Why do you need this book?

Multivariable analysis is confusing! Whether you are performing your first research project or attempting to interpret the output from a multivariable model, you have undoubtedly found this to be true. Basic biostatistics books are of little to no help to you, since their coverage often stops short of multivariable analysis. However, existing multivariable analysis books are too dense with mathematical formulas and derivations and are not designed to answer your most basic questions. Is there a book that steps aside from the math and simply explains how to understand, perform, and interpret multivariable analyses?

Yes. *Multivariable Analysis: A Practical Guide for Clinicians* is precisely the reference that will lead your way. In fact, Dr. Mitchell Katz has asked and answered all of your questions for you!

Why should I do multivariable analysis?

How do I choose which type of multivariable to use?

How many subjects do I need to do multivariable analysis?

What if I have repeated observations of the same persons?

Answers and detailed explanations to these questions and more are found in this book. Also, it is loaded with useful tips, summary charts, figures, and references.

If you are a medical student, resident, or clinician, *Multivariable Analysis: A Practical Guide for Clinicians* will prove an indispensable guide through the confusing terrain of statistical analysis.

This new edition has been fully revised to build on the enormous success of its predecessor. New features include an extensive review of analysis of clustered data, including the use of generalized estimating equations and mixed-effects models, a new chapter on propensity scores, and more detail on Poisson regression and analysis of variance.

Praise for first edition

“This is the first nonmathematical book on multivariable analysis addressed to clinicians. Its range, organization, brevity, and clarity make it useful as a reference, a text, and a guide for self-study. This book is ‘a practical guide for clinicians.’”

Leonard E. Braitman, Ph.D., *Annals of Internal Medicine*

Mitchell H. Katz is Clinical Professor of Medicine, Epidemiology, and Biostatistics at the University of California, San Francisco; he is also Director of the San Francisco Department of Public Health.
Multivariable Analysis
A Practical Guide for Clinicians
Second Edition

Mitchell H. Katz
Department of Medicine, Epidemiology, and Biostatistics, University of California, USA
To my parents, for their unwavering support
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preface</td>
<td>xiii</td>
</tr>
<tr>
<td><strong>1</strong> Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Why should I do multivariable analysis?</td>
<td>1</td>
</tr>
<tr>
<td>1.2 What are confounders and how does multivariable analysis help me to deal with them?</td>
<td>6</td>
</tr>
<tr>
<td>1.3 What are suppressers and how does multivariable analysis help me to deal with them?</td>
<td>9</td>
</tr>
<tr>
<td>1.4 What are interactions and how does multivariable analysis help me to deal with them?</td>
<td>11</td>
</tr>
<tr>
<td><strong>2</strong> Common uses of multivariable models</td>
<td>14</td>
</tr>
<tr>
<td>2.1 What are the most common uses of multivariable models in clinical research?</td>
<td>14</td>
</tr>
<tr>
<td>2.2 How do I choose what type of multivariable analysis to use?</td>
<td>23</td>
</tr>
<tr>
<td><strong>3</strong> Outcome variables in multivariable analysis</td>
<td>24</td>
</tr>
<tr>
<td>3.1 How does the nature of my outcome variable influence my choice of which type of multivariable analysis to do?</td>
<td>24</td>
</tr>
<tr>
<td>3.2 What type of multivariable analysis should I use with an interval outcome?</td>
<td>24</td>
</tr>
<tr>
<td>3.3 What are the different types of analysis of variance and when are they used?</td>
<td>25</td>
</tr>
<tr>
<td>3.4 What should I do if my outcome variable is ordinal or nominal?</td>
<td>27</td>
</tr>
<tr>
<td>3.5 What type of multivariable analysis should I use with a dichotomous outcome?</td>
<td>28</td>
</tr>
</tbody>
</table>
3.6 What type of multivariable analysis should I use with a
time-to-outcome variable?  28
3.7 What type of multivariable analysis should I use with a rare
outcome or a count?  32

4
Type of independent variables in multivariable analysis  35
4.1 What type of independent variables can I use with
multivariable analysis?  35
4.2 What should I do with my ordinal and nominal independent
variables?  35

5
Assumptions of multiple linear regression, multiple
logistic regression, and proportional hazards analysis  38
5.1 What are the assumptions of multiple linear regression,
multiple logistic regression, and proportional hazards
analysis?  38
5.2 What is being modeled in multiple linear regression, multiple
logistic regression, and proportional hazards analysis?  38
5.3 What is the relationship of multiple independent variables to
outcome in multiple linear regression, multiple logistic
regression, and proportional hazards analysis?  42
5.4 What is the relationship of an interval-independent variable to
the outcome in multiple linear regression, multiple logistic
regression, and proportional hazards analysis?  43
5.5 What if my interval-independent variable does not have a
linear relationship with my outcome?  46
5.6 Assuming that my interval-independent variable fits a linear
assumption, is there any reason to group it into interval
categories or create multiple dichotomous variables?  51
5.7 What are the assumptions about the distribution of the
outcome and the variance?  52
5.8 What should I do if I find significant violations of the
assumptions of normal distribution and equal variance in my
multiple linear regression analysis?  55
5.9 What are the assumptions of censoring?  56
5.10 How likely is it that the censoring assumption is valid
in my study?  59
5.11 How can I test the validity of the censoring assumption for my data? 64

6 Relationship of independent variables to one another 68

6.1 Does it matter if my independent variables are related to each other? 68
6.2 How do I assess whether my variables are multicollinear? 69
6.3 What should I do with multicollinear variables? 71

7 Setting up a multivariable analysis 73

7.1 What independent variables should I include in my multivariable model? 73
7.2 How do I decide what confounders to include in my model? 73
7.3 What independent variables should I exclude from my multivariable model? 74
7.4 How many subjects do I need to do multivariable analysis? 77
7.5 What if I have too many independent variables given my sample size? 81
7.6 What should I do about missing data on my independent variables? 87
7.7 What should I do about missing data on my outcome variable? 94

8 Performing the analysis 96

8.1 What numbers should I assign for dichotomous or ordinal variables in my analysis? 96
8.2 Does it matter what I choose as my reference category for multiple dichotomous ("dummied") variables? 97
8.3 How do I enter interaction terms into my analysis? 98
8.4 How do I enter time into my proportional hazards or other survival analysis? 101
8.5 What about subjects who experience their outcome on their start date? 106
8.6 What about subjects who have a survival time shorter than physiologically possible? 107
8.7 What are variable selection techniques? 109
8.8 What value should I specify for tolerance in my logistic regression or proportional hazards model? 114
### Contents

8.9 How many iterations (attempts to solve) should I specify for my logistic regression or proportional hazards model? 114
8.10 What value should I specify for the convergence criteria for my logistic regression or proportional hazards model? 115
8.11 My model won’t converge. What should I do? 115

9 Interpreting the analysis 117

9.1 What information will the printout from my analysis provide? 117
9.2 How do I assess how well my model accounts for the outcome? 117
9.3 What do the coefficients tell me about the relationship between each variable and the outcome? 124
9.4 How do I get odds ratios and relative hazards from the multivariable analysis? What do they mean? 126
9.5 How do I interpret the odds ratio and relative hazard when the independent variable is interval? 129
9.6 How do I compute the confidence intervals for the odds ratios and relative hazards? 130
9.7 What are standardized coefficients and should I use them? 131
9.8 How do I test the statistical significance of my coefficients? 131
9.9 How do I interpret the results of interaction terms? 134
9.10 Do I have to adjust my multivariable regression coefficients for multiple comparisons? 134

10 Checking the assumptions of the analysis 137

10.1 How do I know if my data fit the assumptions of my multivariable model? 137
10.2 How do I assess the linearity, normal distribution, and equal variance assumptions of multiple linear regression? 138
10.3 How do I assess the linearity assumption of multiple logistic regression and proportional hazards analysis? 139
10.4 What are outliers and how do I detect them in my multiple linear regression model? 139
10.5 How do I detect outliers in my multiple logistic regression model? 141
10.6 What about analysis of residuals with proportional hazards analysis? 142
10.7 What should I do when I detect outliers? 142
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.8</td>
<td>What is the additive assumption and how do I assess whether my multiple independent variables fit this assumption?</td>
<td>143</td>
</tr>
<tr>
<td>10.9</td>
<td>What does the additive assumption mean for interval-independent variables?</td>
<td>145</td>
</tr>
<tr>
<td>10.10</td>
<td>What is the proportionality assumption?</td>
<td>146</td>
</tr>
<tr>
<td>10.11</td>
<td>How do I test the proportionality assumption?</td>
<td>148</td>
</tr>
<tr>
<td>10.12</td>
<td>What if the proportionality assumption does not hold for my data?</td>
<td>150</td>
</tr>
<tr>
<td>11</td>
<td>Propensity scores</td>
<td>153</td>
</tr>
<tr>
<td>11.1</td>
<td>What are propensity scores? Why are they used?</td>
<td>153</td>
</tr>
<tr>
<td>12</td>
<td>Correlated observations</td>
<td>158</td>
</tr>
<tr>
<td>12.1</td>
<td>How do I analyze correlated observations?</td>
<td>158</td>
</tr>
<tr>
<td>12.2</td>
<td>How do I calculate the needed sample size for studies with correlated observations?</td>
<td>177</td>
</tr>
<tr>
<td>13</td>
<td>Validation of models</td>
<td>179</td>
</tr>
<tr>
<td>13.1</td>
<td>How can I validate my models?</td>
<td>179</td>
</tr>
<tr>
<td>14</td>
<td>Special topics</td>
<td>184</td>
</tr>
<tr>
<td>14.1</td>
<td>What if the independent variable changes value during the course of the study?</td>
<td>184</td>
</tr>
<tr>
<td>14.2</td>
<td>What are the advantages and disadvantages of time-dependent covariates?</td>
<td>185</td>
</tr>
<tr>
<td>14.3</td>
<td>What are classification and regression trees (CART) and should I use them?</td>
<td>187</td>
</tr>
<tr>
<td>14.4</td>
<td>How can I get best use of my biostatistician?</td>
<td>190</td>
</tr>
<tr>
<td>14.5</td>
<td>How do I choose which software package to use?</td>
<td>190</td>
</tr>
<tr>
<td>15</td>
<td>Publishing your study</td>
<td>192</td>
</tr>
<tr>
<td>15.1</td>
<td>How much information about how I constructed my multivariable models should I put in the Methods section?</td>
<td>192</td>
</tr>
<tr>
<td>15.2</td>
<td>Do I need to cite a statistical reference for my choice of multivariable model?</td>
<td>194</td>
</tr>
<tr>
<td>15.3 Which parts of my multivariable analysis should I report in the Results section?</td>
<td>194</td>
<td></td>
</tr>
<tr>
<td>16 Summary: Steps for constructing a multivariable model</td>
<td>197</td>
<td></td>
</tr>
<tr>
<td>Index</td>
<td>199</td>
<td></td>
</tr>
</tbody>
</table>
Preface

I’ve been very gratified by the success of the first edition of this book. Although the positive reviews from biostatisticians have meant a lot to me, the real payoff has been the response from novice clinical investigators. Comments such as “easy to read,” “easy to understand,” and “helpful and useful” have greatly warmed my heart. In one case, the book even led me to collaborate with a reader (entirely by email) on a project of his. This is exactly why I wrote the book: to promote the work of clinical researchers early in their careers.

Writing a second edition has enabled me to make some important additions to the book. Since the time I wrote the first edition, there has been a major increase in the use of generalized estimating equations and mixed-effects models to analyze correlated (clustered) observations. Such data arise from longitudinal studies that evaluate subjects repeatedly for a particular outcome. Clustered data also arise from other types of studies where patients are randomized or sampled from established groups such as physician practices or hospital. In addition to generalized estimating equations and mixed-effects models, I also explain how to use repeated measures analysis of variance, conditional logistic regression, and extensions of the Cox proportional hazard model to analyze clustered data (Chapter 12).

Another recent development in the field of clinical research is the increased use of propensity scores. These scores allow better adjustment for baseline differences between nonrandomized groups than solely adjusting for potential confounders using a multivariable model. I have therefore added a chapter on the use of propensity scores (Chapter 11). Also, the use of splines to incorporate nonlinear relationships between independent variables and outcomes has increased and I now include instructions on how to use them (Section 5.5). Finally, I beefed up the sections on Poisson regression (Section 3.7) and on performing sample size calculations for multivariable models (Section 7.4).

In revising the book, I have followed the suggestions of readers of the first edition. One pointed out that I barely mentioned analysis of variance (ANOVA) and related procedures (e.g., analysis of covariance [ANCOVA], multivariate analysis of variance [MANOVA]), even though these techniques are widely used in the analysis of interval outcomes. I had downplayed analysis of variance in the first edition because multiple linear regression is easier to explain, easier to set up correctly, and easier to interpret than analysis of variance and is more commonly used in the medical literature. Since both analyses give the same result (assuming you construct the models in comparable ways) I had decided to focus on the simpler technique. However, the reader convinced me that this important technique deserved further discussion in this book. Therefore, I have included a section describing analysis of variance and related procedures (Section 3.3), but have done so in a way that readers uninterested in this technique can skip without losing the meaning of the rest of the chapter.

Writing a second edition has given me the privilege of updating my thinking on multivariable analysis. The biggest change from the prior edition is that I have gone from being “agnostic” on the topic of using automatic variable selection algorithms (e.g., forward stepwise selection) to being against using them for explanatory models. Recent discussions with Frank Harrell, Jr. and Leonard Braitman were especially influential in this regard.

While making these additions and changes I have tried to preserve those features that made the first edition a success. Specifically, I have maintained the question-and-answer format because I wanted to keep the focus on the practical aspects of multivariable analysis. I have resisted the suggestions of some to go to a more traditional topical approach (e.g., separate sections on linear regression, logistic regression, proportional hazards analysis) because beginning researchers may not know which procedure would be best to use. Only by constantly comparing and contrasting the different procedures can you appreciate the differences – some subtle, some substantial – between the different methods.

This book assumes that you are familiar with basic biostatistics. If not, I recommend: S. Glantz’s Primer of Biostatistics (5th edn, McGraw-Hill, 2002). It was my first biostatistics book (then in its first edition!). I have also written a basic statistics book using a question-and-answer approach similar to that used in this book: Study Design and Statistical Analysis: A Practical Guide for Clinicians, Cambridge University Press, forthcoming. I think of it as a “prequel” to this book (in the sense that The Phantom Menace is a prequel in the Star Wars movie series: released later but covering earlier material). As with this text, I focus on conceptual explanations of statistics and minimize the use of mathematics or derivations of formulas.
As was true of the first edition, I owe a great deal to the writers of several biostatistics articles and books. I cite their works throughout the text and recommend them enthusiastically. My greatest debts are to my teachers, students, and colleagues. Several years of students in the University of California, San Francisco, Clinical Research Program have contributed to this book through their insightful questions and observations. Susan Buchbinder, Rani Marx and Eric Vittingoff recommended a number of important changes to the first edition. I am also especially thankful to Joan Hilton who reviewed the new section on correlated observations in this edition. If any errors crept in despite her review, I am only to blame.

I greatly appreciate the support of my editor Peter Silver and the staff at Cambridge University Press for encouraging me to do this second edition.

If you have questions or suggestions for future editions, email me at mhkatz59@yahoo.com.