DATA ANALYSIS *with* STATA

A Comprehensive Guide for Data Analysis and Interpretation of Outputs

First Edition

Stata Version 13

MOHAMMAD TAJUL ISLAM

RUSSELL KABIR

MONJURA NISHA

Data Analysis with
Stata

A Comprehensive Guide for Data Analysis and Interpretation of Outputs

First Edition
Stata Version 13

Mohammad Tajul Islam
MBBS, DTM&H, MSc (CTM), MPH
Professor (Adjunct), Department of Public Health
North South University and State University of Bangladesh

Russell Kabir
BDS, MPH, MSc, PhD
Senior Lecturer (Research Methods), Course Leader (MSc Public Health and Community Wellbeing and MPH Global Public Health)
Anglia Ruskin University, UK

Monjura Nisha
BDS, MPH, PhD
Postdoctoral Research Fellow
The Daffodil Centre, The University of Sydney,
A joint venture with Cancer Council NSW
NSW, Australia

ASA Publications
Dhaka, Bangladesh

ISBN 978-984-35-3165-0 [First Edition]

Copyright © 2022 by Authors

First published 2022

This book is copyrighted by the authors. However, the e-book is free for everyone. Anyone can download it from the links, print it out for personal use, and share it with others, but it is strictly prohibited to use it for any kind of profit-making venture without the written permission of the first author. Its contents may be used and incorporated into other materials with proper acknowledgements and citations. The datasets provided in the links and used in this book are hypothetical and can be used for practice.

Publisher

ASA Publications
Dhaka, Bangladesh
Email: asapublications7@gmail.com

Distributor

Altaf Medical Book Center
Shop: 121 & 128; Lane: 3; Islamia Market, Nilkhet, Dhaka 1205
Cell: +880-1711-985991; +880-1611-985991; +880-1511-985991

Price: BDT 500 (GBP 15; US$ 20)

Production credits

Editor: Mohammad Tajul Islam
Proofreading: Russell Kabir and Md. Golam Kibria
Composition and Cover Design: Zariath Al-Mamun Badhon
Production Director: Md Altaf Hossain
Printing: ASA Publications
Binding: Rahim Bindings

Suggested Citation

Islam MT, Kabir R, Nisha M, eds. Data Analysis with Stata: A Comprehensive Guide for Data Analysis and Interpretation of Outputs. Dhaka, Bangladesh: Altaf Publications; December 2022.

Printed in Bangladesh

To all our family members, students, and promising researchers in health and social sciences

Preface

Stata is a popular and versatile data analysis software. Many books and guidelines on Stata are available online or in the market. Stata books focusing on the analysis of health-related data are scarce. Most of the books used data from sources that are related to business or psychology, with which health researchers are mostly unfamiliar. All these factors, and the inspiration of students, encouraged us to write this book. In this book, we have used simple variables that are commonly used in health and social science-related research as examples to make it easy and understandable for users.

This book is intended for students (MPH, FCPS, MD, MS, MPhil, PhD, and others), teachers, and young researchers in health and social sciences. It is written in very simple language. This book answers three basic questions about data analysis. These are: a) what to do (what statistics to use for data analysis in order to achieve the objectives); b) how to do (how to analyze data using Stata); and c) what do the outputs mean (how to interpret the outputs). All these questions are answered in an understandable manner with examples.

This book covers more than the basic statistical methods of data analysis that are commonly used in health and social sciences research. It is the gateway to learning statistics and Stata together, and will help the users go further. This book covers data management, descriptive statistics, hypothesis testing using bivariate and multivariable analysis, and others. It is easier to learn through exploration than only reading. Users are encouraged to explore further once the basics are known. From our understanding, using the statistics covered in this book, students and researchers will be able to analyze most of their data from epidemiological studies and publish them in international peer-reviewed journals.

We are optimistic that students and researchers will find this book useful while analyzing their data and interpreting the outputs. If you have any comments or suggestions about this book, feel free to write to the e-mail address below.

M. Tajul Islam
abc.taj@gmail.com

Foreword

The book titled "Data Analysis with Stata", written by Dr. M. Tajul Islam, Dr. Russell Kabir, and Dr. Monjura Nisha, is an immense contribution to teaching, training, and research in the field of medical sciences and relevant disciplines. The authors of the book have provided a step-by-step approach from data management to data analysis using Stata in this book.

The book is a useful resource for data analysis using Stata for undergraduate, postgraduate, and doctoral students of medical sciences and other health-related disciplines, such as Dental, Pharmacy, Nursing, Occupational Health, Physical Medicine, Biomedical and Social Sciences.

Each chapter is written in a simple manner, focusing on the needs of students. Contents are systematically presented and described so that students can easily grasp the process of data analysis and interpret the outputs.

This book will serve as a helpful training tool for university lecturers, dissertation supervisors, and data analysis instructors.

Readers will benefit from using this book, and I wish for the extensive circulation of the book.

Professor Hafiz T.A. Khan
PhD (Edin Napier), PostDoc (Oxon), FRSPH, CStat (UK)
Professor of Public Health and Statistics
College of Nursing, Midwifery and Healthcare
University of West London
Middlesex TW8 9GB
United Kingdom
Email: hafiz.khan@uwl.ac.uk

Acknowledgements

We are particularly indebted and grateful to Dr. Shakil Ahmed and Mr. Md. Golam Kibria, who reviewed all the sections, provided their critical views and constructive suggestions, and took the lead role in publishing this book. We are particularly thankful to Professor Hafiz T. A. Khan of the University of West London for his careful review and writing the foreword for this book.

We would like to say thanks to Dr. Ali Davod Parsa, Dr. Richard Hayhoe, Dr. Tim Hayes, Dr. Rachael Cubberley, Dr. Oonagh Corrigan, and Dr. Shannon Doherty of Anglia Ruskin University, UK; Mrs. Madhini Sivasubramanian and Dr. Divya Vinnakota of University of Sunderland, London; Dr. Julia Morgan of Greenwich University, UK; Dr. Md. Anwarul Azim Majumder of University of West Indies; Dr. Brijesh Sathian of Hamad Medical Corporation, Qatar; Dr. Sujita Kumar Kar of King George's Medical University, Lucknow, India; Dr. Shah Jalal Sarker of UCL, UK; and Dr. S. M. Yasir Arafat of Enam Medical College and Hospital, Bangladesh.

Many of our students, who continually pushed and encouraged us to write this book, deserve a share of the credit, including Dr. SM Anwar Sadat and Dr. Md. Naymul Hasan. We, as authors, accept full responsibility for any deficiencies or errors that the book may have. Finally, we express our sincerest thanks for the assistance that we have received from ASA Publications in publishing this book.

Links for the e-book and datasets

Users may download the e-book and datasets used in this book from the links below.

- https://drive.google.com/drive/folders/1tdTEoBuMwObThyuFjehfbMkHlyUcoIp_?usp=sharing
- https://drive.google.com/drive/folders/1D2LSmaz7eYJQwhpy5vZ0TTx7p_nGcTOD?usp=sharing

To the users

This e-book is free for all. However, if you can afford, please donate BDT 100 (US$ 2 for users outside Bangladesh) to a charity or a needy person. This little amount is sufficient to offer a meal for an orphan in developing countries.

Table of Contents

Chapter 1	**Introduction**	**1**
1.1	Version dilemma	1
1.2	Stata interface	2
1.3	Steps of data analysis	2
Chapter 2	**Generating Data Files**	**7**
2.1	Generating data files	7
2.1.1	Generating a data file by typing data in the data editor	7
2.1.2	Generating a data file using copy and paste commands	15
2.1.3	Importing a data file from other programs	15
2.1.4	Deleting and inserting variables	16
2.1.4.1	Deleting a variable	16
2.1.4.2	Inserting a new variable	16
2.1.4.3	Copy a variable into the same data file	17
Chapter 3	**Basics of Stata**	**19**
3.1	Stata files	19
3.1.1	Data file	19
3.1.1.1	Opening (retrieving) an existing data file	19
3.1.2	Output file or log-file	20
3.1.2.1	Saving outputs in a log-file	20
3.1.2.2	Opening an existing log-file	21
3.1.2.3	Browsing Stata outputs in Results window	22
3.1.2.4	Copy tables from Stata outputs to MS Word	22
3.1.2.5	Transformation of a log-file from smcl to ASCII format	22
3.1.3	Do-file or Stata commands file	23
3.1.3.1	Generating a Do-file	23
3.1.3.2	Saving commands in a Do-file	24
3.1.3.3	Executing the commands in a Do-file	24
3.2	Basic command syntax	25
3.3	Knowing the dataset	27
3.3.1	Generating brief description of a dataset	27
3.3.2	Codebook	27
3.4	Sorting of data	29

3.5	Stata operator symbols	30
3.6	Getting help in Stata	30
Chapter 4	**Data Cleaning and Data Screening**	**33**
4.1	Checking for out-of-range errors	33
4.2	Checking for outliers	35
4.3	Assessing normality of data	37
Chapter 5	**Data Management**	**39**
5.1	Converting string variables to numeric variables	39
5.1.1	Converting a string variable with non-numeric values into a numeric variable	40
5.1.2	Converting a string variable with numeric values into a numeric variable	40
5.2	Recoding of data	41
5.2.1	Recoding a string variable into the same variable	41
5.2.2	Recoding a string variable into a different string variable	41
5.2.3	Recoding a numeric variable into the same variable	41
5.2.4	Recoding a numeric variable into a different variable	41
5.3	Making class intervals	43
5.4	Combining data into a new variable	45
5.5	Data transformation	47
5.6	Calculation of total score	48
5.7	Extraction of duration from dates	49
5.8	Relocating variables in the dataset	50
5.9	Selecting a sub-group for analysis	51
5.10	The “egen” command	52
5.11	Creating dummy variables	53
5.12	Transformation of data formats	53
Chapter 6	**Data Analysis: Descriptive Statistics**	**57**
6.1	Frequency distribution	57
6.2	Central tendency and dispersion	59
6.2.1	Interpretation	63
6.3	Descriptive statistics disaggregated by a categorical variable	66

Chapter 7	**Generating Graphs**	**69**
7.1	Histogram	69
7.1.1	Saving graphs	70
7.3	Box and plot chart	73
7.4	Bar graph	76
7.4.1	Bar graph for the mean of a quantitative variable across a categorical variable	76
7.4.2	Bar graph for the frequencies of a categorical variable	76
7.5	Line graph	77
7.6	Pie chart	78
Chapter 8	**Checking Data for Normality**	**81**
8.1	Assessing normality of data	81
8.1.1	Interpretation	83
Chapter 9	**Testing of Hypothesis**	**85**
Chapter 10	**Student's t-test for Hypothesis Testing**	**91**
10.1	One-sample t-test	91
10.1.1	Interpretation	92
10.2	Independent samples t-test	93
10.2.1	Test for equality of variances: Levene's test	93
10.2.2	Commands for independent samples t-test	95
10.2.3	Interpretation	95
10.3	Paired t-test	97
10.3.1	Commands	97
10.3.2	Interpretation	98
Chapter 11	**Analysis of Variance (ANOVA)**	**99**
11.1	One-way ANOVA	99
11.1.1	Commands	100
11.1.2	Interpretation	101
11.1.3	Post hoc test	101
11.1.4	Interpretation of post hoc test results	102
11.1.5	One-way ANOVA for unequal variances	103
11.2	Two-way ANOVA	103

11.2.1	Commands for two-way ANOVA	105
11.2.2	Interpretation	106
11.2.3	Post hoc test for two-way ANOVA	106
Chapter 12	**Repeated Measures ANOVA**	**109**
12.1	One-way repeated measures ANOVA	109
12.1.1	Commands	110
12.1.2	Interpretation	113
Chapter 13	**Association Between Two Categorical Variables: Chi-square Test of Independence**	**117**
13.1	Chi-square test of independence	117
13.1.1	Commands	118
13.1.2	Interpretation	120
13.2	Relative risk and odds ratio	121
13.2.1	Interpretation	125
13.3	Stratified analysis	126
13.3.1	Interpretation	127
Chapter 14	**Hypothesis Test of Proportions**	**131**
14.1	One-sample test of proportion	131
14.2	Two-sample test of proportions	133
Chapter 15	**Association Between Two Continuous Variables: Correlation**	**135**
15.1	Pearson's correlation	135
15.1.1	Scatter plot	136
15.1.2	Commands for Pearson's correlation	136
15.1.3	Interpretation	138
15.2	Spearman and Kendall's tau-b correlations	139
15.2.1	Interpretation	140
15.3	Partial correlation	141
15.3.1	Interpretation	141
Chapter 16	**Linear Regression**	**143**
16.1	Simple linear regression	144

16.1.1	Commands for simple linear regression	145
16.1.2	Interpretation	146
16.2	Multiple linear regression	149
16.2.1	Sample size for multiple regression	150
16.2.2	Commands for multiple linear regression analysis	151
16.2.3	Interpretation	153
16.2.4	Regression diagnostics	154
16.2.4.1	Checking for multicollinearity	155
16.2.4.2	Checking for linearity	156
16.2.4.3	Checking for normality of residuals	158
16.2.4.4	Checking for homoscedasticity	161
16.2.4.5	Checking for outliers	163
16.2.4.6	Test for independence	164
16.2.5	Variable selection for a model	166
16.2.5.1	Backward selection method	167
16.2.5.2	Forward selection method	168
16.2.5.3	Interpretation	169
Chapter 17	**Logistic Regression**	**171**
17.1	Mathematical concept of logistic regression model	172
17.2	Binary logistic regression	173
17.2.1	Unconditional binary logistic regression	173
17.2.1.1	Interpretation	176
17.2.2	Logistic regression diagnostics	178
17.2.2.1	Checking for multicollinearity	179
17.2.2.2	Checking for model fit	180
17.2.2.3	ROC curve	182
17.2.2.4	Other postestimation commands	184
17.2.3	Variable selection for a model	185
17.2.4	Incorporating interaction terms in the model	186
17.2.5	Sample size for logistic regression	188
17.2.6	Conditional logistic regression	188
17.2.6.1	Interpretation	190
17.3	Analysis of cross-sectional data: Estimation of prevalence ratio	191
17.3.1	Poisson regression	193

17.3.2	Cox regression with constant time	193
17.3.3	Generalized linear model	194
Chapter 18	**Multinominal Logistic Regression**	**197**
18.1	Interpretation	198
18.2	Post-estimation commands	200
Chapter 19	**Survival Analysis**	**203**
19.1	Survival analysis: Kaplan-Meier method	204
19.1.1	Preparing data for analysis	205
19.1.2	Commands for Kaplan-Meyer method	206
19.1.3	Interpretation	209
19.2	Cox regression	213
19.2.1	Commands	213
19.2.2	Interpretation	215
19.2.3	Checking for assumptions	217
Chapter 20	**Nonparametric Methods**	**221**
20.1	Mann-Whitney U test	221
20.2	Median test	223
20.3	Wilcoxon signed ranks test	223
20.4	Kruskal-Wallis test	224
20.5	Friedman test	225
Chapter 21	**Analysis of Covariance (ANCOVA)**	**229**
21.1	One-way ANCOVA	229
21.1.1	Commands	231
21.1.2	Interpretation: One-way ANCOVA	233
21.2	Two-way ANCOVA	235
21.2.1	Commands	236
21.2.2	Interpretation: Two-way ANCOVA	238
Chapter 22	**Miscellaneous**	**241**
22.1	Reliability of scales: Cronbach's alpha	241
22.1.1	Interpretation	242
22.2	Constructing wealth quintiles	244
References		**249**

1
Introduction

Stata is a data analysis software that allows us to perform a wide range of statistical analyses. Stata can be operated by using its drop-down menu as well as by writing the commands directly in the command window. The commands of Stata are straightforward and simple. Therefore, data analysis by writing commands is more popular among users. This book is written focusing on the use of Stata's commands for data analysis.

This book is for students, young researchers, and teachers. It provides more than a basic understanding and techniques for statistical analysis of data related to health and social sciences research. This book is based on Stata version 13. There are upper as well as lower versions of Stata. Stata commands are sufficiently similar, and users can use this book for both upper and lower versions. The newer versions are mainly for the more advanced and recently developed techniques that the users most likely do not require.

This book focuses on statistical decision-making, data analysis, and interpretation of the outputs. It covers commonly used data analysis techniques in health and social sciences research. The topics covered in this book include data management, descriptive statistics, and bivariate and multivariable analysis for hypothesis testing, including nonparametric methods and others.

1.1 Version dilemma

Stata is continuously evolving over time. That means that the commands, options, language elements, and others may change in future versions. However, Stata ensures

that the higher versions execute the commands regardless of the version of Stata in which they are written. It is, therefore, expected that all the commands (syntax) used in this book will run in higher (or lower) versions.

This book is based on Stata version 13. If you are using a different version (e.g., version 17), you can still use the commands. If you find any problem in executing a command provided in this book in version 17 (or other versions), type the following command at the top of every Do-file (Chapter 3) that you create:

version 13

This simple step will ensure that your Do-file or program will continue to run not only in version 17 but also in all future versions of Stata, even if that future version has changes in the syntax of some of the commands or programming constructs.

You can also use the above command as a prefix while writing a command in the command window. For example, if you want to execute an ANOVA syntax in version 17, which is written in version 13, use the command:

version 13: anova …..

This command will set Stata's version to 13, run the anova command, and then reset Stata's version to whatever it was before the command was executed. For further information, visit: https://www.stata.com/manuals/pversion.pdf.

1.2 Stata interface

Once you open the Stata (double-clicking the Stata icon), it looks like Figure 1.1. It has six windows, as shown in Figure 1.1 by the numbers 1 to 6. The purposes of the windows are described in Table 1.1. You will also find some useful icons on the Stata toolbar (Fig 1.2). The functions of the icons are provided in Table 1.2. We will discuss Stata in further detail in the subsequent chapters.

1.3 Steps of data analysis

We collect data for our studies using various tools and methods. The most commonly used tools for data collection are questionnaires and record sheets, while the commonly used data collection methods are interviews (face-to-face, telephonic or online), observations, physical examinations, and lab tests. Sometimes we use available data (secondary data) for our research studies, such as hospital records or data from other

studies (e.g., Bangladesh Demographic and Health Survey data). Once data is collected, the steps of data analysis are:

- Data coding, if a pre-coded questionnaire or record sheet is not used
- Development of a data file and data entry
- Data cleaning (checking for errors during data entry)
- Data screening (checking assumptions for statistical tests)
- Data analysis
- Interpretation of results

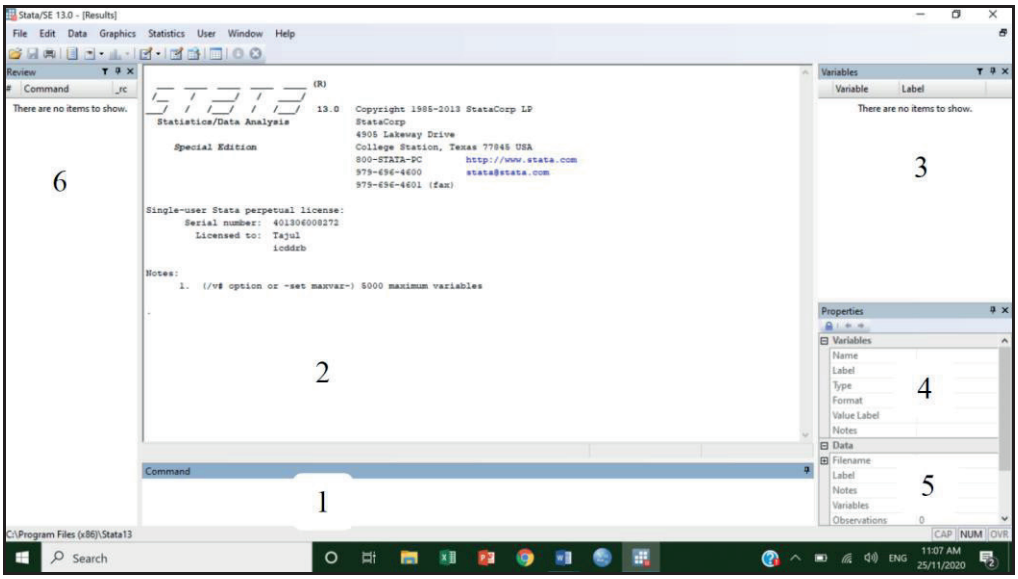


Figure 1.1 Stata interface and windows

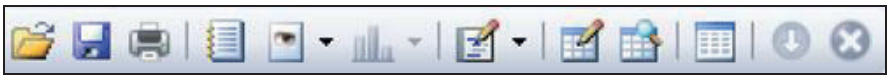


Figure 1.2 Stata toolbar icons

Table 1.1 Stata windows and purposes

Window	Purpose
Command window [1]	This is for writing the commands. We write commands in this window. The commands are executed when the “Enter” key is pressed. You can use the “Page Up” and “Page Down” keys on the keyboard to recall commands from the “Review” window (window 6).
Results or outputs window [2]	This window displays the outputs along with the commands executed. It also shows error messages if there is any problem with the commands. The Results window keeps about 500 lines of the outputs. When this limit is exceeded, Stata deletes the earlier outputs. If you want to save the outputs, you must generate a log-file to store the outputs. You can browse the results using the mouse or <Shift+Page Up/Page Down or Arrow> buttons.
Variables window [3]	Displays the variable names of the dataset currently active in memory (i.e., currently being used). If you double-click on a variable in this window, the variable will appear in the "Command" window.
Properties window: Variables window [4]	Under the “Properties” window, there is a “Variables” window (4) and a “Data” window (5). The “Variables” window shows the variable properties, such as variable name, variable label, etc.
Data window [5]	This window shows the file name, path, number of variables in the dataset, and total observations.
Review window [6]	This window displays the commands already executed during an analysis session. If you click on a command in this window, the command will appear in the “Command” window and can be executed by pressing the “Enter” key. If you double-click on a command in the “Review” window, the command will be directly executed, and the outputs can be seen in the “Results” window.

Table 1.2 Stata toolbar icons and functions

Icon	Function
	To open a data file from the desired location.
	To save the data file.
	To save the log-file (results or outputs file). You can begin, close, suspend, or resume the log-file using this icon.
	This is the new Do-file Editor icon. The Do-file is for writing, editing, and saving the commands. You can generate a Do-file (command file) and edit it using this icon.
	This icon is for going to the Data Editor (Edit mode). In this mode, you can edit or change data in the data file.
	This icon is for going to the Data Editor (Browse mode). You can only browse and see data in this mode, but you cannot change them.
	This icon is for going to the Variables Manager, where you can edit (change) the variable names, variable labels, and value labels.

2

Generating Data Files

Like other data analysis programs, Stata has to read a data file to analyze the data. It is, therefore, necessary to develop a data file or import it from other programs for use by Stata. Data files can be generated by Stata, but it is not a popular practice. Researchers mostly convert or import data files generated in other programs for use by Stata. Data files generated in other programs can be easily transformed into Stata format for analysis. In this chapter, we will discuss how to generate a data file in Stata.

2.1 Generating data files

2.1.1 Generating a data file by typing data in the data editor

For generating a data file, the first and basic step is to decide on a name for each of the variables included in the questionnaire or record sheet. To name a variable, we need to follow certain rules. They are:

- The variable names must be unique (i.e., all the variables should have different names)
- A variable name must be between 1 and 32 characters long. But try to keep it as short as possible.
- Variable names must begin with a letter (small or capital) or an underscore. Variable names cannot begin with a number. Although underscore can be used to begin a variable name, it is strongly discouraged because such variable names are used to indicate temporary variables in Stata.
- Variables cannot include full stop (.), space, or symbols like, ?, *, μ, λ, ~, !, -, @, and #.

- Stata is case-sensitive. For example, "Gender", "gender", and "GENDER" will not be considered as the same variable by Stata. During analysis, you need to type the variable names correctly for execution. We recommend using the variable names with all lowercase letters (e.g., gender).

Once the variable names are decided, the next step is to generate a data file. In Stata, for generating a data file, we start by entering data before inserting the variable names. Suppose that we have collected data using a pre-coded questionnaire (codes are shown in parenthesis) with the following variables:

Categorical variables:

- Sex (m= male; f= female)
- Religion (1= Islam/Muslim; 2= Hindu; 3= Others)
- Occupation (1= Business; 2= Government job; 3= Private job; 4= Others)
- Marital status (1= Married; 2= Unmarried; 3= Others)
- Have diabetes mellitus (1= Yes; 2= No)

Quantitative variables (numerical variables):

- ID number
- Age of the respondent
- Monthly family income
- Systolic blood pressure (BP)
- Diastolic BP

Assume that we have decided to use "age" as the variable name for age, "sex" for sex, and "religion" for religion. Instead of age, sex, and religion, you can use any other names for the variables, such as v1, v2, and v3. It is always convenient to develop a codebook in MS Word or MS Excel where the Stata variable names, actual variable names (variable labels), and variable codes (value labels) are recorded (Table 2.1). The codebook is helpful during data analysis.

It is convenient to use numeric variables instead of string (character) variables for a data file. The numeric variables have numeric codes (e.g., 1= male; 2= female). The string variables may or may not be coded. If a string variable is coded, it is coded with letters (e.g., m= male; f= female). When a string variable is not coded, the data is directly entered into the data file. For example, the data of gender (male/female), religion (Islam/Hindu/Others), and occupation (business/job holder/others) may be entered directly into the data file. Note that Stata does not allow value labels for the

coded string variables (e.g., m= male; f= female).

Table 2.2 shows some data (as an example) that has been collected using a questionnaire (Table 2.1). We will use this data to generate a data file in Stata.

Table 2.1 Questionnaire codebook

Stata variable name	Actual variable name/ variable label	Variable code/ value labels
idno	Identification number	Actual value
age	Age in years	Actual value
sex	Sex*	m= male f= female
religion	Religion	1= Islam/Muslim 2= Hindu 3= Others
occu	Occupation	1= Business 2= Government job 3= Private job 4= Others
income	Monthly family income in Tk.	Actual value
marital	Marital status	1= Married 2= Unmarried 3= Others
diabetes	Have diabetes mellitus	1= Yes 2= No
sbp	Systolic blood pressure in mmHg	Actual value
dbp	Diastolic blood pressure in mmHg	Actual value

*For practical purposes, it is better to consider a numeric variable (e.g., 1= male; 2= female) rather than a string (character) variable for sex (e.g., m= male; f= female) as well as for other variables.

Table 2.2 Data collected from the study subjects (only a portion is shown)

idno	age	sex	religion	occu	income	marital
1	26	m	1	2	25000	1
2	28	f	2	2	35000	1
3	29	f	1	1	60000	1
4	34	m	1	3	20000	2

Open the Stata program by double-clicking the Stata icon. You will see the Stata interface (Stata/SE 13.0) as shown in Figure 1.1 (Chapter 1). The simplest way to generate a data file is through the Data Editor. To go to the Data Editor, select from the menu bar:

Window > Data Editor
Or,
Data > Data Editor > Data Editor (Edit)

Or,

Click the icon on the toolbar.

You will see the “Data Editor (Edit) – Untitled” as shown in Figure 2.1. This is the window for defining the variables as well as for data entry. Use the following steps to generate a data file:

Step 1: Our first variable is "idno" (Table 2.2). When the cursor is placed in the first column of the first row, it will show "var1[1]" in the box above. Type the first value of the variable "idno" as shown in Table 2.2 (which is 1, i.e., just type 1 and press the "Enter" button). You will notice that "var1" appears at the top of the first column (Fig 2.2).

Now, type the value (which is 26) of the second variable "age" in the first box of the second column and press "Enter". You will see that "var2" appears at the top of the second column. Like this, enter the values of other variables into the Stata Data Editor spreadsheet.

If the first value entered for a variable is a number, Stata will consider it a numeric variable and will permit only numbers as its values subsequently. The numeric values may begin with a plus or minus sign, and include decimal points. However, the numbers should not have any commas (,), such as 10,000 or 1,000,000.

If the first value entered for a variable is a non-numeric character (such as m, f, or any other letter), Stata will consider it a string (text) variable. A string variable may have values up to 244 characters long and may have any combination of letters, numbers, symbols, and spaces.

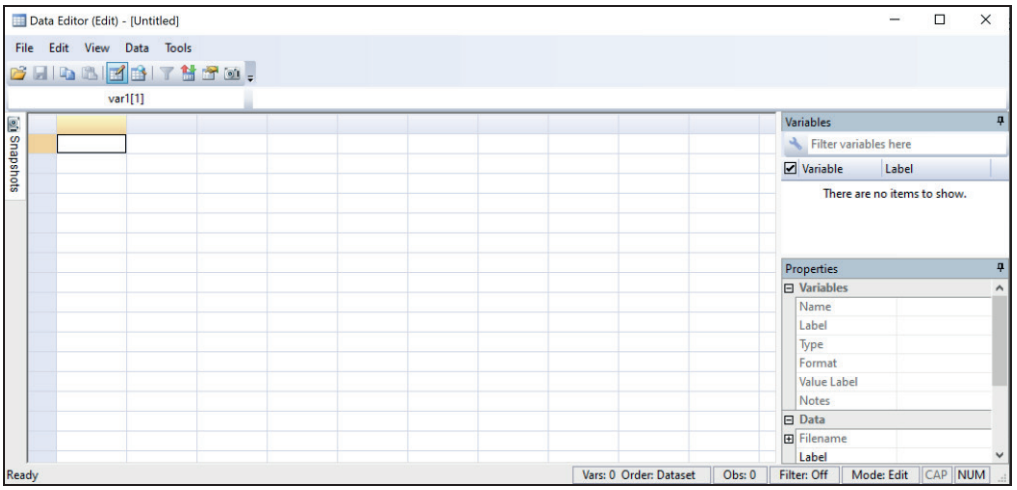


Figure 2.1 Stata data editor spreadsheet

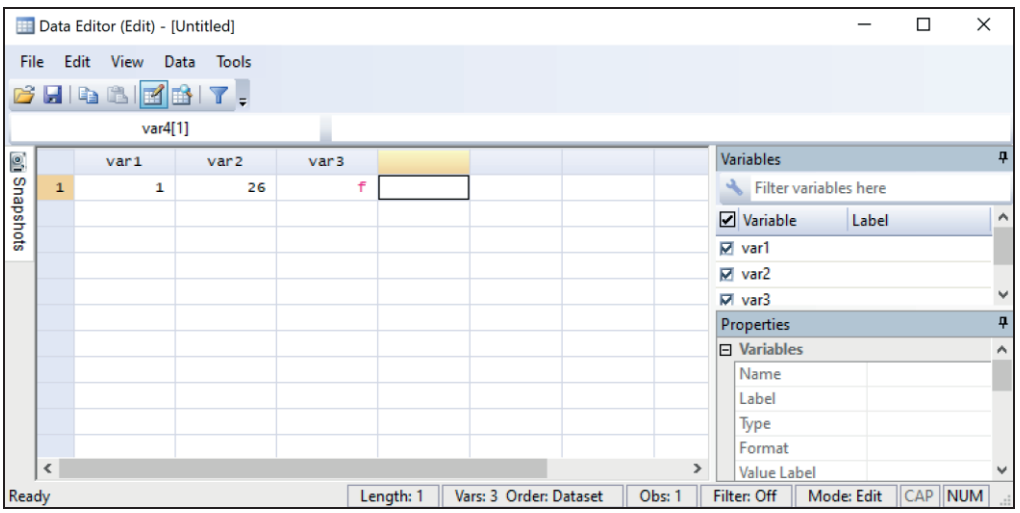


Figure 2.2 Stata data editor with variable names

In the Data Editor or Data Browser, the string variable values appear as red, numeric variable values appear as black, and labeled (coded) numeric variable values appear as blue.

Step 2: In this step, we will rename (replace) the variable names that have been generated automatically by Stata (like var1, var2, and var3) with the variable names as shown in the codebook (Table 2.1). For example, we need to rename the first variable "var1" as "idno", the second variable "var2" as "age", and so on.

You can see that there are three windows on the right side of the Data Editor (Fig 2.1). These are the "Variables", "Properties", and "Data" windows. In the "Variables" window, all the variable names have appeared, like var1, var2, and var3 (Fig 2.2). Click on "var1" in the "Variables" window. Now look at the "Properties" window. It shows "var1" against "Name" in the window. Double-click on "Name" in the "Properties" window, delete "var1" and write "idno". This will replace "var1" with "idno". You can see the new variable name (idno) in the spreadsheet as well as in the "Variables" window. Like this, rename all the variables generated automatically by Stata with the variable names of your choice. You can also use the following command to change the variable name:

rename var1 idno

This command will rename the variable “var1” to “idno”.

Step 3: We will now label the variables one by one. To write the variable label for “idno”, select (click on) the variable “idno” in the “Variables” window (Fig 2.2). Double-click on “Label” in the “Properties” window and write “serial no”. This will be the variable label for “idno” as shown in the “Variables” window for the variable “idno” against “Label”. In this way, complete the variable labels of all variables as shown in the codebook. An alternative to labeling a variable is by using the following command in the command window:

label var idno “serial no”

This will label the variable “idno” as “serial no”.

Step 4: Step 4 is assigning the value labels. Since the variables "idno" and "age" are not categorical variables (i.e., these variables are not coded), they do not need to have value labels. Value labels are only needed for the categorical variables that are coded numerically. *Stata does not allow value labels for string variables*. Therefore, we need

to assign value labels only for the numerically coded variables, such as "religion", "occupation", and others.

In our example, the variable "religion" is a numerically coded variable, and the code numbers are 1= Muslim; 2= Hindu; 3= Others. We need to assign value labels to this variable. To assign value labels, either use the following command or the following steps:

label define religion 1”Muslim” 2”Hindu” 3”Others”

Or simply,

la de religion 1”Muslim” 2”Hindu” 3”Others”

Or, use the following steps:

- Select the variable “religion’ in the “Variables” window (in Data Editor) (Fig 2.2)
- Click on “Value Label” in “Properties” window (under “variables”)
- You will see a dropdown arrow and a small box with 3 dots
- Click on the 3 dot box
- Click on “Create Label”
- Write “religion” in the box “Label name”
- Write 1 in the “Value” box; write “Muslim” in the “Label” box; and click Add
- Write 2 in the “Value” box; write “Hindu” in the “Label” box; and click Add
- Write 3 in the “Value” box; write “Others” in the “Label” box; and click Add
- Click “OK” and then “Close”
- Click on “Value Label” in “Properties” window and using the dropdown arrow, select “religion”

The above steps will insert value labels for the variable "religion". In this way (or using the command), provide value labels to all the coded variables in the data file and enter the data one by one. If there is any missing value, just keep the cell blank (or type the assigned missing value code). Stata will consider it (when the cell is kept blank) a missing value and will indicate it with a dot (.).

Closing the Data Editor window at any time during data entry will keep the data in memory (data will not be lost) unless you exit the Stata program. Once data entry is completed (or partially completed), we need to save the data file (Stata data files have the extension ".dta"). To save a data file, select:

File (from the menu bar) > Save As… > Select the folder where you want to save

the file > Give a file name (.dta will appear by default) > Save

Or,

You can save the data file using the command "**save**"

For example, if you want to save the data file by the name "practice" on your desktop, use the following command (you need to specify the path):

save "C:\Users\HP\Desktop\practice.dta"

When you are in use of a data file for analysis and want to save the data file with the same name, use the command:

save, replace

This will save the data file with the same name and in the same location. The *replace* option in the command replaces the old file with the same name.

So far, we have discussed how to rename and label a variable and how to assign value labels using the Stata Data Editor as well as by using the commands. However, it is convenient to rename and label variables and assign value labels by writing commands in the command window of Stata. The following is the summary of how to use Stata commands for the above functions:

- If you want to change (rename) the variable name "var1" to "idno", use the following command:

  rename var1 idno

- If you want to label the variable "idno" as "serial no.", use the following com mand:

  label var idno "serial no."

- To assign the value labels (1= Muslim, 2= Hindu, 3= Others) to the variable "religion", use the following command:

  label define religion 1"Muslim" 2"Hindu" 3"Others"

After executing the above command (label define), you need to select "religion" for the "Value Label" in the "Properties" window using the dropdown arrow, or simply use the following command:

label values religion religion

- If you want to change (rename) the variable label of a variable (e.g., variable

label of “age” by “age in years”), use the following command:

label var age “age in years”

- If you want to change the value labels, for example, you want to change the value label of religion as 3= Christian (instead of Others), use the following command:

  label define religion 1”Muslim” 2”Hindu” 3”Christian”, replace

- By default, the data file displays the value labels of the variables in the Data Editor (Edit or Browse mode). If you want to see the code numbers (values) of the variables, use the following steps (in data editor mode) or command:

  Tools > Value Labels > Hide All Value Labels
  Or,
  Browse, nolabel

2.1.2 Generating a data file using copy and paste commands

The easiest way to generate a data file in Stata is to copy and paste data from another data file, e.g., from Excel, dBase, SPSS, or others. For example, suppose that you have a data file generated in SPSS, Excel, or other program. You can bring all the data from these files to Stata. To do this, open the "Data Editor" window in Stata (Window > Data Editor). Go to the data file from where you want to copy data into Stata. Select and copy (Control-C) data you want to import and just paste (Control-V) them in the spreadsheet of Stata's "Data Editor". Stata will provide the variable names automatically (by default) as var1, var2, and var3, etc. Rename the variables (var1, var2, var3 and others) with the variable names of your choice. Also provide the variable labels and value labels as discussed in section 2.1.1. Finally, save the data file at your desired location (File > Save as > …). This method is suitable for a small dataset with few variables.

2.1.3 Importing a data file from other programs

Stata version 16 and above has the option to directly import SPSS data into Stata by using the dropdown menu [File > Import > SPSS data (*.sav)]. The lower versions do not have this option. However, the best way to get a SPSS data file for use in Stata is to save the data file into Stata format in SPSS.

For example, if you want to convert the SPSS data file “wealth.sav” into “wealth.dta” in Stata format, first open the data file in SPSS. Then use the following steps:

File > Save as… > Select Stata Version 13 SE (*.dta) for the "Save as type" box using the dropdown arrow > Write "wealth" in the file name box > Save

This will convert and save the SPSS data file "wealth" into Stata format. You can also use a suitable data conversion program to import data from SPSS or other formats into Stata format.

2.1.4 Deleting and inserting variables

2.1.4.1 Deleting a variable

You can delete unwanted variables from a data file. To delete a variable, use the following steps:

- Select a variable in the "Variables" window of Stata
- Right-click the mouse and select "Drop selected variables"

You can also use the command "drop" to delete a variable. For example, if you want to delete the variable "sex" from the data file, use the following command:

drop sex

2.1.4.2 Inserting a new variable

You cannot insert a new variable in the data file without any value or missing value. Therefore, you need to select a value (or missing value) for the variable. For example, you want to insert a new variable "gender", all the values of which will be 1. To insert the new variable (gender), use the following command:

generate gender=1
Or,
gen gender=1

If you want to insert the variable "gender" with the missing values, use the following command:

gen gender=.

The new variable will appear as the last variable in the data file. You also need to provide the variable label and value labels for the new variable as discussed earlier (Section 2.1.1).

2.1.4.3 Copy a variable into the same data file

You may be interested in copying a variable in the same data file that you are using. Suppose that you want to have a copy of the variable "religion" already present in the dataset. If you want this, you need to generate a new variable since the data file cannot take the same variable name as its copy. Let us name the new variable "religion2". Use the following command to get an exact copy (clone) of the variable "religion":

clonevar religion2=religion

You can also use the following command:

gen religion2=religion

The advantages of using the "clonevar" command are that it keeps the variable label and value labels the same, including the missing value code in the same way as they were in the old variable. On the other hand, if the command "gen" is used, it will not keep the variable label and value labels in the new variable. You need to provide them separately.

3

Basics of Stata

In this chapter, we will discuss some of the basic functions related to the use of Stata program, such as Stata files, command syntax, and others. Before we proceed to data analysis, it is important to know all these basic things.

3.1 Stata files

Stata generally involves/generates three types of files. They are:

- **Data file**: The data file contains data for the analysis by Stata and has the extension ".dta";
- **Output file**: The output file contains the results of data analysis and the commands. It is also called a "log" file and has an extension of either ".smcl" or ".log";
- **Command file (Do-file)**: The command file is called a "Do-file" in Stata and has the extension ".do". This file is for storing a collection of commands for data analysis. The commands saved in a Do-file can be reused or executed at any time as required. The commands can be written directly and/or edited in a Do-file for analysis.

3.1.1 Data file

3.1.1.1 Opening (retrieving) an existing data file

In the previous chapter (Chapter 2), we discussed how to generate a data file in Stata. If you already have a Stata data file saved on your computer, you can open it in different

ways, such as using the drop-down menu, writing the command, or using the icon.

Suppose that you have a Stata data file named "Data_3.dta" on your computer and its location is C:\Users\HP\Desktop. If you want to open the data file, use the following steps or command:

File (on the menu bar) > Open > Go to Desktop and select the data file “Data_3” > Open

Or,

use C:\Users\HP\Desktop\Data_3, clear

Or,

Click on the icon > Go to Desktop > Select “Data_3” > Open

Or,

Double click on the data file that you want to open

3.1.2 Output file or log-file

When you analyze data, the outputs (results) are shown in the results window of Stata. The outputs will not be saved automatically. You need to save the outputs in a file. In Stata, the output files are called “log” files. The log-files will have the outputs and commands, but not the graphs. Graphs generated in Stata need to be saved separately, as discussed in Section 7.1.1.

3.1.2.1 Saving outputs in a log-file

A log-file can be saved in two different formats.

- **Stata format (also called *smcl* format):** In Stata format, the log-file will have the extension “.smcl”. The Stata format log-files preserve the formats that we see in the results window and are the defaults.
- **ASCII format:** In ASCII (ordinary text) format, the file will have the extension “.log”.

We generally use the Stata format (smcl format) to save the outputs. However, the “smcl” format can be converted to “ASCII” format whenever necessary.

Suppose that you want to analyze your data and save the outputs in a file named “Results_3”. To save the outputs, create a log-file at the beginning of the analysis (you can also create the log-file in between an analysis) using the menu bar options or command as described below:

File (on the menu bar) > Log > Begin… > Select the location where you want to save the file > Write "Results_3" in the "File name" box > Select the format .smcl (usually the default) > Save
Or,
log using C:\Users\HP\Desktop\Results_3

The above command will open/generate (begin) a log-file with the name "Results_3.smcl" on the desktop.

Once a log-file is opened, you can *temporarily stop* (suspend) saving the outputs at any time during analysis by using the following steps or command:

File > Log > Suspend
Or,
log off

You can restart (resume) saving the outputs at any point in your analysis session by using the following steps or command:

File > Log > Resume
Or,
log on

The log-file is saved and closed automatically at the end of an analysis session when you exit Stata. You can, however, save and close the log-file anytime during analysis by using the following steps or command:

File > Log > Close
Or,
log close

3.1.2.2 Opening an existing log-file

To open an existing log-file already saved on your computer (e.g., you want to open the Results_3.smcl file, which is saved on the desktop), use the following steps:

File > Log > View > Browse > Select the file "Results_3" from Desktop > Open > Ok

If you want to add (append) the results of a new analysis to a previously saved log-file (say, Results_3.smcl), use the following steps or command:

File > Log > Begin > Select the file "Results_3" from Desktop (or from the location where it is saved) > Save > Select "Append to existing file" > Ok

Or,

log using C:\Users\HP\Desktop\Results_3, append

You need to specify the file's path, otherwise the command will not be executed.

If you want to overwrite (replace the contents of a log-file with the outputs of a new analysis) the outputs in a log-file (e.g., Results_3.smcl), use the following command:

log using C:\Users\HP\Desktop\Results_3, replace

This command will delete all the contents of the log-file previously saved and will save the outputs of the new (subsequent) analysis session.

3.1.2.3 Browsing Stata outputs in Results window

You can browse the outputs of analyses in the results window by using the mouse or keyboard buttons (<Shift+Page Up> or <Shift+Arrow button>).

Stata normally (by default) pauses each time the results window is full of information while executing a command unless you press any key on the keyboard. We can ask Stata to continue scrolling (i.e., providing the outputs without pausing) till the completion of an output by using the following command:

set more off

To go back to the pause mode, use the following command:

set more on

3.1.2.4 Copy tables from Stata outputs to MS Word

You can copy a table (or commands or other information) from the Stata outputs (results) window to MS Word. To do this:

Select the table from the output window by dragging down the mouse > Click the right button of the mouse > Select copy > Go to MS Word file where you want to paste the table > Right click the mouse > Paste

Note: Use the Courier New or Consolas font to preserve the table alignments. We have used the Consolas font in this book for better resolution.

3.1.2.5 Transformation of a log-file from smcl to ASCII format

You can convert a Stata log-file from smcl (.smcl) format to ASCII (.log) format (and vice versa). Suppose that you want to transform the log-file "Results_3.smcl" located on your desktop to "Results_3.log". Use the following command or steps:

translate C:\Users\HP\Desktop\Results_3.smcl C:\Users\HP\Desktop\Results_3.log
Or,
File > Log > Translate > Select the file "Results_3.smcl" from Desktop clicking the "Browse" tab of "Input File" > Open > Click the "Browse" tab of "Output File" > Give the file name "Results_3.log" > Save > Translate

3.1.3 Do-file or Stata commands file

3.1.3.1 Generating a Do-file

One can generate a Do-file and write the commands in the Do-file Editor for subsequent use. To open the Do-file Editor (a new Do-file), use the following steps or command:

Window > Do-file Editor > New Do-file Editor
Or,
doedit
Or,
Click on the icon

The above command is for opening a new Do-file. You can save the Do-file like we save a file in the MS word:

File (in Do-file editor) > Save as > Select location and file name > Save

To open a saved Do-file (e.g., to open the Do-file "Test.do"), use either of the following commands. You need to specify the path of the file and file name to open the Do-file, otherwise the command will not work.

doedit Test.do
doedit C:\Users\HP\Desktop\Test.do

This command will open the Do-file "Test.do" saved on the desktop. The alternative way is:

Click on the icon , then

File (in Do-file editor) > Open > File… ctrl+O > Go to the file location and select the file > Open

All the commands must be written in lowercase letters in the Do-file. While writing the commands, Stata considers the end of a command line as the end of that command. If your command exceeds a single line, use three back slashes (///) at the end of the line

before continuing to the next line. Then Stata will consider that the command is continued to the next line.

Once a new Do-file is generated, you can copy the commands from Stata's Review window and paste them into the Do-file.

3.1.3.2 Saving commands in a Do-file

Suppose that you have used some commands to analyze your data. You can see those commands pooled in the Review window. If you want to save the commands in the Review window into a Do-file, use the following steps:

- Select the command(s) you want to copy
- Right click the mouse button
- Select "Send to Do-file Editor" (this will automatically open a Do-file with the selected commands in it)
- Save the file (File > Save as…)

You will see that all the selected commands are in a separate window (Do-file Editor). You can edit the commands in this file and save the file for future execution. Commands can also be copied and pasted into a Do-file from the Review window, Results window, or log-file using the "Copy" and "Paste" options.

3.1.3.3 Executing the commands in a Do-file

It is simple to execute the commands saved in a Do-file. First, open the Do-file in the Do-file Editor (Section 3.1.3.1) containing the commands that you want to execute. To execute a single command or several connected commands at a time written in the Do-file, use the following steps:

Select the command(s) that you want to execute by using the mouse

Tools > Execute (Do), or click on the icon in the Do-file Editor

This will execute the commands selected in the Do-file. If you use the following steps without selecting any command in the Do-file (after opening the Do-file in the Do-file Editor), Stata will execute all the commands stored in the Do-file.

Tools > Execute (do)

To execute all the commands in a Do-file (say, the Do-file "Test.do" saved on the desktop), use the command:

do C:\Users\HP\Desktop\Test.do

This will execute all the commands saved in the Do-file "Test.do" without opening the Do-file in the Do-file Editor.

3.2 Basic command syntax

Most researchers use Stata commands for the analysis of data because of their simplicity and ease of use. A typical form of Stata's command syntax is like:

command [varlist] [if exp] [in] [weight] [, options]
Or,
[prefix:] **command** [varlist] [if exp] [in] [weight] [, options]

command: Indicates Stata's command for the analysis of data. It tells us what Stata is supposed to analyze. Stata commands are case sensitive. All the commands must be written in *lowercase* format, otherwise they will not work.

varlist: "varlist" stands for "variable list". It indicates the list of variables needed for a command to execute. The variable list is optional in many commands. If "varlist" is not specified, the command runs on all the variables in the dataset. For example, if you use the command:

summarize age

Stata will provide the summary statistics of the variable "age". If you use only the command "summarize" without any variable name, Stata will provide the summary statistics of all the variables in the dataset. Instead of writing the command "summarize", you can use only the first three initial letters, such as "sum" to get the summary statistics.

if exp: "if exp" means "if expression". It specifies the conditions to be considered during an analysis. It is optional. For example, if you want to get the summary statistics of age for males only (supposing that males are coded as 1 for the variable "sex"), use the following command:

sum age **if** sex==1

in: "in" indicates the range restrictions in terms of observation numbers. It is optional. For example, if you want to list the first (or last) 10 values of the variable "age" in the dataset, use the following command:

list age in 1/10
list age in -10/-1

The first command will list the first 10 values (1/10 indicates 1 to 10), while the second command will display the last 10 values of the variable "age".

[ ]: All the syntax in [ ] is optional. You may not need to select anything. For example, you can use the following command (without using anything for "if", "in", and "weight") to get the summary statistics for age.

sum age

weight: "weight" indicates the "weight variable". If there is any weight variable (frequency or sampling weight) that you want to include in an analysis, put the variable after "in". For example,

sum age [fweight = v2]

Here, "fweight" indicates frequency weight ("pweight" indicates sampling weight) and "v2" is the weight variable that you want to consider.

options: ", options" indicate an optional instruction for data analysis. Note that there is a comma (,) before the options, which must be used. For example,

sum age, detail

Here, we have used the option "detail". Once we use this option (detail), Stata will give the detailed summary statistics (mean, SD, skewness, kurtosis, percentile, and others) of the variable.

prefix: The "prefix" is not mandatory for an analysis. The prefix is used to get the results by sub-groups, such as by sex, occupation, or other variables. For example, if you want to get the summary statistics of age by sex (i.e., disaggregated by males and females), the prefix is needed. Then the commands would be like:

sort sex
by sex: **sum** age
Or,
bysort sex: sum age
Or,
By sex, **sort: sum** age

Here, “by sex”, “bysort sex”, and “by sex, sort” are the prefixes, while “sum” is the main command to get the summary statistics of age. We have used the command “sort sex” to sort (in ascending order) the variable “sex”. Stata needs sorting (in ascending order) of the prefix variable (in this example, sex) before executing the main command “sum”. That’s why we used the first command, “sort sex”. You can, however, use a single command like “bysort sex” that would first sort the “by variable” (here it is sex) before executing the main command.

3.3 Knowing the dataset

3.3.1 Generating brief description of a dataset

Before data analysis, we need to know about the data, such as the number of observations and variables in the dataset, the nature of the variables, variable labels, and value labels. Once the data file is loaded in Stata, you can get a brief description of the dataset (variables) by using the following command:

describe
Or,
Data > Describe data > Describe data in memory > Ok

The above command (instead of “describe”, you can simply use “des”) will display the characteristics of all the variables in the dataset, since we did not specify any variable in the command (Table 3.1). Table 3.1 shows that there are 210 observations and 17 variables in the dataset. There is also other information. We can, however, get a description of the variables specified with the command, such as:

des age sex occupation religion

This command will display a description of the variables “age”, sex”, “occupation”, and “religion”.

3.3.2 Codebook

Before we start data analysis, it is important to know the variables in the dataset, especially how they are coded and how the missing values are identified. A good practice is to either develop a codebook for the data file (as discussed in Chapter 2) or to look at the data before analysis, so that you understand the structure of the information. This can be done by using the command “codebook”.

Table 3.1 Description of a data file

```
. describe

Contains data from
C:\Users\HP\Documents\Taj_1_desktop_Home\STATA_Book_Taj\Data\Data_3.dta
  obs:           210
 vars:            17
 size:        25,410
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
ID_no           double  %10.0g                Id number
age             double  %10.0g                Age
sex             str1    %1s                   Sex: string
sex_1           double  %10.0g     sex_1      Sex: numeric
religion        double  %10.0g     religion   Religion
religion_2      double  %10.0g     religion_2 Religion 2
occupation      double  %10.0g     occupation Occupation
income          double  %10.0g                Monthly income
sbp             double  %10.0g                Systolic BP
dbp             double  %10.0g                Diastolic BP
f_history       double  %10.0g     f_history  Family history of diabetes
pepticulcer     double  %10.0g     pepticulcer Have peptic ulcer
diabetes        double  %10.0g     diabetes   Have diabetes mellitus
post_test       double  %10.0g                Post test score
pre_test        double  %10.0g                Pre test score
date_ad         long    %dD_m_Y               Date of admission
date_dis        long    %dD_m_Y               Date of discharge
```

codebook

codebook age sex_1 diabetes

codebook age sex_1 diabetes, compact

Or,

label list

label list religion occupation diabetes

The first command listed above (codebook), where the variables are not specified, will provide the codebook of all the variables in the dataset. The second command will provide the codebook of the variables included in the command (age, sex_1, and diabetes) (Table 3.2). The “label list” command considers only the numerically coded variables.

Table 3.2 Outputs of codebook command

```
. codebook age sex_1 diabetes

---------------------------------------------------------------------------
age                                                                     Age
---------------------------------------------------------------------------
                  type:  numeric (double)

                 range:  [6,45]                       units:  1
         unique values:  35                       missing .:  0/210

                  mean:   26.5143
             std. dev:   7.49049

           percentiles:        10%       25%       50%       75%       90%
                              16.5        21        27        32      36.5

---------------------------------------------------------------------------
sex_1                                                         Sex: numeric
---------------------------------------------------------------------------
                  type:  numeric (double)
                 label:  sex_1

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  0/210

            tabulation:  Freq.   Numeric  Label
                           133         0  Female
                            77         1  Male

---------------------------------------------------------------------------
diabetes                                              Have diabetes mellitus
---------------------------------------------------------------------------
                  type:  numeric (double)
                 label:  diabetes

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  4/210

            tabulation:  Freq.   Numeric  Label
                           162         0  No
                            44         1  Yes
                             4         .
```

3.4 Sorting of data

Sorting of data is sometimes required for browsing or editing. Sorting is also needed before executing some commands (Section 3.2). Stata has the option to sort variables both in ascending (default) or descending order. If you want to sort the variable “age” in ascending order (lowest to highest value), use the first of the following commands. If you want to sort it in descending order (highest to lowest value), use the second command:

sort age
gsort –age

3.5 Stata operator symbols

The arithmetic operator symbols and logical expressions are sometimes used in data analysis. Logical expressions are used mostly with the “if” qualifier. Table 3.3 shows the commonly used arithmetic operators and logical symbols. We will see their applications in the subsequent chapters.

3.6 Getting help in Stata

The basic commands for getting help in Stata are "help" and "search". The primary use of the “help” command is to learn about a command or function whose name you already know. For example, if you want to get the help files for the command “generate”, use:

help generate

On the other hand, the primary use of the command "search" is to learn about a subject. It is used, especially when you do not know the command for a function. For example, if you want to know about the ANOVA test and its commands, use:

search anova

This command will provide a search list of several official Stata entries. There are also unofficial websites that include a wealth of commands. The "search" command can locate information outside the official Stata files. Such an important site is the Statistical Software Components (SSC). To get into this site, use:

help ssc

If the “help” command does not find a command or function, it will automatically continue with the “search” command. Say, you are looking for a command that calculates the sensitivity and specificity. Use the keywords with the command “search” to get help, such as:

search sensitivity specificity

For more details about Stata’s help, users may use other resources such as Stata’s YouTube channel (help youtube), Stata list (help statalist), Stata technical support

(search technical support) and others.

Table 3.3 Key operator symbols and logical expressions used in Stata

Operator symbols:	
+	Addition
-	Subtraction
*	Multiplication
/	Division
^	Power
Logical expressions:	
<	Less than
<=	Less than or equal to
==	Equal to
>	Greater than
>=	Greater than or equal to
!=	Not equal to (~= can also be used; ! or ~ indicates not)
&	And
|	Or

4

Data Cleaning and Data Screening

Once data is entered into Stata or imported from other programs, we need to be sure that the data is free from errors. Data cleaning is commonly done by generating frequency distribution tables of all the variables to find the out-of-range values and by cross-tabulations (or by other means) to check the conditional values. If errors are identified, they need to be corrected.

Simultaneously, we need to check the data if it fulfils the assumptions of the desired statistical test (data screening). For example, is the data of a quantitative variable normally distributed to do a t-test? There are other assumptions that need to be checked before using statistical techniques, especially for hypothesis testing. In this chapter, we will primarily discuss data cleaning. *For the time being, users may skip this chapter and proceed to Chapter 5*. Once the users develop some skills in data analysis, they can come back to this chapter. Use the data file <**Data_3.dta**> for practice. The codebook for this data file can be seen in Chapter 6 (Table 6.1).

4.1 Checking for out-of-range errors

We can check the out-of-range errors by generating a frequency distribution table of a variable or by checking the minimum and maximum values. For instance, suppose that you want to check if there are any out-of-range errors in the variable "religion" (the variable "religion" has 3 levels/values: 1= Muslim; 2= Hindu; 3= Others). To check the out-of-range errors, we will find the minimum and maximum values of the variable "religion" in the dataset by using the following command (outputs are shown in Table 4.1):

tabstat religion, stat(min max)

Or,

sum religion

Table 4.1 shows that the value labels of the variable "religion" range from 1 to 3 (minimum 1 and maximum 3), which is within the range of our code numbers. Therefore, there is no out-of-range error in this variable. If the maximum value was more than 3 or the minimum value was less than 1, this would indicate that there were errors in data entry. In such a situation, identify the subjects (by ID no.) and correct the errors by checking the questionnaire. You can also check the same for age and other variables.

Table 4.1 Minimum and maximum values of religion

```
. tabstat religion, stats(min max)

    variable |       min       max
-------------+--------------------
    religion |         1         3
----------------------------------

. sum religion

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    religion |       210     1.52381    .7067039          1          3
```

If there is any value that is outside of the valid range of a variable, we can identify that value against the identification (ID) number (variable name of the identification number in our dataset is "ID_no"). Let us generate a frequency distribution table of "diabetes" to check the values by using the following command:

tab diabetes

This command will produce Table 4.2. The table shows that there is data with a value of "2" for diabetes, which is outside the data range (the valid codes for diabetes are “0” and “1”). Now, to identify the subject (ID_no) who has this value, use the following command (Table 4.2):

list ID_no diabetes if diabetes==2

This command will provide the identification number (ID_no) of the subject for whom the value of diabetes is “2” (Table 4.2). The table shows that the subject with the serial number (ID_no) 9 has a value of “2”. Correct this value in the data file in the Data Editor (Edit) mode and save the file.

Table 4.2 Frequency distribution of diabetes

```
. tab diabetes

       Have |
   diabetes |
   mellitus |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |        164       78.10       78.10
        Yes |         45       21.43       99.52
          2 |          1        0.48      100.00
------------+-----------------------------------
      Total |        210      100.00

. list ID_no diabetes if diabetes==2

     +-----------------+
     | ID_no  diabetes |
     |-----------------|
  9. |     9          2 |
     +-----------------+
```

4.2 Checking for outliers

Outliers are extreme values that deviate from other observations. Outliers may appear in a dataset because of measurement errors, variability in the measurements, errors in data entry, or other reasons. A commonly used rule says that a data point is an outlier if it is more than 1.5×IQR (inter-quartile range) above the 3rd quartile (i.e., Q3 + 1.5×IQR) or below the 1st quartile (i.e., Q1 – 1.5×IQR). There are several ways to identify outliers in a dataset. The most commonly used method is the visualization of data. The visualization methods that can be used are the box-plot chart, scatter plot, and histogram. Statistical methods like Z-score and other methods are also available.

We will check the potential outliers by constructing a box and plot chart. Let us check if there are any outliers in the variable systolic blood pressure (BP) [variable name is "sbp"]. To get the box and plot chart for systolic BP, use the following command (Figs 4.1 and 4.2):

graph box spb
Or,
graph box sbp, marker (1, mlabel (ID_no))

Both these commands will generate box and plot charts for systolic BP (Figs 4.1 and 4.2). Figure 4.1 generated by the first command, shows that there are three dots above the upper whisker without showing the case numbers (ID_no). The second command

has produced Figure 4.2, which shows the case numbers (ID_no) of the dots. These dots indicate that there are 3 potential outliers in systolic BP and their case numbers (ID numbers) are 20, 54, and 193.

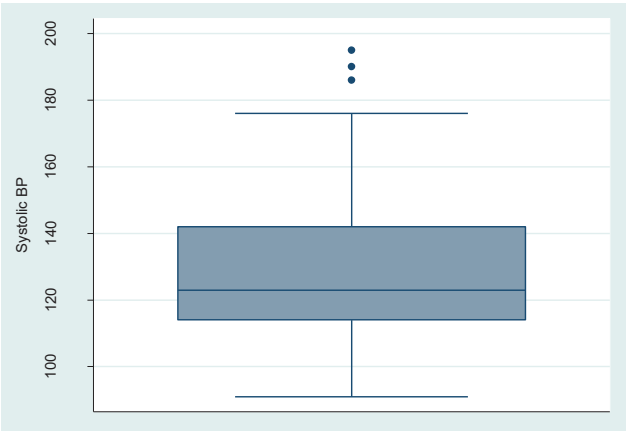


Figure 4.1 Box and plot chart of systolic BP with outliers

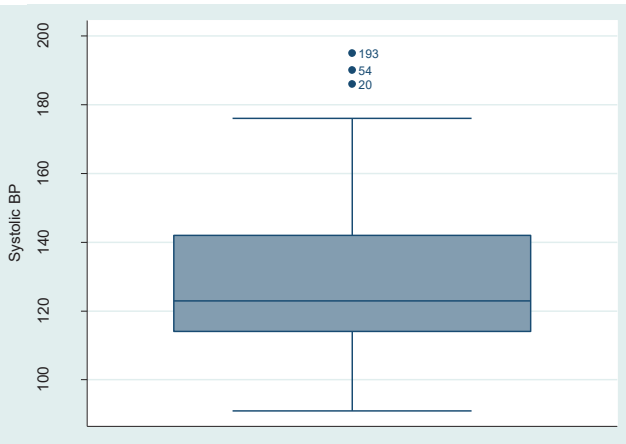


Figure 4.2 Box and plot chart of systolic BP with id no. of outliers

4.3 Assessing normality of data

One of the major assumptions for using parametric tests is that the dependent continuous variable is normally distributed. Whether the data has come from a normally distributed population or not, can be checked in different ways. The commonest methods of checking the normality of a dataset are through:

- Histogram
- Q-Q plot
- Formal statistical tests [Skewness-Kurtosis test, Shapiro Wilk test, or Kolmogorov Smirnov (K-S) test]

This topic is further discussed in Chapter 8.

5

Data Management

Data analysis may require data manipulations, such as making class intervals, classifying a group of people with a specific characteristic using a cut-off value (e.g., you may want to classify people who have hypertension using a cut-off value of either systolic or diastolic BP), and recoding of data for summarization and other purposes. In this chapter, we will discuss data manipulations that are commonly needed during data analysis, like:

- Converting string variables to numeric variables
- Recoding of data
- Making class intervals
- Combine data to generate an additional variable
- Data transformation
- Calculation of total score
- Extraction of duration from dates
- Relocation of variables
- Selection of a subgroup for data analysis
- Transformation of data from wide format to long format

Use the data file <**Data_3.dta**> for practice.

5.1 Converting string variables to numeric variables

Stata data files may have both string and numeric variables. It is always preferable to have numeric variables in the dataset. Numeric variables are easier to manipulate and can be used in various statistical analyses. A string variable may be coded or have

a direct response (e.g., names of districts or provinces). The codes of a string variable may be a character (e.g., m= male; f= female) or a number (e.g., 1= male; 2= female). Even though the codes are made up of numbers, they are actually characters. A detailed overview of how to convert a string variable into a numeric variable is discussed below.

5.1.1 Converting a string variable with non-numeric values into a numeric variable

In our dataset, the variable “sex” is a string variable with non-numeric codes (m= male; f= female). We can convert the string variable “sex” into a numeric variable (say, sex1) by using the command “encode”.

encode sex, **generate**(sex1)

The above command (you can use "gen" instead of "generate") will generate a new numeric variable "sex1" with the codes/values of 1 for female and 2 for male. Stata provides the values in alphabetical order, beginning with 1 (i.e., 1 would be given to the alphabetically first value of the original variable).

5.1.2 Converting a string variable with numeric values into a numeric variable

A string variable may have string numeric values (such as 1, 2, 3, etc.). The following command can be used to change a string variable with numeric string values into a numeric variable, where the numeric values will be retained as numeric.

For example, suppose that the variable "sex" is a string variable with values/codes of 0= female and 1= male. We want to convert "sex" into a numeric variable (say, sex1), retaining the values of the string variable "sex". Use the following command:

gen sex1= real(sex)

This command will generate a new numeric variable “sex1” while keeping the values of the string variable same (0= female; 1= male).

You can also convert a string variable with numeric values without generating a new variable, i.e., into the same variable. To do this, use the following command:

destring sex, replace

This will convert the string variable "sex" into a numeric variable, keeping the name of the variable and its values the same as before.

5.2 Recoding of data

Suppose that you have the string variable “sex” coded as “m” and “f” in the dataset. You want to replace the existing code “m” by 1 and “f” by 2. There are two options for the recoding of data:

a) Recoding into the same variable, and
b) Recoding into a different variable

5.2.1 Recoding a string variable into the same variable

Suppose that we have a string variable “sex” in our dataset, which is coded as m/f. To recode “m” with 1 and “f” with 2 of the string variable into the same variable, use the commands:

replace sex=”1” if sex==”m”
replace sex=”2” if sex==”f”

5.2.2 Recoding a string variable into a different string variable

First, we need to generate a new string variable (say, sex1) before we recode the existing string variable “sex”. Then, we will change the codes. Use the following commands:

gen sex1=” “
replace sex1=”1” if sex==”m”
replace sex1=”2” if sex==”f”

This will generate a new string variable “sex1” with the codes 1= male and 2= female.

5.2.3 Recoding a numeric variable into the same variable

Suppose that you have a numeric variable "diabetes" in your dataset, which is coded as 1= don’t have diabetes and 2= have diabetes. You want to replace the codes "1" with "0", and "2" with "1". Use the following command:

recode diabetes (1=0) (2=1)

5.2.4 Recoding a numeric variable into a different variable

In this case, first, we will generate a new variable that is identical to the original variable using the command “generate or gen”. Then we will recode the values of the

new variable using the command “recode”. Suppose that you have the numeric variable “diabetes”, which is coded as “1= no diabetes” and “2= have diabetes”. You want to recode the variable into a new variable “diabetes1”, where “1” will be coded as “0 (no diabetes)” and “2” will be coded as “1 (have diabetes)”. Use the following commands:

gen diabetes1=diabetes
recode diabetes1 (1=0) (2=1)
Or,
gen diabetes1=0
replace diabetes1=1 if diabetes==2

The first command (gen) will generate a new variable “diabetes1” equivalent to the original variable “diabetes” (i.e., the values of diabetes1 and diabetes will be the same). The second command will change the values of the new variable “diabetes1” from “1” to “0” and from “2” to “1”.

The alternative approach is to generate a new variable “diabetes1” with all the values equal to “0” by using the command “gen” (third command). Then replace the value “0” with “1” if the value of the variable “diabetes” is “2” by using the command “replace” (fourth command).

You can check the coding scheme of the new variable by using the commands “tab” or “codebook”, such as:

tab diabetes1
codebook diabetes1

Note that the new variable will always appear as the last variable in the dataset that you can check in browser mode. When you change the codes, you need to give the variable label (if coded into a new variable) and value labels of the new codes by using the following commands (also discussed in Chapter 2):

label var diabetes1 "have diabetes"
la de diabetes1 0”no diabetes” 1”have diabetes”
label values diabetes1 diabetes1

The outputs of the above commands are displayed in Table 5.1. Other examples of using the “recode” command are provided in Table 5.2.

Table 5.1 Recoding of diabetes into a new variable

```
. gen diabetes1=diabetes
. tab diabetes1

  diabetes1 |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        165       78.57       78.57
          2 |         45       21.43      100.00
------------+-----------------------------------
      Total |        210      100.00

. recode diabetes1 (1=0) (2=1)
(diabetes1: 210 changes made)

. tab diabetes1

  diabetes1 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        165       78.57       78.57
          1 |         45       21.43      100.00
------------+-----------------------------------
      Total |        210      100.00

. label var diabetes1 "have diabetes"

. la de diabetes1 0"no diabetes" 1"have diabetes"

. label values diabetes1 diabetes1

. tab diabetes1

have diabetes |      Freq.     Percent        Cum.
--------------+-----------------------------------
  no diabetes |        165       78.57       78.57
have diabetes |         45       21.43      100.00
--------------+-----------------------------------
        Total |        210      100.00
```

5.3 Making class intervals

The central tendency (such as mean and median) and dispersion (such as SD) of quantitative data provide meaningful information. Further useful summarization may be achieved by grouping the data into class intervals or categories. Suppose that you want to categorize the variable "age" into the following categories or class intervals:

- ≤20 years (to be coded as 1),
- 21-30 years (to be coded as 2),
- 31-40 years (to be coded as 3), and
- >40 years (to be coded as 4)

For this exercise, we will recode the variable "age" into a new variable "age2". We recommend that users always recode into a new variable. If you recode into the same variable, the original data will be lost, and it cannot be recovered once the data file is saved. To recode into a different variable (age2), use the following commands:

gen age2=age
recode age2 (0/20=1) (21/30=2) (31/40=3) (*=4)
label var age2 "age group"
label define age2 1"<=20 yrs" 2"21-30 yrs" 3"31-40 yrs" 4">40 yrs"
label values age2 age2
tab age2

The above commands will generate a new variable "age2" with four categories of age as mentioned above (Table 5.3). The "*" in (*=4) indicates all other values.

Table 5.2 Some examples of the use of recode command

Command	Outputs
recode var1 (0=1)	Will change values of the variable “var1”, 0 to 1
recode var1 (1=0) (2=1)	Will change values of the variable “var1”, 1 to 0 and 2 to 1
recode var1 (0=1) (*=2)	Will change values of the variable “var1”, 0 to 1 and all other values to 2
recode var1 2/4=2	Will change values of the variable “var1”, (2 to 4) to 2
recode var1 (1 3 5 = 1)	Will change values 1, 3 & 5 of the variable “var1” to 1
recode var1 (.=9)	Will change the missing values to 9
recode var1 (9=.)	Will change the value “9” as missing value

Using the command "recode", you can also classify people who have hypertension and those who do not have hypertension (for example). To do this, you need to use a cut-off value to define hypertension. For example, we have collected data on diastolic BP (variable name is "dbp"). We want to classify those as "hypertensive", if the diastolic BP is >90 mmHg. Now, generate a new variable (say, htn) equivalent to the variable "dbp". Recode the variable "htn" as ≤90= 0 (normal BP) and >90=1 (hypertensive). We hope you can do it now. If you cannot, use the following commands:

gen htn=dbp
recode htn (0/90=0) (*=1)

Table 5.3 Age categories

```
. gen age2=age
. recode age2 (0/20=1) (21/30=2) (31/40=3) (*=4)
. label var age2"age group"
. label define age2 1"<=20 yrs" 2"21-30 yrs" 3"31-40 yrs" 4">40 yrs"
. label values age2 age2
. tab age2

  age group |      Freq.     Percent        Cum.
------------+-----------------------------------
   <=20 yrs |         50       23.81       23.81
  21-30 yrs |         97       46.19       70.00
  31-40 yrs |         58       27.62       97.62
    >40 yrs |          5        2.38      100.00
------------+-----------------------------------
      Total |        210      100.00
```

The above two commands will generate a new variable "htn" with values/codes "0" and "1" (the last variable in the data file). The code "0" indicates people who do not have hypertension (diastolic BP ≤ 90) and "1" indicates people who have hypertension (diastolic BP >90).

To give the variable label and value labels, use the following commands:

label var htn "hypertension"
la de htn 0"no hypertension" 1"have hypertension"
label values htn htn

To check whether the commands have been executed properly, make a frequency distribution table of the new variable "htn". This will also provide you with the proportion of subjects with diastolic hypertension. Use the following command to get the frequency distribution table for “htn” (Table 5.4):

tab htn

5.4 Combining data into a new variable

Sometimes, the cut-off point of a measurement (e.g., hemoglobin and blood pressure) for defining a condition (e.g., anemia and hypertension) may vary according to gender or other characteristics. In such a situation, a single cut-off point for defining a condition may not be appropriate.

For example, we have collected data on diastolic BP (variable name is “dbp”) both for

Table 5.4 Frequency distribution of hypertension

```
. gen htn=dbp
. recode htn (0/90=0) (*=1)
(htn: 210 changes made)

. lab var htn "hypertension"
. la de htn 0"no hypertension" 1"have hypertension"
. la values htn htn
. tab htn

     hypertension |      Freq.     Percent        Cum.
------------------+-----------------------------------
  no hypertension |        162       77.14       77.14
have hypertension |         48       22.86      100.00
------------------+-----------------------------------
            Total |        210      100.00
```

males and females (variable name is “sex_1”). We have defined hypertension as a diastolic BP greater than 85 mmHg if it is a female, and a diastolic BP greater than 90 mmHg if it is a male. Now, how to classify those who have hypertension considering gender and cut-off values?

To do this, first we will generate a new variable, say “htn1”, for which all the values will be “0” (zero) by using the command “gen”.

gen htn1=0

This will generate a new variable "htn1" with all the values of “0”. Now, use the following commands to recode the new variable:

replace htn1=1 if dbp>85 & sex_1==0
replace htn1=1 if dbp>90 & sex_1==1

The above commands will replace some of the values of the variable "htn1" from “0” to “1” based on the conditions provided in the commands. In this example, code “1” indicates individuals with hypertension. Note that the variable "sex_1" used in the commands is a numeric variable with codes 0= female and 1= male.

Provide the variable label and value labels as done before and make a frequency distribution table for the variable “htn1” to check whether the commands have worked properly. You will also get the proportion of subjects with hypertension from the table (Table 5.5). Data shows that the prevalence of diastolic hypertension is 30.0% in the sample.

Table 5.5 Data combined into a new variable

```
. gen htn1=0

. replace htn1=1 if dbp>85 & sex_1==0
(55 real changes made)

. replace htn1=1 if dbp>90 & sex_1==1
(8 real changes made)

. lab var htn1 "hypertension"

. la de htn1 0"no hypertension" 1"have hypertension"

. la values htn1 htn1

. tab htn1

     hypertension |      Freq.     Percent        Cum.
------------------+-----------------------------------
  no hypertension |        147       70.00       70.00
have hypertension |         63       30.00      100.00
------------------+-----------------------------------
            Total |        210      100.00
```

5.5 Data transformation

In many situations, the data that has been collected on a quantitative variable for a study is not normally distributed. Since the parametric methods (in general) for testing hypotheses are more effecient than the nonparametric methods, data transformations are occasionally needed to make the distribution normal and meet the assumptions for a parametric test. Depending on the shape of the data distribution, there are several options for data transformation. The following table (Table 5.6) shows some of the options for data transformation.

The commonly used method of data transformation is log transformation. Let us see how to transform data into a log value. Suppose that you want to transform diastolic BP (variable name is “dbp”) into a log of diastolic BP. Use the command "gen", as shown in Table 5.7 (the first and second commands), to transform the data into a new variable "dbp1" with the log values of diastolic BP. The table also shows the commands for other data transformation options.

Table 5.6 Data transformation options

Method	Good for	Bad for
Log	Right skewed data	Zero values and negative values
Square root	Right skewed data	Negative values
Square	Left skewed data	Negative values
Reciprocal	Making small values bigger and big values smaller	Zero values and negative values

Table 5.7 Data transformation commands

Command	Function
	Will generate a new variable:
gen dbp1=log(dbp)	"dbp1" with the values of "log of dbp"
gen dbp2=ln(dbp)	"dbp2" with the values of "natural log of dbp"
gen dbp3=dbp^2	"dbp3" with the values of "square of dbp"
gen dbp4=sqrt(dbp)	"dbp4" with the values of "square root of dbp"
gen dbp5=(1/dbp)	"dbp5" with the values of "reciprocal of dbp"

5.6 Calculation of total score

A study was conducted by a researcher to assess the knowledge of secondary school students about how HIV is transmitted. To assess their knowledge, the researcher collected data on the following questions (Table 5.8; data file: **Data_HIV.dta**).

When data is coded like 1= correct and 2= incorrect, first recode the data of the knowledge variables as 2 = 0 (i.e., 0 indicates incorrect). In such a coding scheme (0/1), the total score would range from 0–4, since there are four knowledge questions. Then, calculate the total score by generating a new variable "tknow". Use the following commands to do this:

recode k1-k4 (2=0)
gen tknow=(k1+k2+k3+k4)
Or,
egen tknow= rowtotal(k1 k2 k3 k4)

Table 5.8 Questions for assessing the knowledge on HIV transmission

HIV is transmitted through:	**Codes**	
1. Sexual contact (variable name: **k1**)	1. Yes	2. No
2. Transfusion of unscreened blood (variable name: **k2**)	1. Yes	2. No
3. Sharing of injection needle (variable name: **k3**)	1. Yes	2. No
4. Accidental needle stick injury (variable name: **k4**)	1. Yes	2. No

The first command will recode the knowledge variables k1 to k4 from “2” to “0”, while the second (or third) command will calculate the total correct answers in the new variable “tknow” (total knowledge on HIV). You can check the data by generating the descriptive statistics and a frequency distribution table of the variable “tknow” by using the following commands (Table 5.9):

sum tknow
tab tknow

Table 5.9 displays the descriptive statistics (mean, standard deviation, and others) and frequency distribution of the students' overall knowledge on HIV transmission. The table shows that the mean of total knowledge is 2.18 (SD 0.63), while the minimum value is 0 and the maximum value is 4. The table also indicates that there are 2 (1.0%) students who do not have any knowledge on HIV transmission (since the score is 0, i.e., they could not answer any questions correctly). One hundred and twenty-five (63.8%) students know two ways of HIV transmission, while only 1.5% of the students know all the ways of HIV transmission. You can also classify the students as having "good" or "poor" knowledge by using a cut-off value based on the total knowledge score.

5.7 Extraction of duration from dates

Stata can extract the difference between two dates. Suppose that you have a dataset with variables for the date of admission (variable name is date_ad) and date of discharge (date_dis) of patients admitted to a hospital (data file Data_3.dta). If you want to calculate the duration of hospital stay (date of discharge minus date of admission), use the following command to extract the difference in days:

gen dura = date_dis – date_ad

Table 5.9 Total knowledge on HIV

```
. sum tknow

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       tknow |       196    2.183673     .638057          0          4

. tab tknow

      tknow |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |          2        1.02        1.02
          1 |         16        8.16        9.18
          2 |        125       63.78       72.96
          3 |         50       25.51       98.47
          4 |          3        1.53      100.00
------------+-----------------------------------
      Total |        196      100.00
```

This command will generate a new variable “dura” (the last variable in the dataset) that contains the duration of hospital stay in days. You can check the data by getting the summary statistics of the variable “dura” using the command “**sum**” (Table 5.10).

sum dura, detail

5.8 Relocating variables in the dataset

You can order the variables according to alphabetical order in the variables window. Suppose that you want to arrange the variables according to ascending order. Use the following command:

order _all, first alphabetic

You can also move a variable to your desired position. In Stata, when a new variable is generated, it appears (by default) as the last variable in the dataset. Suppose that you have generated a new variable "age1" and you want it to be after the existing variable "age". Use the following command:

order age1, after(age)

If you want to send a variable (say, sex) at the end of the dataset, use the first command below. If you want it to be the first variable, use the second command:

order sex, last
order sex, first

Table 5.10 Summary statistics of duration of hospital stay

```
. sum dura, detail
                              dura
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1              1
 5%            3              3
10%            3              3       Obs                  21
25%            4              4       Sum of Wgt.          21

50%            6                      Mean           5.761905
                        Largest       Std. Dev.      2.343177
75%            7              8
90%            8              8       Variance       5.490476
95%            9              9       Skewness       .1741077
99%           11             11       Kurtosis       2.813007
```

5.9 Selecting a sub-group for analysis

You may have a large dataset with many variables. But, you may not need all the data or all the variables for your research. Stata has the option to select a small dataset (a subset) for analysis. Remember that the bigger the dataset, the harder it is for Stata to manage the data. It is, therefore, good to have a dataset as small as possible while keeping all the relevant information.

The data file <**Data_3.dta**> has several variables, including the variable "diabetes1". In the dataset, the variable "diabetes1" is coded as "1= Yes (have diabetes)" and "0= No (do not have diabetes)". Suppose that you want to analyze the data only for those who are more than 30 years old and have diabetes. We can select the data with this criteria using the command either "drop" or "keep". Make sure that you have saved your original dataset before using these commands. To select the group who are more than 30 years old and have diabetes (i.e., diabetes1= 1), use the following commands:

drop if diabetes1==0
drop if age<=30
Or,
drop if diabetes1==0 | age<=30

The first two commands or the third command (alone) will exclude subjects who do not have diabetes and are aged ≤30 years from the subsequent analyses. Instead of the command "drop", you can use the command "keep" with the same effects as follows:

keep if diabetes1==1
keep if age>30
Or,
keep if diabetes1==1 & age>30

This will limit your analyses to subjects with diabetes who are over the age of 30. You can check the data by making a frequency distribution table (using the command "tab diabetes1") of "diabetes1" and the summary statistics (using the command "sum age, detail") of "age".

To delete a variable from the dataset, use the command “drop”. For example, if you want to delete the variables “income” and “sex”, use the following command:

drop income sex

Note that to have analyses for all the data after you use the commands “drop” or “keep”, you need to read the data file again, either by using the command “use” or by using the drop-down menu.

You can also select a subset of data or variables for analysis while opening the data file by using the following commands:

use C:\Users\HP\Desktop\Data_3.dta if diabetes1==1 & age<=30
use age sex diabetes1 **using** C:\Users\HP\Desktop\Data_3.dta

The first command will select a subset of subjects who are less than or equal to 30 years old and have diabetes. The second command will select only the variables “age”, “sex”, and “diabetes1” while opening the data file.

5.10 The “egen” command

The command "egen" (which indicates extended generate) is an extended version of the command "generate" to generate new variables. The "egen" command is used to generate new variables that require some additional functions, such as mean, median, z-score, or others. For example, if you want to generate a new variable (say, zage) with the z-scores of age, use the following command:

egen zage=std(age)

Other examples of the use of the "egen" command are:

egen v5= rowmean(v1 v2 v3 v4)
egen v10= rowtotal(v1 v2 v3 v4)

The first command will generate a new variable "v5" that will have the row-mean of the variables v1 to v4. The second command will generate a new variable "v10", which will have the row-total of the variables v1 to v4.

5.11 Creating dummy variables

A dummy variable is a dichotomous variable that takes values of “0” and “1”, where the values indicate the presence or absence of a characteristic (e.g., “0” may indicate a non-Muslim and “1” may indicate a Muslim). Dummy variables are also known as indicator variables. Where a categorical variable has more than two categories, it can be represented by a set of dummy variables, with one variable for each category.

Sometimes it is necessary to use dummy variables during analysis, e.g., in multiple linear regression analysis (Section 16.2). In Stata, you can generate the dummy variables easily. Suppose that you want to generate dummy variables for religion. The variable "religion" has three levels (categories), 1= Muslim, 2= Hindu, and 3= Christian. To generate dummy variables for religion, use the following command:

tab religion, **gen**(reli)

This command will generate three dummy variables – “reli1”, “reli2”, and “reli3” in the data file. All these variables will be coded as 0/1, where for “reli1”, 1 is Muslim and 0 is Others (Hindu and Christian); for “reli2”, 1 is Hindu and 0 is Others (Muslim and Christian); and for “reli3”, 1 is Christian and 0 is Others (Muslim and Hindu) (Table 5.11). Provide the variable and value labels to the dummy variables as discussed earlier.

You can check the dummy variables by making frequency distribution tables using the following command (Table 5.11):

tab1 reli1 reli2 reli3

5.12 Transformation of data formats

Repeated measures data may come in two different formats – wide format or long format. For example, we have measured the blood sugar levels of five study subjects at three time points, such as at the baseline and at 7 and 14 hours after giving a drug. The variables generated for the blood sugar levels at each time point are "time0", "time7", and "time14", respectively. If the data (blood sugar levels) of the individuals are plotted under each time variable, as shown in Table 5.12, it is called a wide data format.

Table 5.11 Dummy variables of religion

```
. tab religion, gen(reli)

   Religion |      Freq.     Percent        Cum.
------------+-----------------------------------
     MUSLIM |        126       60.00       60.00
      HINDU |         58       27.62       87.62
  Christian |         26       12.38      100.00
------------+-----------------------------------
      Total |        210      100.00

. tab1 reli1 reli2 reli3

-> tabulation of reli1

religion==M |
      USLIM |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         84       40.00       40.00
          1 |        126       60.00      100.00
------------+-----------------------------------
      Total |        210      100.00

-> tabulation of reli2

religion==H |
       INDU |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        152       72.38       72.38
          1 |         58       27.62      100.00
------------+-----------------------------------
      Total |        210      100.00

-> tabulation of reli3

religion==C |
   hristian |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        184       87.62       87.62
          1 |         26       12.38      100.00
------------+-----------------------------------
      Total |        210      100.00
```

Here, the data for each subject is plotted once under each time variable.

In long format, there is a time variable (say, time) with three or more levels/ categories (in our example, we need three levels, such as 0= time0, 7= time7, and 14= time14) and there is a separate variable for blood sugar level (say, sugar), as shown in Table 5.13. In this format, the blood sugar levels of the study subjects are plotted under the variable "sugar" against the categories of the time variable.

Table 5.12 Wide data format

Subject	Time0	Time7	Time14
1	110	108	107
2	115	112	110
3	112	110	115
4	106	105	104
5	109	108	107

Table 5.13 Long data format

Subject	Time	Sugar
1	0	110
2	0	115
3	0	112
4	0	106
5	0	109
1	7	108
2	7	112
3	7	110
4	7	105
5	7	108
1	14	107
2	14	110
3	14	115
4	14	104
5	14	107

Occasionally, it may be necessary to transform the data from a wide format to a long format. For example, Stata cannot use wide data format for the analysis of repeated measures ANOVA. We need to change the wide data format to a long format, which can be done by Stata. For this exercise, let us use the data file <**Data_repeat_2.dta**>.

In the data file, there are seven variables — “sl (serial number)”, “subject (study subject)”, “treatment (treatment group)”, “sugar0 (baseline blood sugar level)”, “sugar7 (blood sugar level at 7 hours after treatment)”, “sugar14 (blood sugar level at 14 hours after treatment)”, and “sugar24 (blood sugar level at 24 hours after treatment)”. This data file is in wide format for the data of blood sugar levels. We can transform this data file into a long data format (for blood sugar) by using the following command:

reshape long sugar, i(subject) j(time)

This command will provide Table 5.14. The table shows that with this command, Stata has generated a new variable "sugar", which contains the blood sugar levels of all the subjects at different time points. The Stata also generated another new variable "time" with four levels – 0 (baseline blood sugar level), 7 (blood sugar level at 7 hours after treatment), 14 (blood sugar level at 14 hours after treatment) and 24 (blood sugar level at 24 hours after treatment) [You can check it by using the command “tab time”]. Since there are 15 study subjects and blood sugar levels are measured four times on each

subject, the total number of observations is 60. The number of variables in the dataset has also been reduced from 7 to 5.

Table 5.14 Transformation of data from wide to long format

```
. reshape long sugar, i(subject) j(time)
(note: j = 0 7 14 24)

Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                       15   ->      60
Number of variables                   7   ->       5
j variable (4 values)                     ->   time
xij variables:
              sugar0 sugar7 ... sugar24   ->   sugar
-----------------------------------------------------------------------------
```

After the transformation of the dataset, provide the variable label and value labels of the new variables "time" and "sugar" by using the following commands:

la var time "time of blood sugar measurement"
la de time 0"baseline" 7"sugar at 7 hrs" 14"sugar at 14 hrs" 24"sugar at 24 hrs"
la values time time
la var sugar "blood sugar level"

You can also transform a long data format into a wide data format by using the following command:

reshape wide sugar, i(subject) j(time)

6

Data Analysis: Descriptive Statistics

Descriptive statistics are always used at the beginning of data analysis. The objectives of using descriptive statistics are to organize and summarize data. Commonly used descriptive statistics are frequency distribution, measures of central tendency (mean, median, and mode), and measures of dispersion (range, standard deviation, and variance). Measures of central tendency convey information about the average value of a dataset, while the measures of dispersion provide information about the amount of variation present in the dataset. Other descriptive statistics that are used during data analysis include quartile and percentile. In this chapter, we will discuss how to analyze data for descriptive statistics. Use the data file <**Data_3.dta**> for practice. The codebook for the dataset is given in Table 6.1.

6.1 Frequency distribution

Frequency distribution of variables, especially for categorical variables, is commonly done during data analysis. The command for the frequency distribution table is "tabulate" or simply "tab". If you want to find the frequency distribution of the variable "sex", use the following command:

tabulate sex
Or,
tab sex

This command will give you the frequency distribution table for sex (Table 6.2). If you want the frequency distribution tables of two or more variables by a single command, the command “tab” will not produce separate tables for the variables. The use of two

variables after the command "tab" will produce a cross-table for the variables included in the command. For example, if you use the variables sex and religion with the

Table 6.1 Codebook for the data file "Data_3.dta"

Stata variable name	Actual variable name	Variable code
ID_no	Identification number	Actual value
age	Age in years	Actual value
sex	Sex: string	m= Male f= Female
sex_1	Sex: numeric	0= Female 1= Male
religion	Religion	1= Islam 2= Hindu 3= Others
religion_2	Religion 2	1= Islam 2= Hindu 3= Christian 4= Buddha
occupation	Occupation	1= Government job 2= Private job 3= Business 4= Others
income	Monthly family income in Tk.	Actual value
sbp	Systolic blood pressure in mmHg	Actual value
dbp	Diastolic blood pressure in mmHg	Actual value
f_history	Family history of diabetes	1= Yes 2= No
peptic	Have peptic ulcer	1= Yes 2= No
diabetes	Have diabetes mellitus	1= No 2= Yes
diabetes1	Have diabetes mellitus	0= No 1= Yes
posttest	Post-test score	Actual value
pretest	Pre-test score	Actual value
datead	Date of hospital admission	Actual date
datedis	Date of discharge	Actual date

command “tab” (i.e., tab sex religion), Stata will produce a cross-table of sex by religion. However, if you want to get the frequency distribution of several variables at a time (e.g., sex and religion), use the command “tab1”. To get the frequency distributions of sex and religion, use the following command:

tab1 sex religion

This command will produce two separate tables, one each for sex and religion, as shown in Table 6.2. The table shows that there are a total of 210 subjects in the dataset, out of which 133 (63.33%) are female and 77 (36.67%) are male. The frequency distribution of religion is provided after the frequency distribution table of sex.

If there is any missing data, the output of the analysis will not show it (by default). If there are missing values in the data, the total number of subjects will be less than the number of data collected. For example, if we look at Table 6.2 (frequency distribution of religion), there are 210 subjects, while Table 6.3 shows that there are 205 subjects in the frequency distribution table of sex. This indicates that there are five missing values in the data for "sex". To get the frequency distribution with missing cases (Table 6.3), use the following command:

tab sex, miss

To get the identification numbers (variable name is "ID_no") of the missing subjects (for example, for sex), use the following command.

list ID_no if missing(sex)

This will provide the id numbers of the missing cases for the variable “sex” (Table 6.3).

6.2 Central tendency and dispersion

We calculate the central tendency and dispersion for the quantitative variables. Suppose that you want to find the mean, standard deviation (SD), minimum, and maximum values of the variable “age”. The command to get these measures is “summarize” or simply “sum”. Use the following command:

sum age

The above command will provide the mean, SD, minimum (min) and maximum (max) values of the variable "age" (Table 6.4). However, to get more detailed information, use the following command:

sum age, detail

Table 6.2 Frequency distribution of sex and religion

```
. tab1 sex religion

-> tabulation of sex

Sex: string |      Freq.     Percent        Cum.
------------+-----------------------------------
          f |        133       63.33       63.33
          m |         77       36.67      100.00
------------+-----------------------------------
      Total |        210      100.00

-> tabulation of religion

   Religion |      Freq.     Percent        Cum.
------------+-----------------------------------
     MUSLIM |        126       60.00       60.00
      HINDU |         58       27.62       87.62
  Christian |         26       12.38      100.00
------------+-----------------------------------
      Total |        210      100.00
```

Table 6.3 Frequency distribution of sex with missing cases and their id numbers

```
. tab sex

Sex: string |      Freq.     Percent        Cum.
------------+-----------------------------------
          f |        130       63.41       63.41
          m |         75       36.59      100.00
------------+-----------------------------------
      Total |        205      100.00

. tab sex, miss

Sex: string |      Freq.     Percent        Cum.
------------+-----------------------------------
            |          5        2.38        2.38
          f |        130       61.90       64.29
          m |         75       35.71      100.00
------------+-----------------------------------
      Total |        210      100.00

. list ID_no if missing(sex)

     +------+
     | ID_no |
     |------|
  6. |    6 |
  9. |    9 |
 15. |   15 |
 20. |   20 |
 26. |   26 |
     +------+
```

This command will provide the percentiles, mean, SD, variance, skewness, kurtosis, and four lower and four upper values of the variable “age” (Table 6.4).

You can also get the specific statistics that you may want to have for a variable. For example, if you want to have the mean, median, SD, and 10th percentile of age (you can also use multiple variables), use the following command:

tabstat age, stat(mean median sd p10)
tabstat age sbp dbp, stat(n mean sd median p10) col(stat) long

You can also get all those statistics at each level of another categorical variable (e.g., sex). Use the following commands (Table 6.5):

tabstat age, stat(n mean sd median) by(sex)
tabstat age sbp, stat(n mean sd p50) by(sex) long format
tabstat age sbp, stat(n mean sd p50) by(sex) long col(stat)
tabstat age sbp, stat(n mean sd q) by(sex) long nototal

Sometimes we need to know the confidence interval (CI) for the mean. For instance, if you want to get the 95% CI for the mean of age, use the following command:

ci age
ci age, level(99)

The first command will provide the 95% CI (default), while the second command will provide the 99% CI (Table 6.4) for the mean age. You can use multiple variables with the command to get the CIs. For example, if you want to get the 95% CIs for the means of age, income, and systolic BP (variable name is sbp), use the following command (Table 6.4).

ci age income sbp

When a variable is coded as 0/1, *Stata considers it a dichotomous categorical variable*. For example, the variable "diabetes1" in the dataset is coded as 0/1 (Table 6.1), where code "0" indicates the subjects who do not have diabetes and code "1" indicates the subjects who have diabetes. With this coding scheme of a variable, if we use the command "sum", the mean actually indicates the proportion of the subjects coded as “1”, i.e., the prevalence of diabetes in the sampled population when it is a cross-sectional data (Table 6.6). We can also get the prevalence (proportion) of diabetes using the command “tab”.

sum diabetes1
tab diabetes1

Table 6.4 Summary statistics of age including the confidence intervals

```
. sum age

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |       210    26.51429    7.490491          6         45

. sum age, detail
                              Age
-------------------------------------------------------------
      Percentiles      Smallest
 1%           10              6
 5%           14              6
10%         16.5             10       Obs                 210
25%           21             11       Sum of Wgt.         210

50%           27                      Mean           26.51429
                        Largest       Std. Dev.      7.490491
75%           32             41
90%         36.5             43       Variance       56.10745
95%           38             43       Skewness      -.0917613
99%           43             45       Kurtosis       2.690733

. ci age

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
         age |        210    26.51429     .516893        25.49529    27.53328

. ci age, level(99)

    Variable |        Obs        Mean    Std. Err.       [99% Conf. Interval]
-------------+---------------------------------------------------------------
         age |        210    26.51429     .516893        25.17059    27.85798

. ci age income sbp

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
         age |        210    26.51429     .516893        25.49529    27.53328
      income |        210    85194.49    1223.074        82783.34    87605.63
         sbp |        210    127.7333    1.384129        125.0047     130.462
```

The first command (sum) will provide the mean, while the command "tab" will provide a frequency distribution table of the variable (Table 6.6). To get the CI for a proportion, use the command “proportion”. For example, to get the 95% CI for the prevalence of diabetes, use the following command (you can use multiple variables together with this command) (Table 6.6):

proportion diabetes1

Table 6.6 shows the 95% CI for the proportion (%) of people who have diabetes, which

is 0.163 (16.3%) to 0.275 (27.5%). The 95% CI indicates we are 95% confident that the prevalence of diabetes in the population is between 16.3% and 27.5%. You can use other options with this command, such as:

proportion diabetes1, level(99)
proportion diabetes1, over(religion) level(99)

The second command will display the 99% CI of diabetes in different religious groups.

Table 6.5 Outputs of "tabstat" command

```
. tabstat age sbp dbp, stat(n mean sd median p5) col(stat)

    variable |         N      mean        sd       p50        p5
-------------+--------------------------------------------------
         age |       210  26.51429  7.490491        27        14
         sbp |       210  127.7333  20.05794       123       101
         dbp |       210  82.76667  11.74929        82        66

. tabstat age sbp, stat(n mean sd p50) by(sex) long col(stat)

sex        variable |         N      mean        sd       p50
--------------------+--------------------------------------
f               age |       133  26.88722  6.802021        27
                sbp |       133  129.5714  21.37695       124
--------------------+--------------------------------------
m               age |        77  25.87013  8.559929        26
                sbp |        77  124.5584  17.22108       122
--------------------+--------------------------------------
Total           age |       210  26.51429  7.490491        27
                sbp |       210  127.7333  20.05794       123
------------------------------------------------------------
```

6.2.1 Interpretation

In Table 6.4, we can see all the descriptive statistics (central tendency and dispersion) of the variable "age", including the statistics for Skewness and Kurtosis.

We believe you understand the meanings of mean (average), SD (average difference of individual observations from the mean) and variance (square of SD). The analysis did not give the median directly. As indicated in Table 6.4, the 50th percentile (P_{50}) is the median (middle value of the data set).

As presented in Table 6.4, the mean age is 26.5 years and the SD is 7.49 years. Let us discuss the other statistics provided in Table 6.4, especially the skewness, kurtosis, percentile, and quartile (not provided directly in the analysis).

Table 6.6 Prevalence and confidence interval (CI) of diabetes

```
. sum diabetes1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    diabetes |       210    .2142857    .4113064          0          1

. tab diabetes1

       Have |
   diabetes |
   mellitus |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |       165       78.57       78.57
        Yes |        45       21.43      100.00
------------+-----------------------------------
      Total |       210      100.00

. proportion diabetes1

Proportion estimation                Number of obs    =     210

--------------------------------------------------------------
             | Proportion   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
diabetes1    |
          no |   .7857143   .0283828       .724512    .8363903
         yes |   .2142857   .0283828      .1636097     .275488
--------------------------------------------------------------
```

Skewness and Kurtosis: These two statistics are used to evaluate whether the data has come from a normally distributed population or not. In Table 6.4, we can see the statistics for skewness (-0.091) and kurtosis (2.69). Skewness indicates the spreadness of distribution (a measure of symmetry). Skewness "greater than 0" indicates data is skewed to the right; skewness "less than 0" indicates data is skewed to the left, while skewness "near to 0" indicates data is symmetrical (normally distributed).

However, the normality of data should not be evaluated based on skewness alone. We also need to consider the statistics for kurtosis. Kurtosis indicates the "peakness" or "flatness" of the distribution. Kurtosis is a measure to understand whether the data is heavy-tailed or light-tailed relative to the normal distribution. Data with high kurtosis tends to have heavy tails. The heavy-tailed distributions usually have outliers. Data with low kurtosis tends to have light tails, or a lack of outliers.

The value of kurtosis for a normal distribution is 3. The data for "age" has a skewness of -0.091 and a kurtosis of 2.69. Since the value of skewness is close to zero and the value of kurtosis is close to 3, we may consider that the variable "age" has come from

a normally distributed population. We can also check the normality of a dataset by other methods, which are discussed in Chapter 8.

Percentile and Quartile: Stata has provided the percentiles (1%, 5%, 10%, 25%, etc.) for the variable "age" (Table 6.4). It did not show the quartiles directly. When a dataset is divided into four equal parts after being arranged in ascending order, each part is called a quartile. It is expressed as Q1 (first quartile or 25th percentile), Q2 (second quartile or median or 50th percentile), and Q3 (third quartile or 75th percentile).

On the other hand, when the data is divided into 100 equal parts (after the ordered array – from lowest to highest), each part is called a percentile (P). We can see in Table 6.4 that the 10th percentile (P_{10}) is 16.5, the 25th percentile or P_{25} (Q1) is 21.0, the P_{50} (median or Q2) is 27.0, and the P_{75} (Q3) is 32.0 years. The Q1 (the first quartile) is 21 years, indicating that 25% of the study subjects' age is less than or equal to 21 years. On the other hand, the 75th percentile (P_{75} or Q3) which is 32 years, indicates that 75% of the study subjects' age is less than or equal to 32 years. There is another measurement, called interquartile range (IQR), which is not shown in the analysis. IQR is the difference between Q3 and Q1 (Q3 – Q1). In this example, the IQR is 11 years (32 – 21).

Sum of weights (wgt): When the "detail" option is used with the command "sum", Stata provides the statistics "sum of wgt.", indicating the sum of weights (Table 6.4). The table shows that the sum of weights for age is 210. Since we did not use any weight variable in the analysis, by default, each subject is given a weight of 1. When the option "detail" is used, the sum of the weights will therefore be equal to the number of observations, as shown in our example. We do not need this information to interpret the descriptive statistics.

95% confidence interval (CI): Table 6.4 shows the 95% CIs for age and other variables (income and systolic BP). The 95% CI for mean age is 25.4 to 27.5 years.This means that we are 95% confident that the mean age of the population from which the sample is drawn is between 25.4 and 27.5 years.

Table 6.6 shows the mean and 95% CI of the dichotomous categorical variable "diabetes1", which is coded as 0/1 (0= don't have diabetes; 1= have diabetes). Here, the mean value is 0.2142. This indicates that 21.42% (mean value multiplied by 100) of the subjects have diabetes. If it is a cross-sectional data, we can also say that the prevalence of diabetes is 21.42%. The 95% CI of the prevalence is 16.36% – 27.54%. Here, the 95% CI indicates that the prevalence of diabetes in the population would be between 16.36% and 27.54%, and we are 95% confident about it.

6.3 Descriptive statistics disaggregated by a categorical variable

You can get the descriptive statistics and other measures disaggregated (separately) by categorical variables. For example, if you want to get the frequency distribution of religion by sex (i.e., by males and females, separately), use any of the following commands (we need to sort the variable "sex" first by using the command "sort"). Outputs are provided in Table 6.7.

sort sex
by sex: **tab** religion
Or,
bysort sex: **tab** religion
Or,
by sex, **sort: tab** religion

If you want to get the measures of central tendency and dispersion of systolic BP (variable name is "sbp") by diabetes, use the following command. Stata will generate separate outputs for those with and without diabetes (Table 6.8). Other alternative commands are provided in Section 6.2.

bysort diabetes: sum sbp
bysort diabetes: sum sbp, detail

Table 6.7 Frequency distribution of religion by sex

```
. bysort sex: tab religion

-------------------------------------------------------------------------------
-> sex = f

   Religion |      Freq.     Percent        Cum.
------------+-----------------------------------
     MUSLIM |         76       57.14       57.14
      HINDU |         35       26.32       83.46
  Christian |         22       16.54      100.00
------------+-----------------------------------
      Total |        133      100.00

-------------------------------------------------------------------------------
-> sex = m

   Religion |      Freq.     Percent        Cum.
------------+-----------------------------------
     MUSLIM |         50       64.94       64.94
      HINDU |         23       29.87       94.81
  Christian |          4        5.19      100.00
------------+-----------------------------------
      Total |         77      100.00
```

Table 6.8 Descriptive statistics of systolic BP by diabetes

```
. bysort diabetes: sum sbp

-------------------------------------------------------------------------------
-> diabetes = No

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         sbp |       165    127.6061    20.87994         91        195

-------------------------------------------------------------------------------
-> diabetes = Yes

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         sbp |        45       128.2    16.90428        100        176
```

7
Generating Graphs

The information derived from data analysis needs to be presented in an effective and understandable manner. Data and information can be presented in textual, tabular, or graphical forms. Tables and graphs are powerful communication tools for the presentation of information. They can make an article easy to understand for the readers. While a table is suitable for presenting quantitative and qualitative information, a graph is an effective visual method for data presentation. A graph displays data at a glance, facilitates comparison, and can reveal trends and relationships.

The researchers need to carefully decide which type of graph or chart will be the best way to present the information. The type of graph or chart to be used depends on the data type, analysis outputs, and the objective of communicating the information. Inappropriate use of graphs may fail to convey the right information and may sometimes confuse the readers, leading to misinterpretation of data. Therefore, the graphs for data presentation must be chosen carefully.

The graphs commonly used for the presentation of data include histograms, scatter plots, box and plot charts, bar graphs, line graphs, and pie charts. In this chapter, we will discuss how to generate these graphs using Stata. Use the data file <**Data_3.dta**>.

7.1 Histogram

We usually generate a histogram to assess the distribution of a continuous variable. The histogram provides information about: a) the distribution of a dataset (whether symmetrical or not); b) the concentration of values; and c) the range of values. To generate a histogram (say, for the variable "age"), use the following command:

histogram age

This will display a histogram for age (Fig 7.1 A). You can specify the Y-axis scale in the command, such as frequency or percentage, by adding the following options:

histogram age, frequency
histogram age, percent
histogram age, norm percent

The first command will display a histogram where the Y-axis value is the number (frequency; Fig 7.1 B), while the second command will display the Y-axis value in percentage. The third command will produce a histogram with the overlying normal curve (Fig 7.1 C).

You can specify the interval width to construct a histogram. For example, if you want to have the histogram of age at a class interval of 3, use the following command:

histogram age, width(3) frequency
Or,
histogram age, width(3) start(2) frequency

You can also get the histograms dissertated by a categorical variable (say, by sex). To do this, use the following command:

histogram age, by(sex) frequency

This command will display histograms of age for females and males, separately (Fig 7.1 D). You can use the graph edit options to give a title, subtitle, and change the font, background, and color schemes of the histogram. These options can also be specified through commands.

Looking at the histogram (Fig 7.1), it seems that the data is more or less symmetrical. This indicates that age may be normally (approximately) distributed in the population.

7.1.1 Saving graphs

The graphs generated by Stata cannot be saved in the output files. They need to be saved separately. The graphs can be saved in various formats, such as Stata Graph (.gph), Enhanced Metafile (.emf), or in other formats. The disadvantage of a .gph format is that it can only be read by Stata. On the other hand, the .emf format is usually the best for use in Microsoft Word documents. Suppose that you have generated a histogram for age. To save a histogram (or other graphs) generated by Stata with the name “graph1”, use the following steps in the graph window:

File > Save As… > Select location of the file to be saved > Give a file name in the "File name" box (e.g., Graph1) > Select a format in the "Save as type" box (e.g., *.gph) > Save

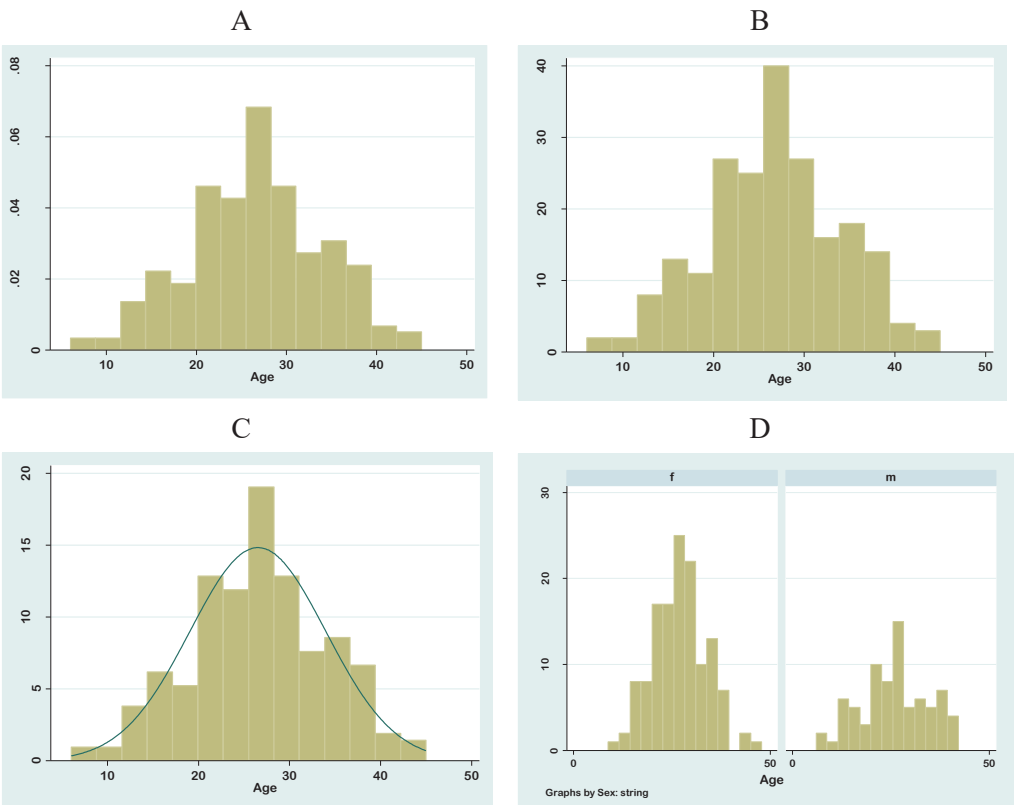


Figure 7.1 Histograms of age

You can also use the following commands to save the graph as "Graph1" on your desktop (or other locations):

graph save C:\Users\HP\Desktop\Graph1
graph save C:\Users\HP\Desktop\Graph1.emf, replace

The first command will save the graph on the desktop in .gph format, while the second command will save the graph on the desktop in .emf format.

7.2 Scatter plot

A scatter diagram provides useful information about the relationship between two continuous variables. The scatter diagram provides information/ideas about:

- Whether there is any correlation between the variables;
- Whether the relationship (if there is any) is linear or non-linear;
- The direction of the relationship, i.e., whether it is positive (if the value of one variable increases with the increase of the other variable) or negative (if the value of one variable decreases with the increase of the other variable);
- The presence of potential outliers in the dataset, i.e., the values that differ significantly from other observations.

Stata can generate scatter plots with varieties of options, such as with a fit line (regression line), 95% CI for the fit line, disaggregated by a categorical variable, and others. Several commands can be used to generate scatter plots. Use the following commands to generate a scatter plot for systolic (variable name is "sbp") and diastolic (variable name is "dbp") BP:

twoway scatter sbp dbp
twoway lfit sbp dbp || scatter sbp dbp
twoway lfitci sbp dbp || scatter sbp dbp

The first command will display the basic scatter plot of systolic BP against diastolic BP (Fig 7.2 A). The first variable after the command is considered for the Y-axis. The second command will display the regression line (fit line; Fig 7.2 B) for systolic BP on diastolic BP, while the third command will provide the 95% CI of the regression line (Fig 7.2 C).

The other commands and options that can be used to generate a scatter plot are:

twoway scatter sbp dbp, mlabel(ID_no)
twoway scatter sbp dbp, by(diabetes)
graph matrix age sbp dbp

The first command will display a scatter plot with the data points labelled by the case numbers (ID numbers) (Fig 7.3 A), while the second command will display a scatter plot by the variable "diabetes" (Fig 7.3 B). The last command will display a scatter plot matrix for the variables listed (age, systolic BP, and diastolic BP) with the command (Fig 7.4).

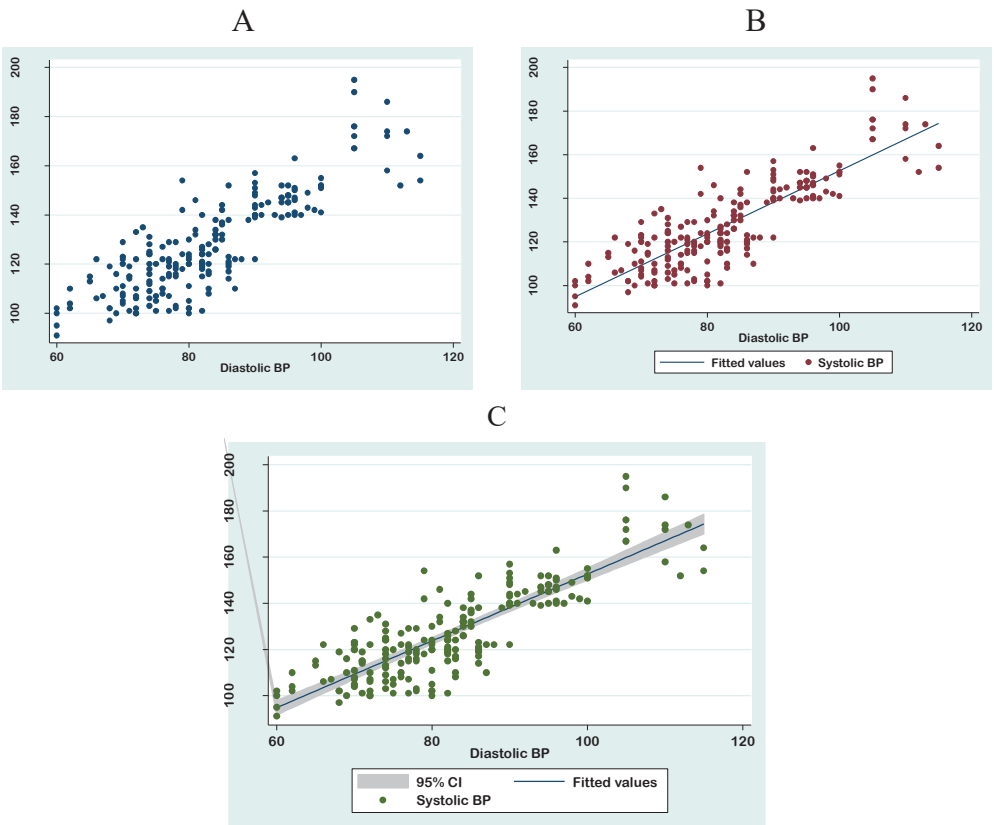


Figure 7.2 Scatter plots of systolic and diastolic BP

7.3 Box and plot chart

We have shown how to generate the basic box and plot charts in Chapter 4. Here, we will discuss the other options for generating the box and plot charts. The commands that may be used for generating a box and plot chart are as follows:

graph box sbp
graph box dbp, over(religion)
graph box age sbp dbp

The first command is the basic command for generating a box and plot chart. The first command will generate a box and plot chart of systolic BP as shown in Figure 7.5. The second command will display box and plot charts of diastolic BP for different catego-

ries of religion (Fig 7.6). You can also generate box and plot charts for multiple variables simultaneously by using the last command. Horizontal box and plot charts can also be generated by using the following command:

graph hbox sbp

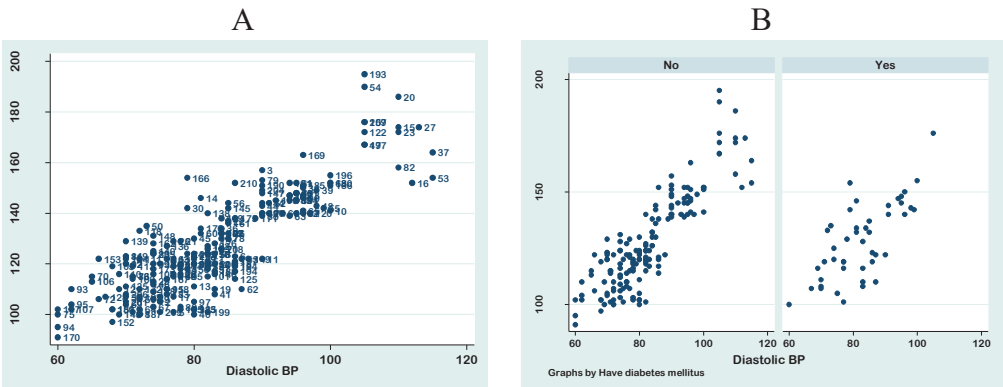


Figure 7.3 Scatter plots of systolic BP and diastolic BP with ID numbers (A) and by diabetes (B)

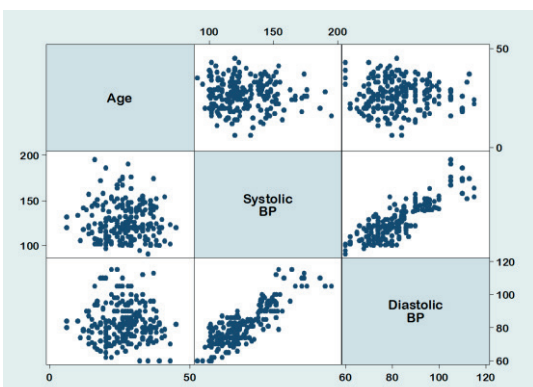


Figure 7.4 Scatter plots matrix for age, systolic BP and diastolic BP

The box and plot chart provides information about the distribution of a dataset. It also provides summary statistics of a variable, like Q1 (first quartile or 25th percentile), median (second quartile or Q2) and Q3 (third quartile or 75th percentile) as well as information about outliers or extreme values. The lower boundary of the box indicates

the value for Q1, while the upper boundary indicates the value for Q3. The median is represented by the horizontal line within the box. The minimum and maximum values are indicated by the horizontal lines of the whiskers (Fig 7.5).

In the box and plot chart, the presence of *outliers* is indicated by the dots (Fig 7.5). The outliers are the values greater than 1.5 box length distance (i.e., interquartile range or IQR) from the edge (upper or lower) of the box [i.e., greater than (1.5×IQR + Q3) or less than (Q1 – 1.5×IQR)]. You can see that there are 3 outliers in the data of systolic BP (Fig 7.5). If you want to get the case numbers of the outliers, use the following command ("ID_no" is the variable name for identification number in our dataset):

graph box sbp, marker (1, mlabel (ID_no))

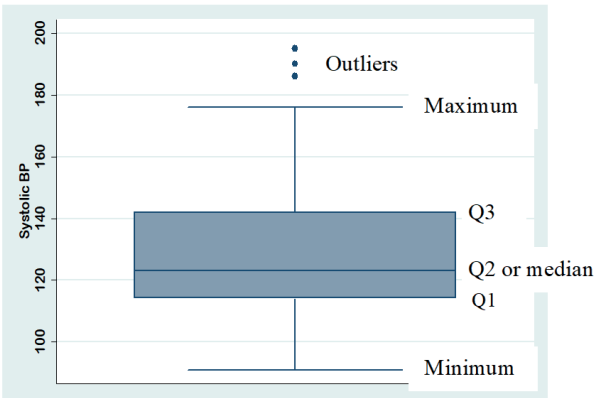


Figure 7.5 Box and plot chart of systolic BP

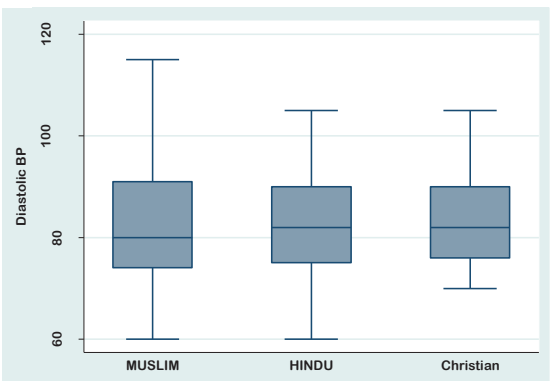


Figure 7.6 Box and plot chart of systolic BP by religion

7.4 Bar graph

Bar graphs are very powerful tools for presenting summary statistics, which make it easier for the readers to understand the relationship between the various values. Here, we will discuss how to generate simple bar graphs, such as: a) the mean of a quantitative variable across a categorical variable; and b) the frequencies of a categorical variable.

7.4.1 Bar graph for the mean of a quantitative variable across a categorical variable

You may be interested in presenting the mean income of the respondents by religion. To do this, use the following command:

graph bar income, over(religion)
graph hbar income, over(religion)

The first command will display a bar graph showing mean income across religious groups (Fig 7.7A). The second command will display a horizontal bar with the same information. If you want to get the median, instead of the mean income by religion, use the following command:

graph bar (median) income, over(religion)

Use the following commands if you want to get the value labels on each bar:

graph bar age, over(religion) blabel(bar)
graph bar age, over(religion) blabel(bar, format (%3.1f))

The first command will provide the value labels (mean age) on the bars. The addition of the option "format (%3.1f)" in the command will provide the mean up to 3 digits with one decimal point (Fig 7.7B).

7.4.2 Bar graph for the frequencies of a categorical variable

There is no specific and direct command for generating a bar chart to graphically represent a frequency table. The “fbar” command can be used to make a bar graph. However, before using the “fbar” command, we need to install a module using the following command:

ssc **install fbar**

Once the module is installed, use any of the following commands to make a bar graph:

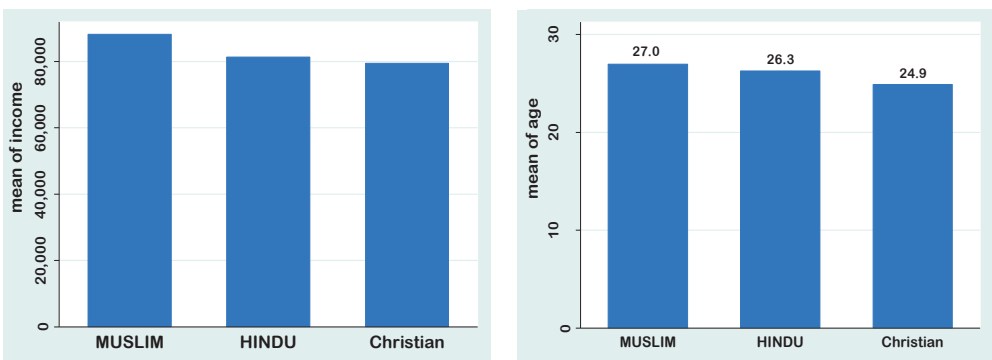
fbar religion
fbar religion, percent
fbar religion, by(sex)

The first command will generate a bar graph of religion with a frequency of occurrence and the second one with a percentage. The last command will generate a frequency bar graph of religion by sex. *The bar graphs generated by the "fbar" command cannot be edited in graph edits*.

An alternative way to generate a bar graph is to use the modified command for histogram. To generate a bar graph for religion with the modified command of histogram, use the following command:

histogram religion, discrete freq gap(20) xlabel(1 2 3, valuelabel)
histogram religion, discrete percent gap(20) xlabel(1 2 3, valuelabel)

The first command will generate a bar graph with the frequency of the bars, while the second one will provide the bar with a percentage.



A B

Fig 7.7 Bar graph of mean income (A) and mean age (B) by religion

7.5 Line graph

A line graph is particularly suitable to demonstrate the trend, i.e., changes over time. We can check the trend for two or more variables at a time. The basic command for generating a line graph is "graph twoway line" or simply "line". Suppose that you want to generate a line graph to demonstrate the change in TFR (total fertility rate; variable

name is "tfr") over the period of 1975 to 2018 (variable name is "year"). To do this, use the following command (use the data file <**Line.dta**>):

line tfr year
twoway connect tfr year

The first command will produce the line graph A without the markers of the scatter points, while the second command will produce the line graph B with the markers as shown in Figure 7.8 (graphs have been edited for better resolution). The command "connect" actually combines the features of scatter with a line, i.e., connects the scatter points with a line segment. The Y-axis variable (dependent variable) must be written first after the basic command.

You can also include two dependent variables to generate a line graph. Suppose that you want to demonstrate the trend of TFR as well as CPR (contraceptive prevalence rate; variable name is "cpr") over the years. Use the following command to get the line graph (Fig 7.8 C).

twoway connect tfr cpr year

Though you can give a title, Y-axis and X-axis labels, and others by using the Stata commands, it is easier to write them using the graph edits. Users can easily explore the graph editing options in Stata.

If you have a dataset in wide format (use the data file **"Repeated Anova_2",** which is in wide format), you can make a line graph by using the following command:

profileplot sugar_0 sugar_7 sugar_14 sugar_24, by(treatment)
Or,
profileplot sugar_0 - sugar_24, by(treatment)

Either of the above commands will generate a line graph of mean blood sugar levels at 4 different time points for three treatment groups ("sugar_0 - sugar_24" indicates the variables sugar_0 to sugar_24).

7.6 Pie chart

Pie charts are also commonly used to present data. Stata can generate pie charts. The "graph pie" command with the "over" option generates a pie chart representing the frequency of each group. The "plabel" option places the labels (names or values) of the categories inside each slice of the pie chart. *Note that Stata does not allow a string variable for making a pie chart.* The following commands can be used to make a pie

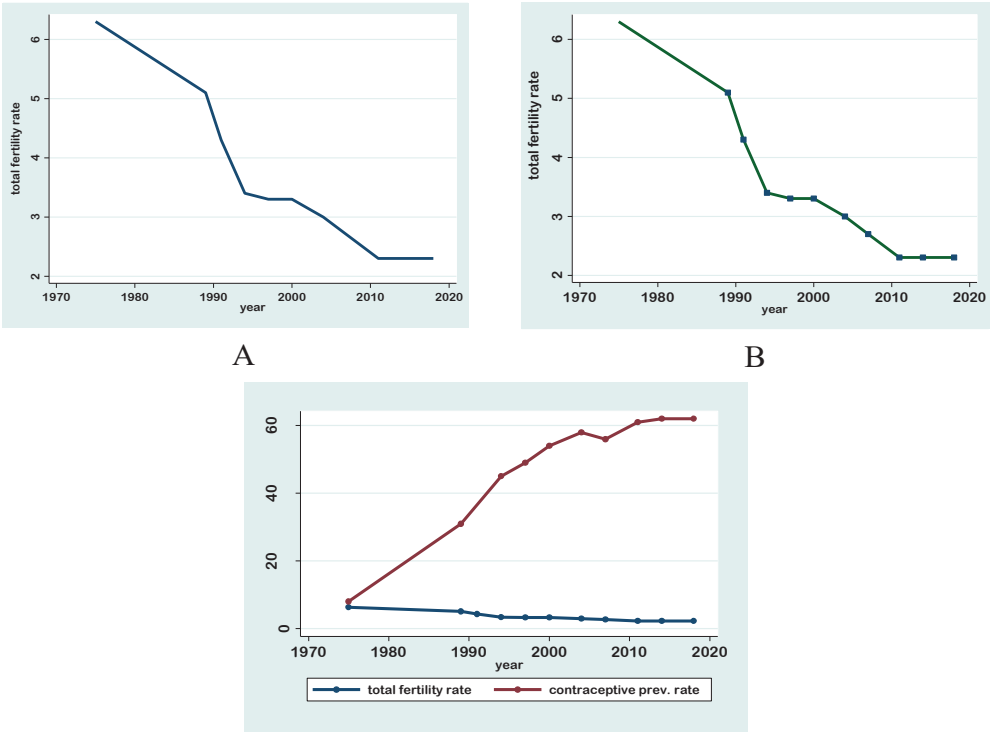


Fig 7.8 Line graphs of TFR (A & B) and trends of TFR and CPR (C) by year

chart for the variable "religion".

graph pie, over(religion)
graph pie, over(religion) plabel(_all name)
graph pie, over(religion) plabel(_all percent)
graph pie, over(religion) pie(1, explode) plabel(_all percent)

The first command will produce a simple pie chart for the variable religion (Fig 7.9 A). The second command will display a pie chart with category names (Fig 7.9 B), while the third command will display category percentages (Fig 7.9 C). To explode a slice of a pie graph, use the last command (Fig 7.7 D).

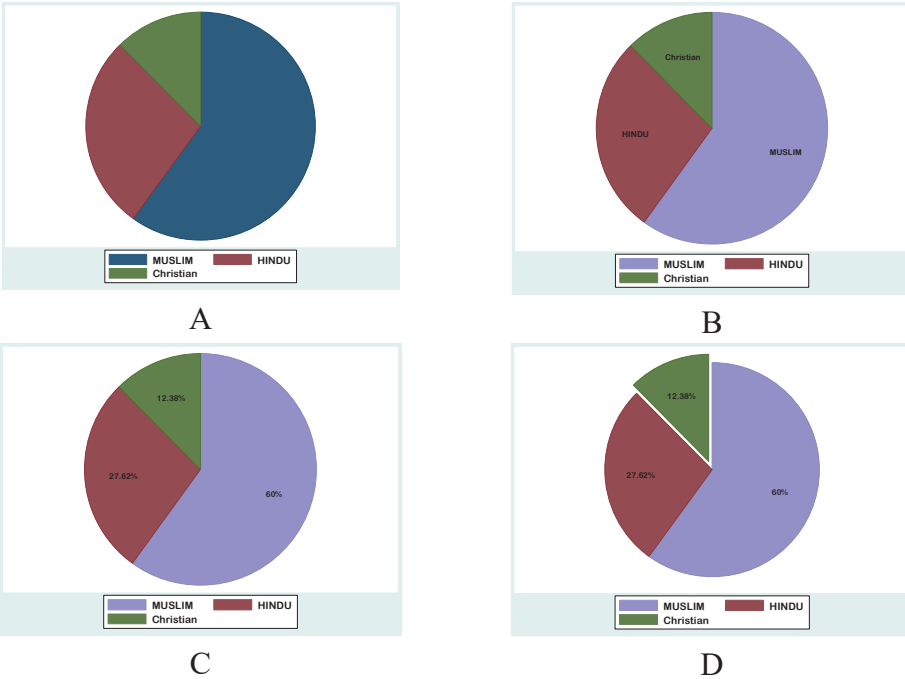


Fig 7.9 Pie charts of religion

8

Checking Data for Normality

It is important to know the nature of the distribution of a continuous random variable before using the statistical tests for hypotheses testing. To use the parametric methods for hypothesis testing (such as t-test, ANOVA, correlations, and others), one of the important assumptions is that the data of a continuous dependent variable is normally distributed. It is, therefore, necessary to check whether the data has come from a normally distributed population or not before we use the parametric methods. Use the data file <**Data_3.dta**> for practice.

8.1 Assessing normality of data

The normality of data is an important assumption for using a parametric method of testing a hypothesis. Whether the data is from a normally distributed population or not, can be checked in several different ways. The most commonly used methods are:

a) Graphs, such as histograms and Q-Q (quantile-quantile) plots,
b) Descriptive statistics, such as skewness and kurtosis, and
c) Formal statistical tests, such as the Shapiro Wilk test (commonly used) and the Skewness-Kurtosis (S-K) test.

In this section, we will discuss how to get a histogram and a Q-Q plot and how to perform the formal statistical tests (Shapiro Wilk test and S-K test) to check the normality of data.

We want to assess whether or not the variable "systolic BP" in our dataset is normally distributed in the population. For this, we will first construct a histogram and a Q-Q plot for systolic BP (variable name is “sbp”). To construct a histogram for systolic BP,

use the following command (also see Chapter 7):

histogram sbp, norm
histogram sbp, by(diabetes) norm

The first command will generate a histogram of systolic BP with an overlying normal curve (Fig 8.1). The second command will display the same by diabetes.

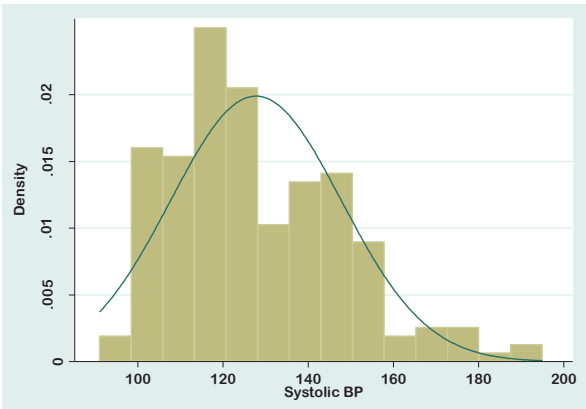


Figure 8.1 Histogram of systolic BP

To construct the Q-Q plot for systolic BP, use the first command (Fig 8.2). The second command will generate a Q-Q plot for males.

qnorm sbp
qnorm sbp if sex==”m”

Formal statistical tests to assess the normality of data can also be used. The statistical tests for checking the normality of data are the Shapiro Wilk test (commonly used) and the Skewness-Kurtosis test. To do these tests for systolic BP, use the following commands:

swilk sbp
sktest sbp
swilk sbp if diabetes==1
sktest sbp if diabetes==1

The outputs of first two commands are provided in Table 8.1. You can also use these tests for multiple variables at a time.

8.1.1 Interpretation

With Stata commands, we have generated the histogram and Q-Q plot (Figs 8.1 and 8.2) for systolic BP to visually check the distribution of data. We have also used the formal statistical tests (Shapiro Wilk test and Skewness-Kurtosis test) to assess the normality of data (Table 8.1).

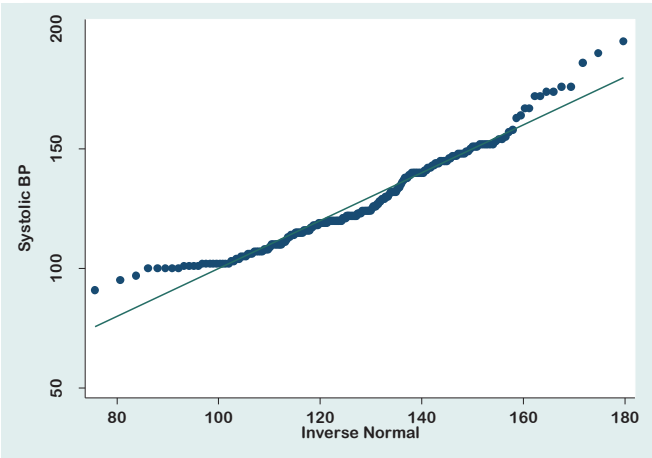


Figure 8.2 Q-Q plot of systolic BP

Table 8.1 Normality tests for systolic BP

```
. swilk sbp
                    Shapiro-Wilk W test for normal data

    Variable |    Obs       W           V         z       Prob>z
-------------+-------------------------------------------------
         sbp |    210    0.95618      6.821     4.429    0.00000

. sktest sbp
                     Skewness/Kurtosis tests for Normality
                                                       ------- joint ------
    Variable |    Obs   Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------------
         sbp |    210      0.0001         0.2925        14.72         0.0006
```

The histogram (Fig 8.1) provides an impression about the distribution of the data (whether the distribution is symmetrical or not). If we look at the histogram of systolic BP, it seems that the data is slightly skewed to the right (i.e., the distribution is not symmetrical).

The Q-Q plot (Fig 8.2) also provides information on whether the data has come from a normally distributed population or not. The Q-Q plot compares the distribution of data with the standardized theoretical distribution from a specified family of distribution (in this case, from a normal distribution). *If the data is normally distributed, all the points (dots) lie on the diagonal straight line.* Our interest is in the central portion of the line as well as in the tails. Deviation from the central portion of the line means non-normality. Deviations at the ends of the plot indicate the existence of outliers. We can see (Fig 8.2) that there is a slight deviation of the dots at the central portion as well as at the two ends. This may indicate that the data may not have come from a normally distributed population.

The specific tests (objective tests) that we have used to assess if the data has come from a normally distributed population are the *Shapiro Wilk test and the S-K test*. The results of these two tests are provided in Table 8.1.

Look at the "Prob > z" (for the Shapiro Wild test) and "Prob>chi2" (for the S-K test) columns of Table 8.1. These columns indicate the p-values of the tests. A p-value of <0.05 indicates that the data *has not* come from a normally distributed population. In our example, the p-value is 0.000 for both the tests. This indicates that the data for systolic BP has not come from a normally distributed population. Here, the *null* hypothesis is "data has come from a normally distributed population" and the alternative hypothesis is "data has not come from a normally distributed population". We will reject the null hypothesis since the p-values of the tests are <0.05.

The formal tests are very sensitive to sample size. These tests may be significant for slight deviations in large sample data (n>100). Similarly, the likelihood of getting a p-value <0.05 for a small sample of data (n<20, for example) is low. Therefore, the rules of thumb for normality checking are:

1) For a sample size of <30: Assume non-normal;
2) For a moderate sample size (30-100): If the formal statistical test is significant (p<0.05), consider a non-normal distribution; otherwise, check the normality using other methods, such as histograms and Q-Q plots;
3) For a large sample size (n>100): If the formal statistical test is not significant (p>0.05), accept normality; otherwise, check with other methods.

However, for practical purposes, just look at the histogram. If it seems that the distribution is approximately symmetrical, consider that the data has come from a normally distributed population and use a parametric test. If the sample size is less than 30, use a nonparametric test.

9

Testing of Hypothesis

The present and the following chapters provide basic information on how to select statistical tests for testing hypotheses, perform the statistical tests with Stata, and interpret the results of common problems related to health and social sciences research. Before we proceed, let us discuss some of the basic concepts of hypothesis testing.

A hypothesis is a statement about one or more populations. A hypothesis is concerned with the parameter of a population about which the statement is made. A hospital manager may hypothesize that the average length of stay at his hospital is seven days, or a researcher may hypothesize that the rate of recovery with drug A is better than that of drug B. By means of hypothesis testing, one determines whether or not such statements are compatible with the available data.

There are two types of statistical hypotheses: null (H_0) and alternative (H_A) hypotheses. The null hypothesis is the hypothesis of equality or no difference. The null hypothesis always says that two or more quantities (parameters) are equal. *We always test the null hypothesis, not the alternative hypothesis.* Using a statistical test, we either reject or do not reject the null hypothesis. If we can reject the null hypothesis, then we can only accept the alternative hypothesis. It is, therefore, necessary to have a very clear understanding of the null hypothesis.

Suppose that we are interested in determining the association between coffee drinking and stomach cancer. In this situation, the null hypothesis is "there is no association between coffee drinking and stomach cancer (or, coffee drinking and stomach cancer are independent)", while the alternative hypothesis is "there is an association between coffee drinking and stomach cancer (or, coffee drinking and stomach cancer are not independent) ". If we can reject the null hypothesis with a statistical test (i.e., if the test

is significant; p-value <0.05), then we can only say that there is an association between coffee drinking and stomach cancer.

Various statistical tests are available to test hypotheses. Selecting an appropriate statistical test is the key to analyzing the data. What statistic is to be used to test a hypothesis depends on the study design, data type, distribution of data, and objective of the study. It is, therefore, important to understand the nature of the variable (categorical or quantitative), measurement type (nominal, ordinal, interval, or ratio scale), as well as the study design. The following tables (Tables 9.1 to 9.4) provide basic guidelines about the selection of statistical tests depending on the type of data and situation.

Table 9.1 Association between quantitative and categorical or quantitative variables

	Situation for hypothesis testing	Data normally distributed	Data non-normal
1.	**Comparison with a single population mean (with a fixed value)** Example: You have taken a random sample from a population of diabetic patients to assess the mean age. Now, you want to test the null hypothesis that the mean age of the diabetic patients in the population is 55 years.	1-sample t-test	Sign test/ Wilcoxon Signed Rank test
2.	**Comparison of the means of two related samples (pre- and post-test comparison)** Example: You want to test the hypothesis of whether the drug "Inderal" is effective in reducing blood pressure (BP) or not. To test the hypothesis, you selected a group of subjects and measured their BP before giving the drug (measurements before treatment or pre-test). Then, you administered the drug to all of them and measured their BP after one hour (measurements after treatment or post-test). Now, you want to compare if the mean BP before (pre-test) and after (post-test)	Paired t-test	Sign test/ Wilcoxon Signed Rank test

Table 9.1 Association between quantitative and categorical or quantitative variables

	Situation for hypothesis testing	Data normally distributed	Data non-normal
	administration of the drug is the same or not.		
3.	**Comparison between two independent sample means [association between a quantitative and a categorical variable with *two levels*]** Example: You have taken a random sample of students from a university. Now, you want to test the hypothesis if the mean systolic BP of male and female students is the same or not.	Independent samples t-test	Mann-Whitney U test (also called the Wilcoxon Rank Sum test)
4.	**Comparison of more than two independent sample means [association between a quantitative and a categorical variable with *more than two levels*]** Example: You have taken a random sample from a population. You want to test the hypothesis if the mean income of different religious groups (e.g., Muslims, Hindus, and Christians) is the same or not. Another example, you have three drugs, A, B, and C. You want to investigate whether all these three drugs are equally effective in reducing blood pressure or not.	One-way ANOVA	Kruskal Wallis test
5.	**Association between two quantitative variables** Example: You want to test the hypothesis if there is a correlation between systolic BP and age.	Pearson's correlation	Spearman's correlation (Can also be used for ordinal variables)

Table 9.1 Association between quantitative and categorical or quantitative variables

	Situation for hypothesis testing	Test Statistics
6.	**Association between a quantitative and an ordinal variable** Example: You want to test the hypothesis if there is a correlation between systolic BP and severity of anemia.	Spearman's correlation, if the ordinal variable has 5 or more levels. Otherwise, use the Kendall's Tau-B statistics
7.	**Association between two ordinal variables** Example: You want to test the hypothesis if there is a correlation between severity of pain and stage of cancer.	Spearman's correlation, if both the ordinal variables have 5 or more levels. Otherwise, use the Kendall's Tau-B statistics

Table 9.2 Association between two categorical variables

	Situation for hypothesis testing	Test statistics
1.	**Association between two categorical variables (independent samples)** Example: You have taken a random sample from a population and want to test the hypothesis that there is an association between sex and asthma. Another example, you want to assess the association between smoking and stomach cancer.	Chi-square test/ Fisher's Exact test
2.	**Association between two categorical variables of related samples, such as data from a matched case-control study** Example: You want to test the hypothesis if there is an association between diabetes mellitus and heart disease when the data is matched for smoking or other variables (a matched case-control study design).	McNemar test

Table 9.3 Multivariable analysis

	Type of outcome/dependent variable	Type of multivariable analysis
1.	The outcome variable (also called the dependent variable) is on an interval or ratio scale – e.g., blood pressure, birth weight, and blood sugar.	Multiple linear regression; Analysis of variance (ANOVA); Analysis of covariance (ANCOVA)
2.	The dependent variable is a *dichotomous* categorical variable (i.e., a nominal categorical variable with two levels) – e.g., disease (present or absent); ANC (taken or not taken); and outcome (cured or not cured).	Binary logistic regression
3.	The dependent variable is a nominal categorical variable with *more than two levels* – e.g., treatment seeking behavior (such as treatment not received; received homeopathic treatment; received allopathic treatment); and cause of death (cancer, heart disease, pneumonia).	Multi-nominal logistic regression
4.	The dependent variable is an ordinal categorical variable – e.g., severity of anemia (no anemia, mild to moderate anemia, severe anemia); stage of cancer (stage 1, stage 2, stage 3); severity of pain (mild, moderate, severe), etc.	Proportional odds regression (Ordinal regression)
5.	The dependent variable is time-to-outcome/event, such as time-to-death, time-to-recurrence, and time-to-cure.	Proportional hazards analysis (Cox regression)
6.	The dependent variable is a count – e.g., the number of post-operative infections; the number of patients admitted with heart disease to a hospital; and the number of road traffic accident cases treated in the emergency department.	Poisson regression

Table 9.3 Multivariable analysis

	Type of outcome/dependent variable	Type of multivariable analysis
7.	Incidence rates, such as incidence rates of tuberculosis; incidence rates of pneumonia; incidence rates of car accidents, etc.	Poisson regression

Table 9.4 Agreement analysis

	Situation for hypothesis testing	Test statistics
1.	**Agreement between two quantitative variables** Example: You want to test the hypothesis that the two methods of blood sugar measurement agree with each other.	Bland Altman test/plots
2.	**Agreement between two categorical variables** Example: You want to test the hypothesis that the diagnosis of cataract in patients is agreed upon by two physicians.	Kappa estimates

10

Student's t-test for Hypothesis Testing

The student's t-test is commonly known as the t-test. It is a frequently used parametric statistical method to test a hypothesis. There are several types of t-tests used in different situations (Table 9.1), such as: a) One-sample t-test; b) Independent samples t-test; and c) Paired t-test. In this chapter, we will discuss all these t-tests and the interpretation of their outputs. Use the data file <**Data_3.dta**> for practice.

10.1 One-sample t-test

The one-sample t-test is done to compare the mean with a hypothetical value. For example, we have collected data on diastolic BP (variable name is "dbp") of students of the State University of Bangladesh by taking a random sample. We are interested in knowing if the mean diastolic BP of students is different from 80 mmHg.

Hypothesis

Null hypothesis (H_0): The mean diastolic BP of students is equal to 80 mmHg in the population (study population is the students of the State University of Bangladesh).

Alternative hypothesis (H_A): The mean diastolic BP of students is different from (not equal to) 80 mmHg in the population.

Assumptions

1. The distribution of diastolic BP in the population is normal;
2. The sample is a random sample from the population.

The first job, before testing the hypothesis, is to check whether or not the distribution of diastolic BP is normal in the population (assumption 1). To do this, check the histogram and/or Q-Q plot of diastolic BP or do the formal statistical test of normality (Shapiro Wilk test or S-K test) as discussed in Chapter 8. If the assumption is met (diastolic BP is at least approximately normal), do the 1-sample t-test; otherwise, use the nonparametric method (Wilcoxon Signed Rank test, as discussed in Chapter 20). Suppose that the diastolic BP is normally distributed in the population. Use the following command to do the 1-sample t-test:

ttest dbp=80

With this command, Stata will generate Table 10.1.

Table 10.1 One-sample t-test

```
. ttest dbp=80

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     dbp |     210    82.76667    .8107779    11.74929    81.16832    84.36502
------------------------------------------------------------------------------
    mean = mean(dbp)                                              t =   3.4124
Ho: mean = 80                                   degrees of freedom =      209

    Ha: mean < 80               Ha: mean != 80                 Ha: mean > 80
 Pr(T < t) = 0.9996         Pr(|T| > |t|) = 0.0008          Pr(T > t) = 0.0004
```

10.1.1 Interpretation

In this example, we have tested the null hypothesis "the mean diastolic BP of the students is equal to 80 mmHg in the population". Data shows that the mean diastolic BP of the sample is 82.77 mmHg (95% CI: 81.16 - 84.36) with a SD of 11.75 mmHg (Table 10.1). The results of the one-sample t-test show that the calculated value of "t" is 3.412 and the p-value [(Pr(|T| > |t|); 2-tailed] is 0.0008. Since the p-value is <0.05, we will reject the null hypothesis at the 95% confidence level. This means that the mean diastolic BP of the students (in the population) from where the sample is drawn is different from 80 mmHg (p<0.001).

10.2 Independent samples t-test

The independent samples t-test involves one quantitative variable (dependent variable) and a categorical variable with *two levels* (categories). This test is done to compare the mean of a dependent variable between two categories of the categorical variable.

For example, we are interested in knowing if the mean age of diabetic and non-diabetic patients in the population is the same or not. Here, the test-variable (dependent variable) is age (a quantitative variable) and the categorical variable is diabetes, which has two levels/categories (have diabetes and do not have diabetes). Before doing this test, we need to check assumption 1 [i.e., age is normally distributed at each level of diabetes], as discussed earlier.

Hypothesis

H_0: The mean age of diabetic and non-diabetic patients is the same in the population.

H_A: The mean age of diabetic and non-diabetic patients is different (not the same) in the population.

Assumptions

1. The dependent variable (age) is normally distributed at each level of the independent variable (diabetes);
2. The variances of the dependent variable (age) at each level of the independent variable (diabetes) are the same/equal;
3. Subjects represent random samples from the population.

10.2.1 Test for equality of variances: Levene's test

Before doing the independent samples t-test, we need to check if the variances of the dependent variable (age) at each level of the independent variable (diabetes) are the same or not (assumption 2). Levene's test is the most commonly used statistical test to determine whether two or more groups have equal variances. To do the Levene's test to determine if the variances of age are equal among the diabetic and non-diabetic patients, use the following command (Table 10.2):

robvar age, by(diabetes)

You can also use the following command to check the equality of variances. It will provide the results of variance ratio statistics (Table 10.2).

sdtest age, by(diabetes)

Table 10.2 Test results for equality of variances

```
. robvar age, by(diabetes)

       Have |
   diabetes |           Summary of Age
   mellitus |        Mean   Std. Dev.       Freq.
------------+------------------------------------
        Yes |   27.911111   8.4633494          45
         No |   26.133333   7.1835969         165
------------+------------------------------------
      Total |   26.514286   7.4904907         210

W0  =  3.2178863   df(1, 208)     Pr > F = 0.0742901
W50 =  3.0099114   df(1, 208)     Pr > F = 0.08423809
W10 =  3.0595865   df(1, 208)     Pr > F = 0.08173708


. sdtest age, by(diabetes)

Variance ratio test
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     yes |      45    27.91111    1.261642    8.463349    25.36844    30.45378
      no |     165    26.13333    .5592423    7.183597    25.02909    27.23758
---------+--------------------------------------------------------------------
combined |     210    26.51429     .516893    7.490491    25.49529    27.53328
------------------------------------------------------------------------------
    ratio = sd(yes) / sd(no)                                      f =   1.3880
Ho: ratio = 1                                     degrees of freedom =  44, 164

    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
  Pr(F < f) = 0.9265         2*Pr(F > f) = 0.1471           Pr(F > f) = 0.0735
```

The results of Levene's test (using the command "robvar") are provided in Table 10.2. The table shows the mean and standard deviation of age as well as the total observations (Freq) for both diabetic (Yes) and non-diabetic (No) patents. We can see that the standard deviation of age is higher among diabetic patients (8.46) compared to non-diabetic (7.18) patients, but the Levene's test will tell us whether or not this difference is statistically significant. The results of the "sdtest" command also provide the same information as shown after Levene's test results in Table 10.2.

Stata has provided three options for Levene's test results:

- **W0**, which is the test statistic of Levene's test centered at the mean;
- **W50**, which is the test statistic centered at the median; and
- **W10**, which is the test statistic centered at a 10% trimmed mean (i.e., the top 5% and bottom 5% of the values are trimmed out).

We will consider the test statistics centered at the median (W50) and the p-value for this is 0.084 (>0.05). This means that the variance of age for diabetic and non-diabetic patients is not different (equal). Similarly, the p-value (0.147) of the "sdtest" indicates the same, i.e., the variances for diabetic and non-diabetic patients with regard to age are not different.

10.2.2 Commands for independent samples t-test

The independent samples t-test can be done in two situations:

- When the variances of the dependent variable (e.g., age) are **equal** at each level of the categorical variable (e.g., diabetes); and
- When the variances of the dependent variable are **unequal** at each level of the categorical variable.

The commands for the independent samples t-test are:

ttest age, by(diabetes)
ttest age, by(diabetes) unequal

The first command is for the t-test when the variances are equal, while the second command is for the t-test when the variances are unequal. The outputs of the t-tests are provided in Tables 10.3 (with equal variances) and 10.4 (with unequal variances).

10.2.3 Interpretation

Table 10.3 shows the descriptive measures of age by diabetes. We can see that there are 45 subjects with diabetes and 165 subjects without diabetes. The mean age of the diabetic patients is 27.9 (SD 8.46) and that of the non-diabetic patients is 26.1 (SD 7.18) years.

The t-test results are provided at the bottom of the descriptive statistics. The calculated t-value is -1.41 (with the degrees of freedom 208). Since we are interested in understanding if the mean age is the same for both diabetic and nondiabetic patients, we will consider the 2-tailed test. Look at the p-value, provided under "Ha: diff !=0", which is

0.158 (>0.05). We cannot, therefore, reject the null hypothesis. This means that the mean age of diabetic and non-diabetic patients in the population from where samples are drawn is not different (p=0.158).

The results of the t-test for unequal variances are provided in Table 10.4. The interpretation of the results is the same as above.

Table 10.3 Independent samples t-test with equal variances

```
. ttest age, by(diabetes)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
      No |     165    26.13333    .5592423    7.183597    25.02909    27.23758
     Yes |      45    27.91111    1.261642    8.463349    25.36844    30.45378
---------+--------------------------------------------------------------------
combined |     210    26.51429     .516893    7.490491    25.49529    27.53328
---------+--------------------------------------------------------------------
    diff |           -1.777778    1.256707               -4.255293    .6997375
------------------------------------------------------------------------------
    diff = mean(No) - mean(Yes)                                   t =  -1.4146
Ho: diff = 0                                     degrees of freedom =      208

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0793         Pr(|T| > |t|) = 0.1587          Pr(T > t) = 0.9207
```

Table 10.4 Independent samples t-test with unequal variances

```
. ttest age, by(diabetes) unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
      No |     165    26.13333    .5592423    7.183597    25.02909    27.23758
     Yes |      45    27.91111    1.261642    8.463349    25.36844    30.45378
---------+--------------------------------------------------------------------
combined |     210    26.51429     .516893    7.490491    25.49529    27.53328
---------+--------------------------------------------------------------------
    diff |           -1.777778    1.380033               -4.536122    .9805668
------------------------------------------------------------------------------
    diff = mean(No) - mean(Yes)                                   t =  -1.2882
Ho: diff = 0                     Satterthwaite's degrees of freedom =  62.3436

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.1012         Pr(|T| > |t|) = 0.2024          Pr(T > t) = 0.8988
```

10.3 Paired t-test

The paired t-test is done to compare the difference between the two means of related samples. Related samples indicate the measurements taken from the same subjects at two or more different times or situations. For example, you have organized a training session for 32 staff members of your organization. To evaluate the effectiveness of the training, you have taken a pre-test before the training to assess the current status of knowledge of the participants. At the end of the training, you have again taken a test (post-test). Now, you want to compare if the training has significantly increased the knowledge or not.

Another example is, suppose that you want to determine the effectiveness of a drug (e.g., Inderal) in reducing the systolic blood pressure (BP). To do this, you have selected a random sample from a population. You have measured the systolic BP of all the individuals before giving the drug (pre-test or baseline measurement). You have again measured their BP one-hour after giving the drug (post-test or endline measurement)". The paired t-test is the appropriate test to compare the means in both situations.

Hypothesis

$\mathbf{H_0}$: There is no difference in mean scores before and after the training (for example 1).

$\mathbf{H_A}$: The mean scores are different before and after the training.

Assumptions

1. The difference between two measurements (pre- and post-test) of the dependent variable (examination scores) is normally distributed;
2. Subjects represent a random sample from the population.

10.3.1 Commands

We will use the first example (to compare the pre- and post-test scores) to do the paired t-test. Use the following command (the variable names are: pre_test and post_test) for the paired t-test (Table 10.5):

ttest post_test = pre_test

Table 10.5 Paired t-test results

```
. ttest post_test= pre_test

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
post_t~t |      32    90.98438    1.492164    8.440957    87.94109    94.02766
pre_test |      32    53.57813    2.727373    15.42835    48.01561    59.14064
---------+--------------------------------------------------------------------
    diff |      32    37.40625     2.47848     14.0204    32.35136    42.46114
------------------------------------------------------------------------------
     mean(diff) = mean(post_test - pre_test)                      t =  15.0924
 Ho: mean(diff) = 0                              degrees of freedom =       31

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

10.3.2 Interpretation

Table 10.5 shows the descriptive statistics of both the pre- and post-test results. Looking at the mean scores, we can get an impression of whether the training has increased the mean score or not. We can see that the post-test mean is 90.9, while the pre-test mean is 53.5, and the difference is large (37.4). To understand if the difference between post-test mean and pre-test mean is significant or not, we need to check the paired t-test results given at the bottom of the descriptive statistics. The results show that the calculated t-value is 15.09 (degrees of freedom: 31).

Three different p-values for the paired t-test are displayed at the bottom of the table. Since we are interested in understanding if there is any difference in the mean scores before and after the training (a two-tailed test), we will consider the p-value provided at the middle [Ha: mean(diff) != 0], which is 0.000. Since this value is smaller than 0.05 (our level of significance), we will reject the null hypothesis. We can, therefore, conclude that there is sufficient evidence to say that there is a significant difference in the mean scores before and after the training; i.e., the mean knowledge score has increased significantly after the training ($p<0.001$).

11

Analysis of Variance (ANOVA)

Analysis of variance (ANOVA) is a commonly used statistical method for testing hypothesis. An ANOVA test is done to compare the means when the categorical independent variable has more than two levels (categories). There are several types of ANOVA tests, such as one-way ANOVA, two-way ANOVA, repeated-measures ANOVA, and others. In this chapter, one-way and two-way ANOVA are discussed. The repeated measures ANOVA is discussed in Chapter 12. Use the data file <**Data_3.dta**> for practice.

11.1 One-way ANOVA

One-way ANOVA is done to compare the means of more than two groups, while the t-test compares the means of two groups. One-way ANOVA involves two variables: one categorical variable with more than two levels or categories (for example, the variable "religion_2", which has 4 categories – Muslim, Hindu, Christian, and Buddhist); and a quantitative variable (e.g., income, age, and blood pressure). Suppose that you want to assess if the mean income (variable name is "income") of all the religious groups (variable name is "religion_2") is the same or not in the population. One-way ANOVA is the appropriate test for this, if the assumptions are met.

Hypothesis

$\mathbf{H_0}$: The mean income of all the religious groups is equal.

$\mathbf{H_A}$: Not all means (of income) among the religious groups are equal.

Assumptions

1. The dependent variable (income) is normally distributed at each level of the independent variable (religion);
2. The variances of the dependent variable (income) at each level of the independent variable (religion) are the same (homogeneity of variances); and
3. Subjects represent random samples from the population.

If the variances of the dependent variable in all the categories are not equal (violation of assumption 2), but the sample size in all the groups is large and similar, ANOVA can be used.

11.1.1 Commands

All the following commands will provide the ANOVA test results. We recommend using the first command, which provides the descriptive statistics along with the ANOVA test results (Table 11.1).

oneway income religion_2, tabulate
oneway income religion_2
anova income religion_2

Table 11.1 One-way ANOVA test results of income and religion

```
. oneway income religion_2, tabulate

            |      Summary of Monthly income
 Religion 2 |        Mean   Std. Dev.       Freq.
------------+------------------------------------
     MUSLIM |   88180.905   17207.614         126
      HINDU |   79166.028   17804.631          36
  Christian |   79405.615   19857.021          26
   BUDDHISM |   84796.591   14447.348          22
------------+------------------------------------
      Total |   85194.486   17724.033         210

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      3.3068e+09      3   1.1023e+09      3.64     0.0136
 Within groups      6.2349e+10    206    302663565
------------------------------------------------------------------------
    Total           6.5656e+10    209    314141354

Bartlett's test for equal variances:  chi2(3) =   2.2804  Prob>chi2 = 0.516
```

11.1.2 Interpretation

The outputs of the one-way ANOVA test are provided in Table 11.1. In this example, we have used "income" as the dependent variable and "religion" as the independent variable. The independent variable (religion) has 4 categories (levels) – Muslim, Hindu, Christian, and Buddhist. The results first provided the descriptive measures (mean, SD, etc.) of income by religion. For example, the mean income of Muslims is BDT 88,180.9, with a SD of 17,207.6.

The ANOVA test (F-test) results are provided at the bottom of the descriptive statistics. The value of the F-statistic is 3.64 and the corresponding p-value (Prob > F) is 0.0136. Since the p-value is <0.05, we can reject the null hypothesis. This means that not all group means of income are equal.

However, before interpreting the ANOVA test results, we need to check the Bartlett's test for homogeneity of variances (equal variances) provided at the bottom of the table. This test is done to assess if all the group-variances in income are equal (assumption 2). The p-value (Prob>chi2) of the Bartlett's test is 0.516. Since the p-value is >0.05, the variances of income at all levels of the religious group are equal (i.e., assumption 2 is not violated). The assumption would have been violated if the p-value was <0.05.

11.1.3 Post hoc test

If the ANOVA test is significant, it indicates that not all group means are equal. But it does not provide information about which group-means are significantly different. To identify which group means are significantly different, we need to use a post hoc multiple comparison test, such as Bonferroni, Tukey, or Scheffe's test. Use any of the following commands (preferably the first one) to get the post hoc test results (Tables 11.2 and 11.3). If the ANOVA test (F-test) is not significant (i.e., p-value is >0.05), we do not need the post-hoc test.

pwmean income, over(religion_2) mcompare(bonferroni) effects
oneway income religion_2, bonferroni tabulate

You can generate the box and plot charts to display the distribution of medians/means of the dependent variable across the groups (at each level of the independent variable, i.e., in different religious groups). To get the box and plot charts (Fig 11.1) of income for different religious groups, use the following command:

graph box income, over(religion_2)

Table 11.2 Multiple comparisons of mean income by religion

```
. pwmean income, over(religion_2) mcompare(bonferroni) effects

Pairwise comparisons of means with equal variances

over         : religion_2
----------------------------
             |    Number of
             |  Comparisons
-------------+--------------
  religion_2 |            6
----------------------------
------------------------------------------------------------------------------------
                     |                           Bonferroni            Bonferroni
                income |   Contrast   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------------+-------------------------------------------------------------
           religion_2 |
      HINDU vs MUSLIM  |  -9014.877   3287.767    -2.74   0.040    -17773.41   -256.3402
  Christian vs MUSLIM  |  -8775.289   3747.399    -2.34   0.121    -18758.27    1207.696
   BUDDHISM vs MUSLIM  |  -3384.314   4019.891    -0.84   1.000    -14093.21    7324.585
   Christian vs HINDU  |   239.5876   4477.525     0.05   1.000    -11688.44    12167.61
    BUDDHISM vs HINDU  |   5630.563   4707.946     1.20   1.000    -6911.298    18172.42
BUDDHISM vs Christian  |   5390.976   5039.677     1.07   1.000    -8034.608    18816.56
------------------------------------------------------------------------------------
```

You can also get the error-bar chart of mean income for different religious groups. Use the following three commands consecutively to get the error-bar (Fig 11.2). If you use the command "oneway" instead of “anova” for the ANOVA test, the commands “margins” and “marginsplot” for error-bar will not work.

anova income religion_2
margins religion_2
marginsplot

11.1.4 Interpretation of post hoc test results

Both the commands for multiple comparisons will provide the Bonferroni test results, as shown in Tables 11.2 and 11.3. Each row of Table 11.2 represents a comparison between two specific religious groups. For example, the first row compares the mean income between Hindus and Muslims. We can see that the mean difference in income between Hindus and Muslims is -9,014.87 and the corresponding p-value is 0.040 (<0.05). This indicates that the mean income of Muslims and Hindus is different in the population (Hindus have a significantly lower income than Muslims). The differences in mean income of other religious groups are not significant as the p-values are >0.05. The table has also provided the 95% CIs of mean differences. Table 11.3 has provided the same information except for the 95% CIs of mean differences.

Table 11.3 Multiple comparisons of mean income by religion

```
. oneway income religion_2, bonferroni tabulate

            |      Summary of Monthly income
 Religion 2 |        Mean   Std. Dev.       Freq.
------------+------------------------------------
     MUSLIM |   88180.905   17207.614         126
      HINDU |   79166.028   17804.631          36
  Christian |   79405.615   19857.021          26
   BUDDHISM |   84796.591   14447.348          22
------------+------------------------------------
      Total |   85194.486   17724.033         210

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      3.3068e+09      3   1.1023e+09      3.64     0.0136
 Within groups      6.2349e+10    206    302663565
------------------------------------------------------------------------
    Total           6.5656e+10    209    314141354

Bartlett's test for equal variances:  chi2(3) =   2.2804  Prob>chi2 = 0.516

                  Comparison of Monthly income by Religion 2
                                (Bonferroni)
Row Mean-|
Col Mean |     MUSLIM      HINDU   Christia
---------+--------------------------------
   HINDU |   -9014.88
         |      0.040
         |
Christia |   -8775.29    239.588
         |      0.121      1.000
         |
BUDDHISM |   -3384.31    5630.56    5390.98
         |      1.000      1.000      1.000
```

11.1.5 One-way ANOVA for unequal variances

When the group variances are not homogeneous (i.e., when Bartlett's test p-value is <0.05), we cannot use the F-test for comparison of group means. Instead, we need to use the *Welch test*. Similarly, for the comparison of individual group means (post hoc test), instead of Bonferroni's (or Tukey's) test, we need to use the *Games-Howell test*. There is no straight-forward way to get these tests done in Stata. The users can do these tests easily by using SPSS [19].

11.2 Two-way ANOVA

Two-way ANOVA is similar to one-way ANOVA except that it examines an additional independent categorical variable. The two-way ANOVA involves three variables: one quantitative (dependent variable) and two categorical variables. This test is not com-

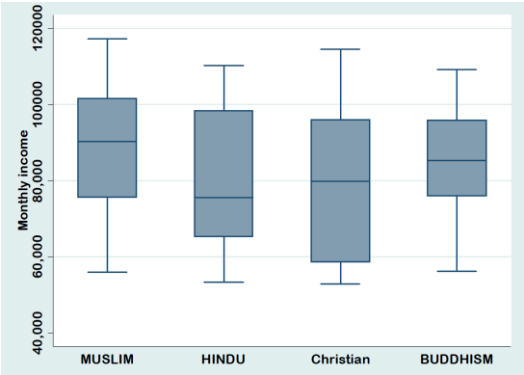


Figure 11.1 Box and plot chart of income by religious groups

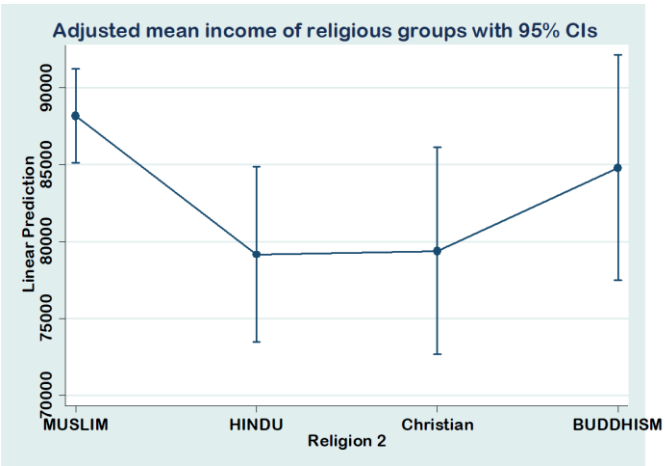


Figure 11.2 Error-bar chart of mean income for religious groups

monly used in health research. Use the data file <**Data_3.dta**> for practice.

Suppose that we want to compare the mean systolic BP (variable name is "sbp") in different occupation and sex groups. Here, the dependent variable is systolic BP and the independent variables are occupation and sex.

Since there are four levels (categories) of occupation (govt. job; private job; business; and others) and two categories of sex (male and female) in the data, we will have a factorial design with eight (4×2) data cells. The two-way ANOVA test answers the following three questions:

1. Does occupation influence the systolic BP (i.e., is the mean systolic BP equal among the occupation groups)?
2. Does sex influence the systolic BP (i.e., is the mean systolic BP equal for males and females)?
3. Does the influence of occupation on systolic BP depend on sex (i.e., is there an interaction between occupation and sex)?

Questions one and two refer to the main effect, while the question three explains the interaction of two independent variables (occupation and sex) on the dependent variable (systolic BP). The primary objective of two-way ANOVA is to assess if there is an interaction between the independent categorical variables on the dependent variable.

Assumptions

1. The dependent variable (systolic BP) is normally distributed at each level of the independent variables (occupation and sex);
2. The variances of the dependent variable (systolic BP) at each level of the independent variables are equal; and
3. Subjects represent random samples from the population.

First, we need to check if the systolic BP is normally distributed in different categories of occupation and sex separately. We can check this by constructing the histograms and Q-Q plots, or by doing the Shapiro Wilk test (Chapter 8). We also need to check the homogeneity of variances in each group of the independent variables (occupation and sex) by using the Levene's test (Section 10.2.1).

11.2.1 Commands for two-way ANOVA

To get the two-way ANOVA test results for systolic BP (variable name is "sbp"), occupation (variable name is "occupation"), and sex (variable name is "sex_1"), use any of the following commands. ANOVA does not allow string variables in the analysis. You need to change the format of a string variable to a numeric variable before including it in the analysis.

anova sbp occupation##sex_1
Or,
anova sbp occupation sex_1 occupation#sex_1

Either of the above commands will generate Table 11.4.

Table 11.4 Two-way ANOVA table

```
. anova sbp occupation##sex_1

                        Number of obs =     210     R-squared     =  0.0386
                        Root MSE      = 20.0049     Adj R-squared =  0.0053

                  Source |  Partial SS    df       MS           F     Prob > F
------------------+----------------------------------------------------
                   Model |  3245.48737     7  463.641053       1.16     0.3283
                         |
              occupation |  470.647297     3  156.882432       0.39     0.7589
                   sex_1 |  1308.03373     1  1308.03373       3.27     0.0721
        occupation#sex_1 |  1735.11998     3  578.373327       1.45     0.2308
                         |
                Residual |  80839.5793   202  400.195937
------------------+----------------------------------------------------
                   Total |  84085.0667   209  402.320893
```

You can get the error graph of the adjusted mean of systolic BP for sex (males and females) by occupation. To get the plot of mean systolic BP with error bars of occupation by sex, use the following commands successively (Fig 11.3):

margins sex_1, at(occupation=(1(1)4))
marginsplot

11.2.2 Interpretation

Table 11.4 shows the outputs of the two-way ANOVA test. The table shows the main effects of the independent variables as well as their interaction. Look at the p-values (Prob > F) for occupation and sex. They are 0.758 and 0.072, respectively. The findings indicate that the mean systolic BP is not different in different occupation groups as well as sex (males and females). Now, look at the p-value for "occupation#sex_1", which indicates the significance of the interaction between these two variables (occupation and sex) on systolic BP. A p-value of <0.05 indicates the presence of interaction. The presence of interaction indicates that the systolic BP in different occupation groups is influenced by (depends on) sex. In our example, the p-value for interaction is 0.230 (>0.05), which means that there is no interaction between occupation and sex to influence the systolic BP.

11.2.3 Post hoc test for two-way ANOVA

The post-hoc test (also discussed under one-way ANOVA) is performed if the main effect is significant (i.e., the p-values for occupation and/or sex are <0.05), otherwise

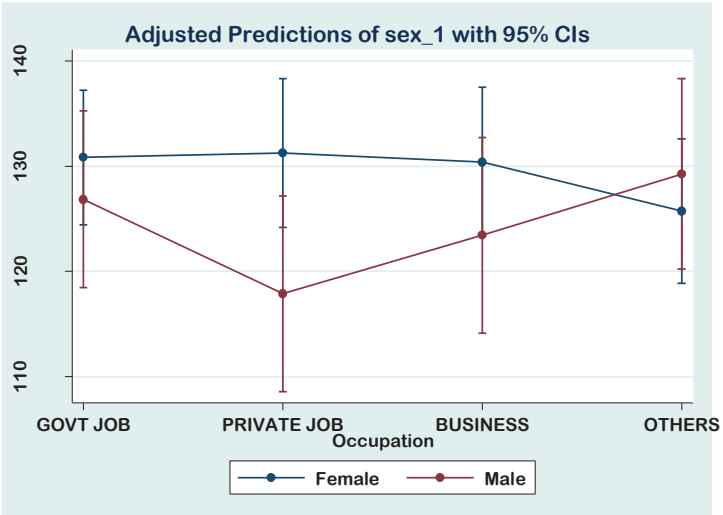


Figure 11.3 Error-bar chart of mean systolic BP of sex by occupation

Table 11.5 Pairwise comparisons table for mean systolic BP by occupation

```
. pwmean sbp, over(occupation) mcompare(bonferroni) effects

Pairwise comparisons of means with equal variances

over         : occupation

----------------------------
             |    Number of
             |  Comparisons
-------------+--------------
  occupation |            6
----------------------------

-------------------------------------------------------------------------------------
                         |                          Bonferroni           Bonferroni
                     sbp |   Contrast   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------------+-----------------------------------------------------------
              occupation |
PRIVATE JOB vs GOVT JOB  |  -3.036395   3.883548    -0.78   1.000    -13.38208    7.309289
   BUSINESS vs GOVT JOB  |   -1.52619   3.883548    -0.39   1.000    -11.87187    8.819493
     OTHERS vs GOVT JOB  |  -2.364103   3.821385    -0.62   1.000    -12.54419     7.81598
BUSINESS vs PRIVATE JOB  |   1.510204   4.074798     0.37   1.000    -9.344964    12.36537
  OTHERS vs PRIVATE JOB  |    .672292   4.015596     0.17   1.000    -10.02517    11.36975
     OTHERS vs BUSINESS  |  -.8379121   4.015596    -0.21   1.000    -11.53537    9.859545
-------------------------------------------------------------------------------------
```

it is not necessary. Our data shows that there is no significant effect of occupation and sex on systolic BP, since the p-values are 0.758 and 0.072, respectively. However, if

you want to perform the multiple comparisons test (say, Bonferroni) for "occupation" after the two-way ANOVA, use the first command, while use the second command to get the same for "sex_1".

pwmean sbp, over(occupation) mcompare(bonferroni) effects
pwmean sbp, over(sex_1) mcompare(bonferroni) effects

The first command will provide Table 11.5. Interpretation of the results is the same as discussed in Section 11.1.4.

12

Repeated Measures ANOVA

Repeated measures design is commonly used in experimental studies. In repeated measures design, measurements of the same variable are made on each subject on two or more different occasions (either at different points in time or under different conditions, such as different treatments). It is similar to a paired t-test, except that there are more than two measurements in the repeated measures ANOVA. The repeated measures ANOVA test compares the means across one or more variables that are based on repeated observations. In this chapter, one-way (within-subjects) repeated measures ANOVA is discussed.

12.1 One-way repeated measures ANOVA

The one-way repeated measures ANOVA test is analogous to the paired samples t-test that we have discussed earlier (Chapter 10). The main difference is that, in a paired samples t-test, we have two measurements on the same subjects at different times (e.g., before and after giving a drug, or pre-test and post-test results), while in a one-way repeated measures ANOVA, there are three or more measurements on the same subjects at different points in time (i.e., the subjects are exposed to multiple measurements over a period of time or conditions). One-way repeated measures ANOVA is also called *one-way within-subjects ANOVA*. Use the data file <**Repeat anova_3.dta**> for practice.

Suppose that we are interested in assessing the mean blood sugar levels at four different time intervals (e.g., at hour-0, hour-7, hour-14, and hour-24) after administration of a drug on 15 study subjects. The objective is to assess whether the drug reduces blood

sugar levels over time (i.e., whether the mean blood sugar levels over time are the same or different).

To conduct this study, we have randomly selected 15 individuals from a population and measured their blood sugar levels at the baseline, i.e., before administration of a drug (hour-0). All the individuals are then administered a drug (say, drug A), and their blood sugar levels are measured again after 7 hours, 14 hours, and 24 hours. We are interested in knowing if the blood sugar levels over time, after giving the drug, are the same or not (in other words, whether the drug is effective in reducing the blood sugar levels over time). The variable "time" in the dataset indicates the times of measurement of blood sugar levels in the subjects. *In this example, we have only one treatment group (received drug A) but have the outcome measurements (blood sugar) at four different points in time on the same subjects (i.e., we have one treatment group with four levels of measurement on the same subjects).*

Hypothesis

$\mathbf{H_0}$: The mean blood sugar levels are equal (same) at each level of measurement (i.e., the mean blood sugar levels at 0, 7, 14, and 24 hours in the population are the same).

$\mathbf{H_A}$: The mean blood sugar levels are not equal at different levels of measurement (i.e., the mean blood sugar levels at 0, 7, 14, and 24 hours in the population are different).

Assumptions

1. The dependent variable (blood sugar level) is normally distributed in the population at each level of within-subjects factor;
2. The population variances of the differences between all combinations of related groups/levels are equal (called the *Sphericity assumption*); and
3. The subjects represent a random sample from the population.

12.1.1 Commands

The data file "**Repeat anova_3.dta**" has the following variables:

Subject: It indicates the study participants, like participants 1, 2, 3, and so on. You will notice that there are 15 subjects enrolled in this study (you can check it by using the command "tab", like "tab subject").

Time: The variable "time" indicates the times of measurement of blood sugar levels. Blood sugar levels were measured on the same subjects at four different times, such as: a) at the baseline, i.e., before treatment (coded as 0); b) 7 hours after treatment (coded as 1); c) 14 hours after treatment (coded as 2); and d) 24 hours after treatment (coded as 3). You can check the times of measurement of blood sugar levels on the subjects by using the following command.

tab time

This will provide Table 12.1. The table shows that there are four categories of the variable "time", and they are: a) before treatment; b) 7 hours after treatment; c) 14 hours after treatment; and d) 24 hours after treatment.

Table 12.1 Frequency distribution of the variable “time”

```
. tab time

   Time of measurement |      Freq.     Percent        Cum.
-----------------------+-----------------------------------
      before treatment |         15       25.00       25.00
 7 hrs after treatment |         15       25.00       50.00
14 hrs after treatment |         15       25.00       75.00
24 hrs after treatment |         15       25.00      100.00
-----------------------+-----------------------------------
                 Total |         60      100.00
```

Sugar: The variable “sugar” indicates the blood sugar levels of the study subjects at different times of measurement.

There is another variable named “treatment” in the dataset that indicates in which treatment group the subjects were enrolled in. This variable is not needed for the analysis of one-way repeated measures ANOVA because we need only one treatment group for the analysis.

To perform the one-way repeated measures ANOVA test, use the following command:

anova sugar subject time, repeated(time)

Once you use this command, the Stata results may show “matsize too small”. If you see this message in the output window, you need to increase the “matsize” by using the following command. Matsize indicates “maximum matrix size”, which influences the number of variables that can be included in any of Stata's estimation commands.

set matsize 10000

The above command will increase the "matsize" to 11,000 in Stata (you can only increase the "matsize" up to 11,000). Now, use the following command to perform the repeated measures ANOVA test:

anova sugar subject time, repeated(time)

While using the command, the computer may take some time to analyze the data, depending on the speed and memory of your computer and the size of the data. It may take 3-7 minutes to calculate the ANOVA test results. The results of the analysis are shown in Table 12.2.

Table 12.2 Repeated measures ANOVA

```
. anova sugar subject time, repeated(time)

Panel 1.
                           Number of obs =      60     R-squared     =  0.7643
                           Root MSE      = 3.94868     Adj R-squared =  0.6689

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  2123.86667    17  124.933333       8.01     0.0000
                         |
                 subject |  1194.73333    14  85.3380952       5.47     0.0000
                    time |  929.133333     3  309.711111      19.86     0.0000
                         |
                Residual |  654.866667    42  15.5920635
              -----------+----------------------------------------------------
                   Total |  2778.73333    59  47.0971751

Between-subjects error term:  subject
                     Levels:  15        (14 df)
     Lowest b.s.e. variable:  subject

Panel 2.

Repeated variable: time
                                          Huynh-Feldt epsilon        =  0.4463
                                          Greenhouse-Geisser epsilon =  0.4232
                                          Box's conservative epsilon =  0.3333

                                            ------------ Prob > F ------------
                  Source |     df      F    Regular    H-F      G-G      Box
              -----------+----------------------------------------------------
                    time |      3    19.86   0.0000   0.0001   0.0001   0.0005
                Residual |     42
              ------------------------------------------------------------
```

12.1.2 Interpretation

Look at the “time” row in Table 12.2 (Panel 1). The p-value (Prob > F) is 0.000. This indicates that the mean blood sugar levels significantly differ at different time points (we have four time points). However, before explaining this table, we need to check whether the sphericity assumption (assumption 2) has been violated or not. To check whether the sphericity assumption is violated or not, we need to do the "Mauchly’s test" for the sphericity assumption. If the assumption is violated (i.e., Mauchly’s test p-value is <0.05), we commonly use the p-value of the Greenhouse-Geisser (G-G) test as shown in the second panel of Table 12.2 to interpret the results of the repeated measures ANOVA test.

If the within-subjects factor has more than two levels, three types of tests are available in repeated measures ANOVA. In our dataset, the within-subjects factor is the times of measurement of blood sugar levels (variable name is "time"), which has four levels — before treatment, 7 hours after treatment, 14 hours after treatment, and 24 hours after treatment (Table 12.1). The tests are:

1. Standard univariate test (when the sphericity assumption is not violated);
2. Alternative univariate tests (Greenhouse-Geisser, Huynh-Feldt, and Lower-bound); and
3. Multivariate tests (Pillai's Trace, Wilks' Lambda, Hotelling's Trace, and Roy's Largest Root).

All these tests evaluate the same hypothesis that the population means are equal at all levels of measurement. The standard univariate test is based on the sphericity assumption, i.e., the standard univariate test result is considered if the *sphericity assumption* is not violated. *In reality, and in most cases, the sphericity assumption is violated, and we cannot use the standard univariate test results as provided in panel 1 of Table 12.2 (time row).* It is, therefore, recommended to use the alternative univariate test, such as the "Greenhouse-Geisser (G-G)" test (or the others), as provided under panel 2 of Table 12.2. The results show that the p-value of the G-G test is 0.0001 (in the time row of panel 2), which is <0.05. This indicates that the mean blood sugar levels differ significantly at different time points. To check the mean of which time points are statistically different, see below (Table 12.5).

However, if you want to check the sphericity assumption, you need to do the Mauchly’s test. To get the Mauchly’s test, you need to install the modules "Mauchly and moremata" (commonly, they are not the built-in commands in Stata). To install the modules, use the following commands:

ssc **install** mauchly
ssc **install** moremata

After installation of the modules, use the following commands to get the Mauchly's test results:

xtset subject
mauchly sugar, m(time)

This will provide the results of Mauchly's test of sphericity (Table 12.3). Our interest is in the p-value of the test. The p-value of the test, as shown in the table, is zero. If the p-value is >0.05, the sphericity assumption is not violated, and you can use the results provided in the first panel of Table 12.2 (in the time row). If the p-value is <0.05 (i.e., the sphericity assumption is violated), use the results of the Greenhouse-Geisser (G-G) or the other test [e.g., Huynh-Feldt (H-F)] as provided under panel 2 of Table 12.2.

Table 12.3 Mauchly's test of sphericity assumption

```
. xtset subject
       panel variable:  subject (balanced)

. mauchly sugar, m(time)

Mauchly's Test of Sphericity
```

Mauchly's W.	Chi2.	d.f.	P-value.	Epsilon_gg.	Epsilon_ff.	Lower-bound
0.0689	34.0332	5	0	0.4232	0.4463	0.3333

To get the means and standard deviations of blood sugar levels at four time points, use the following command:

tabstat sugar, stat(n mean sd) by(time)

This will provide Table 12.4 with the means and standard deviations of blood sugar levels at different time points, including the number of subjects (n). You can get the pairwise comparison of mean blood sugar levels by using the following command (with Bonferroni's test) (Table 12.5):

pwmean sugar, over(time) mcompare(bonferroni) pveffects

Finally, to get the line graph of mean blood sugar levels over time with 95% CI, use the

following commands successively after the primary analysis (Fig 12.1):

anova sugar subject time, repeated(time)
margins time
marginsplot

Table 12.4 Mean and SD of blood sugar levels at different time points

```
. tabstat sugar, stat(n mean sd) by(time)

Summary for variables: sugar
     by categories of: time (Time of measurement)

            time |         N      mean        sd
-----------------+------------------------------
before treatment |        15  110.5333   4.73387
7 hrs after trea |        15     105.2  4.427189
14 hrs after tre |        15  101.5333  6.300416
24 hrs after tre |        15  100.4667  7.099966
-----------------+------------------------------
           Total |        60  104.4333  6.862738
--------------------------------------------------------------------------------
```

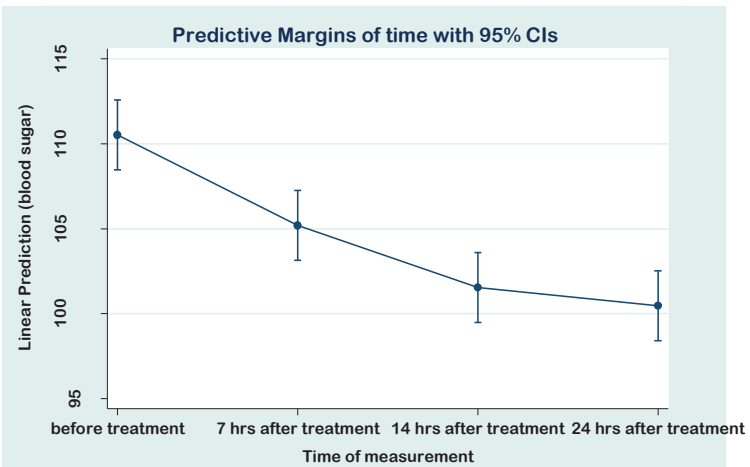


Figure 12.1 Mean blood sugar levels with 95% CIs at different time points

Table 12.5 Pairwise comparison of mean blood sugar levels

```
. pwmean sugar, over(time) mcompare(bonferroni) pveffects

Pairwise comparisons of means with equal variances

over         : time

---------------------------
             |    Number of
             |  Comparisons
-------------+-------------
        time |            6
---------------------------

-----------------------------------------------------------------------------------------
                                                |                           Bonferroni
                                          sugar |   Contrast   Std. Err.      t    P>|t|
------------------------------------------------+----------------------------------------
                                           time |
       7 hrs after treatment vs before treatment  |  -5.333333   2.098526    -2.54   0.083
      14 hrs after treatment vs before treatment  |         -9   2.098526    -4.29   0.000
      24 hrs after treatment vs before treatment  |  -10.06667   2.098526    -4.80   0.000
 14 hrs after treatment vs 7 hrs after treatment  |  -3.666667   2.098526    -1.75   0.516
 24 hrs after treatment vs 7 hrs after treatment  |  -4.733333   2.098526    -2.26   0.168
24 hrs after treatment vs 14 hrs after treatment  |  -1.066667   2.098526    -0.51   1.000
-----------------------------------------------------------------------------------------
```

13

Association Between Two Categorical Variables: Chi-square Test of Independence

The chi-square test is a commonly used statistical test for testing a hypothesis in health and social sciences research. The chi-square value ranges from 0 to ∞ (infinity), and it does not take any negative value. This test is suitable to determine the association between two categorical variables, whether the data are from cross-sectional, case-control, or cohort studies. On the other hand, in epidemiology, cross-tabulations are commonly done to calculate the odds ratio (OR) [for a case-control study] and relative risk (RR) [for a cohort study] with their 95% confidence intervals (CI). The OR and RR are the measures of strength of association between two variables. In this chapter, we have discussed the chi-square and the Fisher's exact tests for hypothesis testing and how to get the RR and OR using Stata. We have also demonstrated how to perform a stratified analysis in this chapter. Use the data file <**Data_3.dta**> for practice.

13.1 Chi-square test of independence

The Chi-square test of independence is used to determine the association between two categorical variables. Suppose that you have collected data on gender (sex) and diabetes from a group of individuals selected randomly from a population. You are interested in knowing if there is an association between gender and diabetes. In such a situation, the chi-square test is the appropriate test for testing the hypothesis.

Hypothesis

H_0: There is no association between gender and diabetes (it can also be stated as gender and diabetes are independent).

H_A: There is an association between gender and diabetes (or, gender and diabetes are not independent).

Assumption

1. The data is a random sample drawn from a population.

13.1.1 Commands

The basic command to get the chi-square test results is to generate a cross-table using the command "tab" with the option of chi-square statistics. Use any of the following commands to find an association (chi-square test) between sex and diabetes.

tab sex diabetes, chi2
tab sex diabetes, row col chi2

The first command will generate a cross-table of sex (the first variable is placed on the row) and diabetes with only the observed frequencies and the chi-square test results. The second command will provide the row and column percentages in the cross-table, including the observed frequencies and chi-square test results (Table 13.1).

You can also use the option “all” to get all the relevant statistics (chi-square, Cramer’s V, and others) for the association between two categorical variables (Table 13.2), such as:

tab sex diabetes, all
tab sex diabetes, col row all
tab sex diabetes, expected all

The last command will provide all the relevant statistics with the expected cell values. The chi-square test is valid if no more than 20% of cells have expected values of less than 5. If the expected value is less than 5 in more than 20% of the cells, we need to use the *Fisher’s exact test* instead of the chi-square test. The command for the Fisher’s exact test is (Table 13.3):

tab sex diabetes, exact
tab sex diabetes, expected exact

Table 13.1 Chi-square test results with row and column percentages

```
. tab sex diabetes, row col chi2

           |     Have diabetes
      Sex: |       mellitus
    string |       Yes         No |     Total
-----------+----------------------+----------
         f |        20        113 |       133
           |     15.04      84.96 |    100.00
           |     44.44      68.48 |     63.33
-----------+----------------------+----------
         m |        25         52 |        77
           |     32.47      67.53 |    100.00
           |     55.56      31.52 |     36.67
-----------+----------------------+----------
     Total |        45        165 |       210
           |     21.43      78.57 |    100.00
           |    100.00     100.00 |    100.00

          Pearson chi2(1) =   8.7995   Pr = 0.003
```

Table 13.2 Cross tabulation with chi-square and other test results

```
. tab sex diabetes, col row all

           |     Have diabetes
      Sex: |       mellitus
    string |       Yes         No |     Total
-----------+----------------------+----------
         f |        20        113 |       133
           |     15.04      84.96 |    100.00
           |     44.44      68.48 |     63.33
-----------+----------------------+----------
         m |        25         52 |        77
           |     32.47      67.53 |    100.00
           |     55.56      31.52 |     36.67
-----------+----------------------+----------
     Total |        45        165 |       210
           |     21.43      78.57 |    100.00
           |    100.00     100.00 |    100.00

          Pearson chi2(1) =   8.7995   Pr = 0.003
 likelihood-ratio chi2(1) =   8.5367   Pr = 0.003
               Cramér's V =  -0.2047
                    gamma =  -0.4618  ASE = 0.135
          Kendall's tau-b =  -0.2047  ASE = 0.071
```

13.1.2 Interpretation

Table 13.1 is a 2 by 2 table of sex and diabetes with row and column percentages and the chi-square test results. The question is, which percentage should you report? It depends on the situation/study design and what you want to report. For the data of a *cross-sectional study*, it may provide better information to the readers if row percentages are reported. In that case, the row percentages indicate the prevalence of the condition (in this example, diabetes).

Table 13.3 Fisher's exact test results with expected cell values

```
. tab sex diabetes, expected exact

           |     Have diabetes
      Sex: |       mellitus
    string |       Yes         No |     Total
-----------+----------------------+----------
         f |        20        113 |       133
           |      28.5      104.5 |     133.0
-----------+----------------------+----------
         m |        25         52 |        77
           |      16.5       60.5 |      77.0
-----------+----------------------+----------
     Total |        45        165 |       210
           |      45.0      165.0 |     210.0

           Fisher's exact =                 0.005
   1-sided Fisher's exact =                 0.003
```

For example, one can understand from Table 13.1 that the prevalence of diabetes among males is 32.47% and that of females is 15.04% when row percentages are used. However, the column percentages can also be reported for a cross-sectional data (most of the publications use column percentages). If column percentages are used, the meaning will be different. In our example (Table 13.1), the results indicate that 55.56% of the diabetic patients are males, compared to 31.52% of the non-diabetic individuals. If the data is from a *case-control study*, you must report column percentages (you cannot use row percentages for case-control studies). On the other hand, for the data of a cohort study, one should report the row percentages. In such a situation, it would indicate the incidence (instead of the prevalence) of the disease among males and females.

We can see in Table 13.1 (in the row of "Total") that the overall (irrespective of gender) prevalence of diabetes is 21.43% (considering the data is from a cross-sectional study).

Table 13.1 also shows that 32.47% of males have diabetes compared to only 15.04% of females (i.e., the prevalence among males and females). The chi-square test actually tests the hypothesis of whether the prevalence of diabetes among males and females in the population is the same or not.

Now, look at the Pearson's chi-square test results provided at the bottom of the table [(Pearson chi2(1) = 8.7995 Pr = 0.003)]. The "Pearson chi2(1)" indicates the Pearson's chi-square test result with the degree of freedom (df) of 1, while "Pr" indicates the p-value.

Before we conclude the chi-square test results, it is important to check if there is any cell in the 2 by 2 table that has an expected value of less than 5. This can be checked using the option "expected". Table 13.3 displays the expected cell values and the Fisher's exact test p-value. The table shows that there is no cell in the 2 by 2 table with an expected value of less than 5 (the minimum expected value found is 16.5). For the use of the chi-square test results, it is desirable to have no cell in a 2 by 2 table with an expected value of less than 5. If this is not fulfilled, we need to use the *Fisher's exact test p-value* to interpret the results.

Since we do not have the problem of an expected value of less than 5 in the 2 by 2 table, we will consider the chi-square test results given at the bottom of Tables 13.1 and 13.2 for conclusion. The results show that the calculated chi-square value is 8.79 (df= 1) and the corresponding p-value is 0.003. Since the p-value is <0.05, we will reject the null hypothesis. This indicates that there is a significant association between gender and diabetes. It can also be concluded that the prevalence of diabetes among males is significantly higher than that of females, which is statistically significant at 95% confidence level (p=0.003).

13.2 Relative risk and odds ratio

We calculate relative risk (RR) for the data of cohort studies and odds ratio (OR) for case-control studies (OR is also sometimes calculated for cohort studies and cross-sectional studies). To calculate the RR and OR with their confidence intervals (CI), we need to use the "epidemiology and related" option of data analysis. We also need to recode the outcome and exposure variables as 0/1 (0 for no disease/unexposed; 1 for have disease/exposed) if they are not coded like this.

We want to calculate the RR from the cross-tabulation between diabetes (outcome variable) and sex_1 (exposure variable). Since the variable "diabetes" (in our dataset)

is coded as 1/2 (1= have diabetes and 2= don't have diabetes), we need to change the coding structure of the variable "diabetes" as 0= don't have diabetes and 1= have diabetes using the command "recode" as discussed in Section 5.2. There is no problem with the variable "sex_1" in the dataset since it is already coded as 0= female and 1= male. Once it is done (there is a variable in the dataset "diabetes1", which is coded as 0/1), use the following command to get the RR and OR with their 95% CIs:

cs diabetes1 sex_1
cs diabetes1 sex_1, or
cs diabetes1 sex_1, or level(99)

The command "cs" indicates "cohort study". This command will automatically provide the RR and its 95% CI (by default) along with other statistics (Table 13.4). If you want to get the OR, in addition to RR, use the second command. The third command is for getting the 99% CIs for both the RR and OR.

Table 13.4 Relative risk and 95% confidence interval

```
. cs diabetes1 sex_1

                 | Sex: numeric           |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        25          20  |         45
        Noncases |        52         113  |        165
-----------------+------------------------+------------
           Total |        77         133  |        210
                 |                        |
            Risk |  .3246753    .1503759  |   .2142857
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |         .1742994       |    .0533492    .2952495
      Risk ratio |         2.159091       |    1.287892    3.619616
 Attr. frac. ex. |         .5368421       |    .2235371    .7237276
 Attr. frac. pop |         .2982456       |
                 +-------------------------------------------------
                                  chi2(1) =     8.80  Pr>chi2 = 0.0030
```

Instead of using the "cs" command, you can use the following commands to get the OR without a 2 by 2 table:

tabodds diabetes1 sex_1, or
tabodds diabetes1 sex_1, base(2) or

The first command will provide the OR without the 2 by 2 table, and with the first

category of sex (females; indicated by Odds Ratio = 1.0) as the comparison group (Table 13.5). The second command will provide the same but the second category of sex (males) as the comparison group (Table 13.5).

Table 13.5 Odds ratio with 95% CI and p-value

```
tabodds diabetes1 sex_1, or

--------------------------------------------------------------------------
       sex_1 |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+------------------------------------------------------------
      Female |    1.000000          .           .              .          .
        Male |    2.716346       8.76       0.0031      1.362845   5.414068
--------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(1)  =     8.76
                                  Pr>chi2  =   0.0031

Score test for trend of odds:     chi2(1)  =     8.76
                                  Pr>chi2  =   0.0031

tabodds diabetes1 sex_1, base(2) or

--------------------------------------------------------------------------
       sex_1 |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+------------------------------------------------------------
      Female |    0.368142       8.76       0.0031      0.184704   0.733759
        Male |    1.000000          .           .              .          .
--------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(1)  =     8.76
                                  Pr>chi2  =   0.0031

Score test for trend of odds:     chi2(1)  =     8.76
                                  Pr>chi2  =   0.0031
```

If you want to calculate the OR with 95% CI (default) for a case-control study, use the first command (Table 13.6) as shown below. Use the second command if you want to get the 99% CI.

cc diabetes1 sex_1
cc diabetes1 sex_1, level(99)

Another way of getting the OR is to use the command "logistic", which is used for logistic regression analysis. The command is:

logistic diabetes1 sex_1

The results of the above command (called univariate logistic regression analysis) are shown in Table 13.7.

Table 13.6 Odds ratio and 95% confidence interval

```
. cc diabetes1 sex_1
                                                        Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
           Cases |        25          20  |         45      0.5556
        Controls |        52         113  |        165      0.3152
-----------------+------------------------+------------------------
           Total |        77         133  |        210      0.3667
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |         2.716346       |    1.311301    5.641061 (exact)
 Attr. frac. ex. |         .6318584       |    .2373985    .8227284 (exact)
 Attr. frac. pop |         .3510324       |
                 +-------------------------------------------------
                                  chi2(1) =     8.80  Pr>chi2 = 0.0030
```

Table 13.7 Odds ratio by using the "logistic" command

```
. logistic diabetes1 sex_1

Logistic regression                               Number of obs   =        210
                                                  LR chi2(1)      =       8.54
                                                  Prob > chi2     =     0.0035
Log likelihood = -104.84341                       Pseudo R2       =     0.0391

------------------------------------------------------------------------------
   diabetes1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       sex_1 |   2.716346   .9334131     2.91   0.004     1.385123    5.326991
       _cons |   .1769912   .0429362    -7.14   0.000     .1100168     .284737
------------------------------------------------------------------------------
```

In the above examples, both the outcome (dependent variable) and exposure (independent variable; sex) variables are dichotomous variables with the coding schemes of 0/1. If the exposure variable has more than two levels/categories (e.g., the variable "religion" has 3 categories in our dataset; 1= Muslim, 2= Hindu, and 3= Christian), the commands "cc" or "cs" will not calculate the ORs. In that case (when the exposure variable has more than two categories), use the command "logistic" to get the ORs, such as (outputs are provided in Table 13.8):

logistic diabetes1 i.religion

The prefix "i." used for the variable "religion" (i.e., i.religion) is necessary when it is a categorical variable with more than two levels. When the "i." prefix is used before the exposure variable, Stata considers it a categorical variable. If the prefix "i." is not used,

Stata will consider the variable as a continuous variable and the results will be misleading. With the above command, Stata will consider the first category (Muslims) as the comparison group by default. If you want the other category of religion as the comparison group (say, last category or Christians), use any of the following commands:

logistic diabetes1 ib3.religion
tabodds diabetes1 religion, base(3) or

Table 13.8 Odds ratios for an exposure variable with more than 2 levels

```
. logistic diabetes1 i.religion

Logistic regression                               Number of obs   =        210
                                                  LR chi2(2)      =       3.01
                                                  Prob > chi2     =     0.2215
Log likelihood = -107.60445                       Pseudo R2       =     0.0138

------------------------------------------------------------------------------
   diabetes1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    religion |
      HINDU  |   1.465201   .5378927     1.04   0.298     .7135283    3.008732
  Christian  |   .5016722   .3271554    -1.06   0.290     .1397418    1.801001
             |
       _cons |        .26   .0572364    -6.12   0.000     .1688846    .4002734
------------------------------------------------------------------------------
```

13.2.1 Interpretation

Table 13.4 shows the cross-tabulation of diabetes and sex. The table shows that of the 45 cases of diabetes (cases), 25 are males (males are exposed since they are coded as 1). On the other hand, of the 165 non-cases (who do not have diabetes), 52 are males. The RR (also called risk ratio) calculated from the data is 2.15, and the 95% CI is 1.28 - 3.61. Stata has also provided the chi-square test results at the bottom of the table, including the p-value of the test (0.003). From the data, we can conclude that males are at 2.15 times higher risk of having diabetes compared to females, which is statistically significant at 95% confidence level (RR: 2.15; 95% CI of RR: 1.28 - 3.61; p=0.003).

Table 13.6 shows the OR (of being male) with the 95% CI. The OR, as calculated, is 2.71 and its 95% CI is 1.31 - 5.64. The analysis also provided the p-value of the chi-square test (Pr>chi2), which is 0.003. We can, therefore, conclude that males are more likely to have diabetes compared to females, which is statistically significant at 95% confidence level (OR: 2.71; 95% CI of OR: 1.31 - 5.64; p=0.003).

Table 13.7 shows the analysis of the data using the "logistic" command (univariate logistic regression analysis). The table shows the OR (2.71), the 95% CI of OR (1.38 - 5.32) and the p-value (0.004). This is the OR for males compared to females (lower value is considered as the comparison group).

Table 13.8 shows the association between diabetes and religion, which has three levels. Note that Stata considers the lowest value/code (Muslims) of the exposure variable as the comparison group during such an analysis (by default). The table shows the ORs, 95% CIs, and p-values for Hindus and Christians. The OR of diabetes for being a Hindu is 1.46 (95% CI: 0.71 - 3.00; p=0.298) compared to Muslims, which is not statistically significant since the p-value is greater than 0.05. Similarly, the OR of diabetes for being a Christian is 0.50 (95% CI: 0.13 - 1.80; p=0.290) compared to Muslims, which is also not statistically significant.

13.3 Stratified analysis

The stratified analysis is a statistical method that allows us to test for confounding and interaction effects. It also provides the adjusted RR (or OR) for the association between an exposure and an outcome after controlling for one or more variables.

For example, you may be interested in estimating the adjusted value of RR (or OR) for the association between sex and diabetes after controlling (adjusting) for the variable “family history of diabetes” (the variable name is “f_history”). To get the adjusted RR (or OR), we will do the stratified analysis using the following commands (Tables 13.9 and 13.10):

cs diabetes1 sex_1, by(f_history)
cc diabetes1 sex_1, by(f_history)

The first command will provide the adjusted RR (Table 13.9), while the second command will provide the adjusted OR (Table 13.10). From these tables, we can also get information on whether the family history of diabetes is a confounding factor in the relationship between sex and diabetes, and whether or not there is an interaction between sex and family history of diabetes on the outcome variable (diabetes).

Table 13.9 Results of stratified analysis (RR)

```
. cs diabetes1 sex_1, by(f_history)

Family history o |       RR       [95% Conf. Interval]   M-H Weight
-----------------+-------------------------------------------------
             Yes |     2.907692     1.330149   6.356186     3.554688
              No |     1.794872     .7021495   4.588146     1.902439
-----------------+-------------------------------------------------
           Crude |     2.159091     1.287892   3.619616
    M-H combined |     2.519746      1.36505     4.6512
-----------------------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =    0.631  Pr>chi2 = 0.4271
```

Table 13.10 Results of stratified analysis (OR)

```
. cc diabetes1 sex_1, by(f_history)

Family history o |       OR       [95% Conf. Interval]   M-H Weight
-----------------+-------------------------------------------------
             Yes |    3.818182      1.390158   11.51741      2.40625 (exact)
              No |    2.192308      .4142173   9.721011     1.268293 (exact)
-----------------+-------------------------------------------------
           Crude |    2.716346      1.311301   5.641061              (exact)
    M-H combined |    3.257001      1.520939   6.974675
-----------------------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     0.44  Pr>chi2 = 0.5056

                    Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =      9.53
                                                Pr>chi2 =    0.0020
```

13.3.1 Interpretation

The stratified analysis in epidemiology is done to estimate the strength of association (RR or OR) between an exposure (e.g., sex) and an outcome (diabetes) after controlling (adjusting) for a third categorical variable (e.g., family history of diabetes). It also enables us to examine whether: a) the third variable is a confounding factor, or b) there is an interaction (also called effect modification) between exposure and the third factor.

The stratified analysis can be done for the data of cross-sectional, cohort, or case-control studies. It is suitable to adjust for a single stratified variable, though more than one variable can be used in the analysis.

In our example, we have examined the relationship between sex (exposure) and diabetes (outcome) at two levels of the categorical stratified variable "family history of diabetes".

Table 13.9 shows the results of stratified analysis considering that the data is from a cohort study. The results show that among those who have a family history of diabetes, the RR for males (compared to females) is 2.90 (95% CI: 1.33 - 6.35), while the RR is 1.79 (95% CI: 0.70 - 4.58) when there is no family history of diabetes. The table also shows the crude (Crude) and adjusted RR (M-H combined). The crude (or unadjusted) RR is calculated without considering the third (stratified) variable (i.e., family history of diabetes). The crude RR, as calculated by Stata, is 2.15 (95% CI: 1.28 - 3.61), while the adjusted RR (M-H combined) is 2.51 (95% CI: 1.36 - 4.65). The adjusted RR indicates the RR after controlling for the family history of diabetes.

In our data (Table 13.9), we can see that there is a difference between the crude (2.15) and adjusted (2.51) RR. This (i.e., when there is a difference between crude and adjusted RR) may indicate that the family history of diabetes has some confounding effect (influence) on the relationship between sex and diabetes. Though there is no set rule for deciding what amount of difference is considered significant, in general, more than 20% change is considered important (in our example, it is less than 20%). However, if the crude (unadjusted) and adjusted RRs (or ORs) are close together, the third variable (stratified variable) is not a confounding factor in the relationship between an exposure and an outcome.

Now, the question is whether there is an interaction between sex and family history of diabetes on the outcome. *If there is an interaction, the RRs (or ORs) in two strata of the third variable will be different*. In our example, for males, if there is a family history of diabetes, the RR is 2.90, while the RR is 1.79 if there is no family history of diabetes, and they are different. This indicates that there may have an interaction between sex and family history of diabetes on the outcome. However, before we conclude like this, we need to check the statistical significance of the difference. Stata has provided the statistical test (Test of homogeneity (M-H); chi2(1) = 0.631; Pr>chi2 = 0.4271) to understand the significance of the difference at the bottom of the table (Table 13.9). It shows that the p-value (Pr>chi2 = 0.4271) is 0.427. Since the p-value is >0.05, the difference in RRs in the two strata is not statistically significant. We may, therefore, conclude that there is no interaction between sex and family history of diabetes on outcome even though the RRs are different in two strata of family history of diabetes. If the homogeneity test is significant (less than 0.05), it is not appropriate to report the adjusted RR (or OR). The results should be presented for each stratum separately.

Table 13.10 has provided similar information, where we have calculated the OR (instead of RR) considering that the data is from a case-control study. We can see that

the crude and adjusted ORs are 2.71 and 3.25, respectively. Since the ORs are substantially different (exactly 20%), the family history of diabetes is a confounding factor in the relationship between sex and diabetes. The p-value for the adjusted OR is provided at the bottom of the table (Mantel-Haenszel chi2(1) = 9.53; Pr>chi2 = 0.0020). It shows that the p-value is 0.002. We, therefore, conclude that males are 3.25 (M-H combined OR) times more likely to have diabetes compared to females after controlling for family history of diabetes, which is statistically significant (95% CI: 1.42 – 6.97; p=0.002).

On the other hand, the ORs of diabetes for males with and without a family history of diabetes are 3.81 and 2.19, respectively. Though they are different, the difference is not statistically significant (p=0.505) as shown in the table [Test of homogeneity (M-H); chi2(1) = 0.44; Pr>chi2 = 0.5056]. Therefore, there is no interaction between sex and family history of diabetes on the outcome variable.

14

Hypothesis Test of Proportions

Data is frequently collected in health and social sciences research to estimate the proportions. We may have a situation where there is a single group of individuals and a certain proportion of them have a particular characteristic (e.g., a disease). For example, a researcher has collected data from a population by taking a random sample and found that a certain percentage (proportion) of individuals have diabetes. The researcher may be interested in testing the null hypothesis that the population proportion is equal to a pre-specified value/proportion (one-sample test of proportion).

On the other hand, for the comparison of proportions of two independent samples or the proportions of two groups of individuals, a two-sample test of proportions is used. In this chapter, we have discussed how to test the null hypothesis for one-sample and two-sample proportions. Use the data file "**Data_3.dat**".

14.1 One-sample test of proportion

There is a variable "diabetes1" in the dataset. In the variable, there are individuals who have diabetes (coded as 1) and those who do not have diabetes (coded as 0) [*the variable's coding scheme must be 0/1, otherwise the command for the statistical test of proportion will not work*]. Let us assume that the data is a random sample from a district. We can calculate the prevalence of diabetes from the data by using the following command:

tab diabetes1

This command will provide Table 14.1. The table shows that there are 45 individuals

who have diabetes out of 210. Therefore, the prevalence of diabetes is 21.43%.

We are interested in testing the hypothesis of whether or not the prevalence of diabetes in the district is different from the national prevalence of 19.0%. Here, the null hypothesis is "the prevalence of diabetes in the district is not different from 19.0%", while the alternative hypothesis is "the prevalence of diabetes in the district is different from 19.0%". This is a situation where we can apply the one-sample test of proportion.

Table 14.1 Prevalence of diabetes

```
. tab diabetes1

       have |
diabetes 01 |      Freq.     Percent        Cum.
------------+-----------------------------------
         no |        165       78.57       78.57
        yes |         45       21.43      100.00
------------+-----------------------------------
      Total |        210      100.00
```

Table 14.2 One-sample test of proportion

```
. prtest diabetes1==.19

One-sample test of proportion              diabetes1: Number of obs =      210
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.                     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   diabetes1 |   .2142857   .0283152                       .158789    .2697824
------------------------------------------------------------------------------
    p = proportion(diabetes1)                                         z =   0.8971
Ho: p = 0.19

    Ha: p < 0.19                 Ha: p != 0.19                 Ha: p > 0.19
 Pr(Z < z) = 0.8152         Pr(|Z| > |z|) = 0.3697          Pr(Z > z) = 0.1848
```

The one-sample test of proportion is analogous to the one-sample t-test. The one-sample proportion test is used to compare the observed proportion with a hypothetical value. To do the one-sample test of proportion, use the following command:

prtest diabetes1==.19

The results are shown in Table 14.2. The hypothesis that we have tested is a two-tailed hypothesis. The p-value of the two-tailed test is provided under "Ha: p != 0.19", which is 0.369 [Pr(|Z| > |z|) = 0.3697]. Since the p-value is >0.05, we cannot reject the null

hypothesis. It can, therefore, be concluded that the prevalence of diabetes in the district may not be different from the national prevalence of 19.0%.

14.2 Two-sample test of proportions

In section 14.1, we analyzed the variable "diabetes1" and found that the overall prevalence of diabetes is 21.43%. There is another variable in the dataset named "sex_1", which is coded as "0" for females and "1" for males. For the analysis of two-sample proportions, the categorical variable must be a numeric variable. If it is a string variable, the command will not be executed. From the data, we can calculate the prevalence of diabetes among males and females by generating a cross-table. To generate a cross-table of sex (variable name: sex_1) and diabetes (variable name: diabetes1) with row percentages, use the following command:

tab sex_1 diabetes1, row

The above command will generate Table 14.3. The table shows that the prevalence of diabetes among females and males is 15.04% and 32.47%, respectively.

Table 14.3 Cross-tabulation of sex and diabetes

```
. tab sex_1 diabetes1, row

+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

      Sex: |   have diabetes 01
   numeric |        no        yes |     Total
-----------+----------------------+----------
    Female |       113         20 |       133
           |     84.96      15.04 |    100.00
-----------+----------------------+----------
      Male |        52         25 |        77
           |     67.53      32.47 |    100.00
-----------+----------------------+----------
     Total |       165         45 |       210
           |     78.57      21.43 |    100.00
```

Suppose that we are interested in examining whether the same proportion of males and females has diabetes in the population, i.e., whether the prevalence of diabetes is the

same among females and males. This is a situation where we can apply the two-sample test of proportions. For this example, the null hypothesis is "the proportion of males who have diabetes is equal to the proportion of females who have diabetes in the population". The alternative hypothesis is that the two proportions are not the same (different) in the population.

To test the null hypothesis that the two proportions are the same in the population, use the following command:

prtest diabetes1, by (sex_1)

Table 14.4 Two-sample test of proportions

```
. prtest diabetes1, by (sex_1)

Two-sample test of proportions                Female: Number of obs =      133
                                                Male: Number of obs =       77
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Female |   .1503759   .0309939                      .0896289    .2111229
        Male |   .3246753   .0533624                      .2200869    .4292638
-------------+----------------------------------------------------------------
        diff |  -.1742994   .0617104                     -.2952495   -.0533492
             |  under Ho:   .0587581    -2.97   0.003
------------------------------------------------------------------------------
        diff = prop(Female) - prop(Male)                          z =  -2.9664
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.0015         Pr(|Z| < |z|) = 0.0030          Pr(Z > z) = 0.9985
```

The results are shown in Table 14.4. The table shows that the difference between the two proportions (female to male) is -.174 (-17.4%). This means that the prevalence of diabetes among females is 17.4% less than that of males (15.03% vs. 32.46%). The 95% CI of the difference is also given in the table, which is -.295 (29.5%) to -.053 (5.3%) ["diff" row in the table]. Our interest is in the two-sided p-value of the test, which is 0.003 ($Pr(|Z| < |z|) = 0.0030$). Since the p-value is <0.05, we will reject the null hypothesis and conclude that the proportion of males who have diabetes is different from that of females (i.e., the prevalence of diabetes among males is significantly higher than that of females). The chi-square test of independence also tests the same hypothesis.

15

Association Between Two Continuous Variables: Correlation

The nature and strength of the relationships between two or more continuous variables can be examined by regression and correlation analysis. Correlation is concerned with measuring the strength of a relationship between continuous variables. The correlation model provides information about the relationship between two variables without distinguishing which one is the dependent and which one is the independent variable. But the basic procedure for regression and correlation models is the same.

Under the correlation model, we calculate the "r" value. The "r" is called the sample correlation coefficient. It indicates the degree of linear relationship between the dependent (Y) and independent (X) variables. The value of "r" ranges between "–1" and "+1". This chapter will cover the correlation model. Use the data file <**Data_3.dta**> for practice.

15.1 Pearson's correlation

Pearson's correlation is used when the normality assumptions are met, i.e., when both the variables involved in the correlation analysis are normally distributed. Suppose that we want to explore if there is a correlation between systolic blood pressure (BP) (variable name is "sbp") and diastolic BP (variable name is "dbp").

Hypothesis

$\mathbf{H_0}$: There is no correlation between systolic and diastolic BP.

$\mathbf{H_A}$: There is a correlation between systolic and diastolic BP.

Assumptions

1. The variables (systolic and diastolic BP) are normally distributed in the population;
2. The subjects represent a random sample from the population.

15.1.1 Scatter plot

The first step, before carrying out the correlation analysis, is to generate a scatter plot. The scatter plot provides information about:

- Whether there is a correlation (relationship) between two variables;
- Whether the relationship (if there is any) is linear or non-linear; and
- Direction of the relationship, i.e., whether the relationship is positive (if the value of one variable increases with the increase of the other variable) or negative (if the value of one variable decreases with the increase of the other variable).

To generate a scatter plot of systolic and diastolic BP, use the following commands (also see Section 7.2):

scatter dbp sbp
scatter dbp sbp || lfit dbp sbp

The first command will display a simple scatter plot of systolic and diastolic BP (Fig 15.1), while the second command will display the regression line (fit line) on the scatter plot (Fig 15.2).

15.1.2 Commands for Pearson's correlation

To determine the Pearson's correlation coefficient between two continuous quantitative variables, we use the command "correlate" or simply "corr" or "pwcorr". The commands "correlate" or "corr" provide correlations of the listed variables based on the non-missing observations. On the other hand, the command "pwcorr (pairwise correlation)" reports on all the observations available for each variable pair.

corr sbp dbp
corr sbp dbp age income
pwcorr sbp dbp, sig
pwcorr sbp dbp age income, sig

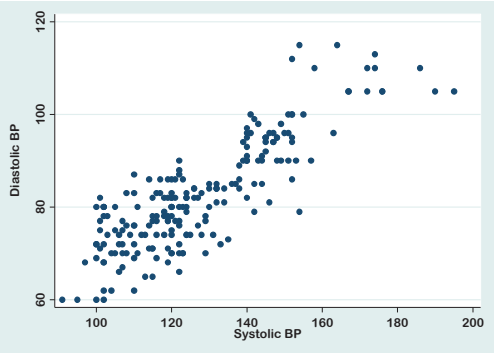


Figure 15.1 Scatter plot of systolic and diastolic BP

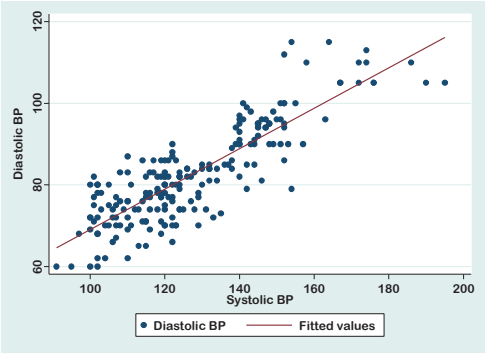


Figure 15.2 Scatter plot of systolic and diastolic BP with the regression line

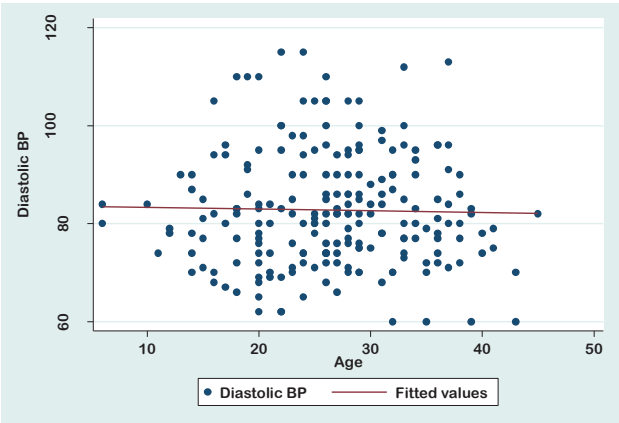


Figure 15.3 Scatter diagram of diastolic BP and age

The first command will provide the Pearson's correlation coefficient (r value) of systolic and diastolic BP without the p-value (Table 15.1). The second command will generate the correlation matrix of the variables (sbp, dbp, age, and income) included in the command (Table 15.2). If you want to obtain the correlation coefficient of systolic and diastolic BP along with the p-value, use the third command (Table 15.3). The fourth command will provide a correlation matrix table for the variables included in the command with the corresponding p-values (Table 15.4). The commands for the correlation matrix are provided as an example.

Table 15.1 Pearson's correlation between sbp and dbp

```
. corr sbp dbp
(obs=210)
             |      sbp      dbp
-------------+------------------
         sbp |   1.0000
         dbp |   0.8468   1.0000
```

15.1.3 Interpretation

In the first step, we have constructed the scatter plots of systolic and diastolic BP (Figs 15.1 and 15.2). Figure 15.1 shows that the data points are scattered around an invisible straight line (indicating linear relationship) and there is an increase in the diastolic BP (Y) as the systolic BP increases (X). This indicates that there may have a positive correlation between these two variables.

Table 15.2 Correlation matrix of systolic BP, diastolic BP, age and income

```
. corr sbp dbp age income
(obs=210)
             |      sbp      dbp      age   income
-------------+------------------------------------
         sbp |   1.0000
         dbp |   0.8468   1.0000
         age |  -0.0378  -0.0224   1.0000
      income |   0.0163   0.0697  -0.1324   1.0000
```

Table 15.3 Pearson's correlation between systolic and diastolic BP with the p-value

```
. pwcorr sbp dbp, sig

             |      sbp      dbp
-------------+------------------
         sbp |   1.0000
             |
         dbp |   0.8468   1.0000
             |   0.0000
```

Look at Fig 15.2, which shows the regression line on the scatter plot. The regression line has passed from near the lower left corner to the upper right corner, indicating a positive correlation between systolic and diastolic BP. If the relationship was negative (inverse), the regression line would have passed from the upper left corner to the lower right corner.

Figure 15.3 shows the scatter plot of diastolic BP and age. It does not indicate any correlation between diastolic BP and age since the dots are scattered around the regression line, which is more or less parallel to the X-axis.

Table 15.4 Correlation matrix with the corresponding p-values

```
. pwcorr sbp dbp age income, sig

             |      sbp      dbp      age   income
-------------+------------------------------------
         sbp |   1.0000
             |
         dbp |   0.8468   1.0000
             |   0.0000
             |
         age |  -0.0378  -0.0224   1.0000
             |   0.5856   0.7465
             |
      income |   0.0163   0.0697  -0.1324   1.0000
             |   0.8144   0.3146   0.0554
```

Table 15.1 shows the Pearson's correlation coefficient (r value) of systolic and diastolic BP, which is 0.846. Table 15.3 shows the same result, but with the p-value, which is 0.000. The correlation coefficient [r] indicates the *strength or degree of the linear relationship* between two variables (systolic and diastolic BP). As the value of "r" is positive and the p-value is <0.05, we can conclude that there is a significant positive correlation between systolic and diastolic BP.

The value of "r" lies between "–1" and "+1". Values near to "zero" indicate no correlation, while values near to "+1" or "–1" indicate a strong correlation. The negative value of "r" (– r) indicates an inverse relationship. A value of r ≥ 0.8 indicates a very strong correlation; an "r" value between 0.6 and <0.8 indicates a moderately strong correlation; an "r" value between 0.3 and <0.6 indicates a fair correlation; and an "r" value of <0.3 indicates a poor correlation [8].

15.2 Spearman and Kendall's tau-b correlations

The Spearman and Kendall's tau-b are the nonparametric methods of obtaining the correlation coefficients. The Spearman correlation is performed when the normality assumption is violated (i.e., if the distribution of either the dependent or independent, or both the variables, is not normally distributed). Spearman correlation is also applicable for two categorical ordinal variables if they have ≥5 levels, such as intensity of pain

(no, mild, moderate, severe, and very severe pain) and grade of cancer (stage 1, stage 2, stage 3, stage 4, and stage 5).

To obtain the Spearman correlation coefficient of systolic BP (variable name "sbp") and income, where income is not normally distributed, use the following command:

spearman sbp income

This command will report the Spearman correlation coefficient of systolic BP and income along with the p-value as shown in Table 15.5.

The Kendall's tau-b statistic is used to determine the correlation between two ordinal variables, or an ordinal and a continuous variable (provided the ordinal variable has less than 5 levels). To determine the correlation between age group (variable name is age2) and systolic BP, use the following command (Table 15.6):

ktau age2 sbp

Table 15.5 Spearman correlation between systolic BP and income

```
. spearman sbp income

 Number of obs =     210
Spearman's rho =     0.0070

Test of Ho: sbp and income are independent
    Prob > |t| =     0.9192
```

15.2.1 Interpretation

Table 15.5 shows the number of pairwise observations (n=210) used to calculate the Spearman correlation coefficient. Spearman's rho indicates the Spearman correlation coefficient. In our example, Spearman's rho is 0.007, which is very small. The p-value of this test is indicated by "Prob > |t|". The p-value of this test is 0.9192, which is >0.05. We cannot, therefore, reject the null hypothesis. This indicates that there is no statistically significant correlation between systolic BP and income.

Table 15.6 shows the results of the Kendall's tau correlation between age group and systolic BP. The correlation coefficient (Kendall's tau-b) of the variables is 0.0114 and the corresponding p-value is 0.83. This indicates that there is no significant correlation between age group and systolic BP.

Table 15.6 Kendall's tau-b correlation between age group and systolic BP

```
. ktau age2 sbp

  Number of obs =      210
Kendall's tau-a =        0.0092
Kendall's tau-b =        0.0114
Kendall's score =      201
    SE of score =      946.641   (corrected for ties)

Test of Ho: age2 and sbp are independent
     Prob > |z| =        0.8327  (continuity corrected)
```

15.3 Partial correlation

The purpose of performing partial correlation is to determine the correlation between two variables after controlling for one or more other variables (continuous or categorical). This means that, through partial correlation, we get the adjusted "r" value after controlling for other variables included in the analysis.

For example, if we suspect that the relationship between diastolic and systolic BP may be influenced (confounded) by other variables (such as age and diabetes), we should use the partial correlation to exclude the influences of other variables (age and diabetes). The partial correlation gives us the adjusted correlation coefficient (r value). If we want to get the correlation coefficient of diastolic and systolic BP after adjusting for age and diabetes, use the following command:

pcorr dbp sbp age diabetes

The command above will provide the correlation coefficient of diastolic and systolic BP after controlling for age and diabetes, including the p-value as shown in Table 15.7.

15.3.1 Interpretation

The results of the partial correlation of diastolic and systolic BP after adjusting for age and diabetes mellitus are displayed in Table 15.7. We can observe that the partial correlation coefficient of diastolic and systolic BP is 0.847 and the p-value is 0.000. This means that these two variables (diastolic and systolic BP) are significantly correlated (p=0.000) even after controlling for age and diabetes mellitus. The table also provides the results of semipartial correlation. In semipartial correlation, the correlation coefficient is calculated holding the other variables (age and diabetes, in our example) constant either for X or Y, but not for both. In partial correlation, the other

Table 15.7 Correlation between systolic and diastolic BP after controlling for age and diabetes

```
. pcorr dbp sbp age diabetes
(obs=210)

Partial and semipartial correlations of dbp with

               Partial   Semipartial      Partial   Semipartial   Significance
   Variable |    Corr.         Corr.      Corr.^2       Corr.^2          Value
------------+-------------------------------------------------------------------
        sbp |   0.8470        0.8468       0.7175        0.7171         0.0000
        age |   0.0220        0.0117       0.0005        0.0001         0.7524
   diabetes |   0.0409        0.0218       0.0017        0.0005         0.5574
```

variables are held constant for both X and Y. We should consider the partial correlation for reporting.

The partial correlation provided for the age indicates the partial correlation coefficient of age and diastolic BP (0.022; p=0.752) after controlling for systolic BP and diabetes.

16

Linear Regression

Regression analysis is a useful statistical method of data analysis in health and social sciences disciplines. The nature and strength of relationships between two or more continuous variables can be ascertained by regression and correlation analyses.

We have discussed correlation in Chapter 15. While correlation is concerned with measuring the strength of the linear relationship between variables, *regression* analysis is useful in predicting or estimating the value of one variable corresponding to a given value of another variable. For instance, we can use regression analysis to examine whether systolic BP is a good predictor of diastolic BP and also to get the estimated (predicted) value of diastolic BP corresponding to a value of systolic BP. In regression analysis, our main interest is in *regression coefficient* (also called slope or β). The regression coefficient indicates the strength of association between dependent (Y) and independent (X) variables. Linear regression analyses can be done either as *simple linear regression or multiple linear regression* methods.

In this chapter, both simple and multiple linear regression methods are discussed. Multiple linear regression is a type of *multivariable analysis*. The multivariable analysis is a statistical tool where multiple independent variables are considered for a single outcome variable. The terms "multivariate analysis" and "multivariable analysis" are often used interchangeably in health and social sciences research. In fact, multivariate analysis refers to a statistical method for the analysis of multiple outcome variables.

Multivariable analyses are widely used in observational studies, intervention studies (randomized or nonrandomized trials), and studies of diagnosis and prognosis. The main purposes of multivariable analyses are to:

- Adjust for the confounding factors;
- Predict the probability of an outcome when several characteristics are present in an individual;
- Determine the relative contribution of independent variables to the outcome variable; and
- Assess the interaction of multiple variables for the outcome.

There are several types of multivariable analysis methods. The choice of multivariable analysis for the type of outcome variable is summarized in Table 9.3 (Chapter 9). The commonly used multivariable analysis methods in health and social sciences research include multiple linear regression, logistic regression, and proportional hazards regression (Cox regression), which are discussed in this book. Use the data file <**Data_4.dta**> for practice.

16.1 Simple linear regression

In simple linear regression, there is one dependent (outcome) and one independent (explanatory or predictor) variable. The objective of simple linear regression is to find the *population regression equation*, which describes the true relationship between a dependent variable (Y) and an independent variable (X). In a simple linear regression model, two variables are involved – one is an independent variable (X) placed on the X-axis, and the other is a dependent variable (Y) placed on the Y-axis. Then, we call it "regression of Y on X".

Suppose that we want to perform a simple linear regression analysis of diastolic BP (dependent variable) on systolic BP (independent variable). The objective is to find the population regression equation to predict diastolic BP by systolic BP.

Assumptions

1. **Normality**: For any fixed value of X (systolic BP), the sub-population of Y values (diastolic BP) is normally distributed;
2. **Homoscedasticity**: The variances of the sub-populations of “Y” are all equal;
3. **Linearity**: The means of the sub-populations of “Y” lie on the same straight line; and
4. **Independence**: Observations are independent of each other.

The first step in analyzing the data for regression is to construct a scatter diagram to

visualize the relationship between the two variables, which is discussed in Chapter 15. The scatterplot will provide an indication of the linear relationship between the variables, diastolic and systolic BP. For example, to get the scatter plot of diastolic and systolic BP with the fit-line, use the following command:

graph twoway lfit dbp sbp || **scatter** dbp sbp

16.1.1 Commands for simple linear regression

The command for linear regression analysis is “regress”. To do the regression of diastolic BP (variable name is “dbp”) on systolic BP (variable name is “sbp”), use the first of the following commands (Table 16.1):

regress dbp sbp
regress dbp sbp, vce(robust)

The first variable immediately after the command “regress” is the dependent variable. The second command will provide robust estimates of the standard error of the regression coefficient (Table 16.1). Robust regression is an alternative to ordinary regression (least squares regression; first command). It provides better estimates of regression coefficients, especially when outliers are present in the data.

You can conduct the regression analysis on a subset of your data. For example, if you want to perform the regression analysis only for the females (variable name is “sex” and females are coded as 0), use the following command (Table 16.2):

regress dbp sbp if sex==0

You can get the predicted values of diastolic BP (outcome variable) of the individuals after performing the regression analysis. Use the following command to get the predicted values:

predict predbp

The above command will generate a new variable "predbp" containing the predicted values of diastolic BP for all the individuals in the dataset. You can see the new variable at the bottom of the variables window. You can also calculate the estimated (predicted) diastolic BP of an individual whose systolic BP is 110 mmHg by using the following command (outputs not shown):

margins, at(sbp=(110))

16.1.2 Interpretation

The results of simple linear regression analysis are provided in Table 16.1. The table shows the coefficient of determination (R-squared) of diastolic BP on systolic BP. The coefficient of determination is the square of the correlation coefficient value (r value). The table shows that the coefficient of determination (R-squared) of diastolic BP on systolic BP is 0.717 and the p-value (Prob > F) is 0.000. The table also shows the value for the adjusted coefficient of determination (Adj R-squared).

Table 16.1 Regression of diastolic BP on systolic BP

```
. regress dbp sbp

      Source |       SS       df       MS              Number of obs =     210
-------------+------------------------------           F(  1,   208) =  527.20
       Model |  20688.9902     1  20688.9902           Prob > F      =  0.0000
    Residual |  8162.57647   208  39.2431561           R-squared     =  0.7171
-------------+------------------------------           Adj R-squared =  0.7157
       Total |  28851.5667   209  138.045774           Root MSE      =  6.2644

------------------------------------------------------------------------------
         dbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sbp |   .4960326   .0216034    22.96   0.000     .4534429    .5386223
       _cons |   19.40677   2.793132     6.95   0.000     13.90029    24.91325
------------------------------------------------------------------------------

. regress dbp sbp, vce(robust)

Linear regression                                      Number of obs =     210
                                                       F(  1,   208) =  440.31
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.7171
                                                       Root MSE      =  6.2644
------------------------------------------------------------------------------
             |               Robust
         dbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sbp |   .4960326   .0236391    20.98   0.000     .4494297    .5426355
       _cons |   19.40677    3.01498     6.44   0.000     13.46293    25.35061
------------------------------------------------------------------------------
```

The coefficient of determination, or the R-squared value, indicates the amount of variation in "Y" due to "X" that can be explained by the regression line. Here, the R-squared value is 0.717 (~0.72), which indicates that 72% of the variation in diastolic BP can be explained by systolic BP. The rest of the variation (28%) is due to other factors (unexplained variation). The adjusted R-squared value (0.715), as shown in the table, is the adjusted value for better population estimation. The significance of the R-squared

value is assessed by the F-test [F (1, 208) = 527.20] as shown in the table. Since the p-value (Prob > F) of the coefficient of determination is less than 0.05, it is statistically significant. This finding indicates that the linear regression model is useful in predicting the dependent variable (diastolic BP) by the independent variable (systolic BP) in the model.

Table 16.2 Regression of diastolic BP on systolic BP among females

```
. regress dbp sbp if sex==0

      Source |       SS       df       MS              Number of obs =     133
-------------+------------------------------           F(  1,   131) =  404.31
       Model |  16168.3835     1  16168.3835           Prob > F      =  0.0000
    Residual |  5238.71425   131  39.9901851           R-squared     =  0.7553
-------------+------------------------------           Adj R-squared =  0.7534
       Total |  21407.0977   132  162.174983           Root MSE      =  6.3238

------------------------------------------------------------------------------
         dbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sbp |   .5177267    .025748    20.11   0.000      .466791    .5686625
       _cons |   17.38358   3.380971     5.14   0.000     10.69521    24.07194
------------------------------------------------------------------------------
```

We can conclude from the data that there is a significant positive correlation between diastolic and systolic BP since the regression coefficient (0.496; “Coef.” in Table 16.1) is positive and statistically significant (p= 0.000)], and we can, therefore, use the regression equation for prediction. Table 16.1 shows the regression (also called explained) sum of squares (20688.99; in the row "Model" and column "SS") and the residual (also called error) sum of squares (8162.57; in the row "Residual" and column "SS"). The residual is the difference between the observed value and the predicted value (i.e., the observed value and the value on the regression line). The residual sum of squares provides an idea of how well the regression line actually fits into the data. The smaller the value, the better the fit.

The strength of the relationship between variables is measured by the regression coefficient (β or b) or slope. Table 16.1 shows the regression coefficient (row "sbp" and column "Coef.”) of diastolic BP on systolic BP, which is 0.496 with a p-value of 0.000 (P> | t |). The table also shows the value for "a" (_cons) or Y-intercept, which is 19.40. These values (regression coefficient and constant) are needed to construct the linear regression equation.

The value of “b” (regression coefficient) indicates the amount of change in “Y” for each unit change in “X”. In our example, the value of “b” is 0.496. It indicates that if the systolic BP increases (or decreases) by 1 mmHg, the diastolic BP will increase (or decrease) by 0.496 mmHg. The table shows the significance (p-value) of “b”, which is 0.000. A p-value of <0.05 indicates that “b” is not equal to zero in the population (the null hypothesis is that “b” is equal to zero in the population). For simple linear regression, if R-square is significant, “b” will also be significant.

On the other hand, the value of "a" (constant or Y-intercept) in our example is 19.407. The value, a= +19.407, indicates that the regression line crosses or cuts the Y-axis above the origin (zero) and at the point of 19.407 (a negative value indicates the regression line cuts the Y-axis below the origin). The value of "a" does not have any practical meaning since it indicates the average diastolic BP of individuals if the systolic BP is zero.

We know that the equation for simple linear regression is $Y = a + bX$, where "Y" is the predicted value of the dependent variable; "a" is the Y-intercept or constant; "b" is the regression coefficient or slope; and "X" is a value of the independent variable. Therefore, the regression or prediction equation for this regression model is:

$$Y = 19.407 + 0.496 \times X$$

With this equation, we can estimate the diastolic BP of an individual by his/her systolic BP. For example, what will be the estimated diastolic BP of an individual whose systolic BP is 130 mmHg? Using the above equation, the answer is that the estimated diastolic BP will be equal to 83.89 mmHg (19.407 + 0.496 × 130). Stata can calculate this for you if you use the following command after performing the regression analysis:

margins, at(sbp=(130))

If we want to use the regression equation for the purpose of prediction, “b” needs to be statistically significant (p<0.05). In our example, the p-value for “b” is 0.000. We can, therefore, use the equation for the prediction of diastolic BP by systolic BP.

The analysis (Table 16.1) has actually evaluated whether "b" is zero or not in the population by the t-test (*Null hypothesis*: the regression coefficient (b) is equal to "zero" in the population; *Alternative hypothesis*: the population regression coefficient is not equal to "zero"). We will reject the null hypothesis since the p-value is 0.000 (<0.05). We can, therefore, conclude that the systolic BP can be considered in estimating the diastolic BP by using the following regression equation:

$$Y = 19.407 + 0.496 \times X.$$

16.2 Multiple linear regression

In simple linear regression, two variables are involved—one dependent (Y) and one independent (X) variable. The independent variable is also called the *explanatory or predictor* variable. In multiple linear regression, there is more than one explanatory (independent) variable in the model. The explanatory variables may be quantitative or categorical variables. The main purposes of multiple regression analysis are to:

- Obtain the adjusted estimates of the regression coefficients of the explanatory variables in the model;
- Predict or estimate the value of the dependent variable by the explanatory vari ables in the model; and
- Understand the amount of variation in the dependent variable explained by the explanatory variables together in the model.

Suppose that we want to assess the contribution of four variables (age, systolic BP, sex, and religion) in estimating (or predicting) the diastolic BP in a sample of individuals selected randomly from a population. Here, the dependent variable is diastolic BP, and the explanatory variables (independent variables) are age, systolic BP, sex, and religion. Of the explanatory variables, two are quantitative (age and systolic BP) and two are categorical variables (sex and religion). Of the categorical variables, sex has two levels (0= female and 1= male) and religion has three levels (1= Muslim, 2= Hindu, and 3= Christian). Regression does not allow string variables in the analysis. If you want to include any string variable in the model, you need to convert the variable into a numeric variable. How to convert a string variable into a numerical variable is discussed in Chapter 5 (Section 5.1).

When an independent variable is a categorical variable with *more than two levels* (like religion), we cannot simply include it in regression analysis because the code numbers are arbitrary (Stata will consider it as a quantitative variable) and the regression estimates will be misleading. We need to create dummy variables for such categorical variables before including them in the analysis. The dummy variables are the dichotomous indicator variables (coded as 0/1) representing the categories of a categorical variable. The number of dummy variables generated for a categorical variable is equal to the number of levels minus one. For example, if we want to include the variable "religion" in the analysis, we need to generate two dummy variables since it has three levels (Muslim, Hindu, and Christian).

We can generate the dummy variables in Stata as described in Section 5.11. Stata can

can also generate the dummy variables automatically during regression analysis if the prefix "i." is used before a variable that has more than two levels (in general, if the prefix "i." is used for any variable, Stata considers it a categorical variable during analysis).

For example, if we include religion in the regression analysis as "i.religion", Stata will automatically generate two dummy variables for religion during the analysis, with the *first category* (Muslim) as the comparison group by default. You can change the comparison group by using the prefix ".ib". For example, if you want the second category of religion (Hindu) as the comparison group, use the prefix ".ib2" (since Hindus are coded as 2).

We always need to decide on a comparison group (e.g., a comparison group for religion or other variables) before generating the dummy variables or entering a categorical variable in the model with the prefix ".i". Stata, by default, considers the first category (e.g., Muslims for the variable "religion" since Muslims are in the first category) of a variable as the comparison group if the variable is entered with the prefix "i." (e.g., i.religion). If we want to consider the other category (e.g., the third category or Christians) as the comparison group, we should use the prefix ".ib3" since Christians are coded as 3.

In our regression analysis, we will use the variable "religion" with Christians (coded as 3) as our comparison group by using the prefix ".ib3". When a variable is coded as 0/1 (e.g., the variables "sex" and "diabetes"), the regression estimates in multiple regression analysis will be for the higher value, and the lower value will be the comparison group.

16.2.1 Sample size for multiple regression

Multiple regression analysis should be done if the sample size is fairly large. The minimum sample size needed for the analysis depends on how many independent variables we want to include in the model. Different authors provided different guidelines. One author recommends a minimum of 15 subjects for each of the independent variables in the model. Other authors provided a formula (n= 50 + 8m) to estimate the number of subjects required for the analysis. For example, if we intend to include five independent variables in the model, we need to have at least 90 subjects (50 + 8×5). For stepwise regression, there should be 40 cases for each of the independent variables in the model.

16.2.2 Commands for multiple linear regression analysis

The basic command for regression analysis is "regress". Use the following commands for the multiple regression analysis where the dependent variable is diastolic BP (variable name is "dbp") and the explanatory (independent) variables are age, systolic BP (variable name is "sbp"), sex, and religion. We will use the prefix ".ib3" before the variable "religion" so that Stata automatically generates the dummy variables and considers the third category (Christians) as the comparison group during analysis. We will also use the prefix "i." for sex to indicate it as a categorical variable. Note that the variable immediately after the command "regress" is the dependent variable.

regress dbp age sbp i.sex ib3.religion
regress dbp age sbp i.sex i.religion

The first command will provide the outputs with Christians as the comparison group (Table 16.3), while the second command will provide the outputs where Muslims are the comparison group. If you want the second category (Hindus) to be the comparison group, use the prefix "ib2" for religion.

You can get the standardized coefficients (beta) by using the following command:

regress dbp age sbp i.sex ib3.religion, beta

The outputs of the above command are provided at the bottom of Table 16.3. The standardized coefficients are used to understand the relative influence of the independent variables on the dependent variable. The higher the value, the greater the influence.

You can introduce interaction terms into the model. Suppose that you want to check if there is an interaction between sex and religion. For this, use the first command:

regress dbp age sbp i.sex i.religion i.sex#i.religion
regress dbp age sbp i.sex i.religion i.sex#c.sbp

The second command will demonstrate the interaction between sex and systolic BP. The prefix "c." used for "spb" is to indicate that it is a continuous variable.

You can use the "margins" command, after the regression analysis, to get the predicted values of dependent variable for the independent variables in the model (also see Section 16.1.1). For instance, to get the predicted values of diastolic BP (i.e., adjusted mean of diastolic BP) for religion, sex, and religion#sex, use the following commands (results not shown):

Table 16.3 Results of the multiple regression analysis

```
. regress dbp age sbp i.sex ib3.religion

      Source |       SS       df       MS              Number of obs =     210
-------------+------------------------------           F(  5,   204) =  107.89
       Model |  20934.8733     5  4186.97466           Prob > F      =  0.0000
    Residual |  7916.69336   204  38.8073204           R-squared     =  0.7256
-------------+------------------------------           Adj R-squared =  0.7189
       Total |  28851.5667   209  138.045774           Root MSE      =  6.2296

------------------------------------------------------------------------------
         dbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0283828   .0562385    -0.50   0.614     -.139266    .0825004
         sbp |   .4892801   .0216686    22.58   0.000      .446557    .5320032
             |
         sex |
       Male  |  -2.164135   .9128191    -2.37   0.019    -3.963905   -.3643654
             |
    religion |
     MUSLIM  |    .212462   1.363372     0.16   0.876    -2.475645    2.900569
      HINDU  |   -.229616   1.488835    -0.15   0.878    -3.165094    2.705862
             |
       _cons |   21.82239   3.426989     6.37   0.000     15.06553    28.57925
------------------------------------------------------------------------------

. regress dbp age sbp i.sex ib3.religion, beta

      Source |       SS       df       MS              Number of obs =     210
-------------+------------------------------           F(  5,   204) =  107.89
       Model |  20934.8733     5  4186.97466           Prob > F      =  0.0000
    Residual |  7916.69336   204  38.8073204           R-squared     =  0.7256
-------------+------------------------------           Adj R-squared =  0.7189
       Total |  28851.5667   209  138.045774           Root MSE      =  6.2296

------------------------------------------------------------------------------
         dbp |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
         age |  -.0283828   .0562385    -0.50   0.614               -.0185771
         sbp |   .4892801   .0216686    22.58   0.000                .8352803
             |
         sex |
       male  |  -2.164135   .9128191    -2.37   0.019               -.0889736
             |
    religion |
     MUSLIM  |    .212462   1.363372     0.16   0.876                  .00888
      HINDU  |   -.229616   1.488835    -0.15   0.878               -.0087588
             |
       _cons |   21.82239   3.426989     6.37   0.000                       .
------------------------------------------------------------------------------
```

margins religion sex, atmeans

margins religion#sex, atmeans

margins sex, at(religion=(1 2 3)) atmeans

The first command will provide the average diastolic BP for different categories of religion and sex, separately. The second and third command will provide the average diastolic BP for religion and sex together (e.g., Muslim males, Muslim females, Hindu males, Hindu females, Christian males, and Christian females).

16.2.3 Interpretation

In the analysis, we used diastolic BP (dbp) as the dependent variable and age, systolic BP (sbp), sex, and religion as the explanatory variables.

Table 16.3 shows the outputs of the multiple regression analysis. The table shows that data from 210 subjects were analyzed. It shows the values for R-squared (0.725) and adjusted R-squared (0.718), including the p-value (Prob > F; 0.000).

The R-squared value of 0.725 indicates that all the independent variables (age, systolic BP, sex, and religion) together in the model explain 72.5% of the variation in diastolic BP, which is statistically significant (p= 0.000). If the sample size is small, the R-squared value may overestimate the population value. The adjusted R-squared (0.718) gives the R-squared value for better population estimation.

Table 16.3 also shows the regression coefficients (Coef.), p-values (P > | t |) and 95% confidence intervals (95% Conf. Interval) for all the explanatory variables in the model, along with the constant (_cons). The regression coefficients as shown in the table are for age (-0.028; p= 0.614), systolic BP (0.489; p<0.001), sex (-2.164; p= 0.019 for males compared to females), Muslims (0.212; p= 0.876 compared to Christians) and Hindus (-0.229; p= 0.878 compared to Christians).

From the analysis, we can conclude that the systolic BP and sex are the factors significantly influencing the diastolic BP (since the p-values are <0.05). The other variables in the model (age and religion) do not have any significant influence in explaining (or predicting) the diastolic BP. The regression coefficient (Coef.) [also called *multiple regression coefficient*] for systolic BP, in this example, is 0.489 (95% CI: 0.44 to 0.53; p<0.001). This indicates that the average increase (or decrease) in diastolic BP is 0.489 mmHg if the systolic BP increases (or decreases) by 1 mmHg after adjusting for all other variables (age, sex, and religion) in the model. On the other hand, the regression coefficient for sex is –2.164 (95% CI: -3.96 to -0.36; p= 0.019), which means that the average diastolic BP of males is 2.16 mmHg less (since the coefficient is negative) than that of females after adjusting for all other variables (age, systolic BP, and religion) in the model. If the regression coefficient was positive (e.g., +2.164), the

average diastolic BP of males would be 2.16 mmHg higher than that of females, given the other variables constant in the model.

Table 16.3 (at the bottom) shows the standardized coefficients (beta). The standardized coefficients are used to understand the magnitude of the influence of independent variables on the dependent variable. The higher the value, the greater the influence. The table shows that the beta for systolic BP and sex are 0.835 and -0.088, respectively. Therefore, systolic BP has the highest influence (also greater than sex) in predicting diastolic BP.

Regression equation

The regression equation to estimate the average value of the dependent variable with the explanatory variables is given below:

$$Y = a + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 \ldots\ldots B_nX_n$$

Here, "Y" represents the estimated mean value (predicted value) of the dependent variable; "a" represents the constant (or Y-intercept); "B" represents the regression coefficient(s) of the variables in the model; and "X" represents the value of the variable(s) in the model.

Suppose that we want to estimate the diastolic BP of an individual who is 40 years old, male, Muslim, and has a systolic BP of 120 mmHg. In Table 16.3, we can find the regression coefficients for age [= -0.028 (B_1)], systolic BP [= 0.489 (B_2)], sex [= -2.164 (B_3) for being male] and Muslims [= 0.212 (B_4) for being Muslim] and the constant (= 21.82). Therefore, the estimated diastolic BP of the individual will be:

$$Y = 21.82 + (-0.028 \times 40) + (0.489 \times 120) + (-2.164 \times 1) + (0.212 \times 1) = \mathbf{79.67}.$$

16.2.4 Regression diagnostics

The regression analysis is based on certain assumptions. It is important to check the underlying assumptions before considering whether the regression analysis is useful or valid. In regression analysis, the assumptions are commonly checked on the residuals. Residual is the difference between an observed value and a predicted value (value given by the fitted line or regression line). Regression diagnostics are used to evaluate whether the assumptions are true or not.

For practical purposes, after checking for multicollinearity of independent variables and the assumption of linearity (the relationship between X and the mean of Y is

linear), we need to check the following four assumptions on residuals for a linear regression model to be valid. The assumptions to be checked on the residuals are:

1) The residuals are normally distributed with the mean equal to zero;
2) The residuals have constant variance (homoscedasticity);
3) There is no outlier; and
4) The data points are independent.

16.2.4.1 Checking for multicollinearity

Before deciding about the multiple regression model, we need to check for *multicollinearity* (inter-correlations among the independent variables) of the independent variables. If there are moderate to high inter-correlations among the independent variables, two situations may occur. *Firstly*, the importance of a given explanatory variable may be difficult to determine because of a biased (distorted) p-value; and *secondly*, a dubious relationship may be obtained. For example, if there is multicollinearity among the independent variables, we may observe that the regression coefficient for sex is not significant (though it is actually significant) and that the systolic BP has a negative relationship (though the relationship is positive) with the diastolic BP. Another important sign of multicollinearity is a *severe reduction of the adjusted R squared value*.

To determine the correlations among the independent variables, we can generate the Pearson's correlation matrix. For example, if we want to see the correlations among the systolic BP, age, sex, and religion, use the following command to get the correlation matrix (Table 16.4).

corr sbp age sex religion

Table 16.4 shows the correlation coefficients (r values) among the variables included in the analysis. The highest correlation coefficient that we see in the table is -0.13, which is for religion and sex. In general, if the r value is greater than 0.5, it is considered that the correlation may interfere with the regression analysis. However, in our analysis, there is no such problem as shown in the table.

Pearson's correlation can only check for collinearity between any two variables. Sometimes a variable may be multicollinear with a combination of variables. It is, therefore, preferable to use the *tolerance* (or variance inflation factor) measure, which indicates the strength of the linear relationships among the independent variables.

Table 16.4 Correlation matrix of systolic BP, age, sex, and religion

```
. corr sbp age sex religion
(obs=210)

             |      sbp      age      sex religion
-------------+------------------------------------
         sbp |   1.0000
         age |  -0.0281   1.0000
         sex |  -0.1207   0.0586   1.0000
    religion |  -0.0249  -0.0582  -0.1308   1.0000
```

To get the measures of tolerance and variance inflation factor (VIF), use the command "vif" *after performing the multiple regression analysis*. A tolerance value indicates the degree of collinearity. The tolerance value is the inverse of the VIF measure (1/VIF). The tolerance value ranges from 0 to 1. A value of "zero" indicates that the variable is almost in a linear combination (i.e., has a very strong correlation) with other independent variables. To get the values for VIF and tolerance, use the following command after performing the regression analysis:

vif

This command will provide both the VIF and Tolerance (1/VIF) values of the independent variables included in the regression analysis (Table 16.5).

The table (Table 16.5) shows that the tolerance (1/VIF) values for sex, systolic BP, and age are greater than 0.95. The tolerance values for both religion 1 (Muslims) and 2 (Hindus) are 0.41. Usually, the dummy variables have lower tolerance values. *The recommended tolerance level is greater than 0.6 before we include the variables in the multiple regression model*. However, a tolerance value of 0.40 and above is also acceptable, especially if it is a dummy variable.

If there are variables that are highly correlated (tolerance value is <0.4), one way to solve the problem is to exclude one of the correlated variables from the model. The other way is to combine the explanatory variables together (e.g., by taking their sum). Finally, to develop a model for multiple regression, we should first check for multicollinearity and then the other assumptions (see below). If the requirements are fulfilled, then only we can finalize the regression model.

16.2.4.2 Checking for linearity

When we do a linear regression analysis, it is assumed that the relationship between the response (dependent) variable and the predictor variables (independent variables) is

Table 16.5 Multicollinearity test results

```
vif

    Variable |       VIF       1/VIF
-------------+----------------------
         age |      1.01    0.992729
         sbp |      1.02    0.982955
       1.sex |      1.05    0.955039
    religion |
           1 |      2.41    0.414242
           2 |      2.40    0.417031
-------------+----------------------
    Mean VIF |      1.58
```

linear. This is called the assumption of linearity. The best way to check the linearity assumption is to construct a scatterplot and visually inspect the plot for linearity.

Checking linearity for a simple linear regression is straightforward since there is only one predictor (independent) variable against the response (dependent or outcome) variable. For a simple linear regression, we can check the linearity by generating a scatterplot of the dependent variable against the independent variable, which is discussed in Section 7.2. However, to check the linear relationship between diastolic and systolic BP, use the following commands to get a scatterplot of diastolic BP against systolic BP:

twoway (**scatter** dbp sbp)
twoway (**scatter** dbp sbp) (**lfit** dbp sbp)
twoway (**scatter** dbp sbp) (**lfit** dbp sbp) (**lowess** dbp sbp)

The first command will display a scatterplot of diastolic BP and systolic BP without the regression line (fit-line). The second command will provide the scatterplot with the fit-line (Fig 16.1), while the third command will provide all the above with a smooth prediction line. Figure 16.1 shows that the relationship between diastolic and systolic BP is linear since the scatter dots are lying more or less symmetrically around a straight line.

Checking the linearity assumption in multiple linear regression is not straightforward. It can be checked in several different ways. The most straightforward way is to draw scatter plots of standardized residuals (z-residuals) against each of the predictor (independent) variables included in the regression analysis. Standard residuals are the residuals divided by the standard deviation of the residuals.

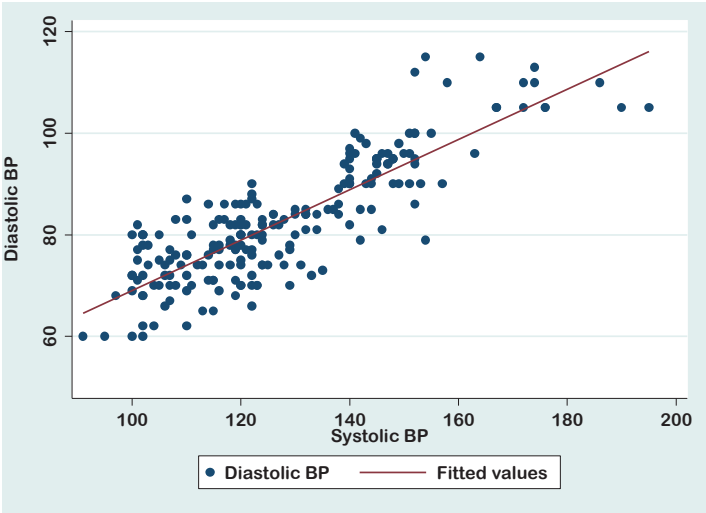


Figure 16.1 Scatterplot of diastolic and systolic BP with the fit line

We have done the regression analysis of diastolic BP (dependent variable) on age, systolic BP, sex, and religion. First, do the regression analysis, then generate the standardized residual variable "zresid" by using the following command:

predict zresid, rstandard

This command will generate a new variable "zresid" with the z-values of the residuals. Now, use the following commands to get the scatterplots of z-residuals against the systolic BP (Fig 16.2) and age (Fig 16.3):

twoway (**scatter** zresid sbp) (lfit zresid sbp)
twoway (**scatter** zresid age) (lfit zresid age)

Both the plots (Figs 16.2 and 16.3) show that the scatter dots are symmetrical above and below the straight lines, indicating that the relationships are linear. You can also construct the same for z-residuals and religion by using the command:

twoway (**scatter** zresid religion) (lfit zresid religion)

16.2.4.3 Checking for normality of residuals

Normality of residuals is required in regression analysis to validate hypothesis testing for R-squared values and regression coefficients, i.e., the normality assumption ensures that the p-values of the F-test and t-test are valid. Normality is not required to

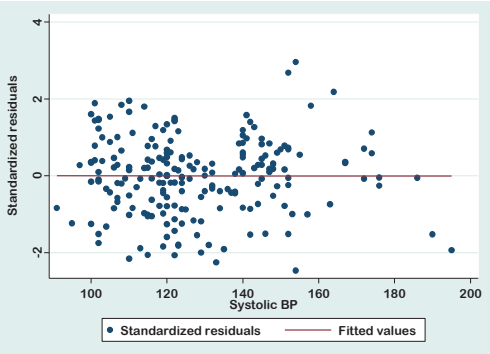


Figure 16.2 Scatterplot of z-residuals against systolic BP

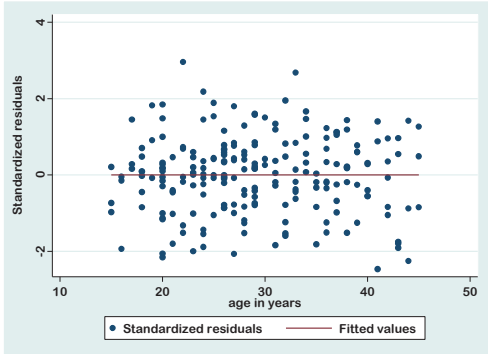


Figure 16.3 Scatterplot of z-residuals against age

obtain unbiased estimates of the regression coefficients (b values). Similarly, the normality assumption of the independent variables is not required for multiple regression analysis.

After running the *regression analysis*, use the following command to generate the residual variable:

predict residual, resid

This command will generate a new variable "residual" with the residuals in the data file. Now, use any of the following commands to check for the normality of residuals.

kdensity residual, normal
histogram residual
pnorm residual
qnorm residual

The command "kdensity" stands for kernel density plot. Using the option "normal" will provide the overlaid normal density curve on the plot (Fig 16.4). The command "histogram" will generate a histogram of the residuals (Fig 16.5), while the commands "pnorm" and "qnorm" will provide the P-P (Fig 16.6) and Q-Q (Fig 16.7) plots, respectively. The P-P plot is sensitive to non-normality in the middle range of data, while the Q-Q plot is sensitive to non-normality near the tails.

You can also use formal statistical tests (Shapiro-Wilk test or Skewness-Kurtosis test) to evaluate the normality of residuals by using any of the following commands (Table 16.6):

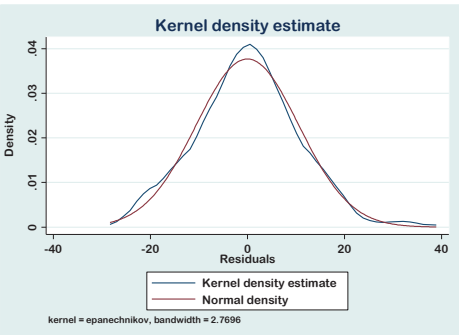


Figure 16.4 Kernel density estimate with overlaying normal density curve

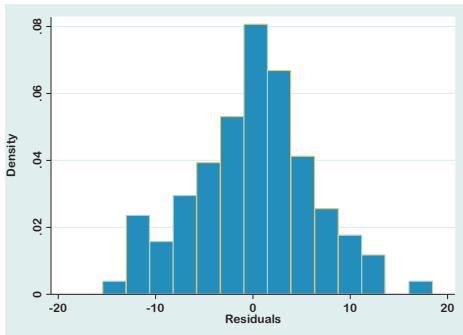


Figure 16.5 Histogram of residuals

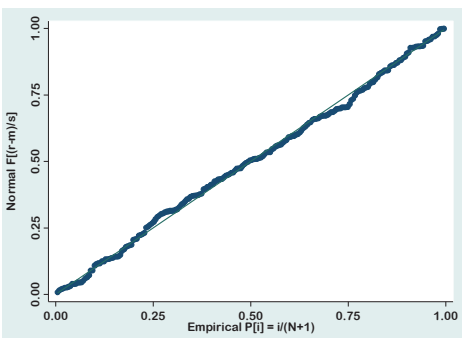


Figure 16.6 P-P plot of residuals

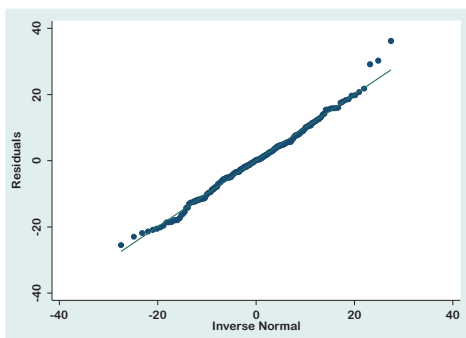


Figure 16.7 Q-Q plot of residuals

swilk residual
sktest residual

Table 16.6 shows that the p-values of both these tests are greater than 0.05, indicating that the distribution of residuals is normal.

To get the mean of the residuals, use the following command (Table 16.7):

sum residual

All the plots (Kernel density, histogram, P-P, and Q-Q plots) indicate that the distribution of the residuals is normal (also see Chapters 5 and 8). The mean of the residuals,

as shown in Table 16.7, is -7.01e^{-09} (-7.01 x e^{-09}), which is equal to "zero". For practical purposes, simply construct the histogram of the residuals and decide whether the distribution is approximately normal or not.

Table 16.6 Normality test of residuals

```
. swilk residual

                   Shapiro-Wilk W test for normal data

    Variable |    Obs       W           V         z       Prob>z
-------------+-------------------------------------------------
    residual |    210    0.99256      1.158     0.339    0.36734


. sktest residual
                     Skewness/Kurtosis tests for Normality
                                                       ------- joint ------
    Variable |    Obs   Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------------
    residual |    210      0.7485         0.9940         0.10         0.9499
```

16.2.4.4 Checking for homoscedasticity

To check for heteroscedasticity (i.e., the variances of the residuals are not homogeneous), after the regression analysis, use either of the following commands. The outputs are shown in Table 16.8.

estat hettest
estat imtest

Table 16.7 Mean of the residuals

```
sum residual

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    residual |       210    -7.01e-09    6.154585  -15.05615   18.24044
```

The first test of heteroscedasticity given by the command "hettest" is the Breusch-Pagan test, and the second test given by the command "imtest" is the White's test. Both these tests test the null hypothesis that the variance of the residuals is homogeneous. Therefore, to fulfill the assumption, we expect a p-value greater than 0.05. Table 16.8 shows that the p-values of both these tests are greater than 0.05 (use the p-value of

Table 16.8 Test for heteroscedasticity

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of dbp

         chi2(1)      =     0.27
         Prob > chi2  =   0.6059

. estat imtest

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      18.98     16    0.2697
            Skewness |       4.25      5    0.5144
            Kurtosis |       0.02      1    0.8930
---------------------+-----------------------------
               Total |      23.24     22    0.3880
---------------------------------------------------
```

either test for interpretation). This indicates that the variances of the residuals are homogeneous.

However, these tests are very sensitive to model assumptions. It is, therefore, a common practice to combine the tests with a diagnostic plot (a plot of residuals against the predicted values) to make a judgment on the severity of heteroscedasticity. To generate the plot of residuals against the fitted (predicted) values of the dependent variable (diastolic BP), use either of the following commands:

rvfplot
rvfplot, yline(0)

The above commands will generate Figures 16.8 and 16.9, respectively. The figures are similar except that Figure 16.9 has a reference line that represents Y= 0. If the scatters of the points show no clear pattern (as seen in Fig 16.8 and Fig 16.9), we can conclude that the variances of the sub-population of "Y" are constant (homoscedastic).

If there is evidence of heteroscedasticity (Fig 16.10), one of the solutions is to run a regression with robust standard errors by using the following command:

regress dbp age sbp i.sex ib3.religion, vce(robust)

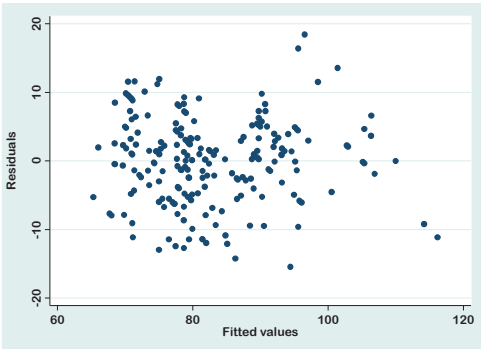


Figure 16.8 No heteroscedasticity

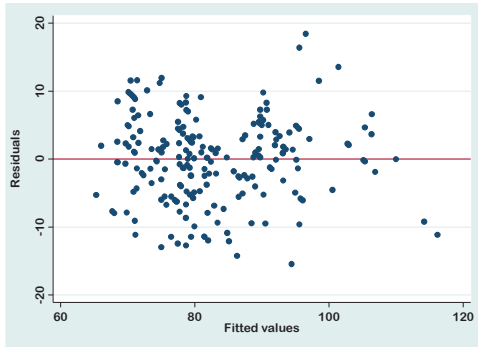


Figure 16.9 No heteroscedasticity

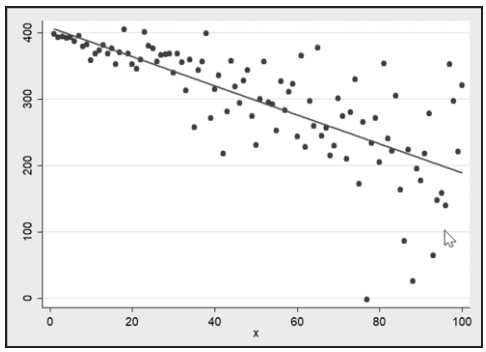


Figure 16.10 Presence of heteroscedasticity

16.2.4.5 Checking for outliers

The presence of outliers is checked on the standardized values (z-values) of the residuals. Outliers are the z-values of the residuals that are either less than -3.0 or greater than +3.0. To assess the outliers, we need to generate (we have done it before; see Section 16.2.4.2) a variable for the z-values of residuals (or transform the variable "residual" generated earlier into z-values; see Section 5.10) by using the following command:

predict zresid, rstandard

This command will generate a new variable "zresid" (if not done earlier) with the z-values of residuals. Now, use the following commands to check if there are any values that are greater than +3.0 or less than -3.0.

list sl zresid if (zresid<-3)

list sl zresid if (zresid>3)
tab zresid

The first command will display the z-residual values if they are less than minus 3.0 along with the serial numbers (ID numbers). Note that if there is no value that is less than minus 3.0, Stata will not show any output. Similarly, the second command will display the z-residual values that are greater than +3.0 along with the ID numbers. If there are any values that are less than -3.0 and/or greater than +3.0, consider them the outliers. If you use the third command, Stata will provide a long table of z-residual values (better to avoid this).

Instead of using the above commands, you can check the highest 10 and lowest 10 values of the variable "zresid" by using the following commands after sorting the variable:

sort zresid
list zresid in 1/10
list zresid in -10/-1

The above commands will display the outputs as shown in Table 16.9. The table shows that there are no values that are greater than +3.0 or less than minus 3.0. This indicates that there is no outlier in the residuals. However, one or more outliers may be present in a dataset. The outliers should be carefully checked for mistakes in data entry. If there is no evidence of mistakes and the value is plausible, then the value should not be altered. It is also discouraged to exclude outliers from analysis. The presence of outliers in the data may influence the results of regression or other statistical analyses. A recommended strategy for handling the outliers is to carry out the analysis with and without the outliers. If there is little difference in the results, the outliers have minimal effect. If the difference is substantial, it may be better to find an alternative method of data analysis, such as data transformation or using the rank method [3]. You may also run the regression with robust standard errors if there are outliers in the data by using the following command:

regress dbp age sbp i.sex ib3.religion, vce(robust)

16.2.4.6 Test for independence

The test for checking the independence of observations (i.e., values of residuals are independent or there is no autocorrelation) is needed for time-series data. The test for independence is done to evaluate if there is any autocorrelation in the residuals. Autocorrelation refers to the degree of correlation of the same variable between two succes-

Table 16.9 Upper and lower 10 z-values of the residuals

```
. list zresid in 1/10

     +-----------+
     |    zresid |
     |-----------|
  1. | -2.462475 |
  2. | -2.258894 |
  3. | -2.158022 |
  4. | -2.071153 |
  5. | -2.065784 |
     |-----------|
  6. | -2.003306 |
  7. |  -1.93732 |
  8. | -1.908815 |
  9. | -1.884941 |
 10. |  -1.83873 |
     +-----------+
```

```
. list zresid in -10/-1

     +----------+
     |   zresid |
     |----------|
201. | 1.600753 |
202. | 1.665619 |
203. | 1.800676 |
204. | 1.823561 |
205. | 1.846748 |
     |----------|
206. | 1.887106 |
207. | 1.951255 |
208. | 2.181215 |
209. | 2.678532 |
210. | 2.960519 |
     +----------+
```

sive time intervals (i.e., the degree of similarity between a given time series). For cross-sectional data, autocorrelation is not an issue.

The independence of residuals (autocorrelation) is assessed by the Durbin-Watson (DW) statistic and is applicable for time-series data. The DW test is done *after executing the multiple regression analysis*. The DW statistic ranges from 0 to 4. A value of 2 indicates that there is no autocorrelation. Values less than 0 to 2 indicate the presence of a positive autocorrelation, while values greater than 2 to 4 indicate the presence of a negative autocorrelation. To perform the DW test, it requires a time variable. In our dataset (Data_4.dta), there is no time variable. However, we can generate a time variable (for the purpose of demonstration) by using the following command before doing the DW test:

gen time = _n

This command will generate a time variable "time". Now, use the following command to let Stata know which variable is the time variable for this analysis (if the time variable in your data file is "year", use the command: tsset year):

tsset time

Finally, use the following command, after multiple regression analysis, to get the DW test statistic (Table 16.10):

Table 16.10 Durbin-Watson test

```
. gen time= _n

. tsset time
        time variable:  time, 1 to 210
                delta:  1 unit

. estat dwatson

Durbin-Watson d-statistic(  6,   210) =  1.700676
```

estat dwatson

If the DW statistic value hovers around 2 (between 1.5 and 2.5), it indicates that the data points are independent (there is no autocorrelation). Table 16.10 shows that the value of the DW statistic is 1.70. Since the value is close to 2, it can be considered that there is no autocorrelation in the residuals.

16.2.5 Variable selection for a model

In general, the independent variables to be selected for a multivariable analysis should include the risk factors of interest and potential confounding factors (based on theory, prior research, and empirical findings), while variables with lots of missing values should be excluded.

We have used the "Enter" method (see Table 16.11) for the regression analysis of data so far in this chapter. The "Enter" method uses all the independent variables in the model as decided by the researcher. It does not remove any variables from the model automatically during analysis. Automatic procedures can be used to determine which independent variable(s) will be included (retained) in the model. The major reason for using the automatic selection procedure (i.e., the stepwise method) is to identify the useful independent variables necessary to estimate or predict the outcome variable. The major limitations of automatic selection procedures (stepwise methods) are that the analysis may provide invalid estimates and confidence intervals. Therefore, the stepwise methods should be used cautiously [28, 40].

Stata and other data analysis software have the option to automatically select the independent variables for a model. They use statistical criteria to select the variables and their order in the model. The commonly used variable selection techniques are the stepwise methods.

In simple terms, stepwise regression is a process that helps us to determine which variables are important and which are not in explaining the outcome variable. Certain variables may have high p-values (e.g., p-values greater than 0.05) and do not meaningfully contribute to predicting the outcome. In stepwise regression, only the important variables that are statistically contributing to the outcome are kept to ensure the best linear model for prediction. Different methods of stepwise approaches are described in Table 16.11.

16.2.5.1 Backward selection method

For a backward selection method of variables in multiple regression, use the following command (outputs are shown in Table 16.12):

xi: stepwise, pr(0.2): **regress** dbp sbp age i.sex i.religion

In this example, "pr(0.2)" indicates the removal (exclusion) criteria of independent variables from the model, i.e., if the p-value of an independent variable is ≥0.02, the variable will be automatically removed from the model. You can change the criteria of removal based on your requirements. For example, you may decide the removal criteria as 0.1 or 0.3 or others. The use of "xi:" before the command "stepwise" will consider both the categorical and quantitative variables for the analysis.

Sometimes it is necessary to keep one or more variables in the model that are considered important for theoretical or practical reasons, though they are not statistically significant (forced-entry of variables). In our example, the variable "age" is not significantly associated with the outcome variable (p=0.614; Table 16.3). However, if you want to force-entry the variable "age" into the model in a stepwise method, use the first of the following commands (Table 16.13):

xi: stepwise, pr(0.2) lockterm1: regress dbp age sbp i.religion
xi: stepwise, pr(0.2) lockterm1: regress dbp (age i.religion) sbp

The first command will force-entry the variable "age" (the first variable immediately after the dependent variable) into the model. The second command will force-entry the variables "age" and "religion" into the model.

For the *backward stepwise* method with a removing criteria of p≥0.2 and adding (entry) criteria of p<0.1, use the following command:

xi: stepwise, pr(.2) pe(.1): regress dbp sbp age i.sex i.religion

Table 16.11 Variable selection methods for modeling

Technique	Method	Advantages and limitations
Backward selection	In this method, all the variables are initially included, and in each step, the most statistically insignificant variable (p>0.05; useless variable) is dropped. This process is repeated until all the variables left over are statistically significant.	Better for assessing (adjusting) for confounding effect than the forward selection method.
Forward selection	In this method, first a single independent variable that has the strongest association (smallest p-value) with the outcome variable is entered into the model. Then (second step), the method identifies the variables among those not in the model that, when added to the model so far obtained, explain the largest amount of the remaining variability. The second step is repeated until the addition of the extra variable is not statistically significant.	Best suited for dealing with the studies where the sample size is small. Does not deal well with suppressor (confounding) effects.
Stepwise/ Remove selection	This is a *combination* of forward and backward selection methods. In the stepwise method, variables that are entered are checked at each step for removal. Likewise, in the removal method, variables that are excluded will be checked for re-entry.	Has the ability to manage large number of potential predictor variables, fine-tuning the model to choose the best predictor variables from the available options.
Enter (all variables)	Enters all the variables at the same time and does not remove any variable automatically from the model.	Including all variables may be problematic, if there are many independent variables and the sample size is small.

16.2.5.2 Forward selection method

For a forward selection method with an entry-term p-value of ≤0.2, use the following command [pe(#) indicates the level of significance for inclusion into the model]:

xi: stepwise, pe(0.2): regress dbp age sbp i.sex i.religion

Table 16.12 Modeling with backward selection method

```
. xi:stepwise, pr(0.2): regress dbp sbp age i.sex i.religion

i.sex             _Isex_0-1         (naturally coded; _Isex_0 omitted)
i.religion        _Ireligion_1-3      (naturally coded; _Ireligion_1 omitted)
                      begin with full model
p = 0.8763 >= 0.2000  removing _Ireligion_3
p = 0.6723 >= 0.2000  removing _Ireligion_2
p = 0.6181 >= 0.2000  removing age

      Source |       SS       df       MS              Number of obs =     210
-------------+------------------------------           F(  2,   207) =  272.86
       Model |   20917.406     2   10458.703           Prob > F      =  0.0000
    Residual |  7934.16062   207  38.3292784           R-squared     =  0.7250
-------------+------------------------------           Adj R-squared =  0.7223
       Total |  28851.5667   209  138.045774           Root MSE      =  6.1911

------------------------------------------------------------------------------
         dbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sbp |    .489694   .0215077    22.77   0.000     .4472918    .5320963
       _Isex |  -2.180167   .8930832    -2.44   0.015    -3.940872   -.4194618
       _cons |   21.01581   2.838019     7.41   0.000     15.42068    26.61094
------------------------------------------------------------------------------
```

For a *forward stepwise* method with a removing criteria of p≥0.2 and adding (entry) criteria of p<0.15, use the following command:

xi: stepwise, pr(.2) pe(.15) forward: regress dbp sbp age i.sex i.religion

16.2.5.3 Interpretation

The interpretation of the outputs of stepwise methods is the same as discussed in Section 16.2.3. In the backward selection method of analysis, our dependent variable was "dbp" (Table 16.12). The independent variables included in the analysis were "sbp, age, sex, and religion". The outputs of the analysis (Table 16.12) show that two variables (systolic BP and sex) are significantly associated with diastolic BP and are retained in the model.

The R-squared value, calculated in the analysis, is 0.725, indicating that the independent variables (systolic BP and sex) together explain 72.5% of the variation in diastolic BP, which is statistically significant [p=0.000 (Prob > F)]. The regression coefficients (Coef.) and p-values (P > | t |) for systolic BP and sex are 0.489 (p=0.000) and -2.180 (for males compared to females; p=0.015), respectively. We can, therefore, conclude from the analysis that the systolic BP and sex are the important factors significantly influencing the diastolic BP (since the p-values are <0.05).

The regression coefficient for systolic BP is 0.489 (95% CI: 0.44 to 0.53; p<0.001). This indicates that the average increase (or decrease) in diastolic BP is 0.49 mmHg if the systolic BP increases (or decreases) by 1 mmHg after adjusting for sex. The regression coefficient for sex, on the other hand, is –2.180 (95% CI: -3.940 to -0.419; p= 0.015), which means that the average diastolic BP of males is 2.18 mmHg less (since the coefficient is negative) than the females after adjusting for systolic BP.

Table 16.13 Backward selection method with forced-entry of age in the model

```
. xi: stepwise, pr(0.2) lockterm1: regress dbp age sbp i.religion

i.religion           _Ireligion_1-3     (naturally coded; _Ireligion_1 omitted)
                      begin with full model
p = 0.8181 >= 0.2000  removing _Ireligion_2
p = 0.6124 >= 0.2000  removing _Ireligion_3

      Source |       SS       df       MS              Number of obs =     210
-------------+------------------------------           F(  2,   207) =  263.03
       Model |  20704.4665     2  10352.2333           Prob > F      =  0.0000
    Residual |  8147.10016   207  39.3579718           R-squared     =  0.7176
-------------+------------------------------           Adj R-squared =  0.7149
       Total |  28851.5667   209  138.045774           Root MSE      =  6.2736

------------------------------------------------------------------------------
         dbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0353995    .056452    -0.63   0.531    -.1466941    .0758952
         sbp |   .4956516   .0216435    22.90   0.000     .4529816    .5383216
       _cons |   20.48269   3.281515     6.24   0.000     14.01322    26.95217
------------------------------------------------------------------------------
```

17

Logistic Regression

Logistic regression is a commonly used multivariable method of data analysis in health and social sciences research. This tool can be applied to analyze the data of cross-sectional, case-control, or cohort studies. Logistic regression analysis is performed when the outcome (dependent) variable is a categorical variable, either dichotomous (also called binary variable, e.g., disease – present/absent), unordered polychotomous (e.g., type of food preferred – rice/bread/meat) or ordinal variable (severity of pain – mild/-moderate/severe). The predictive (independent) variables can be either categorical or continuous. Like other multivariable analyses, the purposes of multivariable logistic regression analysis are to:

- Adjust the estimate of risk (odds ratio) for a number of independent variables included in the model;
- Determine the relative contribution of independent variables to a single outcome;
- Predict the probability of an outcome for a number of independent variables in the model; and
- Assess the interaction of multiple independent variables on the outcome variable.

Logistic regression analysis can be applied in several methods, depending on the type of outcome variable and study design. They are:

1. Binary logistic regression: This method is used when the dependent variable is a dichotomous (binary) categorical variable, such as a disease (present/absent), vaccinated (yes/no), or the outcome of a patient (died/survived). The binary logistic regression can be applied as:

a) *Unconditional binary logistic regression*: This method is used when the depen-

dent variable is a dichotomous categorical variable in an *unmatched* study design (e.g., unmatched case-control studies). The term "*unconditional binary logistic regression*" is commonly expressed as "*unconditional logistic regression or logistic regression*"; and

b) *Conditional binary logistic regression*: This method is applied where the dependent variable is a dichotomous variable and the cases are *matched* with controls for one or more variables (e.g., matched case-control studies). The word *"binary"* is commonly omitted from the terminology and is simply expressed as "*conditional logistic regression*".

2. Multinominal logistic regression: This method of data analysis is used when the outcome (dependent) variable is a nominal categorical variable with *more than two levels*, such as health-seeking behavior (did not seek treatment/ received treatment from village doctors/ received treatment from pharmacists), type of cancer (stomach cancer/ lung cancer/ skin cancer) and others.

3. Ordinal logistic regression (proportional odds regression): This method is used when the outcome variable is an ordinal categorical variable, like severity of pain (mild/ moderate/ severe) and stage of cancer (stage 1/ stage 2/ stage 3/ stage 4).

17.1 Mathematical concept of logistic regression model

In logistic regression analysis, odds are transformed into natural log (ln) of odds, i.e., "ln odds". The "ln" is the log to the base of e. To reverse (antilog) the ln, we take the exponential (e^x) of the log value. When the odds are transformed into ln odds, it is called the *logit transformation*. In logistic regression, ln odds of the outcome variable are put on the Y-axis. The multivariable logistic regression model is given by the equation:

$$\ln\left[\frac{P}{1-P}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Where, P denotes the probability of the outcome, β_0 is the intercept (constant) and β_i is the regression coefficient of the ith variable (i= 1, 2, …, n), and X_i represents the values of the predictor (independent) variables in the model, $X_i = (X_1, X_2, \ldots X_n)$.

The regression coefficients (β) that we get in logistic regression analysis are the ln odds and the exponential of the regression coefficients are the odds ratios (ORs) for the

categorical independent variables after adjusting for other variables in the model. If the independent variable is a continuous variable, the interpretation is different and is discussed in Section 17.2.1.1. We can calculate the probability *(p)* of an outcome by using the following formula:

$$p = \frac{Exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}{1 + Exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}$$

Or

$$p = \frac{1}{1 + Exp[-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)]}$$

A detailed explanation of the model can be found in any standard biostatistics book. In this chapter, we will discuss the unconditional and conditional logistic regression methods, while the multinominal logistic regression method is discussed in Chapter 18.

Assumptions for logistic regression

Logistic regression does not make any assumption concerning the distribution of the predictor (independent) variables. However, it is sensitive to high correlations among the independent variables (multicollinearity). The outliers may also affect the results of logistic regression analysis.

17.2 Binary logistic regression

Binary logistic regression analysis is appropriate when the outcome variable is a *dichotomous* categorical variable (e.g., disease present/absent). It is commonly used for the adjustment of one or more confounding factors or to model (identify) the predictors for a dichotomous categorical outcome. *For logistic regression analysis in Stata, the dichotomous outcome variable must be coded as "0= disease absent" and "1= disease present". Stata will consider the higher value to be the predicted outcome and the lower value as the comparison group*.

17.2.1 Unconditional binary logistic regression

Unconditional binary logistic regression is simply called logistic regression. We will use the term "logistic regression" to mean unconditional logistic regression throughout this chapter.

Suppose that you have conducted an unmatched study (e.g., a cross-sectional or an unmatched case-control study) to identify the factors (or predictors) associated with diabetes mellitus (dependent or outcome variable). The independent (explanatory or predictor) variables/factors that you have considered for logistic regression analysis are sex (variable name: sex), age (variable name: age), peptic ulcer (variable name: pulcer), family history of diabetes (variable name: fhistory), and religion (variable name: religion).

To perform logistic regression analysis, recode all the dichotomous independent variables as "0 for no" and "1 for yes" if they are not coded like this. If they are coded like this, interpretation of the odds ratio (OR) is straightforward, otherwise, interpretation may be complicated (especially if you do not use the prefix "i." to the variable). Use the data file <**Data_4.dta**> for practice.

The coding scheme of the variables to be used in logistic regression analysis is shown in Table 17.1. You can also check the coding scheme (value labels of categorical variables only) of the variables by using the following command:

label list diabetes sex pulcer fhistory religion
Or,
codebook diabetes sex pulcer fhistory religion

To perform the logistic regression analysis, use either of the following commands:

logistic diabetes i.sex age i.pulcer i.fhistory i.religion
logit diabetes i.sex age i.pulcer i.fhistory i.religion, or
logit diabetes i.sex age i.pulcer i.fhistory i.religion

The first two commands will report the results in terms of odds ratios (ORs) with their 95% confidence intervals (CIs) (Table 17.2), while the third command will present the results in terms of logistic regression coefficients with their 95% CIs (Table 17.3). In the second command, the option “or” indicates the odds ratio. If you use this option with the “logit” command, Stata will report the ORs. If the prefix “i.” is used for the independent variables, Stata will consider them as categorical variables during analysis. Note that the first variable after the command is the dependent (outcome) variable. We recommend using the first command for the analysis since our interest is to get the ORs for easy interpretation of the categorical independent variables.

In logistic regression, Stata considers (by default) the first category (lowest value) of the categorical independent variables as the comparison group. For example, we have entered the variable “sex” in the analysis. The coding scheme of sex is 0 for females

Table 17.1 Codebook

Variable name	Variable label	Variable codes
diabetes	Have diabetes (dependent variable)	0= no; 1= yes
sex	Sex	0= female; 1= male
age	Age in years	Actual value
pulcer	Have peptic ulcer	0= no; 1= yes
fhistory	Have family history of diabetes	0= no; 1= yes
religion	Religion of the subjects	1= Muslim; 2= Hindu; 3= Christian

Table 17.2 Logistic regression analysis using the command "logistic"

```
. logistic diabetes i.sex age i.pulcer i.fhistory i.religion

Logistic regression                               Number of obs   =        210
                                                  LR chi2(6)      =     102.61
                                                  Prob > chi2     =     0.0000
Log likelihood = -57.807479                       Pseudo R2       =     0.4702

------------------------------------------------------------------------------
    diabetes | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
        Male |    4.80143   2.555999     2.95   0.003     1.691379    13.63014
         age |   1.262949   .0503621     5.85   0.000        1.168    1.365616
             |
      pulcer |
         Yes |   5.937446   2.947017     3.59   0.000     2.244453    15.70684
             |
    fhistory |
         Yes |    2.85935   1.531351     1.96   0.050     1.000918    8.168382
             |
    religion |
       HINDU |   1.655078   .9067061     0.92   0.358     .5655909    4.843225
   Christian |   .8324862   .7949383    -0.19   0.848     .1281054    5.409867
             |
       _cons |   .0000187   .0000296    -6.90   0.000     8.48e-07    .0004131
------------------------------------------------------------------------------
```

and 1 for males. Therefore, the analysis will give the OR for males compared to females. On the other hand, the independent variable "religion" has three levels (Table 17.1). Since Muslims are coded as 1 (first category), Muslims will be the comparison group for religion. The ORs that the Stata will provide for other religions (Hindus and

Christians) will be compared to Muslims. However, you can change the comparison group by using the following commands:

logistic diabetes i.sex age i.pulcer i.fhistory ib3.religion
logistic diabetes ib1.sex age i.pulcer i.fhistory ib3.religion

The first command will consider the last category of religion (category 3 or Christians) as the comparison group, while the second command will consider males (because males are coded as 1) as the comparison group for sex, and Christians as the comparison group for religion during analysis.

Occasionally, in cross-sectional studies, data is collected through cluster sampling methods. In such a situation, it is necessary to control (adjust) for the cluster effects during analysis. To adjust for cluster effects (say, the name of the cluster variable is "clus"), use the following command:

logistic diabetes i.sex age i.pulcer i.fhistory i.religion, **vce(cluster clus)**

17.2.1.1 Interpretation

In logistic regression analysis, we have entered five independent variables, among which one is a continuous variable (age) and the others are categorical variables (sex, peptic ulcer, family history of diabetes, and religion). Let us interpret the outputs provided in Table 17.2.

The table shows that the data from 210 subjects were analyzed. The likelihood ratio (LR) chi-square [LR Chi2(4)] value is 102.61 and its p-value (Prob > chi2) is 0.000. We want the LR chi-square test to be significant ($p<0.05$). A significant LR chi-square test indicates that the proposed model is better than the null model (i.e., a model without any independent variable) in predicting the outcome variable. Furthermore, if the LR chi-square test p-value is greater than 0.05, it means that the independent variables are unable to predict the outcome variable (a situation where none of the independent variables in the model are significant).

The pseudo R-square value indicates the amount of variation in the outcome variable that can be explained by the independent variables in the model. In this example, the pseudo R-square (Pseudo R2) value is 0.4702. This indicates that 47.02% of the variation in the outcome variable (diabetes) can be explained by all the independent variables (sex, age, peptic ulcer, family history of diabetes, and religion) in the model. However, the pseudo R-square value should be interpreted cautiously because it is not equivalent to the R-squared value that we get in a linear regression model (Sections

Table 17.3 Logistic regression analysis using the command "logit"

```
. logit diabetes i.sex age i.pulcer i.fhistory i.religion

Iteration 0:   log likelihood = -109.11177
Iteration 1:   log likelihood = -65.686705
Iteration 2:   log likelihood = -58.065695
Iteration 3:   log likelihood = -57.807622
Iteration 4:   log likelihood = -57.807479
Iteration 5:   log likelihood = -57.807479

Logistic regression                               Number of obs   =        210
                                                  LR chi2(6)      =     102.61
                                                  Prob > chi2     =     0.0000
Log likelihood = -57.807479                       Pseudo R2       =     0.4702

------------------------------------------------------------------------------
    diabetes |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
        Male |   1.568914   .5323411     2.95   0.003     .5255445    2.612283
         age |   .2334492   .0398766     5.85   0.000     .1552926    .3116058
             |
      pulcer |
         Yes |   1.781279   .4963443     3.59   0.000      .808462    2.754096
             |
    fhistory |
         Yes |   1.050594   .5355591     1.96   0.050     .0009176    2.100271
             |
    religion |
       HINDU |   .5038483   .5478328     0.92   0.358    -.5698842    1.577581
   Christian |  -.1833386   .9548967    -0.19   0.848    -2.054902    1.688225
             |
       _cons |  -10.88592   1.578628    -6.90   0.000    -13.97998    -7.79187
------------------------------------------------------------------------------
```

16.1.2 and 16.2.3). *This information is not needed if the objective of the analysis is to adjust for Odds Ratio*.

The table (Table 17.2) also shows the ORs (Odds Ratios) for the independent variables with their standard errors (Std. Err.), 95% CIs (95% Conf. Interval), and p-values (P>|z|). In our analysis, there are five independent variables in the model. All of them are categorical variables (sex, family history of diabetes, peptic ulcer, and religion), except for age, which is entered as a continuous variable. The interpretation of the OR for a categorical and a continuous variable is not the same.

Let us first interpret the ORs for categorical variables. Table 17.2 shows that the OR for sex (being male) is 4.80 (95% CI: 1.69 - 13.63), which is statistically significant (P=0.003). Here, the comparison group is female. The OR of 4.80 suggests that males are 4.8 times more likely to have diabetes compared to females after adjusting (con-

trolling) for age, family history of diabetes, peptic ulcer, and religion. Note that the ORs provided by logistic regression analysis are the adjusted ORs. Similarly, individuals who have a family history of diabetes are 2.85 times more likely (OR: 2.85; 95% CI: 1.00 to 8.16; p=0.05) to have diabetes compared to those who do not have a family history, after adjusting for age, sex, peptic ulcer, and religion. But the OR is not statistically significant (though the p-value is 0.05) as the null value (one) is included in the 95% CI. Lastly, the ORs for Hindus and Christians provided in the table are compared to Muslims after controlling for age, sex, peptic ulcer, and family history of diabetes.

The interpretation of OR for age is different since the variable is entered as a continuous variable. In our example, the OR for age is 1.262 (95% CI: 1.168 – 1.365; p=0.000). This means that the odds of having diabetes will increase (since the value of OR is greater than one) by 26.2% (calculated as OR – 1; i.e., 1.262 – 1 = 0.262 or 26.2%) (95% CI: 16.8% to 36.5%) with each year increase in age after adjusting for all other variables in the model, which is statistically significant (p<0.001).

All the outputs in Table 17.3 are the same as the outputs in Table 17.2 except that Table 17.3 provides the regression coefficients (with their standard errors and 95% CIs) rather than the ORs. Note that the z-values (Z) and p-values (P>|z|) are the same in both tables. The coefficients are used to calculate the predicted probabilities of the outcome variable with the independent variables in the model. In logistic regression, the odds are transformed into ln odds (logit transformation) during analysis. Therefore, the coefficients reported in the table are the ln odds and their exponentials (e^x) are the ORs. For example, the coefficient for males is 1.568914 and its exponential is 4.80, which is the OR for males (Table 17.2).

The table (Table 17.3) also shows the iterations. The iterations indicate the log likelihood values. The first iteration (iteration 0) indicates the log likelihood of the null model (i.e., a model without any independent variable). The table shows that the log likelihood has increased [from -109.11 (Iteration 0) to -57.80 (Iteration 5)] with the inclusion of independent variables in the model. The improvement in log likelihood value is statistically significant as indicated by the p-value (0.000) of the log likelihood ratio chi-square test (Prob > Chi2). A significant p-value indicates that the model is significantly better than the null model.

17.2.2 Logistic regression diagnostics

In order to check if the analysis is valid, the model has to satisfy certain assumptions of logistic regression. When the assumptions are not met, the analysis may provide

biased estimates of the coefficients, which may lead to misinterpretation of the results. It is, therefore, important to check the underlying assumptions before considering the logistic regression analysis valid. Regression diagnostics are used to evaluate whether the assumptions are true or not. In this section, we will discuss how to assess some of the important assumptions to validate the model, like multicollinearity, model fit, and others.

17.2.2.1 Checking for multicollinearity

Multicollinearity occurs when one or more independent variables are in a linear combination (highly correlated) with other independent variables in the model. The degree of multicollinearity can vary and can have different effects on the model. When multicollinearity is present, the model becomes dubious (i.e., the model may not provide a correct estimate of the regression coefficients). It is, therefore, important to check for multicollinearity of the independent variables in the model.

Multicollinearity can be checked by generating a correlation matrix of the independent variables included in the model. To generate the correlation matrix, use the following command (Table 17.4):

correlate sex age pulcer fhistory religion

Table 17.4 shows the correlation coefficients (r values) of the independent variables included in the analysis. If an r value is greater than 0.5, it is generally considered that the variable has a correlation with another variable that may affect the regression analysis. We don't have any such problem in our analysis because none of the r values is greater than 0.5 (the highest value is 0.22).

Another simple and subjective way to examine multicollinearity is to check the standard errors (SE) of the coefficients as provided in Table 17.3. If multicollinearity is present and affects the model, the magnitude of the standard errors (SEs) of some of the coefficients will be very high (greater than 5.0) or very low (less than 0.001). The existence of multicollinearity means that the model is not statistically stable. To solve the problem (in general), look at the SEs and omit the variable(s) with a very high (or very low) SE until the magnitude of the SEs hovers between 0.001 and 5.0.

Sometimes an independent variable may have a strong correlation with the constant (residuals or errors) in the model. If there is a strong correlation between the constant and any of the independent variables, you can omit the constant from the model. To check the correlation between constant and independent variables in the model, first generate a residual variable (representing the constant) by using the following command:

Table 17.4 Correlation matrix of the independent variables

```
. correlate sex age pulcer fhistory religion
(obs=210)

             |      sex      age   pulcer fhistory religion
-------------+---------------------------------------------
         sex |   1.0000
         age |   0.0586   1.0000
      pulcer |   0.0520   0.2153   1.0000
    fhistory |  -0.2222   0.1585   0.1282   1.0000
    religion |  -0.1308  -0.0582  -0.1038   0.1453   1.0000
```

predict residual, resid

This command will generate a new variable "residual" containing the residual values (errors). Now, to get the correlation matrix of independent and residual variables, use the following command (Table 17.5):

correlate sex age pulcer fhistory religion residual

Table 17.5 shows a moderate correlation (r = 0.68) between age and residual (constant). However, it has not affected the results since none of the SEs of coefficients are greater than 5.0 or less than 0.001 (Table 17.3). If it had affected the results, some of the SEs of the coefficients would have been either greater than 5.0 or less than 0.001. If there is a problem like this, it is suggested to run the logistic regression analysis without (omitting) the constant. The following is the command to do the analysis without the constant in the model:

logistic diabetes i.sex age i.pulcer i.fhistory i.religion, nocons

17.2.2.2 Checking for model fit

Logistic regression is commonly done to adjust the ORs for confounding factors and to identify predictors for the outcome variable. When the intention of analysis is prediction (i.e., to identify the predictors), then the question is "How good is the model for prediction?" This is judged based on the Hosmer-Lemeshow goodness-of-fit test and the positive and negative predictive values given in the classification table.

Hosmer-Lemeshow goodness-of-fit test

The Hosmer-Lemeshow test indicates how well the observed and predicted values fit with each other (i.e., the observed and predicted probabilities match with each other). The null hypothesis is "the model fits" and the p-value is expected to be >0.05

Table 17.5 Correlation matrix of independent variables and residuals

```
. correlate sex age pulcer fhistory religion residual
(obs=210)

             |      sex      age   pulcer fhistory religion residual
-------------+------------------------------------------------------
         sex |   1.0000
         age |   0.0586   1.0000
      pulcer |   0.0520   0.2153   1.0000
    fhistory |  -0.2222   0.1585   0.1282   1.0000
    religion |  -0.1308  -0.0582  -0.1038   0.1453   1.0000
    residual |   0.1227   0.6824   0.3096   0.1749   0.1768   1.0000
```

(non-significant). If the p-value is not significant (>0.05), it suggests that the model is good for prediction of the outcome variable (i.e., the observed and predicted values are close together). *This test is done after running the logistic regression analysis*, and the command is:

lfit, group(10)

The above command will generate Table 17.6. The table shows that the Hosmer-Lemeshow chi-square test p-value is 0.085. Since the p-value is greater than 0.05, we can conclude that the model is useful for prediction of the outcome variable by the independent variables included in the model. If the test is significant (p<0.05), we will conclude that the model is not good enough to predict the outcome variable by the independent variables in the model. *This information is not needed if the objective of doing the logistic regression analysis is to adjust for the confounding factors*.

Classification table

The classification table provides us with the sensitivity, specificity, and positive and negative predictive values, and overall accuracy of the model. The predictive values indicate how well the model is able to predict the correct category of the dependent variable (i.e., have or do not have the disease). To get the sensitivity, specificity, and positive and negative predictive values, we need to generate the classification table by using the following command (after running the logistic regression analysis). Usually, the classification table is generated at a cut-off value of 0.5 (50%). You can, however, change the cut-off value of your choice.

lstat, cutoff(0.5)

This command will generate Table 17.7. The table shows that the overall accuracy of

Table 17.6 Hosmer-Lemeshow goodness-of-fit test

```
. lfit, group(10)

Logistic model for diabetes, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)

       number of observations =       210
             number of groups =        10
     Hosmer-Lemeshow chi2(8) =        13.87
                 Prob > chi2 =         0.0851
```

this model to predict diabetes (with a predicted probability of 0.5 or greater) is 90.5% (shown at the bottom of the table as correctly classified). The sensitivity and specificity of the model are 71.11% [32 ÷ 45] and 96.36% [159 ÷ 165], respectively, while the positive and negative predictive values are 84.2% [32 ÷ 36] and 92.4% (159 ÷ 172], respectively. Interpretation of these findings is a little complicated and needs further explanation, especially the explanation of sensitivity, specificity, and positive and negative predictive values. For a detailed explanation, readers may refer to any standard epidemiology book [10, 16]. You can see the sensitivity and specificity with varying cut-off points (in a graph) by using the following command (not shown):

lsens

To conclude, the information provided in the classification table is needed if the intention of logistic regression analysis is to predict the outcome variable with independent variables in the model. We can ignore the information if the objective of the analysis is to adjust for the confounding factors.

17.2.2.3 ROC curve

We can construct the receiver-operating characteristic (ROC) curve to assess the model discrimination, which is an indication of the accuracy of logistic regression model. Discrimination is defined as the ability of the model to distinguish between those who have the outcome (e.g., a disease) and those who do not have the outcome. Discrimination is evaluated by using the ROC curve analysis. In ROC curve analysis, the area under the curve (called C-statistic) is measured.

The area under the ROC curve ranges from 0 to 1. A value of 0.5 indicates the model is useless. Values between 0.7 and 0.8 are considered acceptable discrimination, values between 0.8 and 0.9 indicate excellent discrimination, and values ≥0.9 indicate

Table 17.7 Classification table

```
. lstat, cutoff(0.5)

Logistic model for diabetes

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        32             6  |         38
     -     |        13           159  |        172
-----------+--------------------------+-----------
   Total   |        45           165  |        210

Classified + if predicted Pr(D) >= .5
True D defined as diabetes != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   71.11%
Specificity                     Pr( -|~D)   96.36%
Positive predictive value       Pr( D| +)   84.21%
Negative predictive value       Pr(~D| -)   92.44%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    3.64%
False - rate for true D         Pr( -| D)   28.89%
False + rate for classified +   Pr(~D| +)   15.79%
False - rate for classified -   Pr( D| -)    7.56%
--------------------------------------------------
Correctly classified                        90.95%
--------------------------------------------------
```

outstanding discrimination.

To get the ROC curve, after performing the logistic regression analysis, use the following command (Fig 17.1):

lroc

Figure 17.1, generated by the command above, shows that the area under the curve is 0.915 (at the bottom of Figure 17.1). Since the value is >0.9, it is an excellent model for prediction. However, we can separately calculate the area under the ROC curve with its 95% CI. To calculate the area under the curve, we need to generate a classifier variable (say, class). Then we will calculate the area under the curve with its 95% CI. Use the following commands:

predict class
roctab diabetes class

The first command will generate the classifier variable "class", while the second command will calculate the area under the curve with its 95% CI (Table 17.8).

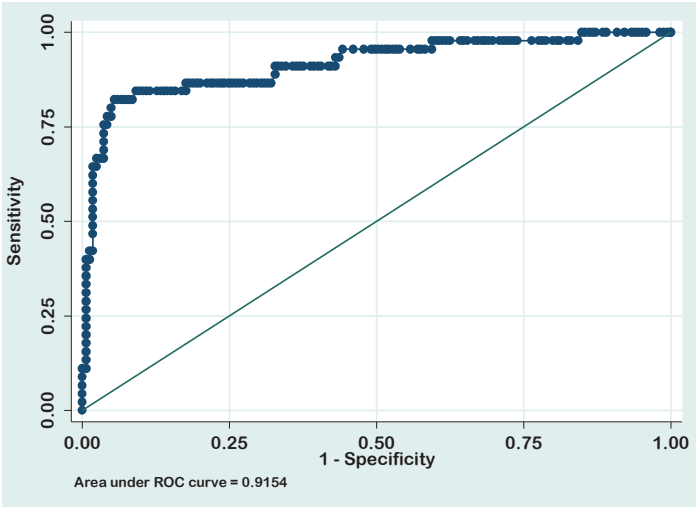


Figure 17.1 ROC curve

17.2.2.4 Other postestimation commands

The postestimation commands are used after performing the logistic regression analysis. Based on the analysis, you can calculate the estimated probability of the dependent variable for each subject in the dataset using the following command:

predict pre1, pr

This command will generate a variable "pre1" with the probability of having the outcome (diabetes) for each subject.

Stata can also provide the probability of having the outcome variable for the independent variables in the model. Use the following command to get the predicted probability (adjusted probability) of having diabetes for each level of sex and religion, considering the average values of the covariates in the model:

margins religion sex, atmeans
marginsplot

The first command will provide the predicted probability of diabetes for each level of religion and sex (Table 17.9). The table shows that the predicted probability (Margin) of diabetes (i.e., the probability of having diabetes) for being a Muslim is 0.0720, while it is 0.113 for Hindus and 0.060 for Christians. The second command will display a graph of predicted probability for all the categories of religion with their 95% CIs (Fig 17.2).

Table 17.8 Area under the ROC curve with 95% CI

```
. predict class
(option pr assumed; Pr(diabetes))

. roctab diabetes class

                      ROC                    -Asymptotic Normal--
          Obs       Area     Std. Err.      [95% Conf. Interval]
        -------------------------------------------------------
          210     0.9154       0.0278        0.86078     0.96993
```

17.2.3 Variable selection for a model

We have discussed different methods of variable selection for a model in the previous chapter in detail (Chapter 16). Like other multivariable analyses, independent variables to be selected for a logistic regression should include the risk factors of interest and potential confounders, while avoiding variables with lots of missing values.

So far in this chapter, we have used the "Enter" method for logistic regression analysis. The "Enter" method uses all the independent variables in the model specified by the researchers. We can use the automatic selection method (stepwise method) for analysis as well. For logistic regression, the commonly used method for the automatic selection of variables for a model is the "*Backward LR*" method. If there is evidence of multicollinearity, you may select the "*Forward LR*" method for analysis. However, stepwise methods are nowadays discouraged from being used because of the biased estimates provided by the analysis [28, 40]. The commands for stepwise logistic regression are provided below:

xi: sw, pr(.1) pe(.05): logistic diabetes i.sex age i.religion i.fhistory
xi: sw, pr(.1) pe(.05) forward: logistic diabetes i.sex age i.religion i.fhistory
xi: sw, pr(.1) pe(.05) forward lockterm1: logistic diabetes (age i.religion) i.sex i.fhistory

The first command is for the backward LR method, while the second one is for the forward LR method. The third command is for the forced entry of variables "age" and "religion" into the model (also see section 16.2.5.1). The "pr(0.1)" indicates the removal criteria of independent variables from the model (i.e., if the p-value of a variable is ≥0.10, the variable will be removed from the model), while the "pe(0.05)" indicates the inclusion (entry) criteria of variables into the model (i.e., if the p-value of a variable is <0.05, the variable will be added to the model). You can change the criteria for removal and inclusion based on your needs. The interpretation of the outputs is the same as discussed in section 17.2.1.1.

Table 17.9 Predicted probabilities for religion and sex

```
. margins religion sex, atmeans

Adjusted predictions                            Number of obs   =        210
Model VCE    : OIM

Expression   : Pr(diabetes), predict()
at           : 0.sex           =    .6333333 (mean)
               1.sex           =    .3666667 (mean)
               age             =    29.01905 (mean)
               0.pulcer        =    .7190476 (mean)
               1.pulcer        =    .2809524 (mean)
               0.fhistory      =    .5428571 (mean)
               1.fhistory      =    .4571429 (mean)
               1.religion      =          .6 (mean)
               2.religion      =    .2761905 (mean)
               3.religion      =    .1238095 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    religion |
     MUSLIM  |   .0720608   .0275834     2.61   0.009     .0179982    .1261233
      HINDU  |     .11389   .0490783     2.32   0.020     .0176983    .2100817
  Christian  |   .0607226   .0527785     1.15   0.250    -.0427214    .1641666
             |
         sex |
     female  |   .0467855   .0199011     2.35   0.019       .00778     .085791
       male  |   .1907179   .0637455     2.99   0.003      .065779    .3156569
------------------------------------------------------------------------------
```

17.2.4 Incorporating interaction terms in the model

Interaction and confounding are not the same. Interaction is also called "effect modification". Interaction is said to be present when the strength of association (OR or RR) of an independent variable with an outcome variable is different at different levels (categories) of a third variable. If the strength of association is the same at different levels of the third variable, there is no interaction.

For example, a researcher is interested in determining an association between smoking and heart disease. He found that there is a causal association between smoking and heart disease. To determine if there was an interaction between smoking and hypertension (the third factor), data were stratified by hypertension (have hypertension and don't have hypertension), and the strength of association (OR) between smoking and heart disease was calculated in each stratum. If the strength of association (OR or RR) is the same in these two strata (hypertensive and non-hypertensive), interaction is

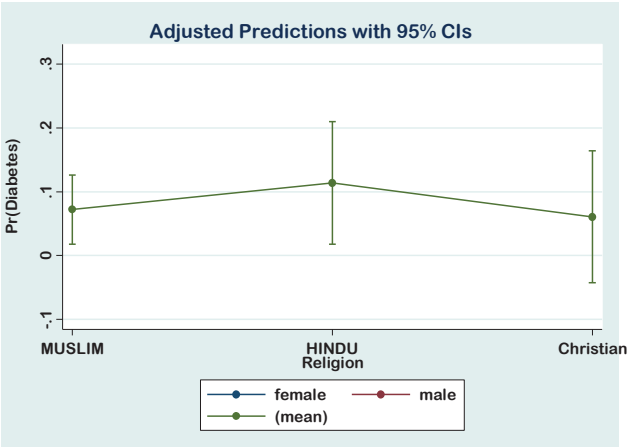


Figure 17.2 Marginal values of religion with the 95% CI

absent. If the strength of association (OR or RR) is different in those two strata, there is an interaction between smoking and hypertension for the causation of heart disease (see Section 13.3).

We can include the interaction terms in logistic regression analysis with other variables for adjustment in the model. Suppose that we are interested in assessing if there is an association of sex and a family history of diabetes with diabetes mellitus. We are also keen to know if there is an interaction between sex and family history of diabetes on the outcome. In such a situation, we need to include both the independent variables (sex and family history of diabetes) and their interaction terms in the model. Note that the variables for which we are looking for interaction must be included in the model independently. To add the interaction term (sex#fhistory) into the model, use the following command:

logistic diabetes i.sex age i.pulcer i.religion i.fhistory i.sex#i.fhistory

The above command will generate Table 17.10. The interaction term included in the model is indicated as "male#Yes". The table shows that the p-value for the interaction is 0.514 (>0.05), indicating that there is no interaction between sex and family history of diabetes for the outcome (i.e., the effect of sex on diabetes is not dependent on the family history of diabetes). If there is an interaction (p-value <0.05), data needs to be analyzed separately at each level of sex or family history of diabetes, depending on the objective.

17.2.5 Sample size for logistic regression

Sample size is always a concern for the analysis of data. The sample size needed for a logistic regression analysis depends on the effect size (OR) you are trying to demonstrate and the variability of the data. It is always better to calculate the sample size during the design phase of the study by using an appropriate formula and selecting the relevant parameters. However, a rule-of-thumb for planning a logistic regression analysis is that for every independent variable in the model, you need to have at least 10 outcomes (some authors recommend a minimum of 15-25 cases for each independent variable) [6, 23].

17.2.6 Conditional logistic regression

In section 17.2.1 we have discussed the unconditional logistic regression analysis, which is used for an unmatched design. The conditional logistic regression analysis is done when the cases and controls are matched for one or more variables (e.g., a matched case-control design).

Suppose that we have conducted a matched case-control study to identify the risk factors for death due to COVID infection. In this study, the cases are matched with the controls by gender. The outcome of interest in this study is death due to COVID (case). The controls are those who survived the COVID infection. Each case (1:1) is matched with a person who survived by gender. The risk factors that will be evaluated in this study are age, religion, diabetes, and hypertension. The data for this matched case-control study is given in the data file <**Data_Ca-Co_matched.dta**>. We will use this data file for conditional logistic regression analysis.

In the dataset, the variable "mID" indicates the matching ID number of cases and controls (we must have this variable), while the variable "death" indicates whether the subject survived (control; coded as 0) or died (case; coded as 1) due to COVID. The codebook of variables to be used in the analysis is provided in Table 17.11. You can also check the coding scheme (value labels) of the variables by using the following command:

codebook mID death gender religion diabetes htn

In logistic regression analysis, we will use the variables death (outcome variable), age, religion, diabetes, and hypertension. The variable mID is needed to specify the groups. Since the cases and controls are matched for gender, it is meaningless to include this variable (gender) in the analysis. To perform the conditional logistic regression, use the

Table 17.10 Logistic regression with interaction terms

```
logistic diabetes i.sex age i.fhistory i.pulcer i.religion i.sex#fhistory

Logistic regression                               Number of obs   =        210
                                                  LR chi2(7)      =     103.03
                                                  Prob > chi2     =     0.0000
Log likelihood = -57.595829                       Pseudo R2       =     0.4721

------------------------------------------------------------------------------
    diabetes | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
        male |   3.421403   2.510496     1.68   0.094     .8121282    14.41398
         age |   1.259393   .0504362     5.76   0.000      1.16432    1.362229
             |
    fhistory |
         Yes |   2.104785     1.4767     1.06   0.289     .5321228    8.325375
             |
      pulcer |
         Yes |   5.802878   2.886573     3.53   0.000     2.188887     15.3838
             |
    religion |
      HINDU  |   1.692301   .9342122     0.95   0.341     .5735658    4.993121
  Christian  |   .8405074   .8056811    -0.18   0.856     .1284131    5.501405
             |
sex#fhistory |
   male#Yes  |   1.978321   2.069448     0.65   0.514     .2546167    15.37117
             |
       _cons |   .0000256   .0000418    -6.49   0.000     1.05e-06    .0006254
------------------------------------------------------------------------------
```

following command:

clogit death age ib3.religion i.htn i.diabetes, group(mID)

The variable immediately after the command (clogit) must be the outcome (dependent) variable. The above command will generate Table 17.12. The table shows the logistic regression coefficients, which are difficult to interpret. We prefer to get the ORs, which are easier to interpret. To get the ORs, use the following command *after the primary analysis:*

clogit, or

This command will give you Table 17.13 with the ORs. We will interpret the outputs provided in this table to draw conclusions.

Table 17.11 Codebook for the data file "Data_Ca-Co_matched.dta"

Variable name	Variable label	Variable codes
mID	Matching ID	Actual value
death	Death due to COVID	0= control/alive; 1= case/died
gender	Sex of the subject	0= female; 1= male
age	Age in years	Actual value
religion	Religion of the subjects	1= Muslim; 2= Hindu; 3= Christian
diabetes	Have diabetes	0= no; 1= yes
htn	Have hypertension	0= no; 1= yes

17.2.6.1 Interpretation

The interpretation of the outputs is similar to unconditional logistic regression analysis as discussed earlier. Table 17.12 shows that the model explains 34% of the variation in the outcome variable by the independent variables (age, religion, hypertension, and diabetes) included in the model (Pseudo R2 = 0.3421).

Our main interest is in ORs, 95% CIs, and p-values. Table 17.13 shows the adjusted ORs for all the variables in the model. The table shows that age (p= 0.004), religion (Muslims compared to Christians) (p= 0.020) and having hypertension (compared to no hypertension) (p= 0.037) are the factors significantly associated with deaths due to COVID. Data indicates that Muslims are more likely to die compared to Christians (adjusted OR: 13.0; 95% CI: 1.49 – 113.2; p= 0.020) after adjusting for age, hypertension, and diabetes. On the other hand, those who have hypertension are 2.56 times more likely to die compared to those who do not have hypertension (adjusted OR: 2.56; 95% CI: 1.06 – 6.22; p= 0.037) after controlling for age, religion, and diabetes. The interpretation of OR for age is different since the variable is entered as a continuous variable (see section 17.2.1.1). In this example, the OR for age is 1.076 (95% CI: 1.02 – 1.13; p=0.004). This means that the odds of dying increase by 7.6% (1.076 – 1.0 = 0.076 or 7.6%) with each year increase in age after adjusting for religion, hypertension, and diabetes, which is statistically significant at 95% confidence level.

Table 17.12 Conditional logistic regression with coefficients

```
. clogit death age ib3.religion i.htn i.diabetes, group(mID)

Iteration 0:   log likelihood = -47.461808
Iteration 1:   log likelihood = -45.703307
Iteration 2:   log likelihood =  -45.60376
Iteration 3:   log likelihood = -45.603507
Iteration 4:   log likelihood = -45.603507

Conditional (fixed-effects) logistic regression   Number of obs   =        200
                                                  LR chi2(5)      =      47.42
                                                  Prob > chi2     =     0.0000
Log likelihood = -45.603507                       Pseudo R2       =     0.3421

------------------------------------------------------------------------------
       death |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0741377   .0258347     2.87   0.004     .0235026    .1247727
             |
    religion |
     MUSLIM  |   2.565603   1.104299     2.32   0.020      .401217    4.729989
      HINDU  |   1.922886   1.091629     1.76   0.078    -.2166663    4.062439
             |
         htn |
        Yes  |   .9436792   .4514828     2.09   0.037     .0587891    1.828569
             |
    diabetes |
        Yes  |    .780711   .4194145     1.86   0.063    -.0413263    1.602748
------------------------------------------------------------------------------
```

17.3 Analysis of cross-sectional data: Estimation of prevalence ratio

It is common practice to analyze data from a cross-sectional study using logistic regression when the outcome variable is dichotomous. Logistic regression analysis, when used for cross-sectional data, provides the adjusted ORs. The ORs provided by the analysis are actually the prevalence odds ratios (PORs). We can use the prevalence ratio (PR) to quantify the association between exposure and outcome in cross-sectional studies.

In cross-sectional studies, when the prevalence of an outcome (e.g., a disease) is more than 10%, POR overestimates the PR if PR is greater than one (i.e., if PR is greater than 1, the POR will be greater than PR). Estimates for the confounding factors are also not equivalent for these two measures. As such, PR should be used in preference to POR while analyzing the cross-sectional data. Therefore, when the prevalence of an outcome of our interest is greater than 10%, our objective should be to calculate the PR rather than the POR for the association of exposure to the outcome.

Table 17.13 Conditional logistic regression with OR

```
. clogit, or

Conditional (fixed-effects) logistic regression   Number of obs   =        200
                                                  LR chi2(5)      =      47.42
                                                  Prob > chi2     =     0.0000
Log likelihood = -45.603507                       Pseudo R2       =     0.3421

------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.076955   .0278228     2.87   0.004     1.023781    1.132891
             |
    religion |
     MUSLIM  |    13.0085   14.36527     2.32   0.020     1.493641    113.2943
      HINDU  |   6.840675   7.467477     1.76   0.078     .8051986     58.1159
             |
         htn |
        Yes  |   2.569417   1.160048     2.09   0.037     1.060552    6.224974
             |
    diabetes |
        Yes  |   2.183024   .9155918     1.86   0.063      .959516    4.966664
------------------------------------------------------------------------------
```

In this section, we will discuss how to analyze cross-sectional data to get the adjusted PR. Several methods are available to obtain the adjusted PR with Stata. They are:

a) Poisson regression;
b) Cox regression (proportional hazards analysis) with constant time; and
c) Generalized linear model (GLM).

Of the above methods, Poisson regression is the easiest. We will, however, demonstrate the use of other methods in this section.

We will use the data file <**Data_4.dta**> for the analysis (consider that the data is from a cross-sectional study). Our interest is to estimate the PR of diabetes for males compared to females after controlling for age, family history of diabetes (fhistory), and peptic ulcer (pulcer).

17.3.1 Poisson regression

We can obtain the adjusted PR using the Poisson regression technique. To get the PR for sex after controlling for age, family history of diabetes (fhistory), and peptic ulcer (pulcer), use the following command:

poisson diabetes i.sex age i.fhistory i.pulcer, irr

This command will generate Table 17.14. The IRR (incidence risk ratio) in the table indicates the adjusted PR since we have analyzed the cross-sectional data.

The analysis reported the IRRs of the independent variables, which are the adjusted PRs (Table 17.14). The table shows that the PR of diabetes for males (compared to females) is 1.93 (95% CI: 1.07 – 3.49; p= 0.028) after controlling for age, family history of diabetes, and peptic ulcer. The findings indicate that the prevalence of diabetes among males is 1.9 times higher than that of females after adjusting for all other variables in the model, which is statistically significant. The interpretation of PR for other categorical variables is similar to this. The interpretation of PR for age is different. The data shows that the IRR (PR) for age is 1.139 (~1.40). This indicates that with each year increase in age, the prevalence ratio of diabetes will increase by 14% (1.14 minus 1) after adjusting for other variables in the model.

17.3.2 Cox regression with constant time

To analyze the data for Cox regression with constant time, first we need to generate a new variable (say, ctime) that will have the same value (say, 1) for all the subjects (constant time). Then we will use the command "stset" to let Stata recognize the time (ctime) and failure (diabetes) variables for the Cox regression analysis. Use the following commands to generate the constant time variable "ctime" and set the data for Cox regression analysis:

gen ctime=1
stset ctime, failure(diabetes)

The first command will generate a new variable "ctime" with all the values equal to 1. The second command will set the data for survival and Cox regression analyses (Chapter 19). Now, use the following command to get the PR:

stcox i.sex age i.fhistory i.pulcer

The command above will generate Table 17.15. You may use the option ", nolog" to suppress the iterations from the outputs. The hazard ratios (Haz. Ratio) reported in the table indicate the adjusted PRs.

Table 17.14 Poisson regression

```
. poisson diabetes i.sex age i.fhistory i.pulcer, irr

Iteration 0:   log likelihood =  -81.30859
Iteration 1:   log likelihood = -81.307121
Iteration 2:   log likelihood = -81.307121

Poisson regression                                Number of obs   =        210
                                                  LR chi2(4)      =      66.03
                                                  Prob > chi2     =     0.0000
Log likelihood = -81.307121                       Pseudo R2       =     0.2888

------------------------------------------------------------------------------
    diabetes |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
        male |   1.938541   .5830272     2.20   0.028     1.075156    3.495254
         age |   1.139111   .0281784     5.27   0.000     1.085199      1.1957
             |
    fhistory |
         Yes |   1.446838   .4649351     1.15   0.250      .770708    2.716125
             |
      pulcer |
         Yes |   2.359057   .7257438     2.79   0.005     1.290843    4.311251
       _cons |   .0011379   .0010385    -7.43   0.000     .0001902    .0068073
------------------------------------------------------------------------------
```

17.3.3 Generalized linear model

You can also use the generalized linear model (GLM). But the GLM with the binomial and log-link functions may suffer from convergence problems, especially when a continuous variable(s) is included in the model for adjustment (i.e., as an independent variable). However, to get the PR for sex after controlling for peptic ulcer and a family history of diabetes, use the following command:

binreg diabetes i.sex i.fhistory i.pulcer, rr nolog

We have used the option "nolog" to suppress iterations. The risk ratios in the table (Table 17.16) indicate the adjusted PRs since we have analyzed the cross-sectional data. Whichever method is used, the interpretation is the same as described in Section 17.3.1.

Table 17.15 Cox regression with constant time

```
. stcox i.sex age i.fhistory i.pulcer

         failure _d:  diabetes
   analysis time _t:  ctime

Iteration 0:   log likelihood = -240.61984
Iteration 1:   log likelihood = -207.96213
Iteration 2:   log likelihood = -207.60698
Iteration 3:   log likelihood = -207.60693
Refining estimates:
Iteration 0:   log likelihood = -207.60693

Cox regression -- Breslow method for ties

No. of subjects =          210                     Number of obs   =       210
No. of failures =           45
Time at risk    =          210
                                                   LR chi2(4)      =     66.03
Log likelihood  =   -207.60693                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
       male  |   1.938541   .5830272     2.20   0.028     1.075156    3.495254
         age |   1.139111   .0281784     5.27   0.000     1.085199      1.1957
             |
    fhistory |
        Yes  |   1.446838   .4649351     1.15   0.250      .770708    2.716125
             |
      pulcer |
        Yes  |   2.359057   .7257438     2.79   0.005     1.290843    4.311251
------------------------------------------------------------------------------
```

Table 17.16 Generalized linear model (GLM)

```
. binreg diabetes i.sex i.fhistory i.pulcer , rr nolog

Generalized linear models                          No. of obs      =       210
Optimization     : MQL Fisher scoring              Residual df     =       206
                   (IRLS EIM)                      Scale parameter =         1
Deviance         =  172.3863397                    (1/df) Deviance =  .8368269
Pearson          =  199.3341071                    (1/df) Pearson  =  .9676413

Variance function: V(u) = u*(1-u)                  [Bernoulli]
Link function    : g(u) = ln(u)                    [Log]
                                                   BIC             = -929.1178
------------------------------------------------------------------------------
             |                 EIM
    diabetes | Risk Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
        male |   2.331071   .5153874     3.83   0.000      1.51133    3.595436
             |
    fhistory |
         Yes |   2.233878   .5522354     3.25   0.001     1.376051    3.626472
             |
      pulcer |
         Yes |   3.044956   .7708366     4.40   0.000     1.853948    5.001086
       _cons |   .0574094   .0163787   -10.02   0.000       .03282    .1004216
------------------------------------------------------------------------------
```

18

Multinominal Logistic Regression

Multinominal logistic regression is an extension of binary logistic regression. It is used when the dependent (outcome) categorical variable has more than two levels (categories) that cannot be arranged into an order, e.g., health-seeking behavior (did not seek treatment, received treatment from the village doctors, received treatment from the pharmacists) or marital status (unmarried, married, divorced, separated).

Suppose that a researcher has conducted a study to identify the factors associated with health-seeking behavior of mothers for diarrhea among their children. Here the dependent variable is "health-seeking behavior", which has three levels (Table 18.1). The independent (explanatory) variables included in the study are maternal age, religion, sex of the child, and severity of diarrhea. The coding scheme of all these variables is provided in Table 18.1.

Data for this study is provided in the data file <**Data_5 multinominal**>. We will use this data to demonstrate multinominal logistic regression analysis.

Table 18.1 Codebook of data file "Data_5 multinominal"

Variable name	Variable label	Variable codes
age	Maternal age in years	Actual value
religion	Religion of the subjects	1= Muslim; 2= Others
behavior	Health seeking behavior	1= Did not receive treatment; 2= Treated by a village doctor; 3= Treated by a pharmacist
sex	Sex of the child	0= female; 1= male
sdiar	Severity of diarrhea	0= Not severe; 1= Severe

For multinominal logistic regression analysis, it is necessary to select a reference group (category) of the outcome variable for comparison. We will select the first category (did not receive treatment) of the outcome variable as the reference category (base) for our analysis. The analysis will, therefore, provide estimates for the categories "treated by village doctors" compared to "did not receive treatment", and "treated by pharmacists" compared to "did not receive treatment". In this example, we will include two categorical variables (religion and severity of diarrhoea) and a quantitative variable (age) as explanatory variables in the model. To do the multinominal logistic regression analysis, use the following command:

mlogit behavior ib2.religion i.sdiar age
mlogit behavior ib2.religion i.sdiar age, rrr
mlogit behavior ib2.religion i.sdiar age, base(2) rrr

The first command will provide the regression coefficients (outputs are not shown). The second command will provide the relative risk ratios (RRRs) rather than the coefficients. For both these commands (first and second), the first category of the outcome variable (did not receive treatment) will be the comparison group (Stata considers the first category as the comparison group by default). The third command will provide the RRRs with the second category of the outcome variable as the comparison group (indicated by the "base(2)" option). The prefix ".ib2", used for religion, is to indicate the second category (other religion; coded as 2) of religion to be the comparison group. Since it is easier to interpret the RRRs, we have shown the outputs of the second command in Table 18.2. The RRRs are the exponentials of the coefficients that we get by using the first command.

18.1 Interpretation

Table 18.2 shows the results of the analysis of the second command. The first iteration (iteration 0) indicates the log likelihood of the null model (i.e., a model without any independent variable). The table shows that the log likelihood has increased [from -223.23 (Iteration 0) to -174.89 (Iteration 5)] with the inclusion of independent variables in the model. The log likelihood chi-square p-value (Prob > Chi2) as provided in the table is significant (p= 0.000). A significant p-value indicates that the model is significantly better than the null model; i.e., the addition of independent variables (religion, severity of diarrhea, and age) in the model has improved the ability to predict the outcome variable compared to the null model.

Table 18.2 Multinominal logistic regression

```
. mlogit behavior ib2.religion i.sdiar age, rrr

Iteration 0:   log likelihood = -223.23547
Iteration 1:   log likelihood =  -181.2718
Iteration 2:   log likelihood = -175.03672
Iteration 3:   log likelihood = -174.89221
Iteration 4:   log likelihood = -174.89173
Iteration 5:   log likelihood = -174.89173

Multinomial logistic regression                   Number of obs   =        210
                                                  LR chi2(6)      =      96.69
                                                  Prob > chi2     =     0.0000
Log likelihood = -174.89173                       Pseudo R2       =     0.2166

------------------------------------------------------------------------------------------
                 behavior |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------------------+---------------------------------------------------------------
Did_not_receive_treatment |  (base outcome)
--------------------------+---------------------------------------------------------------
Treated_by_vill_doc       |
                religion |
                 Muslim   |   2.627298   1.324947     1.92   0.055     .9777956    7.059443
                          |
                   sdiar |
                 severe   |     4.8923   2.469855     3.14   0.002     1.818812    13.15947
                     age  |   1.271572   .0517161     5.91   0.000     1.174145    1.377083
                   _cons  |   .0000683   .0001029    -6.37   0.000     3.56e-06    .0013092
--------------------------+---------------------------------------------------------------
Treated_by_pharmacist     |
                religion |
                 Muslim   |   2.080179   .6647571     2.29   0.022     1.111948    3.891498
                          |
                   sdiar |
                 severe   |   .8920665   .3586054    -0.28   0.776     .4057134    1.961441
                     age  |   1.002805   .0242082     0.12   0.908     .9564624    1.051392
                   _cons  |   .5782204    .395847    -0.80   0.424     .1511349    2.212188
------------------------------------------------------------------------------------------
```

The Pseudo R-squared (Pseudo R2) value indicates how much variation in the dependent variable can be explained by the independent variables in the model. The results show that 21.6% of the variation in the dependent variable can be explained by the independent variables together in the model (religion, severity of diarrhea, and age). However, the readers should interpret the pseudo R-squared value cautiously as it is not equivalent to the R-squared value that we get in linear regression analysis (Sections 16.1.2 and 16.2.3).

The main output has two parts, labelled with the categories of outcome variable (treated by village doctors and treated by pharmacists). The analysis provided the RRRs with their 95% CIs and p-values.

The first half of the table has the results for "treated by village doctors" compared to "did not receive treatment". The results indicate that Muslims (compared to other

religions) are more likely to receive treatment from village doctors after controlling for mothers' age and severity of diarrhea, but the association is not statistically significant (adjusted RRR: 2.62; 95% CI: 0.97 – 7.05; p=0.055). However, severity of diarrhea (i.e., if the baby has severe diarrhea) is significantly associated with seeking treatment from village doctors after controlling for mothers' age and religion (adjusted RRR: 4.89; 95% CI: 1.81 – 13.15; p=0.002).

For the quantitative variables, an RRR greater than one indicates an increased likelihood of the response category (treated by village doctors) compared to the reference category (did not receive treatment). The results show that with the increase in mothers' age, it is significantly more likely to receive treatment from the village doctors after adjusting for religion and severity of diarrhea (adjusted RRR: 1.27; 95% CI: 1.17 – 1.37; p=0.000) [in other words, the likelihood of receiving treatment from village doctors increases by 27% with each year increase in mother's age].

The second half of the table shows the results for "treated by pharmacists" compared to "did not receive treatment". The interpretations are similar to those mentioned above. The results show that only religion (being Muslim) is significantly associated with seeking treatment from pharmacists after adjusting for maternal age and severity of diarrhea (adjusted RRR: 2.08; 95% CI: 1.11 – 3.89; p= 0.022).

18.2 Post-estimation commands

After performing the regression analysis, you can get the predicted probabilities of the outcome variable by using the "margins" command, such as:

margins religion, atmeans predict (outcome(1))
margins religion, atmeans predict (outcome(2))
margins religion, atmeans predict (outcome(3))

The above commands will provide the predicted values for Muslims and other religions for each level of the outcome variable (Table 18.3). The marginal (Margin) values indicate the probabilities of the outcome. For example, the marginal value for Muslims is 0.390 for the first category of the outcome variable (i.e., did not receive treatment). This indicates that the probability of not seeking treatment for being a Muslim is 0.39 (which is 0.58 for other religions), considering the average values of other independent variables in the model. The interpretation of other outputs is similar to this.

Table 18.3 Marginal values

```
. margins religion, atmeans predict (outcome(1))

Adjusted predictions                            Number of obs   =        210
Model VCE    : OIM

Expression   : Pr(behavior==Did_not_receive_treatment), predict(outcome(1))
at           : 1.religion      =    .5571429 (mean)
               2.religion      =    .4428571 (mean)
               0.sdiar         =    .7190476 (mean)
               1.sdiar         =    .2809524 (mean)
               age             =    29.01905 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    religion |
     Muslim  |   .3902659   .0514549     7.58   0.000     .2894162    .4911156
     Others  |   .5810089   .0563714    10.31   0.000      .470523    .6914949
------------------------------------------------------------------------------

. margins religion, atmeans predict (outcome(2))

Adjusted predictions                            Number of obs   =        210
Model VCE    : OIM

Expression   : Pr(behavior==Treated_by_vill_doc), predict(outcome(2))
at           : 1.religion      =    .5571429 (mean)
               2.religion      =    .4428571 (mean)
               0.sdiar         =    .7190476 (mean)
               1.sdiar         =    .2809524 (mean)
               age             =    29.01905 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    religion |
     Muslim  |   .1166562   .0368918     3.16   0.002     .0443495    .1889628
     Others  |   .0661029   .0283829     2.33   0.020     .0104735    .1217323
------------------------------------------------------------------------------

. margins religion, atmeans predict (outcome(3))

Adjusted predictions                            Number of obs   =        210
Model VCE    : OIM

Expression   : Pr(behavior==Treated_by_pharmacist), predict(outcome(3))
at           : 1.religion      =    .5571429 (mean)
               2.religion      =    .4428571 (mean)
               0.sdiar         =    .7190476 (mean)
               1.sdiar         =    .2809524 (mean)
               age             =    29.01905 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    religion |
     Muslim  |   .4930779   .0527806     9.34   0.000     .3896299    .5965259
     Others  |   .3528881   .0548775     6.43   0.000     .2453302     .460446
------------------------------------------------------------------------------
```

We can use the command "margins" to plot the predicted probabilities of each category of the outcome variable (behavior) by religion. Stata will create the plots based on the last margins command used. We can also combine 3 margins plots into a single figure by using the command "graph combine" (Fig 18.1). Use all the following commands successively:

margins religion, atmeans predict (outcome(1))
marginsplot, name (not_treated)
margins religion, atmeans predict (outcome(2))
marginsplot, name (treated_vdoc)
margins religion, atmeans predict (outcome(3))
marginsplot, name (treated_pharm)
graph combine not_treated treated_vdoc treated_pharm

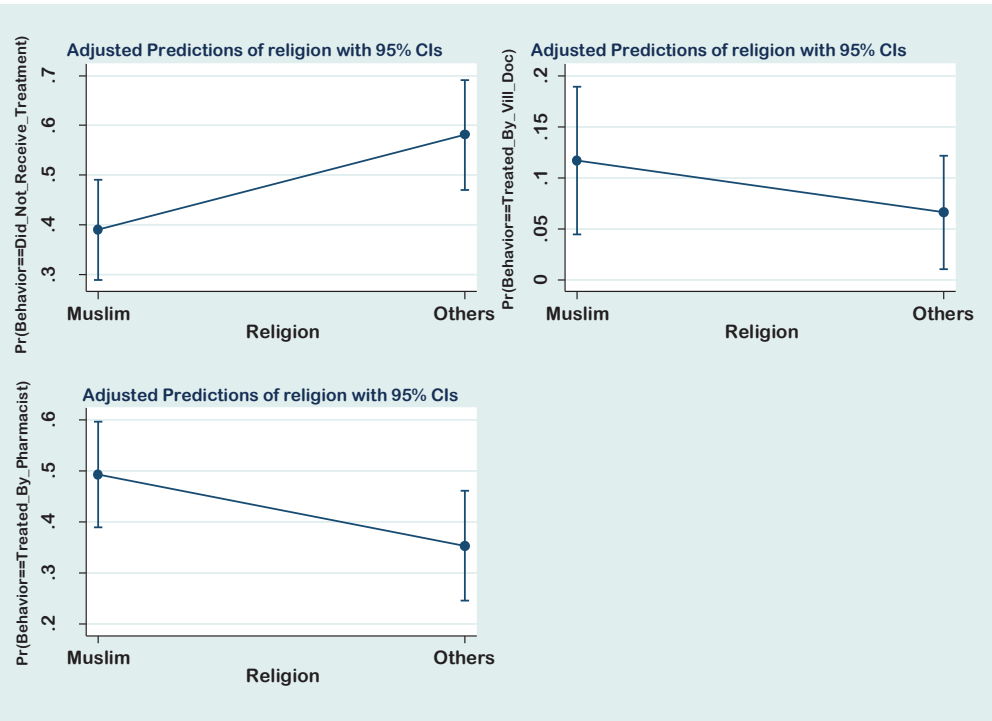


Figure 18.1 Marginal values of different categories of religion for seeking treatment

19

Survival Analysis

There are situations when researchers are interested in knowing the progress of a patient from a specific point in time (e.g., from the point of diagnosis or initiation of treatment) until the occurrence of a certain outcome, such as death or recurrence of any event (e.g., recurrence of cancer). The prognosis of a condition is commonly assessed by estimating the: a) *Median survival time* and b) *Cumulative probability of survival* after a certain time interval (e.g., 5-year or 3-year or others).

For instance, a researcher may be interested in determining the median survival time of colonic cancer if the patient is treated (or not treated), and the estimated probability that a patient with colonic cancer may survive for more than 5 years (5-year cumulative survival probability). The methods employed to answer the above questions in a follow-up study are known as survival analysis (or lifetable analysis) methods.

Survival analysis is done in follow-up studies, where subjects are usually followed over a specified period of time and the focus is on the time at which the event of interest occurs. To do the survival analysis, we need to have data (information) from each of the subjects, at least on the following variables:

- *Time*: It is the length of time the patient was observed in the study (called "survival time"). Time can be measured in days, weeks, months, years or other units of time.
- *Outcome*: Whether the patient developed the outcome of interest (event) during the study period, or whether the patient was either lost to follow-up or remained alive at the end of the study (censored); and
- *Treatment group*: Which treatment (e.g., treatment A or B) did the patient receive in the study (optional)?

Survival time is of two types – a) Censored time; and b) Event time. The *censored* time is the amount of time contributed by the:

a) Patients who did not develop the outcome and remained in the study up to the end of the study period, or
b) Patients who were lost to follow-up due to any reason, such as migration or withdraw, or
c) Patients who developed the outcome for reasons other than the disease of interest (e.g., died in a car accident)

On the other hand, the *event* time is the amount of time contributed by the patients who developed the outcome of interest during the study period. In survival analysis, the outcome measure of interest is survival time, which is a mixture of event time and censored time.

If we have the above information, it is possible to estimate the median survival times and cumulative survival probabilities for two or more treatment groups for comparison. Such a comparison allows us to answer the question, "which treatment delays the time of occurrence of an event?" The method commonly used to analyze survival-time data is the *Kaplan-Meier* method, and Stata can be used for the analysis of such data. Use the data file <**Data_survival_4.dta**> for practice. The codebook for the data file is provided in Table 19.1. Note that in this data, the event of our interest is death.

Table 19.1 Codebook for the data file "Data_survival_4.dta"

Variable name	Variable label	Variable codes
time	Survival time in days	Actual value in days
outcome	Survival status	0= Censored; 1= Died/Event
treatment	Treatment group	0= Placebo; 1= New treatment
age	Age in years	Actual value
sex	Sex	0= Male; 1= Female

19.1 Survival analysis: Kaplan-Meier method

Suppose that a researcher has conducted a follow-up study (clinical trial) on patients with acute heart attack (myocardial infraction) to determine the effectiveness of a new drug (n=22) compared to a placebo (n=22) in reducing the time to death. The patients

were followed until time to death or up to six months (180 days), whichever came first. The outcome of interest in this study was the time to death (event) due to acute heart attack. The objective was to assess whether the "new treatment" delays (increases) the time to event (death) compared to placebo among patients with heart attack. Data from this study is provided in the data file <**Data_survival_4.dta**>, which includes the following necessary variables:

- *Time*: The variable "time" indicates the amount of time each patient has spent in the study in days;
- *Treatment*: It specifies which treatment the patient received in this clinical trial (coded as 0= received placebo; 1= received new treatment);
- *Outcome (event)*: Whether the patient developed the event, i.e., died or not (coded as 0= censored/did not die; 1= died).

Assumptions

- The probability of outcome is similar among the censored (lost to follow-up) and under-observation individuals;
- There is no secular trend over the calendar period;
- Risk is uniform during the interval; and
- Loss to follow-up is uniform over the interval.

19.1.1 Preparing data for analysis

Before conducting the survival analysis, we need to prepare the dataset so that Stata can automatically recognize the time variable and the event variable (censored/event) during analysis. In our dataset, the time variable is "time" and the event variable is "outcome". *The event variable must be coded as 1= event and 0= censored*. Use the following command to prepare the data for survival analysis:

stset time, failure(outcome)

When the command "stset" is used, Stata generates some new variables, like "_st, _d, _t, and _t0". You can see them at the bottom of the variables window. If you save the data file after using the "stset" command, you don't need to prepare the data file again for survival analysis in the subsequent sessions. But if you don't save the data file, you need to prepare it before survival analysis every time during the subsequent sessions. In fact, we need to set the data using the "stset" command for any analysis that uses the commands beginning with "st…", for example, stdescribe, stcox, sts test, or others.

19.1.2 Commands for Kaplan-Meyer method

Before analyzing data for survival analysis, let us check our data for the number of subjects in each treatment group, including the number of events that occurred. To get the summary of the data, use the following command:

tab2 treatment outcome, row

This command will generate a two-by-two table (Table 19.2) of the treatment group and outcome. The table shows that there are 44 subjects enrolled in this study (22 in both the placebo and new treatment groups). In total, 27 patients died (events) during the study period, of which 16 (59.26%) were in the placebo group and 11 (40.74%) were in the new treatment group.

Table 19.2 Area under the ROC curve with 95% CI

```
. tab2 treatment outcome, col

-> tabulation of treatment by outcome

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

    Treatment |    Survival status
        group |  Censored       Died |     Total
--------------+----------------------+----------
      Placebo |         6         16 |        22
              |     35.29      59.26 |     50.00
--------------+----------------------+----------
New treatment |        11         11 |        22
              |     64.71      40.74 |     50.00
--------------+----------------------+----------
        Total |        17         27 |        44
              |    100.00     100.00 |    100.00
```

The primary objectives of survival analysis are to get the median survival time and cumulative survival probability (survival functions) for each treatment group and the significance of the difference (log-rank test) in survival functions between the treatment groups. Once the data set is prepared (by the "stset" command), use the following commands to get the median survival time and its 95% CI by treatment group. The option "by(treatment)" will present the results separately for the placebo and new treatment groups.

stsum, by(treatment)
stci, by(treatment)

The first command will provide the median survival times of the placebo and new treatment groups, while the second command will provide the same but with their 95% CIs (Table 19.3).

Table 19.3 Median survival time by treatment group

```
. stsum, by(treatment)

         failure _d:  outcome
   analysis time _t:  time

         |               incidence       no. of    |------ Survival time -----|
treatm~t | time at risk     rate        subjects        25%       50%       75%
---------+-------------------------------------------------------------------
 Placebo |         1422   .0112518            22         22        40         .
New trea |         2409   .0045662            22         89       146         .
---------+-------------------------------------------------------------------
   total |         3831   .0070478            44         29        89         .


. stci, by(treatment)

         failure _d:  outcome
   analysis time _t:  time

             |    no. of
treatment    |  subjects         50%     Std. Err.     [95% Conf. Interval]
-------------+-----------------------------------------------------------
     Placebo |        22          40      12.89864           22         71
    New trea |        22         146      10.79461           89          .
-------------+-----------------------------------------------------------
       total |        44          89      21.23218           41        168
```

To get the survival functions disaggregated by treatment group, use the following commands:

sts list, by(treatment)
sts list, by(treatment) compare

The first command will provide a long table showing the survival functions (cumulative survival probabilities) at different time points (Table 19.4), while the second command will provide a table showing a comparison of survival functions between placebo and new treatment groups (Table 19.5).

Table 19.4 Survival functions at different time points by treatment group

```
. sts list, by (treatment)

         failure _d:  outcome
   analysis time _t:  time
```

Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf. Int.]	
Placebo							
2	22	1	0	0.9545	0.0444	0.7187	0.9935
3	21	1	0	0.9091	0.0613	0.6830	0.9765
4	20	1	0	0.8636	0.0732	0.6344	0.9539
7	19	1	0	0.8182	0.0822	0.5853	0.9276
10	18	1	0	0.7727	0.0893	0.5374	0.8985
22	17	1	0	0.7273	0.0950	0.4910	0.8671
28	16	1	0	0.6818	0.0993	0.4462	0.8338
29	15	1	0	0.6364	0.1026	0.4029	0.7988
32	14	1	0	0.5909	0.1048	0.3610	0.7621
37	13	1	0	0.5455	0.1062	0.3207	0.7239
40	12	1	0	0.5000	0.1066	0.2818	0.6843
41	11	1	0	0.4545	0.1062	0.2444	0.6433
54	10	1	0	0.4091	0.1048	0.2085	0.6007
61	9	1	0	0.3636	0.1026	0.1743	0.5567
63	8	1	0	0.3182	0.0993	0.1418	0.5111
71	7	1	0	0.2727	0.0950	0.1112	0.4637
127	6	0	1	0.2727	0.0950	0.1112	0.4637
140	5	0	1	0.2727	0.0950	0.1112	0.4637
146	4	0	1	0.2727	0.0950	0.1112	0.4637
158	3	0	1	0.2727	0.0950	0.1112	0.4637
167	2	0	1	0.2727	0.0950	0.1112	0.4637
180	1	0	1	0.2727	0.0950	0.1112	0.4637
New treatment							
2	22	1	0	0.9545	0.0444	0.7187	0.9935
6	21	1	0	0.9091	0.0613	0.6830	0.9765
12	20	1	0	0.8636	0.0732	0.6344	0.9539
54	19	1	0	0.8182	0.0822	0.5853	0.9276
56	18	0	1	0.8182	0.0822	0.5853	0.9276
68	17	1	0	0.7701	0.0904	0.5325	0.8973
89	16	1	0	0.7219	0.0967	0.4822	0.8645
96	15	2	0	0.6257	0.1051	0.3883	0.7926
125	13	0	1	0.6257	0.1051	0.3883	0.7926
128	12	0	1	0.6257	0.1051	0.3883	0.7926
131	11	0	1	0.6257	0.1051	0.3883	0.7926
140	10	0	1	0.6257	0.1051	0.3883	0.7926
141	9	0	1	0.6257	0.1051	0.3883	0.7926
143	8	1	0	0.5475	0.1175	0.2979	0.7410
145	7	0	1	0.5475	0.1175	0.2979	0.7410
146	6	1	0	0.4562	0.1285	0.2047	0.6782
148	5	0	1	0.4562	0.1285	0.2047	0.6782
162	4	0	1	0.4562	0.1285	0.2047	0.6782
168	3	1	0	0.3041	0.1509	0.0676	0.5910
173	2	0	1	0.3041	0.1509	0.0676	0.5910
180	1	0	1	0.3041	0.1509	0.0676	0.5910

We need to use a statistical test to objectively assess the significance of the difference in survival functions between the treatment groups. The most commonly used statistical test is the log-rank test. However, there are alternatives to this test. They are the Tarone-Ware and Peto tests. The commands to get these test statistics are:

Table 19.5 Comparison of survival functions at different time points by treatment groups

```
. sts list, by (treatment) compare

         failure _d:  outcome
   analysis time _t:  time

                 Survivor Function
treatment       Placebo  New treat
----------------------------------
time       2     0.9545     0.9545
          24     0.7273     0.8636
          46     0.4545     0.8636
          68     0.3182     0.7701
          90     0.2727     0.7219
         112     0.2727     0.6257
         134     0.2727     0.6257
         156     0.2727     0.4562
         178     0.2727     0.3041
         200          .          .
----------------------------------
```

sts test treatment
sts test treatment, tware
sts test treatment, peto

The above commands will provide the results of the log-rank, Tarone-Ware, and Peto tests, respectively (Table 19.6).

The cumulative survival probabilities are usually portrayed visually by a graph called the survival curve. You can generate the survival curve by using the following command:

sts graph, by(treatment)
sts graph, by(treatment) ci

The first command will display the survival curve by treatment groups, as shown in Figure 19.1. The second command will also display the curve, but with 95% CIs. You can also generate the cumulative incidence curve (1 – cumulative survival) by using the following command (Figure 19.2):

sts graph, by(treatment) failure

19.1.3 Interpretation

In total, 44 subjects were enrolled in this study, of which 22 received a placebo and

Table 19.6 Significance tests for survival functions

```
. sts test treatment

         failure _d:  outcome
   analysis time _t:  time

Log-rank test for equality of survivor functions

              |   Events         Events
treatment     |  observed       expected
--------------+-------------------------
Placebo       |        16          10.62
New treatment |        11          16.38
--------------+-------------------------
Total         |        27          27.00

                     chi2(1) =       4.66
                     Pr>chi2 =     0.0309

. sts test treatment, tware

         failure _d:  outcome
   analysis time _t:  time

Tarone-Ware test for equality of survivor functions

              |   Events         Events        Sum of
treatment     |  observed       expected        ranks
--------------+--------------------------------------
Placebo       |        16          10.62    33.847597
New treatment |        11          16.38   -33.847597
--------------+--------------------------------------
Total         |        27          27.00            0

                     chi2(1) =       6.07
                     Pr>chi2 =     0.0138

. sts test treatment, peto

         failure _d:  outcome
   analysis time _t:  time

Peto-Peto test for equality of survivor functions

              |   Events         Events        Sum of
treatment     |  observed       expected        ranks
--------------+--------------------------------------
Placebo       |        16          10.62    4.3863004
New treatment |        11          16.38   -4.3863004
--------------+--------------------------------------
Total         |        27          27.00            0

                     chi2(1) =       6.03
                     Pr>chi2 =     0.0141
```

another 22 received the new treatment. There were a total of 27 deaths during the study period, among which 16 (59.26%) were in the placebo group and 11 (40.74%) were in the new treatment group (Table 19.2).

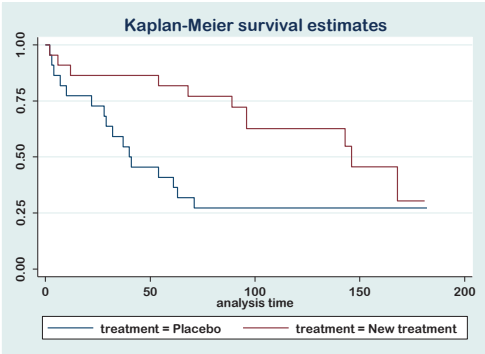


Figure 19.1 Kaplan-Meier survival curve

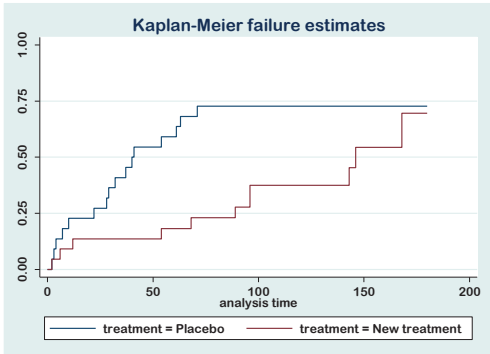


Figure 19.2 Cumulative incidence curve

Table 19.3 shows the median survival times for both the placebo and new treatment groups, including their 95% CIs. The median survival time is the time when the cumulative survival probability is 0.50 (i.e., the time when 50% of the patients develop the event). The table indicates that the median survival time, if a patient is in the placebo group, is 40 days (95% CI: 22 to 71), while it is 146 days (95% CI: 89 to .), if a patient is in the new treatment group. This means that the new treatment increases the survival time, i.e., the new treatment is associated with a longer time-to-event (and the placebo is associated with a shorter time-to-event). We, therefore, conclude that an individual will live longer if s/he receives the new treatment compared to the placebo.

Table 19.4 shows the cumulative survival probability (survivor function) at different points in time in the placebo and new treatment groups. In this table, we can see that the cumulative survival probability at the end of 71 days (in the time column), in the placebo group, is 0.272 (27.2%). Since there is no death after that, the cumulative survival probability at the end of 180 days will be the same (0.272).

On the other hand, the cumulative survival probability is 0.304 (30.4%) at the end of 168 days for the patients who were in the new treatment group. As there is no death after that, the cumulative survival probability at the end of 180 days will be the same (0.304). In the new treatment group, the cumulative survival probability at the end of 71 days (68 days in the table) is about 0.770 (77.0%), which is much higher than the placebo group (0.273). This indicates that the probability of surviving at the end of 71 days is higher among the patients who received the new treatment compared to placebo. This also indicates the benefit of the new treatment (i.e., the new treatment is better than the placebo). A comparison of the survival functions at different points in time is

provided in Table 19.5. For instance, at the end of 90 days, the cumulative probability of survival is 0.2727 in the placebo group compared to 0.7219 in the new treatment group.

However, if we consider the cumulative survival probability of patients in both groups at the end of 180 days, the probabilities are not that different (0.272 in the placebo group and 0.304 in the new treatment group). This information suggests that though survival experiences are significantly different between the treatment groups (as indicated by the log rank test), the difference in survival probability at the end of 180 days is small.

Now, the question is whether the survival experiences of both these groups in the population are different or not. For an objective comparison of survival experiences in two groups, it is desirable to use a statistical method that will tell us whether the difference in survival experiences in the population is statistically significant or not. Here, the null hypothesis is that "the survival experiences in the placebo and new treatment groups are the same in the population". Such a null hypothesis can be tested by the *log-rank* test. We have done the log-rank test and the results are provided in Table 19.6. The p-value (Pr>chi2) of the test is 0.030, which is <0.05. This indicates that survival experiences of both these groups in the population are not the same. In other words, it tells us that the survival probability is better (since the median survival time is higher in the new treatment group) if the patient is in the new treatment group compared to the placebo group (i.e., the new treatment is more effective/better than placebo in improving the patients' survival).

There are alternative procedures for testing the null hypothesis that the two survival curves are identical. They are the *Breslow test, the Tarone-Ware test, and the Peto test*. Here, we have performed the *Tarone-Ware and Peto tests*, and the results are shown in Table 19.6. The log-rank test ranks all the deaths equally, while the other tests give more weight to early deaths.

Survival curve: The cumulative survival probability is usually portrayed visually by a graph called the survival curve (Fig 19.1). The steps in the graph represent the times when events (deaths or any other event of interest) occurred. The graph allows us to represent visually the median survival time and the cumulative survival probability for any specific time period (e.g., 30-day, 60-day, or 90-day cumulative survival probability). In general, the line above indicates the better survival probability. We can see that the line for the new treatment is above the line for the placebo, indicating that the new treatment delays (increases) the time to event compared to the placebo.

19.2 Cox regression

Cox regression is also called *proportional hazards analysis*. In the previous section (Section 19.1), we discussed the survival analysis using the Kaplan-Meier method. Like other regression analyses (e.g., multiple linear regression and logistic regression), Cox regression is a multivariable analysis technique where the dependent measure is a mixture of time-to-event and censored-time observations. Cox regression is commonly done in follow-up studies (e.g., randomized trials) to assess the prognosis. Cox regression with constant time can be used for the analysis of cross-sectional data to estimate the prevalence ratio, which is discussed in Section 17.3.2. For the Cox regression analysis, we will use the same data file (**Data_survival_4.dta**) that was used for the survival analysis.

Returning to our previous example (survival analysis; Section 19.1), where we analyzed the data to assess the effectiveness of a new treatment against a placebo. Our objective was to determine whether the new treatment delays the time-to-death compared to the placebo among patients with heart disease. We found that the new treatment significantly delayed the time-to-death compared to the placebo, as indicated by the higher median survival time and cumulative survival probability, and a significant log-rank test. However, the effectiveness of the new treatment might be influenced (confounded) by other factors, such as age, hypertension, diabetes, or other characteristics. All these factors, therefore, need to be controlled during analysis to assess the effectiveness of the new treatment. Cox regression is a statistical method that is used to control the confounding factors (categorical, continuous, or discrete covariates) that may influence the effectiveness of a treatment.

Cox regression gives us the *hazard ratio (HR) after adjusting the variables included in the model*, which is analogous to relative risk (RR). A hazard ratio (also called a relative hazard) is the ratio of the hazard rate if the individuals are exposed (e.g., to a new treatment) compared to the individuals not exposed (e.g., to placebo). In Cox regression, the dependent variable is the log of hazard.

19.2.1 Commands

In Cox regression analysis, we will consider the variables age and sex for adjustment. The command for Cox regression is "stcox" and it reports the hazard ratios (HRs). *We only use the independent variables with the "stcox" command*. In survival analysis, once the data are set with the "stset" command (see Section 19.1.1), Stata automatically recognizes the time and event variables. To perform the Cox regression analysis, use

the following command:

stcox i.treatment ib1.sex age
stcox ib1.treatment ib1.sex age
stcox ib1.treatment ib1.sex age, nolog nohr

The first command will provide HR for the new treatment compared to the placebo (Table 19.7). The second command will report the HR for the placebo group compared to the new treatment group (Table 19.8). In the third command, we have used the options "nolog" and "nohr" to suppress the iteration history and to get the coefficients instead of HRs, respectively. We have used the prefix ".ib1" for the variables "treatment" (in the second command) and "sex" to indicate the new treatment and females as the comparison groups (since new treatment and females are both coded as 1).

You may prefer using the second command for the analysis. Because if the new treatment is the comparison group, we expect an HR greater than one for the placebo group (since we assume that the new treatment is better than placebo), which is easier to interpret.

Stepwise methods are also available for modeling with Cox regression. For the *backward stepwise* method with a removing criteria of $p \geq 0.2$ and an adding (entry) criteria of $p < 0.1$, use the following command:

xi: stepwise, pr(.2) pe(.1): stcox ib1.treatment ib1.sex age

For the *forward stepwise* method with a removing criteria of $p \geq 0.2$ and an adding (entry) criteria of $p < 0.10$, use the following command:

xi: stepwise, pr(.2) pe(.10) forward: stcox ib1.treatment ib1.sex age

The outputs of these commands are not provided. For further information, see Section 16.2.5. You can generate the survival curve for the treatment groups after adjusting for age and sex by using the following command:

sts graph, by(treatment) adjust(age sex)

This command will display the survival curve for the treatment groups after controlling for age and sex (Figure 19.3). You can also get the cumulative incidence curve adjusted for age and sex by using the following command (Fig 19.4):

sts graph, by(treatment) failure adjust(age sex)

Table 19.7 Cox regression with placebo as comparison group

```
. stcox i.treatment ib1.sex age

         failure _d:  outcome
   analysis time _t:  time

Iteration 0:   log likelihood = -88.962207
Iteration 1:   log likelihood = -84.097071
Iteration 2:   log likelihood = -83.983381
Iteration 3:   log likelihood = -83.983329
Refining estimates:
Iteration 0:   log likelihood = -83.983329

Cox regression -- Breslow method for ties

No. of subjects =           44                     Number of obs   =        44
No. of failures =           27
Time at risk    =         3831
                                                   LR chi2(3)      =      9.96
Log likelihood  =   -83.983329                     Prob > chi2     =    0.0189

------------------------------------------------------------------------------
           _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
     treatment |
New treatment  |   .3665283   .1678311    -2.19   0.028     .1493989    .8992233
               |
           sex |
         male  |   2.433823   1.060213     2.04   0.041     1.036315    5.715921
               |
           age |   1.008859   .0229496     0.39   0.698     .9648665    1.054857
------------------------------------------------------------------------------
```

19.2.2 Interpretation

We have analyzed the data of 44 subjects (22 in the new treatment group and 22 in the placebo group). The variables included in the analysis are “treatment”, “sex”, and “age” to get the estimated effect of the new treatment (compared to placebo) after adjusting for sex and age. Table 19.7 shows the results of the Cox regression analysis.

The table at the beginning shows the iteration history. The first iteration (iteration 0) indicates the log likelihood of the null model (i.e., a model without any independent variable). The table shows that the log likelihood has increased (from -88.96 to -83.98) with the inclusion of independent variables in the model. The log likelihood ratio chi-square p-value (Prob > chi2), as shown in the table, is significant (p= 0.018). A significant p-value indicates that the addition of independent variables in the model has improved the ability to predict the outcome compared to the null model.

Table 19.8 Cox regression with new treatment as comparison group

```
. stcox ib1.treatment ib1.sex age

         failure _d:  outcome
   analysis time _t:  time

Iteration 0:   log likelihood = -88.962207
Iteration 1:   log likelihood = -84.097071
Iteration 2:   log likelihood = -83.983381
Iteration 3:   log likelihood = -83.983329
Refining estimates:
Iteration 0:   log likelihood = -83.983329

Cox regression -- Breslow method for ties

No. of subjects =           44                     Number of obs   =        44
No. of failures =           27
Time at risk    =         3831
                                                   LR chi2(3)      =      9.96
Log likelihood  =   -83.983329                     Prob > chi2     =    0.0189

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   treatment |
    Placebo  |   2.728302   1.249273     2.19   0.028     1.112071    6.693488
             |
         sex |
       male  |   2.433823   1.060213     2.04   0.041     1.036315    5.715921
             |
         age |   1.008859   .0229496     0.39   0.698     .9648665    1.054857
------------------------------------------------------------------------------
```

Table 19.7 shows that out of a total of 44 subjects included in the analysis, 27 died (number of failures). In our analysis, placebo is the comparison group for the variable "treatment" and females are the comparison group for "sex". We will, therefore, get the HRs for the new treatment group compared to placebo and for the males compared to females.

The table (Table 19.7) shows the HRs and their corresponding p-values (P>|z|) with the 95% CIs. The HR for the new treatment is 0.366 (95% CI: 0.149 to 0.899) with a p-value of 0.028. This finding indicates that compared to the placebo, patients in the new treatment group are less likely (63.4%; one minus 0.366) to *have a shorter time to event* (i.e., have greater survival time or survive longer) after controlling for age and sex, which is statistically significant (p=0.028) at 95% confidence level. Furthermore, males are more likely (2.43 times) to have a shorter time to event (have a shorter survival time) compared to females (p=0.041) after controlling for the variables "treat-

ment" and "age". Age, independently, does not have any significant effect on the survival time since the p-value is greater than 0.05.

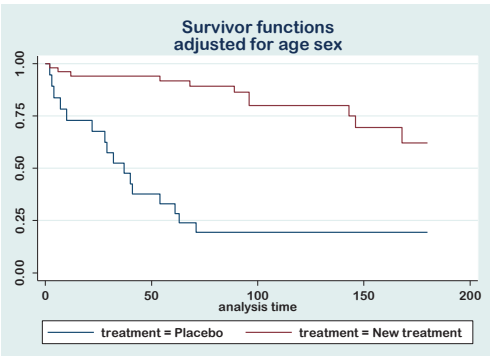


Figure 19.3 Survival curve after controlling for age and sex

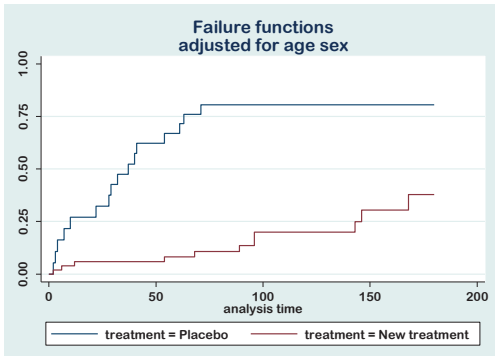


Figure 19.4 Cumulative incidence curve after controlling for age and sex

If you consider the new treatment as the comparison group (second command; Table 19.8), you will get the HR for the placebo group, which is 2.728. This value (2.728) is actually the inverse of the HR (0.366) that we had for the new treatment group compared to the placebo. Interpretation is simple. An HR of 2.7 indicates that the patients in the placebo group are 2.7 times more likely to have a shorter time to event (i.e., the patients in the placebo group are more likely to die early) compared to patients in the new treatment group.

The interpretation of a continuous variable (e.g., age) when included in the model is different. If the HR for age is greater than 1 (say, 1.3), it indicates that the HR will increase (or decrease) by 30% (1.3 minus 1) with each year of increase (or decrease) in age. On the other hand, if the HR is less than 1 (say, 0.7), it means that with each year increase in age, the HR will be reduced by 30% (1 minus 0.7).

19.2.3 Checking for assumptions

Before we conclude the results of Cox regression, we have to check for the important assumptions, such as:

a) There is no multicollinearity among the independent variables; and
b) Relative hazards over time are proportional (also called the proportionality assumption of proportional hazards analysis).

To check for the presence of multicollinearity, look at the standard errors (std. err.) of the coefficients of the variables included in the model (Table 19.9). Since there is no value which is very small (<0.001) or very large (>5.0) (also see Section 17.2.2.1), there is no problem of multicollinearity in the model.

The second assumption (the proportionality assumption) is the major one. If this assumption is violated, the simple Cox regression model is invalid, and more sophisticated analyses are required. Formal statistical tests and graphical methods (log-minus-log plot) can be used for detecting violation of this assumption.

The statistical test for checking the proportional hazards assumption is performed by the following command (this command needs to be used *after performing the Cox regression* analysis since this test is based on the most recent use of the Cox regression results):

estat phtest, detail

This command will provide Table 19.10. The table shows the p-values for the variables "treatment" (new treatment), "sex" (male) and "age". The p-value for age is less than 0.05, while the p-values for other variables are greater than 0.05. A p-value of less than 0.05 indicates that the assumption is violated. This suggests that there are some potential problems with age, while there is no problem with sex and treatment groups (since the p-values are greater than 0.05). The overall test (global test) p-value is also <0.05. In such a situation, you can either omit age from the model (since it is not significant) or use the time-dependent Cox regression method. For further details, readers are referred to any standard text book.

As an alternative, we can also check the assumption by using the graphical method (by a log-minus-log plot). To generate the log-minus-log plot for the treatment groups, after adjusting for age and sex, use the following command:

stphplot, by(treatment) adjust(age sex)

The above command will display the log-minus-log plot for the treatment groups after adjusting for age and sex (Fig 19.5). If there is a constant vertical difference between the two curves (i.e., the curves are parallel to each other), it means that the relative hazards over time are proportional. If the curves cross each other or are much closer together at some points in time and much further apart at other points in time, then the assumption is violated. In our example (Fig 19.5), the two lines are not really parallel, indicating that the assumption may be violated. When the proportional hazards assumption is violated, it is recommended to use the Cox regression with a time-dependent covariate.

Table 19.9 Cox regression with coefficients of HR

```
. stcox ib1.treatment ib1.sex age, nolog nohr

         failure _d:  outcome
   analysis time _t:  time

Cox regression -- Breslow method for ties

No. of subjects =           44                     Number of obs   =        44
No. of failures =           27
Time at risk    =         3831
                                                   LR chi2(3)      =      9.96
Log likelihood  =   -83.983329                     Prob > chi2     =    0.0189

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   treatment |
    Placebo  |   1.003679   .4578939     2.19   0.028     .1062239    1.901135
             |
         sex |
       male  |   .8894631   .4356164     2.04   0.041     .0356707    1.743255
             |
         age |   .0088198    .022748     0.39   0.698    -.0357655    .0534051
------------------------------------------------------------------------------
```

Table 19.10 Test for proportional hazards assumption

```
. estat phtest, detail

      Test of proportional-hazards assumption

      Time:  Time
      ----------------------------------------------------------------
                  |       rho            chi2       df       Prob>chi2
      ------------+---------------------------------------------------
      0b.treatment|            .            .        1             .
      1.treatment |      0.05039         0.07        1          0.7892
      0.sex       |      0.20439         1.12        1          0.2901
      1b.sex      |            .            .        1             .
      age         |     -0.53299         9.22        1          0.0024
      ------------+---------------------------------------------------
      global test |                     12.40        3          0.0061
      ----------------------------------------------------------------
```

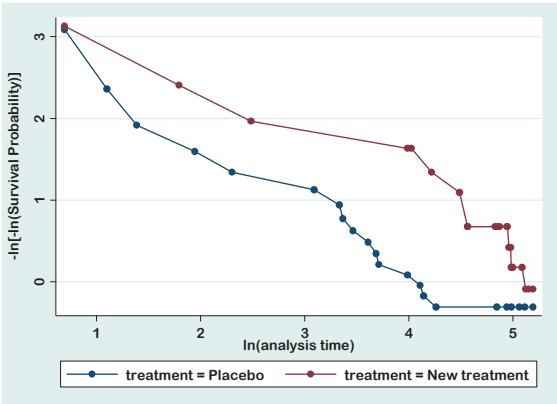


Figure 19.5 Log-minus-log plot for checking the proportionality assumption

20

Nonparametric Methods

Nonparametric methods, in general, are used when the continuous dependent variable is *not* normally distributed. Nonparametric tests are also used when the data is measured on nominal and ordinal scales. Table 20.1 shows the types of nonparametric methods recommended against parametric tests when the dependent variable is not normally distributed in the population. Nonparametric tests are less sensitive compared to parametric tests and may, therefore, fail to detect differences between groups that may actually exist. Use the data file <**Data_4.dta**> for practice.

Table 20.1 Nonparametric methods against the alternative parametric methods

Nonparametric test	Alternative parametric test
Mann-Whitney U test	Independent-samples t-test
Wilcoxon Signed Ranks test	Paired t-test
Kruskal-Wallis test	One-way ANOVA
Friedman test	One-way repeated measures ANOVA
Chi-square test of independence	None
Spearmen's correlation	Pearson's correlation

20.1 Mann-Whitney U test

The Mann-Whitney U test is also called the Wilcoxon rank-sum test. This test is the nonparametric equivalent of the independent samples t-test. This test compares the differences between two samples (groups) on a continuous measure (variable) when

the sample distributions are not normal or the sample size is small (<30). This test is based on ranks of observations and is more efficient than the median test. This test tests the null hypothesis that the two populations have the same median. For example, we may want to know whether the median systolic BP (where the distribution of systolic BP is non-normal) of males and females is the same. To test this hypothesis, use the following command:

ranksum sbp, by(sex)

Table 20.2 Rank-sum (Mann-Whitney) test

```
. ranksum sbp, by(sex)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

         sex |      obs    rank sum    expected
-------------+---------------------------------
      female |      133     14616.5     14031.5
        male |       77      7538.5      8123.5
-------------+---------------------------------
    combined |      210       22155       22155

unadjusted variance   180070.92
adjustment for ties     -130.90
                     ----------
adjusted variance     179940.02

Ho: sbp(sex==female) = sbp(sex==male)
             z =   1.379
    Prob > |z| =   0.1679

. tabstat sbp, by(sex) stat(median)

Summary for variables: sbp
     by categories of: sex (Sex: numeric)

   sex |       p50
-------+----------
female |       124
  male |       122
-------+----------
 Total |       123
------------------
```

This command will provide Table 20.2. The table shows the p-value (Prob > |z|) of the test, which is 0.167 (greater than 0.05). This indicates that the distribution of systolic BP among males and females is not different (or, the median systolic BP of males and females is not different). With this test result, the median systolic BP of females and males should be reported. To get the *median* systolic PB by sex, use the following command:

tabstat sbp, by(sex) stat(median)

This command will report the median (p50) systolic BP for males and females (Table 20.2).

20.2 Median test

The median test is an alternative to the Mann-Whitney U test. Like the Mann-Whitney U test, this test also compares the difference in medians between two categories/levels on a continuous variable. This test is based on the number of observations below and above the common median. Suppose that we want to determine whether or not the median age of diabetics and nondiabetics is the same in the population. Here, the null hypothesis is "the median age of diabetics and nondiabetics is the same in the population". Use the following commands to get the median test results, and the median age for diabetic and nondiabetic individuals:

median age, by(diabetes)
tabstat age, by(diabetes) stat(median)

The results are shown in Table 20.3. The table (at the bottom) shows the median (p50) age of diabetic (39 years) and nondiabetic (26 years) individuals. The table also shows the frequency distribution of diabetic and nondiabetic individuals above and below the common median, which is 28 years (provided as Total under the "tabstat" command). For instance, 41 individuals with diabetes are aged over 28 years (common median), while only 4 are below 28 years. However, our interest is in the p-value given by the median test. The median test p-value is 0.000 (<0.05), which indicates that the median age of diabetics and nondiabetics is different in the population.

20.3 Wilcoxon signed ranks test

This test is the nonparametric alternative to the paired samples t-test. This test compares the distribution of two related samples (e.g., pre-test and post-test results). The Wilcoxon test converts the scores into ranks and then compares them. For example, in order to evaluate the impact of a training session, you have taken the pre- and post-tests before and after the training session. You want to assess if there is an improvement in the post-test scores compared to pre-test scores due to the training session. Use the following command for the Wilcoxon signed rank test:

signrank post_test = pre_test

Table 20.3 Median test

```
. median age, by(diabetes)

Median test

   Greater |
  than the |   Diabetes mellitus
    median |        No        Yes |     Total
-----------+----------------------+----------
        no |       106          4 |       110
       yes |        59         41 |       100
-----------+----------------------+----------
     Total |       165         45 |       210

          Pearson chi2(1) =  43.4324   Pr = 0.000

   Continuity corrected:
          Pearson chi2(1) =  41.2416   Pr = 0.000

. tabstat age, by(diabetes) stat(median)

Summary for variables: age
     by categories of: diabetes (Diabetes mellitus)

diabetes |       p50
---------+----------
      No |        26
     Yes |        39
---------+----------
   Total |        28
--------------------
```

This command will provide Table 20.4. Just look at the p-value of the test, which is 0.000. This indicates that the pre- and post-test scores (medians) are significantly different. We may, therefore, conclude that the training has significantly improved the post-test scores compared to the pre-test scores. You can get the medians of the pre- and post-test scores by using the “tabstat” command as shown earlier (Section 20.2).

20.4 Kruskal-Wallis test

It is the nonparametric equivalent of the one-way ANOVA test. In this test, scores are converted into ranks, and the mean rank of each group is compared. Suppose that we want to test the hypothesis of whether or not the systolic BP (variable name "sbp") is different among religious groups (Muslim, Hindu, and Christian). Here the null hypothesis is "the systolic BP is not different across the religious groups". Use any of the following commands to test the hypothesis:

Table 20.4 Wilcoxon signed ranks test

```
. signrank post_test = pre_test

Wilcoxon signed-rank test

        sign |      obs   sum ranks    expected
-------------+---------------------------------
    positive |       32         528         264
    negative |        0           0         264
        zero |        0           0           0
-------------+---------------------------------
         all |       32         528         528

unadjusted variance     2860.00
adjustment for ties       -1.50
adjustment for zeros       0.00
                      ----------
adjusted variance       2858.50

Ho: post_tes = pre_test
             z =   4.938
    Prob > |z| =   0.0000
```

kwallis sbp, by(religion)

Or,

median sbp, by(religion)

The first command will provide Table 20.5. Since the p-value of the Chi-square test is 0.973 (greater than 0.05), we are unable to reject the null hypothesis. We may, therefore, conclude that the median systolic BP among the religious groups is not significantly different. The median systolic BP in different religious groups can be obtained by using the "tabstat" command.

20.5 Friedman test

The Friedman test is the nonparametric alternative to the one-way repeated measures ANOVA test. Suppose that we are interested in assessing the mean blood sugar levels at four different time intervals (e.g., at hour-0, hour-7, hour-14, and hour-24) after administration of a drug on 15 study subjects. The objective of this study is to determine whether or not the drug reduces blood sugar levels over time (i.e., whether the average blood sugar levels over time are the same or different).

To conduct this study, we randomly selected 15 individuals from a population and

Table 20.5 Kruskal-Wallis test

```
. kwallis sbp, by(religion)

Kruskal-Wallis equality-of-populations rank test

  +----------------------------+
  |  religion | Obs | Rank Sum |
  |-----------+-----+----------|
  |    MUSLIM | 126 | 13298.50 |
  |     HINDU |  58 |  6175.00 |
  | Christian |  26 |  2681.50 |
  +----------------------------+

chi-squared =     0.054 with 2 d.f.
probability =     0.9733

chi-squared with ties =     0.054 with 2 d.f.
probability =     0.9733
```

measured their blood sugar levels at the baseline, i.e., before administration of a drug (hour-0). All the individuals were then administered a drug (say, drug A), and their blood sugar levels were measured again after 7 hours, 14 hours, and 24 hours. We are interested in knowing if the blood sugar levels over time after giving the drug are the same or not (in other words, whether the drug is effective in reducing the blood sugar levels over time). The variable "time" in the dataset indicates the times of measurement of blood sugar levels. In this example, we have only one treatment group (received drug A) but the outcome is measured (blood sugar) at four different points in time on the same subjects (i.e., we have one treatment group with four levels of measurements on the same subjects). Use the data file <**Data_Repeat_anova_3.dta**> for the exercise.

To perform the Friedman test, we need to install a module (emh package), which does not come preinstalled in Stata. To install the module, use the following command:

ssc install emh

For the Friedman test, Stata needs a dataset which is in long format (our dataset is in long format). If your dataset is in wide format, you need to convert it into long format as discussed in Section 5.12. After installing the module, let us first check the number of study subjects and the mean and median blood sugar levels at different time points by using the following command (Table 20.6):

tabstat sugar, by(time) stat(n mean p50)

Table 20.6 Friedman test

```
. tabstat sugar, by(time) stat(n mean p50)

Summary for variables: sugar
     by categories of: time (Time of measurement)

            time |         N      mean       p50
-----------------+------------------------------
before treatment |        15  110.5333       110
7 hrs after trea |        15     105.2       105
14 hrs after tre |        15  101.5333       100
24 hrs after tre |        15  100.4667        98
-----------------+------------------------------
           Total |        60  104.4333       105
----------------------------------------------

. emh sugar time, strata(subject) anova transformation(rank)

Extended Mantel-Haenszel (Cochran-Mantel-Haenszel) Stratified Test of Association

ANOVA (Row Mean Scores) Statistic:
Q (3) = 27.5625, P = 0.0000
Transformation: Ranks
```

Now, use the following command to perform the Friedman test:

emh sugar time, strata(subject) anova transform(rank)

The basic syntax of the command is:

emh outcome_variable explanatory_variable, strata(repeated variable) anova transform(rank).

The output of the Friedman test is provided at the bottom of Table 20.6.

Table 20.6 shows the summary of the blood sugar levels at different time points of measurement. It shows that the mean and median (p50) blood sugar levels have gradually decreased over time. For instance, the median blood sugar at the baseline was 110, while it was 98 at 24 hours after treatment.

The output of the Friedman test is provided at the bottom of Table 20.6. In the table, Q(3) (=27.56) is the test statistic of the Friedman test, and the p-value of the test is 0.000. This indicates that there is a significant difference in blood sugar levels across 4 time points ($p<0.001$). In other words, the findings suggest that the drug is effective in reducing blood sugar levels since the median blood sugar levels have reduced over time.

21
Analysis of Covariance (ANCOVA)

ANCOVA, or analysis of covariance, is a useful technique to statistically control the extraneous variable(s) [called covariate] for the comparison of means of two or more groups. It is similar to ANOVA. In ANOVA, one can incorporate only the categorical independent variables to have the main effect and interaction. But in ANCOVA, one can incorporate both categorical and quantitative variables in the model, including the interaction between categorical and quantitative independent variables. ANCOVA can be performed as a one-way, two-way, or multivariate ANCOVA technique. Use the data file <**Data_3.dta**> for practice.

21.1 One-way ANCOVA

The purpose of using the one-way ANCOVA test is to assess the difference in the mean of the dependent variable (e.g., systolic BP) against a categorical variable (e.g., sex, or effect of a drug) after controlling for one or more quantitative variables [called covariates, such as age and diastolic BP] in the model. The one-way ANCOVA test involves at least three variables, such as:

- One quantitative *dependent* variable (e.g., systolic BP, post-test score, or blood sugar level);
- Only one categorical *independent* variable with two or more levels (e.g., sex, type of intervention, or type of drug); and
- One (or more) *covariate* (continuous quantitative variable), e.g., diastolic BP, age, pre-test score or baseline blood sugar level.

The covariates to be selected for a model should be one or more continuous variables

that are significantly correlated with the dependent variable. One can also include categorical variables as covariates in the model.

Suppose that a researcher is interested in comparing the effectiveness of 3 drugs (drug A, drug B, and drug C) in reducing systolic BP. To conduct the study, the researcher randomly selected three groups of people and assigned a drug to each group. In this scenario, one-way ANOVA could be used. However, it was observed that the mean age and pre-treatment systolic BP of these three groups were not the same. Since age and pre-treatment systolic BP may influence the effectiveness of the drugs in reducing systolic BP, it requires adjustment for these variables (age and pre-treatment systolic BP) to conclude the results. In such a situation, one-way ANCOVA can be used. In ANCOVA, the independent variable must be a categorical variable (here it is "type of drug"). ANCOVA can adjust more than one covariate, either continuous or categorical.

Another example: Assume that you have organized a staff training. You have taken the pre-and post-tests of the participants to evaluate the effectiveness of the training. Now, you want to conclude if males and females (independent variable "sex") have similar performance in the post-test (dependent variable), after controlling for age and pre-test scores (covariates). If the assumptions are met, one-way ANCOVA is the appropriate test for this situation as well.

Hypothesis

Assume that you want to assess if the mean systolic BP (dependent variable; variable name is “sbp”) is the same among males and females (independent variable; variable name is “sex_1”) after controlling for diastolic BP (covariate; variable name is “dbp”).

> $\mathbf{H_0}$: There is no difference in mean systolic BP between males and females in the population, after controlling for diastolic BP.
>
> $\mathbf{H_A}$: The mean systolic BP of males and females is different in the population, after controlling for diastolic BP.

Assumptions

1. The dependent variable is normally distributed at each level of the independent variable;
2. The variances of the dependent variable at each level of the independent variable are the same (homogeneity of variances);
3. The covariates (if more than one) are not strongly correlated with each other ($r<0.8$);

4. There is a linear relationship between the dependent variable and the covariates at each level of the independent variable;
5. There is no interaction between the covariate (diastolic BP) and the independent variable (sex) [called *homogeneity of regression slopes*].

21.1.1 Commands

Before performing the ANCOVA, it is better to check the descriptive statistics of the dependent variable (systolic BP) at each level of the independent variable (sex). Use the following command to get the descriptive statistics of systolic BP by sex (Table 21.1):

tabstat sbp, by(sex_1) stat(n mean sd)

Table 21.1 Descriptive statistics (unadjusted) of systolic BP by sex

```
. tabstat sbp, by(sex_1) stat(n mean sd)

Summary for variables: sbp
     by categories of: sex_1 (Sex: numeric)

 sex_1 |         N      mean        sd
-------+------------------------------
Female |       133  129.5714  21.37695
  Male |        77  124.5584  17.22108
-------+------------------------------
 Total |       210  127.7333  20.05794
--------------------------------------
```

Now, to perform the ANCOVA, use the first of the following commands. Here, the dependent variable is systolic BP (sbp), the independent variable is sex (sex_1), and the covariate is diastolic BP (dbp).

anova sbp i.sex_1 c.dbp i.sex_1#c.dbp
regress
margins sex_1, atmeans

The outputs of the above commands are displayed in Tables 21.2 (first and second commands) and 21.3 (third command). The prefix "i." when used before the name of an independent variable, tells Stata that it is a categorical variable. It also tells Stata to generate dummy variables (a set of dichotomous or indicator variables) if the categorical variable has more than two levels (see Sections 5.11 and 16.2). On the other hand, the prefix "c." is used before a variable to indicate that the variable is a continuous

Table 21.2 ANCOVA table: between-subjects effects

```
. anova sbp i.sex_1 c.dbp i.sex_1#c.dbp

                           Number of obs =     210     R-squared     =  0.7188
                           Root MSE      = 10.7136     Adj R-squared =  0.7147

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  60440.1719     3   20146.724     175.52     0.0000
                         |
                   sex_1 |  1.82547561     1  1.82547561       0.02     0.8998
                     dbp |  41991.2066     1  41991.2066     365.84     0.0000
               sex_1#dbp |  .024723806     1  .024723806       0.00     0.9883
                         |
                Residual |  23644.8947   206  114.781042
              -----------+----------------------------------------------------
                   Total |  84085.0667   209  402.320893
. regress

      Source |       SS       df       MS              Number of obs =     210
-------------+------------------------------           F(  3,   206) =  175.52
       Model |  60440.1719     3   20146.724           Prob > F      =  0.0000
    Residual |  23644.8947   206  114.781042           R-squared     =  0.7188
-------------+------------------------------           Adj R-squared =  0.7147
       Total |  84085.0667   209  402.320893           Root MSE      =  10.714

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       sex_1 |
       Male  |   1.569876   12.44837     0.13   0.900    -22.97267    26.11242
         dbp |   1.458842   .0732245    19.92   0.000     1.314476    1.603207
             |
 sex_1#c.dbp |
       Male  |   .0022405   .1526608     0.01   0.988    -.2987373    .3032184
             |
       _cons |   6.348646   6.254367     1.02   0.311    -5.982131    18.67942
------------------------------------------------------------------------------
```

variable. The follow-up "regress" command without any argument will display results in the form of a regression table (Table 21.2). The last command will provide the predicted (adjusted) mean of systolic BP at each level of sex considering the average value of the covariate (diastolic BP) (Table 21.3).

You can get the pairwise comparison of predicted means of the dependent variable when the main effect of the independent variable (with more than two levels) is statistically significant. If the main effect is not significant (in our example, sex is not significantly associated with systolic BP; Table 21.2), it is not really necessary. However, for the purpose of demonstration, we have performed the pairwise comparison with the Bonferroni option, using the following command:

Table 21.3 Mean systolic BP by sex after adjustment for diastolic BP

```
. margins sex_1, atmeans

Adjusted predictions                            Number of obs   =        210

Expression   : Linear prediction, predict()
at           : 0.sex_1          =    .6333333 (mean)
               1.sex_1          =    .3666667 (mean)
               dbp              =    82.76667 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       sex_1 |
     Female  |   127.0921   .9372843   135.60   0.000     125.2442      128.94
       Male  |   128.8475   1.282687   100.45   0.000     126.3186    131.3763
------------------------------------------------------------------------------
```

pwcompare sex_1, mcompare(bon) effects

This command will provide a comparison of adjusted means of the dependent variable (systolic BP) between the levels (categories) of sex (Table 21.4). The option "effects" used with the command is to get the p-values.

21.1.2 Interpretation: One-way ANCOVA

Table 21.1 shows the mean and SD of systolic BP by sex. The table shows that the unadjusted mean systolic BP of females is 129.5 mmHg and that of males is 124.5 mmHg.

Table 21.2 shows the results of one-way ANCOVA. This is the main table to interpret the results. We have tested the null hypothesis that the population mean of systolic BP in males and females is the same after controlling for diastolic BP. Look at the p-value (Prob > F) for sex (sex_1) in the table, which is 0.899 (main effect). Since the p-value is >0.05, we cannot reject the null hypothesis. We may, therefore, conclude that the mean systolic BP of males and females in the population is not different after controlling for diastolic BP.

We can also assess the influence (association) of the covariate (diastolic BP) on the dependent variable (systolic BP). The p-value for diastolic BP is 0.000, which is highly significant. This indicates that there is a significant association between systolic and diastolic BP after controlling for sex.

However, before we conclude the results, it is important to check whether the assumption of homogeneity of regression slopes is violated or not. To check this, we need to look at the significance level (p-value) of the interaction term (sex_1#dbp) in the table (Table 21.2). We can see that the p-value for interaction is 0.988 (>0.05), suggesting that there is no interaction. This indicates that the *homogeneity of regression slopes* assumption is not violated. A p-value of <0.05 (i.e., if there is an interaction) suggests the regression slopes are not homogeneous and the ANCOVA test is inappropriate.

Table 21.4 Pairwise comparison of adjusted means

```
. pwcompare sex_1, mcompare(bon) effects

Pairwise comparisons of marginal linear predictions

Margins      : asbalanced

note: option bonferroni ignored since there is only one comparison
------------------------------------------------------------------------------
                |                           Unadjusted           Unadjusted
                |   Contrast   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
          sex_1 |
Male vs Female  |   1.569876   12.44837     0.13   0.900    -22.97267    26.11242
------------------------------------------------------------------------------
```

Table 21.3 shows the outputs of the "margins" command. Margins are the statistics calculated from predictions of a previously analyzed model. In our example, we have used the option "atmeans" with the "margins" command (there are other options like "asbalanced" and "asobserved") to get the predicted values (adjusted values) considering the average value of the covariate. Therefore, the "margins" command has provided us with the adjusted (predicted) mean of systolic BP for each level of sex, considering the average value of diastolic BP (82.76 mmHg). We can see that the mean systolic BP of females is 127.09 mmHg, while it is 128.84 mmHg for males after adjusting for the average value of diastolic BP (the adjusted means are different from the unadjusted means as shown in Table 21.1).

Table 21.4 shows the pairwise comparison of the predicted means. This analysis is not necessary in our example since the independent variable (sex) is not significantly associated with the dependent variable. If the independent variable has more than two levels and is statistically significant, then the table for pairwise comparison is important.

Table 21.4 shows that there is no significant difference in mean systolic BP between

males and females as the p-value is 0.90 (>0.05). The analysis ignored the Bonferroni option since the variable "sex" has only two levels. The Bonferroni test is a type of multiple comparison test (there are other tests for pairwise comparisons, such as Scheffe's, Sidak's, and Tukey's tests) used for pairwise comparison of means. The Bonferroni correction is done to adjust the type I errors when multiple pairwise tests are performed on a single variable.

21.2 Two-way ANCOVA

In two-way ANCOVA, there are two independent categorical variables with two or more levels/categories, while in one-way ANCOVA, there is only one independent categorical variable with two or more levels. At least four variables are involved in the analysis of two-way ANCOVA. They are:

- One continuous *dependent* variable (e.g., diastolic BP, blood sugar, or post-test score);
- Two categorical *independent* variables (with two or more levels) [e.g., occupation, diabetes, or type of drug]; and
- One or more continuous *covariates* (e.g., age, systolic BP, or income).

Two-way ANCOVA provides information, after controlling for the covariate(s), on:

- Whether there is a significant main effect of the first independent variable (e.g., occupation) on the dependent variable;
- Whether there is a significant main effect of the second independent variable (e.g., diabetes) on the dependent variable; and
- Whether there is an interaction between the independent variables (e.g., occupation and diabetes).

Suppose that we want to assess, after controlling for age (covariate):

1. Whether or not occupation influences the diastolic BP (i.e., is the mean diastolic BP same in different occupational groups?);
2. Whether or not diabetes influences the diastolic BP (i.e., is the mean diastolic BP same for diabetics and non-diabetics?); and
3. Does the influence of occupation on diastolic BP depend on the presence of diabetes (i.e., is there an interaction between occupation and diabetes)?

Questions 1 and 2 refer to the *main effect*, while question 3 explains the interaction of two independent variables (occupation and diabetes) on the dependent variable

(diastolic BP). For the analysis, we will use the data file <**Data_3.dta**>. The variable names are for diastolic BP is “dbp”, occupation is “occupation (1= govt. job; 2= private job; 3= business; 4= others)”, diabetes is “diabetes1 (0= no diabetes; 1= have diabetes)”, and age is “age”.

Assumptions

All the assumptions for the one-way ANCOVA are applicable to two-way ANCOVA.

21.2.1 Commands

To perform the two-way ANCOVA, use the following command. The variables included in the analysis are dbp (diastolic BP; as dependent variable), occupation, diabetes1, and age.

anova dbp i.occupation i.diabetes1 c.age i.occupation#i.diabetes1
regress

The first command is the basic command for ANCOVA. The follow-up command "regress" will display the outputs in the form of a regression table, as we have seen earlier in one-way ANCOVA. The outputs of both the commands are displayed in Table 21.5.

To get the predicted (adjusted) mean of diastolic BP at each level of occupation and diabetes considering the average values of age (covariate), use the following command:

margins occupation diabetes1, atmeans

The outputs are displayed in Table 21.6. You can also get the marginal means for a combination of occupation and diabetes (e.g., govt. job with diabetes, govt. job without diabetes, etc.), if you use the following command (outputs not shown):

margins occupation#diabetes1
Or,
margins occupation#diabetes1, atmeans

You can generate a plot of predicted values (adjusted means) of the dependent variable for the independent variables in the model. To generate a plot of adjusted mean diastolic BP for occupation and diabetes, use the following commands successively (Fig 21.1):

margins occupation, within(diabetes1) atmeans
marginsplot

Table 21.5 Results of two-way ANCOVA

```
. anova dbp i.occupation i.diabetes1 c.age i.occupation#i.diabetes1

                           Number of obs =     210     R-squared     =  0.0133
                           Root MSE      = 11.9008     Adj R-squared = -0.0260

                  Source |  Partial SS    df       MS           F     Prob > F
    ---------------------+-----------------------------------------------------
                   Model |  384.062807     8  48.0078509       0.34     0.9499
                         |
              occupation |   161.42117     3  53.8070567       0.38     0.7676
                diabetes1 |  5.41875169     1  5.41875169       0.04     0.8451
                     age |   15.420277     1   15.420277       0.11     0.7418
    occupation#diabetes1 |  126.776448     3  42.2588161       0.30     0.8265
                         |
                 Residual |  28467.5039   201  141.629372
    ---------------------+-----------------------------------------------------
                    Total |  28851.5667   209  138.045774

. regress

      Source |       SS       df       MS              Number of obs =     210
-------------+------------------------------           F(  8,   201) =    0.34
       Model |  384.062807     8  48.0078509           Prob > F      =  0.9499
    Residual |  28467.5039   201  141.629372           R-squared     =  0.0133
-------------+------------------------------           Adj R-squared = -0.0260
       Total |  28851.5667   209  138.045774           Root MSE      =  11.901

------------------------------------------------------------------------------------
                 dbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------------+--------------------------------------------------------------
          occupation |
         PRIVATE JOB  |  -1.370739   2.507417    -0.55   0.585    -6.314954    3.573477
            BUSINESS  |  -.9578347   2.602651    -0.37   0.713    -6.089838    4.174168
              OTHERS  |  -3.433968   2.561338    -1.34   0.182    -8.484508    1.616571
                      |
           diabetes1 |
                yes  |  -1.513519   4.127483    -0.37   0.714    -9.652241    6.625204
                 age |  -.0365282   .1107028    -0.33   0.742    -.2548161    .1817597
                      |
occupation#diabetes1 |
    PRIVATE JOB#yes  |  -1.536793   6.179707    -0.25   0.804    -13.72217    10.64858
       BUSINESS#yes  |   2.673676    5.64161     0.47   0.636    -8.450656    13.79801
         OTHERS#yes  |   3.312635   5.554671     0.60   0.552    -7.640267    14.26554
                      |
                _cons |   85.12535   3.322763    25.62   0.000     78.57341    91.67729
------------------------------------------------------------------------------------
```

You can perform the pairwise comparison of predicted means of the dependent variable when the main effects of the independent variables are statistically significant. If the main effect is not significant for an independent variable, it is not necessary to do the pairwise comparison test. In our example, the main effects of both occupation (p= 0.76) and diabetes (p= 0.84) are not statistically significant. However, for the purpose

Table 21.6 Adjusted means (predicted values) of diastolic BP by occupation and diabetes

```
. margins occupation diabetes1, atmeans

Adjusted predictions                            Number of obs   =        210

Expression   : Linear prediction, predict()
at           : 1.occupation    =    .2857143 (mean)
               2.occupation    =    .2333333 (mean)
               3.occupation    =    .2333333 (mean)
               4.occupation    =     .247619 (mean)
               0.diabetes1     =    .7857143 (mean)
               1.diabetes1     =    .2142857 (mean)
               age             =    26.51429 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  occupation |
   GOVT JOB  |   83.83251   1.549116    54.12   0.000      80.7779    86.88711
PRIVATE JOB  |   82.13245   1.717086    47.83   0.000     78.74664    85.51827
   BUSINESS  |    83.4476   1.712003    48.74   0.000     80.07181    86.82339
     OTHERS  |   81.10839   1.663185    48.77   0.000     77.82886    84.38792
             |
   diabetes1 |
         no  |   82.76318     .92934    89.06   0.000     80.93067    84.59569
        yes  |   82.33521   1.830683    44.98   0.000      78.7254    85.94501
------------------------------------------------------------------------------
```

of demonstration, we have performed the pairwise comparisons with the Bonferroni option by using the following command:

pwcompare occupation diabetes1, mcompare(bon) effects

This command will provide a comparison of adjusted mean diastolic BP within the levels (categories) of occupation and diabetes (Table 21.7). The option "effects", used with this command, is for obtaining the p-values of the test.

21.2.2 Interpretation: Two-way ANCOVA

Table 21.5 shows the results of the two-way ANCOVA test. We have tested the null hypothesis that:

- The mean diastolic BP (in the population) among the occupational groups is the same after controlling for age and diabetes;
- The mean diastolic BP (in the population) among diabetics and non-diabetics is the same after controlling for age and occupation; and

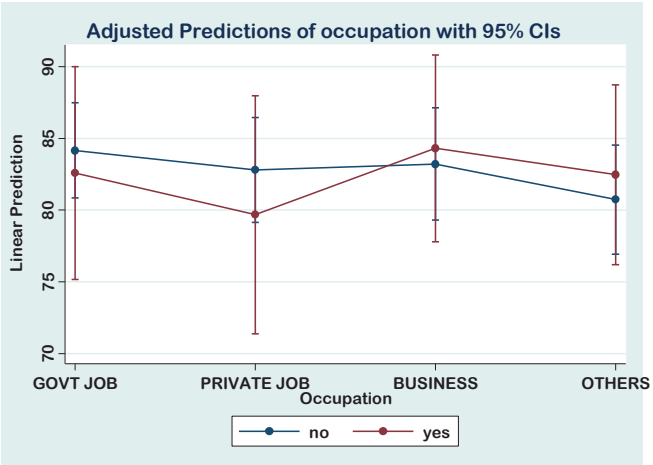


Figure 21.1 Adjusted mean diastolic BP at different levels of occupation by diabetes

- There is no interaction between occupation and diabetes after controlling for age.

Look at the p-values for occupation, diabetes, and the interaction term "occupation#diabetes1" in Table 21.5. They are 0.767, 0.845, and 0.826, respectively, indicating that none of them is statistically significant (we are unable to reject any of the null hypotheses). This means that occupation (after controlling for age and diabetes) and diabetes (after controlling for age and occupation) do not have any significant influence on diastolic BP. There is also no interaction between occupation and diabetes after controlling for age. However, we should always check the p-value of the interaction first. If the interaction is significant (p-value <0.05), then the main effects (of occupation and diabetes) are not important because the effect of one independent variable is dependent on the levels of the other independent variable.

We also have information about the influence of covariates on the dependent variable. We can see (Table 21.5) that the p-value for age is 0.741, which is not statistically significant. This indicates that there is no significant association between age and diastolic BP after controlling for occupation and diabetes.

Table 21.6 shows the adjusted (predicted) mean (Margin column) diastolic BP (dependent variable) at different levels of the independent variables (occupation and diabetes). As an example, the adjusted mean diastolic BP of the government job holders is 83.8 mmHg and that of diabetics (diabetes1 yes) is 82.3 mmHg.

Table 21.7 Pairwise comparison test for occupation and diabetes

```
. pwcompare occupation diabetes1, mcompare(bon) effects

Pairwise comparisons of marginal linear predictions

Margins      : asbalanced

---------------------------
             |    Number of
             |  Comparisons
-------------+-------------
  occupation |            6
   diabetes1 |            1
---------------------------

-------------------------------------------------------------------------------------
                         |                            Bonferroni           Bonferroni
                         |   Contrast   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------------+-----------------------------------------------------------
              occupation |
PRIVATE JOB vs GOVT JOB  |  -2.139135   3.089479    -0.69   1.000    -10.37143    6.093161
   BUSINESS vs GOVT JOB  |   .3790031   2.821173     0.13   1.000    -7.138358    7.896364
     OTHERS vs GOVT JOB  |   -1.77765   2.778096    -0.64   1.000     -9.18023    5.624929
BUSINESS vs PRIVATE JOB  |   2.518138    3.00248     0.84   1.000    -5.482339    10.51862
  OTHERS vs PRIVATE JOB  |   .3614848   2.963006     0.12   1.000     -7.53381    8.256779
     OTHERS vs BUSINESS  |  -2.156654   2.677556    -0.81   1.000    -9.291331    4.978024
                         |
               diabetes1 |
              yes vs no  |  -.4011393   2.050795    -0.20   0.845    -4.444971    3.642693
-------------------------------------------------------------------------------------
```

Table 21.7 depicts the pairwise comparison of adjusted mean diastolic BP within different occupational groups and diabetes. *When the independent variable(s) with more than two levels is significantly associated with the dependent variable, a pairwise comparison is required*. Examine the p-values (P> |t|) in Table 21.7. Since all the p-values are >0.05, there is no significant difference in mean diastolic BP among the occupational groups and diabetes.

Figure 21.1 plotted the adjusted mean diastolic BP with 95% CI of different occupational groups disaggregated by diabetes. Finally, from the data, we can conclude that the diastolic BP is not influenced (there is no association) by occupation and diabetes after controlling for age and the independent variables (diabetes for occupation and occupation for diabetes) included in the model.

22

Miscellaneous

In this chapter, we will discuss two important and useful statistical methods that are frequently needed during data management and analysis, such as how to test the reliability of a scale and develop the wealth quintiles.

22.1 Reliability of scales: Cronbach's alpha

When researchers select a scale (e.g., a scale to measure depression) for their study, it is important to check that the scale is reliable. One of the ways to check the internal consistency (reliability) of a scale is to calculate the Cronbach's alpha coefficient. Cronbach's alpha indicates the degree to which the items that make up the scale correlate with each other in the group.

Ideally, Cronbach's alpha coefficient should have a value above 0.7 to indicate that the scale is reliable. However, this value is sensitive to the number of items on the scale. If the number of items on the scale is less than 10, Cronbach's alpha coefficient tends to be low. In such a situation, it is appropriate to use the "*average interitem correlation*". The optimum range of the average (mean) interitem correlation value is between 0.2 and 0.4. Use the data file <**Data Cronb. dta**> for practice.

Before using the procedure, be sure that all the negatively worded values are "reversed" by recoding (see Section 5.2). If this is not done, it will produce a very low (or negative) value of Cronbach's alpha coefficient. Hopefully, Stata can automatically reverse the negatively worded values if an appropriate option is used with the main command (see below). Assume that a researcher has used a scale to measure depression. The scale has 4 items (questions), q1, q2, q3, and q4. To get the Cronbach's alpha coefficient, use the following commands:

alpha q1-q4
alpha q1-q4, std
alpha q1-q4, std item detail
alpha q1-q4, item reverse(q3 q4)

The first command will provide Cronbach's alpha coefficient based on unstandardized items (Table 22.1) (q1-q4 indicates the variables from q1 to q4). When the option "std" is used (second command), Stata will provide the coefficient based on standardized values of the items (Table 22.1). This option is used when items are not measured on the same scale. It also provides an unbiased estimate. Our suggestion is to use the option "std" to get the alpha coefficient even though items are on the same scale of measurement.

The third command (with the use of options "std", "item", and "detail"), Stata will provide the item-wise values of coefficients in a table and an interitem correlation matrix (Table 22.2). The last command [with the use of option "reverse(q3 q4)"] will reverse the coding of variables q3 and q4. This option is used when one or more negatively worded variables need to be reversed (we don't have such a problem in our data).

Table 22.1 Cronbach's alpha coefficients

```
. alpha q1-q4

Test scale = mean(unstandardized items)

Average interitem covariance:     .6564501
Number of items in the scale:            4
Scale reliability coefficient:      0.8390


. alpha q1-q4, std

Test scale = mean(standardized items)

Average interitem correlation:      0.5675
Number of items in the scale:            4
Scale reliability coefficient:      0.8400
```

22.1.1 Interpretation

Table 22.1 shows the unstandardized (0.839) and standardized (0.840) values of Cronbach's alpha coefficients. The standardized value is especially important when all the items are not on the same scale of measurement. It also provides an unbiased estimate.

In our example, the standardized Cronbach's alpha coefficient is 0.840, which indicates a very good correlation among items on the scale (i.e., the scale is reliable).

Table 22.2 Item-wise values of Cronbach's alpha

```
. alpha q1-q4, std item detail

Test scale = mean(standardized items)
                                                                  average
                                  item-test     item-rest       interitem
Item          |  Obs  Sign    correlation   correlation     correlation      alpha
--------------+-----------------------------------------------------------------
q1            |   60    +       0.7761        0.5991          0.6178       0.8290
q2            |   60    +       0.8444        0.7101          0.5429       0.7808
q3            |   60    +       0.8257        0.6789          0.5633       0.7947
q4            |   60    +       0.8416        0.7054          0.5460       0.7830
--------------+-----------------------------------------------------------------
Test scale    |                                               0.5675       0.8400
-------------------------------------------------------------------------------

Interitem correlations (obs=60 in all pairs)

        q1      q2      q3      q4
q1  1.0000
q2  0.5122  1.0000
q3  0.4911  0.6346  1.0000
q4  0.5483  0.6295  0.5892  1.0000
```

However, before considering the value of Cronbach's alpha coefficient, look at the "Interitem correlations matrix" displayed at the bottom of Table 22.2. All the values in the matrix must be *positive* (all the values are positive in our example). The presence of one or more negative values indicates that some of the items have not been "reverse scored" correctly. This information is also provided in the first part of the table under the "Sign" column (all the items have a positive sign).

The "item-rest correlation" in Table 22.2 indicates the degree to which each item correlates with the total score. In our example, the values for q1 to q4 are 0.60, 0.71, 0.67, and 0.70, respectively. A small value (<0.30) for any item could be a problem. If the Cronbach's alpha coefficient (alpha) and item-rest correlation values for any item are small (<0.7 and <0.3, respectively), one may consider omitting the item from the scale that has a small value. In our example, there is no such problem.

If the number of items is small on the scale (fewer than 10), it may be difficult to get a reasonable Cronbach's alpha coefficient value. In such a situation, consider the average interitem correlation value provided in Table 22.2. In this example, the average

interitem correlation value ranges from 0.542 to 0.617, and the scale mean (average) is 0.567).

22.2 Constructing wealth quintiles

The wealth index is an indicator of the economic status of households or individuals that is commonly used in demographic health and other surveys. The measure of the wealth index may be linked with inequalities in individual characteristics, use of health and other services, and health outcomes. The wealth index is commonly calculated from information on dwelling and household characteristics, access to a variety of consumer goods and services, and assets, which together are used as a measure of the economic status of an individual. The wealth index is constructed using the household asset data via principal component analysis (PCA).

After calculating the composite wealth index scores for each individual, they are categorized into wealth quintiles. A quintile of a dataset represents 20% (one-fifth) of a given sample. Therefore, the calculated wealth index, after arranging them into ascending order, is classified into five categories, or quintiles. The lowest index group (the first quintile) is the poorest section of the sample, while the highest index group (the fifth quintile) is the richest section of the sample. In this section, we will discuss how to construct the wealth quintiles from household information.

We will use the data file <**Wealth.dta**> for this exercise. There are 18 variables in this data file with information on household assets and other characteristics. We will use all this information to construct the wealth quintiles using the PCA technique. Follow the succeeding steps to construct the wealth quintiles.

Step 1: All the variables to be used for constructing wealth quintiles must be dichotomous variables with the coding scheme of 0/1. In our dataset, there are 18 variables, of which two are categorical variables with more than two levels (water and toilet). We need to dichotomize them with a 0/1 coding scheme. Let us first check the value labels and coding schemes of the variables “water” and “toilet” by using the following command:

label list water toilet

Table 22.3 shows the code numbers and value labels of the variables. The table shows that the variable "water" has 14 categories and the variable "toilet" has 11 categories. We need to dichotomize (0/1) both these variables using some guidelines (e.g., demo-

graphic health survey guidelines). Let us consider that the categories (code numbers) 11 to 13, 21, and 91 are the safe sources of drinking water and will code them as 1. The other categories will be coded as 0. Similarly, for the variable "toilet (use of latrine)", let us consider the code numbers 11 to 13, and 22 as hygienic practices, and will code them as 1, while the others will be coded as 0. Use the following commands to recode the variable "water":

```
gen water1=0
replace water1=1 if water==11
replace water1=1 if water==12
replace water1=1 if water==13
replace water1=1 if water==21
replace water1=1 if water==91
```

Table 22.3 Coding schemes of water and toilet

```
. label list water toilet
water:
          11 Piped water into dwelling
          12 Piped to yard/plot
          13 Public tap/stand pipe
          21 Tubewell or borehole
          31 Protected well
          32 Unprotected well
          41 Protected spring
          42 Unprotected spring
          51 Rainwater
          61 Tanker truck
          71 Cart with small tank
          81 Surface water (River/Lake/pond/stream/canal)
          91 Bottled water
          96 Other
toilet:
          11 Flush or pour slush toilet flush to piped sewer system
          12 Flush to septic tank
          13 Flush to pit latrine
          14 Flush to somewhere else
          15 Flush don't know where
          22 Pit latrine with slab
          23 Pit latrine without slab/open pit
          31 Bucket toilet
          41 Hanging toilet/hanging latrine
          51 No facility/bush/field
          96 Other
```

All these commands will generate a new variable “water1” with the coding scheme of 0/1 as stated before. Now to label the new variable (water1) and put the value labels,

use the following commands:

lab var water1 "drinking water source"
la de water1 0"unsafe " 1"safe"
la values water1 water1

In the same manner, recode the second variable “toilet” as “toilet1”.

Step 2: Once the multinominal variables are dichotomized, check the prevalence (relative frequency) of all the variables to be included in the PCA by using the following command:

sum electricity-toilet1

This command will provide Table 22.4 (“electricity-toilet1” as used with the command indicates all the variables from electricity to toilet1). We can see that all the variables are coded as 0/1 (columns Min and Max). The column "Mean" in the table indicates the relative frequency (proportion) of code 1. For example, the mean of the variable "electricity" is 0.42. This indicates that 42% of the subjects use electricity. Now, identify the variables with very small proportions (<0.01 or <1%). In our example, the variables "car" and "boat" have very low proportions (0.005 or 0.5% and 0.004 or 0.4%, respectively) and we will not include them in the PCA.

Step 3: Now we will do the principal component analysis (PCA) of all the variables except for “car” and “boat” and calculate the wealth index by using the following commands:

pca electricity radio tv mobile refrigerator almirah table chair watch ///
cycle motorcycle rikshow hhland firmland water1 toilet1, factor(1)
Or,
pca electricity - motorcycle rikshow - toilet1, factor(1)
predict comp1
ren comp1 w_index

The first or second command will do the PCA (output not shown). The third command (which must be used after performing the PCA) will generate a new variable "comp1" with a wealth index for all the study subjects. The last command will rename "comp1" to "w_index" (we did it for our understanding).

Table 22.4 Relative frequency of the variables

```
. sum electricity-toilet1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
 electricity |       1185    .4236287    .4943416          0          1
       radio |       1185    .1316456    .3382478          0          1
          tv |       1185    .2962025    .4567742          0          1
      mobile |       1185    .8185654    .3855406          0          1
refregerator |       1185    .0270042    .1621641          0          1
-------------+---------------------------------------------------------
     almirah |       1185    .4194093    .4936707          0          1
       table |       1185    .5738397    .4947264          0          1
       chair |       1185    .5848101    .4929628          0          1
       watch |       1185    .5797468    .4938079          0          1
       cycle |       1185    .5409283    .4985325          0          1
-------------+---------------------------------------------------------
  motorcycle |       1185    .0801688     .271669          0          1
         car |       1185    .0059072    .0766631          0          1
        boat |       1185    .0042194    .0648472          0          1
     rikshow |       1185    .1063291    .3083885          0          1
      hhland |       1185    .9611814    .1932439          0          1
-------------+---------------------------------------------------------
    firmland |       1185    .3772152    .4848941          0          1
      water1 |       1185    .8700422     .336399          0          1
     toilet1 |       1185    .8624473    .3445754          0          1
```

Table 22.5 Wealth quintiles

```
. la de w_quintile 1"poorest" 2"poorer" 3"middle" 4"richer" 5"richest"

. la values w_quintile w_quintile

. tab w_quintile

5 quantiles |
 of w_index |      Freq.     Percent        Cum.
------------+-----------------------------------
    poorest |        237       20.00       20.00
     poorer |        237       20.00       40.00
     middle |        238       20.08       60.08
     richer |        236       19.92       80.00
    richest |        237       20.00      100.00
------------+-----------------------------------
      Total |      1,185      100.00
```

Step 4: We will now construct the wealth quintiles from the wealth index variable (w_index) by using the following command:

xtile w_quintile=w_index, nq(5)

The above command will construct the wealth quintiles in a new variable “w_quintile”. Finally, label the new variable and give the value levels by using the first two commands below, and check the wealth quintiles by generating a frequency distribution table using the last command. The outputs are provided in Table 22.5.

lab var w_quintile "wealth quintile"
la de w_quintile 1"poorest" 2"poorer" 3"middle" 4"richer" 5"richest"
la values w_quintile w_quintile
tab w_quintile

References

1. Acock AC. (2014). A Gentle Introduction to Stata. (4th Edition). Stata Press, StataCorp LP, Texas
2. Advanced Research Computing: Statistical Methods and Data Analytics. Regression with Stata chapter 2 – regression diagnostics: UCLA. https://stats.oarc.ucla.edu/stata/webbooks/reg/chapter2/stata-webbooksregressionwith-statachapter-2-regression-diagnostics/
3. Altman DG. (1992). Practical Statistics for Medical Research (1st Edition). Chapman & Hill.
4. Anderson M, Nelson A. Data analysis: Simple statistical tests. FOCUS on Field Epidemiology: UNC School of Public Health. North Carolina Centre for Public Health Preparedness; Vol 3(6).
5. Barros AJD, Hirakata VN. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. BMC Medical Research Methodology 2003; 3(21):1-13. http://www.biomedcentral.com/1471-2288/3/21
6. Bergmire-Sweat D, Nelson A, FOCUS Workgroup. Advanced Data Analysis: Methods to Control for Confounding (Matching and Logistic Regression). Focus on Field Epidemiology: UNC School of Public Health. North Carolina Centre for Public Health Preparedness; Volume 4, Issue 1.
7. Chan YH. Biostatistics 103: Qualitative Data – Tests of Independence. Singapore Med J 2003; Vol 44(10):498-503.
8. Chan YH. Biostatistics 104: Correlational Analysis. Singapore Med J 2003; Vol 44(12):614-619.
9. Chan YH. Biostatistics 201: Linear Regression Analysis. Singapore Med J 2004; Vol 45(2):55-61.
10. Chan YH. Biostatistics 202: Logistic regression analysis. Singapore Med J 2004; Vol 45(4):149-153.
11. Chan YH. Biostatistics 203. Survival analysis. Singapore Med J 2004; Vol 45(6):249-256.
12. Chan YH. Biostatistics 3.5. Multinominal logistic regression. Singapore Med J 2005; 46(6):259-268.
13. Daniel WW. (1999). Biostatistics: A Foundation for Analysis in the Health Science (7th Edition). John Wiley & Sons, Inc.

14. Daniels L, Minot N. An introduction to statistics and data analysis using Stata. (2020). SAGE Publications, Inc: USA.
15. Eric Vittinghoff E, Glidden DV, Shiboski SC, McCulloch CE. Regression methods in biostatistics: Linear, Logistic, Survival, and Repeated Measures Models. (2012). 2nd ed. Springer New York.
16. Gordis L. (2014). Epidemiology (5th Edition). ELSEVIER Sounders.
17. Hamilton LC. (2013). Statistics with Stata (8th Edition). Brooks/Cole Cengage Leaning: USA.
18. Hess KR. Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. Stat Med 1995;14(15):1707-23. doi: 10.1002/sim.4780141510.
19. Islam MT, Kabir R, Nisha M. (2021). Learning SPSS without Pain (2nd Edition. ASA publications, Dhaka.
20. Juul S, Frydenberg M. (2014). An introduction to Stata for health researchers (5th Edition). Stata press, StataCorp, Texas.
21. Katz MH. (2009). Study Design and Statistical Analysis: A Practical Guide for Clinicians. Cambridge University Press.
22. Katz MH. (2010).Evaluating Clinical and Public Health Interventions – A Practical Guide to Study Design and Statistics. Cambridge University Press.
23. Katz MH. (2011). Multivariable Analysis: A Practical Guide for Clinicians and Public Health Researchers (3rd Edition). London, Cambridge University Press.
24. Katz MH. Multivariable Analysis: A Primer for Readers of Medical Research. Ann Intern Med 2003; 138:644–650.
25. Khamis H. Measures of Association: How to Choose? JDMS 2008; 24:155–162.
26. Lee J, Chia KS. Estimation of prevalence rate ratios for cross sectional data: an example in occupational epidemiology. British Journal of Industrial Medicine 1993; 50:861-864.
27. Long JS, Freese J. Regression models for categorical dependent variables using Stata. Stata publication. StataCorp LP, Texas.
28. Malek MH, Berger DE, Coburn JW. On the inappropriateness of stepwise regression analysis for model building and testing. Eur J Appl Physiol 2007; 101: 263–264. https://doi.org/10.1007/s00421-007-0485-9.

29. Pfaff T. A brief introduction to Stata with 50+ basic commands. (2009). Institute for Economic Education, University of Münster.

30. Rabe-Hesketh S, Everitt B. A Handbook of Statistical Analyses using Stata. (2004). 3rd ed. Chapman & Hall/CRC.

31. Reboldi G, Angeli F, Verdecchia P. Multivariable Analysis in Cerebrovascular Research: Practical Notes for the Clinician. Cerebrovasc Dis 2013; 35:187–193. DOI: 10.1159/000345491.

32. Schlesselman JJ, Stolley PD. (1982). Case-Control Studies: Design, Conduct, Analysis. Oxford University Press, Oxford, New York.

33. Stata press. (2009). Stata multivariate statistics reference manual release 11. Stata Publication. StataCorp LP, Texas.

34. Stata press. (2012). Data Analysis Using Stata (3rd Edition). StataCorp LP, Texas.

35. Szklo M, Nieto FJ. (2007). Epidemiology: Beyond the Basics (2nd Edition). Jones and Bartlett Publishers.

36. Tabachnik BG, Fidell LS. (2007). Using multivariate statistics (5th Edition). Boston: Pearson Education.

37. Thompson ML, Myers JE, Kriebel D. Prevalence odds ratio or prevalence ratio in the analysis of cross sectional data: what is to be done? Occup Environ Med 1998; 55:272-277.

38. UCLA. Advanced Research Computing: Statistical Methods and Data Analytics; Repeated Measures Analysis with Stata. https://stats.oarc.ucla.edu/stata/seminars/repeated-measures-analysis-with-stata/

39. UCLA. Institute for Digital Research & Institution: Statistical Consulting. https://stats.idre.ucla.edu/stata/output/descriptive-statistics-using-the-summarize-command/

40. Whittingham M, Stephens P, Bradbury R, Freckleton R. Why do we still use stepwise modelling in ecology and behavior? Journal of Animal Ecology 2006; 75(5):1182–1189.

41. Williams R. Using the margins command to estimate and interpret adjusted predictions and marginal effects. The Stata Journal 2012;12(2):308–331.

42. Zwiener I, Blettner M, Hommel G: Survival analysis— part 15 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2011; 108(10):163–9. DOI: 10.3238/arztebl.201 1.0163

Subject index with chapters and sections

Adjusted odds ratio, 13.3, 13.3.1, 17.2.1.1, 17.2.6.1
Adjusted R-squared, 6.2.4.1, 16.1.2, 16.2.3
Analysis of covariance (ANCOVA), 9, 21
Analysis of variance (ANOVA), 9, 11.1, 11.2
Average interitem correlation, 22.1, 22.1.1

Backward LR, 17.2.3
Backward, 16.2.5.1, 17.2.3, 19.2
Bar graph, 7.4, 7.4.1, 7.4.2
Bartlett's test, 11.1.2, 11.1.5
Beta, 16.2.2, 16.2.3
Binary logistic regression, 17, 17.2
Bland Altman test, 9
Bonferroni's test, 11.1.3, 11.1.5, 11.2.3, 12.1.2, 21.1.1, 21.1.2, 21.2.1
Box and Plot, 4.2, 7.3, 11.1.3
Breslow's test, 19.1.3

Censored time, 19
Central tendency, 6.2
Chi-square test of independence, 9, 13.1
Class interval, 5.3
Classification table, 17.2.2.2
Codebook, 3.3.2
Combine data, 5.4
Conditional logistic regression, 17, 17.2.6
Confidence interval, 6.2
Confounding factor, 13.3, 16.2.5, 17.2, 17.2.4
Converting variables, 5.1
Correlation, 15.1
Correlation coefficient, 15.1, 15.1.2, 16.2.4.1
Correlation matrix, 15.1.2, 16.2.4.1, 17.2.2.1, 22.1
Cox regression, 17.3, 17.3.2, 19.2, 19.2.1, 19.2.2
Cronbach's alpha, 22.1
Cross-tabulation, 6.1, 13.1
Cumulative probability of survival, 19, 19.1.2, 19.1.3, 19.2, 19.2.1

Data cleaning 4
Data file generation, 2.1
Data file import, 2.1.3
Data file, 3.1, 3.1.1
Data screening, 4
Data transformation, 3.1.2.5, 5.5
Description, 3.3.1
Dispersion, 6.2
Do-file execution, 3.1.3.3
Do-file, 3.1, 3.1.3
Dummy variable, 5.11, 16.2, 16.2.2, 16.2.4.1, 21.1.1
Durbin-Watson test, 16.2.4.6

Egen, 5.10
Event time, 19
Exponential, 17.1, 17.2.1.1
Extraction of duration, 5.7

Fisher's exact test, 9, 13.1.1, 13.1.2
Forward LR, 17.2.3
Frequency, 6.1
Friedman test, 9, 20.5

Generalized linear model, 17.3.3
Generating data files, 2.1
Graph, 7
Greenhouse-Geisser test, 12.1.2

Hazard ratio, 19.2, 19.2.1
Help, 3.6
Histogram, 7.1, 8.1, 8.1.1
Homogeneity of regression slopes, 21.1, 21.1.2
Homogeneity of variances, 11, 11.1.2, 11.2, 21.1
Homoscedasticity, 16.1, 16.2.4, 16.2.4.4
Hosmer-Lemeshow test, 17.2.2.2
Huynh-Feldt test, 12.1.2
Hypothesis testing, 9

Independent samples t-test, 9, 10.2, 10.2.2, 10.2.3
Interaction, 11.2, 11.2.2, 13.3, 13.3.1, 16.2.2, 17.2.4, 21.1, 21.1.2, 21.2, 21.2.2
Iteration, 19.2.1, 19.2.2
Interitem correlation, 22.1, 22.1.1

Kaplan-Meier method, 19, 19.1, 19.1.2
Kappa estimates, 9
Kendall's tau-b, 15.2
Kolmogorov-Smirnov test
Kruskal-Wallis test, 9, 20.4
Kurtosis, 6.2, 6.2.1

Levene's test, 10.2.1
Likelihood ratio, 17.2.1.1
Line graph, 7.5
Linear regression, 16
Linearity, 16.2.4.2
Log file, 3.1.2
Log rank test, 19.1.2, 19.1.3
Logistic regression diagnostics, 17.2.2
Logistic regression model, 17.1
Logistic regression, 17
Logit transformation, 17.1
Log-minus-log plot, 19.2.4
Long format, 5.12, 20.5

Mann-Whitney U test, 9, 20.1
Marginal value, 18.2, 21.2.1
Margins plot, 18.2
Mauchly's test, 12.1.2
McNemer test, 9
Median survival time, 19, 19.1.2, 19.1.3, 19.2
Median test, 20.2
Multicollinearity, 16.2.4, 16.2.4.1, 16.2.4.4, 17.1, 17.2.2.1, 19.2.4
Multinominal logistic regression, 9, 17, 18
Multiple comparisons, 11.1.3, 11.1.4, 11.2.3, 21.1.2
Multiple linear regression, 9, 16.2, 16.2.2, 16.2.3

Negative predictive value, 17.2.2.2
Non-parametric methods, 20
Normality of data, 4.3, 5.5, 6.2.1, 7.1, 8.1, 8.1.1
Numeric variable, 5.1.1, 5.1.2

Odds ratio (OR), 13.2
One-sample test of proportion, 14.1
One-sample t-test, 9, 10.1
One-way ANCOVA 21.1
One-way ANOVA 11.1
One-way repeated measures ANOVA, 12.1

Outliers, 4.2, 6.2.1, 7.2, 7.3, 8.1.1, 16.2.4.5
Out-of-range error, 4.1
Output file, 3.1, 3.1.2

Paired t-test, 9, 10.3, 10.3.1, 10.3.2
Partial correlation, 15.3
Pearson's chi-square, 13.1.2
Pearson's correlation coefficient, 9, 15.1
Percentile, 6.2, 6.2.1
Peto test, 19.1.2, 19.1.3
Pie chart, 7.6
Poisson regression, 9, 17.3, 17.3.1
Positive predictive value, 17.2.2.2
Post hoc test, 11.1.3, 11.1.4, 11.2.3, 11.1.5
Post-estimation commands, 17.2.2.4, 18.2,
Prevalence odds ratio, 17.3
Prevalence ratio, 17.3
Proportion test, 14
Proportional hazards analysis, 9, 17.3, 19.2
Proportional odds regression, 9, 17
Proportionality assumption, 19.2.4
Pseudo R-squared, 17.2.1.1, 17.2.6.1, 18.1

Q-Q plot, 8.1, 8.1.1
Quartile, 6.2.1
Quintile, 22.2

Repeated measures ANOVA, 12
Recoding, 5.2
Regression coefficient, 16.2, 16.2.3, 16.2.4.1, 16.2.5.3, 17.2.1.1, 17.2.2.1, 17.2.6
Regression diagnostics, 16.2.4, 17.2.2
Regression equation, 16.1, 16.1.2,
Ordinal logistic regression, 17 16.2.3
Relative risk (RR), 13.2
Reliability of scales, 22.1
Relocation of variables, 5.8
ROC curve, 17.2.2.3
R-squared, 16.1.2, 16.2.3, 16.2.4.3, 16.2.5.3

Sample size, 8.1.1, 11.1, 16.2.1, 16.2.3, 17.2.5, 20.1
Scatterplot, 7.2
Scheffe's test, 11.1.3, 21.1.2
Score calculation, 5.6
Sdtest, 10.2.1
Search, 3.6
Sensitivity, 17.2.2.2
Shapiro Wilk test, 8.1, 8.1.1
Simple linear regression, 16.1
Skewness, 6.2, 6.2.1
Skewness-kurtosis (S-K) test, 8.1, 8.1.1
Sorting, 3.4, 6.3, 16.2.4.5
Spearman correlation, 9, 15.2
Specificity, 17.2.2.2
Sphericity assumption, 12.1
Stata files, 3
Stata windows, 1.2
Stepwise, 16.2.5, 16.2.5.1, 16.2.5.2, 16.2.5.3, 17.2.3, 19.2.1
Stratified analysis, 13.3
String variable, 5.1.1
Sub-group, 5.9
Survival analysis, 19, 19.1
Survival curve, 19.1.2, 19.1.3, 19.2.1
Symbols, 3.5
Syntax, 3.2

Tarone-Ware test, 19.1.2, 19.1.3
Test for independence, 16.2.4.6

Time extraction, 5.7
Tolerance, 16.2.4.1
Total score calculation, 5.6
Transformation of data, 5.5
t-test, 9, 10, 10.1, 10.2, 10.3
Tukey's test, 11.1.3, 11.1.5, 21.1.2
Two-sample test of proportion, 14.2
Two-way ANCOVA 21.2
Two-way ANOVA 11.2
Two-way repeated measures ANOVA

Unconditional logistic regression, 17, 17.2.1

Testing of hypothesis, 9
Unequal variances, 11.1.5

Variable insertion, 2.1.4.2
Variance inflation factor (VIF), 16.2.4.1
Version, 1.1

Wealth quintile, 22.2
Wide format, 5.12, 20.5
Wilcoxon rank-sum test, 20.1
Wilcoxon Signed Ranks test, 9, 20.3
Wilk's lambda, 12.1.2