scores with Identity Numbers	143
7.4. List of Samples Under SRSWOR	148
7.5. List of Samples Under SRSWR	149
7.6. List of Samples Under SRSWOR	152
7.7. Distribution of Sample Variance	152
7.8. The Systematic Sampling	159
7.9. STATA Commands for Simple Random Sampling	162
7.10. STATA Output Simple Random Sampling	162
7.11. STATA Data Browser Showing Repeated Sample Units in SRSWR	163
7.12. STATA Commands for Stratified Random Sampling for Selecting 10 Regions with Less than 50,000 Population	164
7.13. STATA Output Stratified Random Sampling with Less Than 50,000 Population	164
7.14. STATA Commands Stratified Random Sampling for More Than 50,000 Population	165
7.15. STATA Output Stratified Random Sampling with More Than 50,000 Population	165
7.16. STATA Command for Systematic Sampling	165
7.17. STATA Output Systematic Sampling	166

## 8. Index Number

8.1. Calculations of Laspeyres', Paasche's, Marshall-Edgeworth and Fisher's Indices	172
8.2. Calculations of Price Relatives Using Base Year Weights	173
8.3. Calculation of Fisher's Ideal Index and Factor Reversal Test	176
8.4. Calculation of Chain Index Number	178
8.5. Calculation of Food Index	179
8.6. Calculation of CLI	179
8.7. Price Index by Aggregative Method	180
8.8. Prices of Items and Their Weights	180
8.9. Calculation of Price Index by Weighted Aggregative and Weighted Mean of Price Relatives	181
8.10. Prices of Commodities and Their Weights	181
8.11. Price Index Using Weighted Average of Price Relatives	182
8.12. Group Items with Their Weights	182
8.13. Calculation of Whole Sale Price Index	182
8.14. Commodities and Their Production	183
8.15. Calculation of Quantity Index Number	183
8.16. Price and Quantity Sold of Commodities	184
8.17. Calculation of Fisher's Ideal Index	184
8.18. Year Wise Index Numbers	185
8.19. Index Numbers in New Base 1994=100	185
8.20. Series A and B with Two Base Years	186
8.21. Calculations When A Spliced to B Series	186
8.22. Calculations When B Spliced to A Series	187
8.23. Income of a Company and Price Index (Base Year 1991 = 100)	188
8.24. Calculation of Deflated Series of Company Income	188



## 9. Hypothesis Testing and Statistical Inference

9.1. Drive of Golf Ball in Yards	198
9.2. Gender Wise Level of Education	203
9.3. Calculation of Dependence	204
9.4. Acceptance and Rejection of Null hypothesis ( $H_0$ )	207
9.5. Signed Rank Test for Investigating Population Mean Grade is 6.5	218
9.6. College Wise Student Scholarships	218
9.7. Workers' Performance Under Two Training Methods	219
9.8. Calculation of Relative Effectiveness of Training Methods	219

## 10. Linear Correlation and Regression

10.1. Height and Weight of Students in a Class	224
10.2. Calculation of Linear Correlation	224
10.3. Data on $X$ and $Y$ With Zero Covariance	225
10.4. Customer Rankings of Car Manufactures	226
10.5. Calculation of Rank Correlation	226
10.6. Sales and Advertisement Expenditure Data	229
10.7. Calculation of Regression Parameters	229
10.8. Analysis of Variance for Two Variable Regression	234
10.9. Income Versus Consumption of Households	234
10.10. Analysis of Variance for Multiple Regression with Two Independent Variables	236
10.11. Monthly Income, Education and Work Experience of Workers in a Firm	237
10.12. STATA Output Regression Results	237
10.13. Consumption, Income and Wealth of Consumers	239
10.14. Calculation of Regression Parameters	239
10.15. Output Produced and Capital Stock Per Labour Unit	242
10.16. STATA Output Regression Results	243
10.17. STATA Output Heteroscedasticity Test Result	244
10.18. STATA Output Regression Results after Log Transformation	244
10.19. STATA Output Heteroscedasticity Result after Log Transformation	245
10.20. Yearly Data on Dividend Rate and Profit of a Company	246
10.21. Income Consumption of Males and Female workers	249
10.22. Income Consumption of Males and Female with Gender Codes	249
10.23. STATA Output Dummy Variable Regression Results	250

## 11. Time Series, Panel and Non-Linear Regression Model

11.1. Gross National Income Per Capita	258
11.2. STATA Output ADF Test on GNIPC	258
11.3. First Difference Series of Gross National Income Per Capita	259
11.4. STATA Output ADF Test on First Differenced Series of GNIPC	259
11.5. Box Jenkins Methodology	260
11.6. STATA Output ADF Test on $\ln\_price$ Series	262
11.7. STATA Output ARIMA Model	263
11.8. Annual Time Series Data on $X$ and $Y$	265
11.9. STATA Output Using varsoc Command	266
11.10. STATA Output VAR Regression	266
11.11. STATA Output Granger Causality Results	267

11.12.	STATA Output Using varsoc on D_EURO	268
11.13.	STATA Output ADF Test on D_EURO	269
11.14.	STATA Output Auto Regression Results	269
11.15.	STATA Output LM Test	269
11.16.	Panel Data on Income and Consumption	272
11.17.	Panel Data with State and Year Code	272
11.18.	Binary Data on Employment and CGPA	275
11.19.	Calculation of LPM Model	275
11.20.	LPM Regression Results	276
11.21.	STATA Output Odds Ratio of Logit Regression	277
11.22.	STATA Output Coefficients of Logit Regression	278
11.23.	STATA Output Probit Model	279
11.24.	Categorical Data on Employment and Years of Experience	280
11.25.	STATA Output Tobit Regression	281

## 12. Analysis of National Accounts Statistics in RStudio

12.1.	India GDP Agriculture Time Series	291
12.2.	India GDP Agriculture Data	292
12.3.	India GDP Agriculture Seasonal Component	292
12.4.	India GDP Agriculture Trend Component	293
12.5.	India GDP Agriculture Random Component	293
12.6.	RStudio Output ADF Test GDP	296
12.7.	RStudio Output ADF Test First Difference GDP	297
12.8.	RStudio Output ADF Test Second Difference GDP	297
12.9.	RStudio Output of auto.arima Function	299
12.10.	Forecast Values of GDP Data	300

## 13. Analysis of National Sample Survey Data Using STATA

13.1.	Formation of Hamlet Groups	304
13.2.	Formation of Hamlet Groups in Remote Areas	305
13.3.	Second Stage Sampling	305
13.4.	STATA Output Duplicates Report of HHID	310
13.5.	STATA Output Sector Wise Frequency	312
13.6.	STATA Output Sector Wise Cumulative Frequency	312
13.7.	STATA Output Sector Wise Frequency in Sub Rounds	312
13.8.	STATA Output Sector Wise Two Way Frequencies Between Availability of Social Security Benefits and Methods of Payment	313
13.9.	STATA Output Summary Statistics of Log of Consumption and Production	316
13.10.	STATA Output Regression Results of Log of Consumption on Production	317
13.11.	Differential Effect of Sectors on Per Capita Consumption	317
13.12.	Display of Model 1 and Model 2	318
13.13.	Heteroscedasticity Test of Regression Residual	319

## 14. Reserve Bank of India Data Analysis in STATA

14.1.	STATA Output for Selection of Lag Lengths at Levels Using varsoc Command	323
14.2.	STATA Output ADF Test Results of Log US Dollar Using Lag Length One	324
14.3.	STATA Output ADF Test Results of Log US Dollar Using Lag Length Three	324
14.4.	STATA Output for Selection of Lag Lengths for First Difference Series Using varsoc Command	325



14.5.	STATA Output ADF Test for First Difference Series Using Lag Zero	326
14.6.	STATA Output for Selection of Lag Lengths of VAR Model	327
14.7.	STATA Output VAR Model	327
14.8.	STATA Output Granger Causality Test Results	328
14.9.	STATA Output Lagrangian Multiplier Test	329
14.10.	STATA Output Stability of VAR Model	329
14.11.	STATA Output ADF Test of Ins Using Lag One	330
14.12.	STATA Output ADF Test of Ini Using Lag One	331
14.13.	STATA Output ADF Test of First Difference of Ins Using Lag One	331
14.14.	STATA Output ADF Test of First Difference of Ini Using Lag One	331
14.15.	STATA Output Regression of Ini on Ins	332
14.16.	STATA Output ADF Test of Regression Residuals	332
14.17.	STATA Output Regression of Log of Price Index on Log of Broad Money	334
14.18.	STATA Output Selection of Lag Lengths for Regression Residuals	334
14.19.	STATA Output ADF Test of Regression Residuals	335
14.20.	STATA Output ADF Test of Log US Dollar	336
14.21.	STATA Output ADF Test of First Difference of Log US Dollar	336
14.22.	STATA Output Autoregression Results	336
14.23.	STATA Output LM Test Results	337
14.24.	STATA Output ARCH Results	337
14.25.	STATA Output GARCH Results	338

## 15. Analysis of Census Data Using RStudio

15.1.	Number of Sample Units at Different Replacement Periods	339
15.2.	RStudio Output Attributes of Data ORISSA_selected	343
15.3.	RStudio Output Summary Statistics of Data ORISSA_selected	343
15.4.	RStudio Output Showing Variable Names of Data ORISSA_selected	343
15.5.	RStudio Output Showing the Data Frame of ORISSA_selected	343
15.6.	RStudio Output Summary Statistics of Data ORISSA_selected After Converting The Characters to Factors	344
15.7.	RStudio Output Head of Data ORISSA_selected	345
15.8.	RStudio Output Summary Statistics of Z Score of the Variable Primary Education	347
15.9.	RStudio Output Showing Outliers in the Data on Primary Education	347
15.10.	RStudio Output Quantiles of The Variable Primary Education	348
15.11.	RStudio Output IQR of Primary Education	348
15.12.	RStudio Output Lower Quartile of Primary Education	348
15.13.	RStudio Output Upper Quartile of Primary Education	348
15.14.	RStudio Output Head of data-sexratio	360
15.15.	RStudio Output Head data_sum_sexratio	361
15.16.	RStudio Output Head of data_sum_literacy	363
15.17.	RStudio Output Head of data_sum_literacy_sexratio	366
15.18.	RStudio Output Regression of Sexratio on Literacy Rates Coefficients	366
15.19.	RStudio Output Regression of Sexratio on Literacy Rates Results	367

## About the Book

The book is written to provide a strong foundation of data analysis techniques based on statistical and econometric models using STATA and R. The objective is to explain the concepts and their applications with practical illustrations. It covers topics such as data representation, statistical techniques, and regression analysis including non-linear, time series, and panel data models. All the chapters include real-life illustrations and use real-world data sets to provide examples of how to explore data, build models, find results, and evaluate using codes in STATA and R. The book attempts to provide an easier learning experience to the readers. The practical approach would enable readers to develop the required skills to perform data analysis using STATA and R.

## About the Authors

**Dr. Sajal Jana** is currently attached to Dinabandhu Andrews College, Kolkata, in the capacity of Assistant Professor of Economics. He has published more than fourteen research articles in various peer-reviewed Journals of National and International repute. He was awarded the Silver Medal for securing 2nd position in Post Graduate degree from Vidyasagar University. He received his M.Phil degree from Jadavpur University, Calcutta, and obtained his Ph.D. from University of Burdwan. His research interest includes Production Economics, Efficiency Analysis using the stochastic frontier approach, and applied industrial economics.



**Dr. Jhumur Sengupta** is an Assistant Professor of Economics at Dinabandhu Andrews College, Calcutta, India. She has more than 19 years of experience in the fields of Econometrics, Data Analysis, and Quantitative Economics. Her previous affiliations include Assistant Professor at International School of Business, Calcutta; Jaypee Business School, Noida; South City College, Calcutta and Kirorimal College under Delhi University. She got her Master's Degree and M.Phil Degree in Economics from Jawaharlal Nehru University, New Delhi. She completed her Ph.D. at the University of Calcutta. Her research areas include Mathematical Economics, Econometrics, Empirical Economy, and Political Economy. She published the Book, *Introduction to Econometrics* published by Sultan Chand & Sons, New Delhi. She has published several research papers based on empirical research in various peer-reviewed Journals of National and International repute. She has a passion for undertaking research in areas of Empirical and Quantitative Economics.



## Sultan Chand & Sons

*Publishers of Standard Educational Textbooks*

23 Daryaganj, New Delhi-110002  
Phones (S) : 011-23281876, 23266105, 41625022  
(O) : 011-23247051, 40234454  
Email : sultanchand74@yahoo.com  
info@sultanchandandsons.com



Scan to Visit Us

ISBN 978-93-91820-89-3



TC 1281

9 789391 820893