Multivariable Analysis

A Practical Guide for Clinicians

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

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 2RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521840514

First published 1999
Second edition published 2006

Printed in the United Kingdom at the University Press, Cambridge

A catalog record for this book is available from the British Library

Library of Congress Cataloging in Publication data

ISBN-13 978-0-521-84051-4 hardback
ISBN-10 0-521-84051-1 hardback

ISBN-13 978-0-521-54985-1 paperback
ISBN-10 0-521-54985-X paperback

To my parents, for their unwavering support

Contents

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.[1] 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).

[1] Apfel, C. C., Krenke, P., Katz, M. H., *et al.* "Volatile anaesthetics may be the main cause for early but not delayed postoperative nausea and vomiting: a randomized controlled trial of factorial design." *Br. J. Anaesth.* **88** (2002): 659–68.

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.

1

Introduction

1.1 Why should I do multivariable analysis?

DEFINITION

Multivariable analysis is a tool for determining the relative contributions of different causes to a single event.

We live in a multivariable world. Most events, whether medical, political, social, or personal, have multiple causes. And these causes are related to one another. Multivariable analysis[1] is a statistical tool for determining the relative contributions of different causes to a single event or outcome.

Clinical researchers, in particular, need multivariable analysis because most diseases have multiple causes, and prognosis is usually determined by a large number of factors. Even for those infectious diseases that are known to be caused by a single pathogen, a number of factors affect whether an exposed individual becomes ill, including the characteristics of the pathogen (e.g., virulence of strain), the route of exposure (e.g., respiratory route), the intensity of exposure (e.g., size of inoculum), and the host response (e.g., immunologic defense).

Multivariable analysis allows us to sort out the multifaceted nature of risk factors and their relative contribution to outcome. For example, observational epidemiology has taught us that there are a number of risk factors associated with premature mortality, notably smoking, a sedentary lifestyle, obesity, elevated cholesterol, and hypertension. Note that I did not say that these factors *cause* premature mortality. Statistics alone cannot prove that a relationship between a risk factor and an outcome are causal.[2] Causality is established on the basis of biological plausibility and rigorous study designs, such as randomized controlled trials, which eliminate sources of potential bias.

[1] The terms "multivariate analysis" and "multivariable analysis" are often used interchangeably. In the strict sense, multivariate analysis refers to simultaneously predicting multiple outcomes. Since this book deals with techniques that use multiple variables to predict a single outcome, I prefer the more general term multivariable analysis.

[2] Throughout the text I use the terms "associated with" and "related to" interchangeably. Similarly, I use the terms "risk factor," "exposure," and "independent variable," and the terms "outcome" and "dependent variable," interchangeably. Although many use the term "predicts" to refer to the association between an independent variable and an outcome and the term "predictor" to refer to an independent variable, these terms imply causality and I prefer to reserve them for when we are determining how well a model predicts the outcome of individual subjects (Section 9.2C).

Identification of risk factors of premature mortality through observational studies has been particularly important because you cannot randomize people to many of the conditions that cause premature mortality, such as smoking, sedentary lifestyle, or obesity. And yet these conditions tend to occur together; that is, people who smoke tend to exercise less and be more likely to be obese.

How does multivariable analysis separate the *independent* contribution of each of these factors? Let's consider the case of exercise. Numerous studies have shown that persons who exercise live longer than persons with sedentary lifestyles. But if the only reason that persons who exercise live longer is that they are less likely to smoke and more likely to eat low-fat meals leading to lower cholesterol, then initiating an exercise routine would not change a person's life expectancy.

The Aerobics Center Longitudinal Study tackled this important question.[3] They evaluated the relationship between exercise and mortality in 25 341 men and 7080 women. All participants had a baseline examination between 1970 and 1989. The examination included a physical examination, laboratory tests, and a treadmill evaluation to assess physical fitness. Participants were followed for an average of 8.4 years for the men and 7.5 years for the women.

Table 1.1 compares the characteristics of survivors to persons who had died during the follow-up. You can see that there are a number of significant differences between survivors and decedents among men and women. Specifically, survivors were younger, had lower blood pressure, lower cholesterol, were less likely to smoke, and were more physically fit (based on the length of time they stayed on the treadmill and their level of effort).

Although the results are interesting, Table 1.1 does not answer our basic question: Does being physically fit independently increase longevity? It doesn't answer the question because whereas the high-fitness group was less likely to die during the study period, those who were physically fit may just have been younger, been less likely to smoke, or had lower blood pressure.

To determine whether exercise is independently associated with mortality, the authors performed proportional hazards analysis, a type of multivariable analysis. The results are shown in Table 1.2. If you compare the number of deaths per thousand person-years in men, you can see that there were more deaths in the low-fitness group (38.1) than in the moderate/high fitness group (25.0). This difference is reflected in the elevated relative risk for lower fitness ($38.1/25.0 = 1.52$). These results are adjusted for all of the other variables listed in the table. This means that low fitness is associated with higher mortality, independent of the effects of other known risk factors for mortality, such as smoking, elevated

[3] Blair, S. N., Kampert, J. B., Kohl, H. W., *et al.* "Influences of cardiorespiratory fitness and other precursors on cardiovascular disease and all-cause mortality in men and women." *JAMA* **276** (1996): 205–10.

Table 1.1 Baseline characteristics of survivors and decedents, Aerobics Center Longitudinal Study.

	Men		Women	
Characteristics	Survivors ($n = 24\,740$)	Decedents ($n = 601$)	Survivors ($n = 6991$)	Decedents ($n = 89$)
Age, y (SD)	42.7 (9.7)	52.1 (11.4)	42.6 (10.9)	53.3 (11.2)
Body mass index, kg/m^2 (SD)	26.0 (3.6)	26.3 (3.5)	22.6 (3.9)	23.7 (4.5)
Systolic blood pressure, mm Hg (SD)	121.1 (13.5)	130.4 (19.1)	112.6 (14.8)	122.6 (17.3)
Total cholesterol, mg/dL (SD)	213.1 (40.6)	228.9 (45.4)	202.7 (40.5)	228.2 (40.8)
Fasting glucose, mg/dL (SD)	100.4 (16.3)	108.1 (32.0)	94.4 (14.5)	99.9 (25.0)
Fitness, %				
Low	20.1	41.6	18.8	44.9
Moderate	42.0	39.1	40.6	33.7
High	37.9	19.3	40.6	21.3
Current or recent smoker, %	26.3	36.9	18.5	30.3
Family history of coronary heart disease, %	25.4	33.8	25.2	27.0
Abnormal electrocardiogram, %	6.9	26.3	4.8	18.0
Chronic illness, %	18.4	40.3	13.4	20.2

Adapted with permission from Blair, S. N., *et al.* "Influences of cardiorespiratory fitness and other precursors on cardiovascular disease and all-cause mortality in men and women." *JAMA* **276** (1996): 205–10. Copyright 1996, American Medical Association. Additional data provided by authors.

blood pressure, cholesterol, and family history. A similar pattern is seen for women.

DEFINITION

Stratified analysis assesses the effect of a risk factor on outcome while holding another variable constant.

Was there any way to answer this question without multivariable analysis? One could have performed stratified analysis. Stratified analysis assesses the effect of a risk factor on outcome while holding another variable constant. So, for example, we could compare physically fit to unfit persons separately among smokers and nonsmokers. This would allow us to calculate a relative risk for the impact of fitness on mortality, independent of smoking. This analysis is shown in Table 1.3.

Unlike the multivariable analysis in Table 1.2, the analyses in Table 1.3 are bivariate.[4] We see that the mortality rate is greater among those at low fitness compared to those at moderate/high fitness, both among smokers (48.0 vs. 29.4) and among nonsmokers (44.0 vs. 20.1). This stratified analysis shows that the effect of fitness is independent of smoking status.

[4] Some researchers use the term "univariate" to describe the association between two variables. I think it is more informative to restrict the term univariate to analyses of a single variable (e.g., mean, median), while using the term "bivariate" to refer to the association between two variables.

Table 1.2 Multivariable analysis of risk factors for all-cause mortality, Aerobics Center Longitudinal Study.

	Men		Women	
Independent variable	Deaths per 10 000 person-years	Adjusted relative risk (95% CI)	Deaths per 10 000 person-years	Adjusted relative risk (95% CI)
Fitness				
Low	38.1	1.52 (1.28–1.82)	27.8	2.10 (1.36–3.26)
Moderate/High	25.0	1.0 (ref.)	13.2	1.0 (ref.)
Smoking status				
Current or recent smoker	39.4	1.65 (1.39–1.97)	27.8	1.99 (1.25–3.17)
Past or never smoked	23.9	1.0 (ref.)	14.0	1.0 (ref.)
Systolic blood pressure				
≥140 mm Hg	35.6	1.30 (1.08–1.58)	13.0	0.76 (0.41–1.40)
<140 mm Hg	27.3	1.0 (ref.)	17.1	1.0 (ref.)
Cholesterol				
≥240 mg/dL	35.1	1.34 (1.13–1.59)	18.0	1.09 (0.68–1.74)
<240 mg/dL	26.1	1.0 (ref.)	16.6	1.0 (ref.)
Family history of coronary heart disease				
Yes	29.9	1.07 (0.90–1.29)	12.8	0.70 (0.43–1.16)
No	27.8	1.0 (ref.)	18.2	1.0 (ref.)
Body mass index				
≥27 kg/m^2	28.8	1.02 (0.86–1.22)	15.9	0.94 (0.52–1.69)
<27 kg/m^2	28.2	1.0 (ref.)	16.9	1.0 (ref.)
Fasting glucose				
≥120 mg/dL	34.4	1.24 (0.98–1.56)	29.6	1.79 (0.80–4.00)
<120 mg/dL	27.9	1.0 (ref.)	16.5	1.0 (ref.)
Abnormal electrocardiogram				
Yes	44.4	1.64 (1.34–2.01)	25.3	1.55 (0.87–2.77)
No	27.1	1.0 (ref.)	16.3	1.0 (ref.)
Chronic illness				
Yes	41.2	1.63 (1.37–1.95)	17.5	1.05 (0.61–1.82)
No	25.3	1.0 (ref.)	16.7	1.0 (ref.)

Adapted with permission from Blair, S. N., *et al.* "Influences of cardiorespiratory fitness and other precursors on cardiovascular disease and all-cause mortality in men and women." *JAMA* **276** (1996): 205–10. Copyright 1996, American Medical Association. Additional data provided by authors.

But what about all of the other variables that might affect the relationship between fitness and longevity? You could certainly stratify for each one individually, proving that the effect of fitness on longevity is independent not only of smoking status, but also independent of elevated cholesterol, elevated blood

Table 1.3 Stratified analysis of smoking and fitness on all-cause mortality among men, Aerobics Center Longitudinal Study.

	Deaths per 10 000 person-years	Stratum-specific relative risk (95% CI)
Smokers		
Low fitness	48.0	1.63 (1.26–2.13)
Moderate/high fitness	29.4	1.0 (ref.)
Nonsmokers		
Low fitness	44.0	2.19 (1.77–2.70)
Moderate/high fitness	20.1	1.0 (ref.)

Data supplied by Aerobics Center Longitudinal Study.

pressure, and so on. However, this would only prove that the relationship is independent of these variables taken singly.

To stratify by two variables (smoking and cholesterol), you would have to assess the relationship between fitness and mortality in four groups (smokers with high cholesterol; smokers with low cholesterol; nonsmokers with high cholesterol; nonsmokers with low cholesterol). To stratify by three variables (smoking status, cholesterol level, and elevated blood pressure [yes/no]), you would have to assess the relationship between fitness and mortality in eight groups ($2 \times 2 \times 2 = 8$); add elevated glucose (yes/no) and you would have 16 groups ($2 \times 2 \times 2 \times 2 = 16$); add age (in six decades) and you would have 96 groups ($2 \times 2 \times 2 \times 2 \times 6 = 96$); and we haven't even yet taken into account all of the variables in Table 1.1 that are associated with mortality.

With each stratification variable you add, you increase the number of subgroups for which you have to individually assess whether the relationship between fitness and mortality holds. Besides producing mountains of printouts, and requiring a book (rather than a journal article) to report your results, you would likely have an insufficient sample size in some of these subgroups, even if you started with a large sample size. For example, in the Aerobics Center Longitudinal Study there were 25 341 men but only 601 deaths. With 96 subgroups, assuming uniform distributions, you would expect only about six deaths per subgroup. But, in reality, you wouldn't have uniform distributions. Some samples would be very small, and some would have no outcomes at all.

Multivariable analysis overcomes this limitation. It allows you to simultaneously assess the impact of multiple independent variables on outcome. But there is (always) a cost: The model makes certain assumptions about the nature of the

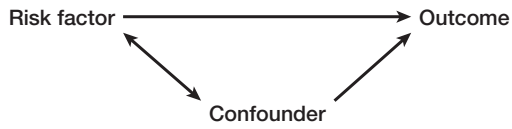


Figure 1.1 Relationships among risk factor, confounder, and outcome.

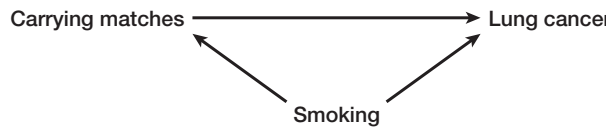


Figure 1.2 Relationships among carrying matches, smoking, and lung cancer.

data. These assumptions are sometimes hard to verify. We will take up these issues in Chapters 5, 6, and 10.

1.2 What are confounders and how does multivariable analysis help me to deal with them?

The ability of multivariable analysis to *simultaneously* assess the independent contribution of a number of risk factors to outcome is particularly important when you have "confounding." Confounding occurs when the apparent association between a risk factor and an outcome is affected by the relationship of a third variable to the risk factor and the outcome; the third variable is called a confounder.

DEFINITION

A *confounder* is associated with the risk factor and causally related to the outcome.

For a variable to be a confounder, the variable must be associated with the risk factor and causally related to the outcome (Figure 1.1).

A classically taught example of confounding is the relationship between carrying matches and developing lung cancer (Figure 1.2). Persons who carry matches have a greater chance of developing lung cancer; the confounder is smoking. This example is often used to illustrate confounding because it is easy to grasp that carrying matches cannot possibly cause lung cancer.

Stratified analysis can be used to assess and eliminate confounding. If you stratify by smoking status you will find that carrying matches is not associated with lung cancer. That is, there will be no relationship between carrying matches and lung cancer when you look separately among smokers and nonsmokers. The statistical evidence of confounding is the difference between the unstratified and the stratified analysis. In the unstratified analysis the chi-square test would be significant and the odds ratio for the impact of matches on lung cancer would be significantly greater than one. In the two stratified analyses (smokers and nonsmokers), carrying matches would not be significantly associated with lung

Table 1.4 Bivariate association between smoking status and risk of death.

Bivariate	Nonsmokers	Former smokers	Recent quitters	Persistent smokers
Relative risk of death	1.0 (ref.)	1.08 (0.92–1.26)	0.56 (0.40–0.77)	0.74 (0.59–0.94)

Adapted from Hasdai, D., *et al.* "Effect of smoking status on the long-term outcome after successful percutaneous coronary revascularization." *N. Engl. J. Med.* **336** (1997): 755–61.

cancer; the odds ratio would be one in both strata. This differs from the example of stratified analysis in Table 1.3 where exercise was significantly associated with mortality for both smokers and nonsmokers.

Most clinical examples of confounding are more subtle and harder to diagnose than the case of matches and lung cancer. Let's look at the relationship between smoking and prognosis in patients with coronary artery disease following angioplasty (the opening of clogged coronary vessels with the use of a wire and a balloon).

Everyone knows (although the cigarette companies long claimed ignorance) that smoking increases the risk of death. Countless studies, including the Aerobics Center Longitudinal Study (Table 1.2), have demonstrated that smoking is associated with increased mortality. How then can we explain the results of Hasdai and colleagues?[5] They followed 5437 patients with coronary artery disease, who had angioplasty. They divided their sample into nonsmokers, former smokers (quit at least six months before procedure), recent quitters (quit immediately following the procedure), and persistent smokers. The relative risk of death with the 95 percent confidence intervals are shown in Table 1.4.

How can the risk of death be lower among persons who persistently smoke than those who never smoked? In the case of recent quitters, you would expect their risk of death to return toward normal only after years of not smoking – and even then you wouldn't actually expect quitters to have a lower risk of death than nonsmokers.

Before you assume that there is something wrong with this study, several other studies have found a similar relationship between smoking and better prognosis among patients with coronary artery disease after thrombolytic therapy. This effect has been named the "smoker's paradox."[6] What is behind the

[5] Hasdai, D., Garratt, K. N., Grill, D. E., *et al.* "Effect of smoking status on the long-term outcome after successful percutaneous coronary revascularization." *N. Engl. J. Med.* **336** (1997): 755–61.

[6] Barbash, G. I., Reiner, J., White, H. D., *et al.* "Evaluation of paradoxical beneficial effects of smoking in patients receiving thrombolytic therapy for acute myocardial infarction: Mechanisms of the 'smoker's paradox' from the GUSTO-I trial, with angiographic insights." *J. Am. Coll. Cardiol.* **26** (1995): 1222–9.

Table 1.5 Association between demographic and clinical factors and smoking status.

	Nonsmokers	Former smokers	Recent quitters	Persistent smokers
Age, year ± SD	67 ± 11	65 ± 10	56 ± 10	55 ± 11
Duration of angina, months ± SD	41 ± 66	51 ± 72	21 ± 46	29 ± 55
Diabetes, %	21%	18%	8%	10%
Hypertension, %	54%	48%	38%	39%
Extent of coronary artery disease, %				
One vessel	50%	51%	57%	55%
Two vessels	36%	36%	34%	36%
Three vessels	14%	13%	10%	9%

Adapted from Hasdai, D., *et al.* "Effect of smoking status on the long-term outcome after successful percutaneous coronary revascularization." *N. Engl. J. Med.* **336** (1997): 755–61.

paradox? Look at Table 1.5. As you can see, compared to nonsmokers and former smokers, quitters and persistent smokers are younger, have had angina for a shorter period of time, are less likely to have diabetes and hypertension, and have less severe coronary artery disease (i.e., more one-vessel disease and less three-vessel disease). Given this, it is not so surprising that the recent quitters and persistent smokers have a lower risk of death than nonsmokers and former smokers: They are younger and have fewer underlying medical problems than the nonsmokers and former smokers.

Compare the bivariate (unadjusted) risk of death to the multivariable risk of death (Table 1.6). Note that in the multivariable analysis the researchers adjusted for those differences, such as age and duration of angina, that existed among the four groups.

With statistical adjustment for the baseline differences between the groups, the former smokers and persistent smokers have a significantly greater risk of death than nonsmokers – a much more sensible result. (The recent quitters also have a greater risk of death than the nonsmokers, but the confidence intervals of the relative risk do not exclude one.) The difference between the bivariate and multivariable analysis indicates that confounding is present. The advantage of multivariable analysis over stratified analysis is that it would have been difficult to stratify for age, duration of angina, diabetes, hypertension, and extent of coronary artery disease.

TIP

Multivariable analysis is preferable to stratified analysis when you have multiple confounders.

Table 1.6 Comparison of bivariate and multivariable association between smoking status and risk of death.

	Nonsmokers	Former smokers	Recent quitters	Persistent smokers
Relative risk of death (bivariate)	1.0 (ref.)	1.08 (0.92–1.26)	0.56 (0.40–0.77)	0.74 (0.59–0.94)
Relative risk of death (multivariable)	1.0 (ref.)	1.34 (1.14–1.57)	1.21 (0.87–1.70)	1.76 (1.37–2.26)

Adapted from Hasdai, D., *et al.* "Effect of smoking status on the long-term outcome after successful percutaneous coronary revascularization." *N. Engl. J. Med.* **336** (1997): 755–61.

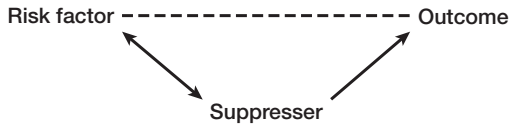


Figure 1.3 Relationships among risk factor, suppresser, and outcome.

1.3 What are suppressers and how does multivariable analysis help me to deal with them?

TIP

Unlike a typical confounder, when you have a suppresser you won't see any bivariate association between the risk factor and the outcome until you adjust for the suppresser.

Suppresser variables are a type of confounder. As with confounders, a suppresser is associated with the risk factor and the outcome (Figure 1.3). The difference is that on bivariate analysis there is no effect seen between the risk factor and the outcome. But when you adjust for the suppresser, the relationship between the risk factor and the outcome becomes significant.

Identifying and adjusting for suppressers can lead to important findings. For example, it was unknown whether taking antiretroviral treatment would prevent HIV seroconversion among healthcare workers who sustained a needle stick from a patient who was HIV-infected. For several years, healthcare workers who had an exposure were offered zidovudine treatment, but they were told that there was no efficacy data to support its use. A randomized controlled trial was attempted, but it was disbanded because healthcare workers did not wish to be randomized.

Since a randomized controlled trial was not possible, a case-control study was performed instead.[7] The cases were healthcare workers who sustained a needle

[7] Cardo, D. M., Culver, D. H., Ciesielski, C. A., *et al.* "A case-control study of HIV seroconversion in health-care workers after percutaneous exposure." *N. Engl. J. Med.* **337** (1997): 1485–90.

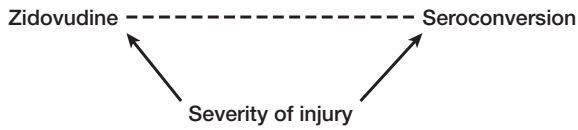


Figure 1.4 Relationships among zidovudine, severity of injury, and seroconversion.

stick and had seroconverted. The controls were healthcare workers who sustained a needle stick but had remained HIV-negative. The question was whether the proportion of persons taking zidovudine would be lower in the group who had seroconverted (the cases) than in the group who had not become infected (the controls). The investigators found that the proportion of cases using zidovudine was lower (9 of 33 cases or 27 percent) than the proportion of controls using zidovudine (247 of 679 controls or 36 percent), but the difference was not statistically significant (probability $[P] = 0.35$). Consistent with this nonsignificant trend, the odds ratio shows that zidovudine was protective (0.7), but the 95 percent confidence intervals were wide and did not exclude one (0.3–1.4).

However, it was known that healthcare workers who sustained an especially serious exposure (e.g., a deep injury or who stuck themselves with a needle that had visible blood on it) were more likely to choose to take zidovudine than healthcare workers who had more minor exposures. Also, healthcare workers who had serious exposures were more likely to seroconvert.

When the researchers adjusted their analysis for severity of injury using multiple logistic regression, zidovudine use was associated with a significantly lower risk of seroconversion (odds ratio [OR] = 0.2; 95 percent confidence interval (CI) = 0.1 − 0.6; $P < 0.01$). Thus, we have an example of a suppresser effect as shown in Figure 1.4. Severity of exposure is associated with zidovudine use and causally related to seroconversion. Zidovudine use is not associated with seroconversion in bivariate analyses but becomes significant when you adjust for severity of injury.

Although this multivariable analysis demonstrated the efficacy of zidovudine on preventing seroconversion by incorporating the suppresser variables, it should be remembered that multivariable analysis cannot adjust for other potential biases in the analysis. For example, the cases and controls for this study were not chosen from the same population, raising the possibility that selection bias may have influenced the results. Nonetheless, on the strength of this study, postexposure prophylaxis with antiretroviral treatment became the standard of care for healthcare workers who sustained needle sticks from HIV-contaminated needles.

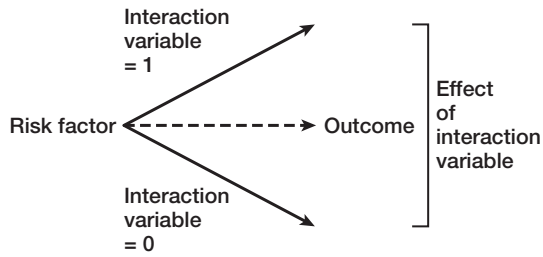


Figure 1.5 Illustration of an interaction effect.

1.4 What are interactions and how does multivariable analysis help me to deal with them?

DEFINITION

An *interaction* occurs when the impact of a risk factor on outcome is changed by the value of a third variable.

An interaction occurs when the impact of a risk factor on outcome is changed by the value of a third variable. Interaction is sometimes referred to as effect modification, since the effect of the risk factor on outcome is modified by another variable.

An interaction is illustrated in Figure 1.5. The risk factor's effect on outcome (solid lines) differs depending on the value of the interaction variable (whether it is 1 or 0). The dotted line indicates the relationship without consideration of the interaction effect.

In extreme cases, an interaction may completely reverse the relationship between the risk factor and the outcome. This would occur when the risk factor increased the likelihood of outcome at one value of the interaction variable but decreased the likelihood of outcome at a different value of the interaction variable. More commonly, the effect of the risk factor on the outcome is stronger (or weaker) at certain values of the third variable.

As with confounding, stratification can be used to identify an interaction. By stratifying by the interaction variable, you can observe the effect of a risk factor on outcome at the different values of the interaction variable. You can statistically test whether the association between a risk factor and an outcome at different levels of the interaction variable are statistically different from one another using a chi-square test for homogeneity.

However, as with the use of stratification to eliminate confounding, use of stratification to demonstrate interaction has limitations. It is cumbersome to stratify by more than one or two variables; yet you may have multiple interactions in your data. Whereas stratification will accurately quantify the effect of the risk factor on the outcome at different levels of the interaction variable, this analysis will not be adjusted for the other variables in your model (e.g., confounders) that may affect the relationship between risk factor and outcome. Multivariable

Table 1.7 Association of independent variables with confirmed diagnosis of acute myocardial infarction based on multiple logistic regression model.

Independent variables	Coefficients	Odds ratio
Male gender	**0.4852**	**1.6**
Age <50	0.1432	1.2
Chest pain	0.8792	2.4
Chief complaint: chest pain	0.4399	1.6
Nausea/vomiting	0.5153	1.7
Congestive heart failure	0.6759	2.0
White race	0.0987	1.1
ST elevation	**2.0948**	**8.1**
ST depression	1.2632	3.5
Q waves	0.5311	1.7
History of diabetes mellitus	0.2781	1.3
History of hypertension	0.2032	1.2
History of angina	−0.2976	0.7
History of peptic ulcers	−0.3210	0.7
Dizziness	−0.4437	0.6
Interactions		
Male gender and congestive heart failure	−0.6899	0.5
Male gender and ST elevation	**−0.5187**	**0.6**
Male gender and white race	0.5206	1.7

Adapted with permission from Zucker, D. R., *et al.* "Presentation of acute myocardial infarction in men and women." *J. Gen. Intern. Med.* **12** (1997): 79–87.

analysis allows you to include interaction terms and assess them while adjusting for other variables.

For example, Zucker and colleagues evaluated whether specific signs or symptoms of myocardial infarction were different in men than in women presenting to the emergency department with chest pain or other symptoms of acute cardiac ischemia.[8]

In Table 1.7 you can see the association between the independent variables and confirmed diagnoses of acute myocardial infarction. The coefficients and odds ratios are from a multiple logistic regression model. The authors found three significant interactions involving gender: male gender and ST elevation (on electrocardiogram), male gender and congestive heart failure, and male gender and white race.

[8] Zucker, D. R., Griffith, J. L., Beshansky, J. R., *et al.* "Presentations of acute myocardial infarction in men and women." *J. Gen. Intern. Med.* **12** (1997): 79–87.

What do these interactions mean? Let's use the interaction involving male gender and ST elevations as an example (I have put these two variables and their interaction term in bold print). Note that men were more likely than women to have an acute myocardial infarction (OR = 1.6), even after adjusting for other variables associated with an infarction. Similarly, ST elevations were more likely to indicate ischemia (OR = 8.1). Given this, you would expect that males with ST elevations would have a markedly higher risk of myocardial infarction ($1.6 \times 8.1 = 13.0$) than women ($1.0 \times 8.1 = 8.1$) (the wonderful property of odds ratios that allows you to multiply them this way is explained in Section 10.8).

The multiplication of the odds ratios of gender and ST elevations would lead you to believe that men with ST elevations would have a significantly higher risk of an acute myocardial infarction than women (13.0 vs. 8.1). In fact, the risk for men and women with ST elevations was similar. This is reflected in the negative coefficient for male gender × ST elevations and the odds ratio of 0.6. If you multiply out the odds ratio for the interaction of male gender with ST elevations, men with ST elevations ($1.6 \times 8.1 \times 0.6 = 7.8$) and women with ST elevations ($1.0 \times 8.1 \times 1.0 = 8.1$) have a similar risk of myocardial infarction.

ST elevations are highly specific for (although not diagnostic of) myocardial infarction. It is not surprising, therefore, that the risks of myocardial infarction are similar in men and women with ST elevations. Had being male made it even worse to have ST elevations the coefficient would have been positive, the odds ratio would have been greater than one, and we would have seen an even greater difference between the risk of an acute myocardial infarction for men and for women in the presence of ST elevations than the difference between 13.0 and 8.1.

Because interaction effects can be difficult to assess and interpret, I will return to this topic in Sections 8.3, 9.9, and 10.8.

2

Common uses of multivariable models

2.1 What are the most common uses of multivariable models in clinical research?

Multivariable models have a variety of uses in clinical research, in both nonrandomized and randomized studies. The four most common uses of multivariable models are to:

A identify risk factors while adjusting for potential confounders
B adjust for differences in baseline characteristics
C determine diagnosis (diagnostic models)
D determine prognosis (prognostic models)

Models for identifying risk factors and adjusting for differences in baseline characteristics (A and B) can be thought of as explanatory or etiologic models (i.e., you are trying to explain or understand an outcome). Models for determining diagnosis and prognosis can be thought of as predictive models (i.e., you are trying to predict outcomes for patients with particular characteristics).

2.1.A Identify risk factors while adjusting for potential confounders

As we learn more about certain multifactorial diseases, such as cardiac disease, we identify a larger and larger number of risk factors for the disease. Because many of these variables are associated with one another, stratification becomes an unwieldy technique for eliminating confounding. For example, when Gardner and colleagues assessed whether the size of low-density lipoprotein particles affected the incidence of coronary artery disease they adjusted their analysis for those risk factors long known to increase the risk of coronary artery disease, such as smoking, blood pressure, and body mass index.[1] But they also adjusted their model for more recently identified risk factors such as HDL cholesterol, nonHDL cholesterol, and triglycerides. One of the most extensive studies of

[1] Gardner, C. D., Fortmann, S. P., and Krauss, R. M. "Association of small low-density lipoprotein particles with the incidence of coronary artery disease in men and women." *JAMA* **276**(1996): 875–81.

cardiovascular disease, the Framingham study, which began in 1948, did not even collect data on HDL cholesterol until 1972.[2]

What if you can prove that the potential confounders are not confounders? What if you test all potential confounders and find that none is related to both your risk factor and outcome? Do you still need to use multivariable analysis to adjust for these factors or could you just report the bivariate association between the risk factor and the outcome? Conceptually, if no variables are associated with both risk factor and outcome, you would not need to use multivariate analysis and you could simply report the bivariate result. However, in practice, most clinical researchers use multivariable analysis when there are other factors associated with outcome, even if these variables are not actual confounders. The reason is that multivariable analysis has become the standard method for proving that there is no confounding. Thus you will often see instances where the association between the risk factor and the outcome are very similar in the bivariate and multivariable analysis, indicating no confounding. Yet, it is the multivariable result that is cited.

TIP

When conducting logistic or proportional hazards analysis the unadjusted results may not be correct, even if your treatment groups are identical on baseline characteristics.

Another reason many investigators prefer citing the multivariable model even when there is no confounding is that if you are conducting nonlinear analysis (logistic or proportional hazards analysis) and the proportion of subjects in each group who have the risk factor/exposure varies, the unadjusted result may not be correct even if the two groups are identical on baseline characteristics.[3]

TIP

Prior to conducting multivariable analysis, use bivariate analysis to verify that there is sufficient overlap of your potential confounders.

Although multivariable models are excellent tools for adjusting for potential confounders, don't assume that just because you have included a potential confounder in your model you have eliminated any bias caused by this confounder. For multivariable models to adequately adjust for confounding, there must be sufficient overlap of the confounders in the different groups or outcomes. For example, if almost all of the smokers are in one group and almost all of the nonsmokers are in another group, adjusting for smoking will not remove confounding caused by smoking. This is why it is important to use bivariate analysis to verify that there is sufficient overlap of your potential confounders prior to conducting a multivariable analysis.

Even if there is sufficient overlap, no adjustment is perfect. Just as there is error in the measurement of your dependent and independent variables, there is error in your confounders. Once you appreciate that your measurement of confounding variables is imperfect, you realize that including a variable in a model cannot completely eliminate confounding. Moreover, the models

[2] Levy, D., Wilson, P. W. F., Anderson, K. M., *et al.* "Stratifying the patient at risk from coronary disease: New insights from the Framingham heart study." *Am. Heart. J.* **119** (1990): 712–17.

[3] Harrell Jr., Frank E. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.* New York: Springer-Verlag, 2001, p. 4.

themselves contain error. Important variables may be omitted, they may be incorrectly specified (Section 5.4), or interactions between the variables may not be appropriately accounted for (Section 10.8). This warning is not meant to be discouraging; rather, it is stated to promote humility about what you can and cannot do with statistical models.

2.1.B Adjust for differences in baseline characteristics

Adjustment for differences in baseline characteristics is especially important in nonrandomized (observational) trials. With studies of this type the groups being compared are not equivalent, since assignment is not determined by randomization.

TIP

When randomization is impossible, use multivariable analysis to statistically approximate equal comparison groups.

Although randomization is the best method of assuring that comparison groups are equal at the start of a study, it is not always feasible to randomize subjects. In instances when expense, logistical, or ethical difficulties preclude randomizing patients, multivariable analysis can be used to statistically approximate equal comparison groups. Of course, multivariable analysis can never adjust for unknown or unmeasured confounders. Only randomization can create groups that are equal with respect to both measured and unmeasured confounders.[4]

Multivariable analysis was used to adjust for known confounders in an important nonrandomized study of statin therapy in patients admitted to hospital with acute coronary syndromes (i.e., symptoms and signs of ischemia).[5] The investigators compared 5959 patients who were begun on statin treatment during hospitalization, to 9522 patients who were not. They found that patients treated with statins were significantly less likely to die in the hospital (OR = 0.19; 95% CI = 0.16 − 0.23).

Does this bivariate analysis prove that statins are effective in decreasing death owing to acute coronary syndromes? No! Since treatment was not randomized, it is likely that there were important differences between patients who received statin treatment in the hospital and those who did not. Indeed, as you can see from Table 2.1, patients receiving statin therapy were different from those who did not receive it, in terms of their demographics, medical history, presenting characteristics, long-term medications, in-hospital medications, and interventions. Perhaps it is these differences that resulted in patients who were treated with statins being less likely to die.

[4] For more on the uses of multivariable analysis for nonrandomized studies see: Anderson, S., Auquier, A., Hauck, W. W., *et al. Statistical Methods for Comparative Studies.* New York: Wiley, 1980; Rosati, R. A., Lee, K. L., Califf, R. M., *et al.* "Problems and advantages of an observational data base approach to evaluating the effect of therapy on outcome." *Circulation* **65** (suppl. II)(1982): 27–32.

[5] Spencer, F. A., Allegrone, J., Goldberg, R. J., *et al.* "Association of statin therapy with outcomes of acute coronary syndromes: The Grace study." *Ann. Intern. Med.* **140** (2004): 857–66.

Table 2.1 Differences between patients who received statin treatment in the hospital and those who did not.

Characteristic	In-hospital statin use ($n = 5959$)	No statin use ($n = 9522$)	*P* Value
Demographic			
Median age, y	62.7	69.8	<0.001
Women, %	29.6	36.7	<0.001
Medical history, %			
Smoking	63.7	52.6	<0.001
Myocardial infarction	19.6	28.2	<0.001
Transient ischemic attack or stroke	5.6	9.4	<0.001
Diabetes	20.1	23.7	<0.001
Positive angiogram	15.5	20.2	<0.001
Peripheral vascular disease	7.5	9.9	<0.001
Hypertension	52.1	58.7	<0.001
Hyperlipidemia	43.0	25.0	<0.001
Percutaneous coronary intervention	8.0	9.6	<0.001
Coronary artery bypass graft surgery	6.4	9.5	<0.001
Presenting characteristics, %			
Killip class			
I	86.2	76.7	<0.001
II	10.8	16.3	
III	2.5	5.3	
IV	0.5	1.7	
Heart rate ≥100 beats/min	13.8	19.5	<0.001
Systolic blood pressure <90 mm Hg	1.6	3.2	<0.001
ST-segment elevation	49	36.7	<0.001
Other long-term medication, %			
Aspirin	26.9	36.2	<0.001
ACE inhibitors	17	24.9	<0.001
β-blockers	17.5	24.1	<0.001
Other lipid-lowering agents	2.4	2.6	>0.2
Other in-hospital medication, %			
Aspirin	96.6	90.9	<0.001
Ticlopidine or clopidogrel	46.9	25.9	<0.001
Unfractionated heparin	53.5	49.9	<0.001
Enoxaparin	48.3	38.5	<0.001
Other low-molecular-weight heparin	12.9	9.6	<0.001
ACE inhibitor	64.8	56.7	<0.001
β-blockers	85.9	71.4	<0.001

(*cont.*)

Table 2.1 (*cont.*)

Characteristic	In-hospital statin use ($n = 5959$)	No statin use ($n = 9522$)	*P* Value
Other lipid-lowering agents	2.1	4.6	<0.001
Glycoprotein IIb/IIIa inhibitors	27.1	13.2	<0.001
Interventions, %			
Cardiac catheterization	59.9	41.7	<0.001
Percutaneous coronary intervention	40.9	22.8	<0.001
Coronary artery bypass graft surgery	5.4	5.3	>0.2

Data from: Spencer, F. A., Allegrone, J., Goldberg, R. J., *et al.* "Association of statin therapy with outcomes of acute coronary syndromes: The Grace study." *Ann. Intern. Med.* **140** (2004): 857–66.

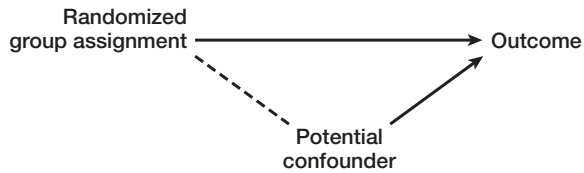


Figure 2.1 **Relationships among randomized group assignment, potential confounder, and outcome.**

To adjust for these baseline differences, the authors performed a multivariable analysis. With adjustment for differences in demographics, medical history, presenting characteristics, long-term medications, in-hospital medications, and interventions, statin use was still significantly associated with a decreased likelihood of death (OR = 0.38; 95 % CI = 0.30 − 0.48). The fact that the adjusted odds ratio is somewhat higher than the bivariate odds ratio indicates that there was confounding. But even after adjustment for multiple confounders, statin use does have an independent effect on reducing mortality.

What about randomized controlled trials? Do you need to use multivariable analysis to adjust for baseline differences? Well, ideally, if the randomization has been conducted in a nonbiased way, your comparison groups will be equal with respect to both known and unknown factors. Therefore, there should be no variables associated with the intervention and thus no confounders (Figure 2.1). One can use a simple bivariate test to compare the outcomes for the two groups and determine if the intervention worked.

That being said, it sometimes happens by chance that, despite randomization, one group is significantly different from the other group. For example, Mittelman and colleagues conducted an intervention to delay nursing-home placement of

patients with Alzheimer's disease.[6] They randomized families to a treatment group (counseling and support for caregivers) and a control group. By chance (and bad luck!), the primary caregiver was significantly more likely to be female among the families randomized to the control group than among the families randomized to the treatment group. Moreover, patients of female caregivers were significantly more likely to be placed in a nursing home (the main outcome of this study). Thus, without adjustment for the gender of the caregiver, the results of the study would have been difficult to interpret. With adjustment for the gender of the caregiver, the authors demonstrated that the treatment was associated with a decrease in nursing-home placement.

TIP

The advantage of randomization is that you can never adjust for what you don't know or can't measure.

Although multivariable analysis made the results of the study by Mittelman and colleagues more interpretable, some researchers are opposed to statistically adjusting randomized clinical trials. Remember that the strength of randomized clinical trials is that randomization produces groups that are equal with respect to both known and unknown factors. This is a tremendous advantage because you can never adjust for what you don't know or can't measure. When you statistically adjust a randomized clinical trial, you are adjusting for the known but not the unknown factors. Whether this partial adjustment makes the groups more comparable or not is a matter of debate. What is clear, however, is that statistical adjustment introduces model error. Thus if your randomized clinical trial shows no or only a borderline effect in the bivariate analysis but a statistically significant effect in the multivariable analysis, some readers will be suspicious that the effect is really because of an error in the model. On the other hand, if your bivariate and multivariable model are consistent, this is stronger evidence of an association than showing only the bivariate or multivariable analysis.

A different and more common use of multivariable analysis in randomized controlled trials is to determine whether other factors besides treatment group are associated with outcome. This use is akin to determining risk factors while adjusting for confounders (Section 2.1.A).

2.1.C Determine diagnosis (diagnostic models)

Multivariable models can identify the best combination of diagnostic information to determine whether a person has a particular disease. The most extensive work on diagnostic algorithms has been done for determining the likelihood of a myocardial infarction in patients presenting to emergency departments with

[6] Mittelman, M. S., Ferris, S. H., Shulman, E., *et al.* "A family intervention to delay nursing home placement of patients with Alzheimer disease." *JAMA* **276** (1996): 1725–31.

chest pain. The reason this question has received so much attention is that the stakes are high. Chest pain is a common presenting symptom in the emergency department. It can be due to something as minor as heartburn or something as serious as a heart attack. Every day, in every emergency department, clinicians decide whom to send home and whom to admit to the coronary care unit. Although coronary care units save lives for patients with acute ischemia, less than half of those receiving this costly intervention actually have ischemia. There is no one test available at the time of an emergency-department visit that distinguishes those patients who should be admitted from those who could be sent home.

TIP

Use multivariable models to determine the best combination of diagnostic information.

Pozen and colleagues developed a diagnostic model for determining the likelihood that a patient presenting to an emergency department with chest pain had acute ischemia.[7] From 59 different clinical features they identified seven clinical features that, when used together in a logistic regression model, produced a prediction of ischemia from 0 to 1.0. To determine the usefulness of the model, the researchers gave the results (during the experimental period) to treating physicians before they had to determine whether to admit patients or send them home. During the control period, the physicians were not given the estimates of the probability of acute ischemia. The researchers found that when the physicians were given the information, their decision-making improved: In particular, the number of coronary care unit admissions decreased by 30 percent without any missed cases of ischemia.

So when confronted with a patient with chest pain, do most emergency department physicians whip out their hand-held calculator and compute the probability of ischemia? Sadly, no. Corey and Merenstein tested the acceptability of this model to physicians by providing a worksheet version of the algorithm in a convenient dispenser in the emergency department, but not requiring the physicians to use them. Physicians used it in only 2.8 percent of the cases.[8] Low use of well-validated, diagnostic rules by physicians in clinical practice has been noted elsewhere.[9]

The reasons that diagnostic rules are not more widely used are complicated. Physicians, especially in emergency departments, are pressed for time. Pozen's algorithm can be computed in less than 20 seconds, but it requires

[7] Pozen, M. W., D'Agostino, R. B., Selker, H. P., *et al.* "A predictive instrument to improve coronary-care-unit admission practices in acute ischemic heart disease: A prospective multicenter clinical trial." *N. Engl. J. Med.* **310** (1984): 1273–8.

[8] Corey, G. A. and Merenstein, J. H. "Applying the acute ischemic heart disease predictive instrument." *J. Fam. Pract.* **25** (1987): 127–33.

[9] Pearson, S. D., Goldman, L., Garcia, T. B., *et al.* "Physician response to a prediction rule for the triage of emergency department patients with chest pain." *J. Gen. Intern. Med.* **9** (1994): 241–7; Wasson, J. H. and Sox, H. C. "Clinical prediction rules: Have they come of age?" *JAMA* **275** (1996): 641–2; Gehlbach, S. H."Commentary." *J. Fam. Pract.* **25** (1987): 132–3.

a preprogrammed hand-held calculator – something most physicians do not carry around with them. Using the worksheet version (if you had one in front of you) it would take you 30–60 seconds to calculate the probability of ischemia. Although this may not seem like a long time, in the emergency department, with many patients in gurneys in front of you, it can seem like an impossibly long task.

Psychological factors also impede the use of diagnostic models by physicians. Medical training has traditionally been akin to apprenticeship. You work with physicians more experienced than yourself until you have enough experience to function on your own. At a certain point, physicians feel that their judgment is accurate (even if studies show that, for some conditions, diagnostic models are more accurate than decisions made by physicians). The physicians in the Corey and Merenstein study complained that they lost confidence in the model when they discovered that two patients with very different characteristics could have the same predicted probability of ischemia. Perhaps, most importantly, as a profession, physicians are not yet comfortable using computer-generated models. But the potential is there. A good diagnostic model can make an intern instantly as good a diagnostician as the head of the department of medicine!

TIP

A good diagnostic algorithm can make an intern as good a diagnostician as the head of the department of medicine!

Before diagnostic models can be used in clinical practice they must be shown to be highly accurate in predicting outcome (Section 13.1). For this reason, developing diagnostic models can be challenging. However, in one respect they are easier to construct than explanatory models. In diagnostic settings, causality is unimportant. For example, diagonal ear lobe creases are associated with coronary events, even with adjustment for known cardiac risk factors including age, left ventricular ejection fraction, cholesterol level, smoking, diabetes, family history, and obesity.[10] No one believes that ear lobe creases cause coronary events. Looked at from a different point of view, lowering a patient's cholesterol level may decrease the risk of a myocardial infarction, but removing an ear lobe crease with plastic surgery would have no effect on the risk of an infarction. The association of ear lobe creases with coronary events is confounded by some yet-to-be determined cardiac risk factor.

But a patient's ear crease still provides useful clinical information. This is especially true if you are a paramedic evaluating patients with chest pain in the field and have little other information about them. Thus in constructing diagnostic algorithms, we are interested in variables that together accurately predict outcome regardless of whether the effect is confounded by some other variable.

[10] Elliott, W. J. and Powell, L. H. "Diagonal earlobe creases and prognosis in patients with suspected coronary artery disease." *Am. J. Med.* **100** (1996): 205–11.

2.1.D Determine prognosis (prognostic models)

How bad is it, Doc? Will the cancer come back? How long do I have to live? These are some of the most difficult questions clinicians face from their patients. Most ethicists and clinicians agree that patients have a right to an honest answer to these questions. While we will never be able to predict the outcome for any one person, multivariable analysis can provide information on the prognosis of a group of patients with a particular set of known prognostic factors.

For example, Schuchter and colleagues developed a prognostic model using logistic regression for estimating 10-year survival in 488 patients with primary melanoma.[11] They prospectively followed patients with primary melanoma. Ten-year survival was 78 percent. Using multiple logistic regression, they identified four factors associated with survival at ten years (yes/no): age, sex, location (extremity versus axis of body), and lesion thickness. At one extreme, women who were 60 years or younger with a lesion <0.76 mm of thickness on their extremity had an estimated 10-year survival of 99 percent. At the other extreme, men who were older than 60 with a lesion >3.6 mm of thickness on their trunk had an estimated 10-year survival of only 10 percent.

This prognostic model illustrates how different survival can be with the same disease but different patient characteristics. Schuchter and colleagues' model correctly predicted outcome in 74 percent of cases. In other words, if you knew age, sex, location, and lesion thickness, you would correctly predict survival or death at 10 years for 74 percent of the sample.

The cynics among you may say: I can do better than that without any prognostic information. If I predict that all patients will be alive at 10 years, I will be correct in 78 percent of the cases (10-year survival = 78 percent). This is true, but you would not have correctly predicted any of the deaths. Methods for judging the success of predictive models are discussed in Section 9.2.C.

TIP

Prognostic models provide valid estimates of risk only for patients with characteristics similar to those in the study population.

Prognostic models provide valid estimates of risk only for patients with similar characteristics to those in the study population. For example, if a prognostic model is based on a sample of males over the age of 50, it will not be helpful in predicting the survival of a 45-year-old woman.

Prognostic models work only when there is a known set of risk factors. They are most useful when they include only those variables readily available to a clinician. If the model requires knowing the genetic markers of the cancer, and testing for those markers is not universally available, the model will be of less help.[12]

[11] Schuchter, L., Schultz, D. J., Synnestvedt, M., *et al.* "A prognostic model for predicting a 10-year survival in patients with primary melanoma." *Ann. Intern. Med.* **125** (1996): 369–75.

[12] For an excellent review of using multivariable models to determine prognosis, see Braitman, L. E. and Davidoff, F. "Predicting clinical states in individual patients." *Ann. Intern. Med.* **125** (1996): 406–12.

Even if all these conditions are met, as is the case with the melanoma study, a model will generally not predict outcome prospectively (with new cases) as well as it does retrospectively (with the cases in the data set from which it was developed). Why? Because the models maximize the correct prediction of the outcome based on the values of the independent and dependent variables in the data set (Section 13.1).

2.2 How do I choose what type of multivariable analysis to use?

There are many types and ways of performing multivariable analysis. The type of multivariable analysis you will use depends primarily on the nature of your outcome variables, your dependent variables, and the hypothesized relationship between your dependent variables and your outcome variable. These issues are discussed in the next three chapters.

3

Outcome variables in multivariable analysis

3.1 How does the nature of my outcome variable influence my choice of which type of multivariable analysis to do?

As shown in Table 3.1, the choice of multivariable analysis depends primarily on the type of outcome variable that you have.

3.2 What type of multivariable analysis should I use with an interval outcome?

DEFINITION

With an *interval* variable each unit (interval) of change on the scale has an equal quantifiable value.

With an interval variable (also called continuous) each unit (interval) of change on the scale has an equal (numerically) quantifiable value. Examples of interval variables are blood pressure, body weight, and temperature. In these examples, a one-unit change at any point on the scale is equal to a millimeter of mercury, a pound (or kilogram), or a degree, respectively.

Interval outcomes can be analyzed with multiple linear regression or with analysis of variance. Either technique will yield the same answer, assuming you set up the models in similar ways. In general, multiple linear regression is more commonly used with observational data, and analysis of variance is more commonly used with experimental designs. Multiple linear regression is more commonly used in the medical literature, and analysis of variance is more commonly used in the behavioral literature.

Because multiple linear regression is easier to explain, easier to set up correctly, and easier to interpret than analysis of variance, I have focused on this technique in this text. Nonetheless, I will devote the next section to explaining the different types of analysis of variance. At a minimum this section will help you to interpret studies in the literature that use these techniques and it should provide enough background to enable you to set up these models correctly. If you are uninterested in analysis of variance you can skip the next section without losing the sense of the chapter.

Table 3.1 Type of outcome variable determines choice of multivariable analysis.

Type of outcome	Example of outcome variable	Type of multivariable analysis*
Interval	Blood pressure, weight, temperature	**Multiple linear regression** Analysis of variance (and related procedures)
Dichotomous	Death, cancer, intensive care unit admission	**Multiple logistic regression**
Time to occurrence of a dichotomous event	Time to death, time to cancer	**Proportional hazards analysis**
Rare outcomes and counts	Time to leukemia, number of infections	Poisson regression

* This text focuses on those procedures that are bolded.

3.3 What are the different types of analysis of variance and when are they used?

The different types of analysis of variance are shown in Table 3.2. These techniques are used to compare two or more groups. The outcome variable must be interval, and normally distributed with equal variance (Section 5.7).

The simplest type of ANOVA is a one-way (or one-factor) design in which two or more groups are compared on a single interval variable.[1] For example, we might compare the blood pressure of persons randomized to receive different treatments (e.g., diuretic, beta-blocker, ACE inhibitor). One-way ANOVA can also be used to compare persons across an observed condition. For example, you can use ANOVA to compare the blood pressure of persons of different ethnicities. If you are comparing only two groups (e.g., men versus women) then an ANOVA is equivalent to a simple *t* test.

Analysis of variance and related procedures can also be used to answer multivariable questions. For example, you can use analysis of variance to determine the effect of (1) group assignment, (2) other categorical variables (e.g., ethnicity, gender) and (3) the interaction between the assigned group and the categorical variable on an interval outcome. For example, you might want to test whether a drug (treatment group) has a different effect on the blood pressure of African-Americans than Caucasians.

If you need to adjust your analysis for interval variables (e.g., age, weight) then you can use an extension of analysis of variance called analysis of covariance (ANCOVA). Using ANCOVA, you can incorporate both interval and categorical

[1] For a review of analysis of variance see: Katz, M. H. *Study Design and Statistical Analysis: A Practical Guide for Clinicians*. Cambridge: Cambridge University Press, forthcoming; for an excellent free statistical book that includes a thorough discussion of analysis of variance, go to: www.statsoft.com/textbook/stanman.html.

Table 3.2 Analysis of variance techniques.

Types	Indication
Analysis of variance (ANOVA)	Compares two or more groups on an interval outcome. Can incorporate categorical independent variables and the interaction of categorical variables with the main effect.
Analysis of covariance (ANCOVA)	Similar to analysis of variance but can incorporate continuous as well as categorical independent variables in the model.
Multivariate analysis of variance (MANOVA)	Similar to analysis of variance but used when there is more than one dependent variable. Use of MANOVA decreases the chance of making a type I error.
Multivariate analysis of covariance (MANCOVA)	Similar to analysis of covariance but used when there is more than one dependent variable. Use of MANCOVA decreases the chance of making a type I error.
Repeated measures analysis of variance/covariance	Similar to analysis of variance/covariance but can incorporate repeated observations of the same subjects (Section 12.1).
Repeated measures multivariate analysis of variance/covariance	Similar to multivariate analysis of variance/covariance but can incorporate repeated observations of the same subjects (Section 12.1).

DEFINITION

Analysis of covariance is an extension of analysis of variance that allows incorporation of interval-independent variables.

independent variables. In addition to the assumptions of normality and homogeneity of variances required by ANOVA, ANCOVA requires that the correlations of the covariates with the dependent variable be similar in the different cells of the design.

Because of the similarities between analysis of variance and analysis of covariance, your software program may automatically choose which of the two to perform, based on whether the covariates you enter into the model include interval variables.

Multivariate analysis of variance (MANOVA) and multivariate analysis of covariance (MANCOVA) are extensions of analysis of variance and analysis of covariance, respectively, that are used when studying more than one dependent variable. The outcome variables are generally correlated. These procedures are used to decrease the chance of making a type I error (falsely rejecting the null hypothesis).

A MANOVA produces a multivariate F (Wilk's lambda). If the multivariate F test were significant, you would then examine the bivariate F test values. Conversely, if the multivariate F test were not significant you would ignore the individual F tests. In addition to the assumptions of normality and homogeneity of variances, multivariate designs require that the intercorrelations of the outcome variables are homogeneous across the cells of the design.[2]

[2] French, A. and Poulsen, J. "Multivariate analysis of variance (MANOVA)." http://online.sfsu.edu/~efc/classes/biol710/manova/manova.htm.

Adaptations of analysis of variance, analysis of covariance, multivariate analysis of variance, and multivariate analysis of covariance, are also available for analyzing repeated observations of the same individuals (Section 12.1).

3.4 What should I do if my outcome variable is ordinal or nominal?

DEFINITION

An *ordinal* variable has multiple categories that can be ordered.

An ordinal variable has multiple categories that can be ordered. An example of an ordinal variable is the New York Heart Association's functional classification of cardiac function. There are four levels of the scale. Although the four levels can be ordered, there is not a numerically quantifiable difference between level I (no limit in physical activity) and level II (slight limitation in physical activity). The difference between level I and II is not equal to the difference between level III (marked limitation of physical activity) and level IV (inability to carry out any physical activity without discomfort).

DEFINITION

A *nominal* variable is a categorical variable with multiple categories that cannot be ordered.

A nominal variable is a categorical variable with multiple categories that cannot be ordered. An example of a nominal variable is cause of death: cancer, heart disease, infection, or other. Unlike an ordinal variable, you cannot numerically order cause of death.

As you can tell by looking at Table 3.1, ordinal or nominal outcomes are not usually used with the three multivariable models discussed in this book. Nonetheless, there are options available for incorporating these variables in a multivariable analysis.

One option for ordinal and nominal variables is to convert them to a dichotomous outcome. For example, New York Heart Association classification is often grouped as level I and II (mild shortness of breath) or level III and IV (severe shortness of breath). Similarly, cause of death can be cancer: yes or no. Obviously, such groupings result in loss of information.

Alternatively, the data can be analyzed using an adaptation of logistic regression. Ordinal outcomes can be analyzed using proportional odds; logistic regression and nominal outcomes can be analyzed using polytomous logistic regression. Because these techniques are not commonly used in medical research, they will not be described here, but readers can obtain more information about these methods from other sources.[3] Another technique available for nominal outcomes is discriminant function analysis, which has both similarities to and differences from the other methods described here.[4]

[3] See Scott, S. C., Goldberg, M. S., and Mayo, N. E. "Statistical assessment of ordinal outcomes in comparative studies." *J. Clin. Epidemiol.* **50** (1997): 45–55. Menard, S. *Applied Logistic Regression Analysis.* Thousand Oaks, CA: Sage Publications, 1995, pp. 80–90.

[4] See Feinstein, A. R. *Multivariable Analysis: An Introduction.* New Haven, CT: Yale University Press, 1996: pp. 431–74.

While I have drawn a distinction between interval and ordinal variables, in practice there is a gray area. Clinical researchers, especially those interested in the behavioral sciences, often use multiple linear regression to analyze outcome variables that do not strictly fit the definition of an interval variable but function like one. Examples include patient satisfaction, patient self-rated health perception, and level of pain or distress. These variables are typically derived by having respondents rate their degree of satisfaction, sense of health, or level of pain on an arbitrary numeric scale of 1 to 4, 1 to 5, or 1 to 100. The scale may include cues such as 1 = excellent, 2 = very good, 3 = fair, and 4 = poor; or 1 = strongly agree, 2 = agree, 3 = no opinion, 4 = disagree, 5 = strongly disagree. These scales are not truly interval because the interval between excellent and very good is not necessarily the same size as the interval between fair and poor. Nonetheless, when these variables are used as dependent variables in multiple linear regression they are likely to work satisfactorily as long as their relationships with the independent variables fulfill the assumptions of multiple linear regression (Sections 5.1–5.4, 5.7).

3.5 What type of multivariable analysis should I use with a dichotomous outcome?

DEFINITION

A *dichotomous* variable has two discrete values.

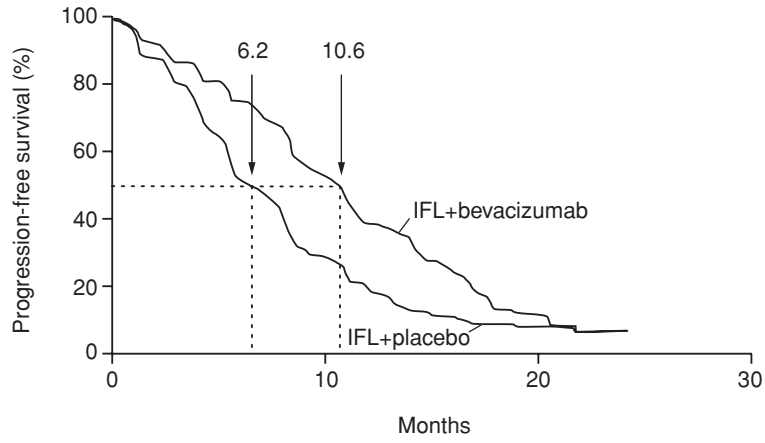
A dichotomous variable (the simplest kind of categorical variable) has two discrete values (categories) at a discrete point in time, for example: alive or dead; development of cancer: yes or no. Dichotomous variables are usually analyzed with logistic regression.

3.6 What type of multivariable analysis should I use with a time-to-outcome variable?

Time to occurrence of a dichotomous outcome refers to events – such as death or development of cancer – that occur over a period of time (e.g., 5 years). They are usually analyzed with proportional hazards analysis.

Sometimes this type of outcome variable is simplified to cumulative outcome at a particular point in time. In other words, instead of having your outcome be "time to myocardial infarction over a five year period," you could change your outcome variable to be "myocardial infarction by five years" (yes/no).

Given that cumulative outcome at a particular point of time is simpler and can be analyzed with logistic regression, an easier to conduct and interpret type of multivariable analysis than proportional hazards analysis, why do so many published clinical trials use time to outcome and proportional hazards analysis? One important reason is that clinical medicine consists more of treatments than



No. at risk						
IFL+bevacizumab	402	269	143	36	6	0
IFL+placebo	411	225	73	17	8	0

Figure 3.1 **Kaplan–Meier estimates of progression-free survival. Although the cancer progresses in almost all patients in both groups by 20 months, the patients who received bevacizumab progressed more slowly. Figure is from Hurwitz, H., Fehrenbacher, L., Novotny, W., *et al*. "Bevacizumab plus irinotecan, fluorouracil, and leucovorin for metastatic colorectal cancer." *N. Engl. J. Med*. **350** (2004): 2235–42. Copyright 2004, Massachusetts Medical Society. All rights reserved.**

TIP

Clinical medicine consists more of treatments than of cures.

of cures. Given this, what matters is how soon the disease occurs or recurs, and how long survival is increased compared to untreated persons.

For example, metastatic colorectal cancer is an extremely deadly disease. Even with chemotherapy the cancer will progress quickly in the majority of patients. Hurwitz and colleagues studied the efficacy of a novel compound, bevacizumab, in patients with metastatic colorectal cancer.[5] Bevacizumab is a monoclonal antibody against vascular endothelial growth factor. The investigators randomized patients to receive either standard chemotherapy with irinotecan, fluorouracil, and leucovorin (IFL) plus placebo, or IFL plus bevacizumab.

As illustrated in Figure 3.1, by 20 months the cancer has progressed in almost all patients in both groups (i.e., the cumulative rate of recurrence at 20 months is the same). Yet, the rate of progression between the groups is significantly different. The median progression-free survival is 6.2 months for those who received IFL + placebo, and 10.6 months for those who received IFL + bevacizumab.

[5] Hurwitz, H., Fehrenbacher, L., Novotny, W., *et al.* "Bevacizumab plus irinotecan, fluorouracil, and leucovorin for metastatic colorectal cancer." *N. Engl. J. Med.* **350** (2004): 2235–42.

One may reasonably ask, especially in these cost-conscious times, how important is it to slow the progression of a disease, if ultimately the same proportion of patients will suffer a recurrence? The answer to this question is more philosophical than statistical. In general, time to outcome matters more for serious outcomes than for minor ones. Patients with life-threatening diseases value additional time, whether of days or months, especially if it will allow them to see a child graduate from college or watch a grandchild take her first steps.

TIP

Time to outcome matters more for serious outcomes than for minor ones.

At the other extreme, for minor outcomes, increased time may not be clinically meaningful. For example, studies have shown that children with chicken pox treated with acyclovir have one day less fever, and experience a decrease in the number of chicken pox lesions one day sooner than those given placebo.[6] If the investigators had used symptoms at seven days as the outcome of their study they would have found no effect for acyclovir because, treated or untreated, immunocompetent children with chicken pox are almost invariably well at seven days. Is it worth the expense of acyclovir for one day less of symptoms? Is it worth the trouble (those of you who have attempted to give a toddler a medicine four times a day know what I mean)? Clearly, this is not a statistical question. In practice, most pediatricians do not prescribe acyclovir for immunocompetent children.

At times, small improvements in the time to outcome may spur scientific progress, even if there is minimal benefit to individual patients. Most medical advances are incremental. Proving that a particular strategy increases survival, if only marginally, may provide a valuable lead to a better treatment.

TIP

Time to occurrence models allow inclusion of subjects with differing lengths of follow-up.

A second reason for preferring time to occurrence models is that they allow you to incorporate subjects with differing lengths of follow-up in your analysis. Differing lengths of follow-up occur commonly in longitudinal studies for a variety of reasons. Subjects may decide they no longer wish to participate. They may move. They may die. They may have to be withdrawn due to a side effect they have developed. Sometimes you will not know why the person is lost to follow-up; they just are.

Coping with subjects who do not finish a study is one major way in which clinical research is different from other types of research. In laboratory research, it is usually possible to control conditions (e.g., laboratory animals, cell cultures) so that no observations are lost. Much of social science research (and some clinical research) is conducted using cross-sectional designs (surveying people about independent variables and outcomes at the same time). Thus while there are always people who choose not to participate, subjects are not lost during

[6] Balfour, H. H., Kelly, J. M., Suarez, C. S., *et al.* "Acyclovir treatment of varicella in otherwise healthy children." *J. Pediatr.* **116** (1990): 633–99.

Table 3.3 Reasons for censoring observations.

Reason for censoring	Examples
1 Lost to follow-up.	Subject moves, doesn't wish to participate, stops attending a particular clinic.
2 Subject has an outcome that precludes the study outcome (also known as alternative outcomes or competing risks).	Death from coronary artery disease in a study of cancer incidence.
3 Subject is withdrawn from study.	Development of side effects, not ethical to continue treatment or placebo.
4 Varying dates of enrollment.	Study enrolls subjects over a two-year period.
5 End of study.	All subjects who have not experienced outcome are considered censored at the end of the study.

the study. However, for longitudinal studies (studies of people over time) we need a way to deal with persons who do not complete the study.

One method of dealing with such subjects is to delete them. Indeed, if you are using a simple cumulative outcome at a particular point in time you have no choice but to drop subjects who leave the study prior to completion. If, for example, your cumulative outcome is breast cancer at three years, and you have a subject who dropped out of your study at two-and-a-half years, you would have to omit her from the study. Omitting subjects decreases the power of a study and potentially introduces bias.

TIP

Censoring allows subjects to contribute information until they leave the study.

What we would ideally like is a technique that allows subjects to contribute information until they leave the study. Such a technique exists. It is called censoring and it is a major element of all types of survival analysis, including proportional hazards analysis.

DEFINITION

Censoring is used to incorporate subjects with differing lengths of follow-up for a variety of reasons.

Besides allowing us to incorporate subjects who are lost to follow-up, censoring has broader implications. It allows us to analyze, within one study, subjects with unequal lengths of follow-up for a variety of reasons (Table 3.3). Indeed, all subjects in a proportional hazards analysis who do not experience the outcome of interest are censored, if not during the course of the study, then at the end of the study.

In summary, time to outcome has two major advantages over cumulative outcome at a particular point: it is a more sensitive measure of efficacy and it allows inclusion of subjects with unequal lengths of follow-up. If you have relatively few outcomes, length of follow-up is relatively short, and few subjects

are lost to follow-up, then using cumulative outcome and logistic regression will provide similar results as using time to outcome and proportional hazards analysis.[7] This was illustrated in a study of mortality among patients who had coronary angiography and were judged to need coronary revascularization.[8] Of 671 patients, 70 (10.4%) were known to have died, the median follow-up was 797 days, and mortality data were available for all subjects. Multiple logistic regression, adjusted for a number of potential confounders, showed that the odds of death at one year were significantly lower among patients who received coronary revascularization (OR = 0.49; 95% CI = 0.30 − 0.84) compared to those who did not receive it. A proportional hazards analysis, which also adjusted for confounders, found that revascularization significantly reduced the risk of death (relative hazard [RH] = 0.59; 95% CI = 0.36 − 0.97).

3.7 What type of multivariable analysis should I use with a rare outcome or a count?

TIP

Use Poisson regression for rare outcomes and outcomes that are expressed as counts.

For outcomes that occur rarely over time (<5%), proportional hazards analysis is not valid. Instead you should use Poisson regression. Poisson regression is a technique for analyzing counts – events that occur one or more times.[9] Although rates and counts seem disparate, remember that a rate is simply a count divided by a period of time.

As implied by the name, Poisson regression assumes that the outcome has a Poisson distribution. The characteristics of a Poisson distribution are that:

- It excludes negative numbers
- Is skewed to the right
- The variance is equal to the mean

Many clinical events (e.g., number of hospitalizations, number of infections) fit this distribution and may be more appropriate for Poisson regression than for other types of multivariable analysis. For example, with Poisson regression the outcome will be estimated to be zero or higher. In contrast with multiple linear regression the outcome may be estimated to have a negative value for certain subgroups of subjects (defined by the independent variables). Clearly, a

[7] Green, M. S. and Symons, M. J. "A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies." *J. Chron. Dis.* **36** (1983): 715–24.

[8] Kravitz, R. L., Laouri, M., Kahan, J. P., *et al.* "Validity of criteria used for detecting under use of coronary revascularization." *JAMA* **274** (1995): 632–8.

[9] For a more extensive review of Poisson regression see: Kleinbaum, D. G., Kupper, L. L., and Muller, K. E. *Applied Regression Analysis and Other Multivariable Methods* (2nd edn). Boston, MA: PWS-Kent, 1988, pp. 497–512; Gardner, W., Mulvey, E. P., and Shaw, E. "Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models." *Psych. Bull.* **118** (1995): 392–404; Simon, S. "Poisson regression model." Available at: http://www.cmh.edu/stats/model/poisson.asp; Grace-Martin, K. "Regression models for count data." StatNews #43, New York, NY: Cornell University, available at: http://www.human.cornell.edu/admin/statcons/statenews/stnews43.htm.

clinical event cannot have a negative value. Second, because negative numbers are not possible with counts and there is no limit on how high a count may go, the distributions of counts tend to be skewed to the right. In contrast, multiple linear regression assumes a normal distribution of the outcome. Finally, since negative numbers are not possible with counts, if you have a subgroup with a very low mean you would expect that almost all the values would be equal to zero. This would result in a very small variance. In a different subgroup with a higher mean, you would expect the distribution to be spread across a larger group of values and therefore have a larger variance.[10] In contrast with multiple linear regression the variance of an outcome is assumed to be equal in all subgroups (Section 5.7).

Poisson regression models the natural log of the outcome as a linear function of the independent variables. It is therefore sometimes referred to as a log-linear model. Poisson regression will provide you with coefficients, rate ratios, and 95% confidence intervals.

Poisson regression assumes that the probability of an occurrence is constant over time. In other words, the incidence of disease is not known to increase over the study period. Therefore it would be inappropriate for studying occurrences of a very highly contagious disease such as chicken pox.

Poisson regression also assumes that occurrences of an outcome are independent of one another. In most clinical situations this assumption is not true; that is, a patient who experiences one event (one infection, one hospitalization) is generally more likely to experience another event than a subject who has not experienced any events. For this reason, with most clinical events you will want to use generalized estimating equations or mixed-effects models (Chapter 12) to adjust for the correlation between observations.

A study of the effect of antidepressants on the risk of falls among nursing home residents illustrates the strengths and weaknesses of Poisson regression.[11] Because residents could not have a negative number of falls, it would be inappropriate to use multiple linear regression which could result in estimating a value of less than zero falls for certain subgroups of patients. Because some residents had more than one fall, the outcome would not be appropriate for multiple logistic regression, for which the outcome must be zero or one. Similarly, although the study was longitudinal, standard proportional hazards analysis can only incorporate events that occur once over time (adaptations to proportional hazards analysis to incorporate events that can occur more than once are discussed in Sections 12.1.F and 12.1.G).

[10] Gardner, W., Mulvey, E. P., and Shaw, E. "Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models." *Psych. Bull.* **118** (1995): 392–404.

[11] Thapa, P. B., Gideon, P., Cost, T. W., *et al.* "Antidepressants and the risk of falls among nursing home residents." *N. Engl. J. Med.* **339** (1998): 875–82.

Using Poisson regression, the authors found that, compared to nonusers, new users of tricyclic antidepressants (rate ratio = 2.0; 95% CI 1.8–2.2), selective serotonin-reuptake inhibitors (rate ratio = 1.8; 95% CI 1.6–2.0), and trazadone (rate ratio = 1.2; 95% CI 1.0–1.4) had higher rates of falls. However, their data do not fulfill the assumption of Poisson regression that the occurrences of the outcome be independent of one another. Elders who fall are more likely to fall again (clustering by subject) and a resident in one nursing home is more likely to have a similar number of falls as a resident in the same nursing home than a resident in a different nursing home (clustering by facility). The investigators therefore reanalyzed their data using a Poisson mixed-effects model (Chapter 14) adjusting for clustering by subject and facility. This analysis had essentially identical results to the simpler Poisson regression.

When studying rare outcomes, such as the development of a rare disease, repeated observations of the events is not an issue. For example, Kittner and colleagues used Poisson regression to examine the risk of stroke caused by pregnancy.[12] In the study catchment area there were 1 051 113 women of reproductive age. The investigators found 31 strokes during 8 011 852 weeks of exposure time (exposure time was the pregnancy and the six-week post-partum period) and 223 strokes during 101 303 016 weeks of nonexposure time (not pregnant). Thus in the study there was a total of 254 strokes in 109 314 868 weeks (incidence of 0.01 strokes per hundred women-years). Despite this rare outcome, Poisson regression showed a significantly elevated risk of stroke associated with pregnancy (rate ratio for pregnancy was 2.4; 95% CI $= 1.6 - 3.6$), after adjustment for age and race.

You will sometimes find that the variance of a count is much larger than the mean. This is referred to as overdispersion. In such cases, Poisson regression underestimates the standard errors for the coefficients. To overcome this problem you can perform an overdispersed Poisson model by including a parameter that estimates how much larger the variance is than the mean. The parameter is used to correct the standard errors. Alternatively, you can use negative binomial regression. This technique is a form of Poisson regression that includes a random component with the result that the model may better represent the relationship between the expected value and the variance of the outcome than a standard Poisson regression.[12]

[12] Kittner, S. J., Stern, B. J., Feeser, B. R., *et al.* "Pregnancy and the risk of stroke." *N. Engl. J. Med.* **335** (1996): 768–74.

4

Type of independent variables in multivariable analysis

4.1 What type of independent variables can I use with multivariable analysis?

Interval and dichotomous independent variables can be used in all three types of multivariable analysis (Table 4.1). Ordinal and nominal variables cannot be used with any of these techniques without transforming the variables.

4.2 What should I do with my ordinal and nominal independent variables?

Don't despair. Ordinal and nominal independent variables can be incorporated into all three multivariable models by transforming them into multiple dichotomous variables. This process is usually called "dummying" by epidemiologists and biostatisticians. However, the terms "dummying" and "dummy variables" are slang. In manuscripts, you should refer to this process as creating multiple categorical variables (if you refer to it as dummying, you may, as I did, receive complaints from the reviewers of your article).

Ethnicity is probably the most common nominal variable in clinical research. Obviously, there is no numeric ordering of different ethnicities, let alone a fixed interval between them. Therefore, ethnicity is either dichotomized (e.g., white/nonwhite) or (better) is represented as several dichotomous variables in multivariable analysis. Below I have represented ethnicity as five dichotomous variables:

African-American (yes/no)
Latino/Hispanic (yes/no)
Asian/Pacific Islander (yes/no)
Native American (yes/no)
Other nonwhite (yes/no)

What happened to persons who are white/Caucasian? When you represent a nominal variable as several dichotomous variables in multivariable analysis you

Table 4.1 Independent variables and multivariable analysis.

Type of independent variable	Example of independent variable	Multiple linear regression	Multiple logistic regression	Proportional hazards analysis
Interval	Age, blood pressure	Yes	Yes	Yes
Dichotomous	Gender	Yes	Yes	Yes
Ordinal	Cancer stage	No	No	No
Nominal	Ethnicity	No	No	No

Table 4.2 Creation of multiple dichotomous variables to represent a nominal independent variable.

	African-American	Latino/Hispanic	Asian/Pacific Islander	Native American	Other nonwhite
African-American	1	0	0	0	0
Latino/Hispanic	0	1	0	0	0
Asian/Pacific Islander	0	0	1	0	0
Native American	0	0	0	1	0
Other nonwhite	0	0	0	0	1
White/Caucasian	0	0	0	0	0

need one variable less than the number of categories of your variable. Why? To answer this question, think about it from the computer's point of view. If you create five dichotomous variables, all of which are either 1 (yes) or 0 (no), the computer will see six patterns as shown in Table 4.2.

We don't create a variable white/Caucasian because it is represented by the other five variables (zero on all five variables). In multivariable analysis, this is called a reference group.

I deliberately chose ethnicity as an example of a nominal variable because how you choose to code it will depend on your study population. For example, in a small clinical study performed in the southeast of the United States, there may be very few Native Americans or Asian/Pacific Islanders. If a group represents less than 5 percent of the total sample, creating a variable for that group may not carry much statistically important information. In this case you might only create variables for the larger ethnic groups and then have a group that is "other." For example, your variables would be African-American (yes/no), Latino/Hispanic (yes/no), and other nonwhite ethnicity (yes/no), with white as the reference group.

Although decreasing the number of groups may prevent having dichotomous variables that convey little information, grouping people of different ethnicities in one group may not adequately represent the data. Even if you retain the category "Asian/Pacific Islander" remember that this category contains more than a dozen disparate cultures, each with their own language, traditions, and genetic composition, all of which could affect the development of disease. As with all really hard questions in multivariable analysis, the question of how to code ethnicity is not a statistical question. The best way to group a nominal independent variable such as ethnicity will depend on the research question, the distribution of the nominal variable (how many people are in each group), and the relationship between the different categories of the nominal variable and the outcome.

TIP

The best way to group a nominal variable will depend on the research question, the distribution of the nominal variable, and the bivariate relationship between the nominal variable and the outcome.

The process of creating multiple dichotomous variables has other uses besides allowing you to incorporate ordinal and nominal variables. It also works in the case of interval variables, for which the relationship between the untransformed interval-independent variable and the outcome does not fit the assumptions of the model. This issue is dealt with in greater detail in Section 5.5.

As discussed in Section 3.4, some variables, while technically ordinal, operate as if they were interval. Just as you can use certain ordinal variables as dependent variables in multiple linear regression, you can use such variables as independent variables in all three types of analysis, as long as they fulfill the assumptions of your model.

5

Assumptions of multiple linear regression, multiple logistic regression, and proportional hazards analysis

5.1 What are the assumptions of multiple linear regression, multiple logistic regression, and proportional hazards analysis?

As shown in Table 5.1, the assumptions underlying the three multivariable models differ somewhat with respect to what is being modeled, the relationship of multiple independent variables to outcome, the relationship of an interval-independent variable to the outcome, the distribution of the outcome variable, and the variance of the outcome variable. These assumptions are explained in this chapter.

Proportional hazards analysis has two additional assumptions with regard to censored observations and relative hazards over time (referred to as the proportionality assumption). These are dealt with in Sections 5.9 and 10.10, respectively.

5.2 What is being modeled in multiple linear regression, multiple logistic regression, and proportional hazards analysis?

In multiple linear regression, as the independent variable increases (or decreases) the mean or expected value of the outcome increases (or decreases) in a linear fashion. Many clinical situations fit this linear assumption.

For example, Figure 5.1 shows the relationship between B_{12} levels and pneumococcal antibody levels following receipt of pneumococcal vaccination among elderly persons.[1] Each square represents an observation (a person) and their vitamin B_{12} level (the independent variable), and their antibody titer after vaccine (the dependent variable). Although arbitrary, the convention is to show the independent variable on the *x*-axis and the dependent variable on the *y*-axis.

[1] Fata, F. T., Herzlich, B. C., Schiffman, G., *et al.* "Impaired antibody responses to pneumococcal polysaccharide in elderly patients with low serum vitamin B_{12} levels." *Ann. Intern. Med.* **124** (1996): 299–304.

Table 5.1 Multivariable model assumptions.

	Multiple linear regression	Multiple logistic regression	Proportional hazards analysis
What is being modeled?	The mean value of the outcome.	The logarithm of the odds of the outcome (referred to as *logit*).	The logarithm of the relative hazard.
Relationship of multiple independent variables to outcome	The mean value of outcome changes *linearly* with multiple independent variables.	The logit of the outcome changes *linearly* with multiple independent variables.	The logarithm of the relative hazard changes *linearly* with multiple independent variables.
Relationship of an interval-independent variable to outcome	The mean value of outcome changes *linearly* with each unit change in interval-independent variable.	The logit of outcome changes *linearly* with each unit change in interval-independent variable.	The logarithm of the relative hazard changes *linearly* with each unit change in interval-independent variable.
Distribution of outcome variable	Normal	Binomial	None specified
Variance of outcome variable	Equal around the mean	Depends only on the mean	None specified
Censored observations	Not applicable	Not applicable	Censored cases have the same time to outcome as noncensored cases
Relative hazards over time	Not applicable	Not applicable	Proportional (Section 10.10)

The linear regression line shows the best single representation of the data. But note that only two points actually fall on the line and many points are not even that close to the line. Statistical models are, at best, approximations of data. The linear assumption does not mean that all observations fit a linear model; it means that a line is a good representation of the fact that as vitamin B_{12} levels increase, the levels of antibodies also increase.

In clinical research the outcome is often a disease or disease state that cannot be measured on an interval scale: cancer, stroke, heart attack, death. In Section 3.5, I indicated that logistic regression is used for dichotomous outcomes. Although this is true, you can also use linear regression for dichotomous outcomes. It just doesn't work as well in most cases. Understanding why will lead us to the advantages of the logistic regression model.

In Figure 5.2, we see the association between skeletal muscle strength (measured in the deltoid muscle) and presence of cardiomyopathy among

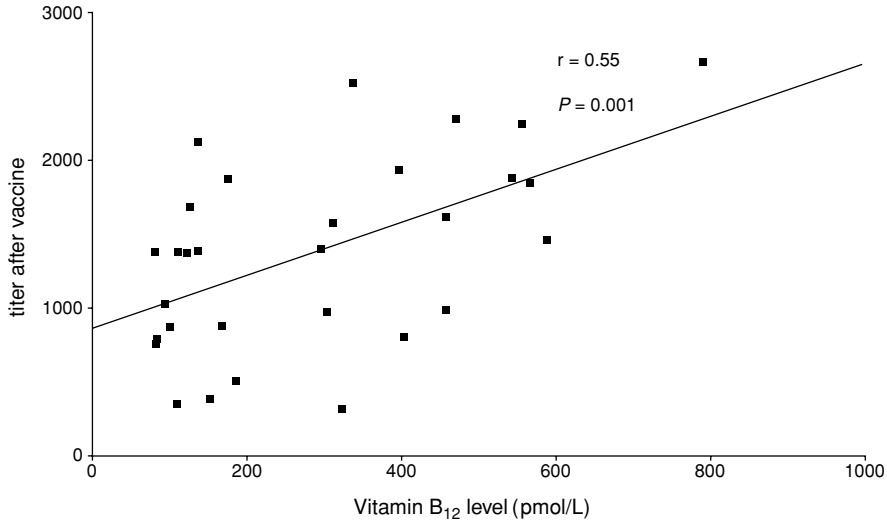


Figure 5.1 **Linear association between vitamin B_{12} levels and pneumococcal antibody titers after pneumococcal vaccination. Reproduced with permission from: Fata, F. T., *et al*. "Impaired antibody responses to pneumococcal polysaccharide in elderly patients with low serum vitamin B_{12} levels." *Ann. Intern. Med*. **124** (1996): 299–304.**

alcoholics.[2] You can see that at low levels of muscle strength (left-hand side of the curve), there are several closed circles (representing patients with cardiomyopathy) while there are no open circles (patients with normal cardiac function). In contrast, at high levels of muscle strength (right-hand side of the curve) there are many open circles and few closed circles, reflecting that patients without cardiomyopathy have normal muscle strength. The triangles indicate the probability of cardiomyopathy at different levels of muscle strength. The curve connecting the triangles shows that at intermediate levels of muscle strength, there is a rapidly decreasing proportion of patients with cardiomyopathy as muscle strength increases.

DEFINITION

The *logit* is the natural logarithm of the odds of the outcome. The *odds* of the outcome is the probability of having the outcome divided by the probability of not having the outcome.

The probability of cardiomyopathy, as with any event, cannot be less than zero or greater than one. The value of the logistic function is that it incorporates this assumption. Logistic regression models the logit of the outcome. The logit is the natural logarithm of the odds of the outcome. The odds of the outcome is the probability of having the outcome divided by the probability of not having the outcome. Whereas the logit can take on any value from minus to plus infinity, the probability, which is the inverse of the logit, can only take on values of zero

[2] Fernandez-Sola, J., Estruch, R., Grau, J. M., *et al.* "The relationship of alcoholic myopathy to cardiomyopathy." *Ann. Intern. Med.* **120** (1994): 529–36.

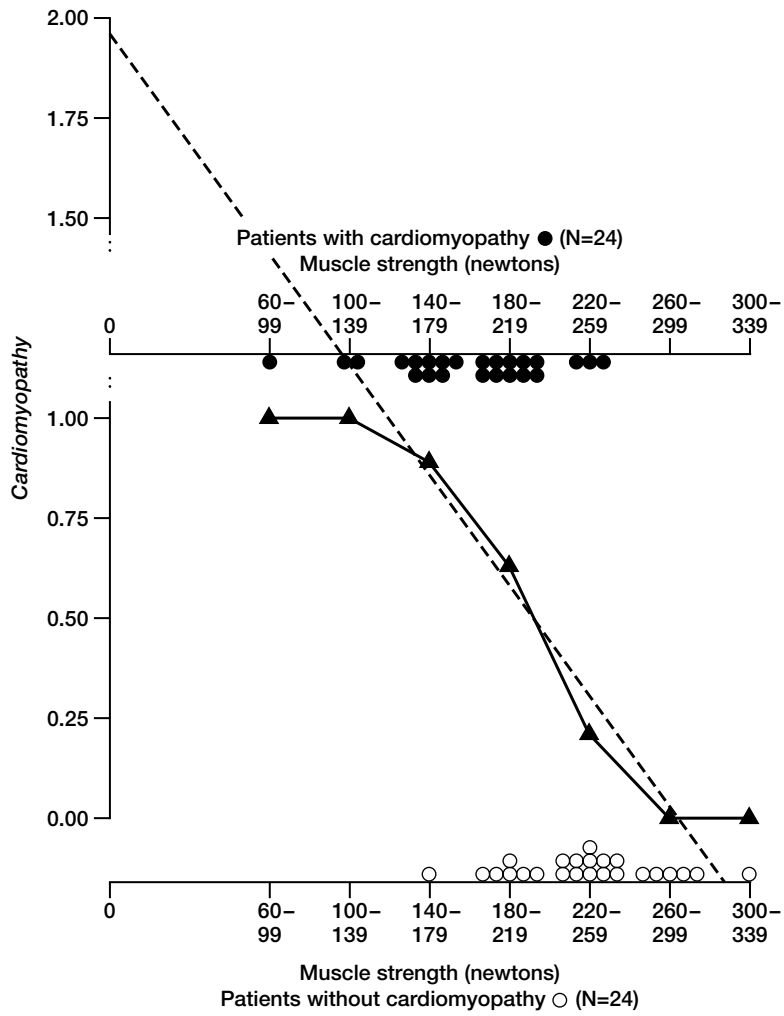


Figure 5.2 Z-shaped association between skeletal muscle strength and presence of cardiomyopathy among alcoholics. The closed circles are patients with cardiomyopathy and the open circles are patients without cardiomyopathy. The triangles show the observed proportion of patients with cardiomyopathy at different levels of muscle strength. Data are from Fernandez-Sola, J., Estruch, R., Grau, J. M., *et al*. "The relationship of alcoholic myopathy to cardiomyopathy." *Ann. Intern. Med*. **120** (1994): 529–36.

to one. This gives the logistic function an S or Z shape (depending on which way the outcome variable is configured). As you can see in Figure 5.2, this shape fits the data. Note that as the probability of outcome approaches zero or one, further increases or decreases in the independent variable have little effect on the outcome.

I have drawn a linear regression line (the dotted line) so that you can appreciate that a linear function could be used to model the data. But there are problems. Cardiomyopathy is either present or not. Yet the line would predict that a patient with muscle strength of less than 100 newtons would have a value greater than one on the cardiomyopathy variable. The line would also indicate that a patient with muscle strength of 300 newtons or greater would have a negative value on the cardiomyopathy variable. Obviously, these values are impossible.

Although this S- or Z-shaped function is useful, remember that it is still just a model. The curve I have drawn in Figure 5.2 is based on the actual data. The estimated curve would be of a similar shape but would not go exactly through the boxes. We will take up the issue of how to assess how well a model fits the observed data in Section 9.2.

DEFINITION

The *relative hazard* is the ratio of time to outcome given a particular risk factor to time to outcome when the risk factor is not present.

Based on Table 3.1 it would be logical to assume that proportional hazards analysis models time to event. While logical, this is incorrect. Proportional hazards analysis models the natural logarithm of the relative hazard. The relative hazard is the ratio of time to outcome given a particular risk factor to time to outcome when the risk factor is not present. For example, persons who smoke may have six times the risk of heart attack as persons who do not smoke. The fact that the underlying time to outcome (in this example, heart attack) is not modeled is one of the features that gives proportional hazards analysis its tremendous flexibility.

5.3 What is the relationship of multiple independent variables to outcome in multiple linear regression, multiple logistic regression, and proportional hazards analysis?

In Section 5.2, we reviewed what was being modeled by each of the three models. However, in the examples, for the sake of simplicity, I illustrated the principles with a single independent variable. Of course, the whole point of multivariable analysis is to include multiple independent variables. What is the relationship of multiple independent variables to outcome? As you can see from Table 5.1 all three models assume a linear component to the relationship of multiple independent variables to outcome, but the relationship is linear on different scales.

TIP

All three multivariable models assume a linear component to the relationship of multiple independent variables to outcome.

With multiple linear regression, the expected value of the outcome changes linearly with the weighted sum of the independent variables. With multiple logistic regression, the logit changes linearly with the weighted sum of the independent variables. With proportional hazards analysis, the logarithm of the relative hazard changes linearly with the weighted sum of the independent variables. In all

three models, the weights are determined by the strength of the independent variables in accounting for the outcome.

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?

As you will remember, each unit of an interval variable is equal and quantifiable. All three multivariable methods assume that a unit change anywhere on the scale of an interval variable will have a linear effect on outcome. However, just as with multiple independent variables, the relationship is linear on different scales.

In the case of multiple linear regression, the change in the mean value of the outcome is modeled as the sum of the unit changes of the interval-independent variable. With logistic regression, the logit of the outcome is modeled as the sum of the unit changes of the interval-independent variable. With proportional hazards analysis, the logarithm of the relative hazard is modeled as the sum of the unit changes of the interval-independent variable.

TIP

The linearity assumption for linear regression is best tested by performing a scatter plot of your raw independent variable versus your outcome variable.

The linearity assumption for linear regression is best tested by constructing a scatter plot of your raw independent variable versus your outcome variable. Look back at Figure 5.1. It is a typical scatter plot showing a linear relationship between two interval variables, with higher values of B_{12} being associated with higher titers after vaccine.

Sometimes a scatter plot will not show a linear relationship. Instead, you may see one of the shapes in Figure 5.3: logarithmic, antilogarithmic, curvilinear (**U**-shape, upside down **U**–shape, **J**-shape), and threshold. If you see a nonlinear relationship, you cannot proceed with linear regression without transforming the variable or using a spline function (Section 5.5). A limitation of the use of a bivariate scatter plot to determine linearity is that it is possible for a relationship between an interval-independent variable and an interval outcome to be linear in a bivariate analysis but not in a multiple linear regression analysis owing to confounding. Methods of testing for linear relationships in multiple linear regression models are described in Section 10.2.

In the case of multiple logistic regression and proportional hazards analysis, you cannot assess the linear assumption by making a simple scatter plot. This is because the linear relationship does not exist on a simple arithmetic scale. Instead, to assess whether an interval variable fits the linear assumption of logistic regression or proportional hazards analysis, categorize the variable into multiple dichotomous variables (Section 4.2) of equal units on the variable's scale. So, for example, if the variable you are testing is age, and the age of your subjects ranges from 20 to 79, have age 20–29 be your reference group. Then create

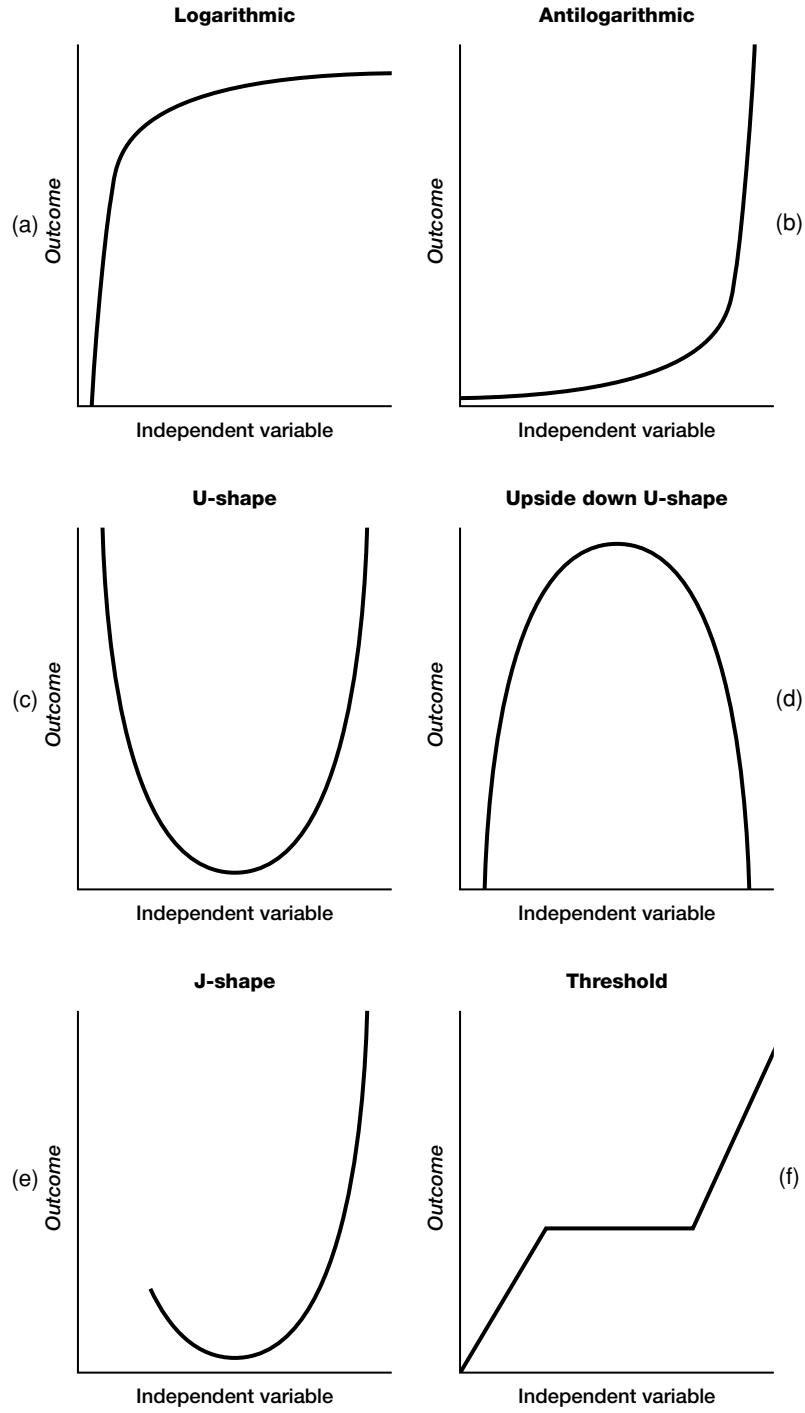


Figure 5.3 Variety of nonlinear relationships between an independent variable and an outcome.

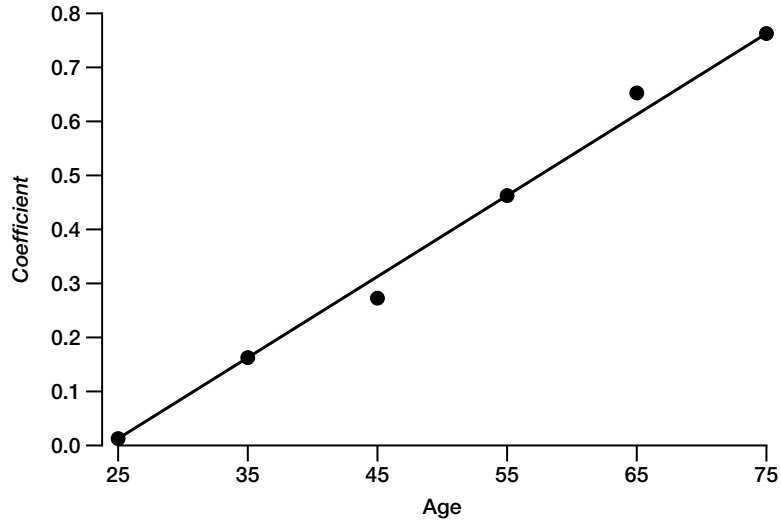


Figure 5.4 Coefficients graphed against age form a straight line.

TIP

To assess whether an interval variable fits the linear assumption of logistic regression or proportional hazards analysis: (1) Categorize the variable into multiple dichotomous variables of equal units; (2) enter these dichotomous variables in the analysis; and (3) graph each variable's coefficient against the midpoint of the variable.

variables 30–39 (yes/no), 40–49 (yes/no), 50–59 (yes/no), 60–69 (yes/no), and 70–79 (yes/no). If you would have too few subjects being yes in these decade categories, you can group them into 20-year periods.

Perform the logistic or proportional hazards analysis with these several dichotomous variables. Each variable will have an estimated coefficient. The coefficient for the reference group is, by definition, 0. Graph the coefficient against the midpoint of each dichotomous variable (e.g., 35 years for the variable that represents the 30–39 group). The graph will show you the relationship between your independent and outcome variable. If you have a linear effect, the coefficients will steadily increase (or decrease) as you go from one age group to another, and you will get a straight line (as shown in Figure 5.4). Alternatively, your graph may appear like one of the nonlinear relationships in Figure 5.3.

Even without graphing, if you have a linear effect, you should be able to see it because the numeric difference between the coefficients of each successive group should be about equal (e.g., the numeric difference between the coefficient of the 30–39 group and the coefficient of the 20–29 group will be equal to the difference between the coefficient of the 40–49 variable and the 30–39 variable). Remember, of course, they are not going to be exactly equal (just as the points do not fall exactly on a straight line). The important issue is whether the data can be reasonably expressed as a linear relationship.

This technique can be used to determine whether a linear relationship exists between an independent variable and an outcome variable after adjustment for potential confounders. Besides being useful for testing a linear association,

this technique is also useful for demonstrating a linear dose–response curve between an independent variable and outcome.[3] Additional tests to see how well a variable fulfills the linear assumption for logistic regression and proportional hazards analysis are described in Section 10.3.

There is another bivariate method of assessing whether an interval-independent variable has a linear association with the outcome that can be done prior to logistic regression. It requires grouping your interval-independent variable data into categories that preserve the interval nature of the variable (e.g., 1 = ages 20–29, 2 = ages 30–39, 3 = ages 40–49, etc.) and are large enough to provide a reasonable number of outcomes in each category. You can then perform a simple cross-tabulation of your independent variable and outcome. The cross-tabulation table should show a steadily increasing (or decreasing) proportion of outcomes as you increase (or decrease) along the levels of your interval-independent variable. The chi-square for trend test should be significant.

5.5 What if my interval-independent variable does not have a linear relationship with my outcome?

In clinical research, interval-independent variables often have the kind of nonlinear relationships with outcome that are shown in Figure 5.3. If you treat these nonlinear relationships as if they were linear, your results may appear sensible. No alarms will go off, your computer will not melt down, and your printout will not look like number salad. Your results may show no relationship between the independent variable and the outcome or may show a weak linear relationship, especially if there are a large number of points along the scale where the relationship is linear.

Modeling such variables as if they were linear clearly is not right. However, don't be discouraged if your variable does not fit the linear relationship assumed by your model. This observation provides valuable information and by transforming the variable you will still be able to keep it in your model.

TIP

If changes in value at the high end of your independent variable have less impact on your outcome variable than changes at the lower end, try a logarithmic transformation of the independent variable to linearize the trend.

If changes in value at the high end of your independent variable have less impact on your outcome variable than changes at the lower end (as indicated by a steadily decreasing slope), with the high end of the independent variable asymptotically approaching a horizontal level as in Figure 5.3(a), a logarithmic transformation of the independent variable (the logarithm of the variable) may

[3] If you are unfamiliar with multiple logistic regression and proportional hazards analysis, this section will make more sense to you after you have read Section 9.3 on interpretation of coefficients. For a more detailed explanation of this technique see Hosmer, D. W. and Lemeshow, S. *Applied Logistic Regression*. New York, NY: Wiley, 1989, pp. 95–7.

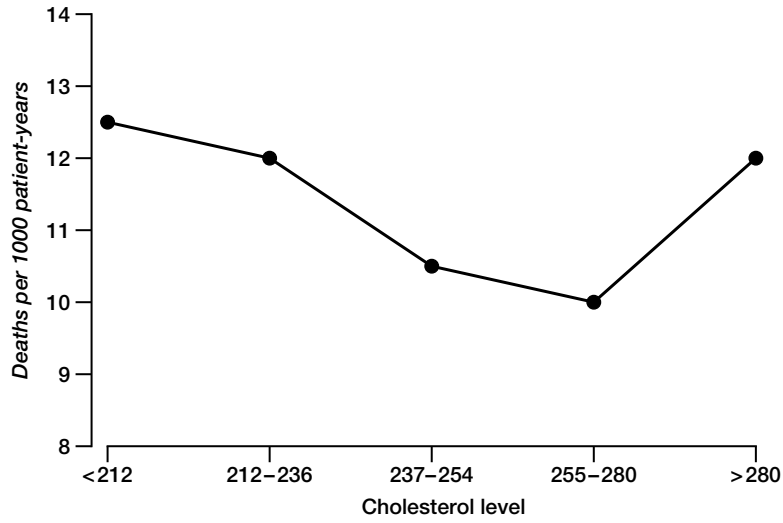


Figure 5.5 Relationship between cholesterol level and all-cause mortality among 1102 women. Adapted with permission from: Isles, C. G., *et al*. "Plasma cholesterol, coronary heart disease, and cancer in the Renfrew and Paisley survey." *Br. Med. J.* **298** (1989): 920–4. Copyright BMJ Publishing Group.

linearize the trend. The natural logarithm is used more often than the logarithm to the base 10, although both may linearize the effect. Remember that with either logarithmic transformation, values for the variable must be positive (i.e., you cannot take the logarithm of zero or negative numbers). If your scale has a true zero you can still use a logarithmic transformation by adding one to all values.

If changes in value at the high end of your independent variable have a greater impact on your outcome variable than changes at the lower end (as indicated by a steadily increasing slope as in Figure 5.3(b)), an antilogarithm transformation (i.e., e^x or 10^x) of the independent variable may linearize the trend. Logarithmic or antilogarithmic transformations can be made of the independent or dependent variable.[4]

TIP

If changes in value at the high end of your independent variable have a greater impact on your outcome variable than changes at the lower end, try an antilogarithm transformation of the independent variable to linearize the trend.

In Figure 5.5, we see the **U**-shaped relationship between cholesterol level and all-cause mortality in a sample of 1 102 women.[5] Mortality is highest for women with the lowest and the highest values of cholesterol. When the investigators treated cholesterol as an interval variable, there was no significant relationship between cholesterol level and mortality, because the two trends statistically cancel each other out. Treating cholesterol as an interval variable misses the vital

[4] Other mathematical transformations are possible: See Armitage, P. and Berry, G. *Statistical Methods in Medical Research* (2nd edn). Oxford: Blackwell Scientific Publications, 1987, pp. 358–68.

[5] Isles, C. G., Hole, D. J., Gillis, C. R., Hawthorne, V. M., Lever, A. F. "Plasma cholesterol, coronary heart disease, and cancer in the Renfrew and Paisley survey." *Br. Med. J.* **298** (1989): 920–4.

information contained in the curvilinear relationship: High cholesterol levels are associated with increased mortality from coronary artery disease, whereas low levels of cholesterol are associated with increased mortality from cancer and other causes.

As you can tell from Figure 5.3, a **J**-shaped curve is just like a **U**-shaped curve with a few missing data points. What can you do when your bivariate plots show a **U**- or **J**-shaped relationship? One solution is to include in your model a quadratic form of the variable in addition to the untransformed value of the variable.

To create the quadratic form of a variable first subtract out the mean of the untransformed variable (X) and then square the result: (value of X – mean of X for sample)^2. The untransformed variable must be in the equation because the quadratic term is comparing extremes to the mean of the untransformed variable. Large differences from the mean in either direction cause the term to be statistically significant. If the relationship is **U**- or **J**-shaped both terms will be statistically significant in your model. A limitation of this technique is that it is hard to describe to your reader how a unit change in the independent variable affects the outcome (because a unit change affects the outcome through two variables – each with a different relationship with the outcome).

DEFINITION

Splines are polynomials linked together; they are used to model complex relationships between interval-independent variables and outcomes.

Another method for incorporating **J**-shaped, **U**-shaped, threshold effects and other complex relationships between interval-independent variables and outcomes is to use splines.[6] The term spline originates from the flexible strip of metal used by draftsmen to draw curves. In the statistical sense, splines are polynomials (an algebraic function of two or more summed terms) that are connected to one another. The points at which they are connected are called knots. Because each piece of the curve is represented by a different polynomial, splines can be used to model a variety of complex relationships.

The simplest type of spline is a linear spline function. Although each piece of the function is linear, because you have multiple pieces, you can model nonlinear relationships between interval-independent variables and outcomes. For example, Figure 5.5 could be modeled as a linear spline function consisting of four linear segments. (If Figure 5.5 were a spline function it would have three knots at the cholesterol levels 212–236, 237–254, and 255–280; these are the points where the segments touch. The number of knots will always be one less than the number of segments.)

When modeling the relationship between a risk factor and an outcome you may find that not all the pieces are linear. Instead, you may have rounded curves.

[6] For a very lucid explanation of splines see: Harrell, F. E. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.* New York: Springer-Verlag, 2001, pp. 18–24.

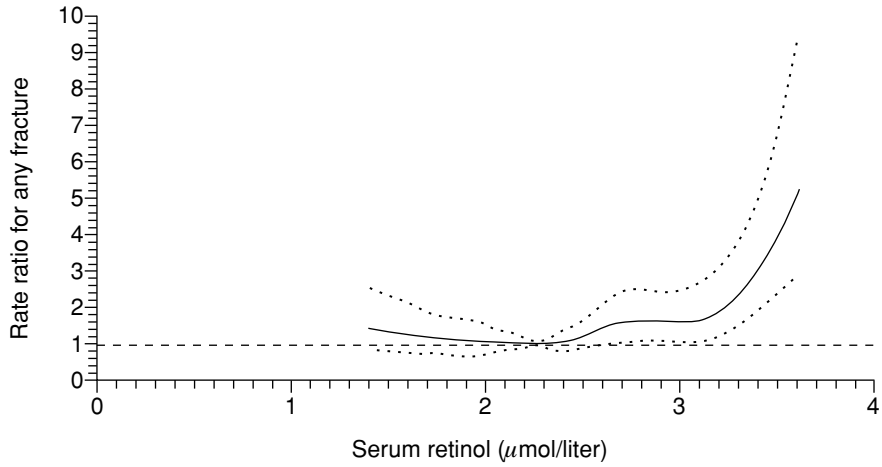


Figure 5.6 Smoothed plot of rate ratios for any fracture according to the serum retinol level. Plot based on use of restricted cubic splines and Cox regression analysis. Figure is from Michaelsson, K., Lithell, H., Vessby, B., *et al*. "Serum retinol levels and the risk of fracture." *N. Engl. J. Med.* **348** (2003): 287–94. Copyright 2003 Massachusetts Medical Society. All rights reserved.

TIP

For rounded curves use a restricted cubic spline function.

In this case you should use a cubic spline function. Because cubic spline functions are higher-order polynomials they better approximate curves.

A weakness of cubic spline functions is that they may not perform well at the tails (before the first knot and after the last knot). To overcome this problem use a restricted cubic spline function (also referred to as a natural cubic spline), which constrains the function to be linear beyond the boundary knots. The restricted cubic spline function also requires fewer parameters to be estimated than the cubic spline function.

Michaelsson and colleagues used a restricted cubic spline function and proportional hazards analysis to assess the relationship between serum retinol levels and the risk of bone fracture.[7] As you can see in Figure 5.6 the relationship between serum retinol level and fracture rate is not linear. In particular, the rate of fractures increases sharply at higher levels of retinol. As was done in this figure, splines can be smoothed.

Although statistically splines are a very satisfying solution, they are only slowly catching on in the medical literature. The major reason is that splines do not result in a single measure (e.g., odds ratio, relative risk) of the association of a risk factor with an outcome (as you would get if you treated the variable as if the association were linear).

[7] Michaelsson, K., Lithell, H., Vessby, B., *et al.* "Serum retinol levels and the risk of fracture." *N. Engl. J. Med.* **348** (2003): 287–94.

Table 5.2 Rate ratio for any fracture according to the base-line serum retinol level.

Retinol quintile	Multivariate RR* (95% CI)
1 (<1.95 μmol/liter)	0.93 (0.62–1.41)
2 (1.95–2.16 μmol/liter)	0.78 (0.50–1.23)
3 (2.17–2.36 μmol/liter)	1.00 (reference)
4 (2.37–2.64 μmol/liter)	0.91 (0.60–1.38)
5 (>2.64 μmol/liter)	1.64 (1.12–2.41)

* The analysis was adjusted for age, weight, height, and serum beta carotene, calcium, and albumin values (all continuous variables); smoking status (never smoked, former smoker, or current smoker); marital status (married or living with a partner vs. single); socioeconomic class (low, middle, or high); and physical activity at work, leisure physical activity, and alcohol consumption (all in three categories).
Data from: Michaelsson, K., *et al.* "Serum retinol levels and the risk of fracture." *N. Engl. J. Med.* **348** (2003): 287–94.

More commonly, although statistically less satisfactory, you will see investigators incorporate nonlinear relationships between an interval risk factor and an outcome by creating multiple dichotomous variables from the interval variable. To do this you would need to use the same procedure as you would use for incorporating a nominal variable into your analysis (Section 4.2). Multiple dichotomous variables allow each category to be its own independent variable and have its own relationship to the outcome.

In fact, in the study of serum retinol levels and fracture risks described above, the authors demonstrated their results using multiple dichotomous variables in addition to the analysis with a restricted cubic spline function. The results are shown in Table 5.2.

The risk of fracture is 1.64 times higher among persons in the highest quintile level compared with those in the middle quintile. You can see that presenting relative risks associated with the different quintiles of the serum retinol level tells the same story as the cubic spline function, but much less elegantly.

If you are going to report your results using multiple dichotomous variables, you will need to determine the cutoffs. If you choose the cutoff points in a way that maximizes the association between your independent variable and your outcome, you will overestimate the true association between these variables in the population. Therefore choose your cutoffs by dividing the sample into equal-sized groups by the independent variable (terciles, quartiles, quintiles of values

TIP

Choose cutoff points by dividing the sample into equal-sized groups or by using natural cutoffs.

of the independent variable), as was done in the study of serum retinol levels and fractures. Alternatively, where they exist, you can use natural cutoffs (e.g., decades of age, systolic blood pressure < 140 mm, 140–159 mm, 160–179 mm, etc.).

A disadvantage of choosing natural cutoffs, such as decades of age, is that the cutoffs may divide the sample into groups with unequal sample sizes. For example, if you divide your sample into decades of age, you may only have 2 percent of your sample as "yes" on the variable 80–89 years. If the number of persons who are yes on this variable is too small, the variable will not be meaningful in the analysis. In comparison, if you choose cutoffs that provide equal sample sizes then the distributions of the multiple dichotomous variables will be equal. For example, let's assume you divide the sample into terciles of age and make the youngest people the reference group. One dichotomous variable will have a "yes" value for a third of the sample (the middle-aged people) and a "no" value for two thirds of the sample (the youngest and oldest persons); the other dichotomous variable will also have a "yes" value for a third of the sample (the oldest persons) and a "no" value for two thirds of the sample (the youngest and the middle-aged persons).

The disadvantage of using cutoffs based on equal sample sizes is that the results may not sound as compelling because you lose the units of the independent variable. Which of the following sounds more compelling? Persons in the highest tercile of age are three times more likely to die than persons in the lowest tercile of age, or persons aged 70–89 years are three times more likely to die than persons aged 30–49 years. The latter sounds more compelling. Don't you think?

A downside of multiple dichotomous variables is that they increase the number of variables in your model. This can be a problem if you do not have a large enough sample size (Section 7.4).

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?

Even when an interval-independent variable, such as age, blood pressure, or cholesterol, fits the linear assumption it is often not left in its original interval form. There are several reasons for this.

For one thing, if you have a small sample size (e.g., 100 persons) it may be difficult to evaluate whether an interval variable (e.g., age) fits the linear assumption unless you group it into categories. Left ungrouped your model will assume that the difference in the likelihood of outcome between a subject age 55 and a subject age 57 is the same as that between a subject age 61 and a subject

age 63. Yet you may have no one or just one or two persons in your entire sample with these ages. Also, your audience is likely to be more interested in the impact of ten years on outcome, than the impact of a single year (which is likely to be very small for most diseases).

TIP

When grouping interval variables maintain the interval nature of the scale.

When grouping an interval variable maintain the interval nature of the scale (e.g., group age by decades). This will allow you to retain the advantages of an interval scale (because the difference between being 20–29, 30–39, and 40–49, etc. is the same – 10 years). Yet, you will be better able to assess whether the variable fits the assumptions of the statistical model and you will be able to report a more meaningful result.

Sometimes researchers create multiple dichotomous variables even though the variable approximates a linear relationship. The reason is that creation of multiple dichotomous variables is a more conservative strategy. Since a linear relationship is not assumed, you do not have to prove to your readers (or your reviewers!) that the linear assumption is fulfilled. However, this strategy results in an increase in the number of variables in your model; this may be a problem if your sample size is small (Section 7.4). Also a statistically significant linear trend between an interval risk factor and an outcome may no longer appear to be statistically significant when the interval variable is represented as multiple dichotomous variables because none of the variables themselves is significantly different from the reference group.

5.7 What are the assumptions about the distribution of the outcome and the variance?

Multiple linear regression assumes a normal distribution and equal variance around the mean.

Multiple linear regression assumes the outcome has a normal distribution and equal variance around the mean for any value(s) of the independent variable(s).[8] A normal distribution has a bell shape. To fulfill this condition your outcome variable should have a bell-shaped curve for any value of your independent variable. This is shown in Figure 5.7(a). Note that for each of the three values of the independent variable (X_1, X_2, and X_3), the range of values of the dependent variable forms a bell-shaped curve. Equal variance means that the spread of dependent variable values (indicated by arrows) from the mean (indicated with a dotted line) is equal for each value of X. This is the case in Figure 5.7a.

[8] The technical term for equal variance for any value of X is homoscedasticity; "homo" meaning same and "scedastic" from the Greek word "to scatter." For a more detailed explanation of these and the other assumptions of linear regression, see Kleinbaum, D. G., Kupper, L. L., and Muller, K. E. *Applied Regression Analysis and Other Multivariable Methods* (2nd edn). Boston: PWS-Kent, 1988, pp. 44–9.

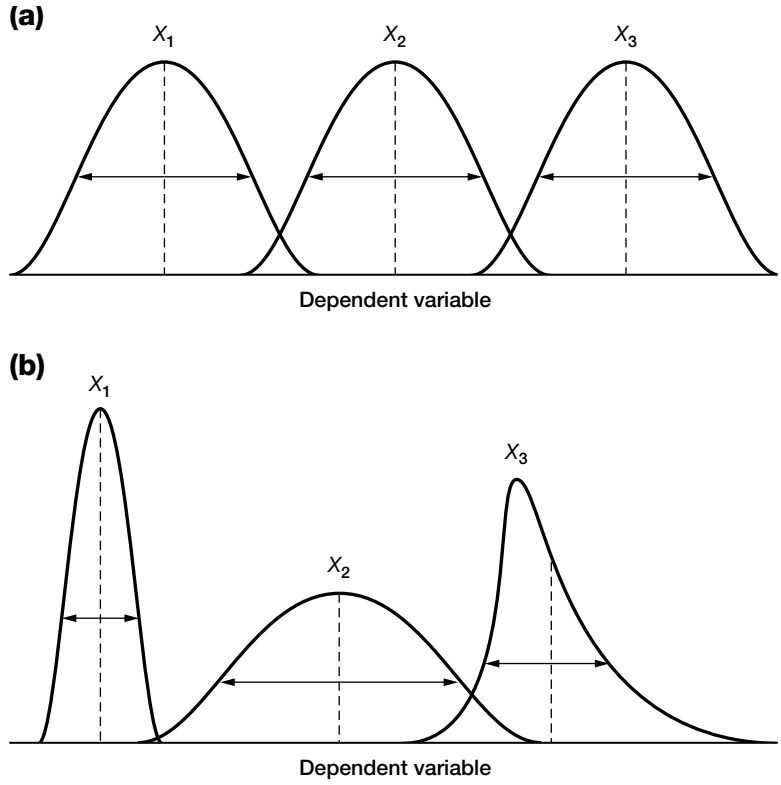


Figure 5.7 Plots of an interval-dependent variable at three different values of the independent variable *X*. In Figure 5.7(a), the assumptions of normal distribution and equal variance are fulfilled because at all three values of *X*, the curves are bell-shaped and the spread from the mean (indicated by arrows) is equal. In Figure 5.7(b) these assumptions are not met. The assumption of normal distribution is violated because the distribution of values for X_3 does not form a bell-shaped curve. The equal-variance assumption is invalid because the spread of values from the mean is different for the three values of *X*.

If you have an interval-independent variable, it is easier to assess these assumptions if you group the independent variable into a few groups. So, for example, in Figure 5.7(a), X_1, X_2, and X_3 may represent a range of values (e.g., age 20–39 years, 40–59 years, 60–79 years).

Figure 5.7(b) shows a bivariate relationship that does not fulfill the assumptions of normal distribution or equal variance. Note that the values of the dependent variable for X_1 and X_2 both produce bell-shaped curves, but the variance is not equal. It is much smaller for values of X_1 than for values of X_2. At X_3 we see a curve with a skewed distribution (long tail). This curve does not have a bell-shaped distribution and the variance is not equal to either of

the other two curves. Therefore, the relationship between this independent and dependent variable does not fit the assumptions of normal distribution or equal variance. Conceptually you should think of the assumptions of normal distribution and equal variance as two separate conditions, although, in practice (as with this example), variables departing from one assumption often depart from the other.

Some investigators mistakenly believe that they can evaluate the assumption of normal distribution by assessing only the univariate characteristics of the variable. In other words, they print a histogram for all values of *X*. If the distribution is bell-shaped they conclude that it fulfills the assumption of normal distribution. However, as explained above, the assumption of normal distribution and equal variance applies to each level of the independent variable, not all values together. Nonetheless, a simple histogram of all your independent variables is a useful first step.

If you find that the univariate distribution of your variable has a significant departure from the bell-shaped curve, it is likely that it will violate the assumption of normal distribution and equal variance in bivariate analysis. Since it is easier to review a univariate distribution than a bivariate association, this procedure alerts you to which variables to watch especially carefully in your analysis. In addition to eyeballing the histogram, you can use one of many statistical packages to provide you with a normal probability plot, which should approximate a line if the data are normally distributed. The statistics skewness and kurtosis when high also indicate that the data do not fit a normal distribution, but these are less informative than looking at the histogram.

TIP

Run histograms for all your variables. They will alert you to: (a) Potential violations of normality and equal variance, (b) implausible values, (c) gaps in your values, and (d) extreme values.

Besides alerting you to potential violations of the assumptions of normal distribution and equal variance, printing histograms of your dependent and independent variables is a necessary step in cleaning your data. Histograms allow you to detect implausible values (e.g., age of 120 years) and help identify gaps in your values. If, for example, you have few observations of persons older than 60 years of age, your results will not necessarily generalize to this older group. Univariate statistics will also help you to detect extreme (but plausible) values (outliers) that might affect your results, such as two octogenarians. If you happen to have two octogenarians and they happen to have extreme values on your outcome variable, they may unduly influence your results (Section 10.4).

TIP

If your sample size is large (greater than 100), you can assume that the assumption of normality is met.

Now that you have read all of the above theory and considered what a pain it would be to perform the recommended analysis for each of your independent variables, I am happy to tell you that if your sample size is large (greater than 100), you can assume that the assumption of normal distribution is met (assuming you

do not have any unduly influential points). We have the central limit theorem to thank for this great gift,[9] the details of which are beyond the scope of this book.

Only significant departures from equal variance are likely to affect your results. The usual effect would be to decrease the power of your analysis to demonstrate an association between your independent and dependent variables.

With logistic regression, the dependent variable is assumed to have a binomial distribution. A binomial distribution describes the number of successes or failures (e.g., yes or no; survival or death) in a series of *independent* trials. Independent trials mean that the outcome for one subject is independent of the outcome for another subject. In other words, the outcomes are not clustered owing to being from the same individuals, families, or medical practices. There are methods for dealing with clustered effects (Chapter 12). The variances in logistic regression are assumed to depend only on the mean.

With proportional hazards analysis the distribution of the outcome and the variance is unspecified (i.e., no assumption is made).

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?

TIP

If you find significant departures from the assumptions of normality and equal variance try transforming the independent or dependent variable.

If you find *significant* departures from the assumptions of normal distribution and equal variance what should you do? You can attempt to transform either the independent or dependent variable so that the relationship fits these assumptions. Commonly performed transformations are the natural logarithm, the square root, the reciprocal, the square, and the arcsine.[10]

Once you have transformed the variable you need to repeat the bivariate relationship to see if indeed the variable more closely fits the assumptions of your model.

Our ultimate concern will be whether these assumptions are met on a multivariable level; in other words, when all independent variables are in the model. Unfortunately, it is harder to illustrate these principles with multiple independent variables. Nonetheless, there are tests that can be performed to assess these assumptions on the multivariable level. They are dealt with in Section 10.2.

[9] For more on the central limit theorem see: Rosner, B. *Fundamentals of Biostatistics* (5th edn). Pacific Grove: Doxbury, 2000, pp. 174–6.

[10] For a fuller discussion of these transformations see Kleinbaum, D. G., Kupper, L. L., and Muller, K. E. *Applied Regression Analysis and Other Multivariable Methods* (2nd edn). Boston: PWS-Kent, 1988, pp. 220–1.

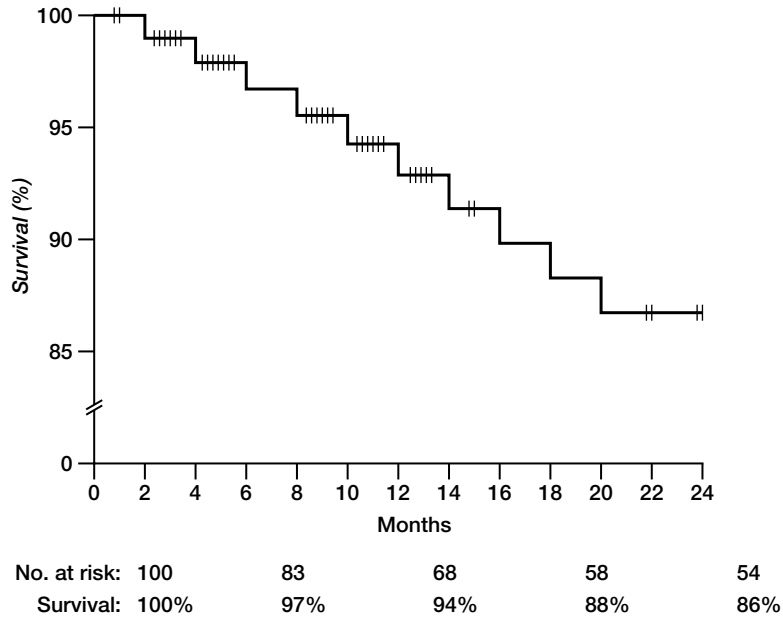


Figure 5.8 Hypothetical survival experience of 100 subjects.

5.9 What are the assumptions of censoring?

Censoring assumes that if subjects could be followed beyond the censor date, they would have the same rate of outcome as those not censored at that time.

Censoring is a technique for incorporating differing lengths of subject follow-up from a longitudinal study (Section 3.6). It assumes that if subjects could be followed beyond the point in time when they are censored, they would have the same rate of outcome as those not censored at that time. Another way of saying this is that the censoring occurs randomly, independent of outcome.

To understand the assumptions of censoring, let's look at the Kaplan–Meier survival graph of 100 persons shown in Figure 5.8. The tick marks on the survival function show where persons are censored – that is, where they leave the analysis. Under the *x-axis* of Figure 5.8, I have shown the number of persons who are at risk for outcome (i.e., have not yet experienced the outcome and have not been censored) and the percent survival.

At time 0, everyone is alive and we have 100% survival. As time passes, people die and percent survival decreases. At two years, survival is 86%. Does this mean that 14 participants died and 86 are still alive? No. In fact, there were only 10 deaths.

When you have censoring, the probability of surviving to the end of the follow-up period is not simply the proportion of the original sample known to be alive at the end of the study. The censored subjects contribute information only until the time that they leave the analysis. To account for this, we compute a current

event rate based on the number of subjects alive and not censored at each point that an event occurs. These current event rates can then be used to compute cumulative survival at the end of the study period (in this case two years). Here's how.

Survival to the end is equivalent to surviving each moment in the entire period. We can write the probability of surviving each moment as a product: the product of surviving the first moment times the conditional probability of surviving the second, given that you survived the first, times the conditional probability of surviving the third, given that you survived the first two, and so on, through the last moment. In turn, we can estimate each of these conditional probabilities as one minus the current event rate.

The cumulative survival at a particular time is simply the conditional probability for that time multiplied by the conditional probability for the prior time at which an event occurred. Because the conditional probabilities are multiplied, this method is sometimes referred to as the product-limit method. For all moments when no event occurs (such as at three months in Figure 5.8 and Table 5.3 when there are six censored subjects but no outcome events), the conditional probability is one, and so it doesn't change the product. These calculations are illustrated in Table 5.3.

Note that censored observations contribute to the analysis until they leave the study, with the provision that they must be in the study at the time that at least one outcome event occurs. Looked at from a different perspective, if observations are censored before any events occur, as is the case with the two censored observations that occur in this example at one month, the censored observations would not contribute to the analysis at all (because to be included in the denominator of the current event rate they must still be in the study when the outcome occurs).

What does survival analysis assume about those persons who are censored? It assumes that their rate of outcome is no higher or lower than subjects who stay in the analysis at that point. So if all subjects could be followed to two years in my hypothetical example, how many would have died? The answer is 14. Why? Because, if no one were censored, then at two years we would have the full sample size of 100. To yield a survival rate of 86%, 14 persons would have died.

I'm sure you can appreciate that this is not an insignificant assumption you are making about censored observations. After all, you cannot prove that censored observations have the same rate of outcome as those that are uncensored. If you actually knew the time to outcome for the censored observations they wouldn't need to be censored! What you as an investigator must address is the likelihood that the censoring assumption is valid based on your understanding of why people have been censored (Section 5.10).

Table 5.3 Calculation of cumulative survival.

Study time (months)	No. of subjects at risk for outcome	No. of outcomes	No. censored	Current event-rate (No. outcomes/No. subjects at risk)	Conditional probability (1– current event rate)	Cumulative survival
1	100	0	2	0/100 = 0	1 – 0 = 1	1
2	98	1	0	1/98 = 0.01	1 – 0.01 = 0.99	(1)(0.99) = 0.99
3	97	0	6	0/97 = 0	1 – 0 = 1	(0.99)(1) = 0.99
4	91	1	0	1/91 = 0.01	1 – 0.01 = 0.99	(0.99)(0.99) = 0.98
5	90	0	7	0/90 = 0	1 – 0 = 1	(0.98)(1) = 0.98
6	83	1	0	1/83 = 0.01	1 – 0.01 = 0.99	(0.98)(0.99) = 0.97
8	82	1	0	1/82 = 0.01	1 – 0.01 = 0.99	(0.97)(0.99) = 0.96
9	81	0	6	0/81 = 0	1 – 0 = 1	(0.96)(1) = 0.96
10	75	1	0	1/75 = 0.01	1 – 0.01 = 0.99	(0.96)(0.99) = 0.95
11	74	0	6	0/74 = 0	1 – 0 = 1	(0.95)(1) = 0.95
12	68	1	0	1/68 = 0.01	1 – 0.01 = 0.99	(0.95)(0.99) = 0.94
13	67	0	5	0/67 = 0	1 – 0 = 1	(0.94)(1) = 0.94
14	62	1	0	1/62 = 0.02	1 – 0.02 = 0.98	(0.94)(0.98) = 0.92
15	61	0	2	0/61 = 0	1 – 0 = 1	(0.92)(1) = 0.92
16	59	1	0	1/59 = 0.02	1 – 0.02 = 0.98	(0.92)(0.98) = 0.90
18	58	1	0	1/58 = 0.02	1 – 0.02 = 0.98	(0.90)(0.98) = 0.88
20	57	1	0	1/57 = 0.02	1 – 0.02 = 0.98	(0.88)(0.98) = 0.86
22	56	0	2	0/56 = 0	1 – 0 = 1	(0.86)(1) = 0.86
24	54	0	2	0/54 = 0	1 – 0 = 1	(0.86)(1) = 0.86

Although I have put tick marks on the survival curve and put the number of subjects at risk for outcome under the graph, most published studies will show you one or the other. The number of persons at risk decreases as persons are censored or experience the outcome.

Sometimes published survival curves do not show either ticks for censored observations or the number of subjects at risk as the study progresses. This is a serious limitation because without this information you cannot assess how much the sample is shrinking as the study progresses. As the sample size shrinks, the data no longer represent the survival experience of the full sample. One tip-off that the sample size is shrinking is large "steps" at the end of the curve. As the sample size decreases, the steps of the curve become larger, because an outcome makes a larger difference in the proportion surviving. This is illustrated in Figure 5.8 and Table 5.3. (Note in the table that a single outcome occurring at the end of the study has a larger current event rate than an outcome occurring earlier on in the study.)

5.10 How likely is it that the censoring assumption is valid in my study?

The likelihood that the censoring assumption is valid depends a great deal on the reason for the censored observations.

5.10.A Loss to follow-up

Subjects may be lost to follow-up because they have grown weary of participating in a study (especially if it involves frequent office visits and blood draws). For this reason, many longitudinal studies go to great effort to keep subjects involved, including providing stipends for each visit with a bonus for completing the study, newsletters to keep subjects informed about the progress of the study, and socials to enhance the cohesion of participants.

In the case of retrospective medical record review studies, subjects will be lost to follow-up and you will not know why. You will know only that after a certain period of time there are no further entries in the chart.

Of all reasons for censoring, losses to follow-up are the most problematic for meeting the censoring assumptions.

Of all reasons for censoring, losses to follow-up are the most problematic for meeting the censoring assumptions. Since the participants are lost, you are unlikely to know what has happened to them after leaving the study. For this reason, it may be problematic to assume that the rate of outcome for the censored observations is the same as that for the uncensored subjects. Also, several studies have found that persons who drop out of studies are different from those who remain in the study. For example, a randomized controlled trial of zidovudine in HIV-infected persons found that persons who were lost to follow-up were more likely to have deteriorating immune function during the trial than those who remained in the trial.[11] (They probably left the trial because they knew they were not doing well and wanted to seek other treatment.)

5.10.B Alternative outcome

Some of your subjects may need to be censored because they experience an outcome that precludes the outcome of interest in your study. This is often referred to as competing risks. For example, consider a study that randomized persons with atrial fibrillation to warfarin or to the standard treatment at that time (aspirin or no treatment at all).[12] Stroke was the outcome of interest. The

[11] Volberding, P. A., Lagakos, S. W., Koch, M. A., *et al.* "Zidovudine in asymptomatic human immunodeficiency virus infection: A controlled trial in persons with fewer than 500 CD4-positive cells per cubic millimeter." *N. Engl. J. Med.* **322** (1990): 941–9.

[12] The Boston Area Anticoagulation Trial for Atrial Fibrillation Investigators. "The effect of low-dose warfarin on the risk of stroke in patients with nonrheumatic atrial fibrillation." *N. Engl. J. Med.* **323** (1990): 1505–11.

investigators therefore censored persons who died from causes other than stroke (e.g., cardiac events, cancer).

Although it is common to censor persons who have an alternative outcome that precludes the outcome of interest, most studies with a main outcome other than death will also report the results for death as an outcome. There are several reasons for this. Even if warfarin prevents stroke in patients with atrial fibrillation as shown in this study, if subjects treated with warfarin do not also live longer, the therapy may not be as strongly recommended. Your view of warfarin as a treatment certainly would be swayed by knowing that the rate of stroke was lower but the rate of death was higher. This could happen if the treatment was effective against stroke but had a life-threatening side effect.

Another reason to include death as an outcome is that there may be some question as to whether an alternative outcome truly excludes the outcome of interest. Even events that appear unrelated to the outcome of interest may be related. In the context of the atrial fibrillation study, consider someone who died in a car accident. Could you be certain that they did not have a stroke while behind the wheel?

TIP

To minimize bias in determining whether an outcome is truly unrelated to the outcome of interest, a study should have objective a-priori criteria as to what constitutes an alternative outcome.

To minimize bias in determining whether an outcome is truly unrelated to the outcome of interest, a study should have objective a-priori criteria as to what constitutes an alternative outcome. In addition, judgments on individual cases should be made by a review committee that is masked to the treatment assignment. This is exactly what was done in the study of warfarin in patients with atrial fibrillation.

However, even with the best criteria and objective reviewers, you may mistakenly assume that some outcomes are unrelated to your outcome of interest when, perhaps, they are related. In the atrial fibrillation study there were 37 deaths; only one was caused by stroke. The warfarin group had a lower rate of death overall, a lower rate of death owing to cardiac causes, and a lower rate of death owing to noncardiac causes. Since warfarin was not anticipated to have any effect on mortality other than decreasing deaths caused by stroke, these findings suggest two possibilities. Warfarin may prevent death owing to other causes, or some of the deaths attributed to other causes may have actually been caused by unrecognized strokes. The possibility of unrecognized strokes does not weaken the findings. If there were missed strokes in the nonwarfarin group it would only strengthen the treatment effect of warfarin. But for the purposes of this discussion, I think this is a good example of how hard it is to correctly categorize nonoutcome-related events.

When death is a common alternative outcome, competing mortality may bias your estimates of time to outcome.

When death is a common alternative outcome, as it is in any long-term study of elderly or very ill people, competing mortality may bias your estimates of time to outcome. This will occur if the likelihood of outcome would have been higher

in those persons who died had they not died. The bias of competing risks can be avoided by not using outcomes other than death (since no outcome precludes death). But in many cases we are interested in these other outcomes. A reasonable compromise is to report both (as was done in the atrial fibrillation trial). More sophisticated methods for dealing with competing risks are beyond the scope of this book.[13]

5.10.C Withdrawal

Persons may withdraw from a study because they do not think the treatment is helping, because they want to receive a treatment and believe they were randomized to placebo, because they have a side effect, or because they find the protocol too demanding of their time. Subjects are usually withdrawn by the investigators for safety reasons (a side effect that makes it dangerous for the subject to continue treatment).

Less commonly, investigators may withdraw a subject because they develop a condition that precludes them from participating in one of the arms of the study. For example, in the study of warfarin in atrial fibrillation, the investigators had to withdraw a participant who required valvular replacement during the study. Why? Although valvular replacement does not preclude the development of stroke nor is it a side effect of treatment (or nontreatment), a patient receiving valvular replacement requires anticoagulation with warfarin. Since treatment with warfarin was mandatory, the person could not be continued in a study comparing warfarin to standard treatment.

At first glance, subjects who voluntarily withdraw or are withdrawn by the investigators may seem like losses to follow-up, but there is an important difference. The difference is that withdrawn participants may be willing to let you passively follow them for outcome, even though they do not want to actively participate in your protocol. (This may not be true if they have a side effect and blame you for it!) If the withdrawn subjects will allow you to follow them for outcome, you do not have to censor them prior to the end of the study. Instead you can perform an intention-to-treat analysis.

Intention-to-treat means that participants are counted as members of their originally assigned group, no matter what happens during the study period. For example, in a study with a treatment and a placebo arm, if you perform an intention-to-treat analysis, persons who were assigned to treatment would be part of the treatment group even if they stopped taking the treatment during

[13] Pepe, M. S., Longton, G., Pettinger, M., *et al.* "Summarizing data on survival, relapse, and chronic graft-versus-host disease after bone marrow transplantation: Motivation for and description of new methods." *Br. J. Haematol.* **83** (1993): 602–7.

DEFINITION

Intention-to-treat means that participants are counted as members of their originally assigned group, no matter what happens during the study period.

the study. Intention-to-treat analysis protects against treatments appearing to be more favorable than they are. If subjects with side effects are less likely to benefit from treatment than those without side effects, removing them from the analysis at the time of their side effect will bias your results. It will make your treatment appear more effective than it is. For example, if 100 people are given a new treatment that is effective but causes nausea so severe that only 50 percent of subjects are able to tolerate the treatment long enough to benefit from it, removing 50 study participants will make the treatment look more effective than it is (because a very high proportion of those subjects left in the trial will benefit from the treatment).

The downside of intention-to-treat analysis is that it dilutes the effect of treatment. Keeping to the same example, if you keep 50 subjects who did not take the full treatment in your treatment arm, the treatment arm will be similar to the nontreatment arm. You might ask: Isn't this justified? What good is a treatment if patients can not tolerate it? But remember there is a difference between a sample and an individual patient. If I had a terrible disease, I might be interested in trying an efficacious medicine that caused unrelenting vomiting for 50 percent of persons. After all, if I am in the half of persons who doesn't develop vomiting, why should I forgo an efficacious treatment? If the authors only report the results of the intention-to-treat analysis you would not know how efficacious the treatment is for the subgroup of people who can tolerate it. Limiting your analysis to persons who tolerate or comply with treatment is sometimes referred to as an efficacy analysis (how efficacious is the treatment for people who take it?) whereas intention-to-treat analysis is referred to as an effectiveness analysis (how effective is the treatment in the real world?).

TIP

Intention-to-treat is a more conservative approach for estimating the efficacy of treatment than censoring subjects when they withdraw from one of the treatment arms.

If the number of subjects who stop treatment or withdraw from your study is small, it won't matter whether you perform intention-to-treat analysis or censor them at the time they leave the study. Since intention-to-treat is a more conservative approach for estimating the efficacy of treatment than censoring subjects when they withdraw from a treatment arm, most researchers prefer it. In studies where a large number of subjects are withdrawn, you may want to report the analysis both ways.

5.10.D Varying time of enrollment

Varying times of enrollment (starting times) is an important issue for large prospective studies and for studies of rare diseases. For studies enrolling thousands of participants, logistic constraints preclude everyone from starting the study on the same day. Most large studies are conducted in multiple centers

and it is rare for all sites to begin enrollment at the same time. Similarly, with studies of rare diseases it may take several years to identify and enroll enough persons.

In observational studies, varying times of enrollment is the rule, rather than the exception. For example, in the Aerobics Center Longitudinal Study discussed in Section 1.1, the investigators enrolled participants over a 19-year period. Indeed, their study is an open cohort with ongoing accrual of subjects. Subjects are enrolled when they complete the baseline medical examination and are then followed prospectively. Similarly, the study of survival with melanoma, discussed in Section 2.1.D, followed patients diagnosed over a nine-year period.

Theoretically, varying start times could be dealt with by following all subjects for a fixed time period (e.g., three years) regardless of when they started the study. In this instance, no one will be censored prior to the end of the study owing to varying lengths of enrollment. However, this method would decrease the power of the analysis because you would lose the additional follow-up time supplied by the persons who began the study early on. In most studies, the greatest cost (in time and funds) is the initial enrollment and evaluation. The cost of continuing follow-up for the outcome is usually minimal, while the gain, in terms of follow-up time, can be great. Also, waiting until the last enrolled participant completes a preset follow-up period will extend the length of the study often beyond the time that the analytic staff is being supported. Finally, we are all impatient to learn the results of our work. Thus, most investigators will censor results at a common point in calendar time. At this point in time, length of follow-up will differ for the participants. If the longest follow-up is three years, all participants who have follow-up less than three years and have not experienced an outcome will be censored at the amount of time of follow-up, which in all cases will be less than three years. Those with three years of follow-up and no outcome will be censored at three years.

5.10.E End of study

TIP

Studies that have a large number of censored observations prior to the end of the study are more problematic than studies that have just a few censored observations.

It may have surprised you that a subject who completes a study (or who in an observational trial has the longest follow-up) without experiencing the outcome is still censored. This is counter intuitive because in common usage we think of censored subjects as those who do not complete the study. Although all subjects who do not experience the study outcome are ultimately censored, there is an important difference between those censored at the end of a trial and those censored for other reasons. The difference is that for those subjects censored at the end of the study, no assumptions need be made about

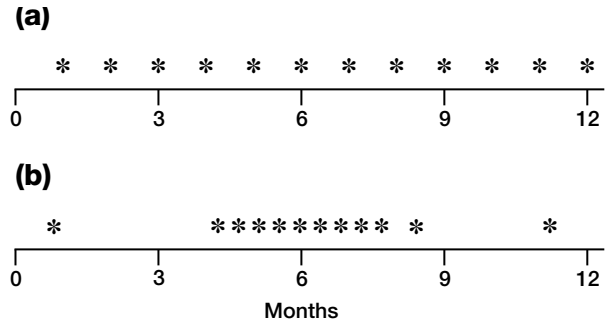


Figure 5.9 Different patterns of censoring. In Figure 5.9(a) censored observations have occurred evenly over the study period whereas in Figure 5.9(b) censored observations are clumped at 6 months.

the future – the study is over. Usually, in the published report, the authors will tell you the date that subjects who completed the study without experiencing an outcome were censored. It is the last date of the study or the last day of observation.

5.11 How can I test the validity of the censoring assumption for my data?

There is no ideal test for assessing the validity of the censoring assumption. It is primarily a judgment call by the investigators, reviewers, and readers of the data. That's why in Section 5.10 I took you through a long discussion of the reasons for censoring and how censored observations may or may not fulfill the assumptions of censoring. Nonetheless, it is possible to make a general assessment about censored observations. First, and foremost, studies that have a large number of censored observations prior to the end of the study are more problematic than studies that have just a few censored observations.

TIP

Censored observations occurring evenly throughout the study are consistent with the censoring assumption, whereas clumps of censored observations suggest nonrandom censoring.

Graphical methods, showing when censoring occurred during the trial, can be used to test the validity of the censoring assumption. Figures 5.9(a) and 5.9(b) show two very different patterns of censoring. In Figure 5.9(a) it would appear that subjects (shown with asterisks) dropped out evenly through the course of the study. In Figure 5.9(b) it would appear that a clump of subjects dropped out around six months. Whereas the latter suggests some event that occurred at around six months, the former suggests the kind of censoring that you would expect if people dropped out for random reasons such as moving out of town, other obligations, etc.

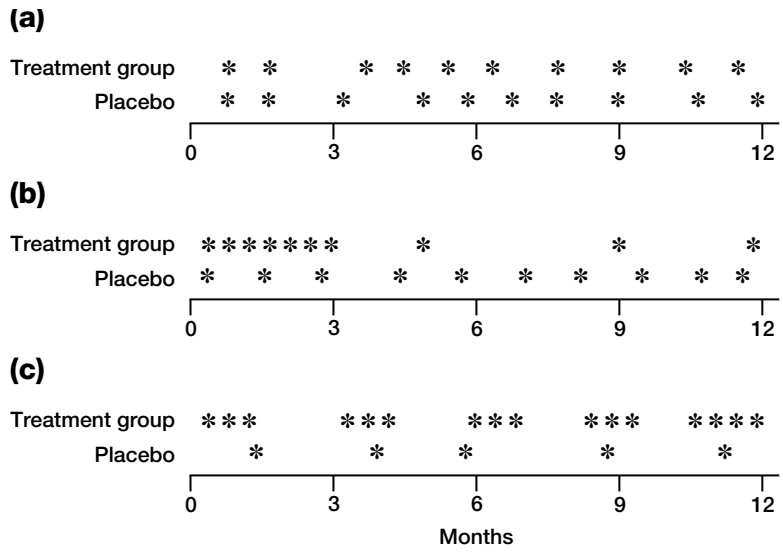


Figure 5.10 Different patterns of censoring between a treatment and a placebo group. In Figure 5.10(a) the patterns of censoring between the two groups are similar. In Figure 5.10(b) there are an equal number of censored observations in the two groups, but the patterns are different. In Figure 5.10(c), there are more censored observations in the treatment group than in the placebo group.

Graphical displays can also be used to compare censoring in two or more different arms of a study. Figures 5.10(a)–(c) show a treatment and a placebo group. Note how in Figure 5.10(a) the number of persons censored between the treatment and placebo groups is the same and the pattern of censoring between the two groups is also similar. In Figure 5.10(b), the number of censored persons is the same but the pattern between the two groups is different. In Figure 5.10(c), there are many more censored observations in the treatment group. Whereas Figure 5.10(a) is consistent with random censoring, Figures 5.10(b) and 5.10(c) suggest that the causes of censored observations are different for the treatment and placebo group.

TIP

Similar patterns of censored observations between treatment groups are consistent with the censoring assumption, whereas different patterns suggest nonrandom censoring.

A useful method for assessing the validity of the censoring assumption is to compare subjects lost to follow-up to those not lost to follow-up. This can be based on baseline characteristics. So, for example, you can examine whether there are differences by age, race, etc. of persons lost to follow-up compared to persons not lost to follow-up. In particular, you might wish to explore whether subjects at high risk of outcome are differentially lost to follow-up. The censoring assumption can also be tested using time-dependent covariates (Section 10.11).

To be certain that censoring owing to alternative outcomes is not affecting your results, report rate of death where possible, in addition to whatever more proximal outcome you are studying. Since there is no alternative outcome to death (no alternative outcome can exclude death), these curves reassure the reader.

For withdrawn cases, it is best to test the censoring assumption by conducting intention-to-treat analyses as described above. Since you can continue to follow withdrawn patients for outcome, you can test whether leaving them in or taking them out makes a difference.

Censoring due to varying starting times is usually assumed to fulfill the censoring assumptions, since subjects who enroll at the start of a study should be the same as subjects who enroll towards the end. I say "should be," because sometimes investigators become more flexible about the enrollment criteria as studies progress, especially if enrollment is running slowly. Changing the enrollment criteria after a study has begun should be avoided.

But even assuming a group of investigators rigidly used the same enrollment criteria over the course of a long study, one should compare subjects enrolled in the early years of a study to those enrolled in the later years. The reason is that subjects enrolled in later years may be more likely to have received technical advances that were not available in the earlier years of the study. If they were, and these advances could affect development of the outcome of interest, then the assumptions of censoring are not valid.

It is also possible to determine how sensitive your results are to different possibilities of what happened to censored observations. For example, Vittinghoff and colleagues evaluated the impact of multiple risk factors on the occurrence of coronary heart disease among women.[14] Of the sample of 2763 women, 60 were lost to follow-up. The investigators considered two extremes for what could have happened to those women: 1) they had a heart disease event on the day they were censored; 2) their censored date was equal to the longest observed follow-up time. Rerunning their proportional hazards analysis both ways, they found no major changes in the importance/nonimportance of most of the risk factors, with two exceptions: there were changes on the impact of HDL-cholesterol level and smoking on coronary heart disease events.

In conclusion, censoring is a very helpful tool for clinical research. It prevents you from having to delete valuable cases. A randomized controlled trial comparing fluoride to placebo for the prevention of fractures among women with

[14] Vittinghoff, E., Shlipak, M. G., Varosy, P. D, *et al.* "Risk factors and secondary prevention in women with heart disease: The heart and estrogen/progestin replacement study." *Ann. Intern. Med.* **138** (2003): 81–9.

osteoporosis greatly benefited from censoring.[15] Only 135 (67 percent) of the 202 enrolled women were able to complete four years of treatment. If the investigators had deleted these cases they would have lost a third of their sample size. As with any powerful statistical tool, however, censoring should be used carefully. Ask yourself (and tell your reader) the circumstances of persons censored. When you know the outcome of censored cases, perform intention-to-treat analyses.

[15] Riggs, B. L., Hodgson, S. F., O'Fallon, W. M., *et al.* "Effect of fluoride treatment on the fracture rate in postmenopausal women with osteoporosis." *N. Engl. J. Med.* **322** (1990): 802–9.

6

Relationship of independent variables to one another

6.1 Does it matter if my independent variables are related to each other?

As discussed in Section 1.1, the strength of multivariable analysis is its ability to determine how multiple independent variables, which are related to one another, are related to an outcome. We would not need multivariable analysis to determine the independent effect of exercise on mortality if it weren't for the fact that exercise, smoking, age, hypertension, and cholesterol level were all related to each other and the outcome of interest. Multivariable analysis helps us to separate the effects of these different variables on outcomes such as mortality.

DEFINITION

Multicollinearity occurs when two or more variables are so closely related to one another that the model may not be able to reliably assess the independent contribution of each variable.

However, if two variables are so closely correlated that if you know the value for one variable you know the value of the other, multivariable analysis cannot separately assess the impact of these two variables on the outcome. This problem is called multicollinearity. I can best illustrate it with an extreme example.

Let's say you were studying factors that affected length of hospital stay among patients with pneumonia. At your hospital, to accommodate the different preferences of the staff, the nurses record patients' temperature in both Fahrenheit and Celsius. When you do your medical abstraction, you record both Fahrenheit and Celsius temperatures. If you entered both variables in a model assessing length of stay, your model would be incorrect, and you would get an error message or unpredictable answers. This is because temperature in Celsius and temperature in Fahrenheit is the same variable even though the numbers are different. There is a simple mathematical conversion from one to the other: Celsius = (Fahrenheit – 32) × 0.56.

Your model cannot possibly assess the independent contribution of temperature in Fahrenheit and in Celsius because they are really the same variable. However, unless you make a mistake and include two variables that really are the same variable (such as temperature in Fahrenheit and in Celsius) it would be very unlikely to have a situation where two variables are exactly correlated with one another. A more likely scenario is to have variables that are not sufficiently

different for the model to distinguish them. For example, Phibbs and colleagues found that birth weight and gestational age were too closely related to include both in their analysis of neonatal mortality.[1]

6.2 How do I assess whether my variables are multicollinear?

DEFINITION

The *correlation coefficient* is a bivariate statistic that measures how strongly two variables are related to one another.

The correlation coefficient (also called Pearson correlation coefficient or r) is a bivariate statistic that measures how strongly two variables are related to one another. The correlation coefficient assumes the relationship between the two variables is linear. It can range from -1 to 1. When the coefficient is -1 or 1 the two variables change together exactly (i.e., knowing one variable tells you the value of the other variable). The only difference between -1 and 1 is that the negative sign indicates that the two variables change exactly together in opposite directions (i.e., as one goes higher the other goes lower). Zero indicates that there is no relationship whatsoever. If you square the correlation coefficient and multiply by 100 you get a measure of how much information the two variables share ranging from 0 percent to 100 percent.

The correlation between temperature in Fahrenheit and in Celsius is 1.0. The two variables share 100 percent of the same information. In contrast, the correlation between vitamin B_{12} level and pneumococcal antibody titer following immunization was found to be 0.55 (Figure 5.1). The two variables share 30 percent ($0.55^2 \times 100 = 30$) of the same information.

TIP

Two variables correlated at > 0.8 may cause multicollinearity problems in your analysis.

To determine how correlated your independent variables are you may run a correlation coefficient matrix with all your proposed independent variables. In general, two variables that are correlated at more than 0.9 will pose problems in your analysis. Variables correlated at less than 0.8 will not pose problems. Variables correlated between 0.8 and 0.9 may cause problems.

The problem with a correlation matrix is that it assesses only the relationship between two variables, without adjustment for the other variables. For this reason, most multivariable analysis programs will print out a correlation matrix for the parameter estimates. Since these estimates are adjusted for one another they are a better measure of whether two variables will result in problems of multicollinearity. As with simple correlations, values greater than 0.9 will cause problems, whereas those between 0.8 and 0.9 are in the gray area.

Astute readers will note that these two techniques deal only with the simplest case of multicollinearity, when one independent variable is highly related to a second independent variable. What if a combination of independent variables is

[1] Phibbs, C. S., Bronstein, J. M., Buston, W., *et al.* "The effects of patient volume and level of care at the hospital of birth on neonatal mortality." *JAMA* **276** (1996): 1054–9.

highly related to another independent variable? Statistically, this is as problematic as the situation where two variables are highly correlated, but as I am sure you will appreciate, it is harder to diagnose.

In a sense we have already dealt with this concept in Section 4.2 on converting nominal variables into multiple dichotomous variables ("dummy variables"). Look back at Table 4.2. You will recall that I said you did not need to create a variable for white ethnicity because if a subject were 0 (no) on the variables for African-American, Latino, Asian/Pacific Islander, Native American, and other nonwhite ethnicity the subject would be of white ethnicity. What if you didn't read this section and entered into your model a yes/no variable for each ethnicity including white? Then you would have a situation where a combination of independent variables completely accounts for the value of a different independent variable. Prove this to yourself, by correctly answering the following questions:

1. A subject is "yes" on any one of the five variables: African-American, Latino, Asian/Pacific Islander, Native American, or other nonwhite ethnicity. What is the subject's value on the white ethnicity variable?
2. A different subject is "no" on all five of the variables: African-American, Latino, Asian/Pacific Islander, Native American, or other nonwhite ethnicity. What is the subject's value on the white ethnicity variable?

You knew that the answer to the first question was **No** and the answer to the second question was **Yes**, even though I didn't tell you anything about the subject's value on the variable of white ethnicity. This is a situation where a combination of variables completely determines the value of another variable. If you entered a variable for white ethnicity, in addition to the others, this would result in spurious results in your multivariable model.

How will you know if a combination of variables accounts for another variable's value? Two related measures of multicollinearity are tolerance and the reciprocal of tolerance: the variance inflation factor. Both measure how much the regression coefficient for a particular variable is determined by the other independent variables in the model. Small tolerance values, including those below 0.25, are worrisome, and those below 0.10 are serious. As you would expect with a reciprocal value, high values of the variance inflation factor, such as values greater than four, are problematic, whereas values greater than ten are serious.[2]

Most linear regression programs will print out tolerance or variable inflation factors for each of the independent variables in your model. If the values of some of your variables are worrisome, you will need to do additional analyses to determine which of the other variables in the model are closely related with

[2] For more on measures of multicollinearity and when to worry, see Glantz, S. A. and Slinker, B. K. *Primer of Applied Regression and Analysis of Variance*. New York, NY: McGraw-Hill, 1990, pp. 181–99.

the problematic variable. This can be done by performing regression analyses using the other variables as independent variables to estimate the value of the problematic variable. This will show you which variables are highly related and enable you to decide which variables to keep in the analysis.

With multiple logistic regression and proportional hazards analysis researchers usually rely on the correlation matrix of the multivariable parameter estimates to determine if there are serious problems with multicollinearity. However, tolerance is a standard for allowing variables to enter into the model; variables with very low tolerance values will not enter (Section 8.8).

6.3 What should I do with multicollinear variables?

If you have variables that are highly related, consider your options:

- omit the variable
- use an "and/or" clause, or
- create a scale

TIP

If you need to omit a variable, omit the one that has more missing data, has more measurement error, or is less satisfactory in some other way.

If you are going to omit one of the variables, how do you decide which one to delete? Omit the one that is theoretically less important, has more missing data, has more measurement error, or is less satisfactory in some other way. In the study of neonatal mortality referred to above, the investigators kept birth weight and excluded gestational age. They excluded gestational age because there were more missing cases on this variable than on birth weight and it was less reliably coded. Ironically, as the authors point out, gestational age is theoretically the more important factor in accounting for mortality (because age, not weight, is the deciding factor in reaching certain fetal developmental milestones). Therefore they included two additional variables in their analysis (small for gestational age [yes/no] and large for gestational age [yes/no]) to adjust for the fact that weight might not be an accurate measure of gestational age in some cases.

Using "and/or" clauses works well for correlated variables that represent the same process. For example, if you asked patients with pneumonia whether they had diaphoresis (sweats) or rigors (shaking), these two variables would be expected to be closely correlated since rigors are a more extreme form of diaphoresis. However, some patients who had rigors may not have noticed that they were first diaphoretic; some who were diaphoretic may have taken aspirin thereby preventing the rigor. The new variable could be diaphoresis and/or rigor. Patients who had one or both would be reported as "yes" on this variable; those who had neither would be reported as "no."

Creating multi-item scales is a strategy often pursued with psychological and sociologic data. In creating scales, the values of multiple variables for each subject are summated or averaged to form a single variable that summarizes the

meaning of the separate variables (Section 7.5.C.3). Researchers may intentionally ask subjects multiple related questions to test the reliability of the subject's responses (i.e., when asked a similar question, using different wording, will subjects answer in the same way?). In this case, researchers usually plan ahead of time which questions will form a scale. Other times researchers will use factor analysis (Section 7.5.C.4) to determine which questions provide similar information.

These techniques for dealing with multicollinear variables also work when you need to decrease the number of independent variables in your analysis because of insufficient sample size. However, they will not work for decreasing sample size if your variables are not highly related. A variety of methods for decreasing the number of independent variables are detailed in Section 7.5.

7

Setting up a multivariable analysis

7.1 What independent variables should I include in my multivariable model?

On the surface this seems like a simple question. You should include the risk factor(s) of interest and any variables that may potentially confound the relationship between the risk factor and the outcome. However deciding which variables may confound your analysis is not always easy. Variables that are extraneous, redundant, have a lot of missing data, or intervene between your risk factor and outcome should be excluded.

Recommendations on what variables to include and exclude in your model are reviewed in Table 7.1 and discussed in the next two sections.

7.2 How do I decide what confounders to include in my model?

TIP

Include in your model those variables that have been theorized or shown in prior research to be confounders of the relationship you are studying.

Ideally researchers should include all those variables that have been theorized or shown in prior research to be confounders. Depending on the outcome you are studying, there may be a large number of variables that have been shown in prior research to be associated with the risk factor and the outcome. For example, studies of cardiovascular outcomes must include a large number of potential outcomes including age, sex, smoking status, hypertension, diabetes, obesity, LDL-cholesterol, HDL-cholesterol, reactive C-protein, aspirin use, and beta-blocker use because all of these variables have already been shown to affect cardiovascular disease.

In addition to including variables that have been theorized or shown in prior research to be confounders, include in your model those variables that fit the empiric definition of confounders in your data. Specifically, include those independent variables that are associated with the risk factor and the outcome in bivariate analysis, unless you believe that those variables are intervening between the risk factor and the outcome (Section 7.3).

Table 7.1 Variables to include and exclude in multivariable models.

What to include?	What to exclude?
Risk factor(s)	Extraneous variables (variables not on the causal pathway to your outcome)
Potential confounders, based on theory, prior research, empirical findings	Redundant variables
	Variables with a lot of missing data
	Intervening variables (variables that are on the causal pathway between a risk factor and an outcome)

TIP

Include in your models those variables that are associated with the risk factor and the outcome in bivariate analysis.

Unfortunately, there is no single standard of how strong an association should be to result in a variable being included as a potential confounder. In general, you want to err on the side of inclusion. Most investigators will include in their multivariable model any variable that is associated with the outcome at a P value of < 0.20 or 0.25, regardless of whether or not the variable has been shown to be associated with the risk factor. (In using empiric criteria for choosing variables to enter into your multivariable analysis, keep in mind that if you have a suppresser effect, the variable may not be even weakly associated with the outcome in bivariate analysis. Recall the example in Section 1.3 of zidovudine treatment, which affected the likelihood of seroconversion in multivariable analysis even though it was not significantly related to seroconversion in bivariate analysis.)

Remember that even if you use an empiric criterion to decide which variables to enter into your model, you should still include those that are theoretically important or have been confounders in prior research even if they did not meet your criterion. For example, Spencer and colleagues in their study of the effectiveness of statin therapy in patients with acute coronary syndrome (described in Section 2.1.B)) included in their multivariable models variables that were associated with the outcome of interest in bivariate analysis at a P value of < 0.25.[1] They then used a stepwise model (Section 8.7) to exclude variables that were not associated with the outcome at a P value of ≤ 0.05. However, they included age, sex, and history of hyperlipidemia in all final models regardless of their statistical significance because of their clinical relevance to coronary artery syndromes.

7.3 What independent variables should I exclude from my multivariable model?

In building accurate models, what variables you exclude is as important as what variables you include (Table 7.1). Exclude from your multivariable model

[1] Spencer, F. A., Allegrone, J., Goldberg, R. J., *et al.* "Association of statin therapy with outcomes of acute coronary syndromes: The Grace study." *Ann. Intern. Med.* **140** (2004): 857–66.

extraneous variables, i.e., variables that are not on the causal pathway to your outcome. For example, if you are developing a model estimating HIV prevalence, and you are using the large multipurpose National Health and Nutrition Examination Survey (NHANES), exclude from your model seat-belt use, even though it may well be associated with safer-sex practices (because people who are health conscious are likely to use both seat belts and condoms). Because seat-belt use is not on the causal pathway between behavior and HIV infection, including it can only add error to your model (since all measurements have error) and confusion to your readers.

It is also important to exclude redundant variables. The strength of multivariable analysis is that it can determine the unique contribution of related variables to outcome. However, if two variables are too closely related, multivariable analysis cannot accurately separate the impact of the two variables on outcome. This problem is called multicollinearity (Chapter 6) and requires that you enter only one of the highly related variables. For example, in a study of factors associated with adherence to combination antiretroviral medication among HIV-infected persons, the investigators found that ethnicity was collinear with acculturation; they chose to put ethnicity in the model and exclude acculturation.[2]

With duplicative variables include the one that is theoretically more important, has less missing data, or has less measurement error.

When deciding which of two duplicative variables to include, choose the one that is theoretically more important, has less missing data, or has less measurement error.

It is also important to exclude variables with a lot of missing data. Missing data is a much greater problem with multivariable analysis than bivariate analysis. This is because subjects with missing values on any of the variables entered into the model are tossed out of the analysis even if the subject has valid values on the other variables. (With a bivariate analysis, you lose only those subjects with missing values on the two variables used, not those with missing values on the other variables.)

TIP

Drop variables rather than subjects when you have missing data.

Dropped subjects decrease the power of your analysis and, perhaps even more problematically, bias your study because subjects missing on a particular variable may be systematically different than subjects not missing on that variable. Because it is generally better to lose variables than subjects, drop variables that have a lot of missing data.

That being said, sometimes you have a variable of such great importance that it must be included in your analysis even if there are many missing cases. If this is the case acknowledge to your readers that the missing cases may bias your results in ways that are hard to assess. You can provide some reassurance to your readers

[2] Golin, C. E., Liu, H., Hays, R. D., *et al.* "A prospective study of predictors of adherence to combination antiretroviral medication." *J. Gen. Intern. Med.* **17** (2002): 756–65.

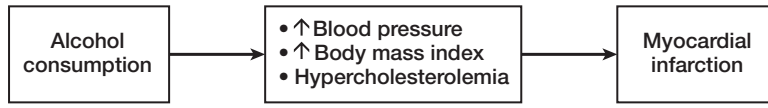


Figure 7.1 HDL is an intervening variable between moderate alcohol consumption and decreased coronary artery disease.

if a comparison of cases with missing data to cases without missing data shows no difference on important characteristics. Alternatively, you may use one of the available methods for assigning values to missing cases, thereby enabling you to use the variable in the analysis (Section 7.6).

DEFINITION

An intervening variable is on the causal pathway to your outcome.

You should also exclude variables that are on the causal pathway to your outcome. Such variables are referred to as intervening variables. It may seem confusing that I recommend excluding intervening variables since I said in Section 7.2 that you should exclude variables *not* on the causal pathway to your outcome. Why am I now telling you to exclude variables that are on the causal pathway?

TIP

Don't adjust for an intervening variable or you may adjust away the effect you are trying to demonstrate.

The reason is that if you statistically adjust for an intervening variable, you may adjust away the very effect you are trying to demonstrate. For example, it is known that moderate alcohol consumption is associated with a lower incidence of coronary artery disease. The mechanism appears to be that moderate alcohol consumption increases the HDL-cholesterol, the "good cholesterol" as shown in Figure 7.1.

If you adjust for HDL level in an analysis of the effect of alcohol consumption on coronary artery disease, it may appear that alcohol has no effect. However, that's not accurate; moderate alcohol use is causally related to coronary heart disease, but the effect is mediated by HDL-cholesterol. This problem is referred to as overadjustment.[3]

TIP

Statistically confounders and intervening variables act the same.

Unfortunately there is no test for whether a variable is a confounder or an intervening variable. Statistically confounders and intervening variables act the same. Therefore the decision on whether to include a variable in a model because you believe it is a confounder, or exclude it because you believe it is an intervening variable, should be made based on prior research and biological plausibility.

One circumstance in which it is appropriate to include an intervening variable is if you are trying to demonstrate that the effect of a risk factor on outcome is mediated by an intervening variable. In such cases, you will want to run the model first without the intervening variable and then a second time with the intervening variable. If it is an intervening variable, the statistical association between the risk

[3] Szklo. M. and Nieto, F. J. *Epidemiology: Beyond the Basics.* Gaithersburg: Aspen Publishers, 2000, p. 333.

factor and the outcome seen in the first model will be eliminated or dampened in the second model. This can be helpful in understanding the mechanism by which a risk factor affects the outcome. (But keep in mind that the association between risk factor and outcome will also diminish with inclusion of a confounding variable.)

7.4 How many subjects do I need to do multivariable analysis?

The sample size needed for multivariable analysis, as with bivariate analysis, depends on the size of the effect you are trying to demonstrate and the variability of the data. It takes a much larger sample size to show that a risk factor is mildly (but statistically) associated with an outcome (e.g., odds ratio of 1.5) than to show that it is strongly associated with an outcome (e.g., odds ratio of 4.0). The reason is that the smaller the sample size, the larger the confidence intervals. The closer the odds ratio is to 1.0 then the more likely wide confidence intervals are to include one. Similarly, although you can never prove the null hypothesis (i.e., no association: odds ratio of 1.0), the larger the sample size the smaller the chance that you have missed an association that was really present. It also takes a larger sample size to demonstrate a difference between groups on a variable that has a great deal of variability (i.e., a large standard deviation).

DEFINITION

A *power* calculation determines the needed sample size to detect a particular effect.

Determining the needed sample size is referred to as a power calculation (the power to detect a result). Power calculations for multivariable analysis are complicated and generally require consultation with a biostatistician. Nonetheless, there are a couple of rules of thumb that will hold you in good stead.

First, determine the sample size needed for a bivariate analysis (i.e., a comparison of two proportions, a comparison of two means, a comparison of two times to outcome without adjustment for other variables). Several free and easy-to-use computer software programs exist to do this.[4] If your power calculation shows that you do not have enough subjects to demonstrate the effect in bivariate analysis, you definitely will not have enough subjects in your multivariable analysis. If it shows that you do have enough subjects, the next question would be whether the sample size is sufficient for your multivariable analysis.

TIP

If your power calculation shows that you do not have enough subjects for demonstrating the effect in bivariate analysis, you definitely will not have enough subjects to demonstrate the effect in multivariable analysis.

[4] Free software packages for doing sample size calculations are available: Statistical Considerations for Clinical Trials and Scientific Experiments by David Schoenfeld (http://hedwig.mgh.harvard.edu/sample_size/quan_measur/defs.html) and Simple Interactive Statistical Analysis (SISA) (http://home.clara.net/sisa/sampshlp.htm). See Katz, M. H. *Study Design and Statistical Analysis: A Practical Guide for Clinicians*. Cambridge University Press, forthcoming, for explanations of the needed ingredients to input into these computer programs. Alternatively, for an easy-to-follow sample-size guide that doesn't require computer software, see Hulley, S. B., Cummings, S. R., Browner, W. S., *et al. Designing Clinical Research* (2nd edn). Baltimore, MD: Lippincott Williams and Wilkins, 2001, pp. 65–91.

The easiest way to perform a multivariable sample-size calculation is to use an available software program. Power and Precision (http://www.power-analysis.com/specifications.htm) provides a free evaluation period and calculates the needed sample size for multiple linear regression and multiple logistic regression.

To compute the needed sample size for a multiple linear regression design using the Power and Precision program you need to specify the number of variables that represent your hypothesis (i.e., the risk factor[s]) and the number of variables that are covariates (i.e., the set of potential confounders). For the set of risk factor(s) and the set of covariates you need to specify the increment of R^2 (Section 9.2.B) you anticipate that each set will explain. You will also state the alpha (usually 0.05 and two-tailed) and the beta (usually 0.80). From this information the program will tell you the needed sample size. The program will also allow you to calculate the needed sample size for more complicated models (e.g., a linear regression model with an interaction term, a linear regression model with a risk factor, the square of the value of the risk factor, and a cubic transformation of the value of the risk factor).

To compute sample size for logistic regression with a dichotomous independent variable, you will need to specify the relative proportion in each group (e.g., 1:1 if equal numbers of subjects in each group; 2:1 if twice the number in one group as the other). A convenient feature of Power and Precision is that it will allow you to enter a categorical variable (e.g., ethnicity in four categories) and it will automatically create multiple dichotomous variables with a reference category, so that you can determine the needed sample size for a logistic regression model with a categorical independent variable.

To compute sample size for multiple logistic regression with an interval-independent variable you will need to specify the mean and standard deviation of the independent variable, the event rate at the mean (the event rate at the mean is the same for all variables), and the event rate at a point other than the mean for the independent variable. The program Power and Precision can accomodate two interval independent variables. If you include two you will need to specify the correlation between them (with the correct sign).

Although not an alternative to formally calculating the needed sample size, a useful rule-of-thumb for planning multiple logistic regression and proportional hazards analysis is that for every independent variable in your model you need at least ten outcomes.[5] The reason I say ten outcomes for each independent

[5] Peduzzi, P., Concato, J., Kemper, E., *et al.* "A simulation study of the number of events per variable in logistic regression analysis." *J. Clin. Epidemiol.* **49** (1996): 1373–9. Peduzzi, P., Concato, J., Feinstein, A. R., *et al.* "Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates." *J. Clin. Epidemiol.* **48** (1995): 1503–10. Harrell, F. E.,

TIP

For multiple logistic regression and proportional hazards analysis you should have at least ten outcomes for each independent variable in your model.

variable, rather than 20 subjects for each independent variable (which would be equivalent if your outcome occurred in half your subjects), is that your model is assessing a particular outcome. In most medical studies, less than half of the subjects experience the outcome (e.g., develop cancer or have a heart attack). If your outcome occurs in only five subjects in your study you may not have enough power to answer your research question even if you have a thousand other subjects in whom the outcome does not occur. This is because the model is determining the likelihood of outcome based on those five people. It does not help you to determine the likelihood of no outcome if this is the larger group. The needed sample size is based on the smaller of the two groups. The reason is that outcome and not outcome are mathematically equivalent: the likelihood of not outcome is simply 1 – (the likelihood of outcome).

The ten outcomes per variable is just a guideline. Just because your sample size does not meet this criterion does not mean that your study won't appear in the *New England Journal of Medicine* (see Chambers and colleagues, example in Section 7.5.A). Conversely, even if you have ten outcomes for each independent variable you still may not have an adequate sample size to answer your study question. For example, you may not have enough subjects in one of the categories of a dichotomous independent variable to study its association with outcome. Just as the sample size depends on the less common outcome state, the sample size to demonstrate that a particular independent variable is associated with your outcome will depend on the less common of the values of the dichotomous independent variable.

An example will help to illustrate this principle. Schwarcz and colleagues performed a logistic regression analysis to assess the factors associated with having received pneumocystis prophylaxis prior to a diagnosis of pneumocystis pneumonia (PCP) among HIV-infected persons.[6] The total sample size was 326 persons diagnosed with PCP. Of these, 114 (35 percent) had received prophylaxis prior to their diagnosis of PCP and 212 (65 percent) had not. The model included six independent variables. Since the smaller group (those who had received prophylaxis) was 114 there would appear to be enough subjects; using our rule of thumb we would need only 60 subjects in the smaller group.

As you can see in Table 7.2, two variables were significantly associated (adjusted P value < 0.05) with a lower likelihood of having received prophylaxis in the logistic regression analysis: being nonwhite and being uninsured. Note that the

Lee, K. L., Matchar, D. B., *et al.* "Regression models for prognostic prediction: Advantages, problems, and suggested solutions." *Cancer Treat. Rep.* **69** (1985): 1071–7.

[6] Schwarcz, S. K., Katz, M. H., Hirozawa, A., *et al.* "Prevention of *Pneumocystis carinii* pneumonia: Who are we missing?" *AIDS* **11** (1997): 1263–8.

Table 7.2 Frequency of primary PCP prophylaxis among patients whose AIDS-defining diagnosis was PCP.

Characteristic	Did not receive PCP prophylaxis *n* (%)	Received primary PCP prophylaxis *n* (%)	Adjusted *P* value	Adjusted odds ratio	95% confidence limits
Total	212 (65.0)	114 (35.0)			
Age group					
$<$ 35 years	61 (69.3)	27 (30.7)		1.0 (ref.)	
$\geq$ 35 years	151 (63.5)	87 (36.6)	0.55	1.19	0.68, 2.07
Ethnicity					
Nonwhite	74 (77.9)	21 (22.1)		0.49	0.28, 0.87
White	138 (59.7)	93 (40.3)	0.01	1.0 (ref.)	
Sex					
Male	206 (64.6)	113 (35.4)		0.81	0.06, 10.12
Female	6 (85.7)	1 (14.3)	0.87	1.0 (ref.)	
Sexual orientation					
Gay/bisexual man	185 (62.7)	110 (37.3)		3.19	0.78, 13.03
Heterosexual	27 (87.1)	4 (12.9)	0.11	1.0 (ref.)	
Injection drug use					
Yes	35 (72.9)	13 (27.1)		1.11	0.49, 2.54
No	177 (63.7)	101 (36.3)	0.80	1.0 (ref.)	
Insurance					
None	52 (82.5)	11 (17.5)		0.35	0.17, 0.73
Public/private	151 (59.5)	103 (40.6)	0.005	1.0 (ref.)	

Adapted with permission from Schwarcz, S. K., *et al.* "Prevention of *Pneumocystis carinii* pneumonia: Who are we missing?" *AIDS* **11** (1997): 1263–8. Copyright Rapid Science Publishers Ltd.

confidence intervals for these two variables exclude one and are reasonably narrow.

Gender was not significantly associated with receipt of prophylaxis: The odds ratio was 0.81. However, before concluding that the study showed that gender was not important, one must note that the 95 percent confidence intervals for gender were 0.06 to 10.12. In other words, it is equally likely that women are a sixteenth as likely or ten times more likely to receive prophylaxis. Obviously, this result is of little scientific value. The study only had seven women (2 percent of sample); the model had little information on which to base an estimate of the likelihood of having received prophylaxis for this group and this is reflected in a large standard error and a broad confidence interval. A similar situation can be seen for the variable of sexual orientation. Because there were only 31 heterosexuals

(10 percent of sample), the confidence intervals for the risk estimate associated with sexual orientation were very broad: 0.78 to 13.03.

In general, large standard errors and large confidence intervals (which are based on standard errors) are clues of an inadequate sample size. Another, more dramatic indication that you do not have a large enough sample size in logistic regression and proportional hazards analysis is if the model does not converge. There is simply not enough information, usually because of too few outcomes, for the computer to solve the equation. Although you can increase the number of attempts the computer makes to solve the equation if your model fails to converge (see Section 8.11), you should consider the possibility that you do not have enough subjects to answer your research question.

TIP

For multiple linear regression you should have 20 subjects for each independent variable in your model.

With multiple linear regression, sample size is not as much an issue as with logistic regression or proportional hazards regression. The reason is that you can consider all subjects as having experienced the outcome (because the outcome is interval). However, just as with these other techniques you will have very large standard errors (and therefore large confidence intervals) if you do not have a large enough sample size. For multiple linear regression 20 subjects (rather than outcome events) per independent variable is recommended.[7] Although this is a very reasonable standard, it does not mean that all analyses with fewer subjects are invalid; you just need to show more caution in interpreting the coefficients.

7.5 What if I have too many independent variables given my sample size?

If your power calculations or analyses show you have too many independent variables for your sample size, you need to increase the number of subjects or decrease the number of independent variables. Although increasing the number of subjects is more desirable, it is usually impossible in the analysis phase. Most researchers will therefore attempt to reduce the number of independent variables in their analysis.

For example, let's say you have determined from Section 7.1–7.2 that you would like to include 20 independent variables in your multivariable model but your sample size is sufficient for inclusion of only ten variables. What should you do?

Assuming that you have already excluded extraneous variables, redundant variables, variables with a lot of missing data, and intervening variables

[7] Feinstein, A. R. *Multivariable Analysis: An Introduction*. New Haven: Yale University Press, 1996, p. 226.

Table 7.3 Methods for decreasing the number of independent variables.

1. Exclude variables that are not empirically operating as confounders.
Variable unrelated to main independent variable and outcome in bivariate analysis
Variable has minimal impact on main effect in multivariable analysis
Variable excluded by variable selection algorithm
2. Choose one variable to represent two or more related variables.
3. Combine variables into a single variable, score or scale.
"And/or" constructions
Scores
Multi-item scales
Factor analysis

(Table 7.1), then you will need additional options. Other methods for reducing the number of variables in your model are shown in Table 7.3.

7.5.A Exclude variables that are not empirically operating as confounders

In Section 7.2 I explained that it is best, where possible, to use an inclusive definition of potential confounders, so as to lessen the chance that you are missing subtle forms of confounding (e.g., several weakly associated variables may together have a substantial impact on the main effect).

However, when you do not have a sufficient sample size you must set more stringent criteria for inclusion of variables in your model. For example, you may want to include only those variables that are associated with both the independent variable of interest and the outcome variable in bivariate analyses. Remember, unless both of these conditions are met, the variable cannot be a confounder or a suppresser.

A variation of this strategy is to include only those independent variables that change the size of your main effect by a predetermined amount. For example, Chambers and colleagues studied whether fluoxetine (Prozac) affected birth outcomes.[8] They compared the rate of prematurity among 98 infants whose mothers had taken fluoxetine early in pregnancy only to 70 infants whose mothers had taken it late in pregnancy. Only 14 infants were born prematurely. To limit the number of variables in the model, they included only those independent variables that changed the estimate of the main effect (early-only versus

[8] Chambers, C. D., Johnson, K. A., Dick, L. M., *et al.* "Birth outcomes in pregnant women taking fluoxetine." *New. Engl. J. Med.* **335** (1996): 1010–15.

late exposure to fluoxetine on prematurity) by more than ten percent. Eleven variables met this criterion and were included in the model, in addition to two other variables (maternal age and dose of fluoxetine), which were included for theoretical reasons.

With a total of 13 independent variables and only 14 outcomes, the investigators were far from fulfilling the guideline of ten outcomes per variable. Late exposure to fluoxetine was associated with an increased risk of prematurity: odds ratio of 4.8. Reflecting the relatively small number of outcomes and the large number of independent variables, the 95 percent confidence intervals for the effect of fluoxetine on prematurity were large, ranging from 1.1 to 20.8. Because the confidence intervals excluded one, the data suggest that late exposure to fluoxetine is associated with an increased risk of prematurity. However, in weighing the benefits versus risks, there is a large difference between an odds ratio of 1.1 and one of 20.8.

Variable selection algorithms (statistical procedures used to select which of your independent variables to include or keep in your multivariable model) are also used to deal with small sample sizes. However, these procedures have substantial drawbacks (Section 8.7).

7.5.B Choose one variable to represent two or more related variables

When you have two or more variables that are *moderately* correlated with one another, you may want to choose one to enter into your model to jointly represent the others. (I have italicized moderately as a reminder that we are not talking about variables that are multicollinear [Sections 6.1–6.3].) For example, income, education, and job status are correlated in the United States, but not to a degree that they would usually cause problems with multicollinearity. Ideally, in an analysis where it were important to adjust for socioeconomic status you would include all three variables. However, if this were impossible because of inadequate sample size, then including one would partially adjust your analyses for socioeconomic status. When variables are correlated, by including one, you include some of the information from the other variables. However, unless they are perfectly correlated, you will lose potentially important information by excluding one or more. The amount of information lost depends on the degree of correlation. In terms of the decision as to which variable to keep and which to exclude, the decision-making process is the same as with multicollinearity. Keep the variable that is theoretically superior, has less missing data, or has less measurement error. When available, a better solution than choosing one variable

to represent a set of variables is to construct a score or scale to represent a group of variables (see next section).

7.5.C Methods of combining multiple variables into a single variable, score or scale

Sometimes you can reduce the number of independent variables in your analysis without omitting a variable. This is done by combining variables into a single variable. Three methods are commonly used: "and/or" constructions, scores, multi-item scales, and factor analysis.

7.5.C.1 Use of "and/or" constructions

Two or more related variables can be combined with the use of an "and/or" clause (Section 6.3). For example, in the study of the effect of perinatal exposure to fluoxetine on birth outcomes, the investigators combined hypertension, pre-eclampsia, and eclampsia as one variable. In other words, women who suffered from hypertension, pre-eclampsia, eclampsia, or more than one of the three, would be "yes" on the variable; women who had none of these conditions would be "no." This fits the pathophysiology in that the three conditions are progressive states of the same underlying condition. And/or clauses may also be helpful when you have variables that are multicollinear (all women with eclampsia by definition have pre-eclampsia) or when you have an independent variable with almost everyone in the same group (as you would if you had a variable of eclampsia yes/no – almost everyone would be no, since eclampsia is relatively rare). Independent variables where almost everyone is in the same group tend to have large standard errors and large confidence intervals.

7.5.C.2 Scores

One straightforward method of reducing the number of independent variables is simply to score the number of risk factors for the disease that each subject has. For example, Turner and Lloyd computed a score to measure the lifetime exposure to adversity.[9] Subjects were assessed as to whether they had been exposed to 33 adverse life events (e.g., failing a grade in school, losing a home due to a natural disaster, being physically abused or injured). The number of adverse events experienced was totaled (i.e., the score could range from 0 to 33). The investigators found that higher scores were associated with an increased risk of developing a depressive and/or anxiety disorder.

[9] Turner, R. J. and Lloyd, D. A. "Stress burden and the lifetime incidence of psychiatric disorder in young adults." *Arch. Gen. Psych.* **61** (2004): 481–8.

7.5.C.3 Multi-item scales

Investigators may use scales (sets of closely related questions) to measure constructs that are difficult to assess by a single question (e.g., attitudes, treatment preferences). The decision to create a scale is usually made in the design phase, although researchers sometimes find, in the analytic phase, that a group of variables measure a reliable construct.

To create a scale, first code all questions in the same direction, so that, for example, a higher score is better on all items. You must also recode variables so that they are all on the same numeric scales. Otherwise variables measured on a 0 to 10 scale will have twice the weight in the total score as variables measured on a 0 to 5 scale. Dividing a 0 to 10 scale by two will put it on the same scale as a 0 to 5 scale. You can then summate the items or average them by dividing the sum by the number of variables in the scale. In creating scales you must pay close attention to the handling of missing data on individual items. If, for a particular subject, the majority of items that constitute the scale have missing values then the value for the scale should be missing for that subject. If at least half of the items, for a particular subject, have a valid response, you can replace the missing values with the mean for the sample on the particular items that are missing. Once you have replaced the missing values, you can then summate the scale and divide by the number of variables in the scale.

There is an alternative method for handling missing cases when constructing scales. You can compute the average for each subject by dividing the total by the number of variables for which the subject has valid responses. For example, if a particular subject had valid responses for four of the five variables that constitute a scale, you could summate their four questions and divide by four. In comparison, for subjects with complete values, you would summate their five questions and divide by five. If you are using this technique, at least half of the variables that constitute the scale for each subject must have complete data. Assign a missing value to subjects who have more than half the variables with missing data. (Some would assign a missing value only if a fourth or more of the questions have missing responses.)[10]

TIP

Alphas greater than 0.65 indicate that the variables form a reliable scale.

This type of multi-item scale will work only if the variables are highly correlated. The usual measure of how correlated the variables are to one another is the alpha (also referred to as a reliability coefficient). Alphas greater than 0.65 generally indicate that the variables form a reliable scale. To achieve alphas of this level, the questions are usually written with the intention of scoring them together.

[10] Hull, C. H. and Nie, N. H. *SPSS Update 7–9*. New York, NY: McGraw-Hill, 1989, p. 257.

The biggest difference between scales and scores is this requirement that scales use related questions. In the case of a score, the items do not need to be related. Using the example of the lifetime exposure to adversity score discussed above, failing a grade in school and witnessing a natural disaster would not be expected to be highly correlated, but are both certainly significant life stresses.

7.5.C.4 Factor analysis

DEFINITION

Factor analysis summarizes multiple related independent variables (say 15) into a few underlying factors (say two or three).

Factor analysis is a popular strategy in the behavioral sciences for reducing the number of variables in cases where you have multiple related independent variables. Factor analysis summarizes multiple related independent variables (say 15) into a few underlying factors (say two or three). The procedure minimizes the correlations between the factors (so that they represent distinct dimensions). Each factor is a weighted combination of the original variables. As such you can develop a factor score for each subject based on that subject's values for each of the variables, thereby reducing the number of variables to the number of factors. Each of the independent variables will be correlated with each of the factors, but to varying degrees. For any one of the factors, a few of the variables will be strongly correlated with it, indicating that the factor primarily represents this cluster of variables. Based on these correlations (referred to as loadings) you can characterize the nature of the factor (i.e., what it represents).

TIP

The major problem with factor analysis for clinical researchers is the loss of the original variables.

The major problem with factor analysis for clinical researchers is the loss of the original variables. Let's say, for example, you used factor analysis to develop factor scores for variables associated with survival from pneumonia. Assume the first factor was characterized by patient co-morbidity (age, underlying lung disease, history of congestive heart failure), the second factor was characterized by virulence of attack (severity of radiological findings, peak of fever, specific organism recovered), and the third factor was characterized by efficacy of treatment (speed of first dose, appropriate choice of antibiotics). This would be a very satisfying characterization of the data from the point of view of the pathophysiology of pneumonia.

Assume that a proportional hazards model demonstrates that the factors of patient co-morbidity and virulence of attack are associated with a worse survival, whereas efficacious treatment is associated with an improved survival. How useful would these results be to a clinician? Certainly, clinicians could not use the information to generate a probability of survival for their patients because it would be mathematically too complicated to generate factor scores. Also, you couldn't tell clinicians how important any one variable was to the outcome (only the importance of the factor to the outcome). Whereas a particular variable may have a strong loading on a factor that is related to outcome, that variable may be

only weakly related to outcome. This is because factor analysis groups independent variables based on their relationship to one another, not their relationship to the outcome variable.

For these reasons, factor analysis has a relatively small place in the analysis of medically oriented data. In the behavioral sciences, where we are dealing with constructs that are complicated to measure (e.g., self-esteem, autonomy), the "loss" of single variables is more than compensated for by the strength of the technique for dealing with multiple related independent variables.[11]

7.6 What should I do about missing data on my independent variables?

The problems caused by missing data in bivariate analysis are magnified in multivariable analysis.

Missing data is a problem in all types of analyses. However, the problems caused by missing data in bivariate analysis are magnified in multivariable analysis. Why? Because different subjects will likely have missing values on different variables. Imagine a study of 300 persons with ten independent variables. Each variable has ten missing subjects. In bivariate analysis, the sample size (n) will be 290 persons or 97 percent of your study population. But in multivariable analysis, there will likely be significantly more than ten missing subjects because cases will be dropped from the analysis if they have a missing value on any of the independent variables. At one extreme, if each of the ten variables is missing on a different ten subjects, you will lose ten cases per variable, or 100 cases for ten variables. Your n will be 200 or only 66 percent of your study population. With this amount of missing data, your power to find a significant result is less. Furthermore, your results may not be generalizable to the study population if the missing cases are systematically different from those cases where the data are not missing.

Usually, subjects who have missing values on one variable are more likely to have missing values on other variables. Therefore, the number of cases that will need to be dropped will likely be less than 100. How much less depends on how many cases have missing values on more than one independent variable. The opposite extreme from my example would be if all ten variables had missing values for the same ten cases, in which case your multivariable analysis would have no more missing data than your bivariate analysis.

In preparation for deciding how you will deal with missing data in your analysis, it is often helpful to know ahead of time how many missing cases you will have in your multivariable analysis. To determine this, create a variable

[11] For more on factor analysis see: Glantz, S. A. and Slinker, B. K. *Primer of Applied Regression and Analysis of Variance*. New York, NY: McGraw-Hill, 1990, pp. 216–36; Kleinbaum, D. G., Kupper, L. L., and Muller, K. E. *Applied Regression Analysis and Other Multivariable Methods* (2nd edn). Boston, MA: PWS-Kent, 1988, pp. 595–640.

Table 7.4 Methods for dealing with missing data in multivariable analysis.

1 Delete cases with any missing data.
2 Create dichotomous variables to represent missing data.
3 Make additional effort to obtain data.
4 Decrease the number of independent variables in the analysis.
5 Estimate the value of the missing cases.

TIP

To determine how many missing cases you will have in your multivariable analysis, create a variable whose value is "1" if the data are missing on any of the independent variables.

whose value is 1 if data are missing on any of the independent variables in your analysis and 0 if all the data are present. A simple frequency will then tell you how many cases will be missing in a multivariable analysis that includes all of these variables.

After you have determined how much missing data you have, what can you do? Table 7.4 shows five methods for dealing with missing data on independent variables in multivariable analysis. These are discussed in what follows.

Deleting cases with missing values on any independent variable is certainly straightforward and remains the most common method of dealing with missing data in clinical research. However, this strategy has two problems: loss of power and introduction of bias. Although it is easy to determine the loss of power, determining whether you have introduced bias by deleting missing data is more complicated. In general, if the cases are missing at random (such as might occur if different subjects missed answering different questions on a long questionnaire) deleting these cases should not bias the results. In contrast, if the cases with missing data are different from cases without missing data (less compliant with filling out forms, less trusting of interviewers, cognitively impaired, etc.) then deleting them will introduce bias into your results.

TIP

To assess whether your data are missing randomly, compare persons with missing data to persons without missing data on the important independent and dependent variables of your study.

To assess whether your data are missing randomly, compare persons with and without missing data on the important independent and dependent variables of your study. If there are no differences, this strengthens the argument that the data are missing randomly, and omitting cases with missing data should not bias your study (although there still may be bias owing to unmeasured factors). If there are significant differences between persons with missing data and those without, you can report these differences, so as to better characterize the potential bias in your study. Characterizing how missing cases differ from nonmissing cases is also useful information prior to estimating missing values (method no. 5).

You will note that assessing bias caused by missing cases is similar to assessing bias introduced by subjects choosing not to participate in a study (response bias). Ironically, although the bias introduced by excluding cases with missing data is of equal importance as the bias introduced by nonparticipation, it is much less

often reported in published reports. Of course, if you are missing only a few cases, it may not be necessary to evaluate the bias introduced by excluding them.

If you plan on deleting cases in your multivariable analysis that have missing values on any independent variable, you will have to decide how you want to deal with these cases in the univariate and bivariate analyses. You have two choices: You can exclude such cases right from the start of your analysis or you can wait until starting the multivariable analysis.

In published clinical research reports investigators tend to exclude such cases right from the start. The advantage of this method is that all analyses (univariate, bivariate, and multivariable) in the report then have the same sample size. The disadvantage is that much can be learned from univariate and bivariate analysis. It seems pointless to delete cases from a univariate or bivariate analysis just because they are missing on some other variable that will be part of the multivariable analysis. However, it does make it harder to follow the published analysis if the sample size changes for each analysis.

TIP

If you don't exclude the cases with missing values right from the start of the analysis, be sure to tell the reader the sample size for each analysis.

If the missing data are scattered over a large number of variables it is reasonable to delete cases with missing data on any independent variable. However, if one or two variables account for most of the missing data, it is not worth the loss of a large number of cases on the univariate and bivariate analysis to have the same sample size on all analyses. Remember, if you don't exclude the cases with missing values right from the start of the analysis, you should be careful to tell the reader the sample size for each analysis.

A second strategy for handling missing data is to create multiple dichotomous variables (as you would with a nominal variable or with an interval-independent variable that has a nonlinear relationship with the outcome), with one variable signifying persons with missing data. This strategy was used for dealing with missing data in a study of determinants of kidney transplant failure.[12] The investigators coded the variable – amount of cold ischemia time – as six dichotomous variables: 9–16 hours (yes/no), 17–24 hours (yes/no), 25–36 hours (yes/no), 37–48 hours (yes/no), greater than 48 hours (yes/no), and missing value (yes/no). The reference group was 0–8 hours.

TIP

Use of a dichotomous variable to indicate missing data allows all subjects to be included in the multivariable analysis without making a strong assumption about the missing subjects' values.

The advantage of using a dichotomous variable to indicate missing data is that it allows all subjects to be included in the multivariable analysis, without making a strong assumption about the missing subjects' values. It has the additional advantage that you get some sense of the bias caused by missing data. In the case of the kidney transplant study, those "yes" on the missing variable had the highest risk of graft failure. This would suggest that those with

[12] Chertow, G. M., Milford, E. L., Mackenzie, H. S., *et al.* "Antigen-independent determinants of cadaveric kidney transplant failure." *JAMA* **276** (1996): 1732–6.

missing values actually had cold ischemia times greater than the other categories (since long cold ischemia time was associated with higher rates of graft failure). The authors also reported that the fit of models that included a dichotomous variable representing those cases with a missing value on cold ischemia time were not significantly different from the fit of models that excluded cases with missing data.

Since this book is primarily about data analysis, it may surprise you that I have listed "additional effort to obtain data" as the third strategy for dealing with missing data. Won't it be too late to go back to obtain additional data once you are already in the data analysis phase? Certainly, the data collection phase is the most appropriate and efficient time to obtain complete data. I mention this strategy here because the impact of missing data is often felt most acutely in the data analytic phase and sometimes researchers are subsequently able to obtain data that were previously missing. In the case of one study I was involved in, the missing data were in another city. When a thoughtful reviewer pointed out the weakness in our study caused by the missing data, we sent a research assistant on a trip to obtain the data. Some of you may complain that we should have sent a research assistant to collect the data from the other city right from the start. But, research, like any enterprise, is a series of trade-offs between costs (e.g., time, travel) and gains (e.g., more data). It was only after the variable proved to be so important to the analysis (and to the odds that the paper would be published!) that it seemed worth the effort and expense to get the data.

TIP

Try to eliminate variables with a large number of missing cases.

The fourth method for dealing with missing data, decreasing the number of independent variables in the analysis, works only if you have variables that can be eliminated without compromising your analysis. In Section 7.5, I discussed strategies for decreasing the number of independent variables in instances where your sample size is insufficient for the number of variables in your analysis. These strategies can also help with missing data. There are a few differences worth mentioning. Usually, some variables have more missing data than other variables in the study. Those variables with a large number of missing observations are the variables you should try to eliminate. If you have two related variables and one has a lot more missing data than the other, exclude the one with the greater number of missing observations. For example, education and income are highly correlated. If you have education level for everyone but income level for only 75 percent of the subjects (people are more sensitive about disclosing income than education level), it may be preferable to drop income and use only educational level in the analysis. Since income is not the same as educational level you will certainly lose information by doing this. Only you as the researcher can answer the question of whether you lose more by dropping the cases or by dropping the variable.

Table 7.5 Methods of estimating missing values.

Assign the sample mean.
Assign the mean by subgroup (conditional mean).
Model the value of the missing data by using the other covariates in the analysis (simple imputation).
Model the value of the missing data by using the other covariates in the analysis and include a random component (multiple imputation).

The fifth method for dealing with missing data, estimating the value of missing cases, is the most satisfying but also the most dangerous method. It is the most satisfying because you don't lose any cases; it is the most dangerous because you may bias your results in ways that are difficult to predict. Several methods of estimating missing values are shown in Table 7.5.

TIP

The simplest method of assigning a missing value for an independent variable is to assign the sample mean (or median) for that variable.

The simplest method of assigning a missing value for an independent variable is to assign the sample mean (or median) for that variable. (Choose the median if the distribution is skewed.) By assigning the mean/median you are saying that you believe the missing data are occurring randomly and therefore the mean/median provides the best estimate. The benefit of this procedure is that you get to keep the case with missing data in your analysis. However, this method is only sensible if the subjects for which you are assigning the mean/median have only one or two independent variables with missing values. If, for example, you have 15 cases that have missing values on only one of ten independent variables, by assigning the mean/median for the missing variable you keep all 15 cases in your analysis. They are useful cases because the information on the other nine independent variables is real. Viewed from the other extreme, if you have 15 cases with missing values on all ten variables, it would serve no purpose to assign them the mean for each of the ten variables. Since all of the data on the independent variables are missing, they contribute no information to your multivariable analysis. Therefore prior to assigning values to missing data make sure that the cases have true values for at least half of the independent variables in your analysis.

TIP

Conditional means will likely yield more accurate estimates of missing values than sample means.

When you assign missing values, you may want to assign mean/median values by subgroups rather than using the mean/median for the entire sample. This procedure is referred to as a conditional mean (conditional on the value of other variables). For example, if you have a number of cases with missing data on income, rather than assigning the mean/median for the whole sample, you may assign these cases the mean/median for other subjects of the same educational level and occupational status. Since income is correlated with educational

attainment and occupational status, assigning the mean/median by subgroup will likely yield more accurate estimates of missing values.

A more sophisticated method of estimating the mean is to perform a multiple linear or logistic regression analysis using the other independent variables to estimate the missing value. This method, usually referred to as imputation, may allow a more precise estimate of the missing value than assigning the mean/median. For example Smith and colleagues followed 383 patients for 24 months to assess the impact of a primary care intervention on depression.[13] Sixty-two of the subjects had a missing value for income. The researchers therefore used multiple linear regression to estimate the missing values. They included eight independent variables in the model: age, gender, race, education, marital status, employment, physical health rating, and mental health rating. Because all of these variables would be expected to be associated with income (older, white, well-educated, married, employed, healthy men would be expected to have a higher income than younger, nonwhite, less educated, unmarried, unemployed women in poor physical and mental health).

However, as with all the above methods of estimating missing values, if you use the estimated values in your multivariable analysis (as if it were an observed value) you will underestimate the error associated with your coefficients. The reason is that once you have filled in the missing values based on your regression analysis, the computer does not know that the filled-in values have more "error" than those values that are actually observed. This results in confidence intervals surrounding the estimates that underestimate the actual variability of those estimates.

TIP

Multiple imputation methods allow you to estimate a missing value, while adding in a random component.

To overcome this problem, multiple imputation methods allow you to add in a random component.[14] With multiple imputation, you fit a multiple regression or logistic model for the variable with missing values using subjects with complete data on this variable and its important correlates. The fitted model provides an estimate of the mean and variance of each missing value, given the data on the correlates available for that subject. Next, for each missing value, you use a random number generator to simulate an observation from the estimated distribution, under the assumption that interval variables are normally distributed and that dichotomous variables have a binomial distribution. Then the primary analysis is carried out using this data set completed by the imputed missing values. This procedure is repeated at least ten times, and the results are

[13] Smith, J. L., Rost, K. M., Nutting, P. A., *et al.* "Impact of ongoing primary care intervention on long term outcomes in uninsured and insured patients with depression." *Med. Care* **40** (2002): 1210–22.

[14] Heitjan, D. F. "What can be done about missing data? Approaches to imputation." *Am. J. Pub. Health* **87** (1997): 548–50.

combined using available formulas.[15] Repeating the procedure makes it possible to compute standard errors that take into account the extra uncertainty induced by the imputation, since each data set is completed with different imputed values for the missing data.

Since each method of dealing with missing values has its advantages and disadvantages, some studies will use a combination of methods. For example, Halfon and colleagues conducted a study on access to health services among Latino children.[16] Income was not reported for 13 percent of the sample. They estimated income by replacing missing values with the sample mean. In addition, they created a dichotomous variable representing the cases with missing data on income (in other words, the variable equals 1 if cases are missing and 0 if the cases are not missing). In this way, they were able to provide a value for income for their entire sample and adjust for the possibility that the cases with missing data were different from those without missing data.

One advantage of trying a variety of methods for dealing with missing data (e.g., eliminate cases, assign the mean, impute values) is that you can see if your choice of method makes a difference in the results. It is reassuring to researchers and to readers when different methods of dealing with missing data produce similar findings.

My general guidance on this complicated issue is:

1. Collect your data to minimize missing information.
2. Assess how much missing data you have on individual independent variables.
3. If you have one or two independent variables that have significantly more missing cases than your other variables, consider deleting the variables rather than the cases. No matter how important the variable is to your theory, if you have a lot of missing data on that variable, your information is likely to be biased.
4. After you have minimized missing data through steps 1 and 3 above, check to see how many cases have missing values on any of the independent variables you are planning to use in your multivariable model. If you have few cases with missing data, delete them right from the start. It is easier to follow a paper that has the same sample size for all analyses.
5. If you have a large number of cases with missing data, determine if cases with missing values differ from cases without missing values.
6. If missing cases do not differ from nonmissing cases, consider assigning means or conditional means. Before you do this, make sure that the cases have true values for at least half of the independent variables in your analysis. If you

[15] Rubin, D. B. *Multiple Imputation for Nonresponse in Surveys*. New York, NY: Wiley, 1987.
[16] Halfon, N., Wood, D. L., Valdez, B., *et al.* "Medicaid enrollment and health services access by Latino children in inner-city Los Angeles." *JAMA* **277** (1997): 636–41.

have cases that are missing data on most of your independent variables, delete them. With the help of a biostatistician consider using a multiple imputation approach.

7. If missing cases differ from nonmissing cases, you are in a tough spot. Go forward as in step 6, but be clear in your own mind, and to your readers, that assigning values based on the other cases is problematic since you know that the cases with missing values are not the same as nonmissing cases. Of course, excluding them is problematic for the same reason.
8. If possible, try more than one method for dealing with missing data.
9. Read more about the theory and practice of dealing with missing data.[17]

7.7 What should I do about missing data on my outcome variable?

Of those strategies listed in Table 7.4 for dealing with missing data on independent variables, only deleting cases and making additional effort to get the data will work reliably for a missing outcome variable. You can't eliminate your outcome variable (that is, what you are studying). You also cannot estimate your outcome variable prior to the multivariable analysis. The whole purpose of multivariable analysis is to estimate the outcome variable based on the independent variables. For longitudinal studies, data from persons lost to follow-up can contribute to the analysis by censoring observations (Section 3.6). However, it is best not to think of censored observations as missing outcome data. For a censored observation you know what the outcome is at a particular time. You just don't know the outcome beyond that time. Also in a longitudinal study where the outcome is measured repeatedly you may have some subjects with both missing and actual observations of the outcome. Methods for dealing with missing data with repeated observations of the same subject are dealt with in Section 12.1.

In this section, I want to focus on a strategy for dealing with a missing outcome measure at a particular point in time: multiple imputation. How does multiple imputation work? Remember that multivariable models estimate outcome based on the relationship of the independent variables to the outcome. Once you have estimated outcome based on cases where you have information on independent variables and outcome, you can estimate the outcome of cases where you only have information on the independent variables. What good does this do? By estimating the outcome variable for cases with missing outcomes and including

[17] Marascuilo, L. A. and Levin, J. R. *Multivariate Statistics in the Social Sciences: A Researcher's Guide.* Monterey, CA: Brooks/Cole Publishing Co., 1983, pp. 64–6. Delucchi, K. L. "Methods for the analysis of binary outcome results in the presence of missing data." *J. Consult. Clin. Psych.* **62** (1994): 569–75. Little, R. J. A. and Rubin, D. B. *Statistical Analysis with Missing Data.* New York, NY: Wiley, 1990. Greenland, S. and Finkle, W. D. "A critical look at methods for handling missing covariates in epidemiologic regression analyses." *Am. J. Epidemiol.* **142** (1995): 1255–64.

a component that takes into account the variability of this estimate, you can repeat your multivariable analysis with the additional cases and see if your results differ. If they do not, it strengthens the validity of your analysis.

This procedure was used in an evaluation of an HIV-prevention intervention tailored for young gay men.[18] The researchers assessed the sexual risk activities pre- and post-intervention. They found significant decreases in HIV-risk activities between the pre- and post-assessment for the intervention group compared to the non-intervention group. However, of 191 young men who received the pre-intervention assessment, only 103 (54 percent) were available for the post-intervention assessment. This substantial loss of the sample raises questions about the validity of the observed differences.

Even more problematic for the researchers, there were significant differences between those subjects lost to follow-up and those not lost to follow-up. Could these differences, rather than the intervention, explain why there were decreases in sexual risk activities following the intervention? There is no way to definitively answer this question since the subjects were lost and we do not know their ultimate outcome. But we do know something about their pre-intervention behavior.

What the researchers did is to estimate outcome for those subjects who had both pre-intervention and post-intervention interviews using logistic regression analysis. They then used these models to estimate outcome for those cases without a post-intervention assessment. Next, using a multiple imputation procedure, they generated 100 data sets in which the missing outcomes were randomly imputed from the distribution of the missing value according to the logistic model and the observed baseline covariates for the subject. The treatment effect was estimated by the average of the effect estimates for each of the 100 data sets. The standard errors were corrected for the multiple imputation by a factor depending on the variance of the 100 effect estimates. The results of this repeated analysis were similar to those of the analysis in which these cases were excluded. While this strengthened the conclusion of the paper, it certainly does not exclude bias as the explanation of their findings.

[18] Kegeles, S. M., Hays, R. B., and Coates, T. J. "The Mpowerment project: A community-level HIV prevention intervention for young gay men." *Am. J. Pub. Health* **86** (1996): 1129–36.

8

Performing the analysis

8.1 What numbers should I assign for dichotomous or ordinal variables in my analysis?

Let's take the simplest case of a dichotomous independent variable based on an interview question: Do you have a history of diabetes: yes or no?

The equations used to solve multivariable analysis need numerical representations of yes and no. Since this scale only has two points, the numeric distance between the two points can be represented by any two numbers that are separated by one: 0 and 1, 1 and 2, 0 and −1, etc. It doesn't really matter. The sign of the coefficient may change depending on whether you assign "yes" the higher or the lower value, but the coefficient and significance level will be the same (Section 9.3). However, you will not get the same answer if you code your variables such that there is more than one point between the two numbers. For example, coding schemes like +1 and −1 will give you a different answer because there is more than one unit between the two points.

Although any two numbers that are one number apart will give you the same answer, a sensible convention, for both independent and dependent variables, is to use 1 and 0, with 1 representing the presence and 0 representing the absence of the condition. This convention is easy to remember and decreases the chance that you will be confused at the direction of the effect. This coding scheme has another advantage: When a variable is coded this way, the mean of the variable represents the prevalence of the condition. For example, if you have 100 subjects and 10 experience the outcome, the mean of the variable (if coded 0,1) will be $([0 \times 90] + [1 \times 10])/100 = 0.10$. This can be handy when you want to know the prevalence of a risk factor or outcome in a particular group of patients.

TIP

Code your variables as 1 = presence of condition and 0 = absence of the condition; then the mean of the variable will be equal to the prevalence of the condition.

For some independent variables, such as gender, where there is no absence or presence of a condition, assign the 1 to the value that will make the most sense in how you will discuss your results. For example, if your hypothesis is that women

Table 8.1 Implications of changing the reference group for dichotomous variables.

	Odds ratio	Odds ratio
White/Caucasian	1.0 (reference)	4.0
African-American	0.25	1.0 (reference)
Latino	0.50	2.0
Asian/Pacific Islander	0.50	2.0
Native American	0.25	1.0
Other nonwhite ethnicity	0.50	2.0

with coronary artery disease receive fewer procedures owing to gender bias, it would be sensible to assign women the 1 and men the 0.

8.2 Does it matter what I choose as my reference category for multiple dichotomous ("dummied") variables?

We have reviewed how to create multiple dichotomous variables from a single variable to represent nominal and ordinal variables (Section 4.2) and how to use multiple dichotomous variables to deal with interval-independent variables that are related to outcome in a nonlinear fashion (Section 5.5).

In Section 4.2 I explained that a variable representing the reference group would not be entered into the analysis; rather that the other categories would be compared to this group. Given that, does it matter which category you choose as your reference group? The answer is that your choice of reference group makes a difference in how you report your results and a small difference in the results themselves.

Table 8.1 illustrates the implications of varying your reference group. Assume the data are from a study on the association between ethnicity and access to healthcare. In column 1, the reference group is white/Caucasian. The odds ratios indicate that African-Americans and Native Americans are a fourth as likely as whites/Caucasians to receive medical care, whereas Latinos, Asian/Pacific Islanders, and other nonwhites are half as likely as whites/Caucasians to receive medical care.

Column 2 lists exactly the same data, but now African-Americans are the reference group. We see that whites are still four times more likely to receive medical care as African-Americans. Latinos, Asian/Pacific Islanders, and other nonwhites are about twice as likely as African-Americans to receive medical

care, whereas Native Americans are equally likely as African-Americans to have received medical care.

Although column 1 and 2 are mathematically equivalent, the reporting of the results is slightly different. If the hypothesis of your study was that persons of color have less access to medical care than persons who are white/Caucasian, it would be sensible to have the white/Caucasian group be the reference group as in column 1. This gives you the ability to report to your readers how access to medical care differs for persons of color compared to persons who are white/Caucasian. If you made African-Americans your reference group as in column 2, you would not be able to directly compare Latinos, Asian/Pacific Islanders, Native Americans, and other nonwhites to the white/Caucasian group because the reference is to African-Americans. If, however, your research question concerns whether African-Americans are less or more likely to receive medical care than other ethnicities, coding African-Americans as the reference category, as in column 2, is sensible.

For this reason, investigators generally choose the reference category based on the main hypothesis being tested. If you have no main hypothesis, and your dummied variables represent an interval variable (such as age), it is generally easier to report your results in a manner consistent with your empirical findings. For example, if age is associated with increasing (or decreasing) rate of outcome, you should use the extreme category (e.g., the youngest or the oldest subjects) as the reference group. This allows you to summarize your results by saying older persons are more likely (or less likely) than younger persons (the reference category) to experience the outcome. Conversely if the variable underlying the dummied variables has a **U**-shaped distribution (Section 5.5), it may be best to code the middle group as the reference group so that you can demonstrate the elevated risk at the two extremes.

TIP

If your hypothesis and empirical findings do not lead you to choose a particular category as your reference group, choose the one with the largest sample size.

Which group you choose as your reference group makes a small statistical difference. If you choose the largest group as your reference category, the standard errors will be slightly smaller and the confidence intervals will be somewhat narrower because the model has a larger comparison group and can therefore make more precise estimates. Although this is not a major factor in most studies, if your hypothesis and empirical findings do not lead you to choose a particular category as your reference group, choose the one with the largest sample size.

8.3 How do I enter interaction terms into my analysis?

In Section 1.4 I explained that an interaction occurs when the association of an independent variable on outcome is changed by the value of a third variable. How do you deal with interaction terms in a multivariable analysis?

Table 8.2 Creation of an interaction (product) term.

	ST elevations	
Male gender	Yes (=1)	No (= 0)
Yes (= 1)	1 × 1 = **1**	1 × 0 = **0**
No (= 0)	0 × 1 = **0**	0 × 0 = **0**

The most common method of incorporating an interaction in a multivariable model is to create a product term.

The most common method of incorporating an interaction in a multivariable model is to create a product term. This is done by creating a variable whose value is the product of two independent variables (i.e., the two variables multiplied by each other). A product term between two independent variables is referred to as a two-way interaction or a primary interaction. A product term between three independent variables is referred to as a three-way interaction or a secondary interaction.

In Section 1.4 I reviewed an example of an interaction between gender and ST elevations. The coding for the product term for male gender and ST elevations is shown in Table 8.2. Note that the two variables male gender (yes/no) and ST elevations (yes/no) divide the sample into four groups: men with ST elevations, men without ST elevations, women with ST elevations, and women without ST elevations. Each cell has its own unique combination of these two variables. In bold is the value of the product term. Note how the product term highlights those subjects who have both risk factors (male and ST elevations).

To determine if an interaction was present between male gender and ST elevations, the authors entered the product term into their multiple logistic regression analysis, along with the two variables constituting the product term (male gender and ST elevations). If there had been no interaction, meaning the effect of the two risk factors on outcome (heart attack) is captured by the two variables, male gender and ST elevations, then the product term would have been nonsignificant. The authors would have established that there was no interaction between male gender and ST elevations. Instead the product term was significant, indicating that there was an interaction. In this case the sign on the product term was negative, indicating that the effect of being male and having ST elevations had significantly less impact on the likelihood of heart attack than you would have expected from the individual effects of male gender and hypertension.

Because a product term describes the relationship between two risk factors and an outcome, it can only be interpreted as an interaction if the two risk factors (in this case male gender and ST elevations) are in the model. If you enter only the product term, without assessing the individual risk factors in the model and the product term is significant, you don't know if the product is significant because

Because a product term describes the relationship between two risk factors and an outcome, it can only be interpreted as an interaction if the two risk factors are in the model.

there is an interaction between the risk factors or because the risk of outcome is significantly higher when both risk factors are present (compared to subjects who do not have both risk factors). In this example, if the investigators did not include the separate variables for male gender and ST elevations, and entered only the product term, the product term certainly would have been statistically significant and positive (since males with ST elevations are at higher risk of heart attack than the rest of the sample). But the importance of the product term is that it is statistically significant and negative when both male gender and ST elevations are in the model.

Although I have stressed the importance of initially including the variables that constitute the product term in the model, it would not be incorrect to have a model that had only the product term. If, for example, the two variables constituting the product term are not on their own statistically associated with the outcome in initial models, it would be acceptable to drop them from subsequent models.

An alternative method for incorporating product terms into your analysis is to create multiple dichotomous variables representing the interaction. Look back at Table 8.2. There are four distinct codings of the variables gender and ST elevations. Rather than entering three variables representing gender, ST elevations, and the product of gender and ST elevations, you could create three dichotomous variables:

men with ST elevations (yes/no)
men without ST elevations (yes/no)
women with ST elevations (yes/no)

The reference group would be women without ST elevations. One advantage of this coding is that it will be easier for you to see and interpret the impact of the combinations of gender and ST elevations on outcome. A second advantage is that you can see the effect of the double-exposed group (male and ST elevations) compared to persons with only one risk factor and persons with neither risk factor (the reference group). (When you use product terms you see the risk of the doubly-exposed persons compared to persons with only one or no risk factors.)

A disadvantage of multiple dichotomous variable coding is that if you are looking at multiple interactions involving a particular variable (e.g., male gender) you will have to create more additional variables than you would if you were using product terms. For example, if in addition to the interaction between male gender and ST elevation you wanted to describe the interaction between gender and congestive heart failure you would need three variables: men with congestive heart failure, men without congestive heart failure, and women with congestive heart failure. If you wanted instead to create an interaction term you would enter

Table 8.3 Starting time for survival analysis.

Type of study	Start time
Randomized controlled trial	Date of randomization
Nonrandomized trial	Enrollment into trial
Observational study	Varies:
	Date of first visit
	Date of first symptom
	Date of diagnosis
	Date of start of treatment

only two variables: (1) congestive heart failure and (2) the product of congestive heart failure and male gender. You would not have to add a variable for male gender because it is already in the model.

8.4 How do I enter time into my proportional hazards or other survival analysis?

For linear and logistic regression you need only enter your independent and dependent variables. For proportional hazards analysis and other types of survival analysis you must also enter a time for each subject. The time is the interval from a subject's participation in the study to the date the subject experienced an outcome, was lost to follow-up, was withdrawn, or completed the study.

The starting point ("zero time") will depend on the kind of study you are performing, as shown in Table 8.3. For a randomized controlled trial the starting time is the date of randomization. For a trial that prospectively enrolls subjects but does not randomize them to a treatment, the starting point is usually the date of enrollment.

In observational studies, the choice of starting point is complicated. The goal is to choose a starting point that best represents the start of the process you are studying. For example, in evaluating the rate of death in patients with coronary artery disease, the starting point should be the onset of coronary artery disease. But how do you determine the date that coronary artery disease began? Is the starting point the date that the patient first developed chest pain? This sounds good, but remember some patients have coronary artery disease without ever having chest pain. Others have chest pain for years from some other cause before they develop coronary artery disease. Also, some patients will not remember their first episode of chest pain; they may report that they have had chest pain for "years."

What can you do to get a more precise starting time? You could use the date coronary angiography first demonstrated coronary artery stenosis. This starting date has the advantage of being the most objective (angiography is the gold standard for diagnosing coronary artery disease). But many patients never require angiography, and access to care and patients' willingness to undergo testing will affect whether and when they have angiography.

Often with observational studies, no one starting point truly represents the onset of the disease process for all participants. You have to choose the best one you have available. In a study of patients seen in a clinical setting this may be the date the patient first presented for medical care. In a prospective cohort study the starting point may be the first cohort visit. Although not ideal, date of first visit has been used in many studies. Notably, many natural-history studies of HIV infection use the date of first visit because this was the first date that the participant was documented to be HIV-antibody positive. The actual disease process had begun months to years earlier when the person actually seroconverted to HIV. Although use of first visit did bias the results from these cohorts and has led to some inaccurate observations, these studies were nonetheless extremely helpful in understanding the nature of HIV disease.[1]

In conclusion, choose the starting point that best represents the start of the process you are studying and clearly state the choice in the methods section of your paper.

The endpoint for survival analysis is the date of the outcome of interest or the censor date (Section 3.6). For subjects lost to follow-up prior to outcome the censor date is the last date of known follow-up. For subjects who did not experience an outcome and were not lost to follow-up, the censor date is the end date of the study (assuming intention-to-treat analysis for patients who are withdrawn).

In some studies there may be ambiguity about the appropriate censor date because the investigators have access to supplemental sources of data about study participants. Analysis of survival time following an AIDS diagnosis provides a good illustration of this principle. Let's say you want to determine whether persons who are lost to follow-up in your study have died. You know that death certificates are part of the public record and it is therefore possible to determine whether subjects who are lost to follow-up have died. However, you are also aware that there is usually a delay between when a subject dies and when you will learn about their death from a local, state, or national death registry. Conversely, matching with a death registry may enable you to learn about the deaths of some participants sooner than you otherwise would have. This can occur if you

[1] For a perspective on the biases of prevalent cohorts of HIV-infected persons see Alcabes, P., Pezzotti, P., Phillips, A. N., *et al.* "Long-term perspective on the prevalent-cohort biases in studies of human immunodeficiency virus progression." *Am. J. Epidemiol.* **146** (1997): 543–51.

interview subjects only periodically (e.g., every six months to a year) but perform frequent reviews with a death registry (e.g., weekly to monthly). How should you deal with this supplementary information about survival?

In San Francisco, we follow all persons who are diagnosed with AIDS by reviewing their medical records every six months. We also review all of the death certificates in San Francisco weekly. Each year we perform a match with the National Death Index.[2] This index covers the deaths of all persons in the United States. Thus, when someone dies we almost certainly find out about it. We need to decide what is the appropriate date to use as the last date of follow-up for someone who is not known to have died. The algorithm we use to determine this date is somewhat complicated but illustrates the types of decisions you must make about the last date of follow-up.

For subjects not known to have died, we check the medical record for the date of the last medical visit or laboratory test. What about people whose records show no recent entries and yet are not listed as being dead? These people either died outside San Francisco (because if they died in San Francisco we would know it since we do weekly reviews of San Francisco's death certificates), moved, switched their site of care, or stopped receiving medical care altogether. What should our last date of follow-up be for these individuals?

For persons not known to have died, with no recent medical follow-up, we use for their censor date the date to which the National Death Index is current at the time we match our database with theirs. The National Death Index receives all state death certificates and updates their computer files for the deaths that occurred in a calendar year within twelve months of the end of the calendar year. For example, if we performed our match in June of 2005, the data would be complete for the calendar year of 2003. For those cases lost to follow-up, we would use December 31, 2003 as the censor date.

Of course, matches with the National Death Index are not perfect. The index is not 100 percent complete (nothing ever is). Also, it is possible that a case is listed in the National Death Index but we are unable to match with it because we have incorrect identifying information (e.g., the wrong date of birth). A more conservative strategy than using the date for which the National Death Index is current would be to use the last dates listed in the medical records of those patients not known to have died. The problem with this strategy is that it underestimates survival because it counts those deaths that we know occurred after the last date of follow-up, but not the survival time beyond the date of follow-up. At the other extreme, we could censor everyone at the date of analysis. Supporting this strategy is the fact that most San Francisco AIDS patients die in San Francisco

[2] For information about the National Death Index (for the United States) along with an application for matching your data with the Index see: www.cdc.gov.nchs/r&d/ndi/ndi.htm.

and we review the death certificates weekly. Therefore, in most cases we will know promptly if someone has died. However, using the date of analysis would overestimate survival because it would count all of the follow-up time but would miss some of the deaths. Our method is something of a compromise.

As you can see, the date of censor can be quite a complicated issue. Your choice will affect the survival time. van Benthem and colleagues illustrate this using a similar example to mine, that of AIDS incubation time (time from HIV seroconversion to AIDS diagnosis).[3] In their example, they show that AIDS incubation time varies based on when participants are censored. When participants with no known AIDS diagnoses are censored at the date of their last visit, incubation time is underestimated (because information about deaths from registry-matches is included but additional AIDS-free time is not included). When participants with no known AIDS diagnoses are censored at the date of analysis, incubation time is overestimated (because it assumes that the information from the registries is complete, which it is not). In their example, they advocate for an alternative method: Persons seen in the year prior to the analysis, who are not in the AIDS registry, are censored at the date of analysis; persons not seen in the year prior to the date of analysis and persons who developed AIDS more than a year after their last visit are censored one year after their last visit.

TIP

Choose a censor date that balances information about outcomes with information about outcome-free time.

The most interesting thing about the analysis of van Benthem and colleagues is that they demonstrate that differences in AIDS incubation time reported by different studies may actually be caused by differences in the censoring techniques. Thus, in handling supplementary information, I recommend you balance information about outcomes with information about outcome-free time and try to be consistent with how others in your field have dealt with this issue.

Once you have settled on the start date and end date for each subject, the difference between these dates represents the survival time for each subject in your analysis.

Table 8.4 illustrates calculations of time for different types of subjects. Subjects were enrolled between May 1, 2004 and August 1, 2004 and were followed until August 1, 2005 unless they dropped out or were withdrawn. The outcome of interest is heart attack.

Subject 1 experienced a heart attack one year (365 days) after enrollment but continued to be followed after the outcome. This is common in clinical studies. You might follow someone beyond their main outcome of interest because you are assessing the development of side effects or a secondary outcome (e.g., death). However, note that to determine time to heart attack for this analysis, you subtract

[3] van Benthem, B. H. B., Veugelers, P. J., Schecter, M. T., *et al.* "Modelling the AIDS incubation time: Evaluation of three right censoring strategies." *AIDS* **11** (1997): 834–5.

Table 8.4 Illustration of time calculations for individual subjects.

Subject	Start Date	Did heart attack occur?	Date of heart attack	Date of last follow-up	Time (days)
1	May 1, 2004	Yes	May 1, 2005	August 1, 2005	365
2	May 1, 2004	No	Not applicable	August 1, 2005	457
3	August 1, 2004	No	Not applicable	August 1, 2005	365
4	May 1, 2004	No	Not applicable	July 1, 2004	61

the start date from the date of outcome, not the date of last follow-up. Let's contrast this with subject 2. This subject did not have a heart attack. Therefore time is the difference between the start date and the date of last follow-up. Subject 3, like subject 2, did not experience a heart attack. But this subject enrolled in the study later than subjects 1 and 2. Therefore, even though the subject stayed till the end of the study, the subject would be censored at 365 days. Subject 4 dropped out of the study and is censored at 61 days.

The four subjects shown in Table 8.4 illustrate two important points about survival analysis.

- Survival analysis tracks length of time without reference to calendar time. If you changed the decade in which the study occurred by subtracting ten years from all the dates, you would get the same survival time. This is the reason that many analyses adjust for year of diagnosis or birth cohort (i.e., year or period of years of birth).
- There is no special designation for cases that are censored. All subjects that do not experience an outcome are censored. The only difference between subjects 2, 3, and 4, from the computer's point of view, is the amount of time they contribute to the analysis.

There is another method for incorporating time into a proportional hazards model: Use of age of subject rather than study time. Using age instead of study time makes sense in observational studies of healthy persons. This is because the hazard of an outcome such as death for a 55-year-old man observed for 15 years is likely to be more similar to the hazard for a 55-year-old man observed for 5 years, than that for a 40-year-old man observed for 15 years.

Korn and colleagues argue persuasively for the use of age instead of study time for observational studies of healthy persons drawn from national surveys.[4] However, their empirical analysis demonstrates that the more usual method of study time, with adjustment for subject's age, produces unbiased estimates even when age may be a more appropriate time scale.

[4] Korn, E. L., Graubard, B. I., and Midthune, D. "Time-to-event analysis of longitudinal follow-up of a survey: Choice of the time scale." *Am. J. Epidemiol.* **145** (1997): 72–80.

Although the use of age in place of study time has its adherents, it is not commonly done, even for surveys of healthy persons. It would certainly not be appropriate in studies of persons with disease. In persons with an illness (e.g., cancer, heart disease) the amount of time that they have the disease is likely to be more closely related to their rate of outcome (e.g., death) than their age.

If you do choose to use age as your time scale, it is important to adjust for birth cohort. Otherwise, your model will not take into account treatment changes that have occurred during the lives of your participants.

8.5 What about subjects who experience their outcome on their start date?

It sometimes happens that subjects experience their outcome on their start date. If this occurs, the time for such subjects would be zero. Since, at time zero, by definition, none of the subjects have experienced an outcome, persons with time equal to zero must be excluded from the analysis. Is this fair? Can you do anything to prevent this?

To answer this question, you have to distinguish those cases where the outcome truly occurred on the start date from those cases where the start date and the outcome are *recorded* as occurring on the same date, but the start date is really unknown. I will illustrate with a few examples.

Imagine you are studying hospital survival with a rapidly progressive disease, such as adult respiratory distress syndrome (ARDS). A certain number of patients will die on their day of admission to the hospital. In this case, if you computed survival in days, patients who died on their date of admission would appear to have a survival of zero days and would be excluded from the analysis. Clearly, this is not what you would want. The true survival time for these patients is in hours. Our use of a day as the unit of survival analysis is arbitrary. For this example, you should switch your unit of analysis to hours. This will work well for ARDS or other diseases that have a very rapid progression time. Day is the convention for most survival analyses because improvement and worsening of most clinical conditions occur in days not in hours.

TIP

The use of a day as the unit of survival analysis is arbitrary. Hours may be a more appropriate unit for rapidly progressive diseases.

Consider a more complex example: How to categorize patients who are diagnosed with AIDS and die on the same day. If you were to review data from the San Francisco Health Department's AIDS registry, you would discover that some of our cases have the same date for AIDS diagnosis and death. There are two reasons for this. In some cases, HIV-infected patients without an AIDS diagnosis are admitted to the hospital, diagnosed with an AIDS-defining illness for the first time, and die the same day they are admitted. In this case, as with the ARDS example, there is a real survival time, measured in hours. Unfortunately,

our records do not contain the hour of AIDS diagnosis or death. In other cases, the patients' date of diagnosis is the same as their date of death because they are diagnosed by the medical examiner (coroner). In these cases, it is unclear what the true survival time is because you don't know if they had an AIDS illness for a short or a long time before death. How do we deal with these two types of cases, both of whom have a survival time of zero?

For cases diagnosed on the day of admission, we consider their survival to be 0.5 days. The half day acknowledges that the death truly occurred after the diagnosis of AIDS, but after an interval of less than one day. (Some statistical software programs will automatically add 0.5 units to cases with a survival time of zero. However, you as the investigator should determine whether this is a reasonable assumption or not.) For those cases diagnosed by the medical examiner, we exclude the case because we do not know what the true interval is between diagnosis and death.

The AIDS registry of the New York City Health Department has an even more complicated problem. Unlike San Francisco, they only record the month and year of AIDS diagnosis. Thus a case who died in the same calendar month as their AIDS diagnosis would have a survival time of zero. They therefore have a large proportion of cases (11 percent) with a survival time of zero.[5]

How do the investigators deal with New York City AIDS cases with survival time equal to zero? They exclude them from the analysis. This may be problematic. To the extent that such persons truly had a short survival, the investigators' method will artificially lengthen survival by excluding these subjects. Because such cases represented a large group, the investigators assessed whether the survivors were different from other participants. In fact they were: They were more likely to be female, persons of color, and injection drug users. This illustrates another important point. You cannot always eliminate bias whether caused by loss of cases or some other reason. Nonetheless, you should always investigate it and describe it to your readers (as these authors did).

8.6 What about subjects who have a survival time shorter than physiologically possible?

It sometimes happens that subjects experience their outcome so soon after their start date that the survival time is not physiologically possible. This is most likely to pose a dilemma with slowly progressive diseases, for which the physiology of the disease does not support a survival time of a day or a week. For example,

[5] Blum, S., Singh, T. P., Gibbons, J., *et al.* "Trends in survival among persons with acquired immunodeficiency syndrome in New York City." *Am. J. Epidemiol.* **139** (1994): 351–61.

what do you do with a subject enrolled in a study of cancer incidence who is diagnosed with lung cancer a week after enrollment? We know it takes years from the first malignant cell division to the time that the cancer is detectable. Do you exclude the subject who is diagnosed with cancer a week after enrollment? If you say yes, what about the subject diagnosed a month after, or a year after? The longer the time, the murkier the decision.

TIP

Develop rigorous pre-enrollment criteria to ensure that subjects do not have the outcome at the time the study starts.

As with most things, prevention is the best defense. To avoid this problem, develop rigorous pre-enrollment criteria to ensure that subjects do not have the outcome at the time the study starts (at least as best as can be determined). Staying with the example of lung cancer, you may want subjects to have a respiratory symptom review and a pre-enrollment chest x-ray.

Unfortunately, certain diseases are difficult to rule out without subjecting participants to very invasive tests (which would increase the expense of your trial and decrease enrollment). For example, some HIV-infected patients have *pneumocystis carinii* pneumonia (PCP) with minimal or no symptoms and normal chest x-rays. If you wanted to be sure that subjects do not have PCP prior to enrolling them in a PCP-prevention trial, you would have to perform bronchoscopy on all of them. However, it is not feasible or ethical to subject asymptomatic persons to an invasive test prior to enrollment in a trial to prevent the disease. Instead, most investigators performing studies on preventing PCP limit the pre-enrollment evaluation to a chest x-ray and symptom review. Invariably, a few patients are diagnosed with PCP just days after enrollment.

Besides being difficult to diagnose, PCP usually develops slowly, over a period of weeks. If a patient is diagnosed with PCP a week after starting a treatment protocol designed to prevent PCP, should the patient be considered a treatment failure (since the outcome of interest occurred while the patient was on the study) or should the subject be deleted from the analysis (since the subject almost certainly had PCP at the time of enrollment)? This is a judgment call. What most investigators do is to exclude those cases of PCP that occur within twenty-eight days of enrollment.[6] Cases that occur after twenty-eight days are considered treatment failures.

In considering this example you may wonder: Would it not be safer to include people who develop PCP after enrollment in the study no matter how soon after the start date? In support of this, remember that in a randomized controlled trial implausibly early outcomes should be evenly distributed in the different arms of the study. Therefore, including these early-outcome subjects will not bias your

[6] Leoung, G. S., Feigal, D. W., Montgomery, A. B., *et al.* "Aerosolized pentamadine for prophylaxis against *pneumocystis carinii* pneumonia." *N. Engl. J. Med.* **323** (1990): 769–75. Golden, J. A., Katz, M. H., Chernoff, D. N., *et al.* "A randomized comparison of once-monthly or twice-monthly high-dose aerosolized pentamadine prophylaxis." *Chest* **104** (1993): 743–50.

TIP

Even if you have a-priori criteria for exclusion it is best to have the decision to exclude a subject made by a review committee that is blind to the treatment assignment.

analysis, although it will result in your reporting higher treatment failure rates in the different arms of the study. But, in observational studies, improbably early outcomes would not necessarily be evenly distributed in the different arms of your study and could thus be a source of bias in your study.

My general advice in this area is develop pre-enrollment criteria that will lower the chance of implausibly early outcomes. Beyond this, decide ahead of time what you will do if a subject develops the outcome of interest a day after your study begins. If it will be important to you to exclude such early outcomes, develop objective exclusion criteria for subjects prior to the start of a study. Even if you have a-priori criteria for exclusion it is best to have a review committee that is blind to the treatment assignment make the decision to exclude a subject.

At times, it may be worth excluding early outcomes as a way of testing a hypothesis on the cause-and-effect relationship between your risk factor and outcome. For example, in the study of cholesterol level and mortality discussed in Section 5.5, the investigators excluded cancers that occurred in the first four years of the study. They did this to test whether low cholesterol levels might be a consequence of cancer that was present but unsuspected at the time of entry into the cohort. When they excluded these cases, the relationship between low cholesterol level and cancer persisted, suggesting that the relationship between the low cholesterol level and mortality was not a consequence of unsuspected cancer at the time of enrollment.

8.7 What are variable selection techniques?

DEFINITION

Variable selection techniques are procedures that determine which independent variables will be included in the model.

Variable selection techniques are automatic procedures that determine which independent variables will be included in a multivariable model. They also can determine the order in which the variables enter the model. The parameters of the algorithms are determined by the investigator.

I have already referred to variable selection techniques as a flawed strategy for decreasing the number of independent variables in your analysis. This may be necessary because of an insufficient sample size for the number of independent variables in your model (Section 7.5). The other major reason for using selection procedures is that you want to determine the minimum number of independent variables necessary to accurately estimate outcome. This is particularly important in the development of diagnostic and prognostic models (Sections 2.1.C and 2.1.D) because the fewer the variables the more likely clinicians are to remember and use them.

In Section 2.1.C, I detailed a decision rule for determining which patients presenting with chest pain to an emergency room were probably having acute ischemia. The investigators used forward stepwise-regression to create the

Table 8.5 Methods of variable selection.

Type of selection technique	Method	Advantages and disadvantages
Forward	Enters variables into the model sequentially. The order is determined by the variable's association with outcome (variables with strongest association enter first) after adjustment for any variables already in the model.	Best suited for dealing with studies where the sample size is small. Does not deal well with suppresser effects.
Backward	Deletes variables from the model sequentially. The order is determined by the variable's association with outcome (variables with weakest association leave first) after adjustment for any variables already in the model.	Better for assessing suppresser effects than forward selection.
Best subset	Determines the subset of variables that maximizes a specified measure.	Computationally difficult.
None (all variables)	Enters all variables at the same time.	Including all variables may be problematic if there are many independent variables and you have a small sample size.

TIP

The fewer the variables included in a clinical prediction rule the more likely clinicians are to remember and use it.

prediction rule. Using forward selection they evaluated a total of fifty-nine clinical characteristics; the selection algorithm chose the seven variables that best accounted for ischemia. If instead the investigators developed a model using all fifty-nine characteristics, it would undoubtedly have had better diagnostic capability than the seven-variable model. But what clinician would use a fifty-nine variable model in a clinical setting? Patients would require hospital admission just so that their physician would have enough time to record the values of the fifty-nine clinical characteristics and compute each patient's probability of ischemia!

Most statistical software packages offer a variety of variable selection techniques (Table 8.5). What all selection methods have in common is that they use statistical criteria to decide which variables should enter the model and the order of the variables entering the model.

Using forward selection, the model will select the variable most strongly related to the outcome and enter it first into the model. In fact, you can predict which variable will enter first in a forward-selection model by looking at your bivariate analysis. The variable with the strongest association with your outcome in the bivariate analysis will enter first. You will not be able to predict the second variable that will enter simply by looking at the bivariate analysis because the

model will choose the variable that best improves the fit of the model after adjusting for the first variable. This may not be the variable with the second strongest association with outcome in the bivariate analysis. It depends how closely these two independent variables are related to each other. If they are very closely related, it is possible that once you know the value of the first variable, the value of the second variable does not substantially improve the fit of your model. Instead a variable that is less strongly associated with outcome in the bivariate analysis, but unrelated to the first variable that entered, may be the second strongest variable in improving the fit of your model.

Forward selection will continue to evaluate each variable for how it improves the fit of your model. When none of the remaining variables significantly improves the fit, it will stop entering variables. You as the researcher must decide what statistical cutoff to use for determining that the addition of another variable does not significantly improve the fit of the model. With lower cutoffs fewer variables will be included, but you will be more likely to miss important confounders. With higher cutoffs you will be less likely to miss important confounders but you will have a model with more variables in it.

Forward models can be modified to allow you to delete variables that were significant on entry into the model but are not statistically significant after other variables have entered. To do this, you will need to specify a statistical cutoff for removal of a variable that was already entered. You may want to set a less stringent (higher *P* value) cutoff to remove a variable once entered, or use the same cutoff for both. Forward selection models with deletion of entered variables that are no longer significant will produce a model with potentially fewer variables than simple forward selection.

Backward selection is similar to forward selection – except it proceeds backwards! At step one all variables enter into the model. If you have ten independent variables, all ten will enter in this step, no matter how unrelated they are to outcome. The algorithm then assesses which of the ten variables in the model is least important in accounting for the outcome, and deletes it, so that there are now nine variables in the model. The model then assesses which of the nine is least important in accounting for the outcome. It deletes this variable and then repeats the process until all the remaining variables are significantly associated with the outcome. At this point no further variables are deleted. As with forward selection, the researcher determines what statistical cutoff will be used for retaining (or not deleting) a variable.

You may, at first, think that forward or backward selection would arrive in the same place just by different routes, like two cars converging on a city from opposite directions. While it is a sign of a robust model when forward and backward selection give you the same answer, this does not always occur. The

TIP

Backward selection is more likely to demonstrate a suppresser effect than forward selection.

reason that forward and backward selection do not necessarily produce the same answer is that the importance of a particular variable often depends on what other variables are in the model at the time of selection. A variable may be statistically important when a variable (or a group of variables) is in the model and yet not significant when that variable (or group of variables) is not in the model. This is referred to as a suppresser effect (Section 1.3). In forward selection, it is less likely that the variable needed to demonstrate the suppresser effect would be in the model. For this reason, backward selection is more likely to detect a variable that is significant only when the suppresser variable is in the model.

TIP

Forward selection is preferable to backward selection when your sample size is small for the number of independent variables in your analysis or when you have concerns about multicollinearity.

Forward selection is preferable over backward selection when your sample size is small for the number of independent variables in your analysis or when you have concerns about multicollinearity. This is because in backward selection all the variables are in the model initially. If you have doubts about the reliability of a model with all the variables in it, then there is reason to worry about having this model be the starting point for decisions on which variables to delete.

In best subset regression, the computer chooses the best combination of variables from all possible models. In the case of an analysis with only five variables, there are thirty-one possible combinations of variables (including models that have one, two, three, four, and five variables). The number of possible combinations increases exponentially as you increase the number of possible variables. "Best" is determined by a specified measure of the ability of the model to account for the outcome. For example, in multiple linear regression, you could have the computer choose the combination of variables that produces the highest adjusted R^2 (Section 9.2.B).

In some ways, best subset regression is hard to argue with. It is, after all, the best statistical answer to the question. However, because of the computational time involved, this technique cannot always be done. For logistic regression and proportional hazards analysis, best subset regression is usually modified to include the best possible combination of two variables, then of three variables, then of four variables, up to the maximum number of variables in your model (some programs will limit the maximum number of variables to ten). This is a simplification in that the computer is not comparing models of different size to one another (e.g., comparing those that have five variables to those that have four variables). Also, just because it is the best statistical answer does not mean it truly reflects the physiology of what you are studying. Confounders may be included in the model while the main effects are missing. Conversely, confounders that may change the coefficients of certain variables in important ways may be omitted.

Although I have described these selection algorithms as distinct, there are many hybrids. A popular hybrid is to enter certain variables in the model at the start of the analysis and not allow them to be deleted even if they are not

significantly related to the outcome. The computer then enters the remaining independent variables in a forward or backward manner. This strategy works well when there are certain variables that you absolutely want in your analysis for theoretical or practical reasons. For example, if every prior analysis of your outcome shows that age is an important confounder, it makes sense to add age right at the start, and not allow it to leave the model.

Forward and backward selection techniques can be modified to minimize the effect of missing data in your analysis. With forward or backward selection, after you have derived your model, you can rerun it with only the variables that entered (or were not deleted). By rerunning the model with a smaller number of variables, missing data on the excluded variables will no longer result in missing cases in the multivariable model. With backward selection, you can rerun the model with each deletion of a variable, so that each iteration of the model has a larger sample size. With both forward and backward selection techniques you will need to specify what level of statistical significance should result in the inclusion or exclusion of a variable. Most researchers use a P value of <0.05 or, for smaller sample sizes, a P value of <0.10 or <0.15. However, just because a variable does not meet the P value criterion does not mean that it is unimportant. The algorithm does not evaluate whether entry of the variable changes the coefficients of the other variables in the model. While it is unlikely that a variable that has no association with outcome will make a significant impact on the other coefficients, it is possible that a variable that is marginally associated with the outcome will change the coefficients of the other independent variables in important ways. This is one reason many researchers favor higher (less restrictive) cutoffs than $P < 0.05$.

TIP

Don't use selection algorithms!

Now that you understand the different types of selection algorithms, I have one more piece of advice: If at all possible DON'T USE THEM! With all of the variable selection procedures you run the risk of the model eliminating (or not selecting) a variable that is on the causal pathway to your outcome, in favor of a variable that is a confounder. Because forward and backward selection algorithms evaluate variables singly, there is a possibility that your final model will not include two variables that together are important in changing your main effect. Also, a variable may be very important in explaining an outcome, and yet get kicked out of the model because it is related to a variable that is already in the model.

Therefore, you (not some computer algorithm) should determine, based on your theoretical understanding and your empirical findings, what variables to include in your model. Without a variable selection algorithm, all variables that you specify will be entered simultaneously (this is sometimes referred to as forcing all variables into the model). You will not have to worry about the

possibility of missing suppresser effects or important changes in coefficients caused by exclusion of a modest confounder. Another advantage of all variable models is that when you submit your paper for publication, you will not have to explain why certain variables are not included in your model. While one can certainly defend exclusion of a variable in selection models because it was not statistically related to the outcome, the reviewer may cite the possibility that the missing variable is a modest confounder or a suppresser variable. With all your variables entered into your model you can demonstrate that the variable of interest is included and does or does not affect the outcome and the relationship of the other independent variables to the outcome.

As with other rules of thumb, there is an exception. It's okay to use a selection algorithm if your goal is to identify the best possible diagnostic model with the fewest number of variables. With predictive models for diagnosis and prognosis (Sections 2.1.C and 2.1.D) we are not concerned with causality. If knowing the value of a risk factor enables you to diagnose accurately a condition it does not matter whether the variable is or is not causally associated with the condition. Also, diagnostic models are much more likely to be used by busy clinicians if they have a small number of variables.

8.8 What value should I specify for tolerance in my logistic regression or proportional hazards model?

DEFINITION

Tolerance is a measure of multicollinearity.

Tolerance is a measure of multicollinearity (Section 6.2). Very small values of tolerance indicate multicollinearity. If you have highly related independent variables in your analysis, the computer may be unable to solve the equation without deleting one or more of the problematic variables from the model. The default criteria set by your software package should work fine. In practice, this feature mainly works to delete completely redundant variables, such as when you create multiple dichotomous variables and include in the model a variable for the reference category (Sections 4.2 and 6.2).

8.9 How many iterations (attempts to solve) should I specify for my logistic regression or proportional hazards model?

The answer to this question is similar to the answer attributed to Abraham Lincoln when asked how long a man's legs should be: Long enough to reach the floor. You should set the number of iterations high enough to solve your equation (referred to as convergence of your model). A typical default value used by software packages is twenty-five attempts. It may take more attempts than that, especially if you have a small number of cases or skewed distributions

When a higher number of iterations is required to get a model to converge there is usually a problem with the model.

for your independent or dependent variables (e.g., only three people are yes on a variable). Although the default will generally work, if it does not, increase the iterations until the model converges (solves the equation). However, note that when a large number of iterations is required to get the model to converge there may be a problem with the model (Section 8.11).

8.10 What value should I specify for the convergence criteria for my logistic regression or proportional hazards model?

In multiple linear regression the parameter estimates can be found by solving an explicit system of equations. However, with logistic regression and proportional hazards models no such explicit solution exists. Instead, the maximum likelihood parameter estimates are found through a search algorithm. Starting from an initial rough solution, the algorithm modifies the estimates and recomputes the likelihood. The algorithm is determined to have converged on the maximum likelihood parameter estimates when the modifications increase the likelihood by less than the preset convergence criterion, or when the largest modification to any single parameter estimate falls below an analogous criterion. The defaults set by your software package should work fine.

8.11 My model won't converge. What should I do?

You may sometimes get a message that your logistic or proportional hazard models won't converge. What this means is that the computer cannot solve the equation. There are several reasons this can happen. In the simplest cases, you have made an error in your coding. If you have coded an outcome variable, such that everyone has the same outcome (this can happen if you are not careful with your if/then statements), the computer cannot solve the equation. It cannot compute the odds of outcome versus no outcome if everyone has the outcome.

TIP

Models will not converge if you have defined a subgroup in which no (or very few) outcomes have occurred.

Recognizing this, you can probably imagine another reason that a model will not converge: You have too few outcomes for the number of independent variables in your model (Section 7.4). Your independent variables (e.g., smoking status, gender, age) may be defining subgroups for which there are no outcomes. For example it may be that among nonsmoking women under the age of forty-five years no heart attacks have occurred. Because this group has no members, the computer cannot determine the parameters for the variables of smoking status, gender, and age.

What can you do if your model won't converge and your outcome variable is correctly coded? You should check to see if your independent variables define any subgroups with no outcomes. If this is not the case, you can increase the number

TIP

If your model does not converge even after increasing the number of iterations, try decreasing the number of independent variables in your model.

of attempts (iterations) the program makes to solve the equation. If increasing the iterations does not work, try decreasing the number of independent variables in your model so that there are no subgroups with very few outcomes. Removing independent variables that have very skewed distributions, especially less than 5 percent of subjects in a particular category, usually helps the most.

9

Interpreting the analysis

9.1 What information will the printout from my analysis provide?

All three multivariable techniques will provide two kinds of information: Information about the relationship of all independent variables taken together to the outcome, and information about the relationship of each of the independent variables to your outcome variable (with adjustment for all other independent variables in your analysis). Let's review these in turn.

9.2 How do I assess how well my model accounts for the outcome?

As you can see in Table 9.1, there are a variety of methods for assessing how well your model accounts for your outcome.

9.2.A How do I know if my model (all my independent variables together) accounts for outcome better than I would expect by chance?

All three types of analyses provide a test of whether the independent variables, taken together, are more strongly associated with outcome than would be expected by chance. As is the case with all inferential statistics, you seek to disprove the null hypothesis. The null hypothesis in this case is that there is no relationship between the independent variables and the outcome. A significant relationship between the independent variables and the outcome means that you can reject the null hypothesis.

In multiple linear regression, the F test compares the success of the independent variables in accounting for the outcome compared to the success in accounting for the outcome based on assuming that everyone in the study had the mean value for the outcome.[1] If knowing the values of the independent

[1] To see how *F* is calculated, see Glantz, S. A. *Primer of Biostatistics* (4th edn). New York, NY: McGraw-Hill, 1997, pp. 37–52.

Table 9.1 Methods for measuring how well a model accounts for the outcome.

	Multiple linear regression	Multiple logistic regression	Proportional hazards analysis
Model accounts for outcome better than chance (Section 9.2.A)	F test	Likelihood ratio test	Likelihood ratio test
Quantitative/qualitative assessment of how well model accounts for outcome (Section 9.2.B)	R^2	R^2 (rarely used), comparison of estimated to observed value, Hosmer–Lemeshow test	Comparison of estimated to observed value
Prediction of outcome (Section 9.2.C)	Not applicable	Sensitivity, Specificity, Accuracy, c index	Not applicable

variables improves the fit more than would be expected by chance, then the value of F will be large. A large F value for a given sample size and a given number of variables in the model (which determines the degrees of freedom) will result in a small *P* value. This indicates that the null hypothesis of no association between the independent variables and the outcome can be rejected.

If knowing the values of the independent variables improves the fit more than would be expected by chance, the null hypothesis can be rejected.

For multiple logistic regression and proportional hazards analysis, the test for assessing the significance of the overall model is the likelihood ratio test (often referred to as model chi-square). It is analogous to the F test. It has a chi-squared distribution.

For logistic regression the likelihood ratio test answers the question of whether the independent variables account for the outcome better than assuming the mean outcome for your subjects. Since logistic regression has a dichotomous outcome, the mean is simply the proportion of persons who experience an outcome. If knowing the values of the independent variables improves the fit of the model more than you would expect by chance, then the value of the chi-square will be large.

For both logistic regression and proportional hazards analysis, when the chi-square of the likelihood ratio test is large, the *P* value will be small and the null hypothesis can be rejected.

With proportional hazards analysis the likelihood ratio test answers a somewhat different question. With proportional hazards analysis, if the time to outcome of subjects with certain values on their independent variables are different from the baseline rate (more than you would expect by chance), then the chi-square will be large.

For both logistic regression and proportional hazards analysis, when the chi-square of the likelihood ratio test is large, for a given number of parameters in the model (degrees of freedom), the *P* value will be small. As with a large

TIP

If individual variables are significantly related to outcome but your overall model is not statistically significant, delete those variables that are not associated with outcome.

F test value, you will reject the null hypothesis and conclude that the independent variables are related to the outcome. The *P* value associated with the chi-square assumes a "large" sample size; sample sizes greater than 80–100 give a good approximation.[2]

Some limitations of the F and the likelihood ratio test will be apparent to you. They do not tell you which or how many variables in your model are significant. You can have a model with five variables in it, four of which are unassociated with the outcome, and still have a significant overall test. Methods for determining the statistical significance of the individual variables are dealt with in Section 9.3.

Another limitation is that knowing that the variables as a group are more closely associated with outcome than you would expect by chance does not tell you quantitatively how well your independent variables account for outcome. Given these limitations, what are these tests useful for? If you get a global F or likelihood ratio test that is not significant, you should worry that your model is a poor representation of your data. If individual variables are significantly related to outcome, and the overall model is not statistically significant, it suggests that there are many variables included in the model that are unrelated to outcome. With each added variable, the degrees of freedom increase, and it takes a larger chi-square value to achieve a statistically significant result. Consider deleting variables from your model that are not associated with outcome.

9.2.B How do I assess how well my model accounts for the outcome?

DEFINITION

R^2 is a quantitative measure of how well the independent variables account for the outcome.

A quantitative measure of how well the independent variables explain the outcome in multiple linear regression is R^2. Also called the coefficient of variation, R^2 indicates how much better you can account for the outcome by knowing the values of the independent variables than by assuming that everyone had the mean value on the outcome variable.

The value of R^2 ranges from 0 (indicating that the independent variables do not explain the outcome any better than assuming everyone has the sample mean) to 1 (the independent variables completely account for the outcome). When R^2 is multiplied by 100 it can be thought of as the percentage of the variance in the dependent variable explained by the independent variables.

TIP

When R^2 is multiplied by 100 it can be thought of as the percentage of the variance in the dependent variable explained by the independent variables.

While R^2 is generally more informative than F, it has the limitation that its value will increase as you include additional independent variables, even if these variables add only a little bit of information. For example, a model with ten

[2] For a more detailed explanation of the likelihood ratio test, see: Hosmer, D. W. and Lemeshow, S. *Applied Logistic Regression*. New York, NY: Wiley, 1989, pp. 8–18; Menard, S. *Applied Logistic Regression Analysis*. Thousand Oaks, CA: Sage Publications, 1995, pp. 19–21.

Adjusted R^2 charges you a price for each variable in your model.

independent variables will have a higher R^2 than a model with five of these variables, even if the additional five variables add little to the model. To account for this, the statistic adjusted R^2 charges you a price for each variable in your model. As you add variables, adjusted R^2 can increase (the gain in having the variable is greater than the charge), decrease (the charge is greater than the gain), or stay the same (the gain and the charge are equal).

Although there is an R^2 measure for logistic regression, it does not perform as well for this type of analysis and is rarely reported in the literature. However, there are other methods for assessing how well a logistic model accounts for the outcome.

TIP

A useful qualitative method for assessing a logistic regression model is to compare the estimated probability of outcome to the observed probability of outcome.

A useful qualitative method for assessing a logistic regression model is to compare the estimated probability of outcome (according to the model) to the observed probability of outcome (the original data). To do this remember that the estimated probability is based on the pattern of independent variables for each subject. If your model has three variables, gender (male/female), age (in terciles), and hypertension (yes/no), then the number of distinct patterns of these three variables is $2 \times 3 \times 2 = 12$. In other words, whether you have 15 subjects or 15 million there are only twelve distinct patterns. For each of these patterns, there is an observed rate of outcome (the proportion of persons who experienced the outcome based on the data) and an estimated rate of outcome (based on the model).

If you have a large number of distinct covariate patterns, you can still use this technique by grouping patients with similar estimated likelihood of outcomes. Thus, for example, you may divide your sample into ten groups of estimated likelihood of outcome: 0–0.10, 0.11–0.20, etc. One problem with dividing estimated likelihood of outcome into equal divisions of likelihood is that you may still get small groups, if, for example, few persons have very high or very low estimated probabilities of outcome. Another alternative is to divide the probabilities such that there are approximately equal numbers of outcomes in each group.

The latter was done by Gordon and colleagues in a study of racial variation in predicted and observed in-hospital death rates.[3] The investigators used logistic regression to develop an estimated probability of in-hospital death. In their model, they included age, sex, race, type of health insurance, emergency department admission, and a mortality measure based on data from the first forty-eight hours of hospitalization. They divided the estimated risk of death into ten strata so that there would be equal numbers of outcomes in each group (653–654 deaths). Note that since the strata are based on the number

[3] Gordon, H. S., Harper, D. L., and Rosenthal, G. E. "Racial variation in predicted and observed in-hospital death." *JAMA* **276** (1996): 1639–44.

Table 9.2 Comparison of estimated to observed risk of death among hospitalized patients.

Stratum	Estimated risk of death	Observed risk of death
1	0.00–0.03	0.01
2	0.03–0.06	0.05
3	0.06–0.10	0.09
4	0.10–0.17	0.14
5	0.17–0.24	0.24
6	0.24–0.34	0.32
7	0.34–0.47	0.38
8	0.47–0.63	0.52
9	0.63–0.81	0.66
10	0.81–1.00	0.87

Adapted with permission from Gordon, H. S., *et al.* "Racial variation in predicted and observed in-hospital death." *JAMA* **276** (1996): 1639–44. Copyright 1996, American Medical Association.

of outcomes rather than the estimated risk of death, the different strata have varying widths of estimated risks of death (stratum 1 ranges from only 0.00 to 0.03, whereas stratum 10 ranges from 0.81 to 1.00). You can see that the estimated risk of death was similar to the observed probability of death in the ten strata (Table 9.2).

Although the data can be shown in tabular form, as was done by the authors in their article, the fit of the model can be seen a little better in Figure 9.1. I have created the figure from the data by plotting the midpoint of the estimated probability of death on the *x*-axis against the observed probability of death along the *y*-axis. You can see that the points (connected by a solid line) are all close to the dotted diagonal line, which represents perfect calibration.

If the points fall close to the diagonal, as in Figure 9.1, your model is an excellent estimate of outcome. If the points are scattered far from the line, it indicates that the model is not very accurate at estimating observed outcomes. An advantage of this approach is that it also allows you to see if your model performs better at certain probabilities of disease.

For proportional hazards analysis, it is possible to compare estimated and observed time to outcome. This can be done using Kaplan–Meier survival curves for each important subgroup of patients defined by your model. For example, Colford and colleagues found in their proportional hazards analysis that two variables, CD4 count and hematocrit (both split at the median), had the strongest association with survival among HIV-infected patients with

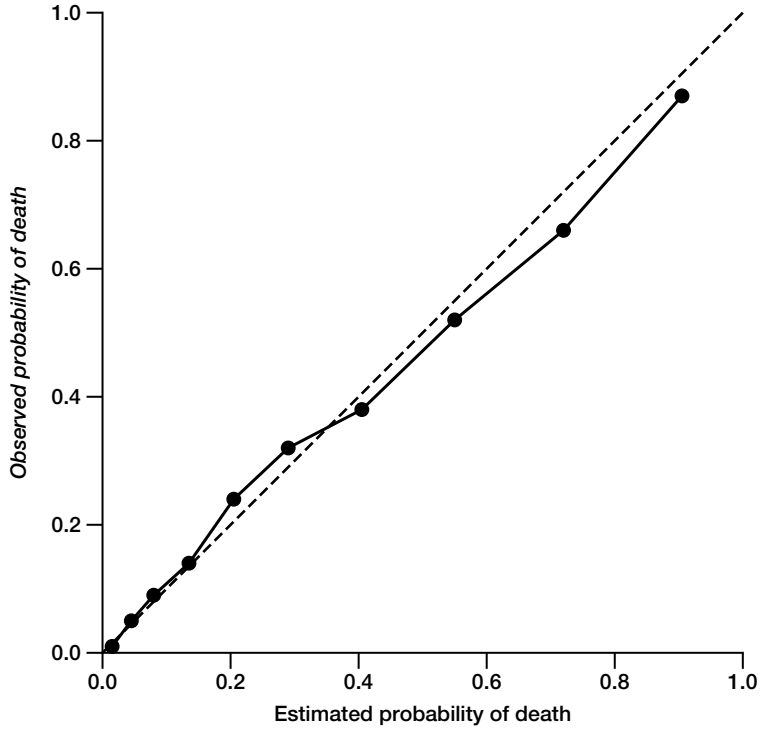


Figure 9.1 Estimated (*x*-axis) versus observed values (*y*-axis) for risk of death among hospitalized patients. Data from Gordon, H. S., *et al.* "Racial variation in predicted and observed in-hospital death." *JAMA* **276** (1996): 1639–44.

cryptosporidiosis.[4] To assess how well their model estimated survival, they stratified their patients into four groups based on CD4 count and hematocrit. As shown in Table 9.3, they found that the estimated and observed median survival times were similar.

Because the underlying survival function is not automatically estimated in proportional hazards analysis, you need to use adjunct estimators to calculate estimated median survival. Also, this procedure will not work if there are few subjects who experienced the outcome. (To calculate an estimated median survival, half of the sample for each covariate pattern must experience the outcome.)

DEFINITION

The *Hosmer–Lemeshow goodness-of-fit test* compares the estimated to observed likelihood of outcome for groups of subjects.

Comparing the estimated to observed probability, whether in tabular form (Table 9.2) or graphed (Figure 9.1), is a qualitative assessment of how well the model accounts for the outcome. There is a statistic that has been developed for logistic regression to assess how similar the estimated probability of

[4] Colford, J. M., Tager, I. B., Hirozawa, A. M., *et al.* "Cryptosporidiosis among patients infected with human immunodeficiency virus: Factors related to symptomatic infection and survival." *Am. J. Epidemiol.* **144** (1996): 807–16.

Table 9.3 Survival among subgroups of HIV-infected patients with Cryptosporidium.

Subgroup	Relative hazard	95% confidence interval	Median survival (days)	
			Estimated	Observed
CD4 count ≤53 cells/ml and hematocrit ≤37%	15.9	6.0–42.2	213	204
CD4 count ≤53 cells/ml and hematocrit > 37%	8.1	2.8–23.6	465	341
CD4 count >53 cells/ml and hematocrit ≤37%	3.1	1.1–8.8	688	878
CD4 count >53 cells/ml and hematocrit > 37%	1.0 (ref.)		1,119	1,119

Adapted with permission from Colford, J. M., *et al.* "Cryptosporidiosis among patients infected with human immunodeficiency virus." *Am. J. Epidemiol.* **144** (1996): 807–16.

outcome is to the observed probability of outcome. This statistic is called the Hosmer–Lemeshow goodness-of-fit test. The statistic compares the estimated to observed likelihood of outcome for groups of subjects. The groups are created by dividing the sample into approximately ten groups based on the range of estimated probability of outcome (the first group contains the 10 percent of subjects with the lowest estimated likelihood of outcome, the second group contains the 10 percent of subjects with the next lowest estimated likelihood of outcome, etc.). In a well-fitting model, the estimated likelihood will be close to the observed likelihood of outcome. This will result in a small chi-square and a non-significant *P* value.

9.2.C How can I assess how well my model predicts the outcome of study subjects?

Because the outcomes in logistic regression are dichotomous, one can ask how well the model predicts the outcome of study subjects. To do this you must dichotomize the estimated outcome. In other words, you must choose a cutoff of what estimated probability of outcome you will consider to be a prediction of outcome. Once you do this, you can compute the sensitivity (proportion of persons who are predicted to have the outcome who really have it), specificity (proportion of persons who do not have the outcome who are predicted not to have it), and the proportion of correctly identified persons.

Choosing the cutoff for probability of outcome is not always easy. Each cutoff has a different sensitivity, specificity, and proportion of correctly identified

persons. One simple cutoff is to assume that anyone with probability of outcome greater than 50 percent is predicted to have the outcome. However, the choice of 0.5 as the cutoff for measuring the predictive ability of your model may not be best. This is especially true for clinical diseases (e.g., ischemic heart disease) where even relatively low probabilities of disease are worrisome because of the seriousness of the disease. For example it would not be appropriate to send a patient with chest pain home who had a 49 percent probability of having acute ischemia. For this reason, models predicting acute ischemia choose a much lower cutoff, such as 7 percent for computing sensitivity and specificity.[5]

DEFINITION

The *c index* is a measure of the concordance between predicted and observed outcomes. The higher the value the greater the ability of your model to predict outcome.

Another useful measure of how well your logistic regression model predicts outcome is the c index.[6] It is a measure of the concordance between predicted and observed outcomes. Here's how it works. In any data set there will be pairs of subjects who have the same observed outcome (e.g., both have had heart attacks, neither have had heart attacks) and some who have different outcomes (e.g., one had a heart attack, one did not). For each pair of subjects with different outcomes, one can ask whether the model predicts a higher likelihood of outcome for the subject in the pair who experiences the outcome or for the subject who does not experience the outcome. If the subject with the higher predicted likelihood of outcome actually experiences the outcome the pair is concordant (with outcome). If the case with the higher predicted likelihood of outcome does not have the outcome, the pair is discordant. If the model predicts the same probability of outcome for both cases, the pair is tied. The c index equals the proportion of concordant cases plus half of the ties. A value of 0.5 would indicate that the model does not discriminate any better than chance. The higher the c value is (maximum 1) the greater the ability of your model to predict outcome.

9.3 What do the coefficients tell me about the relationship between each variable and the outcome?

DEFINITION

A variable's *coefficient* tells you how the outcome changes with changes in the independent variable.

In all three types of models, a variable's coefficient (also called beta) tells you how the outcome changes with changes in the independent variable, while adjusting for the other independent variables in the model. Coefficients can be positive or negative. Because the three different models have different types of outcome variables, there are differences in how these coefficients are interpreted (Table 9.4).

[5] Goldman, L., Cook, E. F., Brand, D. A., *et al.* "A computer protocol to predict myocardial infarction in emergency department patients with chest pain." *N. Engl. J. Med.* **318** (1988): 797–803.

[6] Harrell, F. E., Lee, K. L., Matchar, D. B., *et al.* "Regression models for prognostic prediction: Advantages, problems, and suggested solutions." *Cancer Treat. Rep.* **69** (1985): 1071–7.

Table 9.4 Interpretation of regression coefficients.

	Multiple linear regression	Multiple logistic regression	Proportional hazards analysis
Coefficient is positive	Average value of the outcome increases as independent variable increases	The logit of the outcome increases with increases in the independent variable	The logarithm of the relative hazard increases with increases in the independent variable
Coefficient is negative	Average value of the outcome decreases as independent variable increases	The logit of the outcome decreases with increases in the independent variable	The logarithm of the relative hazard decreases with increases in the independent variable

In multiple linear regression, the dependent variable is interval. This gives the coefficient a special property: It is the slope of the line describing the relationship of the independent variable to the outcome. For example, let's assume that you are interested in the association between the independent variable, age (measured in years), and the outcome variable, cholesterol level (measured in mg/dl), and the coefficient for age is 0.2. The units for the coefficient would be mg/(dl year). This would mean that for each year the cholesterol increases by 0.2 mg/dl. To draw a line showing the best estimated value for all possible values of your dependent variable, all you need is the slope and a point on the *y*-axis (where *x* is zero). This point is provided in linear regression by the intercept (which you will find on your printout).

With linear regression, a positive coefficient indicates that the independent variable and the outcome variable are moving (up or down) together. A negative coefficient indicates that the independent variable and the dependent variable are moving in opposite directions. So, if the coefficient is positive then the independent variable increases (e.g., age goes from 20 years to 30 years) as the average value of the outcome increases (e.g., cholesterol goes from 200 to 220 mg/dl). Conversely, a negative coefficient indicates that as the independent variable increases (e.g., age goes from 20 years to 30 years), the average value of the outcome decreases (e.g., cholesterol goes from 200 to 180 mg/dl).

The meaning of the coefficient in logistic regression is somewhat different from its meaning in linear regression. The coefficient tells you how a one-unit change in the independent variable changes the logit, which you will remember is the natural logarithm of the odds of the outcome (Section 5.2). A positive coefficient means that as the variable increases, the logit increases. A negative coefficient means that as the variable increases, the logit decreases.

To interpret the meaning of the coefficients in logistic regression, you must know which value of the outcome the logit is estimating. The default on most

TIP

Make sure you know for which value of your outcome variable the computer is estimating the logit.

software programs is to determine the logit for the lower numerical value (determined by how you have coded your variables). But you could ask the computer to determine the logit of the higher numerical value. Either way, the results would be the same but the signs of the coefficients would be different. Make sure you know for which value of your outcome variable the computer is estimating the logit.

In proportional hazards analysis, the coefficients tell you how much a one-unit change in the independent variable changes the logarithm of the relative hazard. The relative hazard is the ratio of time to outcome given a particular set of risk factors to time to outcome without these factors (Section 5.2).

The meaning of the signs of the coefficients in proportional hazards analysis is similar to that in logistic regression. A positive coefficient indicates that as the independent variable increases, the logarithm of the relative hazard increases. A negative coefficient indicates that as the independent variable increases, the logarithm of the relative hazard decreases.

9.4 How do I get odds ratios and relative hazards from the multivariable analysis? What do they mean?

DEFINITION

The *odds ratio* (for logistic regression) and the *relative hazard* (for proportional hazards analysis) are equal to the antilogarithm of the coefficient.

The coefficients in logistic regression and proportional hazards analysis have a special value. If you take the antilogarithm of the coefficient, you will obtain the odds ratio (for logistic regression) and the relative hazard (for proportional hazards analysis). This is simply the mathematical constant e raised to the power of the coefficient's value:

$$\text{odds ratio or relative hazard} = e^{\text{coefficient}}$$

For logistic regression, the odds ratio tells you how much the likelihood of the outcome changes with a one-unit change in the independent variable. For proportional hazards analysis, the relative hazard tells you how much the time to outcome changes with a one-unit change in the independent variable. Odds ratios and relative hazards of "1" indicate that there is no change in outcome with changes in the independent variable; values greater than 1 indicate an increase in risk of outcome, while values less than 1 indicate a decrease in risk.

In published work, you will see many other terms used for the odds ratio and the relative hazard. For odds ratio, you will see relative risk, risk, and risk ratio. For relative hazard, you will see all of the terms mentioned for odds ratio plus hazard ratio and rate ratio. Odds ratio (for logistic regression) and relative hazard (for proportional hazards analysis) are the preferred terms because they distinguish results from these two types of analysis. The odds ratio can

Table 9.5 Computing odds ratios and relative hazards.

	Multiple logistic regression coefficient	Odds ratio	Proportional hazards analysis coefficient	Relative hazard
Hypertension (yes/no)	0.693	2.0	0.693	2.0
Female gender (yes/no)	−0.693	0.5	−0.693	0.5
Age (in years)	0.182	1.2	0.182	1.2

be considered an approximation of the relative risk when the outcome is rare (<15 percent).

The coefficients in Table 9.5 are from a hypothetical logistic regression analysis (first column) and from an equally hypothetical proportional hazards analysis (third column). The outcome for the logistic regression analysis is heart attack at three years; the outcome for the proportional hazards analysis is time to heart attack. The coefficient for the variable history of hypertension is 0.693 in both analyses. If you take 0.693 to the *e* you get 2.0, meaning an odds ratio of 2.0 and a relative hazard of 2.0. (To get 2.0 enter 0.693 in your calculator, and press the button with the little *e* on it. If your calculator has no *e* button, buy a new calculator.)

What do these results mean? For the logistic regression analysis, it means that persons with hypertension are twice as likely as nonhypertensive persons to have a heart attack at three years. For proportional hazards analysis, it means that persons with hypertension are twice as likely to have a heart attack during the three-year study period than persons without hypertension. Note how similar these computations and interpretations are between logistic regression and proportional hazards analysis.

Let's look at the next variable, female gender. It has a negative coefficient. The odds ratio is 0.5. What does this mean? It means that women are half as likely as men to have a heart attack at three years (logistic regression) and over the three-year study period (proportional hazards analysis).

How can you tell that it isn't men who are half as likely as women to have a heart attack? Actually, you cannot tell from Table 9.5. But you may assume (correctly) that I followed my own advice (Section 8.1) and coded 1 (presence of the condition = being female) and 0 (absence of the condition = not being female). Since the coefficient is negative, it means that the outcome is less likely when the independent variable increases (from 0 to 1).

What would happen if you changed the coding of the gender variable such that 0 was for women and 1 was for men? The sign on the coefficient would change from negative to positive and the coefficient would stay the same. The odds ratio

would change from 2.0 to 0.5. Has the meaning changed? First version: Women are half as likely as men to have heart attacks (odds ratio = 0.5). Second version: Men are twice as likely as women to have heart attacks (odds ratio = 2.0). The meaning is the same.

I took you through the example of gender in some detail for two reasons. First, I wanted to illustrate that the sign of the coefficient tells you whether the odds ratio will be > 1 (positive sign) or < 1 (negative sign). Second, I wanted you to see that you could change the coding of your dichotomous variable without changing the meaning of the analysis. Finally, I wanted to warn you how easy it is to misinterpret your results if you forget how you have coded your variables.

At the risk of public embarrassment, I will admit to you that I once completed a manuscript and circulated it to my coauthors before realizing that I had confused the coding of my variable. My finding was the opposite of what I reported in the draft of the paper. Fortunately, I found the mistake before I submitted the manuscript for publication. The reason I made the mistake (and I have seen this in other people's work as well) was that I saw what I wanted to see.

I was assessing the factors associated with receipt of mental health treatment among depressed persons with HIV disease.[7] I assumed that unemployment would be associated with being less likely to receive mental health services (because people wouldn't have the money to pay for treatment). Therefore, when I saw that the odds ratio for employment was 0.6, it made sense to me. It was only while performing additional analyses that I realized I had coded employment as 1 and unemployment as 0. My outcome was coded as mental health services received (1) and no mental health services received (0). My model was estimating the likelihood of receiving mental health services. Therefore, employment was associated with a decreased likelihood of receiving mental health services. It may be that employed persons receive fewer mental health services because they are less likely to be recognized as depressed by their clinicians or themselves (since they are working), or because of greater concerns of confidentiality.

You can see in Table 9.6 that for a dichotomous independent variable and a dichotomous outcome there are actually four possible codings. The results all mean the same statistically, but if you become confused (as I did) as to how you coded your variables, you can report the wrong finding.

Researchers rarely report in their manuscripts how they have coded their variables. Thus it is unlikely that a miscoded variable will be discovered in peer review. It is up to you to make sure you are reporting your results correctly.

[7] Katz, M. H., Douglas, J. M., Bolan, G. A., *et al.* "Depression and use of mental health services among HIV-infected men." *AIDS Care* **8** (1996): 433–42.

Table 9.6 Coding of a dichotomous independent variable and outcome.

Possible coding 1	Possible coding 2
Unemployed = 0; Employed = 1	Unemployed = 1; Employed = 0
No treatment = 0; Treatment = 1	No treatment = 0; Treatment = 1
Possible coding 3	**Possible coding 4**
Unemployed = 0; Employed = 1	Unemployed = 1; Employed = 0
No treatment = 1; Treatment = 0	No treatment = 1; Treatment = 0

TIP

Name your variables as specifically as you can.

Besides reviewing your work carefully, there are other strategies for minimizing the chance of reporting a result opposite to what your data show. First, name your variables as specifically as you can within the limits of what your statistical packages will allow (usually up to eight characters). For example, it is better to name your variable "femgend" than "gender."

It is also important to code your independent and dependent variables in a sensible way. As recommended in Section 8.1, code 0 as absence of the condition and 1 as presence. A variable such as "employ" should be coded as 1 for employment and 0 for unemployment. (I actually did this but still managed to become confused.) Alternatively, create a variable called "unemploy" and code it as 1 for unemployment and 0 for employment.

Another useful strategy is to use value labels. The computer will print out the values you have assigned each time you use the variables in the analysis. You are less likely to make a mistake if you see on your printout "1 = employment, 0 = unemployment" next to the variable. Entering value labels takes a bit of extra time at the start of your analysis, but it is worth the effort.

9.5 How do I interpret the odds ratio and relative hazard when the independent variable is interval?

Look back at Table 9.5. The variable under female gender is age. Age is an interval variable measured in years. If you take the coefficient (0.182) to the *e*, you will see that the odds ratio and relative hazards are 1.2. What does this mean? It means that, for every increase of one year in age, the likelihood of having a heart attack increases by a factor of 1.2. What would the odds ratio or relative hazard be if you coded your variable age in five-year blocks (ages 20–24, 25–29, 30–35, etc.), as many researchers do? It would be 2.5. Does it surprise you that the odds ratio and relative hazard change so much by changing the unit of measurement? Does this large change in odds ratio change the interpretation any? No. If a one-year change in age increases the risk by 1.2, then a five-year change will increase the risk by $(1.2 \times 1.2 \times 1.2 \times 1.2 \times 1.2) = 2.5$.

TIP

To evaluate the importance of an odds ratio or relative hazard for an interval variable you must know the units in which the independent variable is measured.

The take-home lesson here is: You cannot evaluate the importance of an odds ratio or relative hazard for an interval variable without knowing what units the independent variable is measured in. This comes up often with variables such as age and blood pressure, where a single unit change may be associated with an odds ratio or relative hazard very close to 1 (e.g., 1.2), yet the variable has a large effect on outcome (when considered over the range of values for that independent variable).

9.6 How do I compute the confidence intervals for the odds ratios and relative hazards?

Confidence intervals tell you the range of plausible values for odds ratios and for relative hazards. They also give you a measure of the precision of your values. Large confidence intervals suggest that your sample size is insufficient for the analysis you are performing (Section 7.4).

The 95 percent confidence interval for the odds ratio and the relative hazard are easily obtained by extension of the formula for computing the odds ratio and relative hazard. Most statistical packages will automatically compute the confidence intervals, but some don't, and it is handy to know how to calculate it, in case you ever need to. To obtain the upper confidence interval use the addition sign; to obtain the lower confidence interval use the subtraction sign. The standard error is usually next to the coefficient on the computer printout.

$$\text{95\% confidence interval for odds ratio/relative hazard} = e^{\text{coefficient} \pm 1.96(\text{standard error})}$$

Looking at the formula you can also see how the precision of your estimate (measured by the standard error) is reflected in the confidence intervals. If the standard error is large you will be adding (or subtracting) a large number from the coefficient. This will result in the upper limit being much bigger than the odds ratio/relative hazard and the lower limit being much smaller than the odds ratio/relative hazard.

Of course, you don't have to use 95 percent confidence intervals. For some exploratory studies you may wish to report 90 percent confidence intervals. For other studies, where precision is very important, you may wish to report 99 percent confidence intervals. The formula is the same as the one shown above except that instead of 1.96, which is the standard normal deviate for 95 percent confidence intervals, you substitute the standard normal deviate for the confidence intervals you want (1.66 for 90 percent confidence intervals; 2.576 for 99 percent confidence intervals).

9.7 What are standardized coefficients and should I use them?

TIP

To compare the magnitude of coefficients of independent variables that are measured on different scales, standardize them.

Standardized regression coefficients are your regression coefficients multiplied by the standard deviation for that independent variable and divided by the standard deviation of the dependent variable:

$$\text{standardized regression coefficient} = \text{regression coefficient} \times \frac{\text{standard deviation of independent variable}}{\text{standard deviation of dependent variable}}$$

Unless you standardize your coefficients you cannot compare the magnitude of coefficients of independent variables that are measured on different scales. If you divide your independent variable by ten (e.g., switch from age in years to age in decades), it will increase your coefficient by ten. Obviously dividing a variable by ten does not in any way change the association between that variable and the outcome. However, if you divide the variable by ten, not only will the coefficient increase tenfold, so will the standard deviation. By multiplying the coefficient by the ratio of the standard deviation of the independent variable to the standard deviation of the dependent variable, the coefficient becomes unitless (the units cancel each other out). You can therefore compare across different independent variables because they are all on the same scale. Of course, if all of your independent variables are on the same scale, as will be true if all your variables are dichotomous, then standardization is unnecessary. The downside of standardized coefficients is that you lose a sense of the actual effect (in real units) that each variable has on the dependent variable.

The above formula works for both multiple linear regression and logistic regression. The only difference is that the standard deviation of the dependent variable in logistic regression is not a calculated value but is the constant 1.81. With multiple linear regression, the square of the standardized coefficient also gives you a sense of the proportion of variance explained by that variable (like a partial R^2). This is not true with logistic regression. Standardized coefficients are rarely reported with proportional hazards models.[8]

9.8 How do I test the statistical significance of my coefficients?

The smart answer to this question is that you read the *P* value next to the variable on the printout. Although this is perfectly true, it is worth understanding where these *P* values come from.

If you understand the difference between unstandardized and standardized regression coefficients (Section 9.7), then you know that you cannot judge the

[8] For more on standardized coefficients see Feinstein, A. R. *Multivariable Analysis: An Introduction*. New Haven, CT: Yale University Press, 1996, pp. 222–5, 322, 330, 391.

Table 9.7 Coefficients and standard errors from study of factors associated with receipt of PCP prophylaxis.

Variable	Coefficient	Standard error
Age ≥ 35 years	0.1696	0.2846
Nonwhite ethnicity	−0.7173	0.2922
Male gender	−0.2162	1.2910
Gay men	1.1592	0.7183
Injection drug users	0.1068	0.4221
No insurance	−1.0356	0.3656

Adopted with permission from Schwarcz, S. K., *et al.* "Prevention of *Pneumocystis carinii* pneumonia: Who are we missing?" *AIDS* **11** (1997): 1263–8. Copyright Rapid Sciences Publishers Ltd. Additional data supplied by the authors.

strength of the association between an independent variable and an outcome by the size of the coefficient (because the size is determined by the size of your units). But you can eyeball whether a coefficient is significant by looking at the unstandardized coefficient and the standard error of the coefficient. How?

Well, remember that a statistical model only *estimates* the true value of the parameter in the population you are studying. There is an error associated with that estimate (the standard error). Common sense tells you that if the size of the standard error is similar to the (absolute) size of the coefficient (the error is as big as the effect), the effect won't be significant.

However, if the coefficient is much bigger than the standard error, the coefficient may be statistically significant. Look at Table 9.7. The coefficients and the standard errors are from the study of factors associated with receipt of PCP prophylaxis (Section 7.4). Without looking back at the discussion of the study results, can you tell which of the variables are statistically significant?

If you said the variables nonwhite ethnicity and no insurance, you are right. You can tell because the standard errors for these variables are much smaller than the coefficient for these variables. Note that the coefficient is larger for the variable gay men than for the variables nonwhite ethnicity and no insurance. Yet this variable is not significant because the coefficient is not large relative to its standard error. Also note that I did not need to remind you that these coefficients were from logistic regression. That's because it doesn't really matter. With any of the three multivariable techniques, a coefficient that is more that twice the size of the standard error is likely to be statistically significant. In Table 9.8, I have included the *P* values as well as the odds ratios and the confidence intervals.

Table 9.8 Coefficients, standard errors, *P* values, and confidence intervals for receipt of PCP prophylaxis.

Variable	Coefficient	Standard error	*P* value	Odds ratio	Upper 95% confidence interval	Lower 95% confidence interval
Age $\leq$ 35 years	0.1696	0.2846	0.55	1.19	0.68	2.07
Nonwhite ethnicity	−0.7173	0.2922	0.01	0.49	0.28	0.87
Male gender	−0.2162	1.2910	0.87	0.81	0.06	10.12
Gay men	1.1592	0.7183	0.11	3.19	0.78	13.03
Injection drug users	0.1068	0.4221	0.80	1.11	0.49	2.55
No insurance	−1.0356	0.3656	0.005	0.35	0.17	0.73

Adopted with permission from Schwarcz, S. K., *et al.* "Prevention of *Pneumocystis carinii* pneumonia: Who are we missing?" *AIDS* **11** (1997): 1263–8. Copyright Rapid Science Publishers Ltd. Additional data supplied by the authors.

TIP

If the error is bigger than the effect, the effect cannot be a very reliable estimate.

Some other points are worth noting from Table 9.8. In Section 7.4, I pointed out that because of the small number of women in this study, the confidence intervals for the odds ratio for the gender variable were very large. You can see this just by looking at the coefficient and the standard error. Note that the standard error is about six times the size of the coefficient. It stands to reason that if the error is bigger than the effect, the effect cannot be a very reliable estimate. You can also see that the two variables that are statistically significant are also the ones where the confidence intervals for the odds ratios exclude 1. This should serve as a reminder to you that confidence intervals depend on similar assumptions as for *P* values (i.e., the size of the effect compared to the size of the error).

The methods for computing the *P* values are similar for linear regression, logistic regression, and proportional hazards analysis. For multiple linear regression, the *P* value is based on a *t* test, where *t* is

$$t = \frac{\text{coefficient}}{\text{standard error}}$$

You can determine the significance of *t*, if you know the degrees of freedom (sample size – the number of parameters, where the parameters are the independent variables plus the intercept). You can then look up the significance of the *t* value in the tables that are at the back of most standard statistical books (but not this one). More likely you will read it off your printout. But, of interest, *t* values greater than 2.0 (coefficient is two times the standard error) are statistically significant at the traditional $P < 0.05$ value (as long as the degrees of freedom are at least sixty).

Logistic regression and proportional hazards analysis use a similar test, called

TIP

Use the Wald chi-square for testing the significance of individual coefficients from logistic regression and proportional hazards analysis.

the Wald test, for determining the importance of an individual coefficient. It is based on either the chi-squared or the *z* distribution:

$$\text{chi-squared distribution} = \left\{\frac{\text{coefficient}}{\text{standard error}}\right\}^2 \quad \text{or} \quad z \text{ distribution} = \frac{\text{coefficient}}{\text{standard error}}$$

With either formula, coefficients that are twice their standard error will be significant at $P < 0.05$. The Wald test assumes a large sample size (i.e., 80–100 or more subjects).

For logistic regression and proportional hazards analysis there are two other tests that you can use to determine the statistical significance of a particular coefficient: the likelihood ratio test and the score test.[9] The likelihood ratio test is based on comparing the likelihood when the variable is not in the model to the likelihood when the variable is in the model. To compute the test statistic for models where you have multiple independent variables, each variable is singly dropped while retaining the other variables in the model. The statistic follows a chi-squared distribution. Computation of the score test is based on derivatives of the likelihood ratio. It also follows a chi-squared distribution.

Although there are situations when the likelihood ratio test may perform better,[10] most researchers use the Wald chi-square. If your results are robust you should get similar results with all three tests.

9.9 How do I interpret the results of interaction terms?

Having reviewed the meaning of coefficients, let's return to the question of how to interpret product terms used to represent interaction effects (Section 8.3). If the impact of the two variables together is substantially greater than the additive effect of the two variables, the coefficient will be positive and statistically significant. If the impact of the two variables together is substantially less then the additive effect of the two variables, the coefficient will be negative and statistically significant.

9.10 Do I have to adjust my multivariable regression coefficients for multiple comparisons?

To answer this complicated question it is best to consider first the simpler case of bivariate analyses. Let's say, for example, that you are assessing the association of

[9] You met the likelihood ratio test in Section 9.2.A on evaluating whether your model accounts for outcome better than chance. The difference is that here you are comparing models with a particular variable present to models where that variable is absent (but the other independent variables are present). In Section 9.2.A this test compared models that contained all of the independent variables to models that contained none of the independent variables.

[10] Hauck, W. W. and Donner, A. "Wald's test as applied to hypotheses in logit analysis." *J. Am. Stat. Assoc.* **72** (1977): 851–3.

age (30–59 years, 60–79 years, 80–99 years) on cholesterol levels using analysis of variance. In addition to being interested in the three-way comparisons, you are also interested in a pairwise comparison of the youngest group to the oldest group. Intuitively, it makes sense that if you compare three groups, one group will be highest and one group will be lowest. Therefore, if you compare the highest and the lowest group, you are running a risk that you are capitalizing on chance. One way to deal with this issue is not to consider pairwise comparisons unless the overall F (for the comparison of the three groups) is significant. In addition, you should set a more stringent cutoff for pairwise comparisons before rejecting the null hypothesis. This is usually done using the Bonferroni correction. It "charges" you for the number of pairwise comparisons by requiring a lower P value before concluding that a comparison is statistically significant. To calculate the correction, simply divide the usual P value (e.g., 0.05) by the number of pairwise comparisons you are performing. If you are performing three pairwise comparisons, you would reject the null hypothesis only if $P \leq 0.016$ $(0.05/3 = 0.016)$.

DEFINITION

The *Bonferroni correction* "charges" you for the number of comparisons performed by requiring a lower *P* value before concluding that a comparison is statistically significant.

When you perform multiple bivariate comparisons (for example, comparing two groups on twenty different variables), some statisticians also recommend adjusting for multiple comparisons. The theory is that, by chance, at least one of your twenty comparisons will be statistically significant at the $P < 0.05$ level (that's because $1/20 = 0.05$). However, instead, what most investigators do is perform a multivariable analysis. With multivariable analysis, you need not worry about multiple comparisons when performing tests of the significance of the overall model (F or likelihood ratio test). The reason is that you are performing only a single test to assess whether the independent variables (as a group) are associated with the outcome.

But when you turn to the question of whether the individual independent variables from your multivariable model are statistically associated with your outcome, you are essentially making multiple comparisons. As with bivariate comparisons, some statisticians advocate adjusting your P value for the number of independent variables in your model. If you have ten variables, you would require that the P value be < 0.005 (0.05/10) before concluding that the association between the independent variable and the outcome is significant.

However, there are major disadvantages to adjusting for multiple comparisons. They have been well articulated by the epidemiologist Kenneth Rothman.[11] He points out that the basis of adjustment for multiple comparisons is the assumption that chance is the most common explanation for an association between two things. But this assumption is flawed because the universe is governed by natural (physical) laws. Most associations in the universe have a true

[11] Rothman, K. J. "No adjustments are needed for multiple comparisons." *Epidemiology* **1** (1990): 43–6.

(rather than chance) connection. (Note that a true connection does not mean a causal connection.)

In addition, an individual comparison cannot "know" how many other comparisons you have made. Therefore an individual association cannot be more or less likely to be caused by chance based on how many other associations you have assessed. Rothman illustrates the absurdity of strict adherence to the principle of adjusting for multiple comparisons by asking: If you favor adjusting for multiple comparisons, should you adjust for the number of comparisons you assessed in a single paper, or the number of comparisons assessed in a series of papers analyzing the same data set, or the number of comparisons performed during your career?

TIP

To minimize potential problems with multiple comparisons, tell your reader how many variables were tested in your analysis, report significant and nonsignificant results, and don't be a slave to the cutoff of $P < 0.05$.

Personally, I am swayed by Rothman's arguments and do not adjust for multiple comparisons with multivariable models. Nonetheless, if you are not going to adjust for multiple comparisons, there are measures that you should take to minimize potential problems. Tell your reader how many variables (comparisons) were tested in your analysis. Do not report only the independent variables that were significantly associated with outcome. Nonsignificant results, while much less sexy, are every bit as informative. Do not be a slave to the cutoff of $P < 0.05$, in either direction. Don't assume something is insignificant just because its P value is 0.06 or that something is significant just because its P value is 0.04. Use confidence intervals, whenever possible, instead of P values (although, as discussed above, confidence intervals are also subject to the potential multiple comparison problem, since they are based on the 95 percent probability that 95 percent of repeated samples of the population would produce 95 percent confidence intervals that would contain the true value). Most importantly, evaluate your findings in the light of previous work and biological plausibility. If you anticipate that a reviewer will not be convinced by the above, cite Rothman's article. It sometimes helps.[12]

[12] For more on the debate about multiple comparisons, see: Savitz, D. A. and Olshan, A. F. "Multiple comparisons and related issues in the interpretation of epidemiologic data." *Am. J. Epidemiol.* **142** (1995): 904–8; Thompson, J. "Re: 'Multiple comparisons and related issues in the interpretation of epidemiologic data.'" *Am. J. Epidemiol.* **147** (1997): 801–6; Goodman, S. N. "Multiple comparisons explained." *Am. J. Epidemiol.* **147** (1997): 807–12; Savitz, D. A. and Olshan, A. F. "Describing data requires no adjustment for multiple comparisons: A reply from Savitz and Olshan." *Am. J. Epidemiol.* **147** (1997): 813–14; Thompson, J. R. "A response to 'Describing data requires no adjustment for multiple comparisons.'" *Am. J. Epidemiol.* **147** (1997): 815.

10

Checking the assumptions of the analysis

10.1 How do I know if my data fit the assumptions of my multivariable model?

In Chapter 5, we reviewed the ways to assess, in bivariate analysis, whether the relationship between a single independent variable and outcome fit the assumptions of your model. In this chapter, we will review how to assess whether these assumptions are fulfilled on a multivariable level.

For didactic purposes, I have separated the discussion of whether your data fit the assumptions of your model from the discussion of how well your model accounts for the outcome (Section 9.2). But, in fact, these two topics are closely related. In a model where your independent variables closely account for your outcome, it is likely that your data fit the assumptions of your model. Conversely, if your model does not appear to fit your observed outcome, the reason might be that your data do not fit the assumptions of the model. When this is the case, adding, deleting, or transforming variables may improve the fit of your model. Part of why I separated the discussion of these two issues is that most researchers will interpret their multivariable printouts, make changes, and add or delete variables before going on to verify the assumptions of their final model.

DEFINITION

Residuals are the difference between the observed and the estimated value.

Residual analysis is a helpful tool for assessing whether your data fit the assumptions of your multivariable model. Residuals are the difference between the observed and the estimated value. They can be thought of as the error in estimation. A good model will have most residuals close to 0, meaning that the observed and estimated values are close to one another. When the observed is greater than the estimated value the residual is positive; when the observed value is less, the residual is negative.

Besides helping you to test if your data fit the assumptions of your model, residual analysis can also help you to identify individual subjects whose values on the outcome variable do not fit with the other subjects (outliers).

I mention in passing that while residual analysis is valuable it is more art than science. You may get alarming patterns of residuals even though your data fit

Residual analysis is more art than science.

the assumptions of your model. It is also possible for your residuals to look fine, and yet your data do not fit the assumptions of your model. Small samples are especially likely to yield messy residuals. With large samples, multivariable models are sufficiently robust that departures from the assumptions of the model, seen on residual analysis, are unlikely to cause significant problems.

10.2 How do I assess the linearity, normal distribution, and equal variance assumptions of multiple linear regression?

Multiple linear regression assumes linearity, normal distribution, and equal variance around the mean (Chapter 5). To test whether your interval-independent variables fit a linear relationship with outcome in your multivariable model, plot your residuals against each of your independent variables and the estimated outcome variable. If the relationship is linear, the points will be symmetric above and below a straight line, with roughly equal spread along the line. In contrast, if residuals are particularly large at very high and/or low levels of one of the independent variables or estimated dependent variable, it suggests a significant departure from linearity.[1]

Plots of residuals against any of the independent variables and the estimated dependent variable also demonstrate whether your model fulfills the assumptions of normal distribution and equal variance. When these assumptions are met the residuals should be close to zero and the spread of values should be equal both above and below zero. If, instead, the residuals are far from zero and there is not an equal spread of values above and below zero (e.g., a few of the residual points are far from the zero line indicating outliers), the assumptions of normal distribution and equal variance are not met.

Besides plotting the residuals against each of the independent variables and the estimated dependent variable, it is also possible to test the assumptions of multivariable normal distribution by plotting standardized residuals on a normal probability plot. If the residuals are normally distributed with equal variance, the plot will appear as a straight line. If, instead, you see a nonlinear pattern, such as an S-shape, this suggests that the assumptions of normal distribution are not fulfilled. Points far from the line indicate outliers.

If your plots do not fit the assumptions of linear regression, don't give up. Look back in the section on transforming variables (Sections 5.5 and 5.8). It may be that by transforming one of your independent or dependent variables you will achieve a better model. Also, if your sample size is greater than 100, you can

[1] For a detailed description of analysis of residuals including numerous graphs see Glantz, S. A. and Slinker, B. K. *Primer of Applied Regression and Analysis of Variance*. New York, NY: McGraw-Hill, 1990, pp. 110–80.

assume that the assumption of normal distribution is met for your independent variables (Section 5.7).

10.3 How do I assess the linearity assumption of multiple logistic regression and proportional hazards analysis?

As with multiple linear regression, residual analysis can help you determine if your interval-independent variables show a linear relationship with outcome. To do this, plot your residuals or standardized residuals against the independent variable. If the linear assumption is met, the residuals should be about the same for all values of your independent variable. Larger residuals as the independent variable gets larger (or smaller or both) suggest that your model does not fit the linear assumption.

Another method of assessing the linear assumption of interval variables is to create multiple dichotomous variables of equal intervals of your variable. I explained this technique in Section 5.4. The only difference here is that you will be using it to assess whether your independent variable fits the linearity assumption after adjustment for other independent variables. If the numeric difference between the coefficients of each successive group is approximately equal, this is consistent with a linear gradient.

Finally, some researchers will assess whether an interval variable has a linear gradient by substituting a transformed version of the variable for the variable itself. Commonly, logarithmic and quadratic transformations are tested. If the coefficient for that variable increases and the fit of the model improves with the transformation, this would suggest that the variable more closely fits a logarithmic or quadratic gradient.

You should be aware that with logistic regression you may get residuals that appear disturbing even if your model is sound. This is especially likely to occur when you have strong dichotomous (rather than interval) independent variables.

10.4 What are outliers and how do I detect them in my multiple linear regression model?

Outliers are points that do not follow the pattern of the other points. For example, Figure 10.1 shows a linear relationship with higher values of the independent variable associated with higher values of the outcome. While most of the points conform to this linear relationship, two points (A and B) do not fit this relationship. Point A has a much higher value for outcome than you would expect given the intermediate value on the independent variable. Point B has a much lower

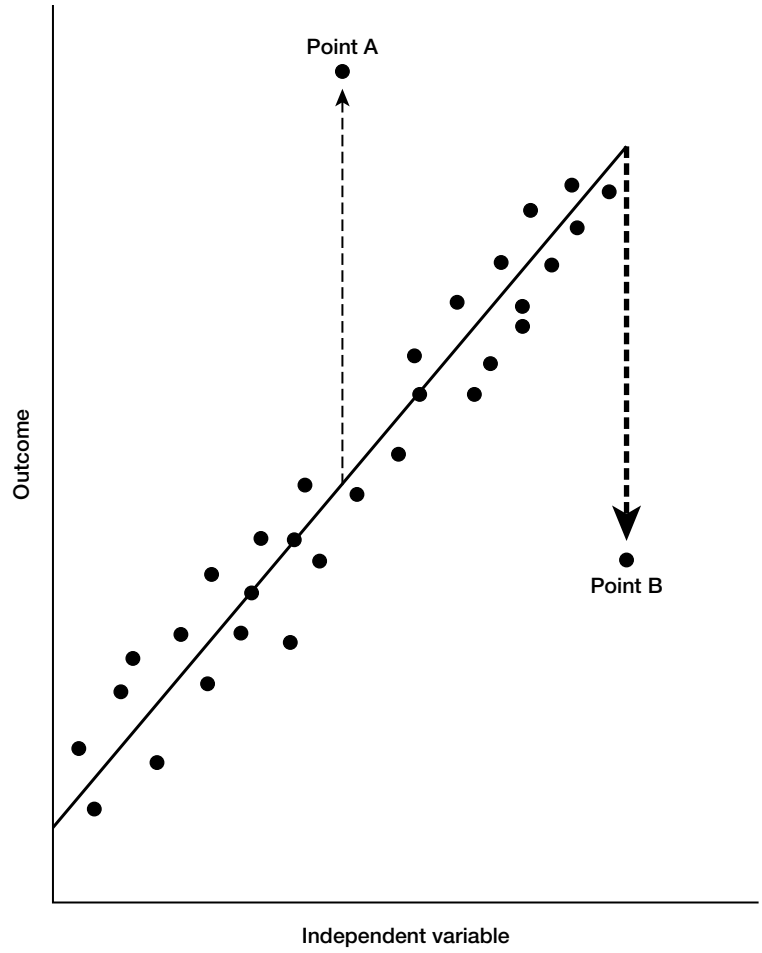


Figure 10.1 Linear relationship between an independent variable and outcome with two outliers (point A and point B). The thicker dotted line pointing to B illustrates the greater influence of point B compared to point A.

value of outcome than you would expect given the high value of the independent variable.

In addition to graphical presentations, outliers can be detected in linear regression by several residual measures that are calculated by most statistical packages: standardized residuals, studentized residuals, leverage, and Cook's distance.

Standardized residuals help pinpoint outliers. Standardized residuals larger than two are the extreme 5 percent of values, whereas those larger than three are the extreme 1 percent of values.

TIP

Outliers at the extremes of the value of the independent value are more influential than those closer to the midpoint.

Studentized residuals adjust for how far each individual value is from the center of the line. The result is that two points an equal distance from the line will have different studentized residuals: The studentized residual will be larger for the value at the extremes of the line than the value at the center. This is because values at the extremes are likely to be influential. They can more easily tilt the line. Think of a seesaw, with the center point as the fulcrum and the extreme points as the ends of the plank. Exerting pressure on the end will cause the entire plank to change slope. Such points are referred to as leverage points and are more influential in the analysis. With studentized residuals, leverage points will get a larger residual than an outlier whose value is close to the center of the line. So in Figure 10.1 point B will have a larger studentized residual than point A even though these points are the same distance from the line. I have used a thicker dotted line in Figure 10.1 to point to B to illustrate the greater influence of point B compared to point A.

The measure leverage quantifies the influence of individual points on the line. Ideally, all your observations should have leverage measures less than two times the expected value. The expected value is the number of variables +1 divided by the sample size. Cook's distance offers another way of telling how influential an individual point is on the overall model. It is equal to the change in the regression coefficients if the observation was deleted.

10.5 How do I detect outliers in my multiple logistic regression model?

Many of the same statistics that detect outliers in multiple linear regression are available for multiple logistic regression, including raw residuals, standardized residuals, studentized residuals, leverage, and Cook's distance. In addition, the Pearson and deviance residuals are particularly useful for assessing multiple logistic regression models.

As with linear regression, standardized or studentized residuals greater than 2–3 suggest outliers. Subjects with leverage values exceeding twice the expected value (number of independent variables divided by the sample size) are very influential. Cook's value greater than 1 also suggests a point of heavy influence.

A helpful test both for assessing outliers as well as the overall fit of the model is to plot the Pearson and deviance residual against the estimated probability of outcome. This will show you whether the model is less successful in estimating the outcome at particular probabilities.[2]

[2] For a detailed review of residuals with logistic regression see Hosmer, D. W. and Lemeshow, S. *Applied Logistic Regression*. New York, NY: Wiley, 1989, pp. 149–70.

10.6 What about analysis of residuals with proportional hazards analysis?

Residual analysis is less often used with proportional hazards analysis than with multiple linear regression and multiple logistic regression. Readers interested in learning more about their use should see cited references.[3]

10.7 What should I do when I detect outliers?

Let's say that your analysis of residuals suggests you have certain extreme values. What do you do? First, check to make sure there is no error in the recording of this data point. You may think that this is an unnecessary step if you have already reviewed your univariate results for extreme values (Section 5.7). But review of outliers from multivariable analysis complements the univariate analysis. For example, residual analysis of a multivariable model may detect an obese diabetic with a heavy fat consumption whose cholesterol is abnormally low (140 mg/dl). This may result in your discovering a data-entry error: The value was really 410 mg/dl. Since a cholesterol level of 140 mg/dl is not an extreme value you would not have noticed a problem in the univariate analysis.

But what if you verify from the original data that this subject's value really is 140 mg/dl? Do you delete it from the data set? No. If you were performing a laboratory experiment, and had a residual value that did not fit your analysis, you might want to repeat the experiment. But in clinical trials this is rarely an option. While it may be tempting to delete such values from the study (especially if they are preventing you from getting the answer you were hoping for) this is rarely justified. Just because certain subjects are outliers, it doesn't mean their values are wrong. In fact, it is "normal" to have a few extreme values.

However, if your residuals indicate leverage points, you may want to consider removing them from the analysis, repeating the analysis, and seeing whether your findings hold. If deletion of a couple of observations changes your entire analysis, the analysis may not be valid. In general, the larger your data set is, the less likely it is that your results will be heavily influenced by one or two points. Therefore, the need to closely examine the residuals of all of your points becomes less important.

[3] Farrington, C. P. "Residuals for proportional hazards models with interval-censored survival data." *Biometrics* **56** (2000): 473–82; Kalbfleisch, J. D. and Prentice, R. L. *The Statistical Analysis of Failure Time Data*. New York, NY: Wiley, 1980, pp. 96–8; Nardi, A. and Schemper, M. "New residuals for Cox regression and their application to outlier screening." *Biometrics* **55** (1999): 523–9.

10.8 What is the additive assumption and how do I assess whether my multiple independent variables fit this assumption?

All three multivariable models assume that your multiple independent variables have an additive effect on the outcome.

In the case of multiple linear regression, the change in the mean value of the outcome is modeled as the sum of the individual effects of the independent variables on outcome (Section 5.3). While both logistic regression and proportional hazards analysis have the same additive assumption, it is more confusing with these two techniques because with logistic regression and proportional hazards analysis we are not modeling the outcome itself. Rather we are modeling the logit of the outcome and the logarithm of the relative hazard, respectively.

TIP

The effect of multiple variables on the odds ratio or the relative hazard is multiplicative.

With logistic regression, the logit of the outcome is modeled as the sum of the individual independent variables. With proportional hazards analysis, the logarithm of the relative hazard is modeled as the sum of the individual independent variables. The consequence of this is that the odds ratio or relative hazard for the effect of multiple variables on outcome is multiplicative, rather than additive. Statisticians refer to this with the somewhat confusing term of "additive on a multiplicative scale." Although I didn't refer to it as the additive assumption, we have dealt with this concept in the discussion of interactions (Sections 1.4, 8.3, and 9.9). Interactions are present when the variables are not additive but rather something greater or less than additive.

If you look back at the discussion of the interaction of gender and ST elevations shown in Section 1.4, you see that I multiplied the odds ratios to show the meaning of the interaction. Specifically, in Table 1.7, the odds ratio for male gender was 1.6 and the odds ratio for ST elevations was 8.1. Since the model is "additive on a multiplicative scale" the odds ratio associated with being male and having ST elevations should be the product of the two or 13.0 ($1.6 \times 8.1 = 13.0$).

But the risk of heart attack for men with ST elevations was not 13 times the risk for women without ST elevations. How do we know this? Because the product term was statistically significant. To find out the true risk for men with ST elevations, you need to include the interaction term, which had an odds ratio of 0.6. When you include the product term you learn that the risk for men with ST elevations is less than 13.0; it is actually only 7.8 ($1.6 \times 8.1 \times 0.6 = 7.8$). After reading Section 9.9, you would have known that the overall risk was lower just from the negative sign of the coefficient of the product term.

What does all this mean? Clinically, as discussed in Section 1.4, it means that the difference in the likelihood of heart attacks between men and women vanishes in the presence of ST elevations. (The risk for women with ST elevations is 8.1 ($1.0 \times 8.1 \times 1.0 = 8.1$), essentially identical to the risk of 7.8 in men with ST

elevations ($1.6 \times 8.1 \times 0.6 = 7.8$). In terms of your model, this means that in the absence of including an interaction term for gender and ST elevations, your model would have been misspecified because the additive assumption would not have been fulfilled.

TIP

The best coefficient for the sample as a whole may not be best for every subgroup of subjects.

You may be asking yourself, why does the model without interaction terms estimate that men with ST elevations have a significantly higher risk of heart attack than women with ST elevations, if the risk for these two groups is similar? To understand why, remember that each variable (e.g., gender, ST elevations) has only one coefficient. Therefore, the best coefficient for the sample as a whole may not be best for every subgroup of subjects as defined by the independent variables. In other words, just because men have a greater risk of heart attack than women and persons with ST elevations have a greater risk than persons without ST elevations does not mean that men with ST elevations have a greater risk than women with ST elevations. This illustrates how models without product terms can be wrong for particular subsets of subjects. Product terms solve this problem by having another variable that can improve the fit of the model for particular subgroups of subjects.

Assessing whether your variables fulfill the additive assumption becomes especially complicated when you realize that in most analyses there are a large number of possible interaction terms. For example there are forty-five possible two-way interactions between just ten independent variables. And there is nothing to prevent interactions from being three-way (e.g., male $\times$ ST elevations $\times$ prior heart attack). Short of trying all possible product terms, there is no way to know for certain if your data contain an important interaction.

One clue indicating that you may need an interaction term is that a variable that you thought, based on clinical grounds, would have an important effect on an outcome variable did not. Could it be that the variable is important only under certain conditions? Other clues may come from your analysis of how well your data fit the assumptions of your model. For example, if the standardized residuals do not form a straight line on the normal probability plot or if your logistic regression residuals are particularly large, the reason may be that you have an interaction.

Besides the fact that interactions are difficult to detect there are other problems. Some statistically significant interactions are very difficult to interpret clinically. In addition, if you test a large number of interactions, it is likely that at least one of them will be statistically significant. Does that mean that this is an important interaction?

In this regard, testing for interaction is similar to performing subgroup analysis. Let's say that a study finds that a new drug is no better than placebo in the sample as a whole. However, when the investigators looked within ten subgroups,

they found one group (e.g., hypertensive women with diabetes) for whom the drug worked. Would you conclude that the drug works for hypertensive women with diabetes?

In general, unless there is some reason to believe that the drug should work only in hypertensive women with diabetes, and the authors had therefore planned this subgroup analysis, one would conclude that the drug does not work and the finding was chance. That is, in some subgroups, the drug worked better than placebo. In other subgroups, the drug worked worse than placebo. Overall, there was no effect. The same issue exists if you test ten product terms. If the main effect is null, and nine of the product terms are insignificant, and one of them is significant, do you conclude that overall the drug works in that special condition? Probably not.

Because of these problems, many clinical researchers do not test for interactions at all. In contrast, it is common in the epidemiologic literature to evaluate all possible two-way interactions. The methodological justification is that if a product term is significant, it is improving the statistical quality of the model. However, this strategy is not usually pursued in the medical literature because of the difficulty of making clinical sense out of product terms.

My own preference is to test only for those interactions that are theoretically important. That is the strategy that was pursued in the study described above of the impact of gender on heart attack risk. The researchers evaluated all possible interactions involving gender because it was known that the variables associated with heart attack are different in men and women and because this was the focus of their study. But they did not assess all possible interactions in their data set (e.g., age $\times$ race).

Whatever strategy you employ for assessing interactions, it is important to tell your reader whether and how you have tested for interactions. Depending on what strategy you have chosen, you can tell your reader that you:

- tested all primary (second-degree) interactions, or
- tested specific interactions (and detail them), or
- chose not to test any interactions

10.9 What does the additive assumption mean for interval-independent variables?

The meaning of the additive assumption for interval-independent variables is analogous to its meaning for multiple independent variables.

For multiple linear regression the change in the mean value of the outcome (on a simple arithmetic scale) is modeled as the sum of the unit changes of the interval-independent variable. The situation is more complicated with logistic

regression and proportional hazards analysis. As with multiple independent variables, changes on an interval-independent variable are "additive on a multiplicative scale." Thus, if you have an interval-independent variable, such as blood pressure, the additive assumption would mean that the odds ratio or relative hazard associated with a 20 mm increase in blood pressure would be twice the odds ratio or relative hazard associated with a 10 mm increase in blood pressure. Numerically, if the increase in risk owing to a 10 mm increase in blood pressure from 140 to 149 mm is 1.4, then the increase in risk owing to an increase from 140 to 159 mm is 1.96 (1.4×1.4).

In Section 5.4, I suggested that one way to assess the linear assumption of an interval variable was to categorize the variable into multiple categorical variables, maintaining the interval nature of the scale. This procedure allows you to see if the differences between sequential coefficients are about the same (as you would expect if the relationship is linear). As an extension of this method, you can include the other independent variables in the model and thereby assess whether the interval-independent variable fits the additive assumption with statistical adjustment for the other covariates.

10.10 What is the proportionality assumption?

DEFINITION

The *proportionality assumption* is that the hazards for persons with different patterns of covariates are constant over time.

The proportionality assumption is relevant only to proportional hazards analysis (Section 5.1). The assumption is that the hazards for persons with different patterns of covariates are constant over time. In other words, if the relative hazard of heart attack among diabetics is three times higher than among nondiabetics in the first year of the study, the relative hazard of heart attack must also be (about) three times higher among diabetics than nondiabetics in the second year of the study. Note that the hazard for a heart attack can be very different in the first year than in the second year (e.g., much higher in the first year than in the second year), but the difference between the hazards for diabetics and nondiabetics must be constant throughout the study period.

While the proportionality assumption may, at first, sound complicated, it is really very simple. Since proportional hazards analysis, like multiple linear and logistic regression, provides you with a single coefficient for each variable, that coefficient, and its associated relative hazard, must represent the risk throughout the time period. If the risk of outcome associated with a particular variable is higher at one point in time and lower at another, a single coefficient cannot represent that relationship.

For example, in Figure 10.2 we see that the risk of death among patients with acute nonlymphoblastic leukemia is not constant over time in the two arms of

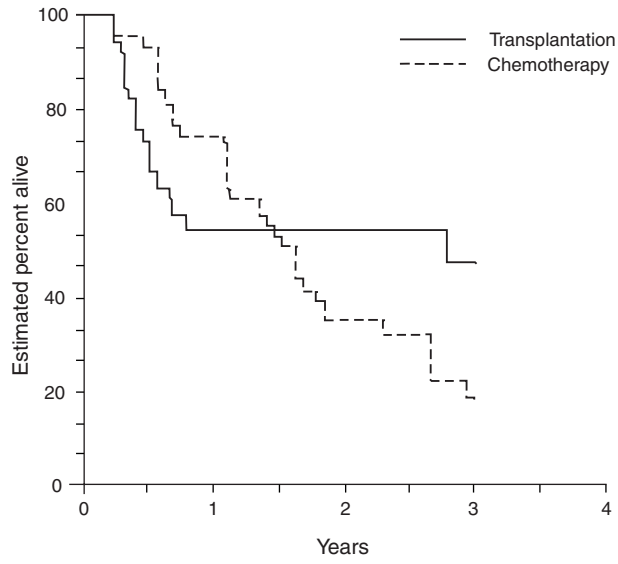


Figure 10.2 Kaplan–Meier curves show the estimated probability of survival for the chemotherapy group (broken line) and the transplantation group (solid line). Reprinted with permission from Appelbaum, F. R., *et al*. "Bone marrow transplantation or chemotherapy after remission induction for adults with acute nonlymphoblastic leukemia." *Ann. Intern. Med*. **101** (1984): 581–8.

the study.[4] Patients who received a bone marrow transplant were more likely to die in the first year. But, thereafter, the risk was lower than with conventional chemotherapy.

If you used proportional hazards analysis to analyze the effect of transplantation on death, the relative hazard would probably be one. The higher risk of death associated with transplantation in the first year-and-a-half and the lower risk of death associated with transplantation in the period between a year-and-a-half and three years would average out. Although this average risk of one is arguably the best single estimate of the difference in risk of death with transplantation compared to chemotherapy, you would not want to tell your patients that the risk of death with the two treatments was the same. It would be more informative to tell them that a bone marrow transplant is a toxic treatment and that there are a significant number of deaths caused by it. However, if they survive the treatment, their survival at three years is significantly better than with conventional therapy.

[4] Appelbaum, F. R., Dahlberg, S., Thomas, E. D., *et al.* "Bone marrow transplantation or chemotherapy after remission induction for adults with acute nonlymphoblastic leukemia." *Ann. Intern. Med.* **101** (1984): 581–8.

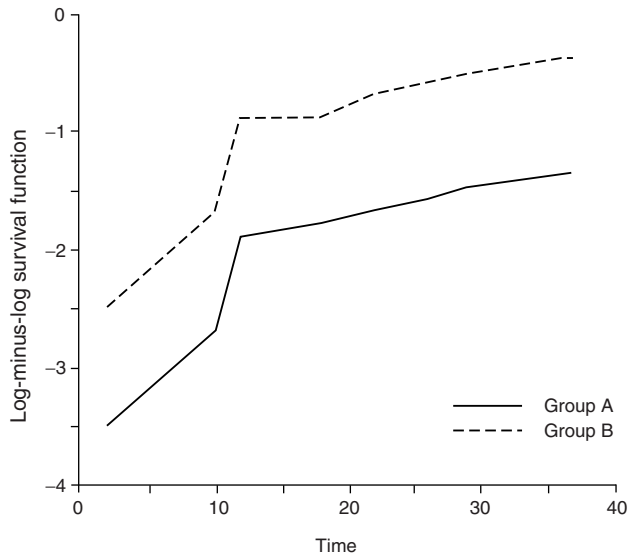


Figure 10.3 Log-minus-log survival plot showing a constant difference between group A (solid line) and group B (broken line). Reprinted with permission from Katz, M. H. and Hauck, W. W. "Proportional hazards (Cox) regression." *J. Gen. Intern. Med*. **8** (1993): 702–11.

Some researchers will assess the proportionality assumption only for the main effect in their study (e.g., treatment versus no treatment). However, if potential confounders do not meet the proportionality assumption you may not be adequately adjusting for them in the analysis. For this reason, it is important to assess whether all your variables fulfill the proportionality assumption, and not just the main effect you are studying.

10.11 How do I test the proportionality assumption?

TIP

If the curves cross, the proportionality assumption is violated.

The proportionality assumption can be assessed for a single independent variable with Kaplan–Meier survival curves. When the proportionality assumption is met, there should be a steadily increasing difference between the two curves. In contrast, if the two curves cross, as in the case of the transplantation versus chemotherapy study, the proportionality assumption is violated (Figure 10.2).

There are more sophisticated methods for assessing whether the proportional hazards assumption is met. Two commonly used methods are: graphical representations and time-dependent variables. The graph is referred to as a log-minus-log survival plot (Figure 10.3). If there is a constant (vertical)

difference between the two curves then you know that the hazards of subjects with different values on a particular independent variable are proportional over time.

A nice feature of log-minus-log survival curves is that they illustrate that proportional hazards analysis makes no assumption about the absolute risks (the lines change direction and slope). Proportional hazards analysis only assumes that as the hazards change, the distance between the two curves stays about the same. The curves do not have to be perfectly equidistant. Especially at the ends of the curves, where there tend to be fewer observations, there may be some coming together or greater splitting of the curves. A limitation of log-minus-log survival curves is that, as with Kaplan–Meier curves, you can assess whether the proportionality assumption is fulfilled for only one variable at a time.

Another way of testing the proportionality assumption is to add interaction terms to the proportional hazards model that allow the relative hazard to vary over time. The interaction terms are called time-dependent covariates (Section 14.1); they require special coding, which varies with different statistical packages. They can be created so that the logarithm of the relative hazards is allowed to vary linearly with time or with the logarithm of time. When the proportionality assumption is valid, the interaction term will have a hazard ratio near 1.0 (no effect) and will not be statistically significant. If the odds ratio is significantly different from 1.0 it means that the effect of the independent variable does vary over time (the proportionality assumption is not met).

The advantage of time-dependent covariates is that you can create more than one such term and assess simultaneously whether the proportionality assumption is met for multiple independent variables. However, if your sample size is small, you may not be able to enter interaction terms for all variables in the model simultaneously (there may be too many variables for the sample size). In that case, you can add each product term individually to your other variables to see if the product term is significant with adjustment for the other variables in the model. A potential disadvantage of time-dependent covariates is that, if your sample is really large, time interaction terms may be too sensitive. You may get statistically significant terms even though the graphical representation does not show a major departure from the proportionality assumption.

Alternatively, you can construct your model such that the relative hazard varies by period, taking on one value in the first X time units, another value in the next Y time units, and so on. This approach is analogous to transforming an interval-independent variable (in this case time) into a categorical variable.

With this approach you can test whether the relative hazards between periods are statistically different.

For example, my colleagues and I were performing an analysis of the impact of socioeconomic status on survival with AIDS.[5] We assessed the proportionality assumption using log-minus-log survival curves. Most of the curves looked fine. But one of the covariates, an initial AIDS diagnosis of cryptosporidiosis, looked suspicious for violating the proportionality assumption. Because we were not sure, we tested the hypothesis. Instead of the single variable cryptosporidiosis (yes/no), we created two covariates: cryptosporidiosis in the first period of study time and cryptosporidiosis in the second period of study time (yes/no). The cutoff for the time period was chosen based on the log-minus-log survival plot (the point where it seemed the relative hazard changed).

Using proportional hazards analysis we tested the hypothesis that the two variables were significantly different from one another. The *P* value was only marginally significant (0.07). Although this was above the conventional cutoff of *P* value <0.05, *P* values less than <0.20 can indicate important changes in risk over time. However, the variable cryptosporidiosis was not the major variable of interest in our study. We were looking for the relationship between socioeconomic status and survival. Our goal was to adequately adjust for AIDS diagnoses that might confound the relationship between socioeconomic status and AIDS survival. Since we were unsure how important this potential departure from the proportionality assumption was for our study, we tested whether a proportional hazards model that included these two time-period variables, instead of the one covariate, changed any of the other coefficients in important ways. It did not. We therefore chose to report the simpler model (to make the study more interpretable for readers), adjusting only for the variable cryptosporidiosis. We explained what we had done in the Methods section (of course!). Had it been a study focused on how cryptosporidiosis affected AIDS survival and/or had the two covariates changed the coefficients of our other variables, we would have included these time-period variables in the final model.

10.12 What if the proportionality assumption does not hold for my data?

You have a few choices (besides abandoning your research career!). You can divide your observation period such that within any one period the odds are

[5] Katz, M. H., Hsu, L., Lingo, M., *et al.* "Impact of socioeconomic status on survival with AIDS." *Am. J. Epidemiol.* **148** (1998): 282–91.

proportional. If you do this, you will have two or more proportional hazards analyses. The factors associated with outcome would differ for the two periods (e.g., diabetics will have a higher risk in the model estimating heart attack during the first three months and a lower risk in the model estimating heart attack in the second three-month period).

A second strategy is to perform a stratified proportional hazards analysis. The sample is stratified by the variable that does not fit the proportionality assumption. Each stratum has its own baseline hazard. Therefore each stratum has a component that can vary over time differently from the other strata. The limitations of this strategy are the limitations of any stratified analysis. You cannot assess the effect of the stratification variable on your outcome. Also, stratification is cumbersome if you have more than one or two variables that do not fit the proportionality assumption.

A third strategy is to switch your analysis to logistic regression. Because time is not taken into account in logistic regression models, the risks do not have to be proportional over time. The researchers assessing the factors associated with survival with AIDS in New York City (Section 8.5) switched their model from proportional hazards analysis to logistic analysis after they discovered the proportionality assumption "was seriously violated and could not be remedied through stratification." They created a dichotomous outcome variable: survival for 15 months or longer (yes/no). They found several variables that were associated with survival at 15 months. Although this solution avoided violating the proportionality assumption, one problem is that their results may have been dependent on the cutoff they chose for their outcome. In other words, they may have found different factors associated with survival if they had chosen a cutoff of six or 24 months. This is especially problematic because the important research question is: What is associated with survival? Not: What is associated with survival at 15 months? Also, any subjects who were lost to follow-up prior to 15 months would have to be excluded from the analysis. Nevertheless, the researchers deserve credit for verifying the proportionality assumption and adapting their analysis; many authors do not report if and how they assessed the proportionality assumption.[6]

TIP

You can account for the lack of proportionality in a model by allowing the relative hazard to vary by period.

A fourth, and probably the best, strategy is to account for the lack of proportionality in the hazards. As explained in Section 10.11 this can be accomplished by setting up the model so that the relative hazard varies by period. You may also be able to transform the independent variable (e.g., polynomial transformation)

[6] Katz, M. H. and Hauck, W. W. "Proportional hazards (Cox) regression." *J. Gen. Intern. Med.* **8** (1993): 702–11; Concato, J., Feinstein, A.R., and Holford, T. R. "The risk of determining risk with multivariable models." *Ann. Intern. Med.* **118** (1993): 201–10.

so that it fits the proportionality assumption. Because these models are complex to set up, it would be best to consult with a biostatistician for help.

No matter what strategy you choose for your analysis, report to the reader how you assessed the assumption and whether it held. If it did not hold, don't feel discouraged. You have learned something, potentially important, about how the risk of your outcome changes over time under certain conditions.

11

Propensity scores

11.1 What are propensity scores? Why are they used?

DEFINITION

Propensity scores are the likelihood of a subject being in a particular treatment group, conditional on that subject's values on those independent variables thought to influence group membership.

Propensity scores are calculations of the likelihood of a subject being in a particular treatment group, conditional on that subject's values on those independent variables thought to influence group membership.[1] They are used to statistically adjust for differences between nonrandomized groups, typically for studies comparing different treatments.

To create a propensity score enter the variables that influence treatment group assignment into a logistic regression model estimating treatment group membership.

To calculate a propensity score, you first identify the variables that influence treatment group membership, including demographics, disease severity, and characteristics of the treatment system (e.g., physician specialty, hospital, etc.). These variables are entered into a model (typically logistic) estimating the likelihood of treatment group membership. This model yields a score for each subject; the score is the estimated likelihood of being in one group versus the other, conditional on a weighted score of that subject's values on the set of independent variables used to create the propensity score.

A propensity score can be used in three ways: (1) as a covariate in a multivariable model estimating outcome; (2) as a variable on which to match subjects; (3) as a variable on which to stratify subjects.

Once calculated there are different ways you can use propensity scores. You can include each subject's propensity score in your multivariable model as an independent variable. Or you can use this score to individually match subjects with different treatment assignments but an equal likelihood of being in a particular group and assess the outcome using a matched analysis (Chapter 12). Finally you can assess the likelihood of outcome within quintiles of the propensity score. Attentive readers will recognize that these three different methods of using propensity scores correspond to three different methods for adjusting for baseline differences: multivariable modeling, matching, and stratification.

But, wait you say! If multivariable analysis (along with matching and stratification) can adjust for baseline differences between nonrandomized groups (Section 2.1.B), why do we need to complicate things further with propensity

[1] Rubin, D. B. "Estimating causal effects from large data sets using propensity scores." *Ann. Intern. Med.* **127** (1997): 757–63; D'Agostino, R. B. "Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group." *Statist. Med.* **17** (1998): 2265–81.

Propensity scores work especially well when outcomes are rare and the proportions of subjects in the treatment groups are relatively equal.

scores? Why not simply enter the variables used to construct a propensity score into a multivariable model?

The answer is that propensity scores often produce a better adjustment for baseline differences than simply including potential confounders in your multivariable model. This is especially true when outcomes are rare and the proportions of subjects in the treatment groups are relatively equal.[2] With rare outcomes you often will not have a sufficient sample size to include all the potential confounders in your analysis estimating outcome. Omitting important variables may result in an incorrectly specified model. On the other hand, including too many variables in your multivariable model may result in an unreliable model (Section 7.4). But assuming the distribution of subjects between your groups is relatively equal, then you will have a sufficient sample size to use all your prognostic variables to estimate group assignment. When you return to your model estimating outcome you can enter your propensity score as a single variable stand-in for your multiple potential confounders.

Another advantage of propensity scores is that they make no assumption about the relationship between the individual confounders and the outcome (Chapter 5). Of course, for the propensity score to be accurate, the relationship of the independent variables to treatment assignment must fit the assumptions of the model you are using.

Propensity scores were important in demonstrating that right-heart catheterization (a procedure used extensively in critically ill patients during the time I was an internal medicine resident) is not a useful procedure, and may be harmful.

Right-heart catheterization involves inserting a monitoring (Swan–Ganz) catheter directly into the right heart. It began to be widely used in the 1970s without any studies proving its efficacy. Many clinicians felt that the readings enabled them to better monitor and treat their patients. Thus, the practice became the standard of care in certain settings, including the intensive care units of the hospitals I trained in. When some studies found higher rates of death in patients who received right-heart catheterization, the validity of the association between right-heart catheterization and death was questioned because the studies were not randomized. In particular, persons who received right-heart catheterization were known to be sicker than persons who did not receive this procedure. This could have confounded the results of the observation trials, resulting in persons who received right-heart catheterization appearing more likely to die because of the catheterization when they were, in reality, more likely to die because of their underlying disease. This relationship is illustrated in Figure 11.1.

[2] Braitman, L. E. and Rosenbaum, P. R. "Rare outcomes, common treatments: Analytic strategies using propensity scores." *Ann. Intern. Med.* **137** (2002): 693–5.

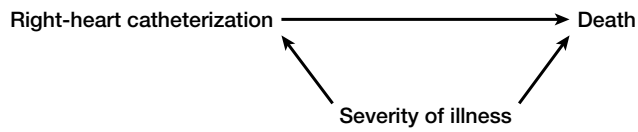


Figure 11.1 Relationships among right-heart catheterization, severity of illness, and death.

When a randomized controlled trial was launched to definitively answer the question, the study was terminated because physicians were unwilling to randomize their patients. They believed that right-heart catheterization was beneficial; therefore, how could they deny the intervention to their patients?

Since randomization was not an option, Connors and colleagues addressed the effectiveness of right-heart catheterization by prospectively following 5735 critically ill adults cared for in an intensive care unit.[3] Of these, 2184 patients received a right-heart catheterization (38 percent) and 3551 (62 percent) did not. Because the decision of whether or not a patient received a right-heart catheterization was not random, the investigators developed a propensity score to assess the likelihood of each patient receiving a right-heart catheterization. To do this, they had a group of seven specialists in critical care specify the variables they would expect to be related to the decision to use or not use a catheter. They identified over 65 variables to include. These variables were included in a logistic regression analysis estimating the outcome of right-heart catheterization in the first 24 hours of hospitalization. When they adjusted for the propensity score for right-heart catheterization in a proportional hazards analysis, along with additional adjustment for potential confounders, patients managed with a right-heart catheterization had an increased risk of death (OR = 1.21; 95% CI = 1.09 – 1.25) at 30 days.

To strengthen their findings, the investigators also used the propensity scores to perform a matched analysis. Patients managed with and without right-heart catheterizations were matched on the basis of disease category and the propensity score. Patients were matched to the patient with the closest propensity score (within 0.03 on a scale of 0 to 1). This procedure resulted in 1008 pairs (note: not all of the 2184 patients who received a right-heart catheterization could be matched). The likelihood of survival at 30 days was higher for those patients who did not receive right-heart catheterization (67.2) than those who did (62.5) with an odds ratio of 1.24 (95% CI 1.03–1.49).

One ironic aspect of their study is that it created enough uncertainty about the use of right-heart catheterization that a randomized clinical trial became

[3] Connors, A. F., Speroff, T., Dawson, N. V., *et al.* "The effectiveness of right-heart catheterization in the initial care of critically ill patients." *JAMA* **276** (1996): 889–97.

feasible.[4] Published in 2003, seven years after the publication of the study by Connors and colleagues, it showed no benefit for right-heart catheterization among high-risk surgical patients and a higher rate of pulmonary embolism in the catheter group.[5]

Propensity scores only adjust for measured confounders.

The limitations of propensity scores are similar to the limitations of all forms of multivariable adjustment. Propensity scores can only adjust for measured confounders.

Although you can never adjust for unknown confounders, it is possible to assess how large such a confounder would have to be to affect your results. With the right-heart catheterization study, the investigators used sensitivity analysis to determine how strong a missing confounder would have to be to change the relationship they found from right-heart catheterization being associated with an increased risk of death (OR = 1.21), to right-heart catheterization being associated with a decreased risk of death (OR = 0.80).[6] They found that the missing covariate would have to increase the risk of death six-fold and increase the probability of right-heart catheterization six-fold for the true relative hazard to be 0.80. While this analysis certainly does not prove that such a covariate does not exist, it seems unlikely. This is substantiated by the investigators determining that singly removing the known confounders that have the largest effect on the probability of right-heart catheterization changed the relative hazard of death by only 0.01.

For propensity scores to be effective in adjusting for baseline differences there must be sufficient overlap between the treatment groups on the independent variables you are using to estimate group assignment. One way to assess this is to look at the propensity scores themselves. In the right-heart catheterization study the mean propensity score of patients who received a right-heart catheterization was 0.577 (95 percent confidence interval 0.108–0.943), while the score for those who did not receive a right-heart catheterization was 0.253 (95 percent confidence interval 0.011–0.779).

A second important test of whether there is sufficient overlap of the groups is to stratify the groups within quintiles[7] of the propensity score. Then compare the independent variables used to create the scores within the quintiles. You

[4] Dalen, J. E. and Bone, R. C. "Is it time to pull the pulmonary artery catheter?" *JAMA* **276** (1996): 916–18.

[5] Sandham, J. D., Hull, R. D., Brant, R. F. *et al.* "A randomized, controlled trial of the use of pulmonary-artery catheters in high-risk surgical patients." *N. Engl. J. Med.* **348** (2003): 5–14.

[6] For more on sensitivity analysis see Rosenbaum, P. R. and Rubin, D. B. "Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome." *J. R. Stat. Soc.* **45** (1983): 212–8.

[7] The reason for using quintiles is because of an analysis by W. G. Cochran showing that stratification into five or six groups will typically remove 90 percent of the bias present in unadjusted analyses. For a more detailed explanation see: Rubin, D. B. "Estimating causal effects from large data sets using propensity scores." *Ann. Intern. Med.* **127** (1997): 757–63.

Test propensity scores by comparing the distribution of independent variables within quintiles of the score.

should find similar distributions of the independent variables (e.g., if age is one of the variables used to create a propensity score then the ages of subjects within each of the quintiles should be similar). If not, then there is insufficient overlap between the groups for the propensity score to satisfactorily adjust for baseline differences. This is analogous to adjusting for a confounder using multivariable modeling without a propensity score: to satisfactorily adjust for a confounder there must be sufficient overlap of the groups on the confounder (Section 2.1.A)

For example, in the right-heart catheterization study, within quintiles of propensities scores there were no differences between those who received and those who did not receive right-heart catheterization on severity of illness, mean blood pressures, heart rate, respiratory rate, pH, PaO_2/FIO_2, $PaCO_2$, disease category, or prognosis.

Despite these limitations propensity scores are the best available method for adjusting for baseline differences in nonrandomized studies of treatments or interventions.

12

Correlated observations

12.1 How do I analyze correlated observations?

The multivariable methods that we have discussed thus far assume that each observation (subject) is independent (i.e., the outcomes of different subjects are not correlated). However, it has become increasingly common to study data where the observations are correlated with one another.

Repeated observations of subjects in longitudinal studies lead to correlated responses.

By far the most common circumstance leading to correlated outcomes is longitudinal studies, where subjects are observed repeatedly (e.g., baseline and every six months thereafter). Because it is the same subject being observed multiple times, the responses are correlated (i.e., the same subject is more likely to have a similar response each time he or she is observed than a different subject would).

However, as you can see from Table 12.1, there are disparate sets of circumstances that may lead to correlated observations.

Although the study designs listed in Table 12.1 seem disparate, what they all have in common is that observations are clustered. In the case of a longitudinal study, the cluster is the subject (subjects are observed multiple times). When subjects receive different treatments or have different body parts observed, the cluster is also the subject. In the case of subjects who have been randomized or recruited, based on an established group (e.g., families, doctors' practices), the cluster is the established group. In a matched design, the cluster is the matched case and control.

Outcomes are more similar within a cluster than between clusters.

When observations are clustered, outcomes are more similar within a cluster than between clusters. The same person is more likely to respond to different circumstances in the same way than different individuals. Different body parts (e.g., the right and left eye) of the same person are more likely to respond similarly than body parts belonging to different persons.

Similarly, when subjects are randomized or observed at an established group level (e.g., family, hospital, city), subjects from the same group are more likely to respond in similar ways than subjects from different groups, leading to correlated outcomes. Finally, when cases have been individually matched to controls, the

Table 12.1 Circumstances leading to correlated observations.

Multiple observations of the same subjects at **different times** (e.g., a longitudinal study that assesses subjects every six months).
Multiple observations of the same subjects after receiving **different treatments (e.g., cross-over studies**).
Multiple observations of **different body parts** of the same subjects.
Study designs where subjects have been randomized by or recruited from established groups (**clusters**) of **related individuals** (e.g., families, doctors' practices, or hospitals).
Matched study designs where cases and controls have been **individually matched**.

case is more likely to respond in a similar fashion to the individually matched control than to other controls.

The analysis of correlated observations is somewhat more complicated than that of independent observations because you have to take into account the within-cluster correlation. Although your first inclination might be to sidestep the problem by not collecting clustered data, there are a number of important reasons for collecting clustered data.

TIP

Multiple observations of subjects increase the power of a study.

First, collecting repeated observations of the same persons can increase the power of your study without increasing the number of subjects. For example, Guskiewicz and colleagues used repeated observations of college football players to assess the effects of recurrent concussion on neurologic function.[1] They followed 2905 players for a total of 4251 player-seasons with an estimated 240 951 exposures. As you would expect, with multiple observations there are more outcomes. Even so, despite the 240 951 exposures, only 184 players had a concussion and only 12 had a repeat concussion within the same season.

TIP

If you are studying a body part that comes in duplicate, increase your sample size by tracking outcomes in both parts.

Similarly, if you are studying an outcome in one of the many body parts that occurs in duplicate (e.g., eyes, most joints) or in higher multiples (e.g., teeth, fingers), using all body parts can increase your sample size with the same number of subjects. For example, McAlindon and colleagues studied the relationship of vitamin D to development of osteoarthritis of the knee using data from the Framingham study.[2] Although this cohort consisted of over 5000 subjects, only 556 participants had x-rays of their knees and assessments of their vitamin D intake and serum levels. Therefore, the investigators needed to maximize their statistical power. They did this by looking at both knees. They found that low intake of vitamin D was associated with progression of osteoarthritis of the knee.

[1] Guskiewicz, K. M., McCrea, M., Marshall, S. W., *et al.* "Cumulative effects associated with recurrent concussion in collegiate football players." *JAMA* **290** (2003): 2549–55.

[2] McAlindon, T. E., Felson, D. T., Zhang, Y., *et al.* "Relation of dietary intake and serum levels of vitamin D to progression of osteoarthritis of the knee among participants in the Framingham study." *Ann. Intern. Med.* **125** (1996): 353–9.

Observing the same subject under different conditions is another way of increasing the power of your study without increasing your sample size. For example, Paz and colleagues studied the effect of ingestion of ethanol on obstruction of the left ventricular outflow tract in patients with hypertrophic obstructive cardiomyopathy.[3] They performed echocardiography on 36 patients before and after ingestion of ethanol or of placebo. They found that compared to placebo even small amounts of ethanol increased obstruction of the outflow tract – an important finding for patients with hypertrophic obstructive cardiomyopathy who are considering drinking alcohol.

TIP

Use repeated observations of subjects to assess time trends.

A second reason for collecting repeated observation of subjects is that it enables us to assess time trends. For example, Conter *et al.* studied the effect of maternal smoking during pregnancy.[4] They found that children born to smoking mothers had significantly lower birth weights, but that the rate of growth between 0 and 6 months was greater for babies born to smoking mothers than those born to nonsmoking mothers. The result was that by 6 months of age the babies of smoking mothers were the same weight as those of non-smoking mothers. (You have to marvel at the ability of nature to triumph over poison!)

A reason you might choose to randomize by group (as in a particular clinic or town) rather than by an individual is that you are interested in interventions that operate on the level of the community, rather than on the level of the individual.[5] For example, you might be interested in studying the effect of the impact of traffic-calming strategies on pedestrian injuries. Also, sometimes it is impossible to randomize individuals within a group because the intervention affects the entire group. For example, Jensen and colleagues randomized nine residential care centers to receive either a multidisciplinary program to reduce falls or standard care.[6] Because the intervention included many center-wide changes such as removing environmental hazards, staff education, and improving transfer techniques, everyone in the center would potentially benefit from the changes. Adjusting their analysis for clustering by center, the investigators found that residents who lived in an intervention center had fewer falls than residents in the centers that did not receive the intervention (OR=0.49; 95% CI 0.37–0.65).

Although analyzing clustered data is more complicated than analyzing independent data, it is sometimes easier to collect data from established clusters

[3] Paz, R., Jortner, R., Tunick, P. A., *et al.* "The effect of the ingestion of ethanol on obstruction of the left ventricular outflow tract in hypertrophic cardiomyopathy." *N. Engl. J. Med.* **335** (1996): 938–41.
[4] Conter, V., Cortinovis, I., Rogari, P., *et al.* "Weight growth in infants born to mothers who smoked during pregnancy. *BMJ* **310** (1995): 768–71.
[5] Murray, D. M., Varnell, S. P., and Blitstein, J. L. "Design and analysis of group-randomized trials: A review of recent methodological developments." *Am. J. Public Health* **94** (2004): 423–32.
[6] Jensen, J., Lundin-Olsson, L., Nyberg, L., *et al.* "Fall and injury prevention in older people living in residential care facilities: A clustered randomized trial." *Ann. Intern. Med.* **136** (2002): 733–41.

than from individuals. For example, Gandhi and colleagues assessed adverse drug events in 661 patients seen by 24 physicians.[7] Because physician practice style would be expected to influence prescribing practices, the investigators needed to adjust for clustering by physicians. However, it would have been a logistic nightmare to review the medical records of 661 patients seen by 661 physicians. At this point you might wonder why not just study 661 patients from the same physician. Then you would not have to adjust for clustering by physician and you would have an easy job with data collection. The problem is that when you collect data from a single practitioner your data are not very generalizable.

Finally, the reason we conduct individually matched studies is to decrease confounding (the key word here is individually matched; it is not a matched design if you choose controls that as a whole are comparable to the cases). Matching cases and controls on potentially confounding variables will enable you to answer the same question with a smaller sample because you will no longer have to adjust for differences between the cases and the controls on the variables for which you have matched (they are the same for each pair of cases and controls).

Failure to account for nonindependent observations will artificially exaggerate the statistical significance of your results.

Regardless of why you have clustered data, you will need multivariable methods that adjust for the correlations within the clusters (Table 12.2).[8] Failure to do so will exaggerate the statistical significance of your results because the analysis will treat each observation as if it were a subject. Although two observations of 100 subjects give you greater confidence that your results reflect "truth" in the population than a single observation of the same 100 subjects, it does not give you as much confidence as observing 200 subjects.

TIP

With correlated observations you need to specify a variable that identifies the clusters.

Although there are major differences in these methods, one commonality is that you will need to have a variable that identifies each cluster. Without such a variable there would be no way for the software to know which observations are repeated or related. In the case of repeated observations of the same person this variable may be the identification number of the subject. In the case of data clustered by group, it will be the identification of the group (e.g., the medical practice or the hospital) from which the case was drawn.

12.1.A Transform repeated observations into a single measure

One straightforward method of analyzing repeated observations of a subject is to transform the repeated observations into a single measure. This can be done

[7] Gandhi, T. K., Weingart, S. N., Borus, J., *et al.* "Adverse drug events in ambulatory care." *N. Engl. J. Med.* **348** (2003): 1556–64.

[8] For a review of bivariate statistics for analyzing correlated observations see: Katz, M. H. *Study Design and Statistical Analysis: A Practical Guide for Clinicians.* Cambridge University Press, forthcoming.

Table 12.2 Methods for studying correlated observations.

Method	Features
Transform into a single measure Change score Slope	May be used with repeated observations of the same subject on an interval outcome. Change scores are useful when there are only two time points. Slopes can accommodate multiple observations, an unequal number of observations per cluster, and unequal time intervals between observations, but can only be used for linear trends.
Generalized estimating equations (also called marginal models or population-averaged models)	Can model a variety of different relationships between risk factors and outcomes; adjustment widens confidence interval but does not change point estimate. Can accommodate unequal number of observations and unequal time intervals between observations.
Mixed-effects models (also called multilevel models, random effects regression models, random coefficient models, and hierarchical models)	Can model a variety of different relationships between risk factors and outcomes; adjustment affects both the point estimate and the confidence interval. Can accommodate unequal number of observations and unequal time intervals between observations.
Repeated measures analysis of variance	Can only be used with interval outcomes, an equal number of observations per subject and fixed periods between observations.
Conditional logistic regression	May only be used with dichotomous outcomes, an equal number of observations per subject, and fixed periods between observations.
Anderson-Gill counting process for proportional hazards analysis	Adaptation of proportional hazards analysis for repeated outcomes of a time to outcome variable.
Marginal approach for proportional hazards analysis	Adaptation of proportional hazards analysis for repeated outcomes of a time to outcome variable.

DEFINITION

A *change score* is the absolute or relative change of your outcome variable over the study period.

using a change score. A change score is the absolute or relative change of an outcome variable over the study period.

For example, Penninx and colleagues were interested in assessing whether depression contributes to subsequent functional decline in older persons.[9] They assessed the physical performance of 1286 elders at baseline and then four years later. To determine the absolute change in performance, they subtracted the baseline score from the follow-up score. They then used change in performance as their outcome measure. Because they used a single change score rather than the two correlated observations, they were able to use multiple linear regression, without adjustment for the correlation between observations. They found

[9] Penninx, B. W. J. H., Guralnik, J. M., Ferrucci, L., *et al.* "Depressive symptoms and physical decline in community-dwelling older persons." *JAMA* **279** (1998): 1720–6.

that older persons who reported depressive symptoms were at higher risk of subsequent physical decline.

Change scores can also be adapted to weigh relative changes more than absolute changes. This is done by dividing the change score by the baseline score. For example, with CD4 lymphocyte counts (an immunologic measure used to assess persons with HIV disease), a 100-cell change in six months would more likely be associated with progression of disease, if it reflected a drop from 200 cells to 100 cells than if it reflected a drop from 1200 cells to 1100 cells. Both result in an absolute change of 100 cells but the relative change of the former is much larger ($100/200 = 0.5$) than the latter ($100/1200 = 0.08$).

TIP

Slopes can be used to incorporate multiple measurements over time.

Change scores will not work when you have more than two observations. However, you can develop a measure of change over the course of the study period. This is usually done by plotting the observations for each case over time and determining the slope for each case. You can then use the slope as a continuous outcome variable. This works well for variables that increase (or decrease) in a linear fashion over time. Continuing with the example of serial CD4 lymphocyte counts, the slope of the CD4 count is often used by investigators as a measure of the rate of progression of disease over time because these counts decrease in a linear fashion (in the absence of therapy).[10]

Analogous to dividing a change score by the baseline value, you can divide the slope by its intercept. The result will be that changes that occur in subjects with high intercepts will be weighted less than changes that occur at the lower values. This method was used by Riggs and colleagues to evaluate change in bone mineral density in their study of fluoride treatment of osteoporosis.[11]

An advantage of slopes is that they can be performed when you have an unequal number of observations between clusters (e.g., one subject has four observations and a different subject has six observations) or when you have unequal time intervals between observations (e.g., one subject has measurements at three months, six months and 18 months and a different subject has measurements at six months, 12 months and 24 months).

To calculate a slope only two points at any time are needed. However, researchers usually will set a minimum of points needed to have a valid slope. (For example, Phillips and colleagues included only subjects who had at least five measurements of their CD4 count.) If the minimum number of measurements is not available, the case is excluded.

[10] Phillips, A. N., Lee, C. A., Elford, J., *et al.* "Serial CD4 lymphocyte counts and development of AIDS." *Lancet* **337** (1991): 389–92.

[11] Riggs, B. L., Hodgson, S. F., O'Fallon, W. M., *et al.* "Effect of fluoride treatment on the fracture rate in postmenopausal women with osteoporosis." *N. Engl. J. Med.* **322** (1990): 802–9.

Although the use of change scores or slopes may seem simplistic, it can be a very powerful method of dealing with repeated observations. For example, D'Amico used slopes to characterize changes in prostate-specific antigen (PSA) levels (PSA velocity) among patients with prostate cancer.[12] Over a thousand men were followed for an average of five years, with PSA levels measured about every six months. They found that men with a PSA velocity of more than 2.0 ng per milliliter per year were significantly more likely to die from prostate cancer than men with who had smaller changes in their serial PSA levels.

A disadvantage of slopes is that they only apply when the outcome changes linearly over time. Any other time trend would yield incorrect results when analyzed by this method. (Any change between two points can be statistically treated as linear.)

12.1.B Generalized estimating equations

Generalized estimating equations represent an extremely flexible method of adjusting for clustered observations.[13] They can be used to model a variety of different relationships between the risk factors and an outcome including linear, logistic, or logarithmic; generalized estimating equations can be used with interval, dichotomous, ordinal, and categorical outcomes;[14] they allow for inclusion of independent variables that do not change (fixed variables such as ethnicity) as well as variables that change at each observation (e.g., blood pressure).

Generalized estimating equations are population-averaged models (also referred to as marginal models). The mean of the dependent variable is modeled as a function of the independent variables, assuming that the variance is a known function of the mean.[15]

Generalized estimating equations can incorporate different numbers of observations for different clusters (e.g., two observations for one subject and four

[12] D'Amico, A. V., Chen M-H., Roehl, K. A., *et al.* "Preoperative PSA velocity and the risk of death from prostate cancer after radical prostatectomy." *N. Engl. J. Med.* **351** (2004): 125–35.

[13] For a comprehensive text of generalized linear models see: Hardin, J. W. and Hilbe, J. M. *Generalized Estimating Equations.* Boca Raton, FL: Chapman & Hall, 2003; for more on the original development of the methods see: Zeger, S. L. and Liang, K.-Y. "Longitudinal data analysis using generalized linear models." *Biometrika* **73** (1986): 13–22; Zeger, S. L. and Liang, K.-Y. "Longitudinal data analysis for discrete and continuous outcomes." *Biometrics* **42** (1986): 121–30.

[14] For an ordinal outcome, the model underlying generalized estimating equations would be proportional odds logistic regression. For a good example of how to set up such a model see: Tishler, P. V., Larkin, E. K., Schluchter, M. D., *et al.* "Incidence of sleep-disordered breathing in an urban adult population." *JAMA* **289** (2003): 2230–7.

[15] Murray, D. M., Varnell, S. P., and Blitstein, J. L. "Design and analysis of group-randomized trials: A review of recent methodological developments." *Am. J. Public Health* **94** (2004): 423–32.

observations for another subject; 20 subjects from one hospital and 40 subjects from a different hospital). This is a major advantage of this method.

In cases where the different number of observations stems from missing data, you have to distinguish between data that are not missing at random, also referred to as "nonignorable" or "informative" missing data, and data that are missing randomly, also referred to as "ignorable" missing data.[16] Missing data is nonignorable if (1) the occurrence of a missing value can be predicted by a prior value of the outcome (e.g., patients with missing CD4 counts are more likely to have had a low CD4 count prior to the missing value); and/or (2) certain groups of patients are more likely to have a missing value on the outcome (e.g., Caucasian women are more likely to have a missing observation for bone density than women of other ethnicities).

TIP

Evaluate your missing data by using logistic regression to test the association between having missing values and the independent variables and the prior values of your outcome variable.

To test whether observations are missing randomly assign each subject a value of 0 or 1, depending on whether or not the subject has one or more missing observations on the outcome. Using multivariable logistic regression, test the association between this variable and the independent variables and the prior values of the outcome. If the data are missing randomly, there should be no association between the independent variables and the prior values of the outcome, and whether or not the subject has missing values.

Generalized estimating equations can accommodate missing responses if the data are missing randomly.[17] Some investigators have found that when the outcome variable is continuous (but not when it is dichotomous), generalized estimating equations may also accommodate missing data that are not missing randomly.[18] My view is that when you have a lot of missing data or nonrandom missing data you need to worry that your analysis is biased. On the other hand, small amounts of data, or data that you can empirically show are ignorable, are unlikely to cause problems with generalized estimating equations.

Generalized estimating equations can handle unequal intervals between observations.

Another advantage of generalized estimating equations is that they can handle unequal intervals between observations.

In adjusting for correlated outcomes, generalized estimating equations change the standard errors but not the point estimates.

The effect of generalized estimating equations is to increase the standard errors (and therefore the confidence intervals) of the point estimates as a reflection of the correlation of clustered observations. However, generalized estimating equations do not change the point estimates (e.g., regression coefficients, odds ratios) themselves.

[16] In their classic test, Little, R. J. A. and Rubin, D. B. *Statistical Analysis with Missing Data.* New York, NY: John Wiley and Sons, 1987, distinguish three types of missing data: "data missing completely at random," "data missing at random," and "data not missing at random." However, in practice, most investigators distinguish just two groups: missing at random or not missing at random.

[17] Diggle, P. J., Heagerty, P., Liang, K-Y., and Zeger, S. L. *Analysis of Longitudinal Data* (second edn). Oxford: Oxford University Press, 2002, pp. 284.

[18] Twisk, J. W. R. *Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide.* Cambridge University Press, 2003, pp. 208–12.

For example, in a study related to the one described at the start of this chapter, McCrea and colleagues used generalized estimating equations to evaluate the effect of concussion on cognitive function among collegiate football players.[19] Injured athletes and uninjured controls were assessed on repeated occasions. Generalized estimating equations were needed to adjust for correlations between repeated observations of the same person. Potential confounders included in the model were baseline neuropsychiatric function, academic year, number of previous concussions, history of learning disability, and collegiate institution. The study found that athletes with concussion exhibited mild impairment in cognitive processing speed and fluency two and seven days after concussion compared to uninjured controls.

To run a model using generalized estimating equations you will need to specify three things:

(1) a link function
(2) a working correlation matrix
(3) a method for estimating the variance–covariance matrix

The link function is the type of model you are using to fit to your data. Common choices are linear, logit, or log. Each model has a corresponding distribution of the random component (Gaussian [normal], binomial or Poisson, respectively).

The working correlation matrix indicates how the observations of each cluster are related to one another. You have several choices as shown in Table 12.3.

The exchangeable working correlation matrix assumes that any two correlated observations are correlated equally.

The most common choice for a working correlation matrix for correlated observations is the exchangeable working correlation matrix (also referred to as a compound symmetric working correlation). This structure assumes that any two correlated observations are correlated equally (e.g., the first and third observations of a particular subject have the same correlation as the first and second observations of that subject, but a [possibly] different correlation than that of the observations of a different subject, or the first and third child in a family have the same correlation as the first and second child in that family, but a [possibly] different correlation than two children from a different family).

In some cases the assumption of equal correlations within clusters will not be true. Certain observations may be more highly correlated than others (e.g., observations of the same individual that are close in time are usually more highly correlated than those that are far apart in time; siblings who are close in age may respond more similarly than siblings born far apart). The correlation matrix may

[19] McCrea, M., Guskiewicz, K. M., Marshall, S. W., *et al.* "Acute effects and recovery time following concussion in collegiate football players." *JAMA* **290** (2003): 2556–63.

Table 12.3 Types of working correlation matrices for generalized estimating equations.

Type	Description	When to use
Exchangeable working correlation matrix (also referred to as compound symmetric working correlation).	Assumes that any two observations within a cluster have the same correlation.	Most common choice for analyzing data for nonindependent observations.
M-dependent structure.	Assumes that the correlation of any two measurements an equal distance apart within a cluster are the same.	May use when correlations between repeated observations are known to decrease as the distance between them increases.
First-order autoregressive correlation model (also known as the exponential correlation model).	Assumes that the correlation within a cluster decreases (in an exponential fashion) as the distance between observations increases.	May use when correlations between repeated observations are known to decrease over time (e.g., longitudinal studies).
Independent correlation model.	Assumes that the repeated observations within a cluster are independent (are not correlated).	May be used for studies where the number of observations per cluster is small relative to the number of clusters.
Unstructured correlation model.	Makes no assumption about the correlation of observations within a cluster.	May use when the number of observations within a cluster is small and balanced. Otherwise computationally difficult.

be different for different groups of subjects (e.g., the observations of patients from small hospitals may be more highly correlated than the observations of patients at larger hospitals).

The M-dependent structure assumes that the correlations of observations measured an equal distance apart within a cluster are equal. In other words, the correlation of any two measurements six months apart is equal, the correlation of any two measurements a year apart is equal, etc. With the M-dependent structure you can also stipulate that the correlation of measurements that are separated far in time (time equals M) is zero.

The first-order autoregressive working correlation matrix (also known as the exponential correlation model) assumes that the correlation between repeated/related observations within a cluster decreases (in an exponential fashion) as observations are further apart. This is often the case with longitudinal studies.

The independent correlation matrix assumes that the repeated measurements within a cluster are independent. You may wonder why I would even include it among your choices for dealing with correlated outcomes. The reason is that

when the number of clusters is large relative to the number of observations per cluster, the impact of the correlation may be small enough to ignore. When using generalized estimating equations with an independent correlation matrix and a normally distributed interval outcome, the procedure is the same as fitting a linear regression model.[20]

As implied by the name, the unstructured working correlation matrix makes no assumption about how the data within a cluster are correlated. This may seem like a major advantage since the true correlation matrix may be unknown. However, when there are many observations and/or a varying number of observations, the unstructured working correlation matrix is hard to estimate accurately.

Although the goal is to choose a working correlation structure that fits your data, it turns out that the generalized estimating equation analysis only needs a rough estimate of the true correlation structure to get started; the final parameter estimates are not generally dependent on the accuracy of the choice of working correlation matrix.[21] Therefore you should take into account the computational difficulty of the different working correlation matrices. For this reason, the exchangeable structure, which is computationally easier than the M-dependent, autoregressive or unstructured correlation structure, is often chosen.

TIP

In most cases you will want to specify an exchangeable working correlation matrix.

The variance–covariance structure is a function of the working correlation and the link function. The two commonly used methods for estimating it are the Huber–White sandwich estimator and a model-based estimate. In general, the Huber–White sandwich estimator is preferred because it makes no assumption about the variance model. Unfortunately, the estimator is biased downward (i.e., the standard errors will not be accurately inflated, resulting in overstating the significance of the results) when the number of clusters is small (< 40).[22] Therefore the model-based estimate is preferred when the number of clusters is small, especially less than 20.[23]

TIP

Unless the number of clusters is small, use the Huber–White sandwich estimator.

Once you have chosen your link function, a working correlation matrix, and a method for estimating the variance–covariance matrix, you can conduct your analysis. Your output will look similar (i.e., coefficients, standard errors, *P* values for your different risk factors) to a standard analysis (e.g., linear regression, logistic regression, or Poisson regression) but the standard errors will be adjusted for the correlation within your clusters.

[20] Davis, C. S. *Statistical Methods for the Analysis of Repeated Measurements.* New York, NY: Springer-Verlag, 2003, p. 297.

[21] Dupont, W. D. *Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data.* Cambridge: Cambridge University Press, 2002, pp. 356–67.

[22] Murray, D. M., Varnell, S. P., and Blitstein, J. L. "Design and analysis of group-randomized trials: A review of recent methodological developments." *Am. J. Pub. Health* **94** (2004): 423–32.

[23] Horton, N. J. and Lipsitz, S. R. "Review of software to fit generalized estimating equation regression models." *American Statistician* **53** (1999): 160–9.

12.1.C Mixed-effects model

Perhaps the most confusing aspect of mixed-effects models is all the different names that this procedure goes by in the literature: mixed models, random effects regression models, random coefficient models, random-regression models, multilevel models, and hierarchical models. Some authors add the word linear to describe them, as in: linear mixed-effects models or hierarchical linear models. The linear is added to distinguish them from nonlinear mixed-effects models, which can also be constructed.

The different names of these models refer to different features of them. Mixed-effects models are referred to as "mixed" because they contain both fixed and random effects. The models assume that individuals deviate randomly from the average (fixed) response. The underlying fixed model may be linear, logistic or Poisson. Because of the random effect, the slope and intercept of each individual subject may be different.[24]

These models are referred to as "multilevel" or "hierarchical" because they incorporate two or more "levels" of "random" variation where one level is "higher" than the other. For example, 1000 patients (level 1) may be cared for by one of 100 treating physicians (level 2) working in one of ten hospitals (level 3). Observations are correlated (clustered) at each of these levels.

Although you might think from this explanation that you cannot use these models to analyze repeated observations of the same subject, this is not true. Repeated observations (level 1) are clustered at the "higher" level of the subject (level 2). Similarly, observations of two body parts (level 1) of the same subject are clustered at the level of the subject (level 2) as well.

An example of clustered data that does not fit a hierarchical model would be surgical complication rates of physicians who work in multiple different hospitals. Because the physicians do not exclusively work in one hospital, they cannot all be fitted within a higher level. Although multiple-level nonhierarchical mixed-effects models are available, they are computationally difficult and their use is controversial.[25]

As with generalized estimating equations, mixed-effects models can incorporate interval, dichotomous, ordinal, and categorical outcomes, as well as independent variables that are fixed and those that may change value at each observation.

Unlike generalized estimating equations, for which you must specify a correlation structure, mixed-effects models assume that the correlation within the

[24] Ker, H.-W., Wardrop, J., and Anderson, C. Applicaton of linear missed-effects models in longitudinal data: A case study. http://www.hiceducation.org/edu_proceedings/hsiang-wei%20ker.pdf.

[25] Panageas, S. S., Schrag, D., Riedel, E., *et al.* "The effect of clustering of outcomes on the association of procedure volume and surgical outcomes." *Ann. Intern. Med.* **139** (2003): 658–65.

cluster arises from the common random effects of the cluster. It is also possible to specify a variance and correlation structure, and use your data to estimate its parameters.[26]

As is true of generalized estimating equations, mixed-effects models can handle unequal intervals between observations, unequal numbers of observations per cluster, and randomly missing data. Again as with generalized estimating equations, some authors report that mixed-effects models can handle data that are not missing randomly when the outcome is interval (but not when the outcome is dichotomous).[27] However, I would be cautious about using these models if you have a lot of missing data (because you can never be sure if the data are missing randomly), or when you know the data are not missing randomly.

When your outcome variable is dichotomous, the meaning of the coefficients differs between generalized estimating equations and mixed-effects models. With generalized estimating equations, the coefficient is the between-person difference in the log odds of the outcome comparing the effect of the intervention to no intervention (or the effect of being in one group to being in a different group) as if the intervention and the no intervention (or the group assignment) had been performed on two separate individuals. The coefficient in the mixed-effects models is the within-person change in the log odds of the outcome comparing the effect of the intervention to no intervention as if the intervention had been performed on the same individual.[28] Unlike generalized estimating equations, mixed-effects models can change the point estimate of effect as well as the confidence intervals.

Kandel and colleagues used a mixed-effects model to study racial and ethnic differences in cigarette smoking among adolescents.[29] The data were drawn from a representative sample of 90 118 adolescents. The investigators used a three-level hierarchical model: adolescents (level 1), school (level 2), and state (level 3). Observations of adolescents within the same school would be expected to be correlated because of characteristics particular to each school (e.g., principal, equipment) and observations from schools within the same state would be expected to be correlated because of similar statewide education policies. A hierarchical structure works because each school is in a particular state and children do not attend more than one school. The study found that transition to

[26] Gueorguieva, R. and Krystal, J. H. "Move over ANOVA: Progress in analyzing repeated-measures data and its reflection in papers published in the Archives of General Psychiatry." *Arch. Gen. Psychiatry* **61** (2004): 310–17.

[27] Twisk, J. W. R. *Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide.* Cambridge University Press, 2003, pp. 208–12.

[28] Murray, D. M., Varnell, S. P., and Blitstein, J. L. "Design and analysis of group-randomized trials: A review of recent methodological developments." *Am. J. Public Health* **94** (2004): 423–32.

[29] Kandel, D. B., Kiros, G.-E., Schaffran, C., *et al.* "Racial/ethnic differences in cigarette smoking initiation and progression to daily smoking: A multilevel analysis." *Am. J. Public Health* **94** (2004): 128–35.

TIP

Choose mixed-effects models over generalized estimating equations when you have a small number of clusters or when your focus is predicting outcomes of specific individuals.

daily smoking was significantly higher among white and Hispanic youth than among black youth.

An advantage of mixed-effects models over generalized estimating equations is that they require no minimum sample size for a particular group.[30]

In addition, if your focus is on predicting the values of individuals rather than the mean of the population, mixed-effects models are a better choice than generalized estimating models.[31] For example, it would be better to use a mixed-effects model for predicting the bone mineral density of a 66-year-old Caucasian woman with a history of smoking and alcohol consumption, and use the generalized estimating model to estimate the impact of age, smoking and alcohol consumption on average bone marrow density in a sample of community elders.

With the exception of situations where you have a small number of groups or where you are predicting the response of particular individuals (two instances where mixed-effects models appear superior) there is no strong reason for choosing mixed-effects models over generalized estimating equations for adjusting for correlations between observations.[32]

Of note, two articles on the same topic (relationship of procedure volume on mortality in cardiac patients) published in the same issue of the same journal used different techniques: Magid and colleagues used generalized estimating equations to adjust for within-hospital clustering and McGrath and colleagues used mixed-effects models to adjust for clustering of physicians.[33]

12.1.D Repeated measures analysis of variance/repeated measures analysis of covariance

DEFINITION

Repeated measures analysis of variance is used to compare means for two or more groups on repeated observations of an interval variable.

Repeated measures analysis of variance is an adaptation of analysis of variance. It is used to compare means of an interval outcome measure (e.g., cholesterol level) when you have two or more groups defined by an experiment or characteristic (e.g., ethnicity). You can also incorporate additional categorical independent variables into repeated measures analysis of variance (e.g., occupation).

[30] Christiansen, C. L. and Morris, C. N. "Improving the statistical approach to health care provider profiling." *Ann. Intern. Med.* **127** (1997): 764–8.

[31] Diggle, P. J., Heagerty, P., Liang, K.-Y., *et al. Analysis of Longitudinal Data* (second edn). Oxford: Oxford University Press, 2002, p. 130; Davis, C. S. *Statistical Methods for the Analysis of Repeated Measurements.* New York, NY: Springer-Verlag, 2003, pp. 294–8.

[32] For more on the comparisons between generalized estimating equations and mixed-effects models see: Twisk, J. W. R. *Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide.* Cambridge University Press, 2003, pp. 91–3.

[33] Magid, D. J., Calonge, B. N., Rumsfeld, J. S., *et al.* "Relation between hospital primary angioplasty volume and mortality for patients with acute MI treated with primary angioplasty vs thrombolytic therapy." *JAMA* **284** (2000): 3131–8; McGrath, P. D., Wennberg, D. E., Dickens, J. D., *et al.* "Relation between operator and hospital volume and outcomes following percutaneous coronary interventions in the era of the coronary stent." *JAMA* **284** (2000): 3139–44.

Unlike generalized estimating equations and mixed-effects models, repeated measures analysis of variance can only accommodate interval outcome variables and independent variables that are fixed (i.e., do not change their value during the course of the study).

Repeated measures analysis of variance requires an equal number of observations for each subject made at the same time.

Another disadvantage of repeated measures analysis of variance is that you must have the same number of observations of each subject and the observations must be made at the same time. This explains why repeated measures analysis of variance and covariance is more likely to be used with small experiments than with larger observational studies. If you have missing data, as is almost always the case with observational data, you must either impute the data point, carry the last observation forward or drop the case from the analysis. Any of these strategies, especially the latter two, are likely to bias your analysis.

TIP

When you have clusters with an unequal number of observations use generalized estimating equations or mixed-effects models.

Although standard repeated measures models require an equal number of observations, repeated measures analysis of variance for an unequal numbers of observations (referred to as unbalanced designs) exist.[34] However, these methods are complicated and require consultation with a biostatistician. When you have clusters with an unequal number of observations, it is generally better to choose generalized estimating equations or mixed-effects models.

Despite these limitations, repeated measures analysis of variance is a perfectly acceptable method of analyzing small experimental studies with an equal number of observations made at the same point in time in each cluster. For example, Schmidt and colleagues used repeated measures analysis of variance to tease out the cause of premenstrual syndrome.[35] Ten women with premenstrual syndrome were compared to 15 women without the syndrome on their response to leuprolide (a gonadatropin-releasing hormone agonist) or leuprolide plus hormone replacement. The investigators found a significant interaction between treatment group (leuprolide alone or with hormone replacement), diagnosis (women with premenstrual syndrome or not), and week of study. Specifically, women with premenstrual syndrome experienced greater sadness and more anxiety at week three after receiving leuprolide and hormone replacement than after receiving only leuprolide. Also at week three, women with the premenstrual syndrome who received leuprolide plus hormone treatment experienced more sadness and anxiety than women without this syndrome. The findings suggest that women with premenstrual syndrome experience an abnormal response to normal hormonal changes.

[34] Jennrich, R. I. and Schluchter, M. D. "Unbalanced repeated-measures models with structured covariance matrices." *Biometrics* **42** (1986): 805–20.

[35] Schmidt, P. J., Nieman, L. K., and Danaceau, M. A. "Differential behavioral effects of gonadal steroids in women with and in those without premenstrual syndrome." *N. Engl. J. Med.* **338** (1998): 209–16.

Table 12.4 Effect of cognitive behavior therapy (treatment group) versus usual therapy (control group) for patients with hypochondriasis.

	Treatment group ($n = 102$) mean (95% CI)	Control group ($n = 85$) mean (95% CI)	P value
Whiteley index baseline	3.58 (3.47–3.68)	3.51 (3.38–3.62)	<0.001
6-month follow-up	2.82 (2.68–2.97)	3.21 (3.05–3.38)	
12-month follow-up	2.65 (2.48–2.81)	3.02 (2.85–3.21)	

Barsky, A. J. and Ahern, D. K. "Cognitive behavior therapy for hypochondriasis: A randomized controlled trial." *JAMA* **291** (2004): 1464–70.

Repeated measures analysis of covariance is similar to repeated measures analysis of variance but it allows for incorporation of continuous independent variables.

For example, Barsky and Ahern studied the efficacy of cognitive behavior therapy for hypochondriasis (a persistent fear that one has a serious undiagnosed medical illness).[36] Clients were randomized to treatment or to usual care. The effect of treatment was measured using the Whiteley index (a measure of hypochrondriacal symptoms). Changes on the Whiteley index over time are shown in Table 12.4.

You can see that the mean Whiteley index shows that symptoms decreased in both groups over time. To test whether the decrease over time was significantly greater in the treatment group than in the control group, the investigators used repeated measures analysis of covariance; this procedure was used because the investigators wished to adjust their results for several variables, including psychiatric comorbidity, a continuous measure. The repeated measures ANCOVA showed that there was a significant interaction effect for group by time.

Repeated measures analysis of variance and covariance require that the data meet the sphericity assumption.

Besides the other assumptions of analysis of variance (Section 3.3), repeated measures analysis of variance and of covariance assumes sphericity. Sphericity in a longitudinal study means that the correlation between any two measurements at different time points is the same and that within subjects there is equal variance of the measurements. (The former is similar to the assumption of an exchangeable working correlation matrix in generalized estimating equations.) The sphericity assumption is always met if you have only two measurements. But when you have three or more measurements, this assumption is often not met. Observations

[36] Barsky, A. J. and Ahern, D. K. "Cognitive behavior therapy for hypochondriasis: A randomized controlled trial." *JAMA* **291** (2004): 1464–70.

close in time to one another tend to be more highly correlated than observations taken far apart. Variability of the measurements tends to increase over time.

TIP

Use the Mauchly test to assess whether your data fit the sphericity assumption.

You can use the Mauchly test to assess whether your data fit the sphericity assumption. The test is available in standard statistical software programs. The null hypothesis is that the data fit the sphericity assumption. If the *P* value is less than 0.05 you would reject the null hypothesis and conclude that your data do not meet the sphericity assumption. Unfortunately, the Mauchly test is very sensitive to sample size. With a large sample size, you may get a significant result even though the departure from the assumption is small, and with a small sample size you may get an insignificant result even though the departure from the assumption is large.[37]

TIP

Use the Greenhouse–Geisser correction when your data do not fulfill the sphericity assumption.

If this assumption is not met, you can use the Greenhouse–Geisser correction. This correction reduces the degrees of freedom for the numerator; this in turn has the effect of increasing the *P* value. It is a conservative adjustment that makes it harder to detect differences between the groups. Therefore, if you find differences despite the adjustment you can have greater confidence that your results are robust to violations of the sphericity assumption. For example, the trial of cognitive behavioral therapy for hypochondriasis reported above used the Greenhouse–Geisser correction in reporting their results, including the results shown in Table 12.4.

When you have more than one dependent variable use multivariate analysis of variance or covariance.

When you have more than one outcome variable, you should use the multivariate forms of repeated measures analysis of variance and repeated measures analysis of covariance: multivariate repeated measures analysis of variance and multivariate repeated measures analysis of covariance. If this analysis is statistically significant, then you can look at the individual outcome variable.

One advantage of the multivariate approach is that the sphericity assumption does not need to be met. However, if you then wish to tease out the effects of the risk factors on each of the outcome variables, you will still need to test for departures from the sphericity assumption and adjust for violations of this assumption. A disadvantage of multivariate designs is that you lose one degree of freedom for each added outcome variable, making your analyses less powerful.

In general, as shown in Table 12.5., repeated measures analysis of variance is much less flexible than generalized estimating equations or mixed-effects models. For this reason, its use is declining.[38]

[37] Twisk, J. W. R. *Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide.* Cambridge University Press, 2003, pp. 25–6.

[38] Gueorguieva, R. and Krystal, J. H. "Move over ANOVA: Progress in analyzing repeated-measures data and its reflection in papers published in the Archives of General Psychiatry." *Arch. Gen. Psychiatry* **61** (2004): 310–7. Besides documenting the decline in the use of ANOVA methods for repeated measures, this article provides a lucid explanation of many aspects of the analysis of clustered data.

Table 12.5 Comparison of generalized estimating equations, mixed-effects models and repeated measures analysis of variance.

Method	Types of outcome variables accommodated	Accommodate covariates that change value during study?	Accommodate unequal number of observations?	Accommodate missing observations?	Accommodate unequally spaced observations?
Generalized estimating equations	Interval, ordinal, dichotomous, and categorical	Yes	Yes	Yes, when data are missing randomly	Yes
Mixed-effects models	Interval, ordinal, dichotomous, and categorical	Yes	Yes	Yes, when data are missing randomly	Yes
Repeated measures analysis of variance	Interval only	No	No	No	No

12.1.E Conditional logistic regression

Clustered data with a dichotomous outcome can be analyzed using conditional logistic regression.

Clustered data with a dichotomous outcome can be analyzed using conditional logistic regression. This procedure is very similar to standard logistic regression. It produces similar outputs (e.g., coefficients, standard errors, odds ratios) and you can use similar diagnostics to assess the fit of your model.[39] In fact, "standard" logistic regression is often referred to as unconditional logistic regression. The difference between these two procedures is that conditional logistic regression takes into account the correlation between the observations.

As with standard logistic regression, conditional logistic regression does not accommodate variables that change their value during the course of the study. Unlike generalized estimating equations and mixed-effects models, it also cannot adjust for correlated outcomes owing to more than one cause (e.g., matching and clustering by sites). But when you have a dichotomous variable and observations correlated for a single reason, conditional logistic regression may be easier to explain and perform than generalized estimating equations or mixed-effects models.

For example, Abenhaim and colleagues were interested in evaluating whether appetite-suppressing drugs cause primary pulmonary hypertension.[40] Because primary pulmonary hypertension is a rare disease with a large number of

[39] Hosmer, D. W. and Lemeshow, S. *Applied Logistic Regression*. New York, NY: Wiley, 1989, pp. 187–215.
[40] Abenhaim, C., Moride, Y., Brenot, F., *et al.* "Appetite-suppressant drugs and the risk of primary pulmonary hypertension." *N. Engl. J. Med.* **335** (1996): 609–16.

suspected but unproved risk factors, the investigators chose a matched design. Ninety-five patients with primary pulmonary hypertension were individually matched to 335 controls by age (within five years), gender, and the number of visits to the physician per year. However, beyond the variables that they used to match cases and controls, they also needed to adjust for other potential confounders including systemic hypertension, use of cocaine, and smoking status. They therefore needed to use a multivariable technique. They found, using conditional logistic regression, that after adjustment for potential confounders, use of appetite suppressants was significantly associated with an increased risk for primary pulmonary hypertension (OR = 6.3; 95% CI = 3.0 − 13.2). They also showed a dose–response relationship, with subjects who used appetite suppressants for more than three months having a risk of development of hypertension 23 times greater than that of persons who did not use appetite suppressants (OR = 23.1; 95% CI = 6.9 − 77.7).

In situations where you have a choice as to which method to use to analyze data with correlated observations, you may want to analyze the data more than one way and see whether or not you get similar answers. For example, Sethi and colleagues looked at the association between the isolation of a new strain of a bacterial pathogen and an exacerbation of chronic obstructive pulmonary disease.[41] Because patients made multiple visits (81 patients made a total of 1975 visits) they needed to adjust for correlations within patient clusters. Their outcome was dichotomous: exacerbation or not. They reported similar results using conditional logistic regression and using generalized estimating equations.

12.1.F Anderson–Gill formation of the proportional hazards model

When you have censored data (Section 3.6) and an outcome that can occur more than once to a subject over time, you can use the Anderson–Gill counting process, which is an adaptation of the proportional hazards model.[42] In this model subjects are considered at risk for the first event from the start of the study to the first event; they are then considered at risk for the second event from the day following the first event until the second event occurs and so on. While this gives you a method of incorporating all of the outcomes as well as all of the person-time, you still must account for the correlation between events in the same individuals. To do this use a robust variance estimate.[43]

[41] Sethi, S., Evans, N., Grant, B. J. B., *et al.* "New strains of bacteria and exacerbations of chronic obstructive pulmonary disease." *N. Engl. J. Med.* **347** (2002): 465–71.

[42] Anderson, P. K. and Gill, R. D. "Cox's regression model for counting processes: a large sample study." *Ann. Stat.* **10** (1982): 1100–20.

[43] Lin, D. Y. and Wei, L. J. "The robust inference for the Cox proportional hazards model." *J. Am. Stat. Assoc.* **84** (1989): 1074–8.

For example, Berl and colleagues compared the incidence of congestive heart failure between patients receiving irbesartan (an angiotensin-receptor blocker) and those receiving placebo among diabetics with nephropathy.[44] Because congestive heart failure can occur more than once in the same individual the investigators used the Anderson–Gill formulation of the proportional hazards model with robust variance estimates. They found that patients receiving irbesartan had a significantly lower incidence of congestive heart failure than placebo recipients (hazard ratio = 0.72; 95% CI = 0.52–1.00; $P = 0.048$).

12.1.G Marginal approach for proportional hazards analysis

Another method for analyzing censored data with outcomes that can occur more than once to a subject over time is to model the marginal distribution of each time to outcome with a proportional hazards analysis.[45] An advantage of this approach is that the nature of the dependence of correlated observations is unspecified.

For example, Gabriel and colleagues used the marginal approach for proportional hazards analysis in a study of complications after breast implantation.[46] Most women in the study had bilateral implants; some had multiple implants in the same breast. The investigators therefore performed follow-up of each breast implant until a complication occurred, the implant was removed, or the end of follow-up occurred. Using a marginal approach to adjust for the correlation between times to implant failure for women who had more than one implant, they found that the rate of complication was significantly higher for women who had an implant for cancer or cancer prophylaxis than for those who had an implant for cosmetic reasons.

12.2 How do I calculate the needed sample size for studies with correlated observations?

The sample size calculations discussed in Section 7.4 assume that the observations are independent. Special methods are needed to determine sample size requirements for multivariable analysis involving clustered data. Although these methods are complicated, two points should be clear to you concerning sample size requirements for nonindependent observations.

[44] Berl, R., Unsicker, L. G., Lewis, J. B. *et al.* "Cardiovascular outcomes in the irbesartan diabetic nephropathy trial of patients with type 2 diabetes and overt nephropathy." *Ann. Intern. Med.* **138** (2003): 542–9.

[45] Wei, L. J., Lin, D. Y. and Weissfeld, L. "Regression analysis of multivariate incomplete failure time data by modeling marginal distributions." *J. Am. Stat. Assoc.* **84** (1989): 1065–73.

[46] Gabriel, S. E., Woods, J. E., O'Fallon, W. M., *et al.* "Complications leading to surgery after breast implantation." *N. Engl. J. Med.* **336** (1997): 677–82.

1. Designs for analyzing nonindependent observations will require larger sample sizes than you would estimate you would need if you assumed that each observation represented a unique subject. Why? Because, as explained in Section 12.1, the effect of adjusting for correlated observations is to increase the size of the standard errors and the resulting confidence intervals.
2. The greater the correlation of nonindependent observations within a cluster, the greater the sample size you will need. Why? Because the inflation of the standard errors is dependent on the size of the correlation.

Although beyond the scope of this text, instructions and software for calculating sample sizes for analyses of correlated outcome data are available.[47]

[47] Delucchi, K. L. "Sample size estimation in research with dependent measures and dichotomous outcomes." *Am. J. Public Health* **94** (2004): 372–7; Twisk, J. W. R. *Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide.* Cambridge University Press, 2003, pp. 280–5.

13

Validation of models

13.1 How can I validate my models?

A valid model is one where the inferences drawn from it are true.

TIP

Models rarely perform as well with new data as with the original data.

A valid model is one where the inferences drawn from it are true. Many factors can threaten the validity of a model including imprecise or inaccurate measurements, bias in study design or in sampling, and misspecification of the model itself.

Because the development of a model maximizes the probability of obtaining the values of the original outcome data, models will not generally perform as well with new data as with the original data. This is a particularly important issue when you are creating models to predict diagnosis or prognosis and a high degree of certainty is needed.

Although predictions based on the original cases will likely not be as accurate in predicting the outcome of new cases, the important question is how large is the decrement in performance. If the decrement is small, the model is said to be validated.

Rather than thinking of validation of models as a distinct activity, think of it as the extreme on a continuum of tests that you perform to evaluate the quality of your model. The continuum is shown in Figure 13.1.

The methods of model validation are:

1. Collect new data
2. Divide your existing data set:
   a. split-group
   b. jackknife method
   c. bootstrap

Without question, the best method of validating an empirical model is to collect more data and test the performance of the initial model with the new data. This was the case with the prognostic model for estimating survival for patients with primary melanoma (described in Section 2.1.D). The investigators found that four factors correctly classified the vital status (alive or dead) of 74 percent of the patients. To validate their model, they studied the success of this four-variable model in predicting outcome among 142 patients who were

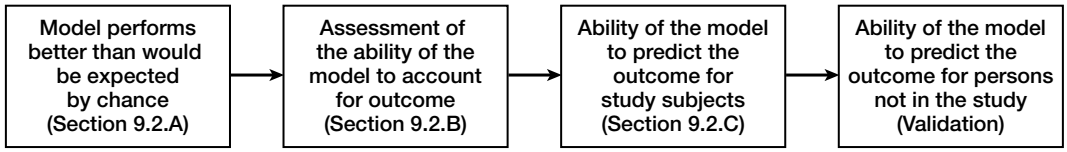


Figure 13.1 Continuum for assessing the quality of your model.

diagnosed with primary melanoma in the same center in the two years following the enrollment of the initial sample. When applied to this new sample, the model correctly classified 69 percent of the patients, a relatively small decrement in performance from the original model.

Although testing the model with a second set of patients strengthens the validity of the model, it is not as strong a validation as testing the model on patients seen at a different center. The reason is that a model may not perform as well under a different set of circumstances (e.g., a different prevalence of disease, referral pattern, patient mix, clinician practice style or temporal changes). In the case of the melanoma prediction rule the only difference between the circumstances of the original and the second data set was the year of diagnosis, which makes it a somewhat less rigorous validation than if the investigators had enrolled subjects from a different institution.

DEFINITION

With a *split-group* validation you randomly divide your data set into two parts – a derivation set and a validation set.

With a split-group validation you randomly divide your data set into two parts – a derivation set (also called a training set) and a validation set (also called a confirmatory set). The parts can be equal halves or you can split the data set such that the derivation set is larger than the confirmatory set. You develop your model on the derivation set and then test it on the confirmatory set.

A split-group validation was used to test the validity of a model designed to predict recurrence of seizures.[1] The sample consisted of 1013 people who were free of seizures on medication for at least two years. The researchers were taking advantage of an existing data set to develop a prognostic model. Collection of additional data was not an option. Instead, the investigators randomly divided their existing sample into two parts, with 60 percent of their sample in the derivation set and 40 percent in the validation sample.

Using the derivation set they developed a proportional hazards model with eight prognostic factors. To validate the model they used it to estimate the probability of a recurrent seizure for each of the subjects in the validation data set. They grouped the estimated probabilities of seizure for the patients in the validation group into eight groups of increasing probability of recurrence. As shown in Figure 13.2, for each of the eight groups they compared the proportion of

[1] Medical Research Council Antiepileptic Drug Withdrawal Study Group. "Prognostic index for recurrence of seizures after remission of epilepsy." *Br. Med. J.* **306** (1993): 1374–8.

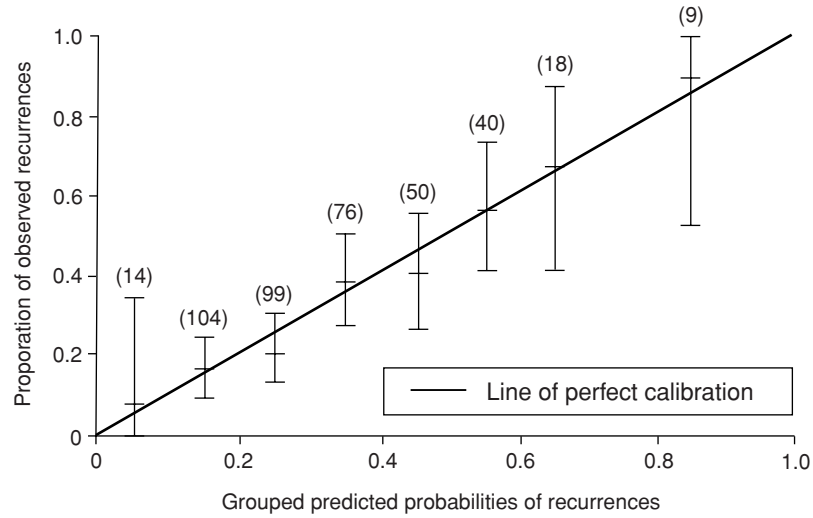


Figure 13.2 **Comparison of the probability of predicted recurrences (probabilities are based on the model generated by the derivation set) to the observed probability of recurrences of seizures in the validation set. The bars show the confidence intervals for the predicted values, and the dashes near the middle of the line show the mean predicted value. Reprinted with permission from Medical Research Council Antiepileptic Drug Withdrawal Study Group. "Prognostic index for recurrence of seizures after remission of epilepsy." *Br. Med. J.* 306 (1993): 1374–8. Copyright BMJ Publishing Group.**

actual recurrences to that predicted by the model. The bars show the confidence intervals for the predicted values, and the dash near the middle of the line shows the mean predicted value. If the validation had been perfect, all dashes would fall exactly on the diagonal line. While the dashes are close to the diagonal line, the confidence intervals are broad, especially for those predicted values where the number of subjects (shown in parentheses) is small.

You will note that Figure 13.2 uses essentially the same technique as Figure 9.1. Figure 9.1 is also based on comparing predicted to observed probabilities. The difference is that in Figure 9.1 the predicted and observed probabilities are based on the same subjects. In Figure 13.2 the predicted probabilities are based on a model that was derived from a different group of subjects.

One final note about validating your model using a second sample or a split sample. Once a model has been validated, investigators will often combine the multiple samples or reunite the split sample for the final model. Investigators do this so that the larger sample size can give their final model tighter confidence intervals.

In cases where it is impractical to collect more data or split your sample, you may use a jackknife procedure (often called cross-validation). With a jackknife

DEFINITION

With a *jackknife* procedure you sequentially delete subjects from your data set, one at a time, and recompute your model with each subject missing once.

procedure you sequentially delete subjects from your data set, one at a time, and recompute your model with each subject missing once. This allows you to assess two things. First, you can assess the importance of any one subject to your results. A model that substantially changes with deletion of a single case is not valid because the results hinge on that one case. Second, once you drop a case, you can predict that subject's outcome from the remaining cases. This is done sequentially such that you are predicting the values of each subject using the rest of the subjects. In this sense, the jackknife is like a split-group validation: The split is the whole sample minus one case (the derivation set) versus the one case (the confirmatory set). When you have a relatively small sample, jackknife procedures are likely to be more sensible than splitting your sample. You should be aware that jackknife procedures are easy to do in multiple linear regression, but they are very computer time intensive in logistic and proportional hazards models.

DEFINITION

Bootstrap procedures take random samples of the subjects in your data set with replacement and average the results obtained from the multiple samples.

The bootstrap procedure provides limited support of the validity of your model. With bootstrap procedures you take random samples of the subjects in your data set with replacement (meaning that after a case is chosen it becomes eligible to be chosen again). Thus, your random samples may include the same subject more than once, whereas some subjects will not be included at all. Once the samples are drawn, you test the strength of the relationships found in your main model in the random samples. The results from these samples can be used to construct 95 percent confidence intervals by excluding the extreme 2.5 percent and 97.5 percent of values. If the confidence intervals are relatively narrow, you can feel more confident in your results.

For example, Hamberg and colleagues used logistic regression to create a model based on clinical and laboratory data that would accurately predict cirrhosis among 303 alcohol-abusing men.[2] Such a model would have clinical significance if it decreased the need for liver biopsies. At a cutoff of 10 percent probability of having cirrhosis, the sensitivity and specificity of their model was 88 percent. They then tested their model using the same cutoff in 1000 samples drawn from their study population. The 95 percent confidence intervals for the sensitivity and specificity were 85–91 percent and 85–95 percent respectively.

Besides producing confidence intervals, bootstrap procedures produce mean coefficients and mean standard errors for your random samples. These confidence intervals and standard errors are likely to be more valid than those created from one simple model.[3]

[2] Hamberg, K. J., Carstensen, B., Sorensen, T. I., *et al.* "Accuracy of clinical diagnosis of cirrhosis among alcohol-abusing men." *J. Clin. Epidemiol.* **49** (1996): 11:1295–301.

[3] Readers interested in learning more about bootstrap techniques should see Efron, B. and Tibshirania, R. J. *An Introduction to the Bootstrap.* New York, NY: Chapman & Hall, 1993.

The bootstrap procedure is a weaker test of the validity of a model than a split-group or a jackknife. This is because the 1000 bootstrap samples of 303 subjects will have the majority of subjects in common. In contrast, a split sample and the jackknife procedures would have no subjects in common. Thus, in the strictest sense, bootstrap does not fit the definition of validation.

The importance of validation varies with the goals of your study. Validation is rarely performed for studies identifying risk factors associated with a particular outcome while adjusting for confounders (Section 2.1.A). Your results will be judged primarily on the strength of your methods, the biologic plausibility of your results, and prior findings in this area. Other investigators may seek to replicate or refute your findings. In contrast, models predicting prognosis or diagnosis of disease (Sections 2.1.C and 2.1.D) are rarely published (at least not in the best journals) without validation.

The reason for the distinction is that models used to determine factors associated with a particular outcome do not need to be highly accurate. For example, while exercise is significantly associated with mortality, the decrease in mortality owing to exercise is relatively small. You certainly would not try to predict a patient's life span based on knowing how much he or she exercises. But that's not the point. The point is that, in a population, exercising will result in increased longevity for the group as a whole. Improving risk factor profiles may have a substantial effect on the development of disease in a large population, even if the absolute effect for an individual is small. This is especially true if the disease and the risk factor are common.

TIP

Clinicians don't trust models that are not validated.

In contrast, with studies designed to predict diagnosis or prognosis, a high degree of certainty is required because you are using the model to predict for an individual patient. Clinicians are not likely to trust a model that has not been validated. (Even then, physicians are notorious for ignoring diagnostic algorithms, preferring instead their gut instinct; see Section 2.1.C.)

14

Special topics

14.1 What if the independent variable changes value during the course of the study?

DEFINITION

Time-dependent covariates allow incorporation of changes in the independent variables that occur during the study.

Let's say that over the course of a longitudinal study a subject's value changes on an independent variable. This may happen because the patient quits (or starts) a habit such as smoking, begins a new medicine, or develops a new symptom or illness. How can you deal with this in your analysis? The answer is that within proportional hazards analysis you can create time-dependent variables. These variables change value at a particular point in time. So, instead of having a variable such as smoking at baseline (yes/no), you create a time-dependent variable, where each subject is 0 (nonsmoker) or 1 (smoker) at a particular point of survival time.

In the simplest case, time-dependent variables change their value only once (e.g., a nonsmoker starts to smoke – the variable is 0, 0, 0, at several points in time before the subject begins smoking and then the variable changes value to 1 at the time the subject begins smoking and remains 1 for the remainder of the observational time). It is also possible to construct time-dependent variables that change their value back and forth multiple times (reflecting what sometimes happens when smokers try to quit). Time-dependent variables need not be dichotomous; that is, the variable may take the value of an interval measure, such as blood pressure, at each point that it is measured.

While the interpretation of time-dependent variables can be complicated, their construction is easy. You need to look up the exact formatting in your statistical package, but in general, the package will have a special designation for time-dependent variables. You tell the computer when (in study time) each subject changes value on the variable.

14.2 What are the advantages and disadvantages of time-dependent covariates?

The advantage of time-dependent covariates is that you can incorporate important events that occur during the course of the study. For example, Mayne and colleagues wanted to determine if depression shortened survival of HIV-infected men.[1] Their subjects were part of a longitudinal study that began in 1984 with more than seven years of follow-up. The simplest design for answering this question would be to measure depression in 1984 and then follow subjects longitudinally for mortality. The problem with this design is that depression may not be present initially (in 1984) but may develop subsequently. Using depression at baseline only will weaken your study (because people who become depressed six months after baseline will not be considered depressed in the analysis, even though this may affect their mortality). Second, using depression only at baseline decreases your power because relatively few persons will be classified as depressed at one time. Third, a single instance of a participant being rated as depressed might not affect mortality, since it might reflect only a short episode of depression, rather than a more chronic condition.

To overcome these issues, the researchers created a time-dependent variable that took the value of the proportion of visits at which the person was depressed. So for each visit (subjects were interviewed every six months) the variable had a value between 0 percent and 100 percent (of visits to date) at which the person was depressed. They found that depression was associated with a higher rate of mortality. They also created time-dependent variables that represented each subject's actual score on the depression index at each visit. The results were similar.

The depression measure was not the only variable that changed value over the course of this study. The subjects' immune function also changed value as patients progressed. The investigators therefore created time-dependent covariates that represented the subjects' CD4 counts, as well as other measures of immune function. They found that depression increased the risk of death, even with adjustment for changes in immune function. This would suggest that the mechanism of depression on mortality is not mediated by more rapid immune function decline (within the limits of the investigators' ability to measure it). This analysis also weakens an "effect–cause" hypothesis (that participants became depressed because they learned of their CD4 count results), since the analysis adjusts for recent CD4 counts.

[1] Mayne, T. J., Vittinghoff, E., Chesney, M. A., *et al.* "Depressive effect and survival among gay and bisexual men infected with HIV." *Arch. Intern. Med.* **156** (1996): 2233–8.

Another advantage of time-dependent covariates is that they do not have to fit the proportionality assumption (Section 10.10). The reason is that they incorporate time and therefore do not have to be independent of time.

The disadvantages of time-dependent covariates are less obvious than the advantages. The two major disadvantages are: overadjustment and decreased usefulness of the model for clinicians.

Adjusting for prognostic markers that are on the pathway to your outcome may prevent you from identifying the effect you are investigating (overadjustment). Let's go back to the example of depression and mortality in HIV-infected persons. Assume that the overall finding is accurate, that is, that depression increases mortality. But, let's assume that the mechanism by which this happens is that depression leads to worsening immune function, which, in turn, leads to more opportunistic infections and death (in other words, change in immune function is an intervening variable between depression and mortality)(Section 7.3). If this is the case, then including time-dependent variables measuring immune function will eliminate the effect you are trying to substantiate. Depression will not be associated with mortality because adjusting for changes in immune function will eliminate the effect. In comparison, if you adjust only for immune function at baseline you will not eliminate the effect.

The use of time-dependent covariates may also decrease the value of your models for clinicians. The reason is that clinicians must advise their patients based on information they have at the time they are counseling the patient. A clinician can't know how a risk factor will change in the future. It would be confusing to a patient to counsel them on their risk of heart attack in case they *were* to develop hypertension at a particular time in the future.

TIP

With time-dependent covariates include only events that have occurred before the outcome.

With time-dependent models it is important to include only events that have occurred before the outcome. Remember that an advantage of a longitudinal design compared to a cross-sectional one is that a longitudinal study is more likely to support causality. That's because if the outcome is not yet present (at least, as best as can be measured), the chance that the "outcome" causes the "risk factor" (i.e., effect–cause) is much less likely. (Remember, in a cross-sectional study you are measuring risk factors and outcomes at the same time.) With time-dependent variables, you are including factors that are more proximal to outcome than baseline measurements. Thus, effect–cause becomes a greater danger. One way to deal with this issue is to "lag" the time-dependent measure substantially before the outcome (but still after the baseline).

A lag for time-dependent variables was used by the investigators to see if depression would still be associated with mortality if the depression measure was lagged by periods of one, two, and three years from the outcome. With

these time lags, depression was still associated with mortality, supporting the hypothesis that depression increases mortality, and weakening the case for the alternative hypothesis that depression reflects worsening health status.

14.3 What are classification and regression trees (CART) and should I use them?

Classification and regression trees (CART), also known as recursive partitioning, is a technique for separating (partitioning) your subjects into distinct subgroups based on the outcome.[2]

The technique is easiest to follow visually. In Figure 14.1, you see an algorithm for assessing the risk of heart attack that was developed using CART.[3] The algorithm is based on 1379 patients, of whom 259 (19 percent) had a heart attack. The investigators assessed the diagnostic ability of 50 variables, including patients' history, physical examination, and electrocardiogram results.

The CART technique attempts to divide the sample into subgroups that have as many patients with the outcome (e.g., heart attack) in one group (high risk) and as few patients with the outcome in the other group (low risk). You can see that at the first branch point of Figure 14.1 (ST elevation or Q waves in two or more leads, not known to be old), CART separates the sample into two groups with very different probabilities of heart attack: 80% and 9%. At the next branch point (chest pain began $\geq$48 hours ago), the sample is also separated into two groups with different probabilities of outcome (10% vs. 3%), although this difference is not as large as the difference in the first branch point. Selecting from the candidate variables, CART will continue partitioning until it reaches a point where it is no longer possible to partition the sample into subgroups with distinctly different risks of outcome. If your CART model partitions your sample into subgroups where the risks are not sufficiently distinct, you can prune your tree back.

What advantage does CART have over other multivariable techniques? It is similar to forward stepwise logistic regression in that you are estimating a dichotomous outcome by sequentially choosing the strongest risk factors for your outcome. The major difference between CART and forward multiple logistic regression is that with CART one branch can have different risk factors for outcome than a different branch. With multiple logistic regression your risk

[2] See the bible of CART: Breiman, L., Friedman, J. H., Olshen, R. A., *et al. Classification and Regression Trees.* Pacific Grove, CA: Wadsworth & Brooks, 1984, p. 189.

[3] Goldman, L., Cook, E. F., Brand, D. A., *et al.* "A computer protocol to predict myocardial infarction in emergency department patients with chest pain." *N. Engl. J. Med.* **318** (1988): 797–803.

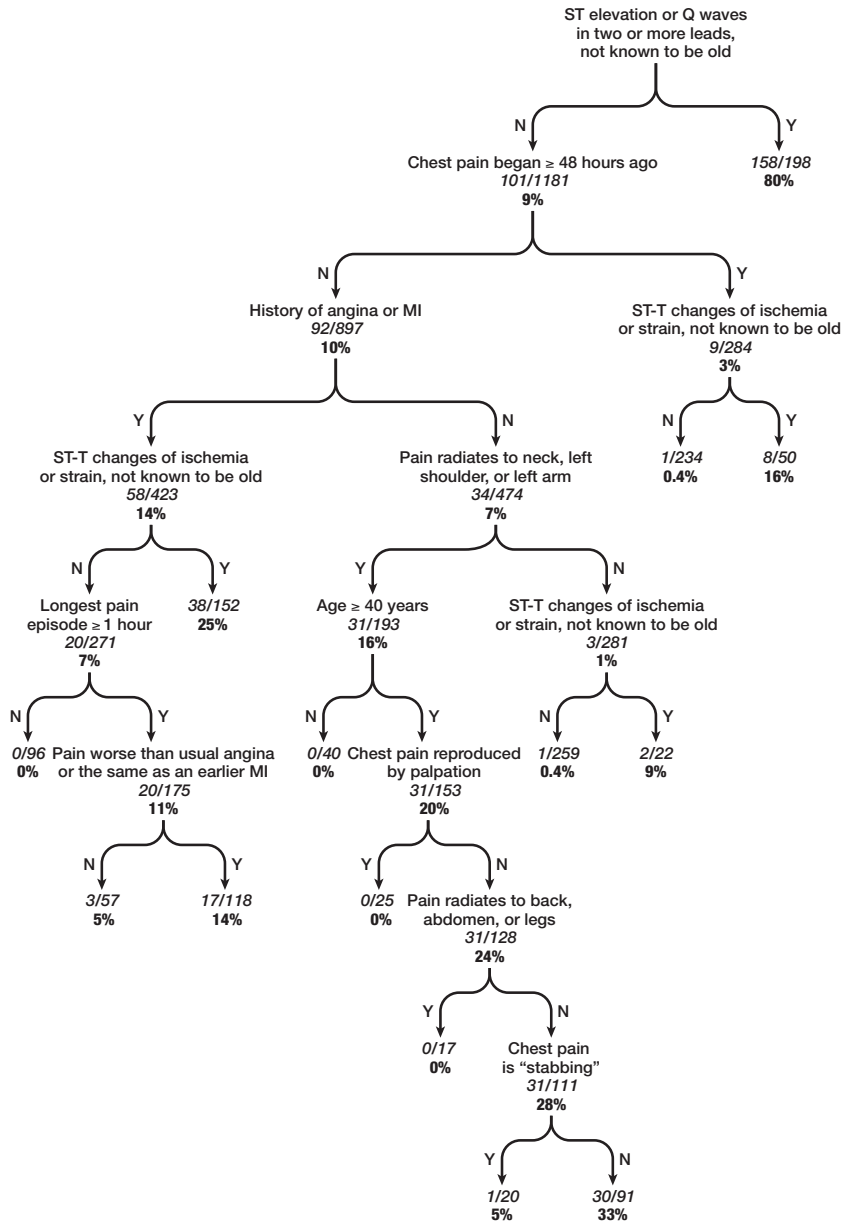


Figure 14.1 Classification and regression tree for predicting the likelihood that the patient has a myocardial infarction. The data are from Goldman, L., *et al.* "A computer, protocol to predict myocardial infarction in emergency department patients with chest pain." *N. Engl. J. Med*. **318** (1988): 797–803. The figure is adapted from Lee, T. H., *et al*. "Ruling out acute myocardial infarction." *N. Engl. J. Med*. **324** (1991): 1239–46.

The major difference between CART and forward logistic regression is that with CART one branch can have different risk factors for outcome than a different branch.

factors are for your entire sample, not one branch of it. For this reason, CART is better suited to data where there are interactions (because with interactions a variable may be important for only a portion of the sample) (see Sections 1.4, 8.3, 9.9, and 10.8).

An advantage of diagnostic trees is that compared with multiple logistic regression they more closely reflect how physicians make decisions. Certain pieces of information take you down a particular diagnostic path; you seek more information to prove or disprove that you are on the right path. Most clinicians do not, in their mind, total up all the information, positive and negative, and make a decision.

Having said that, clinicians have shown no greater willingness to adopt this decision rule than that of Pozen and colleagues (Section 2.1.C). When the authors attached their prediction tree to the back of the patient data forms in their own hospital, physicians looked at it in only 46 percent of the cases; in the 115 cases in which the prediction rule was used, it changed the triage decision only once. Moreover, the likelihood of using the rule decreased with increased level of physician training (i.e., interns used it more than residents, who used it more than attendings). This is despite the fact that the decision model shown in Figure 14.1 was shown to perform better than physicians at university and community hospitals when tested prospectively.

One disadvantage of the chest pain model is the large number of variables it includes. A widely used algorithm that was developed using CART predicts whether or not a patient has an ankle fracture based on only three variables.[4] The model, referred to as the Ottawa ankle rules, has a sensitivity of 100 percent for predicting fracture. Therefore, patients who are negative on the decision rule do not need to be sent for an x-ray film. This saves a great deal of money and time for the patient. The lower specificity of the model (50 percent) is not a major problem because at one time practically all patients with an ankle injury would have received an x-ray.

Because of their convenience and high sensitivity, the Ottawa ankle rules have received much wider acceptance than the heart attack prediction models. This should not, however, be taken as a negative reflection on heart attack prediction models. It is not surprising that it takes more variables to accurately predict a heart attack than a broken ankle, and that even with a large number of variables, there is greater uncertainty in the prediction of a heart attack than a broken ankle. It does highlight, however, that clinicians are more likely to adopt diagnostic rules that are simple and have high sensitivity.

[4] Stiell, I. G., Greenberg, G. H., McKnight, D., *et al.* "Decision rules for the use of radiography in acute ankle injuries: Refinement and prospective validation." *JAMA* **269** (1993): 1127–32; Stiell, I. G., McKnight, R. D., Greenberg, G. H., *et al.* "Implementation of the Ottawa ankle rules." *JAMA* **271** (1994): 827–32.

14.4 How can I get best use of my biostatistician?

Working with a biostatistician should be an iterative process, a dynamic interaction between the clinical details and the statistical realities of your study.

With a complicated study, a biostatistician should be consulted at each phase of the analysis. At the design phase, review with a biostatistician the statistical implications of different study designs and seek their help in conducting or reviewing your power calculation. After conducting your univariate and bivariate data analysis, discuss with your biostatistician strategies for dealing with skewed distributions, nonlinear relationships, multicollinearity, and missing data. Based on the preliminary analysis, determine together the best type of multivariable analysis to perform. Finally, once you have your multivariable model, review it with your biostatistician so as to assess whether the model fits. At this stage a review of the residuals may be particularly helpful.

You will find that all biostatisticians are not alike. Some are primarily interested in developing new methods of analyzing data (data at the service of methods). Others are interested in using methods for improving the analysis of the data (methods at the service of data). In general, you will do better if you have the latter type of biostatistician, although we wouldn't have so many useful statistical techniques if it were not for the former type.

Just as it helps to know more about your car in dealing with car mechanics, the more you know about your research project, and the statistical issues surrounding it, the more helpful your biostatistician will be to you. Or, to switch to a medical metaphor, think of yourself as the primary care doctor and the biostatistician as the specialist.

14.5 How do I choose which software package to use?

Almost all of the popular statistical packages (SAS, SPSS, BMDP, STATA, S-PLUS) perform the same types of analyses. The best one to choose will probably depend on what others in your research group use. Programming questions invariably arise and it is always helpful to have other users nearby.

As with foreign languages, some statistical packages are harder to learn than others, but once you know one it is easier to learn others. While I have not performed any formal polling, most medical researchers use SAS. On the one hand, SAS is somewhat more difficult to learn than the others, in part, because the manuals are poorly organized and confusing. On the other hand, SAS is more flexible and powerful than most of the others. The flexibility is particularly important for longitudinal studies, where you have multiple observations of the same person. The SAS package also allows you to write your own statistical

programs, but it also costs more than some of the others. The STATA package is running a close second to SAS in popularity because it can perform many of the same analyses of complicated longitudinal data as SAS (e.g., generalized estimating equations) but is easier to learn and use.

Some packages are particularly good at certain functions. The S-PLUS package has dramatically increased in popularity because it has fantastic graphing capabilities. R (http://www.r-project.org) is modeled after S-PLUS and is free, making it an excellent choice if you are doing research on your own with a tight budget. The software package MlwiN was specially created to perform multilevel models (Chapter 12). It is available free (http://multilevel.ioe.ac.uk/download/index.html) on the internet. The SUDAAN package is often chosen for analyzing weighted data sets.

15

Publishing your study

15.1 How much information about how I constructed my multivariable models should I put in the Methods section?

The editors of the major biomedical journals have developed guidelines on how much detail of the statistical analysis to include in manuscripts. While the guidelines are general, the editors articulate an important rule of thumb: "Describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to verify the reported results."[1]

Although that goal is important, anyone who has performed statistical analysis knows that it would be impossible to include every detail of the analysis in a manuscript. Imagine writing: "for each independent variable we assessed whether there was any difference in outcome between the 'don't know' category and the 'missing' category" or "for one variable, we found that there was a somewhat increased frequency of outcome in the 'don't know' versus the 'missing' category, so we . . ." I think you get the idea. Research requires thousands of decisions. The readers rely on you to make the right ones. It is your responsibility, however, to report on the important choices you made, especially those that influence the results.

Published articles and journals differ in how they organize the information in the Methods section. I prefer dividing the Methods section into a review of how subjects were enrolled (Subjects), what interventions were used or how data were acquired (Procedures), how the variables were coded (Measures), and how the data were analyzed (Statistical analysis). But some published articles group all of the information under a general heading of Methods. Before writing

[1] International Committee of Medical Journal Editors. "Uniform requirements for manuscripts submitted to biomedical journals." *Ann. Intern. Med.* **126** (1997): 36–47 (also available at www.icmje.org); See also: Moher, D., Schulz, K. F., Altman, D., *et al.* "The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials." *JAMA* **285** (2001): 1987–91; Des Jarlais, D. C., Lyles, C., Crepaz, N., *et al.* "Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: The TREND statement." *Am. J. Public Health.* **94** (2004): 361–6. For an excellent guide on writing up your research for publication see: Browner, W. S. *Publishing and Presenting Clinical Research.* Philadelphia, PA: Lippincott Williams and Wilkins, 1999.

this section for your manuscript, consult a recent issue of the journal to which you plan to submit your paper; you will get a sense of the journal's preferences. At a minimum, include a description of the following in your Methods section:

1. The population from which your subjects were chosen and your method of choosing them (e.g., probability sample of households in low-income census tracts, consecutive sample of patients presenting to a specialty clinic with sinusitis).
2. Your sample size. If any sampled subjects were excluded, explain why (e.g., three subjects were excluded because of insufficient blood samples).
3. Response rate, including differences between persons who participated in your study and those who did not. (Some journals prefer that this information be reported in the Results section.)
4. Nature of intervention (e.g., patients were randomized to one of three arms of drug therapy), if applicable.
5. How data were acquired (e.g., interviews, matches with registries).
6. How your independent and dependent variables were chosen and measured.
7. How your independent and dependent variables are categorized in the analysis (e.g., nominal, multiple dichotomous variables).
8. What bivariate statistics you used (e.g., chi-square statistics for categorical variables, *t* tests for interval variables).
9. What type of multivariable model you used (e.g., multiple linear regression, conditional logistic regression).
10. How you dealt with missing data.
11. What independent variables were eligible for inclusion in the model (e.g., all variables in your Table 15.1, those independent variables associated with the outcome at $P < 0.15$).
12. If you used a variable selection procedure, state type (e.g., forward, backward) and what the inclusion/exclusion criteria were (e.g., $P < 0.10$).
13. If you had censored observations, when you censored them (e.g., alternative outcomes, date of the end of the observation period).
14. How you tested the linearity assumption.
15. How you tested the proportionality assumption (for proportional hazards model).
16. Whether you tested for interactions and, if so, how.
17. What statistical software you used. (The reason for this is that some packages differ in their computational methods for certain statistics.)
18. Whether *P* values were one- or two-tailed.

Within the Methods section, my preference is to report 1–3 in the Subjects subsection, 4 and 5 in the Procedures subsection, 6 and 7 in the Measures subsection,

and 8–18 in the Statistical analysis subsection; but, journals and reviewers vary in their preferences.

15.2 Do I need to cite a statistical reference for my choice of multivariable model?

If you are doing a standard linear or logistic regression analysis it is unnecessary to provide a citation. Some researchers provide a citation for proportional hazards regression. It is always the same classic citation:

Cox, D. R. "Regression models and life tables." *J. R. Stat. Soc.* **34** (1972): 187–220.

Some of the techniques I discussed in Chapter 12 for analyzing correlated observations such as mixed-effects models and the Anderson–Gill counting process for proportional hazards regression are sufficiently unfamiliar that they should be cited explicitly.

15.3 Which parts of my multivariable analysis should I report in the Results section?

As with the question of what to include in the Methods section, there are no absolute rules on what results to report in your published paper.

Unless there are no missing data, you should report the *n* for each analysis. For multiple linear regression models, most investigators report the regression coefficients (standardized or unstandardized, but not both), the standard errors of the coefficients, and the statistical significance levels of the coefficients. As a test of how well the model accounts for the outcome, most researchers report the adjusted R^2.

For logistic regression and proportional hazards analysis, even though the coefficients are similar to those from linear regression, they are not generally reported. Instead, report the odds ratio or relative hazard and the 95 percent confidence interval. The latter incorporates information from the standard error and the *P* value, and so it is not necessary to report these as well. Reporting on how well the model accounts for outcome is variable. Some authors will report the likelihood ratio test or the Hosmer–Lemeshow test, or they may compare the estimated to observed probabilities of outcome in graphical or tabular form, or they may do none of the above (Sections 9.2.A and 9.2.B). For diagnostic or prognostic studies, some measure of prediction (sensitivity, correctly identified cases, c index) will usually be provided (Section 9.2.C).

There is an increasing tendency for clinical researchers to show only the results from the main variables of interest. This is illustrated in Table 15.1, reproduced

Table 15.1 Relative risk of cardiovascular disease among current users of conjugated estrogen alone or with progestin as compared with nonusers, 1978 to 1992.

		Major coronary disease			Stroke (all types)		
		No. of cases	Relative risk (95% CI)		No. of cases	Relative risk (95% CI)	
Hormone use	Person-years		*Age adjusted*	*Multivariate adjusted* †		*Age adjusted*	*Multivariate adjusted* †
Never used	304 744	431	1.0 (ref.)	1.0 (ref.)	270	1.0 (ref.)	1.0 (ref.)
Currently used							
Estrogen alone	82 626	47	0.45 (0.34–0.60)	0.60 (0.43–0.83)	74	1.13 (0.88–1.46)	1.27 (0.95–1.69)
Estrogen with progestin	27 161	8	0.22 (0.12–0.41)	0.39 (0.19–0.78)	17	0.74 (0.45–1.20)	1.09 (0.66–1.80)

† The analysis was adjusted for age (in five-year categories), time (in two-year categories), age at menopause (in two-year categories), body-mass index (in quintiles), diabetes (yes or no), high blood pressure (yes or no), high cholesterol level (yes or no), cigarette smoking (never, formerly, or currently [1 to 14, 15 to 24, or 25 or more cigarettes per day]), past oral-contraceptive use (yes or no), parental history of myocardial infarction before the age of 60 years (yes or no), and type of menopause (natural or surgical).

from a study of the risk of cardiovascular disease and stroke in women.[2] The investigators assessed, using proportional hazards analysis, whether current estrogen or current estrogen/progestin increased the risk of coronary artery disease or stroke.

In the published table, the investigators report the person-years, the number of cases of outcome for the analysis, the relative risk (I would prefer the term relative hazard), and the 95 percent confidence intervals. The relative risk was adjusted first for just age and then for numerous other variables that affect the likelihood of coronary disease and stroke. These variables are listed at the bottom of the table next to the cross sign. Although each of these variables has a relative risk and a 95 percent confidence interval, these values are not presented. Based on other studies, we know that some of the variables listed at the bottom of the table (e.g., age, smoking) are significantly associated with coronary artery disease and stroke.

The advantage of not showing the relative risks and confidence intervals for these variables is that these results would take up a whole page. Also, this study was designed to show the influence of estrogen and estrogen/progestin use on coronary artery disease and stroke. It was not designed to estimate the effect of

[2] Grodstein, F., Stempfer, M. J., Manson, J. E., *et al.* "Postmenopausal estrogen and progestin use and the risk of cardiovascular disease." *N. Engl. J. Med.* **335** (1996): 453–61.

age or cigarette smoking on these outcomes. Other studies have answered these questions.

However, there are disadvantages to showing your data in this way. The reader cannot assess whether the impact of estrogen and estrogen/progestin use on coronary disease and stroke is as strong (or stronger) as other factors, such as smoking. Second, your study may be less helpful to future researchers. For example, if someone was doing a meta-analysis on the effect of past oral-contraceptive use (a variable that has been inconsistently related to outcomes such as coronary disease and stroke) they wouldn't learn anything from the table. You don't know whether oral contraceptives are or are not related to the outcomes in this study.

Finally, when evaluating a published report, one feels greater confidence if the independent variables operate in ways you would expect them to based on prior research. For example, if the investigators reported that cigarette smoking was related to coronary artery disease and stroke it would give you confidence that their model was sound. Conversely, if smoking was not related to these outcomes you would worry about the validity of their model.

All this being said, demands and cost of journal space are likely to dictate more presentations of data like Table 15.1. With large models, and multiple independent variables, it is hard to show all the results. One solution to this dilemma is gaining in popularity: Investigators inform readers where they can obtain the full analysis (usually by contacting the authors or by accessing the journal's website). This seems a good balance between publishing extensive tables and having results available to the public.

TIP

If you are unable to publish your full analysis, make a detailed version of your results available on the web.

16

Summary: Steps for constructing a multivariable model

Step 1. Based on the type of outcome variable you have, use Table 3.1 to determine the type of multivariable model to perform (if you have repeated observations of your outcome see Table 12.2).

Step 2. Perform univariate statistics to understand the distribution of your independent and outcome variables. Assess for implausible values, significant departures from normal distribution of interval variables, gaps in values, and outliers (Section 5.7).

Step 3. Perform bivariate analysis of your independent variables against your outcome variable.

Step 4. If you have any nominal independent variables transform them into multiple dichotomous ("dummied") variables (Section 4.2).

Step 5. Assess whether your data fit the assumptions of your multivariable model (linearity, normal distribution, equal variance) on a bivariate basis (Chapter 5). Transform or group any variables that show significant bivariate departures from the assumptions of your model (Sections 5.4, 5.5, 5.7, and 5.8).

Step 6. Run a correlation matrix. If any pair of independent variables are correlated at $>$ 0.90 (multicollinearity), decide which one to keep and which one to exclude. If any pair of variables are correlated at 0.80 to 0.90 consider dropping one (Chapter 6).

Step 7. Assess how much missing data you will have in your multivariable analysis. Choose a strategy for dealing with missing cases from Table 7.4.

Step 8. Perform the analysis (Chapter 8).

Step 9. Review the multivariable correlation matrix to assess for multicollinearity. If you have evidence of serious multicollinearity, delete a variable (Chapter 6).

Step 10. Assess whether your model accounts for outcome better than would be expected by chance (e.g., F test, likelihood ratio test) (Section 9.2.A).

Step 11. Perform an assessment of the fit of your model (e.g., adjusted R^2, Hosmer–Lemeshow test) (Section 9.2.B) or its ability to predict the outcome for study subjects (e.g., sensitivity, specificity, c index) (Section 9.2.C).

Step 12. Assess the strength of your individual covariates in estimating outcome (Section 9.3).

Step 13. Evaluate whether your data fit the multivariable assumptions of your model (Chapter 10). Use regression diagnostics to test how well your model fits your data on a multivariable level (Sections 10.1–10.6). For proportional hazards models, be sure that the proportionality assumption is met (Sections 10.10–10.12).

Step 14. Decide whether to include interaction terms in your model (Sections 1.4, 8.3, 9.9, and 10.8).

Step 15. Consider whether it would be possible to validate your model (Chapter 13).

Step 16. Publish your results in the *New England Journal of Medicine* and be the envy of your friends and colleagues.

If you have questions or suggestions for future editions send them to me at mhkatz59@yahoo.com.

Index