

Biostatistics in Public Health Using STATA

Biostatistics in Public Health Using STATA

Erick L. Suárez
Cynthia M. Pérez
Graciela M. Nogueras
Camille Moreno-Gorrín



CRC Press

Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2016 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

Printed on acid-free paper
Version Date: 20160201

International Standard Book Number-13: 978-1-4987-2199-8 (Hardback)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
<http://www.taylorandfrancis.com>

and the CRC Press Web site at
<http://www.crcpress.com>

To our loved ones

To those who have enlightened our path throughout their knowledge.

Contents

Preface	xi
Acknowledgments	xiii
Authors	xv
1 Basic Commands	1
1.1 Introduction	1
1.2 Entering Stata	2
1.3 Taskbar	2
1.4 Help	3
1.5 Stata Working Directories	4
1.6 Reading a Data File	6
1.7 <i>insheet</i> Procedure	7
1.8 Types of Files	7
1.9 Data Editor	7
2 Data Description	11
2.1 Most Useful Commands	11
2.2 <i>list</i> Command	12
2.3 Mathematical and Logical Operators	12
2.4 <i>generate</i> Command	14
2.5 <i>recode</i> Command	15
2.6 <i>drop</i> Command	16
2.7 <i>replace</i> Command	16
2.8 <i>label</i> Command	16
2.9 <i>summarize</i> Command	17
2.10 do-file Editor	19
2.11 Descriptive Statistics and Graphs	19
2.12 <i>tabulate</i> Command	20
3 Graph Construction	23
3.1 Introduction	23
3.2 Box Plot	23
3.3 Histogram	25
3.4 Bar Chart	25

4 Significance Tests29

4.1 Introduction29

4.2 Normality Test31

4.3 Variance Homogeneity31

4.4 Student's *t*-Test for Independent Samples.....33

4.5 Confidence Intervals for Testing the Null Hypothesis35

4.6 Nonparametric Tests for Unpaired Groups.....35

4.7 Sample Size and Statistical Power36

5 Linear Regression Models41

5.1 Introduction41

5.2 Model Assumptions..... 42

5.3 Parameter Estimation43

5.4 Hypothesis Testing43

5.5 Coefficient of Determination 44

5.6 Pearson Correlation Coefficient45

5.7 Scatter Plot 46

5.8 Running the Model47

5.9 Centering.....47

5.10 Bootstrapping.....49

5.11 Multiple Linear Regression Model.....50

5.12 Partial Hypothesis52

5.13 Prediction54

5.14 Polynomial Linear Regression Model55

5.15 Sample Size and Statistical Power57

5.16 Considerations for the Assumptions of the Linear Regression Model.... 59

6 Analysis of Variance61

6.1 Introduction61

6.2 Data Structure.....62

6.3 Example for Fixed Effects62

6.4 Linear Model with Fixed Effects.....63

6.5 Analysis of Variance with Fixed Effects 64

6.6 Programming for ANOVA65

6.7 Planned Comparisons (before Observing the Data).....68

6.7.1 Comparison of Two Expected Values.....68

6.7.2 Linear Contrast.....69

6.8 Multiple Comparisons: Unplanned Comparisons.....70

6.9 Random Effects72

6.10 Other Measures Related to the Random Effects Model.....74

6.10.1 Covariance74

6.10.2 Variance and Its Components.....75

6.10.3 Intraclass Correlation Coefficienti û75

6.11	Example of a Random Effects Model.....	75
6.12	Sample Size and Statistical Power	78
7	Categorical Data Analysis	81
7.1	Introduction	81
7.2	Cohort Study.....	82
7.3	Case-Control Study	84
7.4	Sample Size and Statistical Power	86
8	Logistic Regression Model	89
8.1	Model Definition.....	89
8.2	Parameter Estimation	90
8.3	Programming the Logistic Regression Model	91
8.3.1	Using <i>glm</i>	92
8.3.2	Using <i>logit</i>	92
8.3.3	Using <i>logistic</i>	93
8.3.4	Using <i>binreg</i>	93
8.4	Alternative Database.....	94
8.5	Estimating the Odds Ratio	95
8.6	Significance Tests	96
8.6.1	Likelihood Ratio Test	96
8.6.2	Wald Test	96
8.7	Extension of the Logistic Regression Model.....	97
8.8	Adjusted OR and the Confounding Effect.....	100
8.9	Effect Modification.....	101
8.10	Prevalence Ratio	102
8.11	Nominal and Ordinal Outcomes.....	103
8.12	Overdispersion.....	109
8.13	Sample Size and Statistical Power	109
9	Poisson Regression Model	113
9.1	Model Definition.....	113
9.2	Relative Risk.....	114
9.3	Parameter Estimation	115
9.4	Example.....	115
9.5	Programming the Poisson Regression Model.....	116
9.6	Assessing Interaction Terms.....	117
9.7	Overdispersion.....	121
10	Survival Analysis	123
10.1	Introduction	123
10.2	Probability of Survival	126
10.3	Components of the Study Design	126
10.4	Kaplan–Meier Method	127

10.5	Programming of $S(t)$	128
10.6	Hazard Function	132
10.7	Relationship between $S(t)$ and $h(t)$	134
10.8	Cumulative Hazard Function	135
10.9	Median Survival Time and Percentiles	136
10.10	Comparison of Survival Curves	137
10.11	Proportional Hazards Assumption	138
10.12	Significance Assessment	139
10.12.1	Log-Rank Test	140
10.12.2	Wilcoxon–Gehan–Breslow Test	141
10.12.3	Tarone–Ware Test	142
10.13	Cox Proportional Hazards Model	143
10.14	Assessment of the Proportional Hazards Assumption	145
10.15	Survival Function Estimation Using the Cox Proportional Hazards Model	146
10.16	Stratified Cox Proportional Hazards Model	146
11	Analysis of Correlated Data	149
11.1	Regression Models with Correlated Data	149
11.2	Mixed Models	154
11.3	Random Intercept	156
11.4	Using the <i>mixed</i> and <i>gllamm</i> Commands with a Random Intercept	157
11.5	Using the <i>mixed</i> Command with Random Intercept and Slope ..	161
11.6	Mixed Models in a Sampling Design	163
12	Introduction to Advanced Programming in STATA	167
12.1	Introduction	167
12.2	do-files	167
12.3	<i>program</i> Command	168
12.4	Log Files	170
12.5	<i>trace</i> Command	170
12.6	Delimiters	171
12.7	Indexing	172
12.8	Local Macros	174
12.9	Scalars	175
12.10	Loops (<i>foreach</i> and <i>forvalues</i>)	176
12.11	Application of <i>matrix</i> and <i>local</i> Commands for Prevalence Estimation	179
	References	183
	Index	185

Preface

This book is intended to serve as a guide to applied statistical analysis in public health using the Stata program. Our motivation for writing this book lies in our years of experience teaching biostatistics and epidemiology, particularly in the academic programs of biostatistics and epidemiology. The academic material is usually covered in biostatistics courses at the master's and doctoral levels at schools of public health. The main focus of this book is the application of statistics in public health. Because of its user-friendliness, we used the Stata software package in the creation of the database and the statistical analysis that will be seen herein. This 12-chapter book can serve equally well as a textbook or as a source for consultation. Readers will be exposed to the following topics: *Basic Commands*, *Data Description*, *Graph Construction*, *Significance Tests*, *Linear Regression Models*, *Analysis of Variance*, *Categorical Data Analysis*, *Logistic Regression Model*, *Poisson Regression Model*, *Survival Analysis*, *Analysis of Correlated Data*, and *Advanced Programming in Stata*. Each chapter is based on one or more research problems linked to public health. We have started with the assumption that the readers of this book have taken at least a basic course in biostatistics and epidemiology. Further, for those readers who are new to Stata, the first three chapters should be read sequentially, as they form the basis of an introductory course to this software.

Erick L. Suárez

University of Puerto Rico

Cynthia M. Pérez

University of Puerto Rico

Graciela M. Noguera

MD Anderson Cancer Center

Camille Moreno-Gorrín

University of Puerto Rico

Acknowledgments

We thank Dr. Kenneth Hess, professor of biostatistics at MD Anderson Cancer Center in Houston, Texas, for his kind comments and suggestions aimed at improving different aspects of this book. We also thank Bob Ritchie for his excellent work in editing this book. We want to acknowledge the support we received from the Department of Biostatistics and Epidemiology of the Graduate School of Public Health, University of Puerto Rico, in the writing of this book. We are very grateful to the many students—particularly Marc Machín and Kristy Zoé Vélez of the MPH program in Biostatistics—who collaborated by reading material for this book.

This book would not have been possible without the financial support that we received from the following grants: CA096297/CA096300 from the National Cancer Institute of the National Institutes of Health and 2U54MD00758 from the National Institute on Minority Health and Health Disparities of the National Institutes of Health.

Authors

Erick L. Suárez is a professor of biostatistics in the Department of Biostatistics and Epidemiology at the University of Puerto Rico Graduate School of Public Health. He has more than 25 years of experience teaching biostatistics at the graduate level and has co-authored more than 75 peer-reviewed publications in chronic and infectious diseases. Dr. Suárez has been a co-investigator of several NIH-funded grants related to cancer, HPV, HCV, and diabetes. He has extensive experience in statistical consulting with biomedical researchers, particularly in the analysis of microarrays data in breast cancer.

Cynthia M. Pérez is a professor of epidemiology in the Department of Biostatistics and Epidemiology at the University of Puerto Rico Graduate School of Public Health. She has taught epidemiology and biostatistics for over 20 years. She has also directed efforts in mentoring and training to public health and medical students at the University of Puerto Rico. She has been the principal investigator or co-investigator of research grants in diverse areas of public health including diabetes, metabolic syndrome, periodontal disease, viral hepatitis, and HPV infection. She is the author or co-author of more than 75 peer-reviewed publications.

Graciela M. Nogueras is a statistical analyst at the University of Texas MD Anderson Cancer Center in Houston, Texas. She is currently enrolled on the PhD program in biostatistics at the University of Texas—Graduate School of Public Health. She has co-authored more than 30 peer-reviewed publications. For the past nine years, she has been performing statistical analyses for clinical and basic science researchers. She has been assisting with the design of clinical trials and animal research studies, performing sample size calculations, and writing the clinical trial reports of clinical trial progress and interim analyses of efficacy and safety data to the University of Texas MD Anderson Data and Safety Monitoring Board.

Camille Moreno-Gorrín is a graduate of the Master of Science Program in Epidemiology at the University of Puerto Rico Graduate School of Public Health. During her graduate studies, she was a research assistant at the Comprehensive Cancer Center of the University of Puerto Rico where she co-authored several articles in biomedical journals. She also worked as a research coordinator for the HIV/AIDS Surveillance System of the Puerto Rico Department of Health, where she conducted research on intervention programs to link HIV patients to care.