

STATISTICAL METHODS

FOR THE SOCIAL &
BEHAVIOURAL SCIENCES

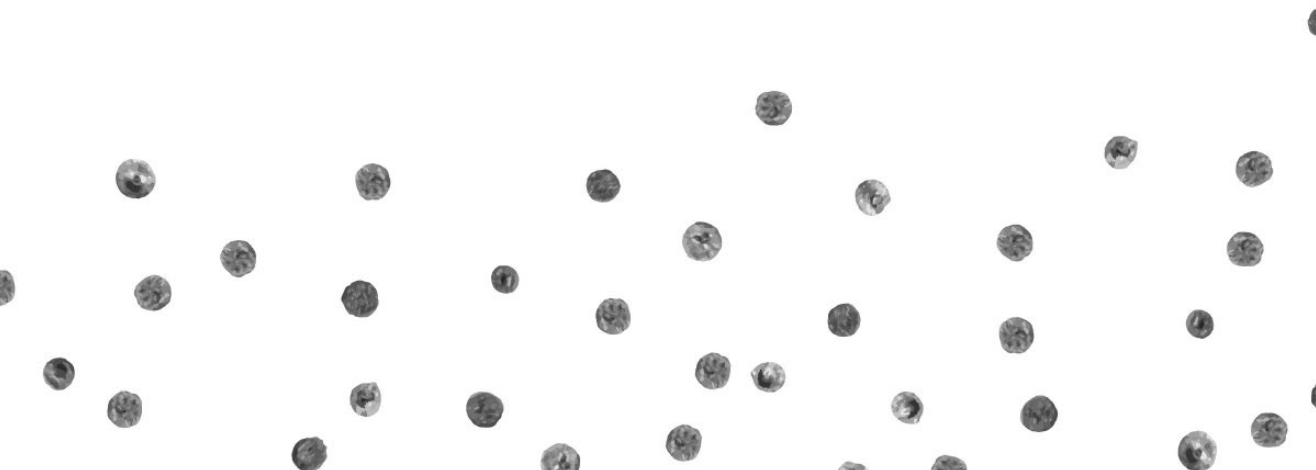
A MODEL-BASED APPROACH

DAVID B. FLORA



STATISTICAL METHODS

FOR THE SOCIAL &
BEHAVIOURAL SCIENCES



Sara Miller McCune founded SAGE Publishing in 1965 to support the dissemination of usable knowledge and educate a global community. SAGE publishes more than 1000 journals and over 800 new books each year, spanning a wide range of subject areas. Our growing selection of library products includes archives, data, case studies and video. SAGE remains majority owned by our founder and after her lifetime will become owned by a charitable trust that secures the company's continued independence.

Los Angeles | London | New Delhi | Singapore | Washington DC | Melbourne

STATISTICAL METHODS

FOR THE SOCIAL &
BEHAVIOURAL SCIENCES

A MODEL-BASED APPROACH

DAVID B. FLORA



Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne



Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

SAGE Publications Ltd
1 Oliver's Yard
55 City Road
London EC1Y 1SP

SAGE Publications Inc.
2455 Teller Road
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road
New Delhi 110 044

SAGE Publications Asia-Pacific Pte Ltd
3 Church Street
#10-04 Samsung Hub
Singapore 049483

Editor: Michael Ainsley
Editorial assistant: John Nightingale
Production editor: Tom Bedford
Copyeditor: Sheree Van Vreede
Proofreader: Thea Watson
Indexer: Cathy Heath
Marketing manager: Susheel Gokarakonda
Cover design: Wendy Scott
Typeset by: C&M Digital (P) Ltd, Chennai, India
Printed in the UK

© David B. Flora 2018

First published 2018

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act, 1988, this publication may be reproduced, stored or transmitted in any form, or by any means, only with the prior permission in writing of the publishers, or in the case of reprographic reproduction, in accordance with the terms of licences issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers.

Library of Congress Control Number: 2017955066

British Library Cataloguing in Publication data

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

ISBN 978-1-4462-6982-4

ISBN 978-1-4462-6983-1 (pbk)

At SAGE we take sustainability seriously. Most of our products are printed in the UK using FSC papers and boards. When we print overseas we ensure sustainable papers are used as measured by the PREPS grading system. We undertake an annual audit to monitor our sustainability.

für Julia

CONTENTS

<i>Online Resources</i>	x
<i>About the Author</i>	xi
<i>Acknowledgements</i>	xii
<i>Preface</i>	xiii

1 Foundations of Statistical Modeling Demonstrated with Simple Regression	1
Chapter overview	1
What is a statistical model?	2
Significance testing and effect sizes	6
Simple regression models	9
Basic regression diagnostic concepts	27
Chapter summary	39
Recommended reading	40
2 Multiple Regression with Continuous Predictors	42
Chapter overview	42
What is multiple regression?	43
Multiple regression with two predictors	44
Multiple regression with P = two or more predictors	62
Regression diagnostics revisited	72
Chapter summary	85
Recommended reading	86
3 Regression with Categorical Predictors	87
Chapter overview	87
What is the general linear model?	88
Using dummy-code variables	92
Using contrast-code variables	97
The analysis of covariance (ANCOVA) model and beyond	104
Regression diagnostics with categorical predictors	108
Chapter summary	111
Recommended reading	112

4 Interactions in Multiple Regression: Models for Moderation	113
Chapter overview	113
What is statistical moderation?	114
Interactions with a categorical moderator	114
Interactions with a continuous moderator	131
Chapter summary	142
Recommended reading	143
5 Using Multiple Regression to Model Mediation and Other Indirect Effects	144
Chapter overview	144
What are mediation and indirect effects?	145
Specification of the simple indirect-effect model	147
Estimation and inference for the indirect effect	151
Chapter summary	161
Recommended reading	161
6 Introduction to Multilevel Modeling	163
Chapter overview	163
What is multilevel modeling?	164
Nonindependent observations	164
The unconditional multilevel model	171
Conditional multilevel models	179
Distinguishing within-cluster effect from between-clusters effect	190
Formal model comparisons	199
Cross-level interactions	202
Three-level models	205
Assumption checking for MLM	205
Chapter summary	208
Recommended reading	212
7 Basic Matrix Algebra for Statistical Modeling	213
Chapter overview	213
Why matrix algebra?	213
Kinds of matrices and simple matrix operations	214
Elementary matrix algebra	218
Matrix calculations for statistical applications	227
Chapter summary	236
Recommended reading	237
8 Exploratory Factor Analysis	238
Chapter overview	238
What is exploratory factor analysis?	239
EFA with continuous observed variables	240


Research example for EFA with continuous observed variables	241
Estimation of the common factor model	247
Determining the optimal number of common factors	257
Using model fit statistics to determine the optimal number of factors	262
Factor rotation	265
Exploratory factor analysis with categorical observed variables	275
Assumptions and diagnostics for EFA	280
Chapter summary	283
Recommended reading	284
9 Structural Equation Modeling I: Path Analysis	285
Chapter overview	285
What is structural equation modeling?	286
Path analysis	288
Path analysis: Model specification	289
Path analysis: Model identification	302
Path analysis: Model estimation	306
Model fit evaluation	316
Assumptions and diagnostics for path analysis models	331
Chapter summary	333
Recommended reading	333
10 Structural Equation Modeling II: Latent Variable Models	335
Chapter overview	335
What are latent variable models?	336
Confirmatory factor analysis	337
Structural regression models	367
Chapter summary	386
Recommended reading	387
11 Growth Curve Modeling	389
Chapter overview	389
What is growth curve modeling?	390
Growth curve models for linear change	397
Growth curve models for nonlinear change	419
Chapter summary	434
Recommended reading	435
<i>References</i>	437
<i>Index</i>	447

ONLINE RESOURCES



Visit <https://study.sagepub.com/flora> to find a range of additional resources for both students and lecturers, to aid study and support teaching.

The online resources are:

- Files containing annotated **input and output** for the most popular statistical analysis software packages (R, SAS®, IBM SPSS Statistics® and MPlus®), allowing you to implement statistical procedures no matter your software preference.
- A series of **datasets**, enabling you to apply statistical techniques to real data. When you see this icon  in the margin, that means there's a dataset online that corresponds to the worked example in the text.

ABOUT THE AUTHOR

David B. Flora, PhD, is an associate professor in the Department of Psychology at York University in Toronto, Canada. Dr Flora has also served as a coordinator of the Quantitative Methods Area in York's graduate program in psychology and as a coordinator of York's Statistical Consulting Service. As a quantitative psychologist, Dr Flora is a co-author on numerous articles focused on quantitative methodology itself as well as on a wide range of articles in which advanced quantitative methods are applied to substantive research topics in psychology. Dr Flora earned his PhD from the Quantitative Psychology program at the University of North Carolina at Chapel Hill. Although he now lives in Toronto, Dr Flora is a Tar Heel born and a Tar Heel bred, and when he dies he'll be a Tar Heel dead.

ACKNOWLEDGEMENTS

If this book is at all successful at presenting and explaining its subject matter, it is because I have had the privilege of learning from many outstanding teachers and mentors. In fact, some of the writing in this text has evolved from their lecture notes. In alphabetical order, these teachers and mentors include Ken Bollen, Laurie Chassin, Patrick Curran (my graduate school advisor, whose enthusiasm about this book I have appreciated), Siek-Toon Khoo, Bud MacCallum, Abigail Panter, Dave Thissen, and Jack Vevea. My undergraduate statistics instructor Brad Hartlaub was the first to emphasize to me the importance of *writing* about data analyses. A few of my colleagues at York University – Rob Cribbie, John Fox, Michael Friendly, Georges Monette, and Jolynn Pek – have enhanced my understanding of this book's subject matter. I also thank Phil Chalmers for helping me create a few of the book's figures and other general computing assistance. Finally, I thank Michael Carmichael, Mila Steele, Michael Ainsley, and John Nightingale at SAGE Publications for steering me through this whole process.

Of course, none of the individuals named here is responsible for any inaccuracies which may appear in the following pages.

PREFACE

This book is intended for graduate students as well as for more seasoned faculty and researchers alike in the behavioural and social sciences; the statistical methods and concepts covered herein are also widely used in education, the health sciences and business and organizational research. As an advanced text, this book was written for PhD-level students and researchers who are already comfortable with the topics typically covered in a first-year sequence of graduate-level applied statistics (see subsequent discussion) who wish to learn about more advanced procedures that are commonly used. As such, this text is suitable for a general second (or advanced) course in statistical methods for the behavioural and social sciences.

The book might be best understood as a survey of the advanced statistical methods most commonly used in modern research in psychology and related fields in the behavioural and social sciences. That is, each chapter gives an overview of a major topic which could be expanded into a full book on its own, and indeed, good book-length treatments are available for each topic (many of these resources are cited in the current text). I hope that the chapters in this text can provide a sufficient foundation for readers to begin using a given statistical modeling method for their data. But I also strongly encourage readers to delve deeper into any statistical topic that is especially pertinent for their research interests; the books and articles cited in the current text (particularly the Recommended Reading section concluding each chapter) should provide direction.

Because this text covers advanced statistical methods and is aimed at readers who have completed coursework in basic statistics and data analysis, it is necessary to assume comfort (more than just familiarity) with a wide range of fundamental topics and principles. If necessary, readers should review basic statistics before tackling this book. The following are the most critical topics readers are assumed to understand:

- Basic high school-level algebra; no experience with calculus is assumed (although concepts from calculus are mentioned in a few places, familiarity with them is not necessary for more general comprehension of the relevant material)
- Major concepts in research methods and design, including principles of sampling
- Describing and representing univariate distributions using frequency tables, descriptive statistics, and graphs
- Basic definitions and rules of probability (this text adheres to the frequentist conception of probability)
- Concepts of sampling distributions and standard error; the central limit theorem
- Logic of null hypothesis significance testing (and limitations of null hypothesis testing), including Z and t tests, χ^2 tests, and definitions of Type I and Type II errors and statistical power

- Basic one-way analysis of variance (ANOVA)
- Definition and interpretation of confidence intervals (CIs); correspondence between confidence intervals and hypothesis tests
- Simple correlation and regression (although these topics are also detailed in Chapter 1)

Readers should also be familiar with American Psychological Association (APA) style (APA, 2010) for presenting statistical results (e.g., $t(272) = 2.91$, $p = .003$); this style guide is used widely in fields other than psychology.

The following topics are extremely important, and they are touched on in this text but not comprehensively detailed:

- The importance of quality measurement, collecting reliable and valid scores for variables of interest
- Multiple comparison control (i.e., accumulation of Type I error probability across a family of null hypothesis tests)
- Concerns regarding ‘data dredging’ or ‘ p hacking’: These problems may be less severe when researchers (1) adhere to rigorous research design principles, (2) take reliable and valid measurements, and (3) choose statistical models that match theories as closely as possible and evaluate models with a wide lens, not abusing null hypothesis significance testing
- Principles for drawing causal inferences from statistical results

PURPOSES AND PERSPECTIVE OF THIS BOOK

Probably the most common reason for setting out to write a textbook is that an instructor of a given course finds all currently available books unsatisfactory. That was my feeling as I planned and taught a course titled ‘Multivariate Analysis’, which is aimed at PhD-level graduate students in psychology. Although there is a wide range of textbooks already available on the general topic of multivariate data analysis, each one of them has certain limitations. Indeed, I have noticed that syllabi for similar courses taught at other universities often list several potential textbooks, implying that no one of them is entirely suitable.

The primary reason that I struggled to find an appropriate textbook for my course was that the particular topics typically emphasized in so-called ‘multivariate’ texts usually did not overlap well with the topics I felt should be emphasized to prepare graduate students for careers in which they are likely to need to understand the data analytic techniques that are most common in modern research in psychology and the social sciences. In particular, even recently published multivariate texts devote entire chapters to traditional multivariate statistical methods such as multivariate analysis of variance (MANOVA), discriminant function analysis, and canonical correlation analysis. Yet, these methods are hardly used in psychological research anymore, and in my roles as a research collaborator and statistical consultant, I almost never encounter studies for which these methods seem ideally suited. Even during my own graduate training in quantitative psychology, we were told that ‘although we are learning about these methods, you will probably never use them!’

Instead, the statistical methods most commonly used in modern research in psychology and the social sciences generally entail developing, estimating, and testing *models* for data. In his landmark paper, Rodgers (2010) explained that a ‘quiet methodological revolution’ has

occurred in which psychology has moved beyond rigid data analytic methods which emphasize statistical tests above all else in favour of more flexible methods for *modeling* data based on substantive theories and expertise regarding the underlying processes which give rise to empirical data. Perhaps most prominent among these modeling procedures is multiple regression, which of course is addressed in most, if not all, multivariate texts. But popular modern applications of multiple regression for modeling hypotheses regarding *moderation* and *mediation* seem to receive little, if any, attention. Another important modeling method also commonly covered in most traditional multivariate texts is *exploratory factor analysis*, but it is usually preceded by a distracting treatment of principal components analysis, or even worse, principal components analysis is falsely presented as if it is a *type* of factor analysis (see Chapter 8). Finally, few multivariate textbooks include chapters on both *multilevel modeling* (also known as hierarchical linear modeling) and *structural equation modeling*, although both have become extremely common in modern research.

Finally, a handful of texts present statistical methods in concert with their implementation using a single particular software package (most commonly SPSS). In my experience, researchers (in graduate school and beyond) often become so wedded to one specific statistical software package that they are later handcuffed by the limitations of that software when they encounter problems. For example, many newly developed statistical techniques become widely available before they are added to SPSS or SAS.

My hope is that if students have a solid understanding of the basic principles underlying a given statistical procedure, then that foundation will allow them to carry out the procedure using whichever software package implements the procedure most appropriately. Nonetheless, I recognize that students come to understand statistical procedures more completely when they can apply them using example data analyses, and for that reason, the website for this text includes annotated input and output files from several prominent software packages (i.e., R, SAS, SPSS, and Mplus) for each of the major statistical modeling procedures presented herein. For the most part, however, statistical software concerns are not addressed in the main body of this text because I do not want computing to become a distraction from the conceptual statistical principles described in the book.

In closing, it is important to disclaim that I am a psychologist, and for that reason, the statistical methods and example data analyses presented in this text primarily draw on psychological research. But the methods and techniques I present (and any accompanying advice for data analysis) are broadly applicable across a wide range of disciplines.

A NOTE ON EQUATIONS

As a text on advanced statistical modeling methods, it is necessary to present statistical models using equations. I strongly believe that having an understanding of the key equations underlying a statistical model is critical to understanding how the model represents empirical data or addresses a substantive research question. Yet, I have tried to keep the text nontechnical. Most equations in the text just involve simple addition and multiplication; occasionally exponentials or logarithms are used. Even the matrix algebra equations presented in Chapter 7 and subsequent chapters can be characterized as organized collections of simple algebraic operations based on addition and multiplication. Nonetheless, throughout the text, I have tried to explain the conceptual meaning of each potentially unfamiliar element within a given

equation rather than assuming that these mathematical expressions are self-explanatory. **Equations are numbered if they are referred to in the text; un-numbered equations tend to be explained immediately without any need to refer to them again.**

Most equations presented in this text are variations of regression equations. As such, just as the conceptual content of later chapters builds on the content introduced in earlier chapters, the equations presented in later chapters usually build on equations first seen in earlier chapters. In many instances throughout the text, a newly introduced equation is compared with an equation presented earlier, either in the same chapter or in a previous chapter. Therefore, as one proceeds through the text, it is wise to keep track of the equations (particularly the numbered equations) as they appear so that one can compare a new equation with previous equations to identify which features are familiar from previous equations and which are new elaborations. Doing so should consequently help one understand the meaning of the particular statistical model that the equation pertains to.

Also, in digesting these equations, it is important not to get overwhelmed by notation. Whenever a symbol is introduced for the first time, its meaning is explained. I have attempted to keep the use of notation consistent throughout the text, but I have also tried to keep my use of notation consistent with the methodological literature on the topics covered in this text so that readers can more easily consult other resources, which does introduce some inconsistency from one chapter to another (especially when moving from multilevel modeling to structural equation modeling). To help prevent confusion about the Greek letters used for notation, at the beginning of each chapter there is a table giving the Greek letters used in that chapter, its English name, and a brief statement of what the Greek letter represents within that chapter.

1

FOUNDATIONS OF STATISTICAL MODELING DEMONSTRATED WITH SIMPLE REGRESSION

CHAPTER OVERVIEW

The major objective of this chapter is to develop an understanding of the principles of statistical modeling in general and the simple linear regression model in particular. These principles provide a conceptual foundation for the remainder of the text. The main topics of this chapter include:

- Definition and description of statistical modeling as a guiding theme for the text
- Perspective on effect-size meaning and significance testing used in this book
- Orientation toward the simple linear regression model
- The intercept-only model as a model against which to compare the simple linear regression model
- Foundational principles for simple linear regression
- Specification and estimation of the simple linear regression model
- Statistical inference with the simple linear regression model
- Dichotomous variables in simple linear regression
- Basic concepts for regression diagnostics as they pertain to simple linear regression
- Outliers and unusual cases from the perspective of simple linear regression

Table 1.0 Greek letter notation used in this chapter

Greek letter	English name	Represents
β	Lowercase 'beta'	Regression model parameter (intercept or slope, depending on subscript)
ε	Lowercase 'epsilon'	Regression model error term
μ	Lowercase 'mu'	Population mean
σ	Lowercase 'sigma'	Population standard deviation
ρ	Lowercase 'rho'	Population correlation
α	Lowercase 'alpha'	Probability of Type I error

WHAT IS A STATISTICAL MODEL?

A trivial example

Before formally defining *statistical model*, I will begin with a trivial example model that demonstrates some of the fundamental ideas about models. Growing up in the United States, I became accustomed to thinking about temperature on the Fahrenheit scale. I know how chilly 40°F is, and I know how warm 75°F is. In Canada, where I now live, temperature is usually reported on the Celsius scale. Unfortunately, I do not automatically have a good sense of what a temperature such as 13°C feels like (should I wear a jacket if I go outside?), so I find that I am constantly converting temperatures reported in Celsius into the approximate Fahrenheit temperature in my head. Of course, there is a known, precise relation between °F and °C, but the conversion isn't always easy for me to calculate in my head, so I use an approximation that I can calculate quickly. Specifically, I multiply the temperature in °C by two and add 30 to arrive at a value that I know is at least near the temperature in °F.

This approximation is my *model* for °F given the reported °C, and it can be expressed using the following mathematical equation:

$$^{\circ}\hat{F} = 2(^{\circ}\text{C}) + 30. \quad (1.1)$$

The hat symbol (^) over F on the left-hand side of the equation indicates that the formula produces a *predicted* value for °F given a particular value for °C. That is, the value for °C is known, or observed, whereas the value for °F is unobserved. (The predicted value is also known as the *model-implied* or *fitted* value.) So if I am told that it is 13°C outside and I am wondering whether I should wear a jacket, then I can quickly calculate

$$^{\circ}\hat{F} = 2(13) + 30 = 56.$$

Thus, my predicted value for the temperature on the Fahrenheit scale is $^{\circ}\hat{F} = 56$, which is not terribly cold but chilly enough that I will probably put on a jacket.

Now, I know that my model does not usually produce the actual, precise value for °F given some temperature in °C. That is, deriving the true °F using this approximation is error-prone, and so another way I can write the model is

$$^{\circ}\text{F} = 2(^{\circ}\text{C}) + 30 + \varepsilon, \quad (1.2)$$

where ε is the error term representing the inaccuracy involved in reproducing the true $^{\circ}\text{F}$ using this formula. Next, with some simple algebra, we see that we can substitute Equation 1.1 into Equation 1.2 such that

$$^{\circ}\text{F} = ^{\circ}\hat{\text{F}} + \varepsilon$$

or

$$\varepsilon = ^{\circ}\text{F} - ^{\circ}\hat{\text{F}}. \quad (1.3)$$

Thus, the error, ε , gives the difference between the true temperature on the Fahrenheit scale ($^{\circ}\text{F}$) and the temperature on the Fahrenheit scale predicted by the model ($^{\circ}\hat{\text{F}}$). Equations 1.1 and 1.2 are different ways of expressing the same model for the relation between $^{\circ}\text{C}$ and $^{\circ}\text{F}$.

All statistical models are like my temperature model in Equation 1.1 in that they generate predicted values for some outcomes but do so with error. Of course there is an established, true relation between the Fahrenheit and Celsius scales, specifically

$$^{\circ}\text{F} = 1.8(^{\circ}\text{C}) + 32. \quad (1.4)$$

Note that Equation 1.4 is not really a *model* because there is no error term; given a value for $^{\circ}\text{C}$, we can use Equation 1.4 to calculate the *exact*, true value for $^{\circ}\text{F}$.

We can also use Equation 1.4 to evaluate the quality of the model expressed in Equations 1.1 and 1.2. That is, we can use Equation 1.4 to calculate values for the model's error term, ε , across different values of $^{\circ}\text{C}$; in other words, we can use Equation 1.4 to find out how well our predicted values, $^{\circ}\hat{\text{F}}$, reproduce the true values, $^{\circ}\text{F}$. For example, if it is 0°C outside (i.e., the temperature at which water freezes), the true $^{\circ}\text{F}$ is $1.8(0) + 32 = 32^{\circ}\text{F}$, but the model's predicted value is $2(0) + 30 = 30^{\circ}\hat{\text{F}}$. Thus, the model is inaccurate by 2°F , or using Equation 1.3, we have $\varepsilon = 32 - 30 = 2$. So although it's not precise, the model does a reasonably good job of predicting $^{\circ}\text{F}$ when $^{\circ}\text{C}$ is 0, or freezing. But how well does the model do when, for example, it's 13°C ? Will the model lead to me being too warm in a light jacket, or will I wish that I had put on something heavier? Again, using Equation 1.4, the true $^{\circ}\text{F}$ corresponding to 13°C is 55.4°F , and now $\varepsilon = -0.6$, which is reasonably accurate given the model's purpose; that is, I am unlikely to regret my decision to wear a jacket.

To get a more complete picture of how good the model is across a wider range of values for $^{\circ}\text{C}$, we can plot Equations 1.1 and 1.4 in the same graph, as shown in Figure 1.1. I have chosen a range of -15°C to 45°C for the x -axis to represent the wide range of outside temperatures experienced in a given year in North America (having lived in Phoenix, Arizona, and Toronto, I am familiar with both extremes). Clearly, Equations 1.1 and 1.4 are both equations for straight lines but with different intercept and slope values. But in the figure, we see that the lines cross above 10°C , where both the predicted value $^{\circ}\hat{\text{F}}$ and the true value $^{\circ}\text{F}$ equal 50. Thus, for 10°C , the model perfectly reproduces the true $^{\circ}\text{F}$ (i.e., $\varepsilon = 0$). To the left of 10°C , the line for the predicted values is below the line for the true values, indicating that when the temperature is below 10°C , the model underestimates the true $^{\circ}\text{F}$ and the corresponding values for the error term ε are all positive. To the right of 10°C , the predicted line is above the true line, indicating that when the temperature is above 10°C , the model overestimates $^{\circ}\text{F}$ and the values for ε are negative. Nonetheless, across the range of $^{\circ}\text{C}$ plotted, the predicted

line never deviates far from the observed line, indicating that the model is good enough for the purpose of predicting degrees Fahrenheit across the range of temperatures most commonly experienced in North America. But if we were to extend the model in either direction, to extremely cold temperatures or extremely hot temperatures, the model's predictions would clearly deteriorate.

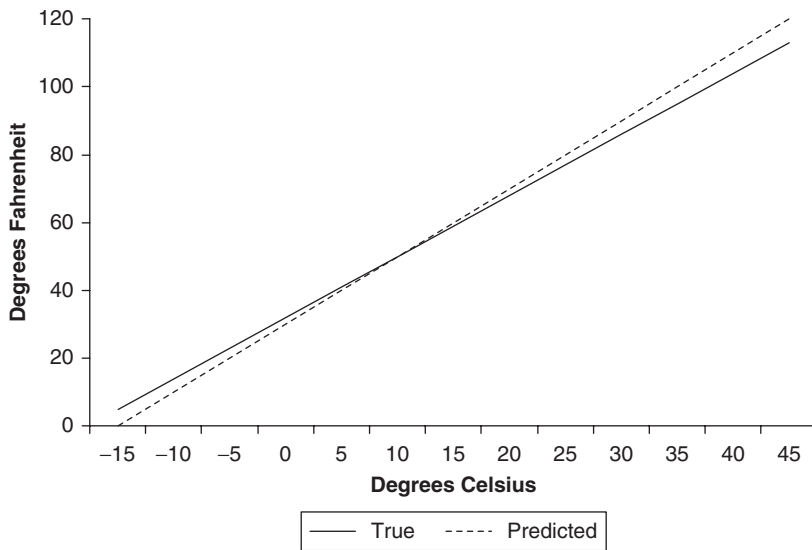


Figure 1.1 True and model-predicted relation between Celsius and Fahrenheit

In behavioural and social research, we can never know the *true* relation between any two variables. Equation 1.4 gives us the known, exact relation between Celsius and Fahrenheit, but there is no analogous version of Equation 1.4 for the types of variables studied in the social and behavioural sciences. Thus, there is no precise way to evaluate the quality of a statistical model in the manner illustrated in Figure 1.1. Instead, other methods for evaluating models must be used, and some of them are discussed throughout this book. Nevertheless, this temperature example demonstrates some key ideas about models:

1. They are developed to give a useful simplification of some natural phenomenon (e.g., an easy way to calculate the relation between Celsius and Fahrenheit).
2. They give predictions, but by virtue of the model being a simplification of nature, these predictions have error (e.g., the difference between model-implied, or predicted, Fahrenheit and true Fahrenheit).
3. They are tailored to serve a particular use, for which the errors are hopefully small, that might not generalize to other uses (e.g., understanding the relation between Celsius and Fahrenheit across the range of temperatures commonly experienced in North America but not the extreme heat near the surface of the sun).

As statistician George Box famously stated, 'All models are wrong but some are useful' (Box, 1979: 208).

Statistical model defined

With these ideas in mind, drawing from Pearl (2000: 202) and Rodgers (2010: 5), a **statistical model** is a set of one or more mathematical expressions (e.g., equations) that provides an idealized representation of reality; this representation represents reality in important ways but is necessarily a simplification that ignores certain features of reality. More formally, a statistical model specifies a univariate or multivariate population probability density function which is hypothesized to generate sample data (Myung, 2003).

The use of such statistical models to analyse quantitative data from behavioural and social empirical research studies is the unifying theme of this text. In particular, the text is focused on modeling procedures that are especially popular for analysing data in behavioural and social research, namely, multiple regression, factor analysis, multilevel modeling, and structural equation modeling (SEM). These modeling procedures are popular in modern research mostly because of their broad capacity to help answer a wide range of sophisticated research questions and, in so doing, to help with the development and evaluation of important substantive theories.

Rodgers (2010) also distinguished between two roles for models. The first is focused on evaluating models (and comparing competing models) for a given dataset using existing statistical modeling methods such as regression and SEM. This first role is the main topic of this text, although many of the principles addressed here also apply to the second role. The second role ‘involves the development of mathematical models to match topics of explicit interest to researchers. Within this second framework, substantive scientists study behaviour and from that process develop mathematical models specific to their research domain’ (Rodgers, 2010: 8). See Rodgers for examples of research based on this second approach. SEM, for example, is a prominent statistical method commonly used for the first role, and its flexibility makes SEM adaptable to many different applications, whereas models developed for the second role are often specific to a particular application.

First and foremost, statistical models are fundamentally **descriptive** in that they provide descriptions of the associations among one or more operational variables (i.e., the actual observed measurements in a research study rather than the more abstract concepts) in terms of a small set of patterns which are summarized with mathematical formulas. But moving beyond basic description, the two other major purposes of statistical models are **explanation** and **prediction**. Building a model for the purpose of explanation means that the model is meant to represent a theoretical account for the variation in some important dependent variable or outcome; typically this account (either explicitly or implicitly) represents the actual causal mechanisms, or a subset of potential causal mechanisms, that produce changes in the outcome. The adage that ‘correlation does not imply causation’ certainly extends to statistical models for observational data, but nonetheless, such models may still represent theoretical causal mechanisms. The models presented in this book are mainly presented with the goal of theoretical explanation in mind. But determining whether the associations among variables in a statistical model truly represent causal effects ultimately is the shared responsibility of researchers producing those results and the consumers of that research, all of whom must carefully consider the quality of the research design (e.g., were observations properly sampled and measured? What are potential confounding effects?) as well as the statistical analysis itself.

Models built for theoretical explanation usually are used only to describe outcomes that are observed within a given dataset (e.g., what is the association between personality and

depression among university students?), whereas models built for the purpose of prediction are meant to provide accurate forecasts for critical outcomes that have not yet been observed (e.g., given a high school student's score on an academic achievement test, what is her likely university grade-point average?). Principles for developing statistical models for the different purposes of explanation and prediction tend to be complementary but not always. For further discussion of these issues, see Pedhazur (1997, especially pp. 195–8, and references therein).

Throughout this text, I frequently use the terms *predictor* and *predicted values* but doing so does not imply that I am referring to pure prediction in this sense of estimating unobserved values or future outcomes. Instead, I use these terms in a more descriptive mathematical sense. In any statistical model fitted to a dataset, the score on an independent or explanatory variable for a given research participant, or case, in that dataset can be used to obtain, or *predict*, values on the dependent or outcome variable for that same case; hence, this explanatory variable may be referred to as a *predictor*. If the predicted value on the outcome variable for a given case is close to the actual observed value for that case in the dataset, then the statistical model has performed well (for that case).



Section recap

Statistical models

A **statistical model** is a set of one or more mathematical expressions that provides an idealized representation of reality; this representation represents reality in important ways but is necessarily a simplification that ignores certain features of reality.

Fundamentally, models describe the variability of an important outcome or dependent variable as a function of one or more predictors or independent variables; these descriptions may reflect (causal) explanation or may simply be used for prediction.

SIGNIFICANCE TESTING AND EFFECT SIZES

Null hypothesis significance testing (NHST) has been the dominant paradigm in data analysis for behavioural and social research since the middle of the 20th century, and criticism of NHST is just as old (e.g., Jones, 1952; Rozeboom, 1960). In psychology, debate over the usefulness (or lack thereof) of NHST bubbled over as a result of a now-famous article by Cohen (1994), leading the American Psychological Association (APA) to create a Task Force on Statistical Inference (TFSI) consisting of a team of eminent quantitative methodologists. The Task Force was charged with evaluating the possibility of banning NHST from psychology journals (or at least those published by APA). They ultimately concluded that although NHST has its flaws, it should remain available as a tool for data analysts but should also be supplemented with (if not subsumed by) other statistical information (see Wilkinson and the Task Force on Statistical Inference, 1999).

What has happened since then? In the APA's flagship journal, Rodgers (2010) argued that a 'quiet methodological revolution, a modeling revolution' has occurred which has made the NHST controversy mostly irrelevant. In particular:

A basic thesis of this article is that the heated (and interesting) NHST controversy during the 1990s was at least partially unnecessary. In certain important ways, a different methodological revolution precluded the need to discuss whether NHST should be abandoned or continued. This quiet revolution, a modeling revolution, is now virtually complete within methodology. But within the perspective of the diffusion of innovations, this revolutionary thinking is only beginning to spread to the applied research arena and graduate training in quantitative methods. Identifying the revolution is one mechanism that will promote its diffusion. The methodological revolution to which I refer has involved the transition from the NHST paradigms developed by Fisher and Neyman-Pearson to a paradigm based on building, comparing, and evaluating statistical/mathematical models. (Rodgers, 2010: 3-4).

This book adheres to the premise by Rodgers that a modeling revolution has occurred, and indeed, it's focused on 'building, comparing, and evaluating' models as the dominant paradigm in data analysis for modern behavioural and social research. But, as Rodgers implied, modeling is hardly a new enterprise in quantitative methodology (in fact, several prominent applied statistics texts already take this perspective, e.g., Maxwell and Delaney, 1990, 2004). Modeling is already a major data-analytic approach for most quantitative research in the behavioural and social sciences; the 'quiet revolution' is 'almost complete.' The task now is to give this epistemological system a more explicit focus in how researchers are trained, which this book aims to help accomplish.

It is important to understand that this focus on modeling does not imply that NHST is no longer used; instead, NHST still plays 'an important though not expansive role' (Rodgers, 2010: 1) in the context of comparing models and evaluating estimates of their parameters. Thus, my perspective for this book is that significance testing through the calculation and reporting of p values is one tool that can be useful for evaluating and comparing models (or parts of models), but other tools [e.g., confidence intervals (CIs)] can be helpful as well. I readily acknowledge that students and researchers often misunderstand the exact meaning of a significance test and related concepts (definitions of p value, Type I and II error, etc.) and that NHST has limitations, and we will keep these issues in mind as we use NHST to examine models. I won't review these definitions and limitations here because they have been thoroughly addressed elsewhere, and frankly I wish to move beyond them (but readers not familiar with these issues should at least consult Cohen, 1994, and Wilkinson and TFSL, 1999).

It is satisfying to recognize that the father of NHST, Sir Ronald Fisher, advocated a model-based approach for statistical inference in the context of nonrandom sampling. Statistics textbooks commonly present statistical inference (i.e., generalizing from sample statistics to population parameters) using NHST as being valid only when the data come from a simple random sample, but behavioural and social research is often conducted using nonrandom samples. Thus, a critical aspect of Fisher's model-based approach is the acknowledgement that there is no basis for statistical inference when observations are nonrandomly sampled from a finite population, but *inference is legitimate under nonrandom sampling from an infinite population* (Fisher, 1922). With this infinite-population inference approach, the researcher first specifies a statistical model that represents the process that generated the outcome variable(s) according to certain population parameters, which are the target of inference. Next, a parametric distributional assumption is imposed on the model to represent the link between the fixed, observed values of the outcome variable and the realizations of a random variable. Finally, it is critical to incorporate model parameters

to account for any meaningful departure from simple random sampling and the sampling design that was used (i.e., sampling based on stratification or clustering or disproportionate sampling). For a detailed discussion of Fisher's model-based inferential framework, please read Sterba (2009). This framework (along with some adaptations discussed by Sterba) represents the perspective used in this text for model-based statistical inference assuming an infinite population. This is also the perspective widely adapted across almost all of behavioural research, even if it's not explicitly acknowledged (although certain areas of social research are more concerned with finite-population inference).

One result of the NHST controversy in psychology is that it has led to a greater emphasis on the importance of **effect-size** calculation and reporting. Put simply, effect size is the extent to which a predictor is associated with an outcome variable. In other words, effect size is the strength of the relation between two variables. As such, effect size is really a simple concept (but see Kelley and Preacher, 2012, for a thorough discussion of defining *effect size*), but it is my impression that the cries for more effect-size reporting have made the issue overly complicated and have needlessly confused students and researchers. These pleas seem to have created an impression among psychologists in particular that effect-size reporting must always consist of a sophisticated-sounding standardized effect-size statistic (e.g., Cohen's *d* or omega squared) when simpler, familiar descriptive statistics (e.g., unstandardized mean differences), graphs, and estimates of *model parameters* are often (if not always) more effective at conveying effect size in meaningful units (Wilkinson and TFSL, 1999; for further discussions of this issue, see Baguley, 2009, and Frick, 1999). To the extent that quality research reports (e.g., journal articles) in the behavioural and social sciences have *always* included descriptive statistics, graphs, and estimates of model parameters, they have therefore also always included effect-size information, even prior to the recent pleas for effect-size reporting (see Pek and Flora, in press, for further discussion).

The models presented in this text share the property that the associations among variables are represented with parameters (e.g., a regression slope coefficient) and, thus, that the size of a given parameter *is* a measure of effect size. This point was emphasized and demonstrated by Steinberg and Thissen (2006), who argued that when results of statistical models are reported, effect sizes 'are *most clearly* expressed in tabular or graphical presentation of parameter estimates' (p. 413; emphasis mine). The parameters need not be standardized; in fact, Wilkinson and TFSL (1999) exclaimed that, 'If the units of measurement are meaningful on a practical level (e.g., number of cigarettes smoked per day), then we usually prefer an unstandardized measure (regression coefficient or mean difference) to a standardized measure (*r* or *d*)' (p. 599). Thus, although there is a potentially overwhelming plethora of standardized effect-size statistics available to researchers, this text is focused on interpretation of the estimates of a model's parameters as the primary mode for effect-size conveyance.



Section recap

Effect size and significance testing

Effect size is simply the extent to which one or more predictor (or explanatory) variables is associated with an outcome (or response) variable. In other words, effect size is the strength of the relation between one or more independent variables and a dependent variable.

In the context of statistical modeling, parameter estimates are effect-size statistics and these effects are usually most clearly conveyed in unstandardized form.

Despite its logical flaws, null hypothesis significance testing (NHST) remains a useful tool for testing the effects in a model and for comparing models.

SIMPLE REGRESSION MODELS

All statistical methods covered in this text involve using parametric models. A parametric model is a mathematical expression that uses parameters to represent hypothetical relations among variables in the population. In other words, the model represents a hypothetical process that generates the outcome variable(s) according to certain population parameters. The first model we will examine extensively is the **simple linear regression model**. I expect that any reader of this text is already familiar with simple regression as it is almost always covered in introductory statistics textbooks and courses. Here, the purpose of studying the simple regression model in some detail is to establish a solid foundation for presenting more complicated and advanced modeling procedures because the same statistical principles and methods continue to apply as the two-variable simple regression model is expanded into larger models for more than two variables. Indeed, a few ideas addressed here even for simple regression may be new for some readers. This chapter also provides a familiar context, simple regression, in which to introduce the reader to the terminology, notation, and style that will be used throughout this text.

To begin nailing down the terminology and notation used in this text, let's take a quick, initial look at the simple linear regression model. It is a model for the score of case, or individual i on some outcome variable Y , given that individuals score on some predictor variable X . 'Individual i ' is a potentially observed unit in the (infinite) population of interest; most often, these units of observation are individual people (i.e., research participants), but of course in certain research areas, the units of observation might instead be animals, cities, parent-child dyads, or business firms, among other possibilities. One way to write the simple regression model is with the linear equation

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i. \quad (1.5)$$

The parameters in the model are the **intercept**, denoted β_0 , and the **slope**, denoted β_1 . The slope term is also often called the **regression coefficient** because in the language of basic algebra, β_1 is the coefficient of X_i in the equation. There are three variables in this simple regression model in that there are three terms with the subscript i , indicating that they vary across observations. The parameters do not have the i subscript because they are constants; they do not vary across observations.

The outcome variable is Y_i ; other terms that are more or less synonymous with *outcome variable* include *dependent variable*, *response*, and *criterion*. The predictor variable is X_i ; other terms that are more or less synonymous with *predictor variable*, depending on context, include *independent variable*, *regressor*, *explanatory variable*, and *covariate*. Often, researchers prefer to use the terms *independent* and *dependent variable* in the context of experimental research, whereas the terms *predictor* and *outcome variable* are used in observational, naturalistic research, and finally the terms *explanatory* and *response* variables might be generalizable

to any research context. Personally, I view these sets of terms as essentially interchangeable regardless of the research context because they are treated the same way mathematically when statistical models are fitted to data. For brevity, and perhaps just out of habit, I primarily use the terms *predictor* and *outcome* variable throughout this text (although, incidentally, most examples come from observational research contexts, but the principles presented in this text apply to models for experimental data as well).

The third variable in Equation 1.5 is the **disturbance** or **error** term, ε_i , which represents the inaccuracy of the model's ability to reproduce the value of the outcome variable perfectly for a given observation. Unlike Y_i and X_i , which are directly measured by the researcher, the error is an unobserved variable that is not directly measured and instead arises as a property of the model. Although it is not explicitly indicated in Equation 1.5, the simple regression model also includes a parameter, σ^2 , to capture the variance of the errors; that is, $\sigma^2 = \text{VAR}(\varepsilon_i)$.

This text follows the standard notational practice in the statistics and methodology literature of using Greek letters to represent *population* parameters. There is an unfortunate tendency in some research areas (as well as the output format of the IBM SPSS Statistics® software package) to use the lowercase Greek letter beta (β) to denote the *sample* statistic estimating the standardized regression slope (often referring to it as the 'beta weight'),¹ whereas a capital Roman letter B is used to denote the sample statistic estimating the unstandardized regression slope.² I feel that the latter notational practice is likely to become confusing as the ordinary linear regression model is expanded into more elaborate models, and it is much clearer if Greek letters such as β are always used to represent population parameters. Estimates of model parameters calculated from sample data are statistics, but they will also be referred to as **parameter estimates** throughout this text. To distinguish parameter estimates from the actual parameter symbolically, we use the 'hat' symbol. For example, $\hat{\mu}$, or 'mu hat', is the estimate of the population mean parameter μ (also commonly denoted as \bar{Y} to represent the sample mean of variable Y), whereas $\hat{\beta}$, or 'beta hat', is the sample estimate of a population regression slope parameter β .



Section recap

Simple linear regression model

The simple linear regression model for the relation between an outcome variable Y and a predictor X is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

The intercept parameter, β_0 , is the predicted value of Y when X equals 0.

The slope parameter, β_1 , is the predicted amount that Y changes when X increases by 1.

The magnitude of β_1 is therefore a (population) effect size for the association between X and Y .

¹Standardized regression slopes are defined and discussed in Chapter 2.

²Contrary to popular impression, this notational practice is also not consistent with the APA style guide; see APA, 2010: 119 and 122.

Research example for the simple regression model

To provide a context in which the simple regression model might be used, let's consider an actual research study: A graduate-student researcher is interested in whether and how certain personality characteristics relate to aggression. The researcher collects a sample of $N = 275$ undergraduate university students who each complete the Buss-Perry Aggression Questionnaire (BPAQ; Buss and Perry, 1992) and the Barratt Impulsiveness Scale (BIS; Barratt, 1994), among other questionnaires. At a basic level, the researcher wants to devise a *model* for aggression (operationalized with BPAQ scores) to represent (or describe or explain) how and why people vary on this important outcome.

This dataset is available on the text's webpage (<https://study.sagepub.com/flora>) along with annotated input and output from several popular statistical software packages showing how to reproduce the analyses presented in this chapter.



Before beginning to fit models to empirical data, it is always wise to examine the data descriptively and especially graphically. Thus, for this example, we begin by looking at the sample distribution of the outcome variable, BPAQ scores. Because the BPAQ score is an (approximately) continuous variable and there are $N = 275$ observations, it's best to examine its distribution graphically rather than with a frequency table, for instance, with the boxplot and histogram in Figure 1.2. These plots show that the sample distribution of BPAQ scores is unimodal and (approximately) symmetric, with the bulk of the scores falling between 2.0 and 3.0. The center of the distribution seems to be just above 2.5, and there are no outliers.

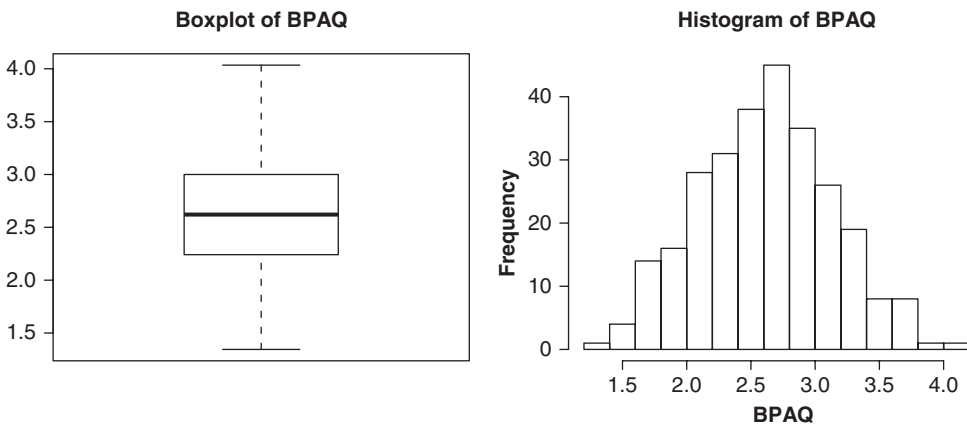


Figure 1.2 Univariate distribution of BPAQ scores

Next, we might apply a density smoother (technically, a *kernel density estimate*) over the histogram of the sample data to approximate the (infinite) population probability distribution (see Wand and Jones, 1995). Likewise, we can superimpose the normal distribution curve that best fits the sample data; the histogram with the smoother and normal curve is shown in Figure 1.3. Now we can see that the sample distribution of BPAQ is remarkably close to a normal distribution; although the estimated density has a few small bumps, it is close to the

superimposed normal curve.³ Having an outcome variable with a distribution that's so close to a normal distribution is nice but unusual. But as we will see in later chapters, it is not always problematic for the outcome to be non-normal, and sometimes we don't need to do anything about it at all. Next, we might wish to have some numerical summaries, or descriptive statistics, to describe the sample distribution of BPAQ more specifically, as shown in Table 1.1. These summary statistics essentially support the observations made earlier from the graphs.

Table 1.1 Univariate descriptive statistics from aggression dataset

Variable	Min	Q1	Mdn	Q3	Max	M	SD	Skewness	Kurtosis
BPAQ	1.35	2.24	2.62	3.00	4.03	2.61	0.52	0.01	-0.41
BIS	1.42	1.42	2.27	2.54	3.15	2.28	0.35	0.36	-0.22
Age	17.00	18.00	18.00	20.00	50.00	20.21	4.96	3.70	15.43
Alcohol	0.00	3.00	12.00	24.00	96.00	16.00	15.87	1.50	3.09

Note. $N = 275$. Min = minimum, Q1 = first quartile (or 25th percentile), Mdn = median (or 50th percentile), Q3 = third quartile (or 75th percentile), Max = maximum, M = mean, SD = standard deviation.

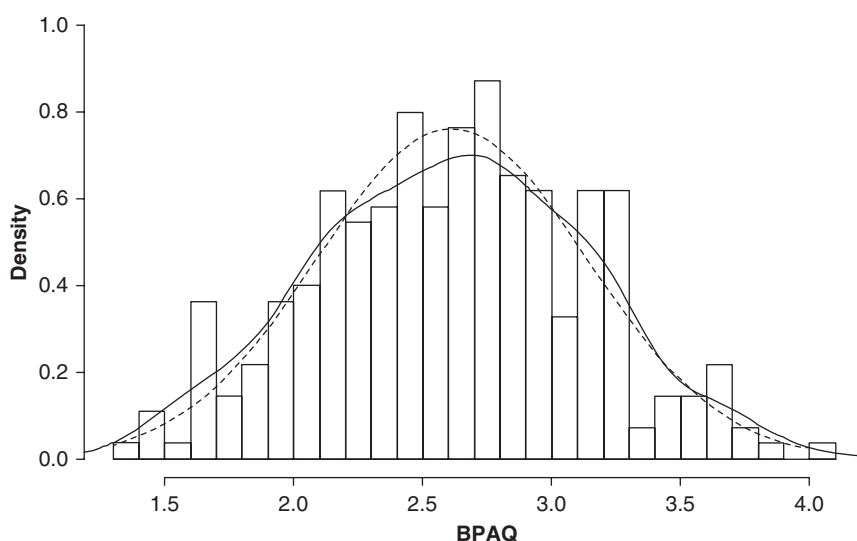


Figure 1.3 Histogram of BPAQ with fitted kernel density smoother (solid curve) and fitted normal distribution (dashed curve)

Initial model for Y: Intercept-only model

Our first model for the BPAQ outcome variable isn't interesting or substantively informative, but it will give us a standard against which to compare the main regression model

³Note that the y-axis for Figure 1.3 is 'density,' or relative frequency on a probability scale, whereas the y-axis of the previous histogram in Figure 1.2 is just raw 'frequency,' on the scale of the number of cases at each observed value of BPAQ. Scaling frequency into relative frequency does not affect the shape of a distribution.

that incorporates the BIS predictor. In the absence of any predictor variable, the best model for BPAQ scores is simply an expression of the distribution's central tendency, such as the population mean, μ . Of course we also know that there is variation around the mean, in that many (if not all) of the observations are not exactly equal to the mean. With this in mind, we can express this initial model like so:

$$Y_i = \mu + \varepsilon_i, \quad (1.6)$$

where Y_i is the BPAQ score for case i and the error term, ε_i , indicates that case i is likely to have a value of Y that deviates from the mean. This model is referred to as the **intercept-only model** because we can view it as a regression model without any predictors; that is, Equation 1.6 can be rewritten as

$$Y_i = \beta_0 + \varepsilon_i, \quad (1.7)$$

where $\beta_0 = \mu$. Note also that Equation 1.7 is equivalent to Equation 1.5 if $\beta_1 = 0$. This intercept-only model also includes the parameter σ^2 for the error variance; that is, $\sigma^2 = \text{VAR}(\varepsilon_i)$. In the intercept-only model, the error variance summarizes the extent to which observations differ from the mean, which is the value of Y predicted by the model for every case in the population.

The parameter estimates for the intercept-only model are the familiar sample mean, \bar{Y} (recall that $\bar{Y} = \hat{\mu}$, and therefore, for this model, $\bar{Y} = \hat{\beta}_0$), and sample variance, $s_Y^2 = \hat{\sigma}^2$. The square root of the sample variance, $\sqrt{s^2} = s$, is of course the sample standard deviation. Recall that the sample variance is calculated from the squared deviations of the mean from each observation:

$$\text{VAR}(Y) = s_Y^2 = \frac{\sum_{i=1}^N (Y_i - \bar{Y})^2}{N - 1} = \frac{SS_Y}{N - 1}, \quad (1.8)$$

where SS stands for **sum of squares**.

Next, if we substitute the sample mean for the population mean (i.e., substitute the parameter estimate for the parameter) in the model, we have

$$Y_i = \hat{\mu} + e_i$$

or

$$Y_i = \bar{Y} + e_i.$$

Thus,

$$Y_i - \bar{Y} = e_i.$$

Note that because we have substituted parameter estimates for the actual population parameters, the deviation between the observed Y_i and the value of Y_i predicted by the model expressed in terms of parameter estimates is a **residual** term denoted e_i . In general, for both

the intercept-only model and more complex regression models, the residual e_i for a given case i based on the estimated model will almost always differ from the error term ε_i based on the true population model.

Next, following from the definition of SS , we have

$$\sum_{i=1}^N (Y_i - \bar{Y})^2 = \sum_{i=1}^N (Y_i - \hat{\mu})^2 = \sum_{i=1}^N e_i^2 = SS_e.$$

This is a simple but important result that generalizes to other, more elaborate models: The sum of squared residuals is the sum of squared deviations of the model's predicted \hat{Y} (here, the predicted \hat{Y} is just the mean) from the individual observed Y values.

From Table 1.1, we see that the sample mean BPAQ score is 2.61 which, absent any predictor variables, gives the estimated predicted value of BPAQ for all cases in the population. Yet, the standard deviation of BPAQ is 0.52, indicating that there are substantial individual differences in the BPAQ outcome variable; soon we will incorporate predictor variables to explain this variability.

Furthermore, given the shape of the BPAQ distribution displayed in Figure 1.3, we should be comfortable that it is a reasonable approximation of a normal distribution such that $Y_i \sim N(\mu, \sigma^2)$; that is, the outcome Y is distributed as a normal variable with population mean μ and variance σ^2 . This statement then implies that the errors from the intercept-only model are also sampled from a normal distribution such that $\varepsilon_i \sim N(0, \sigma^2)$; that is, ε is distributed as a normal variable with mean 0 and variance σ^2 . Therefore, in this basic intercept-only model, the variance of the outcome variable equals the variance of the errors, but that won't be the case as soon as we add other variables to the model. Because the variance of the errors equals the variance of Y , the model has not explained *any* of the variation in the outcome. One way to evaluate the quality of subsequent models is to see how much of the variation in Y is explained; that is, how much smaller is the error variance compared with the observed variance of Y ? In other words, how much better is a model with one or more predictors of Y than a model without any predictors?



Section recap

Intercept-only model

The **intercept-only model** is

$$Y_i = \beta_0 + \varepsilon_i.$$

Because this model has no predictor variables, its parameter estimates are the sample mean $\bar{Y} = \hat{\mu}$ and sample variance $s^2 = \hat{\sigma}^2$.

Because the model does not explain any of the outcome variable's variability, the residual variance is equal to the observed variance of Y .

Subsequent models which do include one or more predictors can be compared with this initial model to determine the amount of observed variance that is explained by the predictor(s).

Focal model for Y: Simple regression with a single predictor

In the current research example, the main goal is to model aggression (measured with BPAQ) as a function of personality traits, such as impulsivity (measured with the BIS), not merely to describe the univariate distribution of BPAQ scores without any predictors. Thus, answering the actual substantive research question depends on devising a model that's more elaborate than the intercept-only model. Hopefully for this hypothetical personality researcher, the new model with BIS will be a statistical improvement over the basic intercept-only model. Again, before estimating the model, we should investigate the data graphically. Now that we are introducing a second variable, BIS scores, we can use a scatterplot to visualize the distribution of BPAQ scores conditioned on BIS scores, as depicted in Figure 1.4. The plot suggests that those with higher impulsivity scores tend to have higher aggression scores. That is, BPAQ scores covary with BIS scores, but the relation is far from perfect.

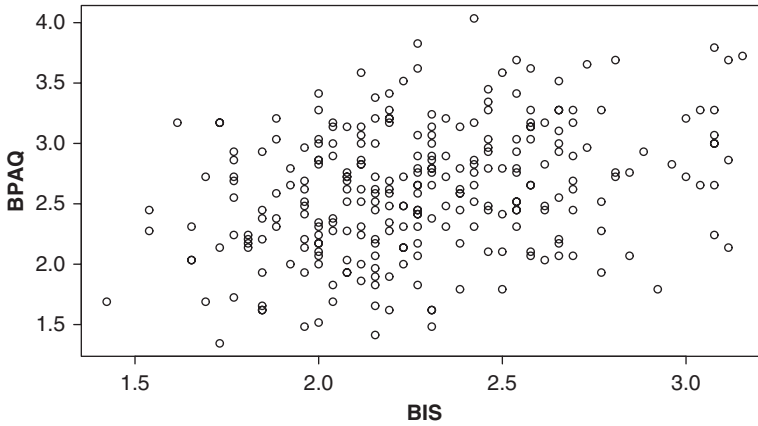


Figure 1.4 Scatterplot of BPAQ scores against BIS scores

The sample **product-moment covariance** between two variables is the fundamental, basic ingredient that allows us to estimate the parameters of linear models (including advanced linear models, such as structural equation models). Consider the formula for the sample variance again:

$$\text{VAR}(Y) = s_Y^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{N-1} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(Y_i - \bar{Y})}{N-1} = \frac{SS_Y}{N-1}.$$

This is an index of the amount that Y varies with itself. The formula for sample covariance is similar, but it incorporates both Y and X :

$$\text{COV}(Y, X) = s_{YX} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{N-1} = \frac{SCP}{N-1},$$

where SCP stands for **sum of cross-products** in that $(Y_i - \bar{Y})(X_i - \bar{X})$ is the *cross-product* between Y and X for a given case. Thus, s_{YX} is an index of the amount that Y covaries with X

in the sample. In our current research example, the covariance between BPAQ scores and BIS scores is $s_{YX} = 0.06$. But the value of a covariance depends on the scales of Y and X ; for this reason, covariance can be difficult to interpret as a basic descriptive statistic measuring the (linear) association between X and Y .

Therefore, to aid interpretation of the association between two variables, the covariance can be standardized into the **Pearson product-moment correlation** (usually simply referred to as the ‘correlation’ or ‘correlation coefficient’). The population correlation is represented with the Greek letter *rho*, ρ , whereas the sample correlation is represented with the Roman letter r . Thus, r estimates ρ . As readers are likely aware, a correlation is *always*⁴ a value between -1 and $+1$, which is obtainable with the formula

$$r = \frac{\text{COV}(Y, X)}{\sqrt{\text{VAR}(Y) * \text{VAR}(X)}} = \frac{s_{YX}}{\sqrt{s_Y^2 s_X^2}}. \quad (1.9)$$

In our current example, the correlation between BPAQ and BIS is $r = .32$, which is consistent with the earlier observation that individuals with higher BIS scores tend to have higher BPAQ scores. More impulsivity is associated with more aggression. A 95% CI estimate of ρ is $(.21, .42)$; that is, with 95% confidence, the interval from $.21$ to $.42$ captures the population correlation. In other words, a population correlation between BPAQ and BIS in the range of $.21$ to $.42$ is likely to have produced these data.

The correlation describes the extent to which two variables are *linearly* associated, implying that a straight line going through the middle of the points in the bivariate scatterplot (i.e., the simple regression line) does an adequate job of representing, or modeling, the pattern of covariation between Y and X . It is easier to assess the quality of the straight-line model if we also add a nonparametric LOWESS (‘LOcally WEighted Scatterplot Smoothing’) regression curve to the plot, which is designed to capture more subtle, nonlinear regularities in the data that might not be well represented with the parametric regression line (see Fox, 2008: 21–4 and 496–507 for details on how LOWESS curves are calculated). Because it’s nonparametric and is susceptible to chance variations in sample data, this LOWESS curve is difficult to use for population inference, but it can help evaluate the adequacy of the parametric linear regression model, given the observed data.⁵ As shown in Figure 1.5 for the current example, the LOWESS curve is consistent with the straight line, so we should be comfortable using the line as a model for the data. Hence, although the relation between BPAQ and BIS is not particularly

⁴Here, the word ‘always’ is emphasized because when estimating certain advanced models, the obtained parameter estimates sometimes imply that a correlation between two variables is greater than $+1$ or less than -1 . Such an estimated model solution must be discarded as *improper* because any correlation outside the -1 to $+1$ range is inherently nonsensical. This result commonly occurs in the context of advanced modeling procedures such as multilevel modeling (Chapter 6) and structural equation modeling (Chapters 9 and 10).

⁵The LOWESS curve is especially susceptible to chance variations in data when the sample size is small; in this case, only a few cases may cause the curve to display dramatic bends. In such a situation, the curve may *overfit* the data rather than smoothing over minor, sample-specific variation. One can control the extent to which a LOWESS curve captures such minor data characteristics by adjusting its *span* (see Fox and Weisberg, 2011: 117). In our current example, the sample size is rather large, and so the LOWESS curve is resistant to the influence of just a few unusual cases.

strong, it does appear as if the ordinary simple regression line is a reasonable model for the relation. That is, there doesn't seem to be any particular pattern in the data that would be grossly misrepresented by the straight line that best fits the data, namely, the **least-squares regression line**, which is defined later in this section.

As a quick aside, when a scatterplot shows a consistent nonlinear trend, the bivariate association may be effectively described using **Spearman's rank-order correlation coefficient**. Spearman's correlation is calculated by first transforming both variables into ranks, and then Spearman's correlation is the product-moment correlation between the two sets of ranks (see Chapter 8 for an example). The discrepancy between this rank-order correlation and the original, raw-score, product-moment correlation can be used as a diagnostic to determine whether a linear regression model is likely to be distorted by nonlinear patterns in the data.

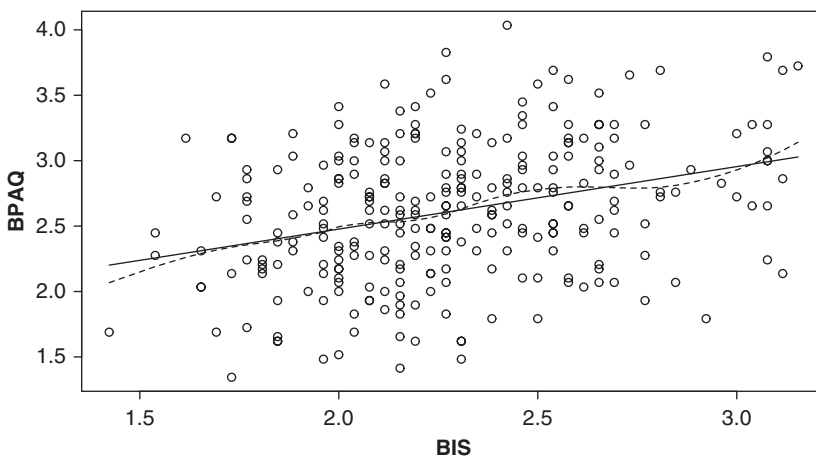


Figure 1.5 Scatterplot of BPAQ scores against BIS scores with fitted linear regression line (solid line) and fitted LOWESS curve (dashed line)

But before considering the actual mathematical formulas for the linear model, it is important to distinguish between **predicted values** of Y , also known as *fitted* or *model-implied* values, and **observed values** of Y . The observed values of Y are the actual measured values for the outcome that are in the dataset. In the current example, there are $N = 275$ participants, or cases, with BPAQ aggression scores. Each of the 275 BPAQ scores is an observed value of Y . We can see in Figure 1.5 that most of the observed values do not fall on the regression line, and some are far from the line. The predicted values of Y , represented as \hat{Y} , are the Y values determined by the regression line across the X continuum. As we are using BIS impulsivity scores to predict or model BPAQ scores, then the \hat{Y} values are the predicted, model-implied BPAQ scores at each possible BIS score (i.e., at each value of X). Because the prediction is not perfect (the correlation between BPAQ and BIS does not equal 1), most of the observed BPAQ scores do not equal the score that is predicted by the linear regression of BPAQ on BIS. Next, the **residual** for the individual, or case, i is the difference between that participant's observed Y value and the corresponding predicted \hat{Y} value, given the *estimated* linear effect of the observed value for the predictor X :

$$e_i = Y_i - \hat{Y}_i$$

So observed scores falling on an estimated regression line have residuals equal to zero, whereas observed scores that are far from the line have large residuals.



Section recap

Foundations for a simple linear regression model

A scatterplot of an outcome variable against a potential predictor helps determine whether it is appropriate to model the relation between the two variables using a straight-line function, i.e., a linear regression equation.

Covariance is a descriptive value measuring the strength of linear association between two variables; the **product-moment correlation** is a covariance which has been standardized to range between -1 and $+1$.

A **predicted value**, \hat{Y}_i , is the outcome variable score for individual i that is predicted from the regression line given that individual's score on the predictor variable, X_i .

An **observed value**, Y_i , is the actual outcome variable score for individual i regardless of that individual's score on the predictor variable, X_i .

Given a population regression equation, the regression error for individual i is $\varepsilon_i = Y_i - \hat{Y}_i$.

Given an estimated regression equation, the **residual** for individual i is $e_i = Y_i - \hat{Y}_i$.

Simple linear regression: Model specification

Model specification simply refers to establishing the model's parameters with one or more equations giving the hypothetical mathematical relations among the variables. Often specification also involves statements about a model's assumptions regarding variance terms or probability distribution, although these assumptions sometimes arise because of estimation method (see later discussion in this section) and are not always a part of specification. Specification of the simple linear regression model was presented earlier, but here it is reiterated with slightly more detail.

The one-predictor linear regression model can be specified in two equivalent ways. The first expression is in terms of the predicted values of the outcome variable:

$$\hat{Y}_i = \beta_0 + \beta_1 X_i, \quad (1.10)$$

where \hat{Y}_i is the predicted value on the outcome for case i and X_i is the observed value of the predictor for the same case i . As described previously, β_0 is the intercept parameter for the line, which is the value of \hat{Y} when $X = 0$. β_1 is the slope parameter of the line, also called the regression coefficient, which is the amount that \hat{Y} differs when X increases by one unit. In other words, a one-unit increase in X is associated with a change in \hat{Y} equal to β_1 . Of course sometimes β_1 is a negative number, indicating the amount that \hat{Y} decreases per unit increase in X , just as a correlation can be positive or negative.

The second way of expressing the same model substitutes the observed Y for the predicted Y on the left-hand side of the equation:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad (1.11)$$

where Y_i is the observed score on the outcome for case i ; β_0 , β_1 , and X_i are the same as that presented earlier, and ε_i is the error term. Because $\varepsilon_i = Y_i - \hat{Y}_i$, it is easy to see that the two expressions of the model (Equations 1.10 and 1.11) are algebraically equivalent. Finally, a third parameter of the model is the variance of the errors, $\sigma^2 = \text{VAR}(\varepsilon_i)$, which captures the extent to which observed values differ from predicted values.

Because β_1 measures the amount or extent to which the predictor variable is related to the outcome, it is an effect-size parameter. More specifically, because β_1 describes the linear effect of X on Y in terms of the scale of Y , it is an *unstandardized* effect-size measure. The correlation is a type of *standardized* effect-size measure because it indicates the strength of the linear relation between two variables on a standard scale from -1 to $+1$ rather than the observed scale of the outcome variable.

Simple linear regression: Model estimation

In statistical modeling, **estimation** refers to the process of calculating estimates of model parameters from sample data. For any particular kind of model, there are many potential methods of estimation, but of course, some are better than others, where ‘better’ typically means that parameter estimates from an optimal estimation method are **unbiased**, **consistent**, and **efficient**, provided the method’s assumptions are met. Briefly, a parameter estimate is unbiased if the mean of its sampling distribution equals the true value of the parameter at a given sample size; an estimate is consistent if its value approaches the parameter value as the sample size increases toward infinity; and an estimate is efficient if, compared with other estimation methods, its sampling distribution has the smallest variance.

In simple linear regression, the model parameters β_0 and β_1 are most commonly estimated from sample data using formulas derived using the **ordinary least-squares (OLS)** method of estimation. When its assumptions are met, parameter estimates calculated with OLS are unbiased, consistent, and efficient. These assumptions are addressed in both this chapter and subsequent chapters on linear regression.

The OLS formulas give values for $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the set of squared residuals in the sample are as small as possible. More specifically, the line defined by the OLS estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ minimizes the **sum of squared residuals**, which is also known as the **error sum of squares**:⁶

$$SS_e = \sum_{i=1}^N e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2.$$

Substituting the simple regression line as the model for \hat{Y} , we see that the residual sum of squares is

$$SS_e = \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^N (Y_i - [\hat{\beta}_0 + \hat{\beta}_1 X_i])^2.$$

⁶But the term ‘error sum of squares’ is misleading because of the distinction between *residuals*, which are deviations from \hat{Y} based on sample estimates of the parameters, and *errors*, which are deviations from \hat{Y} based on the true (but unknown) parameters. In practice, this sum-of-squares term can only be calculated using residuals.

The goal of OLS estimation is then to find parameter estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$, that make the quantity SS_e as small as possible. Calculus is required to show the derivation of the formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$ that lead to the minimization of squared residuals (see Fox, 2008: 78–81). The result of this basic calculus problem gives the OLS estimate of β_1 as

$$\hat{\beta}_1 = \frac{\text{COV}(Y, X)}{\text{VAR}(X)} = r \left(\frac{s_Y}{s_X} \right). \quad (1.12)$$

With this expression, it is easy to see that if both X and Y are standardized (i.e., transformed so that their sample means equal 0 and variances equal 1), then the slope estimate is equal to the correlation.⁷ Next, given the estimate of β_1 , the OLS estimate of β_0 is

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

Incidentally, when both X and Y are standardized, the intercept estimate equals zero.

For the current example modeling BPAQ aggression scores as a function of BIS impulsivity scores, the parameter estimates are $\hat{\beta}_0 = 1.52$ and $\hat{\beta}_1 = 0.48$, and thus, the estimated OLS regression line is

$$\hat{Y}_i = 1.52 + 0.48X_i.$$

This is the equation of the line going through the middle of the points in the scatterplot in Figure 1.5. In fact, an OLS regression line is guaranteed to pass through the point (\bar{X}, \bar{Y}) . The slope estimate $\hat{\beta}_1 = 0.48$ indicates that a one-unit increase in BIS scores predicts an increase of 0.48 in BPAQ scores. Again, this is the effect of BIS on BPAQ; it is an effect-size estimate for this data analysis. The intercept estimate $\hat{\beta}_0 = 1.52$ indicates that with a BIS score equal to zero, the predicted BPAQ score is 1.52. In this example, the intercept parameter is not substantively useful because a BIS score equal to 0 is outside of the range of the data, given the way that the questionnaire is scored. The intercept parameter often is of little interest in ordinary regression modeling, but in certain contexts, the interpretation of the intercept may be extremely important.



Section recap

Specification and estimation of the simple linear regression model

Specification establishes a model's parameters using one or more equations giving the hypothetical mathematical relations among the variables. Often specification also involves statements about a model's assumptions regarding variance terms or probability distributions.

Once a model is specified, **estimation** is the procedure by which the model's parameters are estimated from sample data.

Ordinary least squares (OLS) is the most common estimation method for linear regression. OLS produces the parameter estimates that minimize the sum of squared residuals.

⁷Standardized regression coefficients are discussed in Chapter 2.

Simple linear regression: Statistical inference

The relation between BPAQ and BIS scores is clearly evident in the scatterplot, but the simple regression slope estimate $\hat{\beta}_1 = 0.48$ does not seem like a strong effect size given the observed scales of BPAQ and BIS. Thus, an important aspect of model interpretation does in fact involve significance testing and confidence interval estimation to determine whether this slope estimate is distinguishable from a population slope equaling zero (using a significance test) or, more comprehensively, to determine a plausible range for the value of the population slope (using a confidence interval or CI).

To make such inferences from the OLS parameter estimates to the unknown population parameters, it is necessary to make some assumptions about the error random variable, that is, the deviations from the estimated model. Primarily, we assume that $\varepsilon_i \sim N(0, \sigma^2)$, meaning that the errors come from a normal distribution with its mean equal to zero and variance equaling σ^2 . Thus, in contrast to popular perception, in OLS regression, the normal distribution assumption is about the unmeasured errors and there is *no* explicit assumption that either measured variable Y or X is normal. Additionally, the fact that there is no i subscript on σ^2 is important because it reflects the assumption that the error variance is the same (i.e., *constant variance*) for all observations, regardless of their value for X . This constant variance assumption is also known as **homogeneity of variance** or **homoscedasticity**: The variance of the errors is assumed homogeneous, or constant, across all values of X . Finally, there are a few other assumptions for OLS regression that we address in later chapters.

Usually, we are most interested in testing the null hypothesis that the regression slope equals zero:

$$H_0: \beta_1 = 0.$$

Note that if the null hypothesis is true, the regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

becomes

$$Y_i = \beta_0 + 0 \times X_i + \varepsilon_i$$

or

$$Y_i = \beta_0 + \varepsilon_i,$$

which is the intercept-only model presented earlier (Equation 1.7). Thus, the significance test for the slope $\hat{\beta}_1$ is also a *model comparison* test indicating whether the one-predictor, simple regression model is significantly different from the intercept-only model.

The ratio of the OLS estimate of a simple regression slope to its estimated *standard error* follows a t distribution with $(N - 2)$ degrees of freedom. Thus, the null hypothesis for the slope is evaluated with a t test:

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}, \quad (1.13)$$

where $s_{\hat{\beta}_1}$ is the estimated standard error of $\hat{\beta}_1$. A formula for this standard error estimate is

$$s_{\hat{\beta}_1} = \sqrt{\frac{SS_e / (n - 2)}{SS_x}}.$$

Recall that the **standard error** of a statistic is the standard deviation of its sampling distribution; hence, the standard error reflects the average amount that a statistic (such as a sample mean or a regression slope estimate) randomly drawn from its sampling distribution is expected to differ from the true parameter value.

Additionally, we can get a confidence interval estimate of β_1 using the usual symmetric confidence interval approach based on the t distribution:

$$\hat{\beta}_1 \pm s_{\hat{\beta}_1} \times t_{\alpha},$$

where t_{α} is the appropriate *critical t* value for a $(1 - \alpha)\%$ confidence interval and α is the pre-determined probability of a Type I error (usually $\alpha = .05$, leading to 95% confidence intervals). Of course, there is an exact correspondence between the t test and the confidence interval in that the null hypothesis is not rejected at the given alpha level if the confidence interval overlaps 0 and is rejected if the confidence interval does not contain zero. Ultimately, though, the confidence interval is more informative than the null hypothesis significance test: Not only does the confidence interval indicate whether 0 (the value given by the null hypothesis) is a plausible value for the population slope parameter, but also the confidence interval gives a whole range of plausible values for the plausible parameter values.

This form of the t test and confidence interval construction also applies to the intercept estimate, $\hat{\beta}_0$, but these results are often of little, if any, substantive interest (although standard statistical software does typically include the estimated standard error, t , and p value of $\hat{\beta}_0$ within regression modeling output).

In the current example of the regression of BPAQ aggression scores on BIS impulsivity scores, $s_{\hat{\beta}_1} = 0.085$. Therefore, we have

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{0.4777}{0.0854} = 5.59.$$

The two-tailed p value for this t statistic is less than .0001, so using the conventional Type I error probability $\alpha = .05$, we reject the null hypothesis that the population slope equals zero. Thus, even though the effect-size estimate $\hat{\beta}_1$ is somewhat small, its true population value is likely to differ from zero. Furthermore, this t test implies that the simple regression model with BIS as a predictor explains the data significantly better than the intercept-only model does.

Compared with the null hypothesis test, more specific information regarding the likely value of the population regression slope is given by a confidence interval around the slope estimate. Here, the 95% confidence interval estimate of β_1 is (0.31, 0.65), suggesting that a population slope of any value between 0.31 and 0.65 is likely to have produced these data. Hence, the data suggest that the population effect size could be as large as 0.65. This would seem like a large effect given that the BPAQ and BIS have similar scales; a one-point increase in BIS would predict a BPAQ score that is larger by 0.65 BPAQ units. But the lower end of the

confidence interval suggests that the population effect could be 0.31, less than half as large as that suggested by the upper end.

Additionally, recall that in the intercept-only model, the variance of the residuals equals the variance of Y :

$$\text{VAR}(Y_i) = \text{VAR}(\beta_0 + \varepsilon_i) = \text{VAR}(\varepsilon_i).$$

Therefore, the intercept-only model does not explain any variance in Y . Let's call the intercept-only model for our current example 'Model 0'. The variance of BPAQ, and thus of the residuals from Model 0, equals 0.2746. Then, let Model 1 be the simple linear regression of BPAQ on BIS. The residual variance from Model 1 equals 0.2463. Thus, adding the predictor BIS has reduced the residual variance by $(0.2746 - 0.2463) = 0.0283$. The ratio $(0.0283) / (0.2746) = 0.1031$ then indicates the proportion of variance in BPAQ explained by Model 1. The (positive) square root of this proportion, $\sqrt{0.1031}$, equals the correlation of .32 between BIS and BPAQ given earlier.

More generally, the **coefficient of determination**, R^2 , is a descriptive statistic often reported as a measure of the overall standardized effect size given by a regression model. This statistic may be calculated as

$$R^2 = \frac{\text{VAR}(Y) - \text{VAR}(e)}{\text{VAR}(Y)}.$$

Because the sample-size terms in the numerator and denominator of this equation cancel out, it simplifies to an expression based on sum-of-squares terms:

$$R^2 = \frac{SS_Y - SS_e}{SS_Y}. \quad (1.14)$$

Consequently, in addition to the significance test for $\hat{\beta}_1$, we can also compare the intercept-only model to the one-predictor simple regression model using R^2 , which indicates the proportion of variance in the Y that's accounted for by including X in the model. In simple linear regression, R^2 also equals the squared correlation, r^2 , between X and Y and the p value for the t test for $\hat{\beta}_1$ is identical to the p value for the significance test of the correlation. Hence, in our current example, we can also conclude that the proportion of the variance in Y accounted for by BIS is significantly greater than zero.



Section recap

Statistical inference with the simple linear regression model

The null hypothesis that $\beta_1 = 0$ can be evaluated using a t test in which

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}},$$

where $s_{\hat{\beta}_1}$ is the estimated standard error of $\hat{\beta}_1$.

(Continued)

(Continued)

Even more information about the true parameter value is provided by a confidence interval, which can be calculated with

$$\hat{\beta}_1 \pm s_{\hat{\beta}_1} \times t_{\alpha}$$

where t_{α} is the *critical t* value for a $(1 - \alpha)\%$ confidence interval with α as the predetermined Type I error probability.

The validity of these inferential methods depends on the assumption that the errors are normally distributed with constant (homogenous) variance across the range of X .

The **coefficient of determination**, R^2 , gives the proportion of outcome variable variance explained by the predictor(s) in a regression model.

Simple regression with a dichotomous predictor

Contrary to what is presented in some introductory statistics texts, the predictor variable in a simple correlation or regression analysis need not be continuous. The predictor can also be dichotomous (i.e., categorical with two values or categories, or binary) without violating any assumptions, although the outcome variable should still be continuous. For instance, continuing with our applied research example, we can use simple linear regression to model the relation between BPAQ aggression scores and gender. A scatterplot depicting the relation between BPAQ and gender is in Figure 1.6, with gender *dummy-coded* so that 0 = male and 1 = female. Because gender is a nominal variable, the choice of numerical values for its two categories is completely arbitrary, and the results presented here generalize to any numerical coding scheme for a binary variable. But dummy codes of 0 and 1 produce an especially convenient interpretation of the OLS regression parameters, as

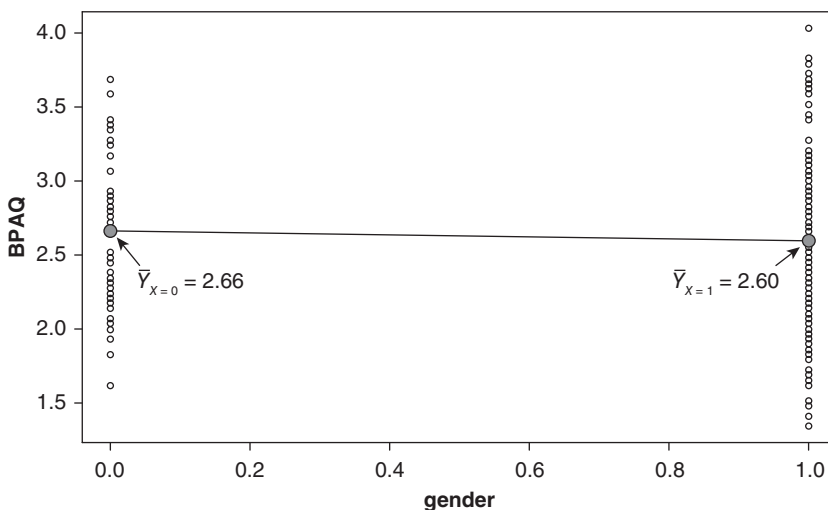


Figure 1.6 Scatterplot of BPAQ scores against gender with fitted regression line

we will see later in the book (dummy coding is addressed further in Chapter 3). In addition to providing further insights about the simple regression model, the primary purpose of this section is to build a foundation for Chapter 3, which describes how categorical predictors with any number of categories (i.e., not just binary variables) can be incorporated in a linear regression model.

Because there are only two possible values of X , 0 and 1, a straight line is the mathematically simplest way to connect the middle of the points in the plot above $X = 0$ to the middle of the points above $X = 1$. Specifically, when X is binary and coded 0 or 1, the OLS regression line is the line connecting the point $(0, \bar{Y}_{X=0})$ to the point $(1, \bar{Y}_{X=1})$. This line is superimposed in Figure 1.6, which provides the scatterplot of BPAQ scores against gender.

Once the binary predictor has been coded into two numerical values (such as the dummy codes used here), the formulas presented earlier for the OLS regression parameter estimates are directly applicable with the two dichotomous values used for X_i . In our current example with gender dummy coded, the estimated OLS regression line modeling BPAQ scores as a function of gender is

$$\hat{Y}_i = 2.66 - 0.06X_i.$$

Because $\hat{\beta}_0$, the estimated intercept, is the predicted value of Y when $X = 0$, then 2.66 is the predicted BPAQ score among males. The intercept estimate is in fact the mean BPAQ score for males:

$$\hat{Y}_{male} = 2.66 - 0.06(0) = 2.66 = \bar{Y}_{male}.$$

Next, because $\hat{\beta}_1$, the estimated slope, is the predicted difference in Y when X_i changes by one unit, then -0.06 is the predicted difference in Y between males and females. This parameter estimate thus gives the difference between the male mean and the female mean:

$$\hat{Y}_{female} = 2.66 - 0.06(1) = 2.60 = \bar{Y}_{female}.$$

Once again, the estimated regression slope represents the unstandardized effect size; the simple mean difference between males and females on the outcome is $\hat{\beta}_1 = 0.06$. Here, this is a small effect given that the range of observed BPAQ scores is approximately 1.0 to 4.0. Additionally, the plot in Figure 1.6 clearly illustrates the weakness of the effect. Thus, even though the scale of the BPAQ operational variable is essentially meaningless, we can easily tell that this simple mean difference is a small effect without converting it to some type of standardized effect size.

Moving on to inference, when the predictor in a simple regression model is binary, the t test of whether the slope significantly differs from 0 (Equation 1.13) is equivalent to the well-known independent-groups t test comparing the outcome variable means of the two groups formed by the dichotomous predictor variable:

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{\bar{Y}_2 - \bar{Y}_1}{s_{\bar{Y}_2 - \bar{Y}_1}},$$

where $s_{\bar{Y}_2 - \bar{Y}_1}$ is the standard error of the difference between means calculated using the familiar pooled-variance statistic formed under the homogeneity of variance assumption. Thus, the independent-groups t test is a special case of significance testing in OLS regression. Likewise, the p value for the correlation between a dichotomous X and a continuous Y will be equal to that for the t test.⁸ In the current example, this t test is not significant, $t(273) = -0.797$, $p = .43$, implying that the gender difference on BPAQ scores may not differ from zero in the population. Finally, the confidence interval for the regression slope is identical to a confidence interval for the difference between two independent means; in the current example, the 95% confidence interval is $(-0.22, 0.09)$, conveying a range of plausible values for the population mean difference $(\mu_{\text{female}} - \mu_{\text{male}})$.

Because the number of male participants ($n = 57$) is much smaller than the number of females ($n = 218$) in this sample, we should pay extra attention to the viability of the homogeneity of variance assumption because it is well known that the consequences of heterogeneous variance for the t test are more serious when the group sample sizes are markedly discrepant. The sample variances of BPAQ are similar across gender ($s^2_{\text{male}} = 0.26$, $s^2_{\text{female}} = 0.28$), which should restore some comfort with the homogeneity of variance assumption despite the unbalanced sample size. But regardless of the homogeneity of variance assumption, as we observed earlier, the estimated effect size is tiny, and therefore, nonsignificance seems to be the appropriate result for this t test.

Dichotomous outcome?

In contrast to the previous section where a continuous outcome variable is regressed on a dichotomous predictor, it is critical to recognize that when the outcome itself is dichotomous, the simple linear regression model is not appropriate for the data. If Y_i is coded as either 0 or 1 for each observation i , then the regression line will typically produce predicted values, \hat{Y} , that are outside the range from 0 to 1 and thus improper. Furthermore, the residuals cannot be normally distributed with homogeneity of variance, which leads to incorrect significance tests and confidence intervals for the parameters (Fox, 2008: 337). Instead, it is more reasonable to model dichotomous outcomes using a nonlinear modeling procedure such as logistic regression (or, similarly, probit regression). More generally, whenever the outcome variable is categorical, whether dichotomous or with multiple categories, the ordinary linear regression model is likely to produce misleading results and alternative nonlinear models for categorical outcomes, such as those within the class of generalized linear models, are more appropriate (see Fox, 2008, for a textbook-length treatment of these models). Fortunately, having a solid understanding of ordinary linear regression provides an excellent foundation for learning about logistic regression and other generalized linear models.

⁸The product-moment correlation between a dichotomous variable and a continuous variable calculated according to the formula presented earlier (Equation 1.9) is a special type of correlation known as the *point-biserial* correlation. Here, the point-biserial correlation between gender and BPAQ is $r = -.05$. In that the correlation is a type of standardized effect measure, the same value is obtained regardless of the numerical coding scheme used for the dichotomous variable (dummy-coded or otherwise).



Section recap

Dichotomous variables in the simple linear regression model

When a continuous outcome variable is regressed on a dummy-coded dichotomous (or binary) predictor, the intercept estimate will equal the mean of the group coded zero, $\hat{\beta}_0 = \bar{Y}_{X=0}$, and the slope estimate will equal the difference between the means of the two groups, $\hat{\beta}_1 = (\bar{Y}_{X=1} - \bar{Y}_{X=0})$.

Furthermore, the t test for the slope estimate is equivalent to the independent-groups t test and a $(1 - \alpha)\%$ confidence interval for the population slope is identical to a $(1 - \alpha)\%$ confidence interval for the difference between the two population means.

It is generally inappropriate to model a dichotomous outcome variable (or any categorical outcome) using a linear regression model.

BASIC REGRESSION DIAGNOSTIC CONCEPTS

Regression diagnostics are graphical and numeric methods for evaluating the extent to which a regression model fitted to data is an adequate representation of that data and for evaluating the trustworthiness of inferential conclusions about the model's parameter estimates. In particular, regression diagnostics are used to check the extent to which the model's assumptions have been violated and to examine whether any unusual or outlying observations may be impacting the results. Although diagnostic methods primarily show their strength when applied to multiple regression models (i.e., models with two or more predictors), I introduce the basic concepts here because they can be more plainly and simply illustrated with the simple regression model.

Linearity

First and foremost, as emphasized earlier, the simple regression model specifies a *linear* relation between the predictor and outcome variables. Linearity is most easily examined with a scatterplot of the observed variables (although this may not be sufficient when we move to multiple regression), which can be enhanced with a superimposed nonparametric regression curve as demonstrated previously in Figure 1.5. In this example, the straight-line regression model did seem to be a good representation of the positive (but weak) relation between BPAQ scores and BIS scores.

But let's also consider an example where things don't work out so well. The student researcher who collected the BPAQ aggression data was also interested in studying alcohol use among university students. Therefore, $N = 270$ participants in her study also answered questions about their quantity and frequency of alcohol use over the past year, which the researcher then combined to produce an overall alcohol-use index.⁹ Next, although most of the undergraduate-student participants were 18 or 19 years old, there was a considerable

⁹Specifically, the study used the Quantity \times Frequency index described in Chassin, Flora, and King (2004).

number of nontraditional students who were in their late-20s or 30s, and even a few as old as 50. The results reported in the literature suggest that heavy alcohol use is most common among young adults and declines thereafter. Thus, the researcher is interested in examining the association between age and alcohol use in her sample.

Although I recommend examining data descriptively before proceeding with model fitting, and we will look at the alcohol-use data momentarily, for didactic purposes, let's first consider the results from the OLS simple regression of alcohol use on age. The estimated regression slope is $\hat{\beta}_1 = -0.46$, indicating that for each one-year increase in age, the predicted amount of alcohol use declines by 0.46 of a point on the alcohol-use scale. This estimated effect seems somewhat weak given that the alcohol-use scale ranges from 0 to as high as 95 in the current sample, but it is significantly different from zero, $t(268) = 2.41$, $p = .02$. Further demonstrating the weakness of the effect, R^2 is only .02. Thus, the researcher might be tempted to conclude that there is a small, but significant, negative relation between age and alcohol use, and stop there. But as we see next, this conclusion, although correct in a broad sense, also may be an oversimplification of the data.

The scatterplot of alcohol use by age in Figure 1.7 (enhanced with the fitted regression line and a nonparametric LOWESS regression curve) supports the notion that high amounts of alcohol use are less common among older participants compared with those in early adulthood. But this simple observation seems to neglect some other complexities in the data. One immediately noticeable issue is that there are far fewer participants at the older ages than at age 20 and younger, implying that these data give us scant information about the population level of alcohol use among older individuals. But more pertinent to our discussion of the adequacy of the linear regression analysis is that the LOWESS curve suggests that the typical amount of alcohol use increases from the youngest age until around age 23, but then decreases steadily until around age 35.¹⁰ This nonlinear pattern is consistent with the results

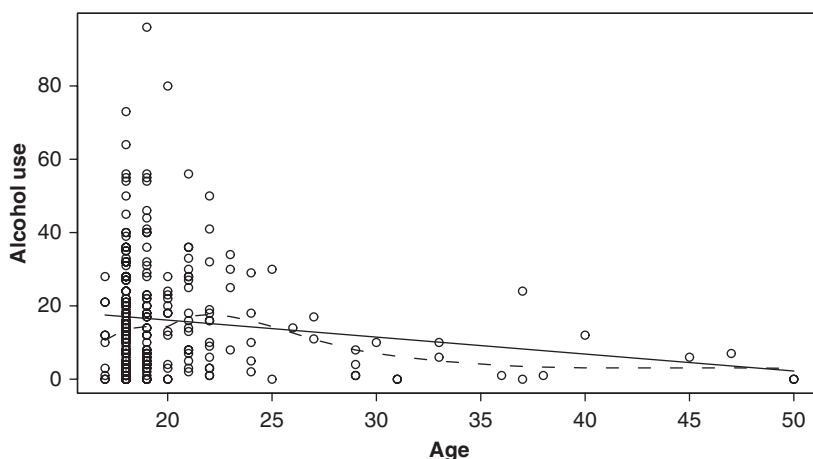


Figure 1.7 Scatterplot of alcohol-use scores against age with fitted linear regression line (solid line) and fitted LOWESS curve (dashed line)

¹⁰As described in Footnote 5, with small samples, the LOWESS curve may overfit the data. Here, though, N is large, especially at the younger ages, implying that the LOWESS curve is likely capturing meaningful trends in the data.

in the published literature on alcohol use among young adults and is therefore important to recognize in these data, but the straight-line model only captures the broad observation that older participants tend to have lower amounts of alcohol use than younger participants. Nonetheless, this subtle nonlinear trend is not the only challenge that these data pose for the linear regression model, as we will see shortly.

Distribution of residuals

Next, recall from earlier that correct inference for the regression parameters assumes that the errors are normally distributed with constant, or homogeneous, variance, $\varepsilon_i \sim N(0, \sigma^2)$. When this assumption is violated, the Type I error rate (if the null hypothesis is true) and power (if the null hypothesis is false) of the t test for the slope parameter estimate is compromised, and the width of its confidence intervals are incorrect. Because of the central limit theorem, if the sample size is large (>100 or so for a simple regression model), then the normality assumption is less critical, and hypothesis tests and confidence intervals are still valid, as long as the other assumptions are met. Nevertheless, non-normality still impacts the efficiency of OLS estimates, meaning that alternative estimators of the regression parameters can produce smaller standard errors and therefore are associated with greater statistical power and more narrow confidence intervals for the estimates. Therefore, it is important to evaluate the tenability of the assumption that the errors are normally distributed with homogeneity of variance.

Recall that the *errors* from the population model (Equation 1.11) differ from the *residuals* from the regression equation formed with the sample-based OLS parameter estimates. Because the errors are unobservable, we must instead work with the residuals. With most statistical software packages, it is possible to extract the residuals from a fitted regression model and then the residuals can be examined like any other variable. Returning to the example regression of BPAQ aggression scores on BIS impulsivity scores, Figure 1.8 presents a histogram of the residuals from this model. Not surprisingly, given the approximate

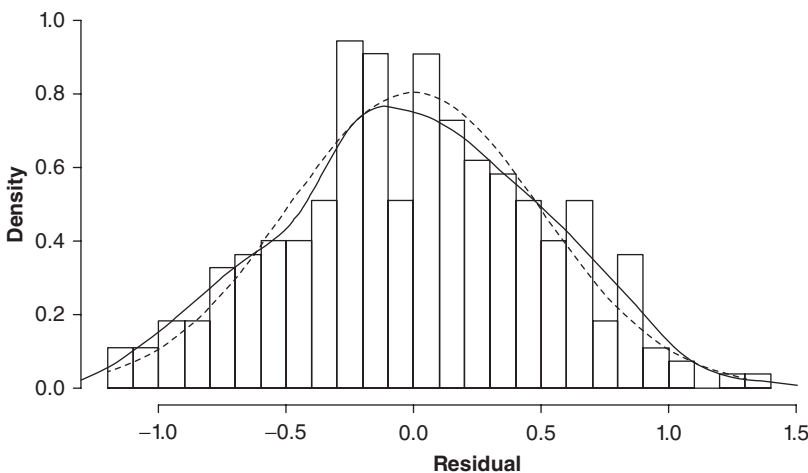


Figure 1.8 Histogram of residuals from OLS regression of BPAQ scores on BIS scores with fitted kernel density smoother (solid curve) and fitted normal distribution (dashed curve)

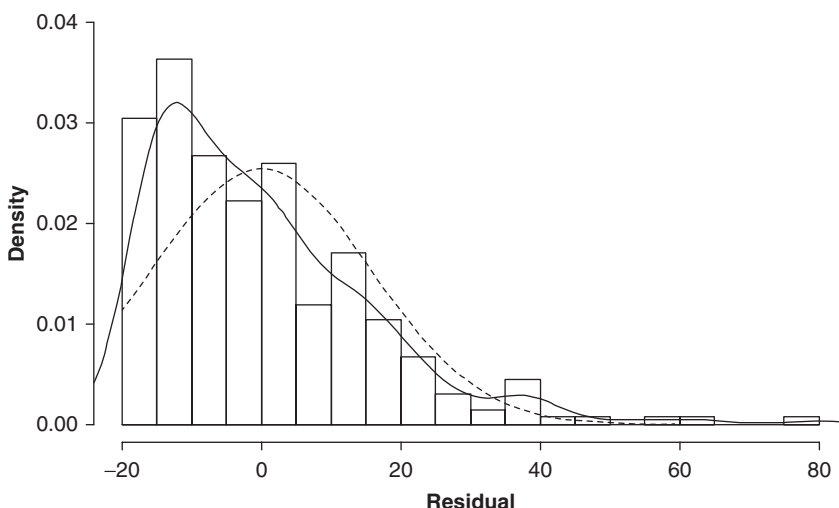


Figure 1.9 Histogram of residuals from OLS regression of alcohol-use scores on age with fitted kernel density smoother (solid curve) and fitted normal distribution (dashed curve)

normal distribution of BPAQ and its reasonably linear (but weak) association with BIS, these residuals appear to provide a reasonable approximation of a normal distribution. In contrast, Figure 1.9 gives a histogram of residuals from the regression model of alcohol use on age, which is clearly non-normal.

Homoscedastic versus heteroscedastic residuals

Perhaps more important than determining whether residuals are normally distributed is determining whether they are **homoscedastic**, meaning that their variance is consistent across the values of the predictor X ; in other words, it's important to examine the residuals for homogeneity of variance. In our example regression of BPAQ scores on BIS scores, because the predictor (BIS) is approximately continuous, it's best to examine the residuals conditioned on BIS using a scatterplot, as in Figure 1.10. To aid interpretation, the scatterplot in this figure is enhanced with a nonparametric LOWESS curve along with smoother applied to the root-mean-square positive and negative residuals from the LOWESS curve (see Fox and Weisberg, 2011: 117–18), which give a graphical summary of the spread of the data in the plot. Because these dashed curves are approximately equidistant from each other as they move from low to high levels of BIS, the residuals in the plot are in fact evidencing homoscedasticity; that is, the spread of the data is constant across the values of the predictor, BIS. Therefore, we can be satisfied that the homogeneity of variance assumption is met for this linear model. Incidentally, the fact that the solid LOWESS curve in the plot is mostly straight and horizontal is further evidence that a linear model is a good representation of the relation between BPAQ and BIS scores.

Unlike the model regressing BPAQ on BIS, the estimated linear model predicting alcohol use from age appears to have **heteroscedastic** residuals, that is, residuals which are unevenly spread across the age predictor. In particular, Figure 1.11 displays a scatterplot of this model's

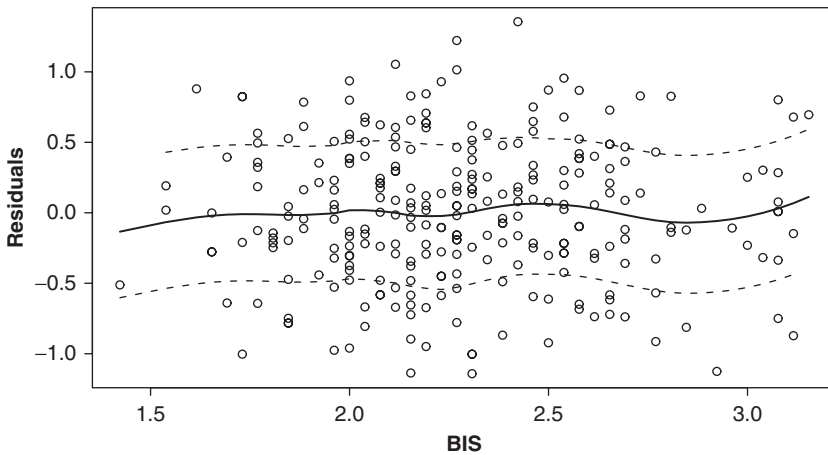


Figure 1.10 Scatterplot of residuals against BIS scores with fitted LOWESS curve (solid line) and smoothed root-mean-square positive and negative residuals from the LOWESS curve (dashed lines)

residuals against age, again enhanced with the nonparametric LOWESS spread. In this figure, the residuals are more spread out at younger ages (especially around age 20) compared with higher values of age. Thus, the homogeneity of variance assumption is violated for this model, which then casts doubt on the inferential conclusion that the linear slope in the regression of alcohol on age significantly differs from zero. Note that Figure 1.11 also further illustrates the nonlinear pattern in the data.

Investigating the variability in a dataset can have value beyond just evaluating the homogeneity of variance assumption for regression. Here, for example, it may be of substantive importance to recognize there is much more variation in the amount of alcohol use among

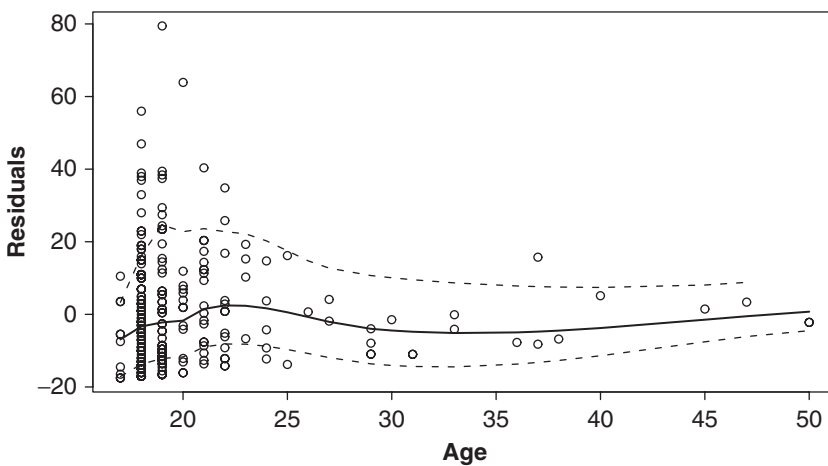


Figure 1.11 Scatterplot of residuals against age with fitted LOWESS curve (solid line) and smoothed root-mean-square positive and negative residuals from the LOWESS curve (dashed lines)