

David Weisburd  
David B. Wilson  
Alese Wooditch  
Chester Britt

# Advanced Statistics in Criminology and Criminal Justice

*Fifth Edition*

# **Advanced Statistics in Criminology and Criminal Justice**

# **Advanced Statistics in Criminology and Criminal Justice**

**Fifth Edition**

**David Weisburd**

*Department of Criminology, Law and Society, George Mason University, Fairfax, VA, USA and Institute of Criminology, Faculty of Law, Hebrew University of Jerusalem, Jerusalem, Israel*

**David B. Wilson**

*Department of Criminology, Law and Society, George Mason University, Fairfax, VA, USA*

**Alese Wooditch**

*Department of Criminal Justice, Temple University, Philadelphia, PA, USA*

and

**Chester Britt**

*Department of Sociology, Iowa State University, Ames, IA, USA*



**Springer**

David Weisburd  
Department of Criminology, Law and Society  
George Mason University  
Fairfax, VA, USA

Institute of Criminology  
Faculty of Law  
Hebrew University of Jerusalem  
Jerusalem, Israel

Alese Wooditch  
Department of Criminal Justice  
Temple University  
Philadelphia, PA, USA

David B. Wilson  
Department of Criminology, Law and Society  
George Mason University  
Fairfax, VA, USA

Chester Britt (deceased)  
Department of Sociology  
Iowa State University  
Ames, IA, USA

ISBN 978-3-030-67737-4      ISBN 978-3-030-67738-1 (eBook)  
<https://doi.org/10.1007/978-3-030-67738-1>

© Springer Nature Switzerland AG 2007, 2014, 2022

1st edition: @ West/Wadsworth Publishing Company, 1988

2nd edition: @ Thomson/Wadsworth, 2003

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

---

# Contents

---

## Chapter one

### Introduction 1

- Proportionality Review and the Supreme Court of New Jersey: A Cautionary Tale 3
- Generalized Linear Models 7
- Special Topics 13
- References 14

---

## Chapter two

### Multiple Regression 15

- Overview of Simple Regression 17
- Extending Simple Regression to Multiple Regression 23
- Assumptions of Multiple Regression 27
- Measurement Error in the Independent Variables 32
- Regression Diagnostics 33
- Dealing with Outliers and Influential Cases 38
- Testing the Significance of Individual Regression Coefficients 40
- Assessing Overall Model Fit and Comparing Nested Models 41
- Comparing Regression Coefficients Within a Single Model: The Standardized Regression Coefficient 46
- Correctly Specifying the Regression Model 48
- Model Specification and Building 50
- An Example of a Multiple Regression Model 53
- Chapter Summary 59
- Key Terms 60
- Symbols and Formulas 61
- Exercises 63
- Computer Exercises 66
- References 72

---

**Chapter three****Multiple Regression: Additional Topics 73**

- Nominal Variables with Three or More Categories in Multiple Regression 76
- Nonlinear Relationships 80
- Interaction Effects 92
- An Example: Race and Punishment Severity 96
- An Example: Punishment Severity 105
- The Problem of Multicollinearity 109
- Chapter Summary 112
- Key Terms 113
- Symbols and Formulas 113
- Exercises 114
- Computer Exercises 118
- References 126

---

**Chapter four****Logistic Regression 127**

- Why Is It Inappropriate to Use OLS Regression for a Dichotomous Dependent Variable? 130
- Logistic Regression 136
- A Substantive Example: Adoption of Compstat in U.S. Police Agencies 146
- Interpreting Logistic Regression Coefficients 151
- Comparing Logistic Regression Coefficients 158
- Evaluating the Logistic Regression Model 166
- Statistical Significance in Logistic Regression 169
- Chapter Summary 173
- Key Terms 175
- Symbols and Formulas 176
- Exercises 178
- Computer Exercises 181
- References 185

---

**Chapter five****Multiple Regression with Multiple Category Nominal or Ordinal Measures 187**

- Multinomial Logistic Regression 190
- Ordinal Logistic Regression 205
- Chapter Summary 219
- Key Terms 220
- Formulas 221
- Exercises 222
- Computer Exercises 225
- References 231

---

## Chapter six

### Count-Based Regression Models 233

- The Poisson Distribution 236
- Poisson Regression 239
- Over-Dispersion in Count Data 249
- Quasi-Poisson and Negative Binomial Regression 251
- Zero-Inflated Poisson and Negative Binomial Regression 255
- Chapter Summary 259
- Key Terms 260
- Symbols and Formulas 261
- Exercises 262
- Computer Exercises 263
- References 271

---

## Chapter seven

### Multilevel Regression Models 273

- A Simple Multilevel Model 277
- Random Intercept Model with Fixed Slopes 287
- Random Coefficient Model 295
- Adding Cluster (Level 2) Characteristics 300
- Chapter Summary 309
- Key Terms 310
- Symbols and Formulas 311
- Exercises 312
- Computer Exercises 315
- References 319

---

## Chapter eight

### Statistical Power 321

- Statistical Power 323
- Components of Statistical Power 326
- Estimating Statistical Power and Sample Size for a Statistically Powerful Study 335
- Summing Up: Avoiding Studies Designed for Failure 346
- Chapter Summary 347
- Key Terms 348
- Symbols and Formulas 348
- Computer Exercises 349
- References 365

---

## Chapter nine

### Randomized Experiments 367

- The Structure of a Randomized Experiment 368
- The Main Advantage of Experiments: Isolating Causal Effects 371

Internal Validity	375
Selected Design Types and Associated Statistical Methods	377
Block Randomized Designs	389
Using Covariates to Increase Statistical Power in Experimental Studies	400
Chapter Summary	402
Key Terms	403
Symbols and Formulas	404
Exercises	408
Computer Exercises	409
References	415

---

## Chapter ten

### Propensity Score Matching 417

The Underlying Logic Behind Propensity Score Matching	419
Selection of Model for Predicting Propensity for Treatment	421
Matching Methods	422
Assessing the Quality of the Matches	427
Sensitivity Analysis for Average Treatment Effects	431
Limitations of Propensity Score Matching	433
Chapter Summary	435
Key Terms	436
Symbols and Formulas	437
Exercises	437
Computer Exercises	438
References	448

---

## Chapter eleven

### Meta-analysis 451

A Historical Note	454
The Logic of Meta-analysis	455
The Effect Size	456
Meta-analysis of Effect Sizes	467
Forest Plots	474
Moderator Analysis	475
Handling Statistically Dependent Effect Sizes: Robust Standard Errors	480
Publication Selection Bias	482
Chapter Summary	485
Key Terms	486
Symbols and Formulas	486
Exercises	490
Computer Exercises	491
References	496

**Chapter twelve****Spatial Regression 499**

- Why Can't We Use OLS Regression with Spatial Data? 501
- How Do We Define Spatial Relationships? 502
- What Is Spatial Regression? 510
- Which Type of Spatial Regression Should I Use? 514
- Spatial Regression Example 518
- Chapter Summary 523
- Key Terms 524
- Symbols and Formulas 525
- Exercises 526
- Computer Exercises 528
- References 535

**Glossary 537****Index 543**

## C h a p t e r   o n e

---

# Introduction

## **Proportionality Review and the Supreme Court of New Jersey: A Cautionary Tale**

---

What is Proportionality Review?

How Were Advanced Statistical Techniques Used in Proportionality Review in New Jersey?

What Were the Underlying Problems that Led to Misuse of Advanced Statistical Techniques in this Study?

## **G e n e r a l i z e d L i n e a r M o d e l s**

---

What is the Generalized Linear Model (GLM)?

How can GLM Models be Linear Even Though the Variables Examined are Nonlinear?

What Does the Link Function in GLM Do?

What is the Importance of the Error Terms in GLM?

What are Examples of Different Types of GLMs?

## **S p e c i a l T o p i c s**

---

What additional statistical additional approaches are examined in the text?

**A**N ADVANCED VOLUME for classes in statistics is often difficult to define. What are advanced and what are basic statistics? In our *Basic Statistics in Criminology and Criminal Justice*, we covered materials that would generally be taught in introductory statistics classes at the undergraduate level and first semester Master's level. We dealt with measurement, basic descriptive statistics, and statistical significance. We also laid out the basic approaches for modeling association between variables. While many important studies use these approaches, researchers often encounter problems that need more advanced statistical solutions. Such solutions are the focus of this volume.

As in our *Basic Statistics* volume, we focus on comprehension and not computation. Nonetheless, the concepts and equations in this volume become more complex as the issues we seek to resolve require innovative statistical solutions. We want to encourage you to do your best to understand the underlying decision making behind these statistical solutions. As we discuss in the next section, more complex statistics can be important in solving problems we encounter in research, but sometimes when statistics become complex, they also can lose transparency and can hide underlying problems with statistical analyses. The better your understanding of such statistics, the better you will be at assessing the quality of research.

We begin this introductory chapter with an example where more complex statistics led to misleading statistical conclusions. It is a cautionary note as we begin our journey to advanced statistical approaches. We then introduce the idea of the generalized linear model (GLM). The GLM is behind most of the statistical methods we introduce in this volume, and we wanted to give the reader a sense at the outset of what GLM is and from where it derives. Regression analyses are key in some form to most of the chapters in our volume. The generalized linear model is the key ingredient of regression analyses, though as we show in later chapters, it can be

adapted to many different types of statistical outcomes and problems. In concluding the chapter, we will focus on additional methodological, statistical, and analytic tools that we have included in our volume, discussing briefly why we think they are important for researchers.

## Proportionality Review and the Supreme Court of New Jersey: A Cautionary Tale<sup>1</sup>

---

Can we sometimes go wrong in using advanced statistics? It is of course always possible that results today will be supplanted by better statistical work later on. And indeed, this is simply part of good science. But sometimes policy makers and even researchers can make the mistake of accepting statistical work as providing valid results when it does not. And this is something we can avoid by understanding the statistical approaches used, and the assumptions that underlie the conclusions that are reached. This is why it is so important for you to not just use advanced statistics but to understand how they are developed and what key assumptions they require.

A good example of a case where advanced statistics was incorrectly interpreted is that of proportionality review of the death penalty in the New Jersey Supreme Court. The Court had spent more than a decade considering the role of statistical models in assessment of the fairness of death penalty sentencing. Its efforts spanned a series of Special Masters who had been appointed to help the court in deciding on the proper use of such methods and had been described by one observer, Leigh Bienen, as a “live experiment in the use of social science data by a court.”<sup>2</sup> Proportionality review may be defined more generally as the examination and comparison of cases that result in a death penalty with other similar cases. Death sentences in this context are proportional if other similar cases have led to similar outcomes. Cases are disproportionate if other similar cases did not result in a death sentence. In New Jersey, as in other states that have instituted proportionality review systems, the legislature had mandated proportionality review.

In an order issued on July 29, 1988, the New Jersey Supreme Court appointed Professor David Baldus of the University of Iowa Law School as Special Master to assist the court in developing such a system. Baldus, importantly, was strongly established as a proponent of social science applications in the law and had recently completed a major empirical

---

<sup>1</sup>We draw heavily in this section from Weisburd (2001).

<sup>2</sup>See Bienen (1996).

study of the death penalty in Georgia. The methodology recommended to the court was developed by Professor Baldus over a 3-year period.

The most sophisticated of the methods suggested and the one deemed most reliable by Professor Baldus were based on logistic multiple regression analyses (see Chap. 4). It was termed by the court in its deliberations as an index-of-outcomes test (State v. Marshall 1992). Using this technique, Baldus claimed he was “able to rank-order the [death penalty] cases according to overall defendant culpability, as measured by the presence or absence in the cases of factors that appear to influence prosecutorial and jury decision-making” (Baldus and New Jersey Administrative Office of the Courts 1991a, p. 93). He noted to the court that the “resulting statistical model conformed to what one would expect from jurors who attempted to base their decisions on a balancing of aggravating and mitigating circumstances” (Baldus and New Jersey Administrative Office of the Courts 1991, p. 94).

At first glance, the results were impressive. A series of regression models were produced for the court that allowed the prediction of culpability levels for individual defendants. In theory, the court could use these culpability levels to define whether someone who received a death penalty would have been expected to gain this sentence on the basis of the factors included in the model. If the predicted culpability level was low and the individual received the death penalty, this would create a basis for challenging the proportionality of the sentence. Certainly, this seems like a scientific application of the multiple regression methods we cover in this book to court sentencing data. Indeed, the approach was described in an article that appeared in *Chance*, a publication of the *American Statistical Association* dedicated to new directions in the practical application of statistics and computing (Baldus and Woodworth 1993).

From these models, estimates were made of the relative culpability of defendants in specific cases. Overall, Baldus argued in the main part of his report that the predicted culpability scores for the cases produced by the regression discriminate “quite well between the majority of cases in which the death-sentencing rates are low, cases with middling death sentencing rates, and those with very high rates” (Baldus and New Jersey Administrative Office of the Courts 1991a, p. 95). However, in a technical appendix, Professor Baldus discussed a series of statistical problems that were encountered in the estimation of the regression models upon which these estimates were based.

Our goal was to develop multivariate models with which to measure defendant culpability on the basis of the case characteristics that appeared to be most important to New Jersey’s prosecutors and jurors. Our vehicle for the task was logistic multiple regression analysis... The first issue was how to include in a model all of the statutory aggravating and mitigating circumstances, let alone any other factors, with such a small sample of cases and especially 39 death sentences. Logistic regression, the preferred technique, we quickly discovered was out of

the question. Logistic analyses run in SAS would not converge. (Baldus and New Jersey Administrative Office of the Courts 1991b, p. 11)

Put in lay terms, there were too few cases and too many variables for a logistic regression to be estimated. In statistical terms, as we note in Chap. 4, the models estimated could not converge, or reach a single statistical solution. As we describe in Chap. 4, lack of convergence is an indicator of serious problems in the specification of a logistic regression model. Professor Baldus goes on to explain what was done in order to overcome this problem:

To deal with this problem we used discriminant analysis, which is capable of estimating regression coefficients with the same properties as logistic regression coefficients. Most importantly, discriminant analysis can handle a much larger number of independent variables. We tested the comparability of the results from the two procedures with small models that both methods could handle. The results were comparable, and the discriminant analysis showed no signs of bias or tendency toward misspecifications. (Baldus and New Jersey Administrative Office of the Courts 1991b, p. 11)

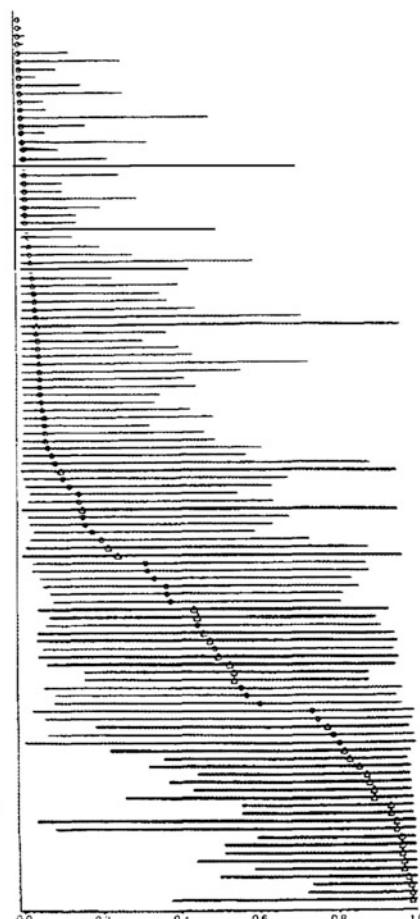
Faced with the reality that the preferred logistic regression technique could not provide a statistical solution, Baldus looked to an alternative approach. He and his colleagues used *discriminant functions*, which are not based on maximum likelihood estimates and therefore do not face problems of convergence (see Chap. 4), as a first step in estimating logistic regression procedures. Importantly, the use of an alternative estimating technique did not purge the models of the problems that caused lack of convergence in the first place. It merely allowed estimation of coefficients even though such problems were present. Baldus, however, notes that a series of diagnostic techniques were used that suggested that significant biases did not develop from taking this alternative approach.

Despite Baldus' assurances, there are elements of these models that make them highly suspect, especially for a decision as important as death penalty sentencing. Many of the coefficients in these models were very large, sometimes referred to as *jumbo coefficients*. One of the measures suggested that the odds of being sentenced to death are 400,000 times higher for those cases in which a public official is killed. This based on a sample of only 132 cases.

An additional suggestion of the problems inherent in this approach, and perhaps a more easily understood indication that the ground was not very solid, is gained when we examine the 95% confidence intervals around probability estimates for individual defendants given by Baldus in his report (Fig. 1.1). A confidence interval in this case provides a basic method for assessing how stable the estimates gained are for each specific defendant. The tighter the interval, the more stable the estimate. The larger the interval, the less confidence we can put in the specific result.

**Figure 1.1**

*Predicted probabilities of death sentence and associated confidence interval for 113 penalty trial cases, 1983-1991*



A 95% confidence interval is commonly used in social science and is also commonly applied to public opinion polls where the upper and lower limits are referred to as the margin of error of the poll. In statistical terms, the interval can be defined as the values within which we are fairly confident that the population or true estimate (as opposed to sample estimate) may be found. As indicated in the figure presented by Baldus in *State v. Marshall* for post-1983 penalty trial cases, many of the confidence intervals are very large and span most of the probability range from 0 to 1. Few of these intervals are below 10 or 20% of the probability range. This means that we can have very little confidence in the reliability of the predictions produced by this method.

Importantly, Professor Baldus provided the Administrative Office of the Courts (AOC) of New Jersey with an appendix in which this information was available. But no one in the office had reviewed this information carefully, until one of us (Weisburd) was invited by the AOC to look at the materials. The Supreme Court had already written a case in which Professor Baldus' analyses were used in the decision. But when the Chief Judge of the Court called Weisburd to understand what the statistical work meant, she clearly was never briefed on the key assumptions and problems of the statistical methods employed. This is a cautionary tale because it emphasizes how important it is to understand the statistics you are using. You cannot just accept on faith that the results of other people's work, or even your own, are believable. You need to make sure that you understand what underlies the conclusions that are reached. The Special Master later appointed by the court to review the problems discussed here noted:

In a society that so fervently values the sanctity of life, it is not surprising that we would turn to the certitude of science and statistics in our attempt to ensure fair application of our capital punishment statutes. In this *Final Report*, Professor Baldus thus recommended adoption of the index-of-outcomes test which employs numerous multivariate logistic regression analyses in order to rate and rank defendant's culpability. This recommendation was adopted by the Court in *State v. Marshall*. The vision of a mechanized approach purported to deliver empirically-based quantitative assessments of criminal culpability. Its promise was to extract human judgment from human decision making. Unfortunately, our experience with the index-of-outcomes test discloses that this was a promise unkept. We have attained only a bitter semblance of efficiency by attempting to rely on these statistics and the calculation of chance. (Baime and Administrative Office of the Courts 1999, p. 76)

## Generalized Linear Models

---

In Chap. 2, we introduce multiple ordinary least squares (OLS) regression. In the following chapters, we address alternative types of regression which solve specific problems that the researcher can encounter. In Chap. 3, we look at nonlinear variables, interaction effects, and dummy variables reflecting multiple nominal categories. In Chap. 4, we examine logistic regression that deals with the case of a binary dependent variable. In Chap. 5, we extend this model to ordinal dependent variables and to nominal dependent variables with three or more categories, and in Chap. 6, to count variables, such as the count of crimes. In Chap. 7, we again extend this model to hierarchical models in which there is nesting, such as the nesting of hot spots within communities or repeated measures within individuals. All of these approaches can be seen as being part of a larger family of statistical models termed generalized linear models (GLMs).

The GLM framework was originally proposed by Nelder and Wedderburn (1972). The framework helps establish the common

underlying statistical method used by the large variety of regression methods that are widely used and many of which are presented in this book. While we will leave discussion of specific GLM models to the following chapters, we think there is value in understanding the basic outlines of this framework.

We can express a basic linear model as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon,$$

where  $Y$  is the dependent variable,  $\beta_0$  is the intercept, each  $\beta_k$  is a regression coefficient or slope associated with an independent variable,  $x_k$ , and  $\epsilon$  reflects the prediction errors. This model has three components: the dependent variable, the structural model, and the random errors.

The structural component of the model is what makes generalized linear models *linear*. It is this component that produces the predicted values of the dependent variable. Thus, we can express the predicted value for  $Y$  as a function of the structural component, as shown here:

$$\hat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k, \quad \text{Equation 1.1}$$

where  $\hat{Y}$  is our prediction based on the regression function. In formal statistics, this is called our *expectation* for  $Y$  given the collection of independent variables  $x$ . This is written as  $E(Y|x)$  and is read as the expected value of  $Y$  given  $x$ . Thus, we can rewrite Eq. (1.1) as

$$E(Y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

The structural component is linear in its elements. We mean by this that each independent variable has a linear relationship with the expected or predicted value for  $Y$  and that relationship is a function of the regression coefficient. For example, if our regression coefficient,  $\beta_k$ , for a specific independent variable,  $k$ , is 0.50, then the expected value for  $Y$  increases by 0.50 for every one-unit increase in  $x_k$ . This is a very important concept for you to keep in mind. Many of the GLM models we examine in this book deal with nonlinear measures either as a dependent variable or an independent variable. GLM is by definition linear in its elements. But GLM approaches can introduce nonlinearities into these models. Such models remain linear in their elements, even though they account for nonlinearities as we describe in later chapters.

To illustrate this, assume we have a simple linear regression model with a single independent variable, as shown here:

$$Y = \beta_0 + \beta_1 x_1 + \epsilon.$$

If the value for the intercept,  $\beta_0$ , is 2 and the value for the regression coefficient for  $x_1$  is 0.5, the expected values for  $Y$  increase linear by 0.50 as  $x_1$  increases in one-unit increments. This is shown below.

$$E(Y|x=1) = 2 + 0.5(1) = 2.5$$

$$E(Y|x=2) = 2 + 0.5(2) = 3.0$$

$$E(Y|x=3) = 2 + 0.5(3) = 3.5$$

$$E(Y|x=4) = 2 + 0.5(4) = 4.0$$

With multiple independent variables, the expected value for  $Y$  is an additive composite of these linear effects.

This structural linearity does not prevent these models from estimating curvilinear effects or even interaction effects (see Chap. 3). You may find this counterintuitive. However, even in these more complex models, each regression coefficient represents a fixed or linear increase in the expected value of  $Y$  for a one-unit change in the associated independent variable. For example, a common way to estimate a curvilinear relationship between an independent variable and the dependent variable is to have two terms in the model related to a single independent variable, one that reflects the variable in its raw form and the other that is the variable squared. This is shown below:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon.$$

Note, however, that  $x_1^2$  is just another variable in the model. It just so happens to be the square of  $x_1$ , but the expected value for  $Y$  still increases by a fixed amount,  $\beta_2$ , as  $x_1^2$  increases by 1. It is only the combination of these two linear effects that produce a curved prediction line for  $Y$  relative to  $x$ . Thus, the structural model is *linear in its parameters* even if it produces a nonlinear prediction for  $Y$  relative to any given  $x$ .

Why does this matter? First, the structural model is linear even if the relationship being modeled is nonlinear. Second, the structural model is the same across all variants of the generalized linear model. Thus, everything that you learn about building a multiple regression model estimated using ordinary least squares regression discussed in Chaps. 2 and 3 applies to all generalized linear model variants throughout this volume. For all of these

models, you can have one or more independent variables. You can include interaction terms, dummy codes for nominal variables, transformed independent variables, and nested models.

The differences across generalized linear models are not only the type of dependent variable, but how the dependent variable is linked or connected to the structural model, and the assumptions regarding the error term. Focusing first on the dependent variable, a generalized linear model links the expected (predicted) value of  $Y$  to the structural model via some function, which we will denote as  $f()$ . For example, if our link function is the natural log, written as  $\ln()$ , then the expected value for  $Y$  is

$$f(E(Y|x)) = \ln(E(Y|x)).$$

Conceptually, this has the effect of log transforming the dependent variable. Thus, the full regression model for a log-linked generalized linear model could be written as

$$\ln(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon.$$

**Equation 1.2**

We can exponentiate both sides of Eq. (1.2) (recall that exponentiation reverses the natural log function) and express the expected value for  $Y$  in its raw form as the exponent of the structural model:

$$E(Y|x) = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k}.$$

The link function used is determined by the nature of the dependent variable. For continuous and normally distributed dependent variables, the link function is simply the identity function. The identity function leaves  $Y$  unchanged. You could also think of this transformation as multiplication by 1. For other dependent variable types, there are other link functions. The link function for count-type dependent variables is the natural log, as shown above (see Chap. 6). For a dichotomous dependent variable, the link function is either the logit or the probit (see Chap. 4). We will dig much deeper into these issues in later chapters, but the critical concept is that the linear structural model is linked to the dependent variable via a function. It is this function that enables generalized linear models to estimate the expected values for dependent variables that are not normally distributed and that may be constrained in some way, such as to positive values (i.e.,  $y \geq 0$ ) or constrained to the value 0 or 1, as is the case with a binary outcome.

Why not just transform the dependent variable and estimate an ordinary least square regression model? We identify such transformations in Chap. 2. In some cases, such as log-transformed counts where all values of  $y$  are greater than 0, this can be an acceptable approach. In many situations, however, it is not possible to transform each observation. For example, the natural log of 0 is undefined (i.e., negative infinity). Estimating an OLS regression model on log-transformed counts would require dealing in some way with the 0s, possibly by adding 1 to all values. Counts equal to 0 are not a problem in a generalized linear model because the estimation method does not transform the dependent variable but rather minimizes the difference between the observed  $Y$  values and the expected  $Y$  values based on the structural model and the link function.

The logistic regression model discussed in Chap. 4 provides a clear example of this. The link function is the logit. A logit is defined as

$$\text{logit}(\pi) = \ln\left(\frac{\pi}{1 - \pi}\right) \quad \text{Equation 1.3}$$

where  $\pi$  is the expected probability of success. For example, if we have a binary outcome for two groups, such as a treatment condition and a control condition, then the expected probability of success for each group is simply the mean of the 0s and 1s, where 0 is a failure and 1 is a success. We can write this model as

$$E(\text{logit}(\pi)|x) = \beta_0 + \beta_1 x_1, \quad \text{Equation 1.4}$$

where  $x_1$  indicates whether an observation is in the treatment or control condition. Notice, however, that we cannot compute the logit for each observation because each observation was coded as either a failure (0) or a success (1). It is the expectation or predicted mean rate of successes for each combination of the independent variables that is modeled. We can invert this function and rewrite Eq. (1.4) as

$$E(Y|x) = \frac{e^{\beta_0 + \beta_1 x_1}}{1 - e^{\beta_0 + \beta_1 x_1}}.$$

In a model with a continuous independent variable or simply with multiple independent variables, the mean rate of success might be 0 or 1 for some combinations of the independent variables, making it impossible to transform  $Y$  manually and estimate the model using OLS regression. Furthermore, it would be far more difficult than estimating a logistic regression model that handles this through a link function without needing to

**Table 1.1**

Selected variants of the generalized linear model

REGRESSION MODEL	LINK FUNCTION	ERROR	DEPENDENT VARIABLE
Linear	Identity	Normal, homoscedastic	Continuous
Logistic	Logit	Binomial	0s and 1s
Probit	Probit	Binomial	0s and 1s
Poisson	Log	Poisson	0, 1, 2, 3, ...
Quasi-Poisson	Log	Poisson with over-dispersion	0, 1, 2, 3, ...
Negative binomial	Log	Negative binomial	0, 1, 2, 3, ...
Multinomial	Generalized logit	Multinomial	Nominal

Note: Linear regression also includes one-way ANOVA, multiway ANOVA, and ANCOVA.

transform the dependent variable. That is, GLM handles these complexities for you.

The final component of a GLM is the error term. In an OLS regression model (see Chaps. 2 and 3), we assume that the errors are normally distributed with a mean of zero and equal variance across the expected values of  $Y$ . Other variants of GLM, however, make different assumptions regarding the errors, consistent with the nature of the dependent variable and assumptions about the variability of the errors. For example, a Poisson regression model assumes that the errors are Poisson distributed (see Chap. 6). An important feature of this is that the variability in the errors is not equal across expected values of  $Y$ ; rather, the variance is assumed to equal the expected value of  $Y$  for any combination of the independent variables.

Table 1.1 lists some of the common variants of the GLM. There are other variants not listed, as we do not present them in this book (Fox 2016). We have not listed multilevel models in this table (see Chap. 7), although they are also a variant on the GLM. Multilevel models can take many forms, including any listed in this table, such as a multilevel linear model or a multilevel logistic model. The primary difference is the inclusion of additional error terms. The structural model may also become more complicated, reflecting the multilevel nature of the data.

We will go over these issues carefully as we cover different types of GLM models, but we wanted to provide a general introduction to GLM at the outset. The value of the generalized linear model framework is that it unifies many aspects of regression modeling that remain the same across model types. For example, except for linear models, these models are estimated using maximum likelihood methods and, as such, the tests for model fit and comparing nested models are the same for each. Also, the methods and issues related to omitted variable bias or how to include interaction terms or nominal independent variables are the same across these model types (see Chaps. 2 and 3). Once you have mastered the fundamentals of multiple regression modeling, expanding your knowledge

to include other variants becomes relatively easy. Furthermore, this framework provides a useful way to think about which modeling approach best fits your data.

## Special Topics

---

We include in our *Advanced Statistics* volume a series of special topics that we think are important to include in the toolbox of students and researchers in our field. Much of our focus in this *Advanced Statistics* volume is in describing and developing the GLM. But we also wanted to include other approaches to describing data and identifying causal effects. A particularly important approach in criminology and criminal justice is the randomized experiment. Randomized studies provide a particularly strong method for identifying a valid effect of a treatment or intervention. As we describe in Chap. 9, this derives from the advantages that accrue from being able to work with experimental data. Experimental data are those that are produced through a random process of allocation of units in a study. For example, a researcher might randomly allocate drug-involved offenders to an innovative treatment condition and a control condition which received a standard protocol. In some experiments, places are randomly allocated. For example, there are a number of experiments that randomly allocate enhanced or intensive police patrols to crime hot spots. Again, the control condition in such studies is generally standard police responses. Randomized experiments can be hard to implement, but they are generally viewed as allowing stronger causal inferences than nonexperimental studies.

Experimental studies are becoming more common in criminology and criminal justice, and for this reason, we have included a full chapter on experimental statistics (see Chap. 9). We also have included a chapter on propensity score matching (PSM, see Chap. 10). PSM has become a very common method for assessing treatments or interventions absent a true randomized experiment. PSM builds on GLM approaches, but in a way that follows the logic of experimental studies.

We also include a chapter on statistical power (Chap. 8). Statistical power is a common tool in experimental and quasi-experimental studies, but can be used across the spectrum of statistical analyses in criminology and criminal justice. Statistical power assesses the extent to which our statistical analyses are designed in ways to observe an impact of a variable or treatment if such an impact exists in the population you are interested in. Statistical power is a tool that prevents researchers from concluding that a variable or treatment has no impact, when the research design of a study is not developed in a way that allows observations of such an impact. You

certainly would not want a study designed for failure at the outset, or one that did not produce a fair test the intervention of variable being tested.

In Chap. 11, we introduce the statistical approach of meta-analysis. Though, it is still rare to see texts in criminology that devote significant attention to meta-analysis, we think that it is a key method and one that has become centrally important in drawing conclusions in criminology and criminal justice. Meta-analysis provides a method for summarizing groups of studies addressing a common research question. This approach allows researchers to make sense of a large group of studies on the one hand and allows them to take advantage of the stronger claims that can be made when studies are combined.

Finally, in Chap. 12, we provide an introduction to geographic statistical analyses. Geography has become an important consideration in understanding crime, and we think it appropriate to include statistical tools that allow the researcher to take into account geographic impacts.

## References

---

- Baime, D., & Administrative Office of the Courts. (1999). *Report to the New Jersey Supreme Court*.
- Baldus, D. C., & New Jersey Administrative Office of the Courts. (1991a). *Death penalty proportionality review project final report to the New Jersey Supreme Court* 93.
- Baldus, D. C., & New Jersey Administrative Office of the Courts. (1991b). *Methodology appendix, death penalty proportionality review project final report to the New Jersey Supreme Court*.
- Baldus, D. C., & Woodworth, G. G. (1993). Proportionality: The view of the special master. *Chance*, 3(6), 9–17.
- Bienen, L. B. (1996). The proportionality review of capital cases by state high courts after Gregg: On the appearance of justice. *Journal of Criminal Law and Criminology*, 87, 130.
- Fox, J. (2016). *Applied regression analysis and generalized linear models*. Thousand Oaks: Sage.
- Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3), 370–384.
- State v. Marshall. (1992). 130 Supreme Court of New Jersey 109.
- Weisburd, D. (2001). Magic and science in multivariate sentencing models: Reflections on the limits of statistical methods. *Israel Law Review*, 35(2), 225–248.

## C h a p t e r   t w o

---

# Multiple Regression

## **S i m p l e R e g r e s s i o n**

---

What is Regression?

How is a Simple Regression Model Estimated and Interpreted?

## **M u l t i p l e R e g r e s s i o n**

---

What is Ordinary Least Squares (OLS) Regression?

How is it Estimated?

What are the Assumptions of OLS Regression?

How Do We Assess the Plausibility of the Assumptions?

How Do We Test for Outliers and What Do We Do with Them?

How Do You Assess Model Fit and Compare Nested Models?

How Do You Compare Regression Coefficients?

## **M o d e l B u i l d i n g**

---

What is Omitted Variable Bias?

How Do You Build a Regression Model?

**R**EGRESSION MODELING IS THE FUNDAMENTAL building block for a vast array of statistical methods, yet these methods often appear to be distinct to the novice social science researcher. As a class of analytic methods, regression in some variant is the most widely used statistical method and the focus of much of this volume. Even methods that appear unrelated to regression modeling, such as *t*-tests and ANOVA, can be represented as regression models and share a common underlying mathematical estimation method. As such, it behooves you as a researcher to become well versed in the methods and uses of this statistical framework. Once you have mastered the foundations of regression modeling, learning new variants, including those not covered in this book, is fairly straightforward.

In its simplest form, a linear regression model determines the amount of predicted change in one variable, the dependent variable, associated with a change in another variable, the independent variable. For example, a regression model could be used to examine the relationship between a youth's impulsiveness and level of delinquency or the relationship between a youth's gender and his or her level of delinquency. More complex models include multiple independent variables and examine the unique contribution of each to the single dependent variable. This is called a **multiple regression** model.<sup>1</sup> Using the prior example, a multiple regression model could examine the unique contribution of impulsiveness

---

<sup>1</sup>There are many authors who label these models as multivariate. This seems intuitive given that the models include multiple variables and multivariate seems like a portmanteau of multiple and variable. However, in statistics, multivariate models are formally defined as having multiple dependent variables or *variates*, the latter being an unknown random variable. In a multiple regression context, only the dependent variable is assumed to be a random variate. Examples of multivariate models include multivariate analysis of variance, factor analysis, structural equation modeling, and canonical correlation. It is also possible to have a multivariate multiple regression model that has both multiple dependent and independent variables. These are beyond the scope of this book and are less frequently used in criminological research.

to predicting delinquency, controlling for gender, and the unique contribution of gender in predicting delinquency, controlling for impulsiveness. It is not uncommon to see regression models in the criminology literature that have up to 15 or more independent variables in a single model. Given a sufficiently large sample size, the number of independent variables can be large. With the emerging availability of big data, the opportunities for exploring highly complex models have increased.

The uses of regression can generally be categorized as relating to prediction or theory testing. As will be explored in more depth later, these two purposes have implications for how one goes about building a regression model and how the model is judged. A correctly specified model for the purpose of prediction may not be an adequate model for the purpose of theory testing and vice versa. The former is judged on its ability to accurately predict the dependent variable with a minimum number of independent variables, particularly as tested on a new set of data or subset of the original dataset aside for that purpose. A model built for theory testing is judged on whether there is a third variable (or variables) omitted from the model that might explain the theoretical relationship of interest. The latter is generally more difficult than the former. In the absence of experimentation (discussed in Chap. 8), we can rarely be certain that a regression model uniquely identifies the causal relationship of interest. As the well-known dictum states, correlation is not causation. That said, we are often striving with our theoretical models to establish the plausibility of drawing a causal inference for one or more of the variables in our model. Our theories posit causal explanations, and we use regression to test the plausibility of these assertions. Incomplete models lead to biased inferences.

In this chapter, we explore the most basic form of a multiple regression model: ordinary least squares (OLS) regression. This model tests for a linear relationship between the dependent variable and a linear combination of the independent variables. The term *least squares* in the name reflects that the mathematics of this method minimizes the sum of the squared difference between the observed and predicted values.

## Overview of Simple Regression

---

We are assuming that you have a basic familiarity with simple linear regression, but we will review the basics of it here as it serves as the building block for multiple regression. Suppose, for example, that a simple

linear regression is defined in which the number of years in prison is identified as influencing the number of arrests after prison:

$$y_{\text{rearrests}} = b_0 + b_1(\text{years in prison}) + e.$$

**Equation 2.1**

This equation is stating that we believe that the number of rearrests, our dependent variable, is a linear function of the number of years in prison (our independent variable). The term of interest is the slope or  $b_1$  (also called a regression coefficient). This term reflects how much change in the dependent variable we expect with each one-unit change in the independent variable. Stated with respect to this particular model, the slope reflects how much of an increase or decrease we would predict in the number of rearrests for every 1-year increase that a person spends in prison. The first term on the right-hand side of this equation,  $b_0$ , is the intercept (also called the constant) and reflects the predicted value for the dependent variable when the independent variable equals zero. In the data we will use below for this model, all individuals have at least 1 year in prison. Thus, the intercept is beyond the range of our observed data and represents an extrapolation. This is common and in most situations the intercept is of no interest.

In a simple linear regression model, the slope and the intercept can easily be computed using Eqs. (2.2) and (2.3), shown below.

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

**Equation 2.2**

$$b_0 = \bar{y} - b_1 \bar{x}$$

**Equation 2.3**

Using Eqs. (2.2) and (2.3), we estimate this relationship from the data presented in Table 2.1 and find that the regression coefficient ( $b_1$ ) based on this model is 1.709. That is, every additional year of imprisonment produces about a 1.709 increase in our prediction of the number of subsequent arrests.

**Table 2.1**

Calculations for the regression coefficient of years of imprisonment and number of subsequent arrests

AGE			ARRESTS		
$x_i$ (1)	$x_i - \bar{x}$ (2)	$(x_i - \bar{x})^2$ (3)	$y_i$ (4)	$y_i - \bar{y}$ (5)	$(x_i - \bar{x})(y_i - \bar{y})$ (6)
1	-2.15	4.62	1	-2.1	4.52
2	-1.15	1.32	0	-3.1	3.57
2	-1.15	1.32	1	-2.1	2.42
2	-1.15	1.32	2	-1.1	1.27
2	-1.15	1.32	2	-1.1	1.27
3	-0.15	0.02	0	-3.1	0.47
3	-0.15	0.02	1	-2.1	0.32
3	-0.15	0.02	1	-2.1	0.32
3	-0.15	0.02	3	-0.1	0.02
3	-0.15	0.02	3	-0.1	0.02
3	-0.15	0.02	3	-0.1	0.02
3	-0.15	0.02	4	0.9	-0.14
4	0.85	0.72	2	-1.1	-0.94
4	0.85	0.72	4	0.9	0.77
4	0.85	0.72	4	0.9	0.77
4	0.85	0.72	4	0.9	0.77
4	0.85	0.72	5	1.9	1.62
4	0.85	0.72	6	2.9	2.47
4	0.85	0.72	9	5.9	5.02
5	1.85	3.42	7	3.9	7.22
			$\sum = 18.55$		
				$\sum = 31.7$	
			$\bar{x} = 3.15$		
			$\bar{y} = 31.7$		

### Working It Out

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$b_1 = \frac{31.7}{18.55}$$

$$b_1 = 1.7089$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_0 = 31.7 - 1.7089(3.15)$$

$$b_0 = -2.28$$

The intercept and slope create a regression line, showing the linear relationship between years in prison and the number of rearrests. A scatterplot showing each observation and the regression line is shown in

**Figure 2.1**

*Scatterplot and regression line showing the relationship between the number of years in prison and the number of rearrests based on the data in Table 2.1.*

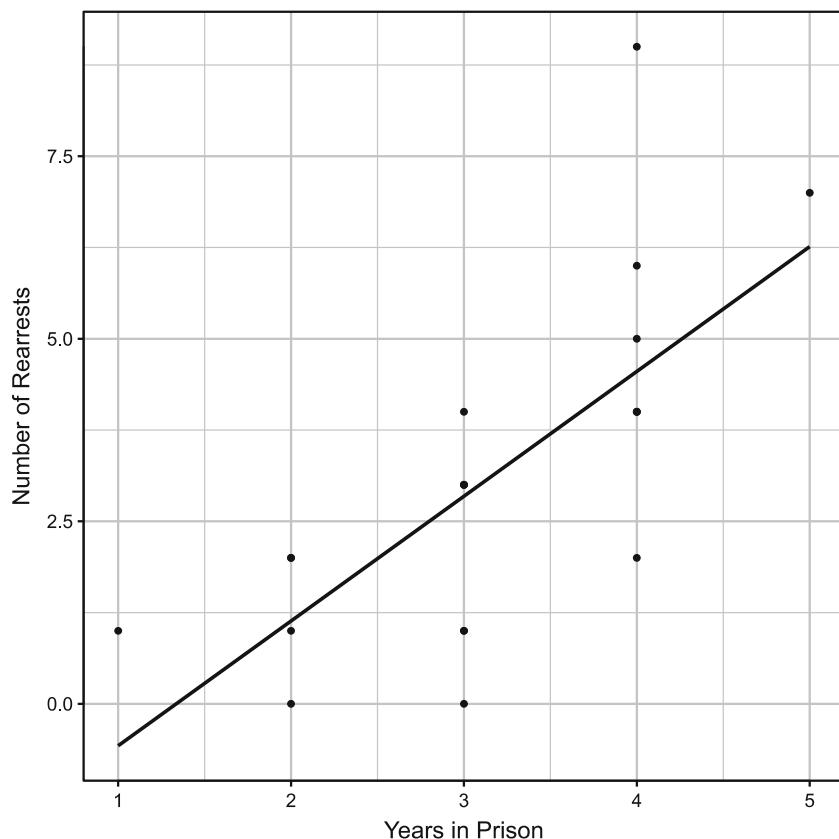


Fig. 2.1. Clearly, the number of years in prison is strongly related to the number of rearrests. Additionally, we can use the regression line to predict the number of rearrests. Visual inspection of Fig. 2.1 indicates that for 5 years in prison (the highest value in our data), we would predict 6.25 rearrests. For 3 years in prison, we would predict a bit over 2.5 rearrests. The lowest value for the number of years in prison in our data is 1 year, and for this value, we would predict less than 0 rearrests. This is clearly an impossible prediction—it is not possible to have fewer than 0 rearrests. What went wrong? Technically, nothing. The problem is that linear regression and linear multiple regression assume that the dependent variable is continuous. Our dependent variable is a count variable that can only be whole numbers (e.g., 0, 1, and 2). Negative values are not logically possible. In this example, the negative value is close to zero for the lowest value

of our independent variable, and as such is not a serious problem. However, in Chap. 3, we will explore solutions to this problem such as transformation of the dependent variable, or in Chap. 6 using a variant of the regression model designed specifically for count-based dependent variables.

Rather than rely on estimating the predicted number of rearrests from the plotted regression line, we can use the regression coefficients explicitly to calculate these values by inserting the various years in prison of interest into the regression equation, as shown in the *Working it Out* box for 1, 3, and 5 years in prison, where  $\hat{y}$  is the predicted value of  $y$  for a specific value of  $x$ .

### Working It Out

$$\begin{aligned}\hat{y} &= b_0 + b_1x \\ -0.5711 &= -2.28 + 1.7089 \times 1 \\ 2.8467 &= -2.28 + 1.7089 \times 3 \\ 6.2645 &= -2.28 + 1.7089 \times 5\end{aligned}$$

This shows that we predict  $-0.5711$ ,  $2.8467$ , and  $6.2645$  rearrests for individuals with 1, 3, and 5 years in prison based on these data. The distance between each predicted value and the observed value is the prediction error or *residual*. The single observation with 5 years in prison had 7 rearrests. The predicted value was  $6.2645$ . In this case, the residual is  $7 - 6.2645 = 0.7355$ . The regression line obtained by Eqs. (2.2) and (2.3) is the line that minimizes the squared residuals. This is called the least squares property of regression.

As should be obvious from the plot, this is a strong positive relationship. We can express this as a correlation coefficient. This should be familiar to you from your course in introductory statistics, but the formula is as follows:

$$r = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\left[ \sum_{i=1}^n (y_i - \bar{y})^2 \right] \left[ \sum_{i=1}^n (x_i - \bar{x})^2 \right]}} \quad \text{Equation 2.4}$$

Applying this to these data produces an  $r$  value of 0.7156, which is a very strong relationship, particularly in the social sciences. We can test the statistical significance for  $b_0$  and  $b_1$  using the  $t$ -distribution, although we are rarely interested in the statistical significance of  $b_0$ . The null hypothesis in each case is that the coefficient equals zero. For  $b_1$ , the  $t$  is shown in Eq. (2.5).

$$t = \frac{b_1}{\sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n-2)}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \quad \text{Equation 2.5}$$

The degrees of freedom for this  $t$ -test is  $n - 2$ , where  $n$  is the sample size. Working this out for our running example produces a  $t$ -value of 4.346, which is statistically significant at  $p < .05$ . Thus, we can reject the null and conclude that these data support the hypothesis that years in prison predicts the number of rearrests. In the case of simple regression,  $t$  can also be computed from the correlation coefficient, as shown below:

$$t = r \sqrt{\frac{n-2}{1-r^2}} \quad \text{Equation 2.6}$$

Equations (2.5) and (2.6) are equivalent and will return the same value of  $t$ . Intuitively, this should make sense as linear regression and linear correlation are two different ways of representing the same relationship. It would be problematic if they produced conflicting tests of statistical significance.

### Working It Out

$$\begin{aligned} t &= r \sqrt{\frac{n-2}{1-r^2}} \\ &= 0.7156 \sqrt{\frac{20-2}{1-0.7156^2}} \\ &= 4.346 \end{aligned}$$

## Extending Simple Regression to Multiple Regression

---

The above model assumes that there are no additional variables that might explain why some individuals with longer prison terms might have more rearrests on release. This, of course, is a questionable assumption, because common sense tells us that there are certainly other factors that influence arrests. Some of those factors, in turn, may also be related to the number of years that an offender serves in prison. If this is true, that is, that relevant factors related to years of imprisonment have been omitted from the model, then the regression coefficient for years in prison may provide a very misleading estimate of the effect of imprisonment on arrests.

Judges, for example, are likely to impose longer prison sentences on offenders with more serious prior records. Using the sample data in Table 2.2, we can look at the correlations among these three variables (see Table 2.3). The number of prior arrests is strongly related ( $r = 0.63$ ) to the length of prison term served. Prior arrests are even more strongly related to subsequent arrests ( $r = 0.76$ ). This suggests, first of all, that prior record is a relevant factor that should be included in our model if it is to be correctly specified. But it also raises a very important concern: How do we know that our finding that *years in prison* increases reoffending is not

**Table 2.2**

Number of rearrests, years spent in prison, and number of prior arrests for 20 former inmates

SUBJECT	REARRESTS	YEARS IN PRISON	PRIOR ARRESTS
1	0	2	4
2	0	3	2
3	1	1	2
4	1	2	3
5	1	3	3
6	1	3	2
7	2	4	3
8	2	2	3
9	2	2	1
10	3	3	2
11	3	3	3
12	3	3	3
13	4	3	4
14	4	4	3
15	4	4	4
16	4	4	5
17	5	4	4
18	6	4	5
19	7	5	5
20	9	4	6
	$\bar{y} = 3.1$	$\bar{x}_1 = 3.15$	$\bar{x}_2 = 3.35$
	$s = 2.300$	$s = 0.9631$	$s = 1.2360$

**Table 2.3**

Correlation coefficients for the variables years in prison, prior arrests, and subsequent rearrests based on data from 20 former inmates

	YEARS IN PRISON	PRIOR ARRESTS
Prior arrests	0.6280	
Subsequent rearrests	0.7156	0.7616

simply a result of the fact that those who serve longer prison terms generally have more serious prior records of offending?

In an ideal world, our comparisons of the impact of imprisonment would be made with subjects who were otherwise similar. That is, we would want to be sure that the offenders with longer and shorter prison sentences were comparable on other characteristics, such as the seriousness of their prior records. In this case, there would be no relationship between prior arrests and length of imprisonment, and thus, we would not have to be concerned with the possibility that the effect of length of imprisonment actually reflects the influence of prior arrests on reoffending.

In criminology and criminal justice, this approach is taken in the development of randomized experiments (see Chap. 9).<sup>2</sup> A randomized study of the impact of length of imprisonment on reoffending would be one in which the researcher took a sample of offenders and assigned them to treatment and control conditions at random. For example, the researcher might define a sentence of 6 months as a control condition and a sentence of 1 year as an experimental condition. In this case, the researcher could examine the effects of longer versus a shorter prison sentence on rearrests without concern about the confounding influences of other variables. In Chap. 9, we focus more directly on the analysis of experimental data. But it is important to note here that random allocation of subjects to treatment and control conditions allows the researcher to assume that other traits, such as prior record, are randomly scattered across the treatment and control conditions and as such are uncorrelated to treatment assignment, at least in the population. Our problem in criminal justice is that it is often impractical to develop experimental research designs. For example, it is highly unlikely that judges would allow a researcher to randomly allocate lengthy prison sanctions. The same is true for many other research problems relating to crime and justice.

Fortunately for criminal justice researchers, a correctly specified regression model will take into account and control for relationships that exist

---

<sup>2</sup>For a discussion of experimental methods in criminal justice, see Babbie and Maxfield (1995). For a comparison of experimental and nonexperimental methods, see Weisburd et al. (2001a).

among the independent variables of theoretical interest and the dependent variable. So, for example, the inclusion of both length of imprisonment and prior arrests in one regression model will provide regression coefficients that more accurately reflect the specific impact of each variable, once the impact of the other has been taken into account. This is part of what makes multiple regression so useful. It allows users to statistically control for potential confounds that provide alternative explanations for theoretically hypothesized causal relationships. In the absence of experimentation, however, it is difficult to know whether all such variables have been included in the model and measured without error. For this reason, caution must be used in drawing causal inferences from multiple regression models.

Equation (2.7) describes the calculation of a multiple regression coefficient in the case of two independent variables ( $x_1$  and  $x_2$ ). Equation (2.7) computes the regression coefficient for  $x_1$ , and Eq. (2.8) computes the regression coefficient for  $x_2$ . The model can be described as follows:

$$y_i = b_0 + b_1x_{1i} + b_2x_{2i} + e_i$$

where  $y_i$  is subsequent arrests,  $x_i$  is years in prison, and  $x_2$  is prior arrests. The population model for this equation would be written as:

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \varepsilon$$

In the model for the sample, the  $e_i$  represents the residuals or prediction errors for each observation. In the population model, these are considered errors and we assume that the errors are normally distributed with a mean of zero. We will discuss this in more depth in the section on the assumptions of multiple regression.

In the case of two independent variables, it is fairly straightforward to compute the regression coefficients, including the two slopes and the single intercept. For models with three or more independent variables, the computations involve matrix algebra and it is best to leave these tedious and complex computations to a computer and a statistical software package, such as SPSS, Stata, SAS, or R.

In our simple example with two independent variables, we calculate the regression coefficient for  $b_1$  for years in prison as:

$$b_{x_1} = \left( \frac{r_{y,x_1} - (r_{y,x_2} r_{x_1,x_2})}{1 - r_{x_1,x_2}^2} \right) \left( \frac{s_y}{s_{x_1}} \right)$$
Equation 2.7

where  $r_{y,x_1}$  is the correlation between rearrests and years in prison,  $r_{y,x_2}$  is the correlation between rearrests and prior arrests,  $r_{x_1,x_2}$  is the correlation between years in prior and prior arrests,  $s_y$  is the standard deviation of  $y$ , and  $s_{x_1}$  is the standard deviation for years in prison. These correlations are shown in Table 2.3, and the standard deviations are in Table 2.2. Working this out produces a  $b_1$  of 0.9358.

### Working It Out

$$\begin{aligned} b_{x_1} &= \left( \frac{r_{y,x_1} - (r_{y,x_2} r_{x_1,x_2})}{1 - r_{x_1,x_2}^2} \right) \left( \frac{s_y}{s_{x_1}} \right) \\ &= \left( \frac{0.7156 - (0.7616)(0.6280)}{1 - 0.6280^2} \right) \left( \frac{2.300}{0.9631} \right) \\ &= 0.9358 \end{aligned}$$

To find  $b_2$ , we simply switch  $x_1$  and  $x_2$  in Eq. (2.7), as shown in Eq. (2.8). Working this out for prior arrests produces a regression coefficient of 0.9593.

$$b_{x_2} = \left( \frac{r_{y,x_2} - (r_{y,x_1} r_{x_1,x_2})}{1 - r_{x_1,x_2}^2} \right) \left( \frac{s_y}{s_{x_2}} \right)$$

**Equation 2.8**

### Working It Out

$$\begin{aligned} b_{x_2} &= \left( \frac{r_{y,x_2} - (r_{y,x_1} r_{x_1,x_2})}{1 - r_{x_1,x_2}^2} \right) \left( \frac{s_y}{s_{x_2}} \right) \\ &= \left( \frac{0.7616 - (0.7156)(0.6280)}{1 - 0.6280^2} \right) \left( \frac{2.300}{1.236} \right) \\ &= 0.9593 \end{aligned}$$

What is most important to note in Eq. (2.7) is that the numerator (in the first part) takes into account the product of the relationship between prior arrests and subsequent rearrests and that of prior arrests and years in prison. This relationship is subtracted from the simple correlation between years in prison and subsequent arrests. In this way, multiple regression provides an estimate of the slope,  $b$ , that takes into account that some of the impact of years in prison may be due to the fact that longer prison terms are associated with more serious prior records. This estimate is now purged of the bias that was introduced when prior record was not included in the regression model. The multiple regression coefficient for years in prison when prior record is included in the regression model (0.936) is considerably smaller than the estimate calculated earlier in the simple regression model (1.709). Similarly, the value of  $b$  when we take into account years in prison (0.96) is much smaller than that in the single variable case (1.4).

The fact that the results are different when we examine the effects of years in prison and prior arrests in the multiple regression model shows that the simple regression coefficient was indeed biased, at least as estimates of the unique effect of each on rearrests. In both cases, the estimate of the effect of  $b$  provided by the simple regression coefficient was much too high. These differences also reflect a difference in interpretation between the multiple regression coefficient and the simple regression coefficient. In the simple case, the regression coefficient represents the estimated change in  $y$  that is produced by a one-unit change in  $x$ . In the multiple regression case,  $b$ , represents the estimated change in  $y$  associated with a one-unit change in  $x$  when all other independent variables in the model are held constant. Holding prior arrests constant leads to a reduction in the impact of years in prison. Holding years in prison constant leads to a reduction in the estimate of the effect of prior arrests. These differences may be seen as the bias introduced by not correctly specifying the regression model through the exclusion of prior arrests.

## Assumptions of Multiple Regression

---

The assumptions of multiple linear regression are fundamentally the same as for simple linear regression, just extended to multiple independent variables. Briefly, the four main assumptions are that the observations are independent, that the errors are normally distributed with a mean of zero, that the variance of the errors is equal across levels of the independent variables (i.e., homoscedasticity of variance), and linearity. We state the assumptions below and then return later to some methods for assessing those assumptions.

### Independence

Multiple regression assumes that each observation is independent of all other observations. Independence means that the values of the dependent and independent variables for one observation, such as an individual, are not influenced by the values of these variables for others in the sample. More technically, we are assuming that the errors are independent. The distinction between errors and residuals in the regression model is the distinction between population values versus sample values. Recall that in inferential statistics, our assumptions are about what is true in the population. Thus, we are assuming that in the population from which this sample was drawn, the errors are independent. A method to ensure independence is simple random sampling from a known population. Thus, some textbooks state the independence assumption as the assumption of random sampling. However, throughout the social sciences, we regularly make use of convenience samples and apply regression methods to the data obtained from such samples. So long as the assumption of independence is reasonable, such modeling is also reasonable, albeit with limited or unknown generalizability.

There are numerous possible sources of nonindependence. Two common sources in criminal justice are repeated measurements and clustered data. For example, if we take repeated measures on our units of analysis, be they persons, organizations, or geographic areas, etc., prior observations are almost certainly related to latter observations. A person's score on impulsivity at time one is likely to be related to their score on impulsivity taken at time two. This is a clear source of dependence in the data. Criminal justice data are also often clustered, that is, have a nested or hierarchical nature. A simple example from education is measuring student performance. Students within the same classroom are likely to be at least slightly more similar to one another than to the general student population. Thus, there is a dependence of students nested within classrooms. More relevant to criminology are neighborhood effects. Individuals within the same neighborhood likely share at least a slight dependence on their levels of fear of crime. Fortunately, statistical methods exist for handling these types of dependencies such as multilevel regression models discussed in Chap. 7.

We have slightly oversimplified the issue of independence and random sampling as it is a thorny issue with legitimate debates among statisticians. Even when we have a random sample from a clearly defined population, we are often interested in drawing inferences beyond that population. For example, we might have a random sample of prison inmates from a specific prison on a specific date. However, we are likely to be interested in drawing an inference to *prison inmates* including persons who have yet to commit a crime. Some scholars call this a *superpopulation* or a hypothetical population from which the specific population (or sample, in the case of a convenience sample) is representative (Gelman and Hill 2007; Hartley

and Sielken 1975). Another common way to conceptualize this is by reference to the underlying *data-generating process* from which our data emerged (Greene 2018). For example, when we flip coins, we are not actually sampling from a population of actual heads or tails but rather using a data generating mechanism of coin flipping. We are assuming that our sample is a random expression of this underlying process even if it represents a convenience sample obtained for the purpose of a given study. However, Berk and Freedman (2003) are highly critical of these frameworks for justifying the use of regression modeling, calling these *imaginary populations* that lead to imaginary inferences. For our purposes, so long as it is reasonable to assume that the data collection methods and research design yielded independent observations, then this assumption of regression modeling is satisfied. However, any generalizations of the results must be made in light of the sampling method used to obtain the data.

### Normally Distributed Errors

A common oversimplification of the assumption that the errors are normally distributed is to state that the dependent variable is normally distributed in the population. If this assumption is true, then the errors will also be normally distributed. However, there are many situations where the dependent variable might not be normally distributed, but the errors are. For example, imagine a characteristic that has a large average difference between men and women. This difference will result in a bimodal distribution for the dependent variable. However, a model that regresses this dependent variable onto the sex of each person will have an error distribution for which the bimodal effect created by sex has been removed, producing a unimodal distribution of errors. Thus, the critical assumption is that after taking into account the independent variables, the error distribution for the dependent variable is normally distributed in the population. We typically assess the plausibility of this assumption by examining the regression residuals but keep in mind that the assumption is about the population errors. Your sample might not look normally distributed, particularly if it has a small sample size, even if it was sampled from a normally distributed population.

The second aspect of the assumption of normally distributed errors is that the mean of the distribution of errors in the population is zero. There is no way to establish the plausibility of this assumption from the data. One method of ensuring this is through random assignment to conditions, such as the use of an experimental design. Other methods include instrumental variable analysis and regression discontinuity design. More generally, this issue is about the possibility that our model is not correctly specified and that there is an omitted variable or variables that might explain the observed regression coefficient for a specific independent variable. In quasi-experimental research designs, we call this selection bias or the possibility

that the errors might be correlated with the treatment condition. This problem goes by several names, including confounding, omitted variable bias, a lack of exogeneity in the independent variables, and the problem of correlated errors. These are all different ways of expressing a failure to satisfy this assumption.

The case of years of prison and prior arrests provides an illustration of this violation of the regression assumptions. When we leave prior arrests out of the regression equation, the variance it would have explained in rearrests becomes part of the error term. Because prior arrests and sentence length are correlated, sentence length is now correlated with the error term. This is a violation of our assumption and leads to a biased estimate of years of prison, as we showed above.

A related violation of the assumption occurs when we incorrectly identify the causal direction of our model. Let us say that we were examining the relationship between number of police and crime. In your regression model, you might place number of police as an independent variable in your model predicting levels of crime. While this is a plausible assumption, it is also the case that crime might impact the number of police in a city. If this is true, a simple regression model with cross-sectional data would violate this assumption as well. This is why it is very important to make sure that you have correctly identified the causal mechanisms in your data.

It is important to emphasize that the assumptions we make about errors in a regression model relate to errors in the population distribution. Often, such violations of assumptions will not be visible in the sample distribution. Indeed, it is useful to keep in mind that the computation of the regression line requires that the residuals (the prediction errors in the sample) add up to zero.

An implication of the assumption of normally distributed errors is that the dependent variable is assumed to be measured on an unbounded continuous scale that is at the interval- or ratio-level of measurement. A true normal distribution represents continuous data and extends from negative to positive infinity. Thus, the values of a normal distribution can be any real number even if in practice most all of the distribution falls within a limited range. In practice, almost all of our measures are discrete, at least to some level of precision, even if the underlying construct that is being measured is continuous. Fortunately, given the *central limit theorem*, we can safely relax this assumption if we have a sufficiently large sample size and enough discrete categories. For example, it is common to use ordinary least squares regression with dependent variables with as few as 5–7 discrete categories. So long as the data are roughly normally distributed without a lot of cases at either extreme, OLS regression performs reasonably well. However, if the data do not meet these criteria, then it is best to use a regression approach designed for the nature of the dependent variables, such as logistic regression (see Chap. 4), multinomial regression (see Chap. 5), or Poisson/negative binomial regression (see Chap. 6).

### Homoscedasticity of Errors

The assumption of **homoscedasticity** states that we assume that the variance of the errors is equal for all combinations of the independent variables. Recall from your introductory statistics course that we assume equal variances between the two samples when computing a *t*-test and between each of the samples when computing a one-way ANOVA *F*-test. With a single independent variable that represents a small number of discrete categories, such as two in the case of the *t*-test, this assumption is easy to understand: We are assuming that the variance in the population for one sample (or group) is equal to the variance in the population for the other sample. Notice that the variance within each sample around its mean is the error variance, as it does not include the variance between the two means. If we were to predict the score for an observation in the first sample, we would predict the mean for that sample. With a continuous independent variable, the assumption is extended to all possible values of the independent variable; that is, we assume that in the population the error variance is equal across all levels of the independent variable. With multiple independent variables, we extend this assumption to reflect all possible combinations of the independent variables. This sounds more complicated than it is. The predicted values from a regression model represent all of these combinations that exist in a particular sample. Thus, the assumption is that the error variance is equal across all predicted values from the regression model. As we will explore in the section on diagnostics, we can assess the plausibility of this assumption by examining the distribution of residuals across predictive values.

### Linearity

The linearity assumption states that the dependent variable is a linear function of the additive effects of the independent variables. In the case of a single independent variable, this assumption is simply that the relationship between the independent and dependent variable is linear. With multiple independent variables, if each has a linear relationship with the dependent variable, then any additive composite will also have a linear relationship. However, it is not essential that each independent variable has a linear relationship with the dependent variable, only that the predicted values of the dependent variable are linearly related to the observed values of the dependent variable. This assumption is fairly easy to assess, and violations of this assumption can be dealt with either through a transformation of the dependent variable, a transformation of one or more of the independent variables, or by adding additional quadratic terms to the model. This topic will be discussed in Chap. 3 in the section on nonlinear relationships.

## Measurement Error in the Independent Variables

---

Regression modeling can handle nominal-, ordinal-, interval-, and ratio-level measures as independent variables, although nominal and ordinal measures require special considerations as will be explored below. Regression modeling does, however, assume that the independent variables are measured without error. This is clearly an unrealistic assumption as almost all measures in the social science have at least some level of measurement error (as do most in the physical sciences as well). Regression modeling assumes measurement error in the dependent variable but not in the independent variables. What is the implication of this for our modeling?

The short answer is that any measurement error in our independent variables will attenuate the observed regression coefficient toward the null value (toward 0). The more the measurement error, the larger the attenuation. When an independent variable is included in a model as a statistical control, that is, to deal with selection bias or omitted variable bias, then the statistical control will be imperfect (Fox 2016). For independent variables of theoretical interest, the result of this will be an underestimate of the true strength of the relationship. Furthermore, any comparison between regression coefficients (such as in the standardized form, discussed below), will be problematic if the amount of measurement error is meaningfully different between the variables. One independent variable might appear to be a better predictor of the dependent variable simply because it is measured more accurately than another variable.

A related assumption of regression is that the levels of the independent variables are fixed. Because we are assuming that they are measured without error, we are also assuming that the levels of these variables represented in our data are the only levels to which we wish to generalize. This should be intuitive. Any generalization about the relationship between the independent and dependent variable to values not in the data represents an extrapolation and in statistics extrapolation is generally risky. There are models, such as mixed-effects models discussed in Chap. 7, that allow the researcher to assume that the levels of one or more of the independent variables represent a random sampling of levels from a population of possible values. However, at this point, it is important to recognize that OLS regression is assuming that we are only interested in drawing generalizations from the findings across the observed values of the independent variables.

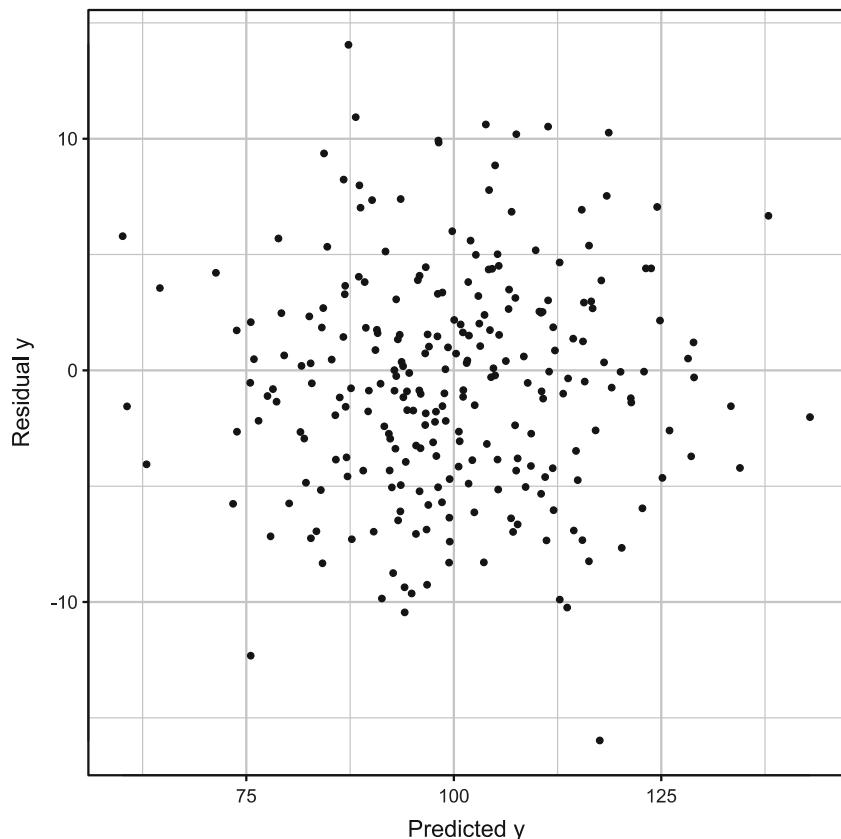
## Regression Diagnostics

We can gain information about the credibility of several of the assumptions of multiple regression through regression diagnostics. One of the most useful diagnostic tools is the scatterplot of the model residuals on the  $y$ -axis and the predicted values of  $y$  on the  $x$ -axis. This single plot provides us with information about the linearity assumption and homoscedasticity assumption. It also can reveal outliers, a topic that will be addressed in the next section.

Figure 2.2 shows a scatterplot for residuals against predicted values for a model that satisfies the assumption of linearity and homoscedasticity. These data were generated as a random draw of 250 cases from 2 uncorrelated

Figure 2.2

*Example of a scatterplot of residual against predicted values from a regression model with  $n = 250$  showing homoscedasticity*



normal distributions. What we are looking for with respect to the homoscedasticity assumption is a roughly rectangular or square shape to the scatter. With homoscedasticity, we are assuming that the variance in the residuals is equal across all levels of the predicted values for  $y$ . That translates into a roughly equal vertical spread in the residuals as you move left to right across the scatterplot. Because there are typically fewer cases at the high and low ends of the predicted values, the scatter may be harder to evaluate at the extremes and the shape may appear to have rounded corners, as is the case with Fig. 2.2. The main density of this scatterplot, ignoring the small number of more extreme cases, is roughly rectangular with somewhat rounded corners. This figure is consistent with the assumption of homoscedasticity. With large sample sizes, the ideal shape becomes easier to identify. Conversely, with small sample sizes, this assumption is difficult to assess.

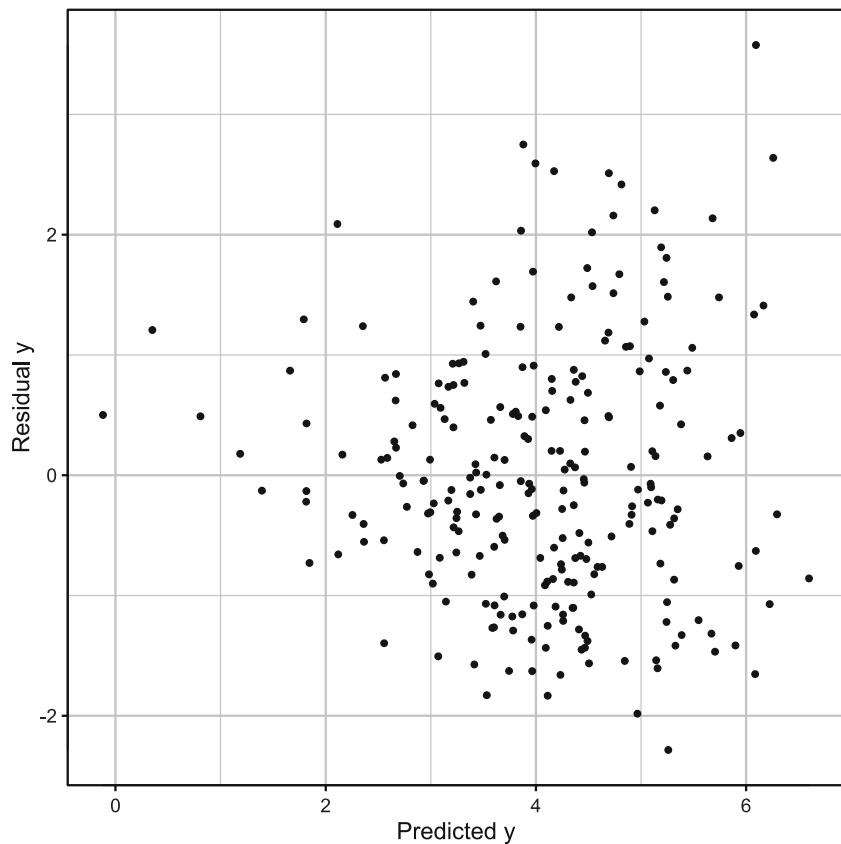
In contrast, Fig. 2.3 shows a clear example of heteroscedasticity. The scatter grows larger as we move left to right. This fan shape is a common form of heteroscedasticity with the variance increasing as the predicted values increase. We typically see this pattern when the dependent variable represents a count or other positively valued variables that grow multiplicatively rather than additively. The solution to this problem is either to transform the dependent variable (discussed in Chap. 3) or use a different regression modeling method, such as Poisson or negative binomial regression (discussed in Chap. 6). That said, OLS regression is robust to mild levels of heteroscedasticity so long as you have a sufficiently large sample size (e.g., greater than 100 cases).

In assessing the linearity assumption, we want to see no curvature in the scatterplot. The residuals should be roughly centered around the mean value of 0 on the  $y$ -axis for all predicted values of  $y$ . A clear curved shape to the residuals, such as seen in Fig. 2.4, indicates that one or more of the independent variables has a curvilinear relationship with the dependent variable. Adjusting the regression model to accommodate curvilinear relationships is discussed in Chap. 3 and as with heteroscedasticity involves either transforming the dependent variable and/or one or more of the independent variables or including quadratic terms (e.g., squared independent variables) in the regression model. The linearity assumption is a strong assumption of OLS regression. Violating it will result in a model that does a poor job of predicting the dependent variable. It will also generate an inaccurate portrait of the impacts of specific independent variables.

In multiple regression, we are also assuming that the errors are normally distributed in the population. The normality of the distribution of residuals provides information about the credibility of this assumption. Both the histogram and the quantile plot (often called the Q–Q plot or quantile–quantile plot) are useful for assessing this assumption. Figure 2.5 shows the histogram for the same residuals shown in Fig. 2.4. We can see that the

**Figure 2.3**

*Example of a scatterplot of residual against predicted values from a regression model with n = 250 showing heteroscedasticity*

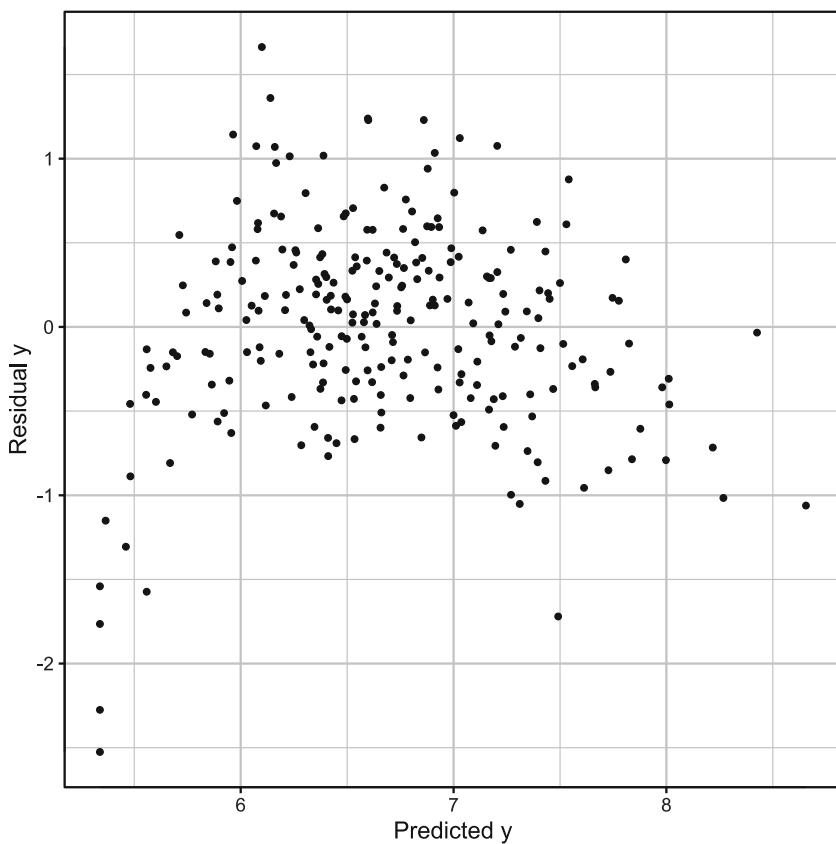


residuals have a roughly normal shape but are negatively skewed, that is, have a longer tail on the left side than we would expect from a truly normal distribution.

The quantile plot tells the same story and is shown in Fig. 2.6. This plot compares the theoretical quantiles to the quantiles in our sample data. Essentially, this is showing whether the percentage of observations below any given  $z$  value is consistent with the percentage of observations below each  $z$  value in our sample data. In a Q-Q plot, if the residuals are perfectly normal, then they will fall along the diagonal reference line. In Fig. 2.6, we see that the lower values of residuals are substantially below the line. This reflects the negative skew shown in Fig. 2.5. The rest of the distribution is consistent with the normal distribution. Keep in mind that we do not expect sample data to land perfectly along this diagonal line. Sampling error results

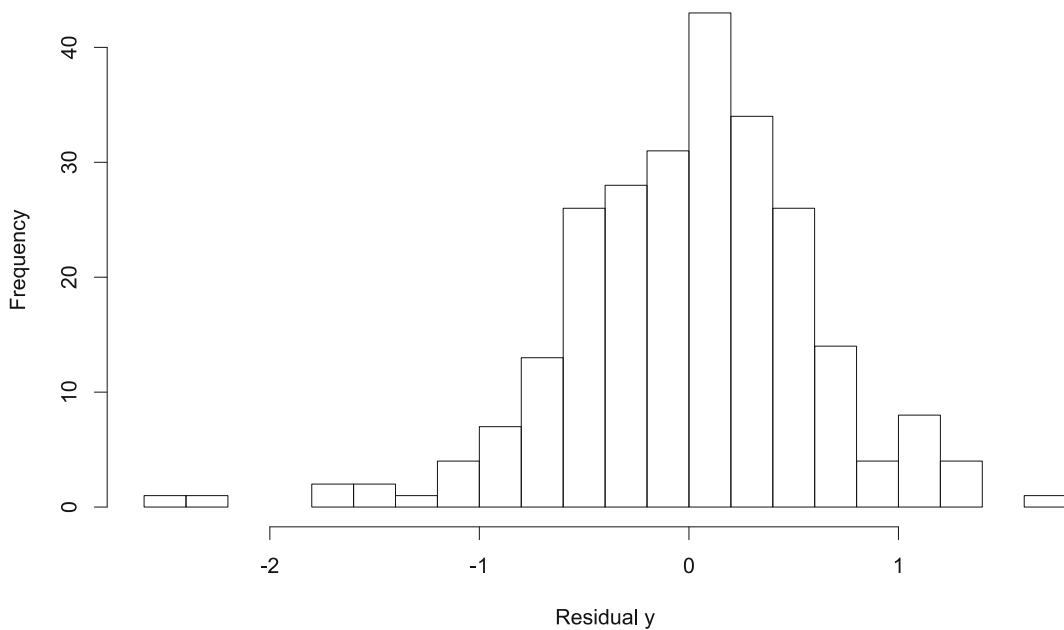
**Figure 2.4**

Example of a scatterplot of residual against predicted values from a regression model with  $n = 250$  showing a curvilinear relationship



in departures from normality even if the sample is drawn from a population that is normally distributed. Thus, we are only concerned with large departures from normality. Furthermore, the *central limit theorem* establishes that OLS regression is robust to modest departures from normality so long as you have a large sample size. The smaller your sample, the more concerned you must be about nonnormally distributed errors. Serious departures from normality may necessitate the transformation of one or more variables in the model (discussed in Chap. 3) or the use of an alternative modeling method, such as those discussed in Chaps. 4–6.

The assumption of independence is generally asserted based on knowledge regarding how the data were generated and/or collected. Simple random sampling from a known population ensures that this assumption is satisfied. With convenience samples, the assumption that each

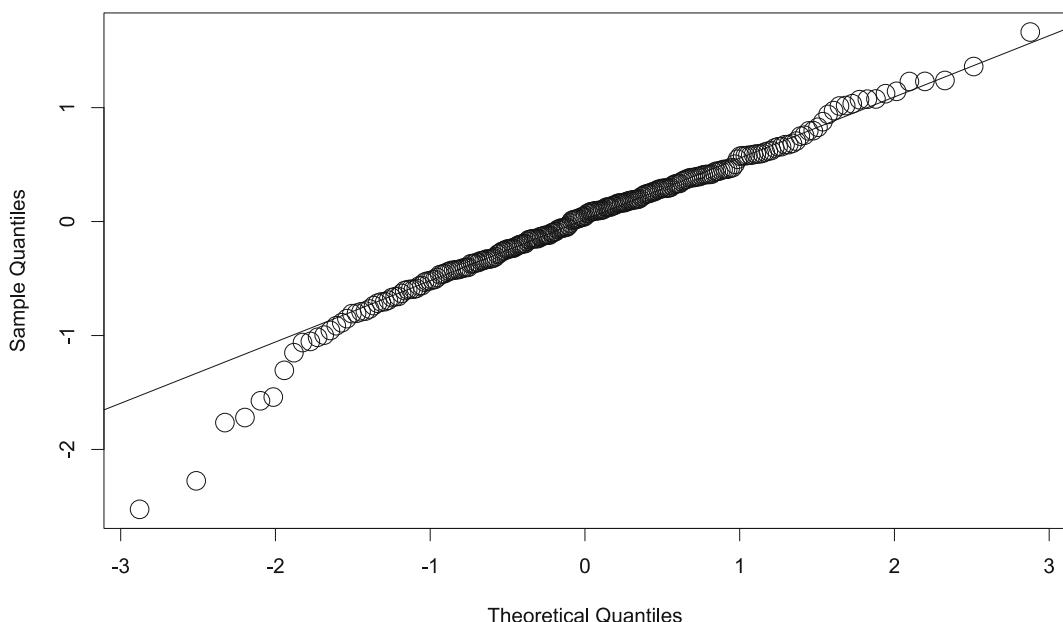
**Figure 2.5***Histogram of the residuals from Fig. 2.4*

observation is independent of all other observations must be thought through carefully based on an understanding of the research design that produced the data. If there are any potential sources of nonindependence, then mixed-effects or multilevel regression modeling methods should be employed (see Chap. 7). Ignoring nonindependence can lead to serious bias in the standard errors, generally resulting in overstating statistical significance (i.e.,  $p$ -values that are too small). Unlike the normality and homoscedasticity assumptions, OLS regression is not robust to departures from independence.

A diagnostic test that is sometimes advocated for assessing independence is plotting the residuals against the unique identifier for each observation. This can be useful but only in limited cases. For this to have diagnostic value, the identifier must reflect some natural sequencing in the data, such as the order in which the observations were made. If the identifiers are simply randomly assigned numbers, then this diagnostic is uninformative. Essentially, this is assuming that observations with adjacent identifiers may be nonindependent. Plotting the residuals against the identifier may reveal patterns supporting this concern, such as clumps of cases above and below the reference line. That said, this only assesses one

**Figure 2.6**

*Quantile–Quantile plot of the residuals from Fig. 2.4*



potential form of nonindependence and assumes that the nonindependence is related to the assigned identifier.

A model that satisfies the above assumptions may still have outliers that have an undue influence on the regression model. Testing for and addressing outliers is critical and discussed in the next section.

## Dealing with Outliers and Influential Cases

In multiple regression, an observation may be an outlier on a single variable or it may be unusual in the combination of values across the variables. Our definition of an outlier is simply an extreme observation that is clearly different than the other observations in the data. More formal definitions of outliers, such as any case that is more than some number of standard deviations above or below the mean on a given measure, tend to perform poorly as a method of detecting which observations may be truly problematic. What is critical is the unusualness of an observation given the data and even more importantly the influence such a case has on the regression model.

There are numerous methods for identifying outliers and influential cases in multiple regression. We will focus on two specific methods, the studentized residual and Cook's  $D$ . The studentized residual provides a measure of how poorly a regression model fits a particular observation and is scaled as a  $t$ -value. More importantly, the residual for each observation is based on a regression model with that observation removed. Outliers pull the regression function toward them, reducing their residual. As such, a better index of whether an observation is an outlier is to compute the residual for each observation based on a regression model that excludes that particular observation. Thus, if we have 250 observations, we (well, the computer) must compute the regression model 250 times in order to compute each of the 250 studentized residuals. These residuals are then standardized by dividing by the standard error of the residuals, producing a  $t$ -test for each residual. The number of studentized residuals equals the sample size, and we would expect some of these to be statistically significant at a conventional significance level (e.g., .05) just by chance, particularly as sample size increases. As such, any significance testing on the studentized residuals should use a Bonferroni correction (Beckman and Cook 1983).

In our example with a sample size of 250, the critical value for a statistically significant  $t$ -test at the two-tailed  $p = .05$  is the  $t$ -value associated with a  $p$  of  $.025/250$  or  $p = .0001$ . Stated in words, we have divided the significance level of .05 by 2 to reflect a two-tailed test and again by 250 to make the Bonferroni adjustment for multiple tests. The degrees of freedom is the sample size minus the number of independent variables minus 2. Given two independent variables and a sample size of 250, this would be 246 degrees of freedom. The critical  $t$  associated with this  $p$ -value and degrees of freedom is 3.78. Thus, any studentized residual from this example model that is greater than 3.78 or less than  $-3.78$  is likely to be an outlier. All modern software programs can easily compute the studentized residuals. A visual inspection of these values is often sufficient to identify outliers although determining which are statistically significant can add strength to this identification.

Just because an observation is an outlier does not mean that it has much influence on the regression model. In a model with a large sample size, a single observation is unlikely to have much influence. With more modest sample sizes, however, it is useful to examine the influence that each observation has on the model. One measure of this is Cook's  $D$ . Like the studentized residual, Cook's  $D$  requires successively dropping each observation from the regression model. However, rather than focusing on the residuals, Cook's  $D$  examines the difference in the regression coefficients, the  $bs$ , between the model with and without each observation. An observation that has a Cook's  $D$  that is large relative to the other observations

exerts a relatively high influence on the regression coefficients. Thus, we would evaluate it as a potential outlier.

There are other methods of identifying potential outliers, but these two approaches will suffice in most situations. Once you have identified a potential outlier, how should you proceed? The first step is to assess whether the outlier reflects a data entry error or otherwise invalid observation. Data entry errors should of course be fixed. If the data represent an invalid value, it should also be dropped. However, if the outlier is a legitimate but extreme case, then it should not simply be removed from the analysis and ignored as if it did not exist. With a large sample size, the outlier (or outliers) is unlikely to alter the substantive conclusions drawn from the model. You can check this by running your model with and without the outliers. If the outlier meaningfully affects the results, then it is reasonable to present two models, one with and one without any outliers. In some cases, researchers will transform the values of the outliers so that they are closer to the rest of the distribution of cases. Again, if you carry out this solution, you should present the results in a transparent way letting the reader of your work know exactly what you have done and how it affects your findings. Running your models using these different alternative solutions is a type of **sensitivity analysis**, since it allows you to assess how sensitive your analysis is to the outliers. More generally, sensitivity analyses are useful in assessing the stability of regression to changes in specifications of the model.

## Testing the Significance of Individual Regression Coefficients

---

In multiple regression, the statistical significance of an individual regression coefficient is assessed using the  $t$ -test, just as it was for testing the significance of the slope from a simple regression model. The null hypothesis for these  $t$ -tests is that the coefficient equals zero. We compute  $t$  as the ratio of the coefficient to its standard error, as shown in Eq. (2.9) below.

$$t = \frac{b}{se_b}$$

**Equation 2.9**

The numerator is simply the regression coefficient of interest. The denominator is the standard error specific to that regression coefficient. Unfortunately, this is no longer the denominator shown in Eq. (2.5) but is rather more complicated and its calculation is best left to a computer as it involves matrix algebra, at least when there are more than two independent variables. The degrees of freedom for each  $t$  in a specific regression model

is  $n - k$ , where  $n$  is the sample size and  $k$  is the number of independent variables. Most software programs for computing multiple regression include the  $t$ -test and the standard errors as part of the default output, along with the associated significance level.

Related to testing for statistical significance is confidence intervals. Rather than simply providing a binary test of whether a coefficient is statistically significant or not, the confidence interval provides useful information about the precision of an estimate. A narrow confidence interval is a more precise estimate of the population parameter than a wide confidence interval. You can also determine whether a coefficient is statistically significant by examining whether the confidence interval includes the null value of 0, at least at the level of significance associated with the confidence interval (i.e., a 95% confidence interval is associated with a 5% significance level). If assumptions of the regression model are satisfied, in the long run, 95 out of 100 confidence intervals at the 5% level of significance will include the true population value. Thus, we are 95% confident that our confidence interval includes the true population value. In this sense, the confidence interval helps us not only identify whether a coefficient is statistically significant but also provides a range within which the true population value might credibly be found.

The confidence interval for OLS regression coefficients is calculated in the typical fashion by adding and subtracting from the coefficient the value of the standard error ( $se$ ) times the  $t$  critical value ( $t_{CV}$ ), as shown in Eq. (2.10).

$$\text{Lower 95\% } b = b - se(t_{CV})$$

$$\text{Upper 95\% } b = b + se(t_{CV})$$

**Equation 2.10**

Both  $t$ -tests and confidence intervals are illustrated in the worked example below.

## Assessing Overall Model Fit and Comparing Nested Models

---

Much of the focus when using multiple regression for theory testing is on the statistical significance of the focal independent variables that were hypothesized to bring about change in the dependent variable, holding the other variables constant. However, there is often value in assessing the overall fit of the regression model as well. This is particularly useful when using multiple regression for prediction and forecasting but can also be useful in a theory-testing context (Weisbord and Piquero 2008). Additionally, model fit indices can be used to compare the fit of competing models.

**$R^2$  and Adjusted  $R^2$** 

A commonly used descriptive index of the fit of a regression model is  $R^2$ . This index reflects the percent of variance explained by the model. Thus, an  $R^2$  of 0.20 indicates that the regression model explains 20% of the variability in the dependent variable. Recall from introductory statistics that the variance of a measure is the average squared deviation of the observations around the mean, as shown in Eq. (2.11).<sup>3</sup>

$$\sigma^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}$$
**Equation 2.11**

The numerator of this equation is called the total sum-of-squares (*TSS*). If we were asked to guess someone's score on  $y$  without additional information, such as that provided by other variables, our best guess would be the mean of  $y$ . This would minimize our prediction errors given the least squares characteristic of the mean. Thus, this is our baseline variance before making use of other variables. We can partition this variability into two pieces: the portion explained by the regression model and the residual portion remaining unexplained. These can be expressed mathematically as Eqs. (2.12) and (2.13).

$$\sigma_{model}^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{n}$$
**Equation 2.12**

$$\sigma_{residual}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}$$
**Equation 2.13**

The ratio of the model variance (or the explained sum-of-squared) to the total variance produces  $R^2$ . Because each of these equations has the same denominator, we can drop the denominator and express  $R^2$  as the ratio of the numerators, or sums of squares, as shown in Eq. (2.14).

---

<sup>3</sup>Note that we are using  $n$  and not  $n - 1$  in the denominator here to illustrate that this is an average. However, with sample data, we would always use  $n - 1$  in the denominator when computing the variance.

$$R^2 = \frac{SS_{model}}{SS_{total}} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$
**Equation 2.14**

As a proportion,  $R^2$  can only take on the values between 0 and 1, inclusively. Ideally, we want  $R^2$  to be as close to 1 as possible. However, much more modest values are typical in the social sciences, and very large values can be suspect. Keep in mind that all, or at least most, of our variables have some degree of measurement error. That is, there is some amount of random variation in our measures that cannot meaningfully be explained or predicted. Thus, we would not expect  $R^2$  to ever exceed the nonmeasurement error proportion of variability in our dependent measure. Furthermore, the dependent variables of interest within criminology, such as delinquency or fear of crime, are multiply determined and unlikely to be fully explained by any single model. This complicates the interpretation of  $R^2$ . There are no absolute values for it that are considered good or poor, per se. Context matters and this index is most useful in assessing the predictive value of a model. In short, caution should be used in interpreting these values.

Another important characteristic of  $R^2$  is that it can only increase with the inclusion of additional variables in the model. The increase may be trivial, but adding another variable to the model can never reduce  $R^2$ . Thus, focusing too much on maximizing  $R^2$  can lead to over-fitting a regression model.  $R^2$  and the validity of the regression model more generally are sensitive to including too many variables relative to the sample size. Although there is no hard rule regarding the ratio of the number of independent variables relative to the sample, anything less than 1/10 (10 observations for each independent variable) runs the risk of over-fitting the data. A model that is over-fit is likely to encounter estimation problems, leading to unstable estimates of the independent variables. The adjusted  $R^2$  was designed to help address this problem.

The adjusted (or shrunken)  $R^2$  penalizes models with a large number of independent variables relative to the sample size. A large difference between  $R^2$  and the adjusted- $R^2$  is suggestive of a model that includes too many independent variables relative to the sample size. Eq. (2.15) shows the computation of the adjusted  $R^2$ .

$$\text{Adjusted } R^2 = 1 - (1 - R^2) \frac{n - 1}{n - k - 1}$$
**Equation 2.15**

Notice that the adjustment is a function of the sample size ( $n$ ) and the number of independent variables in the model ( $k$ ). Applying this to a study with a sample size of 200 and an  $R^2 = 0.25$  produces the following adjusted  $R^2$  values for 5, 10, 20, and 40, independent variables: 0.23, 0.21, 0.17, 0.06. We can see that the adjustment is slight for 5 and 10 independent variables relative to a sample size of 200, but this adjustment becomes larger for 20 independent variables and very large for 40 independent variables. Forty independent variables are simply too many with a sample size of 200. Because the main value of the adjusted  $R^2$  is as an index of how much of a penalty is being applied for the complexity of the model, it should always be reported along with the  $R^2$  value.

Both  $R^2$  and the adjusted- $R^2$  are descriptive indices of the overall fit of the model. We can explicitly test the statistical significance of the model using the analysis of variance (ANOVA) and associated  $F$ -test. The null hypothesis for this test can be stated in three different ways: (1) The population  $R^2$  equals zero; (2) all  $b$ s, other than the intercept, equal zero; (3) the additive linear composites of independent variables are uncorrelated with the dependent variable. These are each a different way of stating the same thing: The regression model is no better than a null model at predicting the dependent variable. Not surprisingly, the  $F$ -test is based on the sums of squares used in computing  $R^2$ . Equation (2.16) shows the computation of  $F$ .

$$F = \frac{\frac{SS_{model}/df_{model}}{SS_{residual}/df_{residual}}}{\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 / (k-1)}{\sum_{i=1}^n (y_i - \hat{y})^2 / (n-k)}} \quad \text{Equation 2.16}$$

where the  $df_{model}$  is the number of independent variable ( $k$ ) minus 1 and the  $df_{residual}$  is the number of observations ( $n$ ) minus the number of independent variables. The critical value for  $F$  is based on these two degrees of freedom. Modern statistical software programs will report the exact  $p$ -value for the observed  $F$ . This value is the test of whether the model as a whole is statistically significant. That is, can we reject the null that this model in the population has no explanatory power. In practice, it is rare that the overall test of the regression model is not statistically significant. It does not tell you whether your model is the correct model or whether the results are important or meaningful.

$F$  can also be computed from  $R^2$ , as shown in Eq. (2.17).

$$F = \frac{R^2(n - k - 1)}{(1 - R^2)k}, \quad \text{Equation 2.17}$$

where  $k$  is the number of independent variables, and  $n$  is the sample size.

Another use for the  $F$ -test in multiple OLS regression modeling is testing the joint significance of a subset of variables. This may arise in two ways. The first is related to testing the overall effect of a nominal independent variable. Nominal variables can be used as independent variables, but if they have more than two categories they must be represented as a set of dummy or indicator variables (see Chap. 3). There will always be one less dummy variable than the number of categories for the nominal variable. For example, if we have a nominal variable indicating the race/ethnicity of the persons in our sample with five categories, we need four independent dummy variables to encode this information in a usable way for multiple regression. The standard output would list the regression coefficient and the associated significance test for each dummy variable. However, we are likely to want to know if race/ethnicity is a significant contributor to our model. We can test this with an  $F$ -test of the difference between the model with the race/ethnicity variables and the model without. This test is shown as Eq. (2.18).

$$F_{\Delta} = \frac{SS_{\text{residual}}(\text{Restricted}) - SS_{\text{residual}}(\text{Full})}{MS_{\text{residual}}(\text{Full})} \quad \text{Equation 2.18}$$

where the numerator is the difference between the sums of squares for the residuals ( $SS_{\text{residuals}}$ ) for the restricted model, that is, the model without the race/ethnicity variables, and the sums of squares for the full model that includes these variables. This difference reflects the sums of squares explained by the set of variables that differentiate these two models. The mean-square error ( $MS_{\text{residual}}$ ) for the full model is the sums of squares of the residuals divided by  $n - k$ , where  $n$  and  $k$  are the sample size and the number of independent variables, respectively. The numerator degrees of freedom and denominator degrees of freedom for this  $F$ -value are the difference in the number of independent variables in the two models and  $n - k$  for the full model. This  $F$  comparing two models can also be computed using the  $R^2$  for each as shown in Eq. (2.19).

$$F = \frac{(R_{\text{Full}}^2 - R_{\text{Restricted}}^2)/(k_{\text{Full}} - k_{\text{Restricted}})}{(1 - R_{\text{Full}}^2)/(n - k_{\text{Full}} - 1)} \quad \text{Equation 2.19}$$

This  $F$  test can be extended to compare any two models where one is nested within the other. A nested model has a subset of the independent variables of the full model and no unique independent variables. We can think of this as a way of testing a *block* or a subset of independent variables in a regression model. This can be useful to assess whether a block of theoretically relevant variables contributes meaningfully to the prediction of the dependent variable, after controlling for the initial set of variables included in the model, rather than simply focusing on whether individual variables are statistically significant.

## Comparing Regression Coefficients Within a Single Model: The Standardized Regression Coefficient

---

A multiple regression model allows us to specify the impact of a specific independent variable while holding constant the impact of other independent variables. This is a very important advantage of multiple regression analysis over simple regression analysis. However, when we include multiple variables in the same model, it is natural to want to compare the impact of the different variables examined. For example, in our case, does years in prison have a stronger effect on subsequent rearrests than number of prior arrests? Or does number of prior arrests have a stronger effect than years in prison? The ordinary regression coefficient  $b$  does not allow us to answer this question, since it reports the effect of a variable in its original units of measurement. Accordingly, the regression coefficient for years in prison reports the predicted change in subsequent rearrests for each year change in years in prison. The regression coefficient for number of prior arrests reports the predicted change in subsequent rearrests for each change in number of prior arrests. Though the interpretation of the regression coefficients in these cases is straightforward, we cannot directly compare them.

Another statistic, called the **standardized regression coefficient** or **Beta**, allows us to make direct comparisons. Beta weights take the regression coefficients in an equation and standardize them according to the ratio of the standard deviation of the variable examined to the standard deviation of the dependent variable. Beta is expressed mathematically in Eq. (2.20):

$$\text{Beta} = b \frac{s_x}{s_y} \quad \text{Equation 2.20}$$

The interpretation of the standardized coefficient is similar to that of  $b$  (the unstandardized coefficient), except that we change the units. We interpret Beta as the expected amount of change in standard deviation units

of the dependent variable, given a one standard deviation change in the independent variable.

For years in prison in our example, we take the regression coefficient of 0.9358 and multiply it by the ratio of the standard deviation of years in prison (0.9631) and subsequent rearrests (2.3000). The result is 0.3919, which tells us that an increase of one standard deviation in years in prison is expected to result in an increase of 0.392 of a standard deviation in rearrests.

### Working It Out

$$\begin{aligned}\text{Beta} &= b \frac{s_x}{s_y} \\ &= 0.9358 \left( \frac{0.9631}{2.3000} \right) \\ &= 0.3919\end{aligned}$$

For prior arrests, we begin with our regression coefficient of 0.9593. Again, we standardize our estimate by taking the ratio of the standard deviation of prior arrests (1.2360) and subsequent rearrests (2.3000). Our estimate of Beta here is 0.5155, which indicates that an increase of one standard deviation in prior arrests is expected to result in an increase of 0.516 of a standard deviation in rearrests.

### Working It Out

$$\begin{aligned}\text{Beta} &= b \frac{s_x}{s_y} \\ &= 0.9593 \left( \frac{1.2360}{2.3000} \right) \\ &= 0.5155\end{aligned}$$

In our example, the Beta weight for prior arrests is larger than that for years in prison. According to this estimate, the number of prior arrests has a

greater impact on subsequent rearrests than does the number of years in prison. The standardized regression coefficient thus provides us with an answer to our original question regarding which of the independent variables examined have the most influence on the dependent variable. As you can see, the standardized regression coefficient is a useful tool for comparing the effects of variables measured differently within a single regression model. However, because standardized regression coefficients are based on the standard deviations of observed samples, they are generally considered inappropriate for making comparisons across samples.

## Correctly Specifying the Regression Model

---

A **correctly specified model** is one where the regression coefficient for the independent variable or subset of variables that are of theoretical interest is unbiased. We are focusing here on the independent variables that are part of our theoretically derived hypotheses and that we believe bring about change in the dependent variable. There are likely other variables in the model that have been included to reduce bias in the regression coefficient(s) of interest. We often conceptualize these as control variables, although the math of regression makes no such distinction.

As noted earlier, a common way to frame this issue is in terms of assumptions related to the error term in regression. It is assumed not only that the errors in the regression are stochastic (i.e., random), but also that there is no specific systematic relationship between the error term and the independent variables included in the regression. If there is such a relationship, the regression coefficient will be **biased**. This is simply a restatement of the omitted variable bias problem and reflects that there is something else that causally affects both the independent and dependent variables.

We will use our model predicting rearrest as a substantive example. We saw that if we estimated the regression coefficient for years in prison without taking into account prior arrests, the regression coefficient would be biased—in this case, overestimated. What happens in theory to the error term in this case? As we discussed earlier in the chapter, when we exclude an independent variable, the effect of that variable moves to the error term. In our case, the population model including both independent variables may be stated as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where  $y$  = subsequent rearrests,  $x_1$  = years in prison, and  $x_2$  = prior arrests. When we take into account only one independent variable, the model includes only the term  $x_1$ :

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

In the latter model, number of prior arrests is included by implication in the error term. But what does this mean regarding the relationship in this model between the error term and years in prison? Our sample data suggest that the number of prior arrests is related to years in prison (as was shown in Table 2.3). By implication, since number of prior arrests is now found in the error term, the error term can now be assumed to be related to years in prison as well. Accordingly, if we leave prior arrests out of our equation, then we violate the assumption that there is no systematic relationship between the error term and the independent variables in the equation. Stated differently, the mean of the errors will not be zero but rather upwardly or downwardly biased.

By looking at bias in terms of the error term, we can also specify when excluding an independent variable will not lead to bias in our estimates of the regression coefficients of other variables. If the excluded variable is unrelated to other variables included in the regression, it will not cause bias in estimates of  $b$  for those specific variables. This is the case because when there is no systematic relationship between the excluded variable and the included variable of interest, its exclusion does not lead to a systematic relationship between the error term and the variable of interest.

For example, if years in prison and prior arrests were not systematically related (e.g., the correlation between the variables was 0), it would not matter whether we took into account prior arrests in estimating the regression coefficient for years in prison.<sup>4</sup> In this case, the exclusion of prior arrests would not lead to a systematic relationship between the error term and years in prison, because there is no systematic relationship between prior arrests and years in prison even if years in prison has a causal effect on the number of rearrests. If our goal is prediction, and not just estimating the effect of years in prison on number of prior arrests, then omitting prior arrests would clearly reduce our predictive accuracy given that it has a strong relationship with rearrests, even if its omission did not bias the causal estimate for years in prison.

In criminology and criminal justice research, we can seldom say with assurance that the models we develop include all relevant predictors of the

---

<sup>4</sup>It is important to note that bias can be caused by a nonlinear relationship between the excluded and the included variable. The assumption is that there is no systematic relationship of any form.

dependent variables examined. The problem is often that our theories are not powerful enough to clearly define the factors that influence criminal justice questions (Weisburd and Piquero 2008). Criminal justice is still a young science, and our theories for explaining crime and justice issues often are not well specified. This fact has important implications for the use of criminal justice research in developing public policy. When our predictions are weak, they do not form a solid basis on which to inform criminal justice policies.<sup>5</sup>

An implication of our failure to develop strongly predictive models in criminal justice is that our estimates of variable effects likely include some degree of bias. We have stressed in this chapter the importance of controlling for relevant predictors in regression modeling. The cost of leaving out important joint causes of the dependent variable and an independent variable of interest is that estimates of a variable's effects may include a spurious component. This fact should make you cautious in reporting regression analyses and critical in evaluating the research of others. Just because regression coefficients are reported to the fifth decimal place on a computer printout, it does not mean that the estimates so obtained are solid ones.

The fact that regression models often include some degree of misspecification when based on observational data (i.e., nonexperimental), however, should not lead you to conclude that the regression approach is not useful for criminology and criminal justice researchers. As in any science, the task is to continue to build on the knowledge that is presently available. The researcher's task in developing regression models is to improve on models that were developed before. With each improvement, the results we gain provide a more solid basis for making decisions about theory and policy. This, of course, makes the practical task of defining the correct model for the problem you are examining extremely important.

## Model Specification and Building

---

How should you begin building a model? Importantly, model specification does not begin with your data. Rather, it starts with theory and familiarizing yourself with the research literature on this topic. To build a regression model, you should first identify what is already known about the dependent variable you have chosen to study. If your interest, for example, is in the factors that influence involvement in criminality, you will need to

---

<sup>5</sup>Mark Moore of Harvard University has argued, for example, that legal and ethical dilemmas make it difficult to base criminal justice policies about crime control on models that still include a substantial degree of statistical error (see Moore 1986).

carefully research what others have said and found regarding the causes of criminality. Your regression model should take into account the main theories and perspectives that have been raised by others.

If you do not take prior research and theory into account, then those reviewing your work will argue that your predictions and your estimates of variable effects are biased in one way or another. Just as the exclusion of prior record from our example led to a misleading estimate of its impact on length of imprisonment, so too the exclusion of relevant causal factors in other models may lead to bias. The only way to refute this potential criticism is to include such variables in your regression model.

Taking into account the theories and perspectives of others is the first step in building a correctly specified regression model. However, in most research, we seek to add something new to existing knowledge. In regression modeling, this usually involves the addition of new variables or new ways of measuring old variables. Sometimes, such new variables are drawn from an innovative change in theory. Other times, they involve improvements in measurement. Often, the finding that these new or transformed variables have an independent impact above and beyond those of variables traditionally examined by researchers leads to important advances in criminology or criminal justice policy.

The implication of the above is that you would specific the model you wish to test based on your theorizing prior to examining the data. Common practice often differs from the ideal, with researchers testing numerous variations of a regression model, dropping and including various combinations of independent variables. This leads to over-fitting of the data and tempt a researcher to select and justify the model that is most consistent with the theoretical propositions they are testing and not the empirical reality under study (it is surprisingly easy to convince yourself that this model really is the most credible model). This is a form of *p*-hacking and is a likely contributor to the replicability crisis in the social sciences.

Ideally, we would specify a model, test it, and live with the consequences for better or worse (Harrell 2015). This would preserve the fundamental logic inherent in null hypothesis significance testing. Once we start tweaking our model, we run the risk of inflated Type I errors. However, real-world data is often messy, and our initial attempt to model a dependent variable often reveals statistical problems that should not be ignored, such as violations to assumptions, multicollinearity, or problematic outliers. Thus, some amount of learning from the data and allowing it to determine how to analyze it seems reasonable and often necessary.

The first protection against this is carefully theorizing as discussed above. The second is thinking through the most appropriate modeling method based on the nature of the dependent variable, such as whether to use OLS regression and logistic regression. Additionally, it is important to think through any complexities related to the independent variables (e.g.,

nonlinearities), and any multicollinearities that might exist among them. However, this will not always ensure that everything will proceed smoothly.

What level of modification or tweaking is acceptable? We will attempt to answer this by first discussing what is not acceptable: including or excluding variables based on statistical significance. If there was a reason to include an independent variable in the model, then it should stay in the model even if it is not statistically significant. Removing it is to assume that its true value in the population equals zero exactly. Unless the confidence interval is very tight around the null value, this is an unjustified assumption and one that will likely lead to bias in the remaining coefficients.

Modifying the initial model is acceptable to address clear violations of assumptions that are likely to bias the model. This may include a transformation of the dependent variable to address heteroscedasticity or nonlinearity. Nonlinearity may also be addressed, depending on the situation, by adding additional terms to the model, such as the square of one of the independent variables (see Chap. 3). In the next chapter, we introduce the problem of multicollinearity which occurs when independent variables in the model become so correlated that the estimation of standard errors in the model become inflated. Either dropping one or more of the offending variables or creating a composite or factor score for the multicollinear variables is a reasonable modification. Identifying and potentially removing an outlier may also be necessary although this should be done with extreme caution and results with and without the outliers reported in any presentation of the results. With large datasets, an outlier is unlikely to have much real effect on the model, as shown in the worked example below.

The above only applies to regression modeling for the purpose of testing theoretically derived hypotheses. If the purpose of the modeling is prediction, such as the development of a risk prediction index or predicting which neighborhoods are likely to see an increase in crime in the coming year, then model tweaking is not only acceptable but an explicit part of the method. The goal is to identify as small of a set of independent variables as possible that will accurately predict the dependent variable. The focus is not on the statistical significance of any given coefficient but rather on the performance of the model as a whole. Furthermore, there is no presumption that any of the independent variables is causally related to the outcome, although that may well be the case. As noted by John Fox (2016), in predictive models, we are not concerned with omitted variable bias or even the causal direction between the independent and dependent variables. For example, in predictive models for medical diagnostics, it is common to include variables that represent symptoms caused by the disease that the model is trying to predict. The bottom line is that model-building methods vary depending on the purpose, and as the researcher, you should be sensitive to these differences and use methods appropriate to your purpose.

## An Example of a Multiple Regression Model

---

To help illustrate the method of multiple regression, we will work through an example with real-world data. The data used are drawn from a national sample of police officers developed by the Police Foundation (Weisburd et al. 2001b). The dependent variable in this analysis is hours worked per week, which represents a ratio-level measure that is continuous but measured in units of whole hours. There are two independent variables. One, years with the department, is measured at the ratio-level. The second, level of education, is on an ordinal scale with eight levels, ranging from some high school to doctoral degree. Finally, we have the officer's gender. This is a binary nominal-level variable.

In regression analysis, a binary nominal-level independent variable is incorporated into the model by creating a **dummy variable**. This dummy variable assigns the values to the categories of the binary variable. We need to do this because multiple regression analysis does not recognize qualitative categories—variable must be numeric. By convention, we give one category a value of 0 and the other a value of 1, although any two numeric values would work. It is completely arbitrary which category you assign to 0 or 1, although it will affect the sign of the coefficient. In this sample, we have coded the dummy variable such that 1 equals female and 0 equals male.

Table 2.4 presents the descriptive statistics for these variables. Note that as a binary variable, the mean for women reflects the proportion of women in the sample because women was coded as 1. Table 2.5 presents the simple correlations among these variables. We can see that all three of these variables are correlated with the dependent variable, but these correlations are small. We have not reported the statistical significance of these correlations as that is not relevant at this stage. It is the regression model and not the individual correlations that we have made hypotheses about. We can also see that the independent variables are slightly intercorrelated.

The regression model with these three independent variables predicting hours worked is shown in Table 2.6. Because the independent variable education is an ordinal variable, how would we interpret the meaning of its

**Table 2.4**

Descriptive statistics for selected variables from the police officer abuse of authority study

VARIABLE	MEAN	SD	MIN	MAX
Hours per week worked	45.84	6.707	4	80
Educational level	3.80	1.269	1	8
Years with department	11.67	8.406	0	35
Female officer	0.08	0.275	0	1
Sample size = 923				

**Table 2.5**

Correlation coefficients among the selected variables

	HOURS WORKED PER WEEK	OFFICER EDUCATIONAL LEVEL	YEARS WITH DEPARTMENT
Officer educational level	0.084		
Years with department	-0.090	0.029	
Female officer (yes = 1, no = 0)	-0.058	0.023	-0.106

**Table 2.6**

Regression model predicting hours worked per week from years with department, female officer, and officer educational level.

	REGRESSION COEFFICIENT	STANDARD ERROR	-95%	+95%	t	p	STANDARDIZED BETA
(Intercept)	45.143	0.752	43.667	46.619	60.030	<.0001	
Years with department	-0.080	0.025	-0.130	-0.028	-3.036	0.003	-0.100
Officer educational level	0.466	0.173	0.126	0.805	2.697	0.007	0.088
Female officer (yes = 1, no = 0)	-1.726	0.801	-3.298	-0.154	-2.155	0.031	-0.071

Multiple R-squared: 0.020, Adjusted R-squared: 0.017, F(3, 919) = 6.389, p = 0.0003

effect? Here, what we have is a group of ordered categories. For the regression, this ordinal-level scale is treated simply as an interval-level scale. It is assumed that the differences across categories must be roughly similar in value or more simply that each one-unit increase in that scale is related in a **linear** manner to the dependent variable. Thus, our interpretation of this regression coefficient is that for every one-level increase in education level, there is, on average, a 0.466 increase in the number of hours worked (once we have taken into account years with the department and whether the officer is female or male). The dummy code for female has a straightforward interpretation: It is the adjusted mean difference in hours worked between men and women with women working roughly 1.7 fewer hours per week than men, adjusting for years with the department and educational level. For each 1-year increase in years with the department, an officer is predicted to work .08 fewer hours per week, adjusting for the other variables in the model.

In this case, the standardized regression coefficient is very useful for comparing the relative effects of these variables. The standardized regression coefficients (represented by Beta) show that the difference among these three variables is small, suggesting that each has roughly a similar effect on hours worked once taking into account the effect of the other variables.

One way to gain a better understanding of the interpretation of a dummy variable regression coefficient is to see how it affects our regression equation. Let us begin by writing out the regression equation for our example without the error term:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

where  $y$  is hours worked per week,  $x_1$  is years with the department,  $x_2$  is education level, and  $x_3$  is officer gender. As a second step, let us insert the coefficients gained in our regression analysis:

$$y = 45.15303 + (-0.07957)x_1 + (0.46594)x_2 + (-1.72620)x_3$$

or

$$y = 45.15303 - (0.07957)x_1 + (0.46594)x_2 - (1.72620)x_3$$

What happens if we try to write out the regression equations for men and women separately? For men, the regression equation is

$$y = 45.15303 - (0.07957)x_1 + (0.46594)x_2 - (1.72620)0$$

or

$$y = 45.15303 - (0.07957)x_1 + (0.46594)x_2$$

**Equation 2.21**

Because men are coded as 0, the third term of the equation falls out. But what about for women? The third term in the equation is a constant because all of the women have a value of 1. If we write it out, we have the following result:

$$y = 45.15303 - (0.07957)x_1 + (0.46594)x_2 - (1.72620)1$$

or

$$y = 45.15303 - (0.07957)x_1 + (0.46594)x_2 - 1.72620$$

We can simplify this formula even more because the two constants at the beginning and the end of the equation can be added together:

$$y = 43.41638 - (0.07957)x_1 + (0.46594)x_2$$

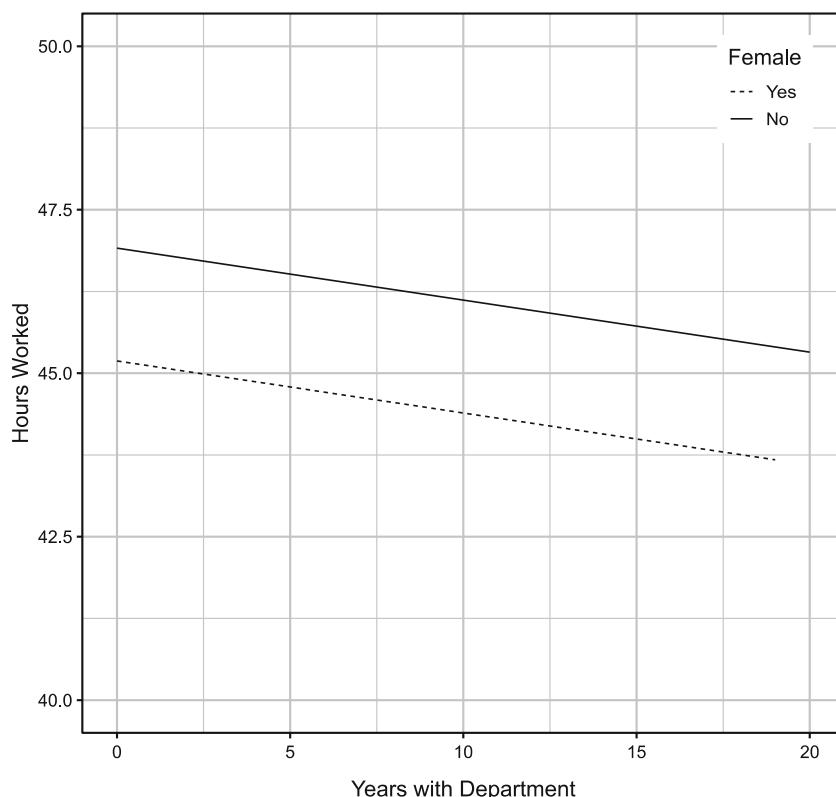
**Equation 2.22**

What then is the difference between the regression equations for men and women (i.e., Eq. (2.21) versus Eq. (2.22))? In both cases, the slope of the regression line for years with the department and education level is the same. The difference between the two equations is in the  $y$ -intercept, as illustrated in Fig. 2.7. This figure is showing only two of the three independent variables. The lines are fit with education level set at its mean value. As you can see, men and women have parallel regression lines (i.e., the model does not include an interaction effect that would allow the lines to be nonparallel). However, the women's line intersects the  $y$ -axis about 1.7 h lower than the men's line. This provides us with the interpretation of our coefficient. Women police officers, on average, work about 1.73 h a week less than men police officers, taking into account years with the department and educational level (Fig. 2.7).

What about the assumptions of multiple regression for this analysis? First off, we are assuming that the responses from each officer are independent of the other officers. Second, we can assess the plausibility of the assumptions of linearity and homoscedasticity by examining the scatterplot of the residuals on the  $y$ -axis against the predicted values on the  $x$ -axis. This is shown in Fig. 2.8. Ignoring the outliers at the moment, we can see a slight increase in the vertical spread as we move from left to right, suggesting mild heteroscedasticity. That is, the variance in the residuals appears to increase as the predicted hours worked increase. However, given that we have a large sample size ( $n = 923$ ), we can relax this assumption. Regression is robust to mild departures from this assumption when you have a large sample. In terms of the linearity assumption, there is no evidence to suggest that this assumption is not reasonable. There is no visible curvature to these residuals.

The striation in Fig. 2.8 (the five slightly sloping vertical lines in the scatter) is common when you have discrete values of the dependent variable, as we do here, and discrete values on one or more of the independent variables, such as our ordinal variable and binary nominal variable. This is not a concern in terms of the assumptions.

Another concern with regression is the possible influence of outliers. Visual inspection of this plot suggests two cases with large negative residuals relative to the other residuals and a few possible positive residuals. A histogram of the residuals, shown in Fig. 2.9, is slightly negatively skewed

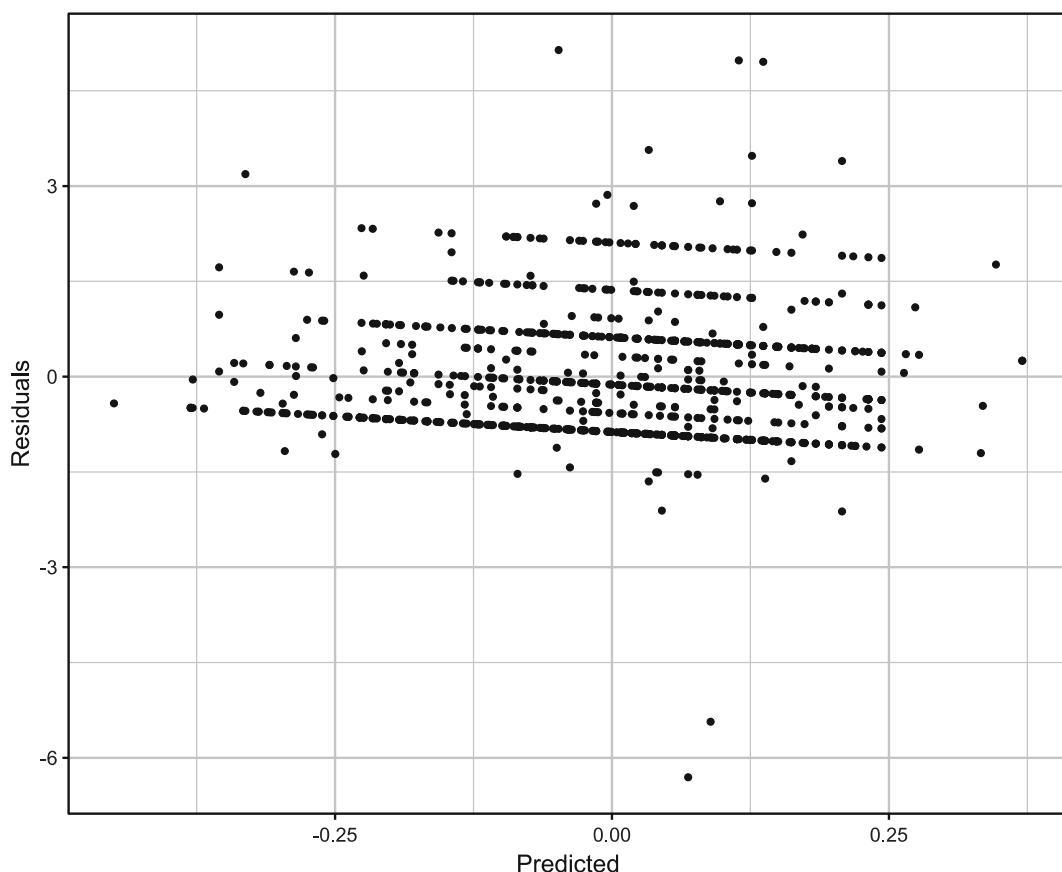
**Figure 2.7***Regression lines for female and male police officers*

but well within the realm that can be handled by the central limit theorem, particularly given the large sample size. Thus, the assumption of normally distributed errors seems reasonable.

A final concern is the potentially biasing effects of extreme outliers. Again, this is mitigated with a large sample size as a single observation is unlikely to have a large effect on the regression results. However, to assess whether this is true, we will start by examining the studentized residuals. We find 5 studentized residuals that exceed the  $t$  critical value of 4.22 with 919 degrees of freedom with a Bonferroni correction for 923 tests. Two of these are negative residuals, and three are positive residuals. These five values are identifiable on the scatterplot (Fig. 2.8). Rerunning the regression model without these five values shows that the model is essentially

**Figure 2.8**

*Scatterplot of residuals against predicted values for model shown in Table 2.6*

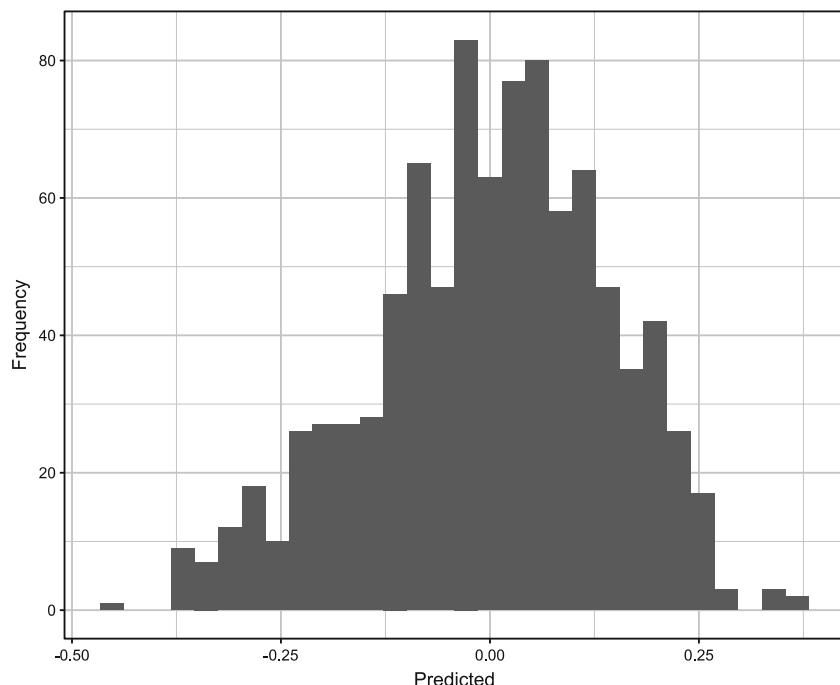


unchanged, as shown in Table 2.7. While there are differences, these are small and have no effect on the substantive conclusions drawn from this model. Thus, the assumptions of regression seem adequately met with these data.

It is also worth noting that although all three independent variables have statistically significant regression coefficients and the overall fit for the regression model is also statistically significant ( $F(3, 919) = 6.389, p = 0.0003$ ), the amount of variance explained in hours worked is small or roughly 2%. Thus, these variables only explain a small amount of the variation in officer hours worked.

**Figure 2.9**

*Histogram of residuals for model shown in Table 2.6*

**Table 2.7**

Regression model with 5 outliers removed predicting hours worked per week from years with department, female officer, and officer educational level

	REGRESSION COEFFICIENT	STANDARD ERROR	-95%	+95%	t	p
(Intercept)	45.025	0.693	43.665	46.385	64.984	<.0001
Years with department	-0.077	0.024	-0.124	-0.030	-3.200	0.001
Officer educational level	0.482	0.159	0.170	0.795	3.033	0.002
Female officer (yes = 1, no = 0)	-1.693	0.736	-3.138	-0.249	-2.301	0.022

Multiple R-squared: 0.024, Adjusted R-squared: 0.021, F(3, 914) = 6.389, p = 0.0003

## Chapter Summary

In a simple regression model, there is only one independent variable. In a **multiple regression** model, there are several independent variables. These independent variables may be multivalued variables ranging from a

binary or **dummy variable**, an ordinal discrete variable, or discrete or continuously scaled interval- and ratio-level variables. Such a model considers the effect of each independent variable on the dependent variable, while holding all the other variables constant. Thus, these models allow researchers to more accurately predict a dependent variable by using the unique information contained in numerous independent variables. Furthermore, it allows researcher to test for theoretically hypothesized relationships unconfounded by other known or measured variables.

In this chapter, we explored the most basic multiple regression model, OLS regression model. This statistical method tests for a linear relationship between the dependent variable and a linear additive combination of the independent variables. The output for such a model produces a linear function of the independent variables that minimizes the predictive errors or residuals. The model assumes that the observations are independent, that the errors are normally distributed, that the error variance is homogeneous, and that the relationship between the regression model and the dependent variable is linear.

When using regression to test hypothesized causal relationships, a major concern is whether the model is correctly specific. A **correctly specified model** will have an error term that is uncorrelated with the independent variables. Thus, there are no omitted variables that are causally related to both the independent variables and the dependent variable. If there are such variables, then the estimate of  $b$  for the included factor will also be **biased**. Randomized experiments, which scatter pre-existing traits at random, offer a solution to this problem, but they are sometimes impractical in criminal justice research. Outside of the context of a randomized experiment, it is often difficult to know if a model is unbiased. Hence, it is important to be cautious in drawing causal inferences from such regression models.

## Key Terms

---

**Biased** Describing a statistic when its estimate of a population parameter does not center on the true value. In regression analysis, the omission of relevant independent variables will lead to bias in the estimate of  $Y$ . When relevant independent variables are omitted and those measures are related to an independent variable included in regression analysis, then the estimate of the effect of that variable will also be biased.

**Correctly specified regression model** A regression model in which the researcher has taken into account all of the potential confounding variables that might account for the relationship between the dependent variable and independent variables of theoretical interest.

**Dummy variable** A binary nominal-level variable that is included in a multiple regression model.

**Homoscedasticity** An assumption of multiple regression. When this assumption is met, the error variance is equal across all combinations of the independent variables.

**Multiple regression** A technique for predicting change in a dependent variable, using more than one independent variable.

**Sensitivity analysis** The running of multiple regression models that allow the

researcher to see how different specifications (such as inclusion and exclusion of outliers) impact the model results.

### Standardized regression coefficient

**(Beta)** Weighted or standardized estimate of  $b$  that takes into account the standard deviation of the independent and the dependent variables. The standardized regression coefficient is used to compare the effects of independent variables measured on different scales in a multiple regression analysis.

## Symbols and Formulas

---

$k$	Number of independent variables in the regression model
$r_{y,x_1}$	Correlation coefficient for $y$ and $x_1$
$r_{y,x_2}$	Correlation coefficient for $y$ and $x_2$
$r_{x_1,x_2}$	Correlation coefficient for $x_1$ and $x_2$
$s_y$	Standard deviation for $y$
$s_x$	Standard deviation for $x$
$R^2$	Variance in the dependent variable explained by the regression model
$R_{fm}^2$	$R^2$ for the full regression model
$R_{rm}^2$	$R^2$ for the reduced regression model
$k_{fm}$	Number of independent variables in the full regression model
$k_{rm}$	Number of independent variables in the reduced regression model

To calculate the correlation coefficient between two variables:

$$r = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\left[ \sum_{i=1}^n (y_i - \bar{y})^2 \right] \left[ \sum_{i=1}^n (x_i - \bar{x})^2 \right]}}$$

To calculate a multiple regression coefficient for two independent variables:

$$b_{x_1} = \left( \frac{r_{y,x_1} - (r_{y,x_2} r_{x_1,x_2})}{1 - r_{x_1,x_2}^2} \right) \left( \frac{s_y}{s_{x_1}} \right)$$

and

$$b_{x_2} = \left( \frac{r_{y,x_2} - (r_{y,x_1} r_{x_1,x_2})}{1 - r_{x_1,x_2}^2} \right) \left( \frac{s_y}{s_{x_2}} \right)$$

A sample multiple regression model with three independent variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

To calculate  $R^2$ :

$$R^2 = \frac{SS_{model}}{SS_{total}} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

To calculate the adjusted  $R^2$ :

$$\text{Adjusted } R^2 = 1 - (1 - R^2) \frac{n - 1}{n - k - 1}$$

To calculate an  $F$ -test for the overall significance of a multiple regression model:

$$\begin{aligned} F &= \frac{SS_{model}/df_{model}}{SS_{residual}/df_{residual}} \\ &= \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 / (k - 1)}{\sum_{i=1}^n (y_i - \hat{y})^2 / (n - k)} \end{aligned}$$

To calculate an  $F$ -test on a subset of variables in a regression model:

$$F_{\Delta} = \frac{SS_{residual}(\text{Restricted}) - SS_{residual}(\text{Full})}{MS_{residual}(\text{Full})}$$

To calculate an  $F$ -test on a subset of variables in a regression model using  $R^2$ :

$$F = \frac{R^2(n - k - 1)}{(1 - R^2)k}$$

To calculate the standardized coefficient (Beta):

$$\text{Beta} = b \frac{s_x}{s_y}$$

To calculate an  $F$ -test on a subset of variables in a regression model:

$$F_{\Delta} = \frac{SS_{residual}(\text{Restricted}) - SS_{residual}(\text{Full})}{MS_{residual}(\text{Full})}$$

To calculate an  $F$ -test on a subset of variables in a regression model using  $R^2$ :

$$F = \frac{(R_{Full}^2 - R_{Restricted}^2)/(k_{Full} - k_{Restricted})}{(1 - R_{Full}^2)/(n - k_{Full} - 1)}$$

## Exercises

---

- 2.1. Consider the following regression model, which purports to predict the length of sentence given to individuals convicted of theft:

$$y = b_0 + b_1x + e$$

where  $y$  = length of sentence, and  $x$  = number of prior sentences.

- (a) List the variables you might wish to include in a more comprehensive model. Include a brief statement about why each additional variable should be included.
- (b) Present your model in equation form.
- 2.2. In an article in the newspaper, a researcher claims that low self-esteem is the cause of crime. Upon closer inspection of the results in the paper, you learn that the researcher has computed a bivariate model using self-reported theft as the dependent variable and self-esteem as the one independent variable.
- (a) List the variables you might wish to include in a more comprehensive model. Include a brief statement about why each additional variable should be included.
- (b) Present your model in equation form.
- 2.3. A researcher has built a multiple regression model to predict the effect of prior offenses and years of education on the length of sentence received by 100 convicted burglars. He feeds the data into a computer package and obtains the following printout:

Dependent Variable ( $y$ ): Length of Sentence (months)

Independent Variable ( $x_1$ ): Number of Prior Offenses

Independent Variable ( $x_2$ ): Years of Education

$F$  Sig. = 0.018

$R^2$  = 0.16

$x_1$ :  $b_1$  = 0.4

Sig.  $t$  = 0.023

$x_2$ :  $b_2$  = -0.3

Sig.  $t$  = 0.310

Evaluate the results, taking care to explain the meaning of each of the statistics produced by the computer.

- 2.4. An analysis of the predictors of physical violence at school produced the following results:

INDEPENDENT VARIABLE	<i>b</i>	BETA
Age (years)	0.21	0.05
Sex (female = 1, male = 0)	-3.78	0.07
Race (white = 1, nonwhite = 0)	-1.34	0.06
Number of Friends Arrested	1.96	0.33
Number of Times Attacked by Others	3.19	0.24
Number of Times Hit by Parents	2.05	0.27

Explain what each regression coefficient ( $b$ ) and standardized regression coefficient (Beta) means in plain English.

- 2.5. Danny has obtained data on the quantity drugs seized per month at a seaport over the course of 2 years. He wishes to explain variations in the quantity of drugs seized per month and runs a regression analysis to check the effect of his independent variable—the total number of customs officers on duty for each month—on the quantity of drugs seized. The resulting regression coefficient is +4.02. Danny is worried, however, that his model might not be correctly specified, and he decides to add another variable—the number of ships that arrive at the port each month. He calculates the correlations between the three pairs of variables, and the results are as follows:

$y$ (drugs seized),  $x_1$ (customs officers): +0.55

$y$ (drugs seized),  $x_2$ (ships arriving): +0.60

$x_1$ (customs officers),  $x_2$ (ships arriving): +0.80

The standard deviations for the three variables are 20 kg (quantity of drugs seized per month), 1.6 (number of customs officers on duty), and 22.5 (number of ships arriving).

- (a) Calculate the regression coefficient for customs officers.
- (b) Calculate the regression coefficient for ships arriving.
- (c) How do you account for the difference between your answer to part a and the regression coefficient of +4.02 that Danny obtained earlier?
- 2.6. A study of prison violence examined the effects of two independent variables—percent of inmates sentenced for a violent crime ( $x_1$ ) and average amount of space per inmate ( $x_2$ )—on the average number of violent acts per day ( $y$ ). All variables were measured for a random selection of cell blocks in three prisons. The researcher reported the following results:

$$r_{y,x_1} = 0.20$$

$$r_{y,x_2} = 0.20$$

$$r_{x_1,x_2} = 0.20$$

$$s_y = 0.35$$

$$s_{x_1} = 10.52$$

$$s_{x_2} = 2.64$$

- (a) Calculate the regression coefficients for the effects of  $x_1$  and  $x_2$  on  $y$ . Explain what these coefficients mean in plain English.
- (b) Calculate the standardized regression coefficients for the effects of  $x_1$  and  $x_2$  on  $y$ . Explain what these coefficients mean in plain English.

- (c) Which one of the variables has the largest effect on prison violence? Explain why.
- 2.7. A study of recidivism classified offenders by type of punishment received: prison, jail, probation, fine, or community service. A researcher interested in the effects of these different punishments analyzes data on a sample of 967 offenders. She computes two regression models. In the first, she includes variables for age, sex, race, number of prior arrests, severity of the last conviction offense, and length of punishment. The  $R^2$  for this model is 0.27. In the second model, she adds four dummy variables for jail, probation, fine, and community service, using prison as the reference category. The  $R^2$  for this model is 0.35. Explain whether the type of punishment had an effect on recidivism (assume a 5% significance level).
- 2.8. A public opinion poll of 471 randomly selected adult respondents asked about their views on the treatment of offenders by the courts. Expecting race/ethnicity to be related to views about the courts, a researcher classifies respondents as African American, Hispanic, and white. To test for the effect of race/ethnicity, he computes one regression using information about the age, sex, income, and education of the respondents and finds the  $R^2$  for this model to be 0.11. In a second regression, he adds two dummy variables for African American and Hispanic, using white as the reference category. The  $R^2$  for this second model is 0.16. Explain whether the race/ethnicity of the respondent had a statistically significant effect on views about the courts (assume a 5% significance level).

## Computer Exercises

In Chap. 15 of *Basic Statistics in Criminal Justice: Volume 1*, we explored the basic features of the regression commands in SPSS, Stata, and R in the computation of a simple regression model. To compute a multiple regression model, we simply add additional independent variable names to the list of independent variables on the command line. The following exercises illustrate some of the additional features of the regression command. Please see the appropriate SPSS (Chapter\_2.sps) or Stata (Chapter\_2.do) syntax file for specific examples.

### SPSS

#### *Standardized Regression Coefficients (Betas)*

The standardized regression coefficients (Betas) are part of the standard output for SPSS' linear regression command. In the table of results presenting the coefficients, the standardized coefficients are located in the column following those presenting the values for the regression coefficients ( $b$ ) and the standard errors of  $b$ . Nothing else is required to obtain the standardized coefficients.

### *F-Test for a Subset of Variables*

The computation of an *F*-test for a subset of variables requires a little planning in setting up a multiple linear regression model. When thinking about your regression model and a test of one or more subsets of variables, you will need to enter these independent variables on separate /METHOD = ENTER lines. In general, if we have one subset that we are interested in, we would use the following syntax:

```
REGRESSION  
/STATISTICS COEFF R ANOVA CHANGE  
/DEPENDENT dep_var_name  
/METHOD = ENTER list_of_variables_in_first_model  
/METHOD = ENTER list_of_variables_in_second_model.
```

The trick here is to keep track of all the independent variables in your regression model and determine whether they belong to the first or the second group—the second group would be the subset of interest. For example, suppose we had an interest in looking at whether demographic characteristics of offenders affected punishment severity. In this case, we would then list the demographic characteristics (however measured) in the second block (i.e., /METHOD = ENTER line).

We have also added the /STATISTICS option line to the REGRESSION command. The reason for this is to force SPSS to compute the *F*-test on the subset of variables and to simultaneously report all of the other results in a linear regression analysis that it usually reports. Specifically, the items on the /STATISTICS line request the coefficient table (COEFF), model summary (R), ANOVA table (ANOVA), and change in  $R^2$  when the second block of variables is added to the regression model (CHANGE). The *F*-test on the subset of variables is produced with the CHANGE option.

The output from running this command is nearly identical to what you have viewed previously. The major difference is that there will be two major rows of results for all of the tables viewed in the output before—one row will be labeled Model 1 and the other row Model 2. In other words, there will be a row for the *reduced* model (Model 1 in SPSS) that contains only those variables included in the first block of variables and a second row for the full model that includes all variables (Model 2 in SPSS).

The *F*-test for the subset of variables can be found in the *Model Summary* table of results under the columns labeled *Change Statistics*. For Model 2, the *F*-statistic for the subset of variables appears in the column labeled *F Change*. The value for the numerator degrees of freedom ( $df_1$ ) will appear in the next column to the right and will equal the number of independent variables included in the subset. The value for the denominator degrees of freedom ( $df_2$ ) appears in the next column, providing you with all the information you need to test whether the subset of variables makes a statistically significant contribution to the overall regression model.

Since the description of the various pieces may be confusing, we encourage you to open and run the accompanying SPSS syntax file for this chapter (Chapter\_2.sps).

### *Residual Plot*

It is also possible with the regression command to analyze residuals in ways ranging from simple to complex. Perhaps, the most straightforward way of analyzing residuals is graphical, using a residual plot that SPSS can produce. There are many kinds of residual plots that SPSS could create—we highlight only one simple example here. A histogram of the residuals from a regression analysis with a normal curve overlaid on the histogram is obtained as follows:

```
REGRESSION
/DEPENDENT dep_var_name
/METHOD=ENTER list_of_indep_vars
/RESIDUALS HISTOGRAM(ZRESID).
```

where the /RESIDUALS line will request a plot of residuals—the HISTOGRAM (ZRESID) option specifies a histogram of what are known as *standardized residuals*. A histogram of the residuals with the overlaid normal curve will give you some idea of how closely the residuals approximate a normal distribution (which is what is to be expected). If the residuals do not resemble a normal distribution, this is often an indication of a problem with the regression model, such as one or more relevant independent variables having been omitted from the analysis.

## **Stata**

### *Standardized Regression Coefficients (Betas)*

To request the standardized regression coefficients (Betas) in a multiple linear regression model in Stata, you will need to add the option **b** to a **regress** command:

```
regress dep_var_name indep_var_names, b
```

The last column of output in the coefficient table will then report the standardized coefficients.

### *F-Test for a Subset of Variables*

In contrast to the cumbersome syntax in SPSS for testing a subset of variables, the syntax required in Stata involves two steps: (1) Estimate the full regression model with the **regress** command, and (2) test the subset of variables using the **testparm** command. The form of the **testparm** command is simply as follows:

```
testparm list_of_subset_variables
```

The output from running this command is an *F*-test on the set of variables listed on the **testparm** command line.

### *Residual Plot*

At the end of Chap. 15 in *Basic Statistics in Criminal Justice: Volume 1*, we illustrated the process for computing residuals from a linear regression analysis. To create a histogram of the residuals, we would use the **histogram** command (discussed at the end of Chap. 3 in *Volume 1*), and if we wanted to overlay a normal curve, we use the **normal** option:

```
regress dep_var_name indep_var_names
predict RES_1, r
histogram RES_1, normal
```

### R

To get started in R, install and load the *haven* package. Then, import one of the datasets from the appropriate folder (*nys\_1.sav*, *nys\_1\_student.sav*, or *nys\_1.dta*). Below, we place the dataset into an object called, *df*, but this can be any name you wish.

```
install.packages("haven")
library(haven)
# Load the datasets using read.dta()
# or you may also use read.sav()
df <- read.dta("nys_1.dta")
```

### *Standardized Regression Coefficients (Betas)*

The estimation of regression models in R uses the **lm()** function, which is short for *linear model*. The **~** in the syntax below is indicating that the variable to the left is a function of what is to the right. We are storing the model in an object called, *regmodel*. This can be any name you wish. To generate output with unstandardized regression coefficients, you may use the **summary()** function and apply the function to our regression model object. We are also going to use the **lm.beta()** function from the *QuantPsy* package because it will provide us the standardized regression coefficients for our model. Do not forget to first install and load the *QuantPsy* package.

```
regmodel <- lm(dep_var_name ~ indep_var_name1 +
    indep_var_name2, data= dataset_name)
summary(regmodel) # Unstandardized coefficients
lm.beta(regmodel) # Standardized coefficients
```

### *F-Test for a Subset of Variables*

The equivalent to Stata's **testparm** command in R is the **regTermTest()** function, which is from the *survey* package. The form of the **regTermTest()** command is simply as follows:

```
#For one independent variable
regTermTest(regmodel, "indep_var_name1")
```

```
#For multiple variables
regTermTest(regmodel, ~dep_var_name +
indep_var_name1 + indep_var_name2)
```

The output from running this command is an *F*-test on the set of variables listed with the **regTermTest()** function.

### *Residual Plot*

Predicted values and residuals are generated with the **predict()** and **resid()** functions. We can store the results either in a named vector or as a new variable in our dataset. Both methods are shown below.

```
ypredict <- predict(regmodel)
yresidual <- resid(regmodel)
dataset_name$ypredict <- predict(regmodel)
dataset_name$yresidual <- resid(regmodel)
```

To plot the residuals in a histogram, you do not need to use the **predict()** or **resid()** functions. You can use the **gg\_reshist()** function from the *lindia* package. It will also even allow you to specify the number of bins on the histogram.

```
gg_reshist(regmodel2)
gg_reshist(regmodel2, bins = 10)
```

### Problems

1. Enter the data from Table 2.2. Run the regression command to reproduce the unstandardized and standardized regression coefficients presented in this chapter.
  - (a) Compute two simple regression models using years in prison as the independent variable in one regression and prior arrests as the independent variable in the second regression. Generate a histogram of the residuals for each regression model. What does the pattern of results in this plot suggest to you about the distribution of error terms?
  - (b) Compute the multiple regression model, and generate a histogram of the residuals for this regression model. How has the pattern of error terms changed relative to the two histograms produced in part (a)?

Open the NYS data file (nys\_1.sav, nys\_1\_student.sav, or nys\_1.dta) to do Exercises 2 through 5.

2. Compute a multiple regression model using number of times the student hit other students as the dependent variable. From the variables included in the data file, select at least five independent variables that you think have some relationship to hitting other students.

- (a) Explain what each regression coefficient ( $b$ ) and standardized regression coefficient (Beta) means in plain English.
  - (b) Generate a histogram of the residuals for this regression model. What does the pattern of results in this plot suggest to you about the distribution of error terms?
3. Compute a multiple regression model using number of times something worth \$5 or less has been stolen as the dependent variable. From the variables included in the data file, select at least five independent variables that you think have some relationship to stealing something worth \$5 or less.
- (a) Explain what each regression coefficient ( $b$ ) and standardized regression coefficient (Beta) means in plain English.
  - (b) Generate a histogram of the residuals for this regression model. What does the pattern of results in this plot suggest to you about the distribution of error terms?
4. Compute a multiple regression model using number of times the student cheated on exams as the dependent variable. From the variables included in the data file, select at least five independent variables that you think have some relationship to cheating on exams.
- (a) Explain what each regression coefficient ( $b$ ) and standardized regression coefficient (Beta) means in plain English.
  - (b) Generate a histogram of the residuals for this regression model. What does the pattern of results in this plot suggest to you about the distribution of error terms?
5. Compute a multiple regression model using number of times drunk as the dependent variable. Use age, sex, race, employment status, hours spent studying per week, grade point average, and number of friends who use alcohol as the independent variables.
- (a) Use an  $F$ -test to test whether demographic characteristics—age, sex, and race—affect drinking behavior.
  - (b) Use an  $F$ -test to test whether academic characteristics—hours spent studying per week and grade point average—affect drinking behavior.

## References

---

- Babbie, E., & Maxfield, M. (1995). *The practice of social research in criminal justice*. Belmont: Wadsworth.
- Beckman, R. J., & Cook, R. D. (1983). Outlier. .... s. *Technometrics*, 25(2), 119–149.
- Berk, R. A., & Freedman, D. A. (2003). Statistical assumptions as empirical commitments. In T. Blomberg & S. Cohen (Eds.), *Law, punishment, and social control: Essays in honor of Sheldon Messinger* (Vol. 2, pp. 235–254). Berlin: Aldine de Gruyter.
- Fox, J. (2016). *Applied regression analysis and generalized linear models*. Thousand Oaks: Sage.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel hierarchical models* (Vol. 1, p. 173). New York: Cambridge University Press.
- Greene, W. H. (2018). *Econometric analysis* (8th ed., pp. 16–17). London: Pearson.
- Harrell, F. E., Jr. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis*. Cham: Springer.
- Hartley, H. O., & Sielken, R. L., Jr. (1975). A “super-population viewpoint” for finite population sampling. *Biometrics*, 31(2), 411–422.
- Moore, M. (1986). Purblind justice: Normative issues in the use of prediction in the criminal justice system. In A. Blumstein, J. Cohen, A. Roth, & C. A. Visher (Eds.), *Criminal Careers and “career criminals”* (p. 2). Washington, DC: National Academy Press.
- Weisburd, D., & Piquero, A. (2008). How well do criminologists explain crime?: Statistical modeling in published studies. *Crime and Justice*, 17, 453–502.
- Weisburd, D., Lum, C., & Petrosino, A. (2001a). Does research design affect study outcomes in criminal justice? *The Annals*, 578, 50–70.
- Weisburd, D., et al. (2001b). *The abuse of authority: A national study of police officers' attitudes*. The Police Foundation: Washington, DC.

## Chapter three

---

# Multiple Regression: Additional Topics

## Nominal Independent Variables

---

- Can You Dummy Code Nominal Variables with Three or More Categories?
- What Alternative Indicator Coding Methods Are There?

## Nonlinear Relationships

---

- What are They?
- How are They Included in Regression Models?
- How are They Interpreted?

## Transformations

---

- What are They?
- When Would We Transform a Dependent Variable?
- When Would We Transform an Independent Variable?

## Interaction Effects

---

- What are They?
- How are They Included in Regression Models?
- How are They Interpreted?

## M u l t i c o l l i n e a r i t y

---

When Does It Arise?

How is It Diagnosed?

How is It Treated?

**I**N THE PREVIOUS CHAPTER, we introduced the basic concepts and methods of ordinary least squares (OLS) regression and illustrated how to incorporate multiple independent variables into such a model. We also explored the assumptions underlying OLS regression and diagnostic methods for assessing the plausibility of these assumptions. The models examined were restricted to a linear relationship between the dependent variable and the independent variables. In the real world, we are sometimes confronted with more complex research questions that require us to make additional modifications to our regression model. These include transformations of variables, modeling nonlinear relationships, modeling interaction effects, and addressing multicollinearity among independent variables. Additionally, the previous chapter also showed how to include a binary nominal variable into a multiple regression model using *dummy coding*. We are often interested in including nominal variables with three or more categories into a regression model, and the dummy coding method can be extended to handle such cases. These additional topics will be addressed in this chapter.

In an OLS multiple regression model, our interpretation of the coefficients is based on the notion that there is a linear relationship between the independent variables and the dependent variable. What if we find evidence of a curvilinear relationship? Or, what if theory suggests that there may be a nonlinear relationship between two variables? Although the OLS regression model is based on the assumption of a linear relationship between the dependent and each of the independent variables, *nonlinear relationships* can be incorporated into an OLS regression model in a straightforward manner.

Another issue in the application of OLS regression is that the interpretation of the coefficients is based on the idea that each independent variable has a constant additive effect on the dependent variable irrespective of the levels of other independent variables. For example, if we include a binary variable in the model, we assume that the effect of every other independent variable is the same for each of the two categories of this variable, such as

men and women. But what if there was a good theoretical or policy reason to suspect that the effect of some variable was different for men and women? How would we incorporate that into our model? In the statistical literature, these are known as *interaction effects*, and they allow us to test whether the effect of one independent variable varies by the levels of another independent variable.

In this chapter, we also introduce an important problem that researchers sometimes face when estimating multiple regression models called *multicollinearity*. Multicollinearity is a situation where independent variables are very highly correlated with each other. This can lead to unstable regression coefficients and associated significance levels. When independent variables are highly intercorrelated, it is difficult for OLS regression to determine the unique effect of each independent variable.

## Nominal Variables with Three or More Categories in Multiple Regression

---

We saw in the previous chapter that it is straightforward to include a binary normal variable, such as a person's gender,<sup>1</sup> into a multiple regression model through **dummy coding**. We can extend dummy coding to nominal variables with three or more categories by creating additional dummy variables. To do so, you must create a separate variable for each category of the nominal variable, less one as will be explained. For example, the Police Foundation study on abuse of authority used in the prior chapter divided the USA into four regions: North Central, Northeast, South, and West. In practice, you would need to create a separate variable for each of these regions. In other words, you would define a variable that would be coded 1 for all those officers in the North Central region and 0 for all other officers. You would repeat this process for each of the other regional categories. This is shown in Table 3.1.

As with the binary independent variable, you must choose one of the categories to be a reference category. That is, with officer gender, we only needed the dummy variable for female and not two variables, one for female and one for male, as these two variables are redundant. Notice that with Table 3.1, once you know an officer's status on any three of these variables, you know their status on the fourth. For example, if you know that an officer has the value of 0 on North Central, Northeast, and

---

<sup>1</sup>We recognize that gender is not strictly binary, with some individuals being intersex at birth, and that gender includes male and female as well as nonbinary, transgender, and other identities. However, it is still common in the social sciences to treat this as a binary construct.

**Table 3.1**

Illustration of dummy coding for a four-level nominal variable

CATEGORY OF REGION VARIABLE	DUMMY VARIABLE			
	NORTH CENTRAL	NORTHEAST	SOUTH	WEST
North Central	1	0	0	0
Northeast	0	1	0	0
South	0	0	1	0
West	0	0	0	1

Note: One of the dummy variables would need to be defined as the reference category and excluded from the regression model, as discussed in the text

South, then you know that they must be an officer from the West region. In this case, when you have three of the dummy variables in the equation, you have already specified the fourth. Thus, in our regression model, we drop the dummy variable for one of the categories which now becomes our reference category. Statistically, it does not matter which category you drop. The overall model will be the same with the same predicted values, residuals, and model fit. However, conceptually, it often makes sense to omit a specific category. For example, if you were examining race and wanted to see how minority groups compared with the majority group, you would choose to exclude the majority group category. In that case, it would allow you to see how much each of the minority groups differed from the majority group. When you do not have a specific theoretical reason for excluding a specific group, it is good practice to choose the category with the largest sample size. This is because the reference category in some sense anchors the measure overall, and using the largest category will add some stability to your estimates. Conversely, it is bad practice to use a category as the reference category that has very few cases in it.

Absent a theoretical reason for choosing a reference category, in our example, the largest number of officers is drawn from the South, so we will use it as the reference category. In our example, we will therefore include a dummy variable for North Central, Northeast, and West but not for South. We will start with a regression model that only includes these dummy variables to help illustrate how they work. We will then build a more extensive model with the other variables used in the prior chapter. The results from this model are shown in Table 3.2. Table 3.3 shows the mean hours worked by region.

Writing out the regression model yields the following:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + e$$

where  $x_1$  is the dummy variable for North Central,  $x_2$  is the dummy variable for Northeast, and  $x_3$  is the dummy variable for West. Writing this out

**Table 3.2**

Regression model predicting hours worked per week based on region

VARIABLE	REGRESSION COEFFICIENT ( $B$ )	STANDARD ERROR	$t$	$p$
(Intercept)	46.984	0.375	125.443	<0.0001
North Central	-2.455	0.610	-4.028	<0.0001
Northeast	-1.839	0.574	-3.205	0.0014
West	-0.853	0.618	-1.381	0.1678

Multiple  $R$ -squared: 0.021, Adjusted  $R$ -squared: 0.018,  $F(3, 919) = 6.556$ ,  $p = 0.0002$

**Table 3.3**

Descriptive statistics for hours worked per week by region

REGION	MEAN	SD	N
North Central	44.529	5.72	191
Northeast	45.145	6.44	234
South	46.984	6.97	183
West	46.131	7.21	315

separately for each region and inserting the values of the dummy variables produce the following:

$$\text{North Central: } y = b_0 + b_1(1) + b_2(0) + b_3(0) + e$$

$$\text{Northeast: } y = b_0 + b_1(0) + b_2(1) + b_3(0) + e$$

$$\text{West: } y = b_0 + b_1(0) + b_2(0) + b_3(1) + e$$

$$\text{South: } y = b_0 + b_1(0) + b_2(0) + b_3(0) + e$$

The multiplication by zero cancels out some of the terms in each equation. If we rewrite these equations, dropping out coefficients multiplied by zero and also dropping the error term, we get:

$$\text{North Central: } y = b_0 + b_1(1)$$

$$\text{Northeast: } y = b_0 + b_2(1)$$

$$\text{West: } y = b_0 + b_3(1)$$

$$\text{South: } y = b_0$$

What this shows is that the intercept is the mean for the South region and the regression coefficient for each of the other regions reflects the difference between its mean and the mean for the South. Substituting in the actual values for the regression coefficient and calculating the results produce the following:

$$\text{North Central: } y = 46.984 - 2.455(1) = 44.529$$

$$\text{Northeast: } y = 46.984 - 1.839(1) = 45.145$$

$$\text{West: } y = 46.984 - 0.853(1) = 46.131$$

$$\text{South: } y = 46.984 = 46.984$$

Notice that this produces the mean hours worked per region, as shown in Table 3.3. Furthermore, if we compute the mean difference between each of the regions and the South, we reproduce the regression coefficient for each, as shown below.

$$\text{North Central} - \text{South} = 44.529 - 46.984 = -2.455$$

$$\text{Northeast} - \text{South} = 45.145 - 46.984 = -1.839$$

$$\text{West} - \text{South} = 46.131 - 46.984 = -0.853$$

This illustrates the meaning of the term *reference* category. The omitted category serves as the references against which the other categories are compared.

The similarity of this model to the one-way ANOVA is also worth pointing out. The  $F$ -statistic for the overall fit for this regression model equals the  $F$ -statistic that you would have obtained by performing a one-way ANOVA comparing these four means. The statistically significant  $F$  allows us to reject the null that these four means are equal and conclude that officers in the four regions work a different number of hours, on average. Furthermore, the significant regression coefficients for North Central and Northeast indicate that these two regions work fewer hours, on average, than the South. Changing the reference category, such as to the West, would change the regression coefficients as they would now represent the difference between the West region and each of the remaining regions. Importantly, the overall  $F$  would remain unchanged.

What happens when we include additional variables? The mean differences between these groups becomes adjusted for any distributional differences on those other variables. In the last chapter, we used these data to build a model predicting hours worked from educational level, years with the department, and the officer's sex. Adding these variables to the model along with the dummy variables for region produces the model shown in Table 3.2. Notice that the coefficients for the region dummy variables have changed slightly, reflecting some level of confounding (correlation) between region and the other independent variables in this model. If we were interested in the overall effect of region after taking into account the effects of educational level, years with the department, and officer sex, we could use the  $F$ -statistic for nested models presented in the prior chapter.

**Table 3.4**

Regression model predicting hours worked per week based on region

VARIABLE	REGRESSION COEFFICIENT (B)	STANDARD ERROR	t	p
(Intercept)	46.222	0.795	58.118	< 0.0001
Region				
North Central	-2.337	0.607	-3.848	0.0001
Northeast	-1.735	0.571	-3.039	0.0024
West	-0.985	0.617	-1.597	0.1107
Educational level	0.447	0.173	2.586	0.0099
Years with department	-0.070	0.026	-2.670	0.0077
Female (1 = yes, 0 = no)	-1.776	0.795	-2.233	0.0258

Multiple R-squared: 0.039, Adjusted R-squared: 0.033, F(3, 916) = 6.165, p &lt; 0.0001

This produces an  $F = 5.84$  with 3 and 916 degrees of freedom. This is significant at  $p < .05$ , indicating that region accounts for some of the variability in hours worked, above and beyond what is explained by the other variables.

Dummy variable coding is not the only method of creating indicator variables. There are several other types of indicator coding as well, each with a particular use. See <https://stats.idre.ucla.edu/spss/faq/coding-systems-for-categorical-variables-in-regression-analysis/> for details on these methods.

## Nonlinear Relationships

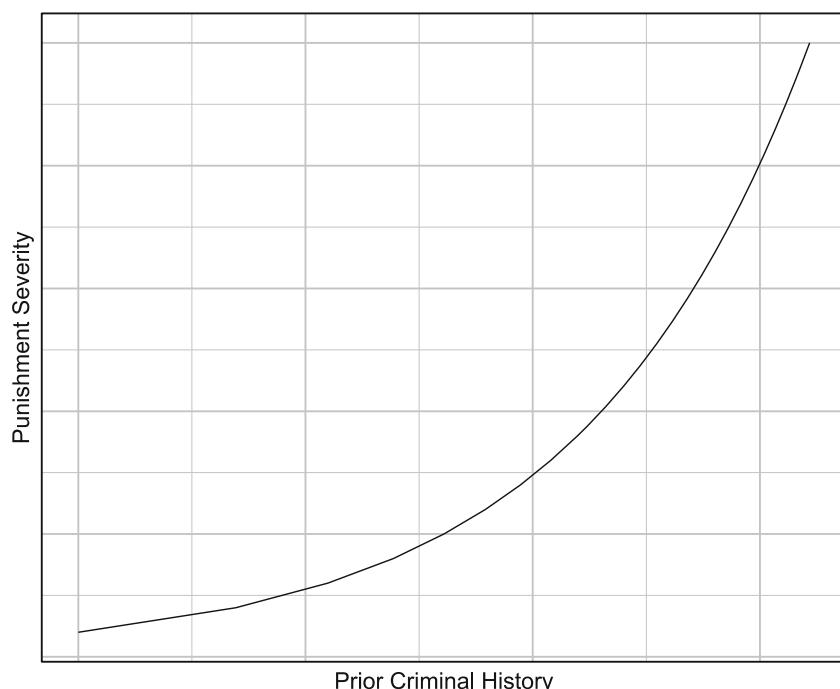
---

Policy-oriented research focused on the severity of punishment for convicted offenders illustrates that as the severity of an offender's prior record and the severity of the conviction offense increase, the severity of the punishment tends to increase (Tonry 1997). In the interpretation of OLS regression results, we would say something, for example, about how each one-unit increase in the severity of an offender's prior record results in the length of a sentence increasing by some fixed time period (e.g., 8 months). Key to the interpretation of OLS regression coefficients is the idea that the level of the independent variable does not matter—each unit change is expected to result in the same change in the dependent variable regardless of whether we are looking at small or large values on the independent variable. But, this is not always the case.

For example, an assumption often made in research on sentencing outcomes is the idea that first-time offenders (i.e., those with no prior record) or those offenders convicted of relatively minor forms of crime will be punished much more leniently than other offenders. Then, as the severity of prior record or of conviction offense increase, there is an expectation of an increasingly punitive response by the criminal justice system. Figure 3.1

**Figure 3.1**

*Hypothetical nonlinear relationship between punishment severity and prior criminal history*



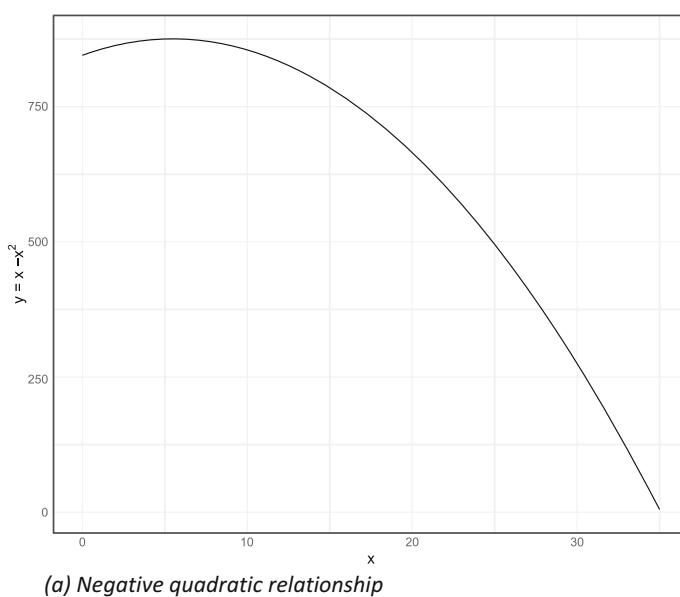
presents a hypothetical plot for punishment severity and prior criminal history that reflects increasingly harsher punishments for offenders with more extensive criminal records.

As can be seen in the figure, there is a gradual increase in the severity of the punishment as the severity of the prior record increases. Then, the increases in the severity of the punishment become larger for the same unit increase in prior criminal history.

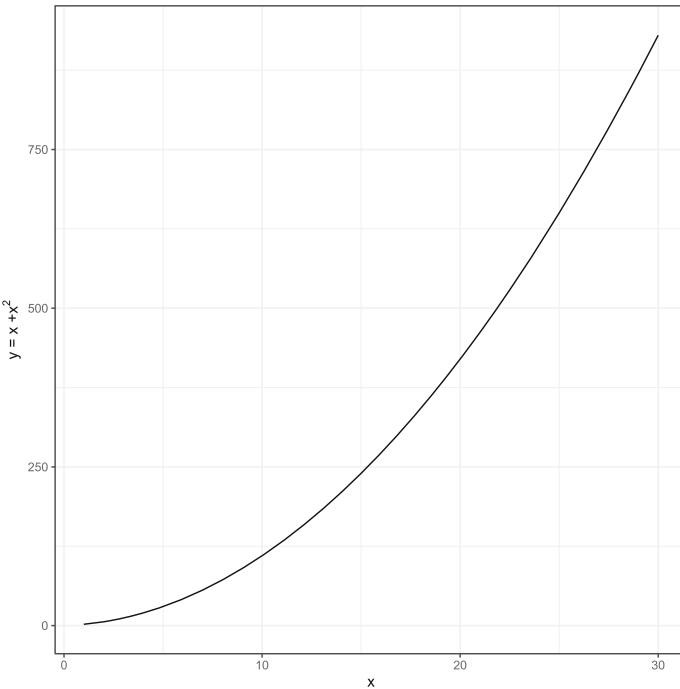
The range of potential **nonlinear relationships** is limitless and is bounded only by the imagination and creativity of the researcher and the theoretical basis for conducting the research. Yet, while there may be a wide range of possible nonlinear relationships, most researchers will confine their analyses to a relatively limited group of nonlinear possibilities, some of which are displayed in Fig. 3.2. Panel (a) presents what is referred to as a quadratic equation. All that this means is that a squared term has been added to the equation to give it a form such as  $y = x - x^2$  or  $y = x + x^2$ . The quadratic equation is one of the more commonly used transformations in criminology and criminal justice and has had frequent application in the

**Figure 3.2**

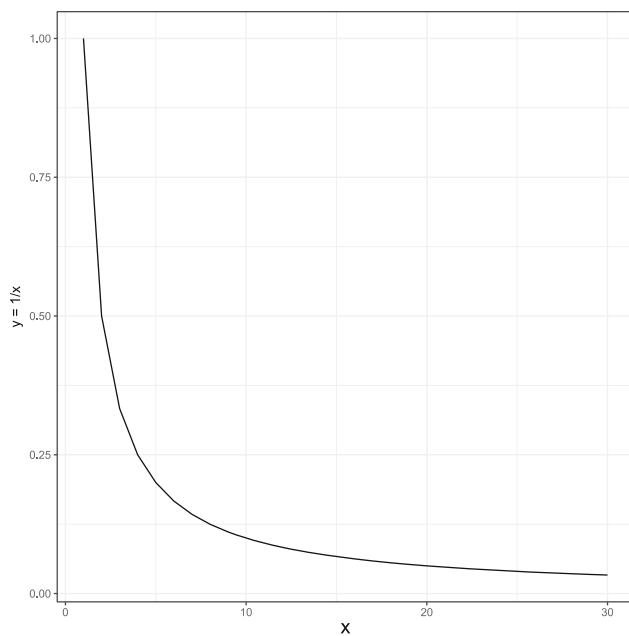
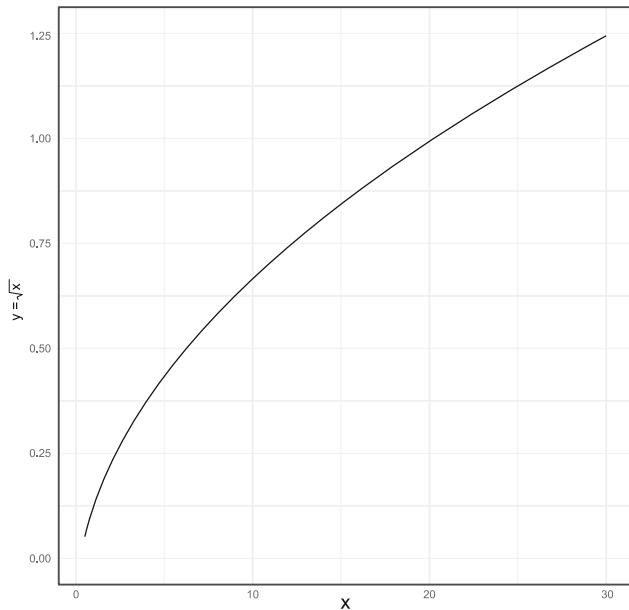
Common nonlinear relationships used in criminal justice research



(a) Negative quadratic relationship



(b) Positive quadratic relationship

**Figure 3.2***(continued)**(c) Inverse relationship**(d) Square root relationship*

study of age-related behavior. Changing the sign in the function changes the direction of the curve. When it is negative, the curve is bent downward, as shown in Panel (a). When it is positive, the curve is bent upward, as shown in Panel (b). Panel (c) presents an inverse function of the form  $y = 1/x$ . This kind of transformation helps to capture relationships where there is a decreasing negative effect of the independent variable on the dependent variable. Panel (d) presents a square-root transformation of the form  $y = \sqrt{x}$ . This kind of transformation is useful when there is a diminishing positive impact of the independent variable on the dependent variable.

### Finding a Nonlinear Relationship: Graphical Assessment

Perhaps the most straightforward way of exploring data for a nonlinear relationship is to use a line graph. A simple scatterplot will often contain so many data points that it is difficult to discern any pattern in the data. A line graph that plots the mean of the dependent variable against the value of the independent variable will likely provide a rough indication of the nature of the relationship between the two variables. For example, Fig. 3.3 presents the mean for length of sentence against the severity of the conviction offense for over 20,000 offenders sentenced in Pennsylvania in 1998.<sup>2</sup> As you look at Fig. 3.3, you can see that there is a gradual, linear increase in length of sentence as offense severity increases to about level 6–7. At that point, the increases in sentence length become larger for each additional increase in offense severity. To highlight the curvilinear nature of the relationship between offense severity and length of punishment, the OLS regression line for these data is overlaid in Fig. 3.3, indicating that the straight-line relationship does not capture the relationship between length of sentence and severity of offense particularly well. When the independent variable is continuous or has a large number of possible values, it is helpful to *bin* the values as you would in a histogram and compute the mean of the dependent variable for intervals of the independent variable. This smooths out the graph. Most modern statistical software programs also have the ability to fit and plot splines of the moving average. This produces a smoothed line that follows the ups and downs of the dependent variable across levels of the independent variable.

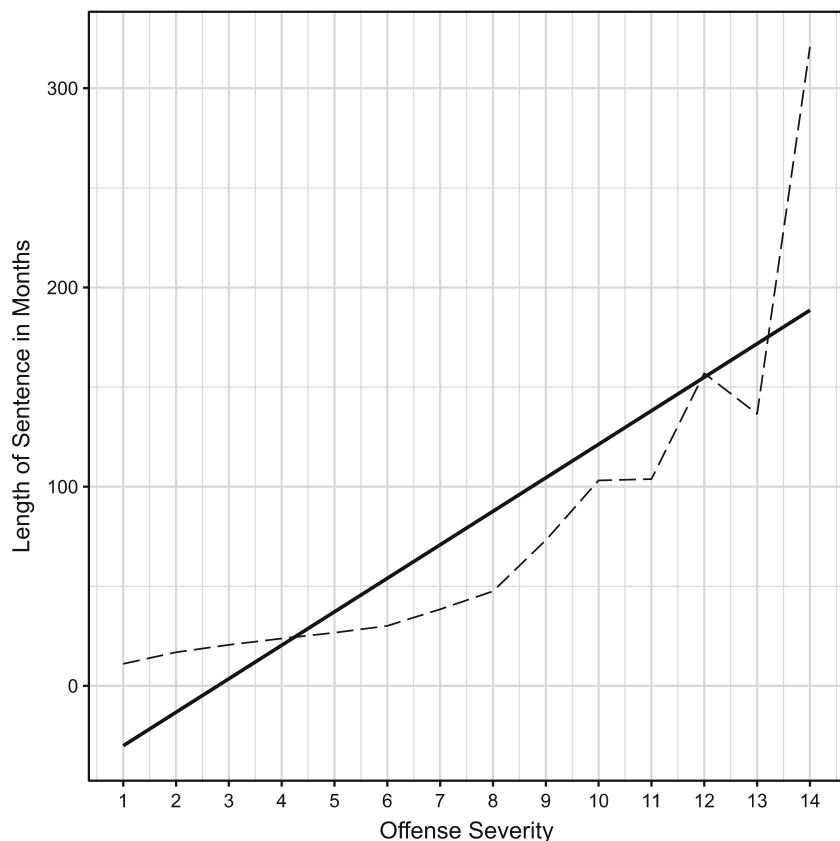
There are two main approaches to addressing a nonlinearity. The first is adding quadratic terms, such as a squared or cubed version of an independent variable into the model. The second is transforming either the dependent variable or an independent variable, or both. We will start by exploring the first of these.

---

<sup>2</sup>These data are available through the National Archive of Criminal Justice Data and can be accessed at <https://www.icpsr.umich.edu/NACJD>.

**Figure 3.3**

*Plot for mean length of sentence by offense severity for offenders in Pennsylvania with linear regression line*



### **Incorporating Nonlinear Relationships into an OLS Model Using a Quadratic Term of an Independent Variable**

Assuming that we have good reason for believing that a nonlinear relationship exists between the dependent variable and one or more of the independent variables, how can we incorporate this information into the OLS regression model? The first step, as noted above, is to try and gain a sense of the relationship graphically. In most circumstances, if theory suggests or if we find evidence of a curvilinear relationship, the most straightforward approach is to add a quadratic term—the squared value of the independent variable—such as that in Panel (a) and Panel (b) of Fig. 3.2. More formally, a quadratic regression equation would have the following form:

$$y = b_0 + b_1x_1 + b_2x_1^2$$

where  $y$  represents the dependent variable, and  $x_1$  represents the independent variable.

In our example presented in Fig. 3.3, we have evidence of a curvilinear relationship that might be accounted for by adding a squared term for offense severity to a regression equation. We begin by noting that the OLS regression line portrayed in Fig. 3.3 is:

$$y = -6.767 + 9.030x_1$$

where  $y$  represents length of sentence (in months), and  $x_1$  represents offense severity.

To incorporate a nonlinear relationship, we begin by computing a new transformed variable—in this case offense severity squared—and then add this transformed variable to the regression equation. When we square offense severity and add it to the regression equation, we obtain the following:

$$y = b_0 + b_1x_1 + b_2(x_1x_1) = b_0 + b_1x_1 + b_2x_1^2$$

If we then estimate this new regression equation that includes both the original measure of offense severity and the squared value of offense severity, we obtain the following results:

$$y = 33.201 - 11.147x_1 + 1.801x_1^2$$

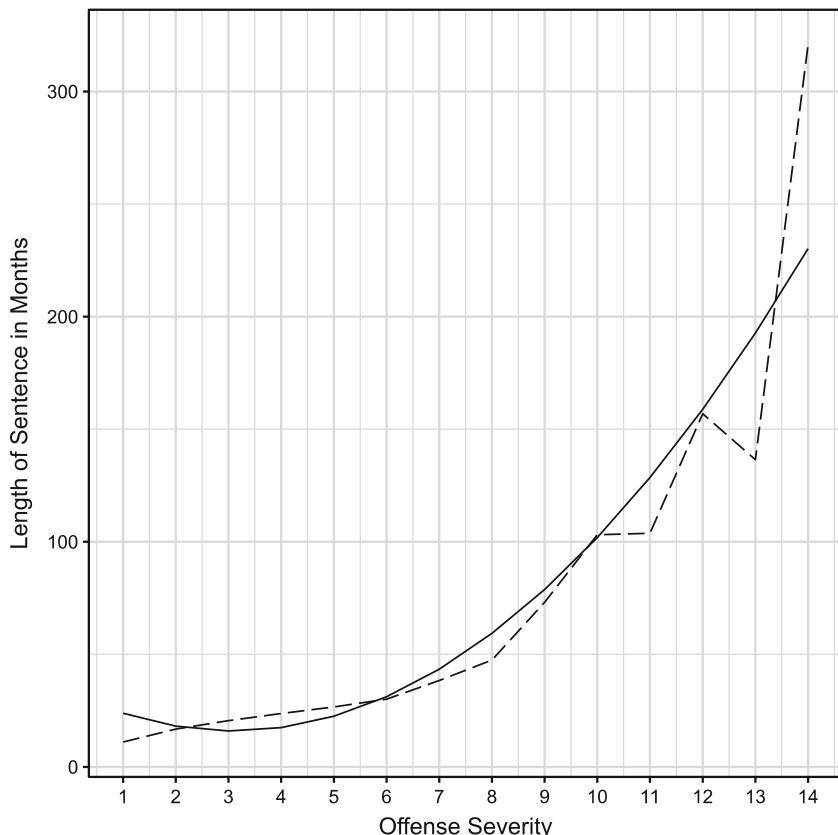
Substantively, this regression equation captures the curvilinear relationship between offense severity and sentence length much better than a straight-line relationship, since there are increasingly larger increases in sentence length for each unit change in offense severity. Figure 3.4 presents the mean sentence length by offense severity (similar to that in Fig. 3.3) along with the new regression line based on including the quadratic term.

### *Interpreting Nonlinear Coefficients*

In many practical applications of adding nonlinear terms to OLS regression models, there is often less emphasis on the interpretation of the individual coefficients that represent transformed variables than on the shape of the relationship. The reason for this is the difficulty in making sense of individual coefficients. For example, in the example using the data on offenders

**Figure 3.4**

*Plot for mean length of sentence by offense severity for offenders in Pennsylvania with quadratic regression line*



sentenced in Pennsylvania, the coefficient for the squared value of offense severity is given as 1.801. This coefficient reflects that the effect of offense severity on length of sentence increases as the offense severity increases, as shown in Fig. 3.4. However, the individual coefficient for offense severity and offense severity squared cannot be interpreted individually. This can clearly be seen by the negative coefficient for offense severity despite the overall upward trend of the line.

Consequently, the description of results using nonlinear transformations in OLS regression models should focus on the general pattern of results, rather than on the specific coefficients. Stated differently, we are interested in the full relationship between this independent variable and the dependent variable that is captured by both coefficients (or all three coefficients if a cubed term is added, which allows for a double curve to the line). A

positive value for the coefficient associated with the squared terms indicates that the regression line curves upwards (the effect gets bigger as  $x$  gets bigger), whereas a negative value indicates that the line curves downward. Just looking at the values, however, gives little insight into the fundamental relationship of interest. Our suggestion is to use graphs, such as that presented in Fig. 3.4, which do provide an effective way to convey evidence of a nonlinear relationship between the dependent and independent variables. If there are other variables in the model, these can be held constant at their mean value when producing such a graph. This allows you to illustrate just the relationship of interest. What makes this kind of plot particularly useful is that it conveys both the pattern in the observed data and the predicted values based on the estimated regression model.

### *Note on Statistical Significance*

Estimating statistical significance for a nonlinear term does not present any new problem to our understanding of multiple regression. The statistical significance of both the individual coefficients and the overall model in an OLS regression model incorporating nonlinear terms is determined in the same way as for any other OLS regression model. For individual coefficients, we use the  $t$ -test and for the overall model, we use the  $F$ -test. To get a test of the two (or more) coefficients that are part of the curvilinear relationship, we can use the  $F$ -test discussed in the previous chapter. In this case, you would compare a model without the independent variable and its squared and/or cubed terms to a model with these terms, thus providing a test for the significance of the curvilinear relationship.

### **Transforming the Dependent Variable**

There are times when the best approach to dealing with a nonlinear relationship is to transform the dependent variable. This is particularly useful when the nonlinearity is not restricted to a particular independent variable but is evident in the scatterplot of the residuals against the predicted values. While numerous **transformations** are possible, one of the most useful transformations of the dependent variable is the natural logarithm. Many positively valued variables in criminal justice research have a log-normal distribution. Such a distribution is positively skewed but becomes normally distributed when log-transformed. More importantly, this often reflects an important characteristic of the variable, that is, that the variable grows proportionately or multiplicatively rather than in an additive manner.

A bit of background may be helpful. Normal distributions emerge from random additive processes. Any characteristic that is the sum of a large number of independent random processes will result in a normal distribution (Bulmer 1979). For example, a person's self-esteem is likely determined in an additive way. Self-esteem is the result, in part, of a person's

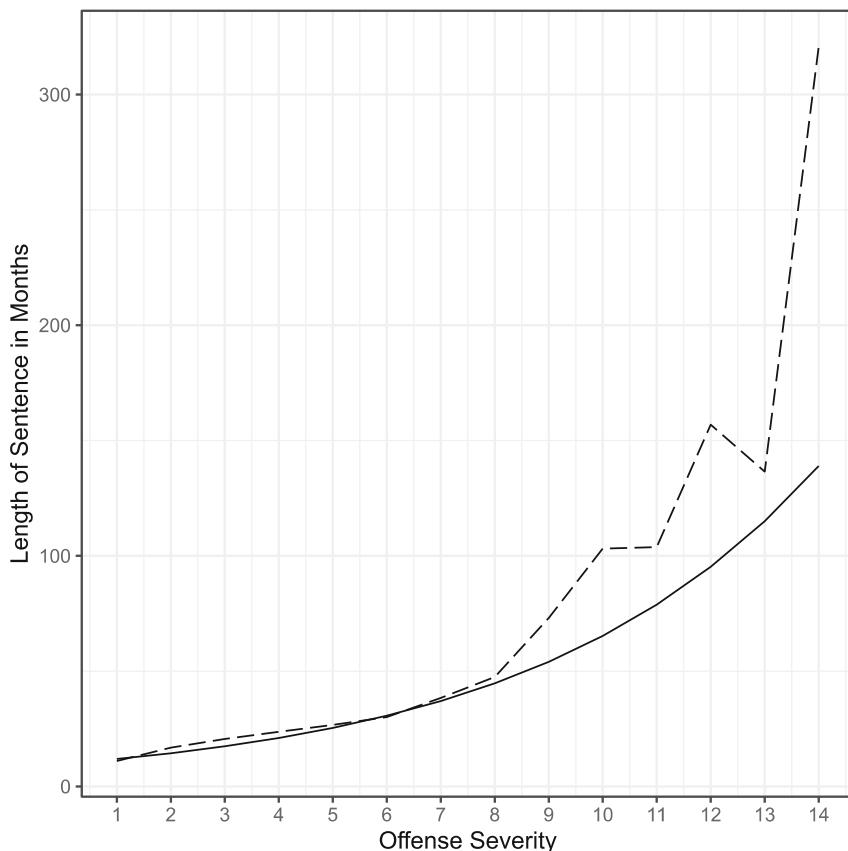
genetic make-up but also a large number of environmental factors, including such things as parenting, family dynamics, peer relationships, teacher relationships, community-level experiences, and exposure to environmental toxins. Each of these factors contributes either positively or negatively to a person's self-esteem and do so in an additive way. In contrast, log-normal distributions are the multiplicative product of a large number of independent random processes. For example, income follows a log-normal distribution as it tends to grow *proportionately*. Cost of living increases and merit raises are almost always based on a percentage of one's salary or hourly wage, rather than set at a fixed (additive) dollar amount. Similarly, investments grow or shrink as a percentage of their current value.

A more relevant example to the field of criminal justice is the dollar amount set for bail. When a judge is considering increases or decreases to a bail amount, she will usually do so proportional to the amount recommended by a prosecutor or bail schedule. For example, she may consider doubling the bail amount. If the recommended bail is \$1000, this would increase it to \$2000, whereas a bail of \$1,000,000 would increase to \$2,000,000. It would make little sense to increase a million-dollar bail by one thousand dollars. Similarly, it would seem cruel to increase a one-thousand-dollar bail by one million dollars. On a natural log scale, a \$1000 increase on a \$1000 initial amount is the same as a \$1,000,000 increase on a \$1,000,000 initial amount. Changes in values of the same proportion have the same difference on the natural log scale. Thus, in many situations, modeling a log-transformed dependent variable has theoretical value in that it may more closely align with the underlying data-generating mechanism, that is, more closely align with how changes occur on that measure. Furthermore, if the dependent variable is positively valued (i.e., negative values are not possible), then a regression model with a log-transformed dependent variable will avoid the problem of negative predicted values. For example, we can reanalyze the data shown in Figs. 3.3 and 3.4 using the log of sentence length and no quadratic term. This is shown in Fig. 3.5. We can see that for lower values of offense severity, this fits the data better than the quadratic model but performs worse for higher values of offense severity. However, an advantage of the natural log transformation of the dependent variable is that it will not predict negative values of the dependent variable (e.g., negative sentence lengths as seen in Fig. 3.3) and it will not predict increased values of the dependent variable for lower levels of the independent variable as shown in Fig. 3.4 when the overall trend is positive. Logically, it does not make sense for an offense severity score of 1 to have a higher predicted sentence length than an offense severity score of 2. This is not the case in Fig. 3.5.

Gelman and Hill (2016) argue that most positively valued dependent variables benefit from the log transformation, so long as there are not a lot of zeros (the log of zero is undefined, creating complications with this

**Figure 3.5**

*Plot for mean length of sentence by offense severity for offenders in Pennsylvania with regression line fit on the natural log of length of sentence*



transformation). OLS regression is often not suitable for count variables and any variable with a large number of zeros and no negative values. Alternative regression modeling methods, such as Poisson-based models, should be considered instead (see Chap. 6). Note, however, that these models use the log transformation to link the dependent variable with the independent variables but in a manner that allows for zeros. Sentence length, however, is not a count but a continuous, positively valued variable and benefits from such a log transformation.

Independent variables that are log-normal also benefit from the log transformation, particularly if they are positively valued and positively skewed. When both the dependent and an independent variable are

log-transformed, economists call the associated regression coefficient an *elasticity*. The coefficient represents the percent change in  $y$  for every 1% change in  $x$ . When only the dependent variable is log-transformed, the exponent of the regression coefficient minus 1 times 100 indicates the percentage change in  $y$  for every one-unit change in  $x$ . In contrast, when only the independent variable is log-transformed, then the regression coefficient, divided by 100, indicates how much change in  $y$  is associated with a 1% change in  $x$ . Thus, the regression coefficients of log-transformed variables in a regression model have useful and intuitive interpretations. The regression coefficient for offense severity for the model shown in Fig. 3.4 is 0.19. Thus, for every 1 unit increase in offense severity, this model predicts a 20% increase in sentence length [ $(\exp(0.19) - 1) \times 100 = 20$ ].

The dependent variable may also be transformed in numerous other ways, such as squaring, taking the square root, and the inverse. These would model a curvilinear relationship that looks like those shown in Fig. 3.2. Additionally, squaring or taking the square root of the dependent variable is often useful to normalize a skewed distribution but is only recommended when the scale has no natural meaning or you are modeling a specific functional form. Many measures of attitudes and psychological traits produce scores that rank people from high to low on those characteristics but are otherwise arbitrary. Thus, squaring or taking the square root of such a measure rarely affects the interpretation as these transformations are monotonic, maintaining the original ranking of individuals. The effect of the transformation is to stretch out or compress portions of the scale to improve the normality of the distribution.

### Review of Nonlinear Relationships

How does one know whether to include a nonlinear term in a regression model? In light of the many different nonlinear relationships that are possible—we could conduct a transformation of any number of our independent variables in an OLS regression model—how do we settle on an appropriate model? The single best guide for the researcher is prior theory and research. If a theoretical perspective claims a nonlinear relationship or prior research has established a nonlinear relationship between two variables, then the researcher may want to examine a nonlinear relationship in the regression analysis. Without the guidance of theory and prior research, the researcher is better off using an OLS model without any nonlinear relationships included. If subsequent analyses, such as a residual analysis (discussed in Chap. 2), indicate a nonlinear relationship, then some kind of transformation of the dependent and/or independent variables may be in order.

## Interaction Effects

---

A number of different theories of crime and delinquency make statements about how the effect of one variable will vary by the level of some other variable. A perspective known as general strain theory hypothesizes that the effects of psychological strain (e.g., having one's parents file for divorce) on delinquency will vary by the ability of a youth to adapt to strain (Agnew 1992). For example, if an individual characteristic, such as self-esteem, helps individuals to adapt to various forms of strain, then the effect of that strain may vary by the level of self-esteem: As the level of self-esteem increases, the effect of strain on the chances of delinquency may become smaller. Alternatively, research on criminal justice decision making has suggested that the effects of offender characteristics, such as the offender's age, may differentially affect the severity of punishment across different racial or ethnic groups (Steffensmeier et al. 1995).

Assuming that we have a rationale for including an **interaction effect**, how do we incorporate it into our regression model? Let us begin with a simple regression model that has two independent variables  $x_1$  and  $x_2$ , as shown here:

$$y = b_0 + b_1x_1 + b_2x_2 + e$$

This model is stating that each independent variable has an *additive* effect on the dependent variable.

To add an interaction effect to a regression model, all that we need to do is to compute the product of the two variables:  $x_1x_2 = x_3$ , introducing a *multiplicative* effect to the model. We then add this term to the regression equation:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + e$$

This additional regression coefficient ( $b_3$ ) represents the interaction of the variables  $x_1$  and  $x_2$ , which we will need to interpret. The interpretation of interaction effects can be complicated, with the degree of complexity based on the level of measurement of the two variables. Also note that including the interaction term alters the meaning and interpretation of the coefficients associated with main (simple) effects for each of the two variables that make up the interaction. In the equation above, this is  $b_1$  and  $b_2$ . We will explore this below, but the general advice is to focus on the interaction effect and to be cautious in interpretation of the two related main effects from such a model.

### Interaction of a Dummy Variable and a Scaled Variable

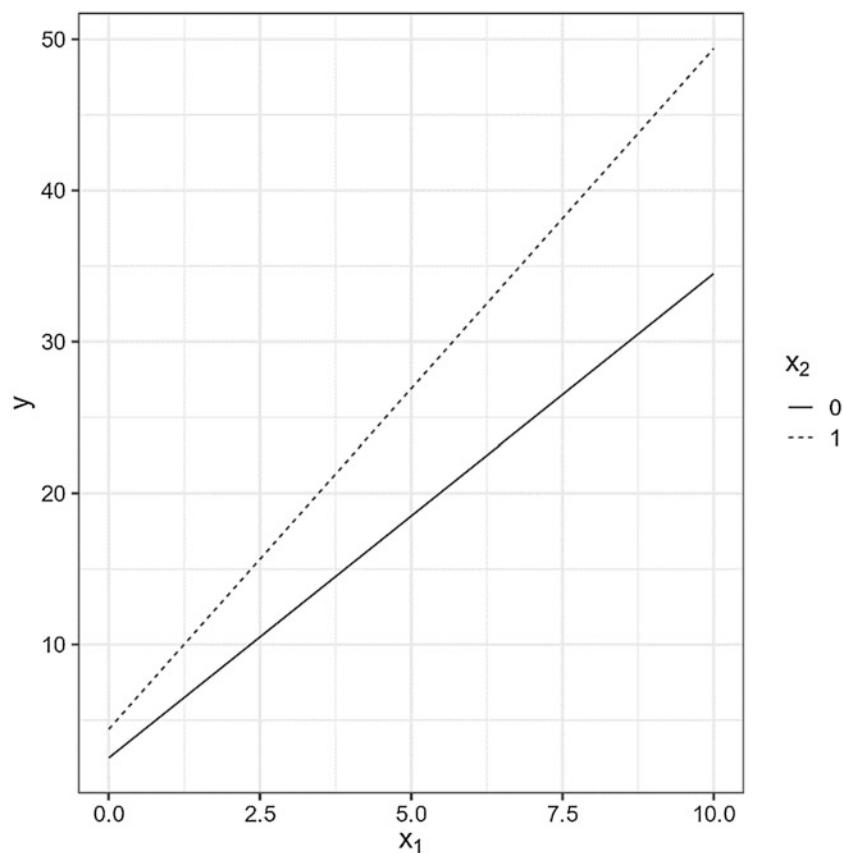
To illustrate the process of interpreting interaction effects, it is useful to begin with a relatively simple case: the interaction of a binary variable with a scaled variable. A scaled variable is simply a multivalued interval- or ratio-level measure. In the regression model above, let us assume that  $x_2$  is a dummy code for a binary variable, where the two categories are coded as either 0 or 1. The regression equation for this model is as follows:

$$y = 2.5 + 3.2x_1 + 1.9x_2 + 1.3x_1x_2$$

Figure 3.6 graphically illustrates this model. In the figure, we can see that there is a modest interaction effect with the lines growing apart as

**Figure 3.6**

*Regression lines for the interaction of  $x_1$  and  $x_2$*



$x_1$  increases. Stated differently, the effect of  $x_2$  increases as  $x_1$  increases. At the lower levels of  $x_1$ , there is only a small difference between the two groups represented by  $x_2$ , but at higher levels of  $x_1$ , category 1 of  $x_2$  has higher values of  $y$  relative to category 0 of  $x_2$ . Assuming that this effect is statistically significant, we can conclude that the effect of  $x_2$  depends on the level of  $x_1$ . Conversely, the slope for  $x_1$  is steeper (a stronger relationship) for category 1 of  $x_2$  than for category 0 of  $x_2$ . Graphical methods are key to understanding interaction effects. However, it is also possible to work through a series of regression equations, much like we did in the previous chapter in our discussion of dummy variables, by inserting different values for  $x_2$  to gain insight into the model. Note that the key difference is we now have more than one place where we need to insert values for the dummy variable.

If we set the value for  $x_2 = 1$ , we have the following regression equation:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3(x_1x_2) = b_0 + b_1x_1 + b_2(1) + b_3(x_11)$$

which reduces to:

$$y = b_0 + b_1x_1 + b_2 + b_3x_1$$

By rearranging our terms, we can rewrite the regression equation as:

$$y = (b_0 + b_2) + b_1x_1 + b_3x_1 = (b_0 + b_2) + (b_1 + b_3)x_1$$

As in the previous chapter, we see that when we focus our attention on the category of  $x_2$  with the value 1, the model intercept changes by the value of the coefficient for that variable (i.e.,  $b_2$ ). What is different in the above equation is that the effect of variable  $x_1$  is now the sum of two different regression coefficients: the original coefficient for  $x_1$  (i.e.,  $b_1$ ) and the coefficient for the interaction term (i.e.,  $b_3$ ).

How do we now interpret the effect of  $x_1$ ? After summing the two regression coefficients  $b_1$  and  $b_3$ , we would say that for cases that had a value of 1 on  $x_2$  (i.e., the cases were in group 1), for each one unit increase in  $x_1$ ,  $y$  is expected to change by  $b_1 + b_3$  units.

When we set the value for  $x_2 = 0$  (the reference category for our dummy variable—group 0), we now have the following regression equation:

$$y = b_0 + b_1x_1 + b_20 + b_3(x_10)$$

which reduces to:

$$y = b_0 + b_1x_1$$

This indicates that the model intercept ( $b_0$ ) and coefficient for  $b_1x_1$  represent the intercept for the reference category on  $x_2$  and the effect of  $x_1$  for cases in the reference category, respectively. Notice that the two categories of  $x_2$  have different slopes for  $x_1$ . For category 0, the slope is  $b_1$ , whereas for category 1, the slope is  $b_1 + b_3$ . Thus, the interaction term tells us how much more or less the regression slope differs for category 1 relative to category 0. For every one-unit increase in  $x_1$ , the difference in  $y$  between categories 0 and 1 increases by  $b_3$ . If this interaction term equaled 0, it would indicate that both categories have the same slope for  $x_1$ . It is also important to note that the main effects ( $b_1$  and  $b_2$ ) no longer reflect the overall or marginal effect of  $x_1$  and  $x_2$  but rather the effect when the other variable equals zero. That is,  $b_1$  is the effect of  $x_1$  when  $x_2$  equals zero and  $b_2$  is the effect of  $x_2$  when  $x_1$  equals zero. In many contexts, these may well be values outside of the range of one or both of the variables and as such, these main effects may be meaningless by themselves. We will make use of this fact later when we discuss centering of the independent variables. To make the example more concrete, suppose that after estimating this regression equation, we find the following results:

$$b_0 = 2.5$$

$$b_1 = 3.2$$

$$b_2 = 1.9$$

$$b_3 = 1.3$$

By inserting the values for the regression coefficients into the regression equation, we have the following:

$$y = 2.5 + 3.2x_1 + 1.9x_2 + 1.3x_1x_2$$

For  $x_2 = 1$ , we have the following:

$$\begin{aligned} y &= 2.5 + 3.2x_1 + 1.9(1) + 1.3x_1(1) \\ &= (2.5 + 1.9) + (3.2 + 1.3)x_1 \\ &= 4.4 + 4.5x_1 \end{aligned}$$

And for  $x_2 = 0$ , we have:

$$y = 2.5 + 3.2x_1 + 1.9(0) + 1.3x_1(0) = 2.5 + 3.2x_1$$

The interpretation of the effect of  $x_1$  is straightforward, but we need to make sure that we are clear about the group for which we are interpreting the effect of  $x_1$ . Thus, for  $x_2 = 1$ , for each one-unit increase in  $x_1$ , we expect  $y$  to increase by 4.5 units. When  $x_2 = 0$ , for each one-unit increase in  $x_1$ ,  $y$  is expected to increase by 3.2 units. Substantively, this type of result would allow a researcher to say that the effect of  $x_1$  varied across the groups measured in  $x_2$ . As a visual aid to understanding these results, we have presented the two regression lines in Fig. 3.6, where group  $x_2 = 0$  is represented by the solid line and group  $x_2 = 1$  is represented by the dashed line.

Up to this point, we have assumed that all of the coefficients are positive. Figure 3.7 presents several additional possibilities for various combinations of positive and negative values for  $b_1$  and  $b_3$  (we assumed that  $b_0$  and  $b_2$  were positive in each plot). (Keep in mind that  $b_1$  represents the effect of  $x_1$  for the reference category of the dummy variable and  $b_3$  represents the value of the interaction effect.) Panel (a) illustrates a hypothetical example when  $b_1$  is positive and  $b_3$  is negative. Notice that slope for category 1 is now less steep than for category 0, the opposite of what we saw for Fig. 3.6. Panels (b) and (c) illustrate possible patterns when  $b_1$  is negative and  $b_3$  is positive (Panel (b)) or negative (Panel (c)). Clearly, there are many other possibilities, but we wanted to provide a few illustrations for different patterns that researchers have had to address in their analyses.

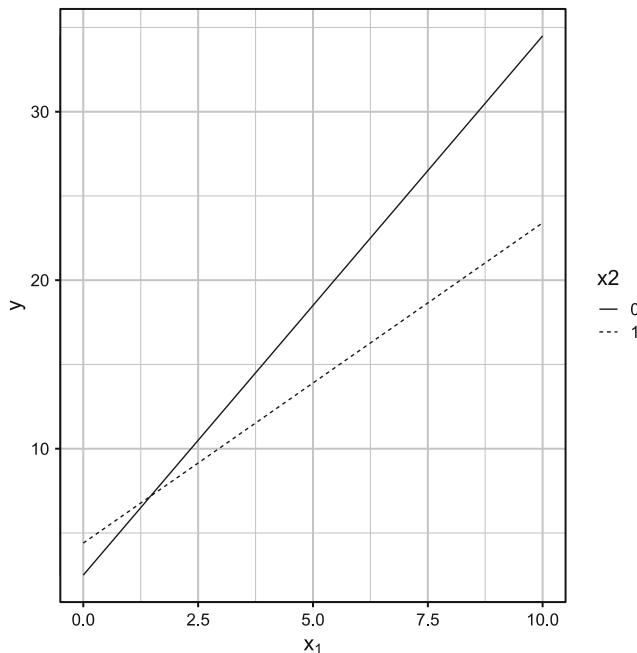
## An Example: Race and Punishment Severity

---

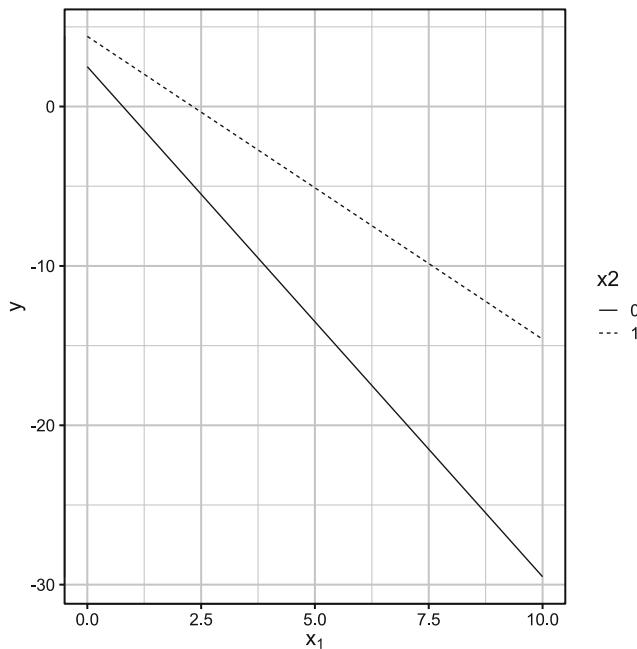
Suppose that we are interested in testing whether the severity of a criminal offense differentially affects the length of time offenders are sentenced to prison by race. Put another way, does the severity of the conviction offense affect the severity of punishment in the same way for offenders of different races? We again use data on the sentences of over 20,000 offenders sentenced to prison in Pennsylvania in 1998 to illustrate the test for an interaction effect between severity of offense and race of offender. To simplify our model here, we measure race as a dummy variable (0 = white, 1 = African American, with other race/ethnic groups dropped from the analysis). Offense severity is scored by the Pennsylvania Sentencing Commission and has values ranging from 1 to 14, and sentence length is measured in

**Figure 3.7**

*Hypothetical interaction effects for different combinations of positive and negative main effects and interaction effect*



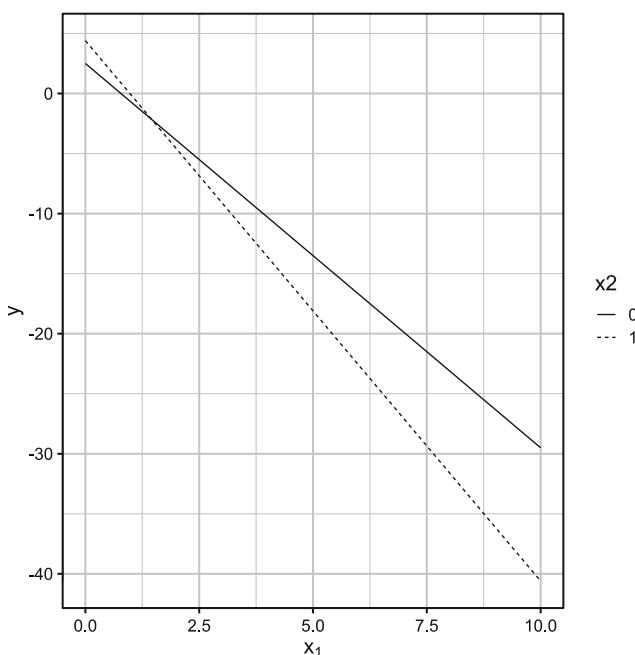
(a)  $b_1$  is positive and  $b_3$  is negative



(b)  $b_1$  is negative and  $b_3$  is positive

**Figure 3.7**

(continued)

(c)  $b_1$  and  $b_3$  are negative

months sentenced to prison. We will analyze the log-transformed sentence length because it is a positively valued scale that is strongly positively skewed and because it is reasonable to assume that the underlying data-generating mechanism is multiplicative rather than additive. The regression model we set out to test can be written as:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2$$

where  $y$  represents the natural log of sentence length in months,  $x_1$  represents offense severity, and  $x_2$  represents race.

When we estimate this regression model, we produce the following set of results:

$$y = 2.322 + 0.185x_1 - 0.115x_2 + 0.21x_1x_2$$

Using the same approach as above, we begin by focusing on African Americans ( $x_2 = 1$ ):

$$\begin{aligned}
 y &= 2.322 + 0.185x_1 - 0.115(1) + 0.021x_1(1) \\
 &= (2.322 - 0.115) + (0.185 + 0.021)x_1 \\
 &= 2.207 + 0.206x_1
 \end{aligned}$$

For whites, the equation is as follows:

$$\begin{aligned}
 y &= 2.322 + 0.185x_1 - 0.115(0) + 0.021x_1(0) \\
 &= 2.322 + 0.185x_1
 \end{aligned}$$

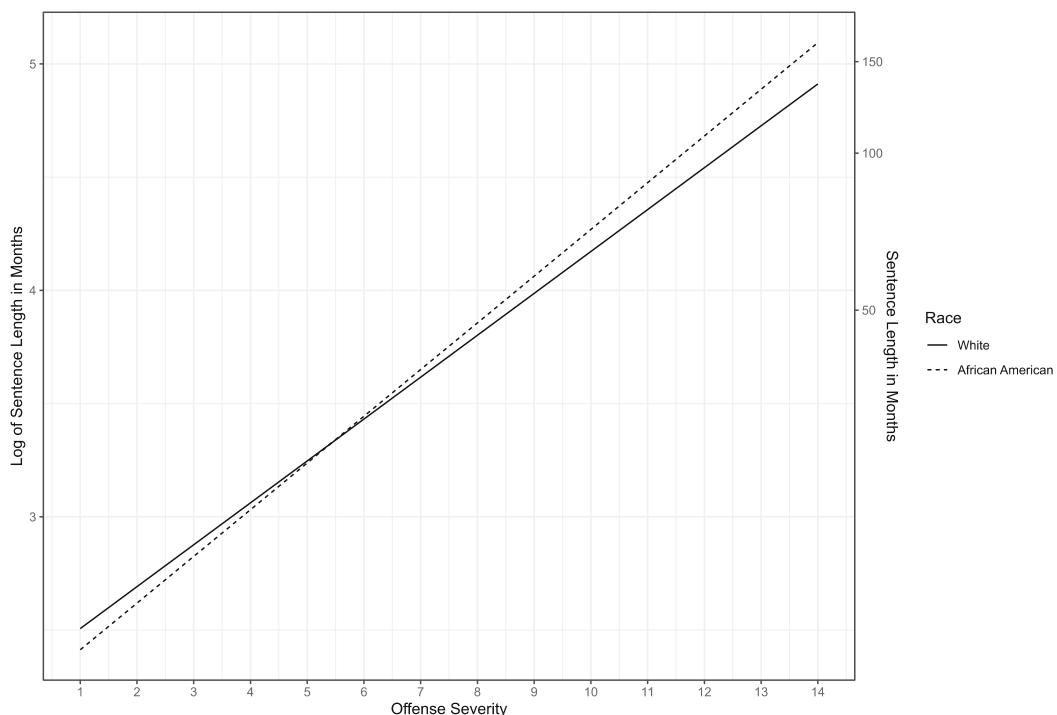
Substantively, we can now directly interpret the effect of offense severity for white and African American offenders separately. Among white offenders, each one-unit increase in offense severity is expected to increase sentence length by 0.185 logged months, while for African American offenders, each one-unit increase in offense severity is expected to increase sentence length by 0.206 logged months. More succinctly, these results suggest that the effect of offense severity on punishment severity is greater for African American offenders than for white offenders. These results are presented graphically in Fig. 3.8. The solid line reflects the slope for offense severity on sentence length for white offenders, while the dashed line reflects the effect for African American offenders. This effect is small compared to the rather large, and expected, effect of severity on sentence length (all effects are statistically significant but that is not surprising given the very large sample size). However, African Americans appear to get slightly longer sentence lengths when they are convicted of more serious crimes, relative to white offenders. Recall from the section on transformations, that we can convert these results into percentages to make them more easily interpretable. For white offenders, the 0.185 converts to a 20% increase in sentence length for every one-unit change in severity [ $(\exp(0.185) - 1) \times 100 = 20$ ] compared to a 23% increase for African Americans [ $(\exp(0.206) - 1) \times 100 = 23$ ]. At the high end of the scale, this is a meaningful difference in sentence lengths, as can be seen in the figure by looking at the right  $y$ -axis that shows the effect in months, not logged months.

### **Interaction Effects Between Two Scaled Variables**

Up to this point, our attention has been focused on interaction effects involving one scaled variable measured at the interval- or ratio-level of measurement and one dummy variable reflecting a binary nominal variable. The inclusion of an interaction effect between two scaled variables in a regression model is done in exactly the same way—we compute a product of the two variables and add the product to the regression

**Figure 3.8**

*Regression lines for the effect of offense severity on logged sentence length by race of offender*



equation. The interpretation of the interaction effect is much more complex, however, since we are no longer able to simplify the regression equation to represent the effect of one variable for two different groups.

A statistically significant interaction term between two measures indicates that their effects cannot be expressed by the additive effects of each measure in the model, but rather there is an additional effect that is measured by the interaction term. Stated differently, this means that the effect of one variable depends on the level of the other.

It may help to conceptualize this issue if we turn to a substantive example. Many sociologists have suggested that extra-legal variables such as income and social status impact upon sentencing outcomes (Black 1976). In a simple additive model, each of these factors would have some defined independent effect on the severity of a sentence measured in months of imprisonment. This model is illustrated in equation form below.

$$y = b_0 + b_1x_1 + b_2x_2$$

where  $y$  represents length of sentence (in months),  $x_1$  represents income, and  $x_2$  represents social status.

But what if the researcher believed that the effect of income and social status was not simply additive but also multiplicative, meaning that there was an added effect that was due to the interaction between the two. This theory is represented in the equation below:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2$$

where the terms are defined as above.

In this case, the researcher is hypothesizing that there is not only the independent effect of income and of social class, but that there is an additional interaction effect that is measured by multiplying income by social class. As we saw in the prior case with the interaction between a binary variable and a scaled variable, the additive or main effects are the localized effects when the other variable of the interaction is at zero. In our example, the regression coefficient for income ( $b_1$ ) would reflect the effect of income when social status is zero. Similarly, the regression coefficient for social status ( $b_2$ ) would reflect the effect of social status when income is zero. Both of these locations may well be outside the bounds of the observed data (i.e., there may be no cases with zero on both variables) and even if there are cases with the value of zero on both measures, this is unlikely to be a meaningful location for examining the main effect each independent variable. Essentially, there is no overall effect for each of these independent variables: The effect is conditional on the value of the other variable, and the regression coefficient represents this effect at a specific point on the other variable that is part of the interaction.

A (partial) solution to this problem is to center the two independent variables that make up the interaction term. Centering is simply subtracting the mean from each observation. This shifts the distribution for a variable such that it is centered around zero; that is, the mean of the centered distribution becomes zero. All other characteristics of the distribution remain the same. Furthermore, in a model with no interaction terms, the only effect this has on the regression is to shift the intercept to reflect the mean of  $y$ . The regression coefficients for other independent variables that are not part of the interaction effect also remain unchanged as does the fit of the model. Thus, it is always safe to center independent variables. The effect is either benign when there are no interaction terms or beneficial when there are. Using centered independent variables produces main effects that are at least somewhat interpretable in the context of an

interaction effect. In our example, the main effect of income ( $b_1$ ) reflects the effect of income on sentence length at the mean value for social status. Similarly, the main effect for social status ( $b_2$ ) would reflect the effect of social status for those with a mean income. In the case of a statistically significant interaction, however, these main effects are of limited value, as they may be quite different, and even in the opposite direction, at different values of the other variable. Thus, the interpretation should focus on how the effect of one of these variables changes over levels of the other, rather than focusing on the main effects of each of the variables.

Assume that our analysis produced the results shown in the equation below and that the interaction is statistically significant. What interpretation can the researcher draw? To illustrate this, we take a hypothetical example of regression results as reported below:

$$y = 7.2 - 2.4x_1 - 1.6x_2 - 1.1x_1x_2$$

The interaction term in this case suggests that there is an additional benefit beyond that of the additive independent effects of income and social status that must be taken into account. In a very simple interpretation, we can say that the effect of income ( $x_1$ ) decreases across levels of social status ( $x_2$ ) and that likewise, the effect of  $x_2$  (social status) decreases across levels of  $x_1$ . We can no longer, however, interpret the two main or additive effects of income and social status because the regression lines are no longer parallel. The effects are now conditional on the other variable. Thus, the value of 2.4 for the regression coefficient for income reflects the effect of income when social status equals zero. For all other values of social status, the effect of income on sentence length is either larger or smaller than this.

Conceptually, when we have an interaction between two scaled variables, we are testing the idea that the effect of one scaled variable varies by the level of the second scaled variable. For example, in the example noted above from general strain theory, the hypothesis is that the effect of strain varies by the level of self-esteem. In practice, the difficulty we often have in the interpretation of interaction effects between two scaled variables is in choosing values for one variable to represent the effect of the other variable. In effect, we have already done this in our example of an interaction between the dummy variable and the scaled variable. Recall that when we include an interaction effect between a dummy variable and a scaled variable, we set the value of the dummy variable to either 0 or 1 and then interpret the effect of the scaled variable for each group represented in the dummy variable.

In trying to determine how to interpret the interaction between two scaled variables, we would encourage you to first consider which variable

is of key importance for a study. The second variable would then be set at a limited number of values, which allows the researcher to see how the effect of the key variable changes across levels of the second variable. For example, if we again refer to the interaction between strain and self-esteem, the key theoretical variable is strain. Following these guidelines, we would then want to interpret the effect of strain for several designated values of self-esteem. Clearly, we could interpret the interaction effect the other way: The effect of self-esteem for specified levels of strain, but this is not a key piece of the theory.

What values do we use for the second scaled variable? For any given scaled variable, there may be hundreds or thousands of realistic possible values that we could use. We think that a useful place to start is to use the mean, one standard deviation above and below the mean, and two standard deviations above and below the mean. This will cover a wide range of possible values of the variable we are using and should be ample for understanding how our key variable changes across values of the second variable. In other cases, where there may be meaningful values on the second independent variable that have more intuitive meaning to the reader, these values should be used. For example, if we were to fix years of education, we might use 8, 12, and 16 to reflect the completion of junior high school, high school, and undergraduate collegiate education, respectively.

For example, suppose that we have estimated a regression model with two interval-level variables  $x_1$  and  $x_2$  and the interaction of  $x_1$  and  $x_2$ :

$$y = 2.3 + 1.7x_1 + 2.0x_2 + 0.5x_1x_2$$

For the purpose of this example, we will consider  $x_1$  the key variable. We find the mean and standard deviation of  $x_2$  to be 3.2 and 1.2, respectively.

The values that are one or two standard deviations above and below the mean of  $x_2$  are as follows:

$$\text{Two standard deviations above: } 3.2 + 2(1.2) = 3.2 + 2.4 = 5.6$$

$$\text{One standard deviations above: } 3.2 + 1(1.2) = 3.2 + 1.2 = 4.4$$

$$\text{One standard deviations below: } 3.2 - 1(1.2) = 3.2 - 1.2 = 2.0$$

$$\text{Two standard deviations below: } 3.2 - 2(1.2) = 3.2 - 2.4 = 0.8$$

We can now input these values for  $x_2$  to determine the effect of  $x_1$  on  $y$  at that level of  $x_2$ . The effect of  $x_1$  at the mean of  $x_2$ :

$$\begin{aligned}
 y &= 2.3 + 1.7x_1 + 2.0(3.2) + 0.5(x_13.2) \\
 &= 2.3 + 1.7x_1 + 6.4 + 1.6x_1 \\
 &= (2.3 + 6.4) + (1.7 + 1.6)x_1 \\
 &= 8.7 + 3.3x_1
 \end{aligned}$$

If we wanted to interpret the effect of  $x_1$  directly, then we would state that at the mean for  $x_2$ , each one-unit increase in  $x_1$  is expected to increase  $y$  by 3.3 units.

Effect of  $x_1$  at one standard deviation above the mean of  $x_2$ :

$$\begin{aligned}
 y &= 2.3 + 1.7x_1 + 1.8(4.4) + 0.5(x_14.4) \\
 &= 2.3 + 1.7x_1 + 7.92 + 2.2x_1 \\
 &= (2.3 + 7.92) + (1.7x_1 + 2.2x_1) \\
 &= 10.22 + 3.9x_1
 \end{aligned}$$

If we wanted to interpret the effect of  $x_1$  directly, then we would state that at one standard deviation above the mean for  $x_2$ , each one-unit increase in  $x_1$  is expected to increase  $y$  by 3.9 units.

Effect of  $x_1$  at one standard deviation below the mean of  $x_2$ :

$$\begin{aligned}
 y &= 2.3 + 1.7x_1 + 2.0(2.0) + 0.5(x_12.0) \\
 &= 2.3 + 1.7x_1 + 4.0 + 1.0x_1 \\
 &= (2.3 + 4.0) + (1.7 + 1.0)x_1 \\
 &= 6.3 + 2.7x_1
 \end{aligned}$$

If we wanted to interpret the effect of  $x_1$  directly, then we would state that at one standard deviation below the mean for  $x_2$ , each one-unit increase in  $x_1$  is expected to increase  $y$  by 2.7 units.

Effect of  $x_1$  at two standard deviations above the mean of  $x_2$ :

$$\begin{aligned}
 y &= 2.3 + 1.7x_1 + 2.0(5.6) + 0.5(x_15.6) \\
 &= 2.3 + 1.7x_1 + 11.2 + 2.8x_1 \\
 &= (2.3 + 11.2) + (1.7 + 2.8)x_1 \\
 &= 13.5 + 4.5x_1
 \end{aligned}$$

If we wanted to interpret the effect of  $x_1$  directly, then we would state that at two standard deviations above the mean for  $x_2$ , each one-unit increase in  $x_1$  is expected to increase  $y$  by 4.5 units.

Effect of  $x_1$  at two standard deviations below the mean of  $x_2$ :

$$\begin{aligned}y &= 2.3 + 1.7x_1 + 2.0(0.8) + 0.5(x_1 - 0.8) \\&= 2.3 + 1.7x_1 + 1.6 + 0.4x_1 \\&= (2.3 + 1.6) + (1.7 + 0.4)x_1 \\&= 3.9 + 2.1x_1\end{aligned}$$

If we wanted to interpret the effect of  $x_1$  directly, then we would state that at two standard deviations below the mean for  $x_2$ , each one-unit increase in  $x_1$  is expected to increase  $y$  by 2.1 units.

Aside from making direct interpretations of the effect of  $x_1$  at these five values of  $x_2$ , what we see is that as the value of  $x_2$  increases, the *effect* of  $x_1$  on  $y$  increases.

We can also approach this graphically by plotting the regression line for each of these levels of self-esteem ( $x_2$ ) across values of strain. Figure 3.9 shows this graph. This graph shows that the positive effect of  $x_1$  on  $y$  increases with increases levels of  $x_2$ . This effect, however, is rather modest compared to the strong positive effects of both  $x_1$  and  $x_2$ .

## An Example: Punishment Severity

---

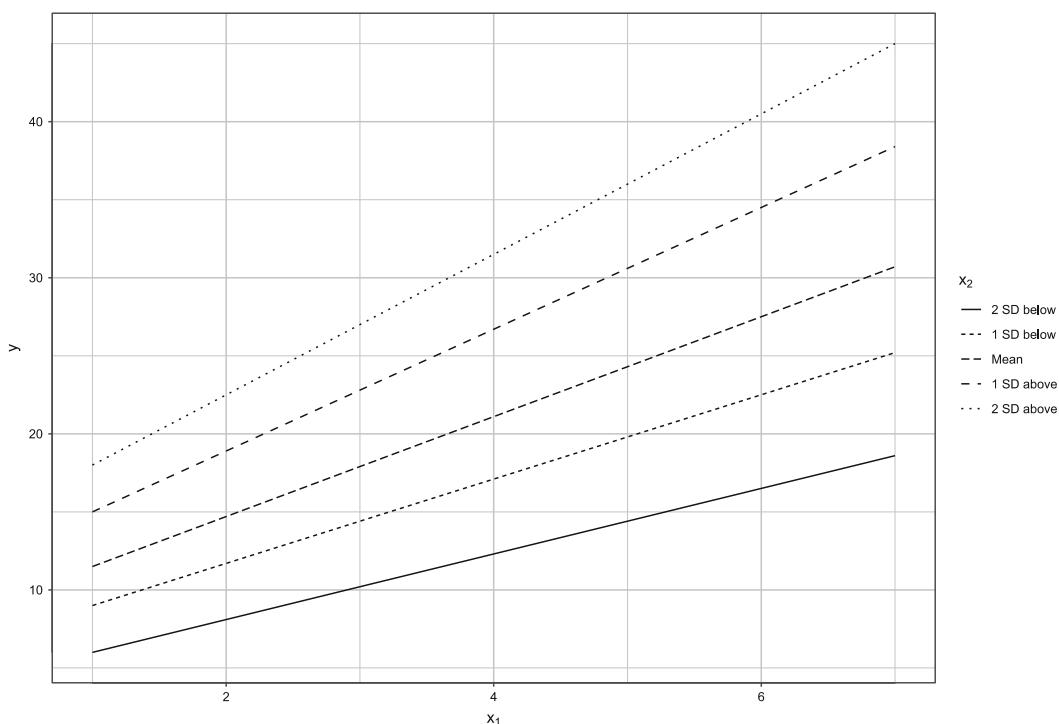
We again use the data on the sentencing of offenders in Pennsylvania in 1998 and modify our regression model slightly. We continue to use logged length of sentence as the dependent variable and severity of the offense as an independent variable. Our second independent variable is a prior record score that is computed by the Pennsylvania Sentencing Commission and can take on values ranging from 0 to 8 and reflects the criminal history of the offender; larger values for prior criminal history reflect both a greater number of prior offenses and more serious prior offenses. For the purposes of this example, we have added an interaction effect between severity of the offense and prior criminal history and are interested in how the effect of offense severity varies across levels of prior criminal history.

After estimating this model, we obtain the following regression equation:

$$y = 2.325 + 0.159x_1 - 0.006x_2 + 0.014x_1x_2$$

**Figure 3.9**

*Regression lines illustrating the interaction effect of  $x_1$  and  $x_2$  on  $y$*



where  $y$  represents length of prison sentence (in logged months),  $x_1$  represents offense severity, and  $x_2$  represents prior criminal history. We have graphed this relationship for three levels of the prior record score, 0, 1.68, and 8. These represent the minimum, mean, and maximum values for this scale. Following the same procedure as in our hypothetical example, we calculate the effect (slope) of offense severity for each of these three values of the prior record score.

The effect of offense severity at a prior record score of 0 is as follows:

$$\begin{aligned} y &= 2.325 + 0.159x_1 - 0.006(0) + 0.014x_1(0) \\ &= 2.325 + 0.159x_1 \end{aligned}$$

The effect of offense severity at a prior record score of 1.68 (the mean) is as follows:

$$\begin{aligned}
 y &= 2.325 + 0.159x_1 - 0.006(1.68) + 0.014x_1(1.68) \\
 &= 2.325 + 0.159x_1 - 0.010 + 0.024x_1 \\
 &= (2.325 - 0.010) + (0.159 + 0.024)x_1 \\
 &= 2.315 + 0.183x_1
 \end{aligned}$$

The effect of offense severity at a prior record score of 8 is as follows:

$$\begin{aligned}
 y &= 2.325 + 0.159x_1 - 0.006(8) + 0.014x_1(8) \\
 &= 2.325 + 0.159x_1 - 0.048 + 0.112x_1 \\
 &= (2.325 - 0.048) + (0.159 + 0.112)x_1 \\
 &= 2.277 + 0.271x_1
 \end{aligned}$$

The three values for the regression coefficient for  $x_1$  reflect the slope of the effect of offense severity at these three selected values of the prior record score. We can see that the slope increases slightly when the prior record score changes from 0 to 1.68 but considerably when this score is at its maximum value of 8. This can clearly be seen in Fig. 3.10.

The negative regression coefficient for the main effect of prior record score may seem counterintuitive as the effect appears to be positive in the figure (i.e., the line for a prior record score of 8 is higher than it is when the prior record score is 0). The regression coefficient is also very small, but the effect appears to be large. What is going on here? Recall that in a model with an interaction term, the main effects are conditional. Thus, this value reflects the effect of the prior record score on sentence length when the offense severity equals 0. The lowest possible value for the offense severity score is 1. If on the figure we extend the regression lines to the left to the zero intercept, we would see that the lines would cross each other ever so slightly. Thus, when the offense severity score is 0, the effect of prior record score is slightly negative. This value by itself is uninteresting. Within the range of our data, the effect of the prior record score is positive and increases with increasing offense severity. Those who committed a serious offense and had a more extensive prior history get the longest sentences.

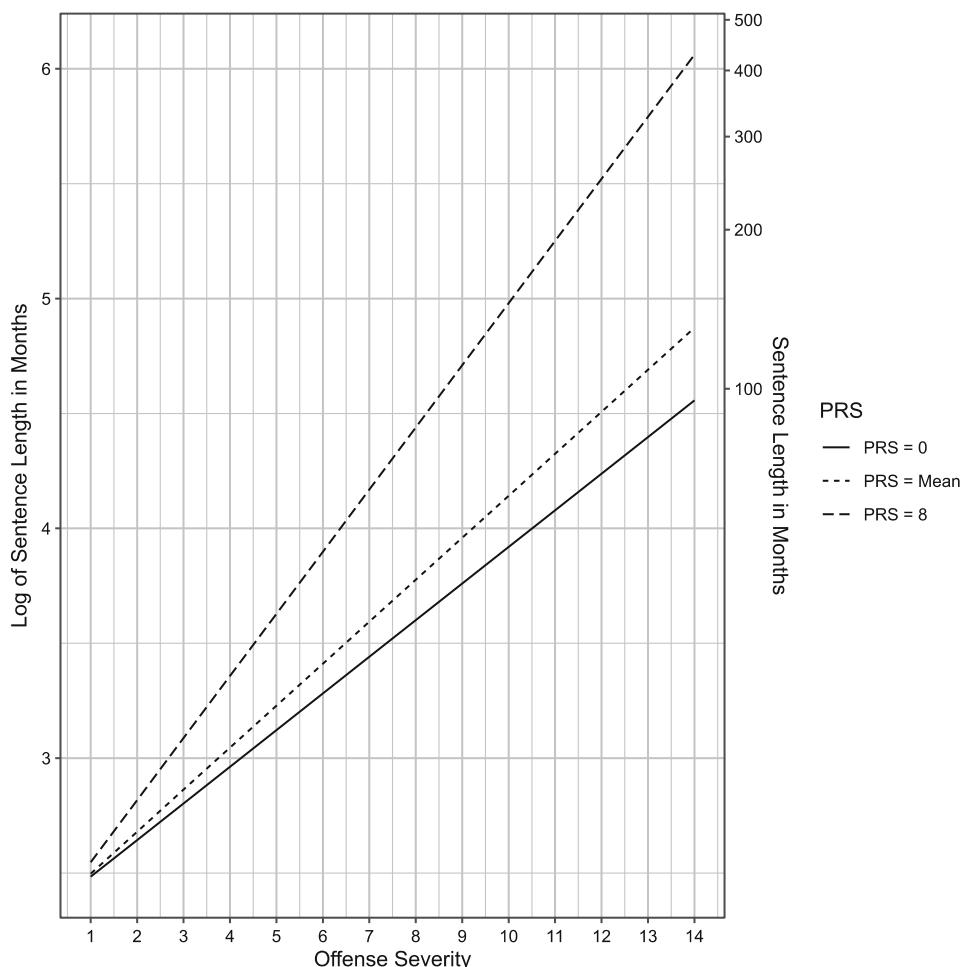
What would happen if we centered both the prior record score and the offense severity score? The model with centered independent variables is shown below.

$$y = 3.015 + 0.183x_1 + 0.047x_2 + 0.014x_1x_2$$

First, notice that the regression coefficient for the interaction remains unchanged and equals 0.014. Second, notice that the main effect for prior record score ( $x_2$ ) is now positive. It is also statistically significant in this

**Figure 3.10**

*Regression lines for the effect of offense severity on logged sentence length by level of prior record score (criminal history)*



model, whereas it was not statistically significant in the prior model. This coefficient now reflects the effect of a one-unit increase in the prior record score when the offense severity is at its mean value of 3.83. This is a more sensible indicator of the effect of the prior record score, but the more interesting and meaningful story has to do with the interaction effect. This also illustrates that great care should be taken if you draw any conclusions regarding the main effects. Many books recommend ignoring them in the presence of an interaction.

## The Problem of Multicollinearity

---

In criminal justice, the independent variables examined are generally multicollinear, or correlated with one another. Indeed, this correlation is one reason it is so important to use multiple regression techniques in criminal justice research. When variables are intercorrelated, it is important to control for the potential confounding influences of one variable on the other. Failure to do so is likely to lead to bias in our estimates of the effects of specific regression coefficients. However, the irony of multicollinearity is that when variables become too correlated, or highly multicollinear, the regression estimates become unstable.

**Multicollinearity** can be identified in one of two ways. A common method is to look at the intercorrelations among the independent variables included in your model. Very high correlations between independent variables are likely to lead to multicollinearity problems. What is considered a very high correlation? As with many other definitions in statistics, there is no absolute number at which multicollinearity is considered serious. As a general rule, a correlation between two independent variables greater than 0.80 should be seen as a warning that serious multicollinearity may be evident in your model.

Multicollinearity between two variables occurs less often than multicollinearity across a series of variables. To diagnose this type of multicollinearity, we use a statistic called **tolerance**. Tolerance measures the extent of the intercorrelations of each independent variable with all other independent variables. It is defined as 1 minus the percent of variance in  $x$  explained by the other independent variable examined (Eq. 3.1).

$$\text{Tolerance} = 1 - R_x^2$$

**Equation 3.1**

Calculation of tolerance is generally provided as an option in standard statistical computing packages, but it also can be calculated by taking each independent variable as the dependent variable in a regression that includes all other independent variables (but not the dependent variable). This value is then subtracted from 1. For example, let us say we defined a model for explaining sentence length among offenders in Pennsylvania that included three independent variables:

$$\begin{aligned}y_{length} &= b_0 + b_1(\text{age}) + b_2(\text{offense severity}) \\&+ b_3(\text{prior criminal history}) + e\end{aligned}$$

**Table 3.5**

Tolerance statistics for regression of sentence length for Pennsylvania offenders

INDEPENDENT VARIABLE	TOLERANCE	VIF
Age	0.940	1.064
Offense severity	0.931	1.074
Prior criminal history	0.975	1.026

The  $R_x^2$  for age would be estimated by calculating a regression in which age was the dependent variable, and offense severity and prior criminal history were the independent variables. You would then take this  $R^2$  and subtract it from 1. Similarly, to get  $R_x^2$  for offense severity, you would regress age and prior criminal history on offense severity and then subtract the resulting  $R^2$  from 1. Table 3.5 presents the tolerances and variance inflation factors (described below) for this regression model.

How do we know if multicollinearity is negatively affecting our model estimates based on the tolerance statistics? A very small tolerance statistic suggests that the model is likely to include a high level of multicollinearity. Again, there is no clear yardstick for defining a level of tolerance that is likely to lead to estimation problems. In general, however, a tolerance level of less than 0.20 should be taken as a warning that serious multicollinearity may exist in your model. We see from the results in Table 3.5 that the smallest tolerance statistic has a value of 0.94, which does not indicate any serious multicollinearity in this regression model.

Another approach commonly used to measure multicollinearity is what is called the **variance inflation factor (VIF)**. The VIF is the inverse of tolerance. By taking the inverse, higher values reflect greater multicollinearity, which is more intuitive than smaller values being problematic, as with tolerance. VIF is computed using the following formula.

$$\text{VIF} = \frac{1}{1 - R_x^2} = \frac{1}{\text{Tolerance}} \quad \text{Equation 3.2}$$

The VIF values for our example are also shown in Table 3.5. VIF values range from a low of 1 (no multicollinearity) to infinity. However, values larger than five should raise concern. This is the same criteria as with tolerance, given that  $1/0.20 = 5$ . For age, VIF is 1.06, suggesting that there is little collinearity between age and the other variables in the model.

A VIF value greater than 10 or a tolerance less than 0.10 suggests very high multicollinearity. But even with VIF values between 5 and 10, and tolerance levels ranging between .10 and .20, the model may not be affected greatly, although it can produce confusing effects on the regression coefficients for the variables involved. For example, if two variables are highly multicollinear and both are included in a model, one might be

statistically significant and the other not. Furthermore, one may have a coefficient that is in the opposite direction relative to its bivariate relationship with the dependent variable.

If there is a strong theoretical reason for including a variable, you may still opt to do that. However, you should make sure that the inclusion of the measure does not create instability in the model overall, as just described. This can be examined by running your model with and without the variable of interest and seeing whether the other variables in the model remain relatively stable in terms of their impact and significance levels. Large changes in this situation would be likely due to multicollinearity.

Beyond these diagnostic procedures for multicollinearity, there are warning signs that can be observed in the regressions that are estimated. Sometimes when multicollinearity is present, the percent of explained variance in a model is high, but the regression coefficients overall fail to reach conventional thresholds of statistical significance. Sometimes multicollinearity inflates coefficients to unrealistic sizes or produces coefficients in a direction contrary to conventional wisdom. One problem in diagnosing multicollinearity is that it may have such varied effects in your model that you may have difficulty distinguishing a misleading result that is due to multicollinearity from one that represents a new and interesting finding.

When there are indications of serious multicollinearity, you can take a number of alternative corrective measures. If interaction terms are multicollinear, it means that the two variables are too highly correlated to produce any meaningful interaction and the interaction term should be dropped from the model. Never keep the interaction term and drop one or both of the lower order main effects. This will produce an uninterpretable model. In the case where a small group of measures are highly collinear, you might choose to exclude the one variable that appears to present the most serious problem (e.g., that has the lowest tolerance value). The drawback of this approach is that the exclusion of such measures might lead to model misspecification and may result in biased estimates of other regression coefficients that remain in the model. This approach makes sense only when other variables that remain in the model measure the same concept. An approach that achieves a similar result, without excluding specific measures, is to create a new index or scale from clusters of variables that are multicollinear. For example, if a group of measures all relating to social status are multicollinear, you may decide to create a new composite measure defined as social status and use it as an independent variable in subsequent regressions. This could either be a simple summation or average of the multiple items, assuming they share a common scale, or some other logically created index. Alternatively, principle components or factor analysis could be used to create a composite scale of multicollinear items. Any up-to-date psychometrics textbook will have details on various methods of creating scales from multiple related items.

## Chapter Summary

---

This chapter addressed several topics related to fitting OLS multiple regression models. These were the following: how to include nominal independent variables with three or more categories into a model, how to estimate nonlinear relationships, how and why to transform variables, how to test for interactions between two independent variables, and how to assess for multicollinearity.

Nominal independent variables with three or more categories can be included in a regression model by creating a series of dummy variables using **dummy coding**. Each dummy variable codes whether an observation is a member of a given category of the nominal variable. You always include one fewer dummy variable than the number of categories of the nominal variable. This omitted category becomes the reference category. The regression coefficients for the dummy variables in the model indicate the predicted difference between the reference category and the category represented by the dummy variable, controlling for any other independent variables in the model.

**Nonlinear relationships** refer to the effect of the independent variable on the dependent variable not being a straight-line (linear) relationship. A linear relationship implies that each one-unit increase in the independent variable will result in the dependent variable increasing or decreasing by some fixed amount, regardless of the level of the independent variable. A nonlinear relationship implies that each one-unit increase in the independent variable does not result in the same amount of change in the dependent variable—it may be larger or smaller and will vary by the level of the independent variable. There are two main approaches to addressing a nonlinearity. The first is adding quadratic terms, such as a squared or cubed version of an independent variable into the model. The second is **transformation**—either the dependent variable or an independent variable, or both can be transformed. Transformations can also be useful to improve the normality of one or more of the distributions.

**Interaction effects** reflect the varying effect of one independent variable on the dependent variable across the levels or values of a second independent variable. When we have an interaction effect between a dummy variable and a scaled variable, we can directly interpret the effect of the scaled variable on the dependent variable for each group measured by the dummy variable. Interpretation of an interaction of two interval-level independent variables is much more difficult. One way of simplifying interpretation is to designate values for one variable, such as the mean, one standard deviation above/below the mean, and two standard deviations above/below the mean, as fixed points to compute the effect of the other scaled variable on the dependent variable. This can either be done

mathematically, producing the slope of one independent variable for the specified levels of the other, or graphically.

**Multicollinearity** occurs when independent variables in a regression model are too strongly related. It leads to unstable results. The problem may be diagnosed by checking the bivariate correlations between the variables and by measuring **tolerance** or **variance inflation factor (VIF)**. Multicollinearity may be dealt with either by excluding specific variables altogether or by merging several similar variables into one composite index or scale.

## Key Terms

---

**Dummy coding** A method for including nominal variables in a regression model that involves the creation of a series of binary 0/1 variables that indicate membership of an observation in each category of the nominal variable.

**Interaction effect** An interaction effect is present when the effect of one independent variable on the dependent variable is conditional on the level of a second independent variable.

**Multicollinearity** Condition in a multiple regression model in which independent variables examined are very strongly intercorrelated. Multicollinearity leads to unstable regression coefficients.

**Nonlinear relationship** Relationship between the dependent and the independent variable that is not captured by a straight-line (linear) relationship.

**Tolerance** A measure of the extent of the intercorrelations of each independent variable with all other independent variables. Tolerance may be used to test for multicollinearity in a multiple regression model.

**Transformation** Dependent and independent variables can be mathematically transformed, such as by taking the natural logarithm, squaring, and taking the square root. This can improve normality and/or be used to fit a nonlinear relationship.

**Variance inflation factor (VIF)** A measure of the extent to which a variable of interest is highly intercorrelated with other variables in the regression equation. The VIF is used to test for multicollinearity in a regression equation and is the inverse of tolerance.

## Symbols and Formulas

---

$$R_x^2 \quad R^2 \text{ obtained when an independent variable is treated as a dependent variable in a test for multicollinearity}$$

To calculate tolerance:

$$1 - R_x^2$$

To calculate the VIF (variance inflation factor):

$$\frac{1}{1 - R_x^2}$$

## Exercises

---

- 3.1. An analysis of shoplifting frequency among youth and young adults included a quadratic term for age of the individual and produced the following results:

INDEPENDENT VARIABLE	<i>b</i>
Age (in years)	0.35
Age <sup>2</sup> (in years <sup>2</sup> )	-0.01

Interpret the effect of age on the frequency of shoplifting.

- 3.2. An analysis linking level of informal social control to frequency of delinquency produced the following results:

INDEPENDENT VARIABLE	<i>b</i>
Age (in years)	-0.12
Sex (1 = female, 0 = male)	-1.50
Race (1 = white, 0 = nonwhite)	0.27
Informal social control (1 = low, 10 = high)	-0.83

After plotting the mean level of delinquency by level of informal social control, the researcher observed what appeared to be an inverse relationship ( $1/x$ ) between delinquency and informal social control. After transforming the measure of informal social control, the researcher estimated a new regression and produced the following results:

INDEPENDENT VARIABLE	<i>b</i>
Age (in years)	-0.11
Sex (1 = female, 0 = male)	-1.61
Race (1 = white, 0 = nonwhite)	0.32
Inverse of informal social control (1 = low, 10 = high)	2.45

- (a) Interpret the effect of the inverse of informal social control.
- (b) Sketch the relationship between delinquency and informal social control using the coefficient for the inverse of informal social control.
- 3.3. A researcher wanted to test the hypothesis that adolescent females were more affected by parental supervision than adolescent males. In a regression analysis incorporating an interaction effect between sex and supervision, the researcher produced the following set of results:

INDEPENDENT VARIABLE	<i>b</i>
Sex (1 = female, 0 = male)	-2.7
Supervision (1 = low, 10 = high)	-1.3
Sex * supervision	-0.5

Interpret the effect of supervision for adolescent females and males.

- 3.4. A study of attitudes about punishment used a scale of punitiveness ranging in value from 1 (Low) to 10 (High). The researcher was particularly interested in whether there was an interaction effect between age and political conservatism. A regression analysis produced the following results:

INDEPENDENT VARIABLE	<i>b</i>	MEAN
Age (years)	1.67	44.95
Political conservatism (1 = low, 10 = high)	0.92	6.50
Age * political conservatism	0.56	

- (a) What is the effect of political conservatism at the mean age of the sample? Interpret this effect.
- (b) What is the effect of age at the mean level of political conservatism for the sample? Interpret this effect.
- (c) What is the effect of political conservatism at each of the following ages?

- 20
- 30
- 50
- 60

Describe how the effect of political conservatism changes as age increases.

- (d) What is the effect of age at each of the following values of political conservatism?

–0  
–2  
–5  
–8  
–10

Describe how the effect of age changes as the level of political conservatism increases.

- 3.5. A study of violence in prison cell blocks was concerned about the amount of space available to each inmate and the proportion of inmates identified as gang members who had been identified as gang members. The researcher tested the hypothesis of an interaction effect between space available and the proportion of inmates identified as gang members. A regression analysis produced the following results:

INDEPENDENT VARIABLE	<i>b</i>	MEAN
Space available (square feet per inmate)	–0.25	10.00
Proportion gang members	0.77	0.77
Space available * proportion gang members	–0.05	

- (a) What is the effect of space available at the mean proportion of gang members for the sample of cell blocks? Interpret this effect.  
 (b) What is the effect of proportion of gang members at the mean level of space available for the sample of cell blocks? Interpret this effect.  
 (c) What is the effect of space available at each of the following proportions of gang membership?

–0.2  
–0.4  
–0.6  
–0.8

Describe how the effect of space available changes as proportion of gang membership increases.

- (d) What is the effect of proportion of gang membership at each of the following values of space available?

–3  
–6

–12

–15

Describe how the effect of proportion of gang membership changes as the level of space available increases.

- 3.6. Rachel collects police data on a series of burglaries and wishes to determine the factors that influence the amount of property stolen in each case. She creates a multiple regression model and runs a test of tolerance for each of the independent variables. Her results are as follows, where  $y$  = Amount of property stolen (\$):

INDEPENDENT VARIABLE	SCALE	TOLERANCE
$x_1$ : Time of robbery (AM or PM)	Nominal	0.98
$x_2$ : Accessibility of property	Ordinal	0.94
$x_3$ : Number of rooms in house	Ratio	0.12
$x_4$ : Size of house	Ratio	0.12
$x_5$ : Joint income of family	Ratio	0.46

Would you advise Rachel to make any changes to her model? Explain your answer.

- 3.7. A researcher examining neighborhood crime rates computes a regression model using the following variables:

$$\begin{aligned}y &= \text{crime rate (per 100,000)} \\x_1 &= \text{percent living in poverty} \\x_2 &= \text{percent unemployed} \\x_3 &= \text{median income} \\x_4 &= \text{percent of homes being rented}\end{aligned}$$

The researcher finds the  $F$ -statistic for the overall model to be statistically significant (with  $\alpha = 0.05$ ), but the results for each variable are as follows:

INDEPENDENT VARIABLE	b	SIG.	TOLERANCE
$x_1$ : Percent living in poverty	52.13	0.17	0.15
$x_2$ : Percent unemployed	39.95	0.23	0.07
$x_3$ : Median income	22.64	0.12	0.19
$x_4$ : Percent of homes being rented	27.89	0.33	0.05

- (a) Explain why the researcher found a statistically significant regression model, but no significant regression coefficients  
 (b) What would you recommend the researcher do in this case?

## Computer Exercises

In Chap. 2, we illustrated the use of the regression command in SPSS, Stata, and R to estimate multiple regression models. The analyses described in this chapter—nonlinear terms, interaction effects, and a test for multicollinearity—are accomplished with the same regression command in SPSS and Stata. The following exercises should help to illustrate how to perform these analyses in SPSS, Stata, and R, as will the sample syntax files for SPSS (Chapter\_3.sps) and Stata (Chapter\_3.do).

### SPSS

#### *Dummy Coding Nominal Variables*

In SPSS version 22 or later, there is a tool to automatically create dummy variables for a given variable. You will find this tool by going to *Transform > Create dummy variables* in the dropdown menus. If you have an SPSS version prior to 22, you will need to rely on the tool to recode variables, which can be found by going to *Transform > Recode into Different variables*. The code for recoding into different variables is below. Here, we create a dummy variable for one of the categories from the *race\_eth* variable, where white is coded as a 1.

```
RECODE race_eth (1=Copy) (ELSE=0) INTO white.  
EXECUTE.
```

Note that you need to carry out the recoding procedure for each dummy code you need to create. For instance, here, we are creating a dummy variable named *black* from the category represented by the value 2 on the *race\_eth* variable.

```
RECODE race_eth (2=Copy) (ELSE=0) INTO black.  
EXECUTE.  
RECODE black (2=1).  
EXECUTE.
```

#### *Computing Nonlinear and Interaction Terms*

To include a nonlinear or an interaction term in a multiple regression model, it is necessary to first compute the nonlinear or the interaction term. This computation is done with the COMPUTE command. The general format for the COMPUTE command is the following:

```
COMPUTE new_var_name = calculation
```

The calculation can be a function of one or more variables, which we illustrate below.

#### *Nonlinear Terms*

Suppose we wanted to compute a squared term for a variable AGE. We might name this new variable AGE\_SQ. The COMPUTE command would appear as

```
COMPUTE AGE_SQ = AGE**2.  
EXECUTE.
```

In SPSS, the double asterisk (\*\*) indicates that we want to take a variable to a power. In this example, we want to square AGE, so we add the value 2. If, for some reason, we had wanted to cube AGE, then we would have typed AGE\*\*3. Also, recall that the addition of the EXECUTE command forces SPSS to perform this calculation immediately.

An alternative that accomplishes the same thing is

```
COMPUTE AGE_SQ = AGE * AGE.  
EXECUTE.
```

where we simply multiply the variable AGE by itself. In either case, once the COMPUTE command has been executed, the new variable will appear in the data file.

### *Interaction Terms*

The computation of an interaction term is as direct as the equations given in this chapter. We again use the COMPUTE command, but our calculation involves multiplying the two variables of interest. We have found that it is often helpful to make the name of the new variable representing an interaction term a combination of fragments from both of the original variables being used in the calculation.

For example, suppose that we want to create an interaction term for two variables that we have named EDUCATION and INCOME. We might call the interaction variable EDUC\_INC:

```
COMPUTE EDUC_INC=EDUCATION * INCOME.  
EXECUTE.
```

### *Estimating the Regression Model*

After computing the nonlinear or the interaction term, we then simply treat the created variable as an additional variable added to our multiple regression model—identical to how we presented these terms in the discussion in this chapter. For situations where we are using nonlinear terms, we may need to drop the original variable. Prior research and theory indicating that a nonlinear term was appropriate will often be the best guide on what the regression equation should look like. In the case of an interaction term, keep in mind that we must include both of the original variables and the interaction term; otherwise, it will be nearly impossible to interpret the coefficients that we do obtain from a regression analysis.

### *Collinearity Diagnostics*

SPSS's regression command will produce a wide assortment of collinearity statistics, including the tolerance statistic discussed above. To obtain the collinearity diagnostics, we include the /STATISTICS option in our REGRESSION command:

```
REGRESSION
/STATISTICS COEFF R ANOVA COLLIN TOL
/DEPENDENT dep_var_name
/METHOD = ENTER list_of_indep_vars.
```

where TOL requests the tolerance statistics for each independent variable and COLLIN requests all other collinearity measures. The tolerance statistics are presented in the coefficients table, where you will also find the regression coefficients. Recall from the discussion in the chapter that a tolerance of less than about 0.20 is indicative of collinearity problems in a regression model.

## **Stata**

### *Dummy Coding Nominal Variables*

You have the option to write syntax in Stata to create indicator variables by recording a given variable, but there is a way to do so automatically that is easier. If you are trying to create dummy variables for a string variable, you can use the **tabulate** and **generate** functions. For example, here we are automatically recoding the variable *race\_eth*, and we are creating dummy variables that start with the prefix *race* that will appear as the far-right columns in your dataset (e.g., *race1*, *race2*, *race3*). When doing so, you will notice that Stata provides a frequency table of the variable you are recoding.

```
tabulate race_eth, generate(race)
```

### *Computing Nonlinear and Interaction Terms*

To include a nonlinear or an interaction term in a multiple regression model, it is necessary to first compute the nonlinear or the interaction term. This computation is done with the **gen** command (short for **generate**). The general format for the **gen** command is

```
gen new_var_name = calculation
```

The calculation can be a function of one or more variables, which we illustrate below. (Note that the structure of the discussion is nearly identical to that in the SPSS section, with slight changes for the Stata syntax.)

### *Nonlinear Terms*

Suppose we wanted to compute a squared term for a variable AGE. We might name this new variable AGE\_SQ. The **gen** command would look like

**gen** AGE\_SQ = AGE<sup>2</sup>

In Stata, the upward pointing arrow (^) indicates that we want to take a variable to a power. In this example, we want to square AGE, so we add the value 2. If, for some reason, we had wanted to cube AGE, then we would have typed AGE<sup>3</sup>.

An alternative that accomplishes the same thing is

**gen** AGE\_SQ = AGE \* AGE

where we simply multiply the variable AGE by itself. In either case, once the **gen** command has been executed, the new variable will appear in the data file.

### *Interaction Terms*

The computation of an interaction term is as direct as the equations given in this chapter. We again use the **gen** command, but our calculation involves multiplying the two variables of interest. We have found that it is often helpful to make the name of the new variable representing an interaction term a combination of fragments from both of the original variables being used in the calculation.

For example, suppose that we want to create an interaction term for two variables that we have named EDUCATION and INCOME. We might call the interaction variable EDUC\_INC:

**gen** EDUC\_INC=EDUCATION \* INCOME

### *Estimating the Regression Model*

After computing the nonlinear or the interaction term, we then simply treat the created variable as an additional variable added to our multiple regression model—identical to how we presented these terms in the discussion in this chapter. For situations where we are using nonlinear terms, we may need to drop the original variable. Prior research and theory indicating that a nonlinear term was appropriate will often be the best guide on what the regression equation should look like. In the case of an interaction term, keep in mind that we must include both of the original variables and the interaction term; otherwise, it will be nearly impossible to interpret the coefficients that we do obtain from a regression analysis.

### *Collinearity Diagnostics*

Within Stata, the **vif** command provides collinearity diagnostics, including VIF and tolerance (1/VIF). It can be performed by running the **vif** command after the **regress** command:

**regress** dep\_var list\_of\_independent\_variables  
**vif**

Each measure of collinearity is presented in a table that lists each independent variable (by row) and collinearity statistic (VIF/tolerance) across columns.

Additionally, a correlation matrix of all model covariates can be conducted by using the **vce**, **corr** command after the **regress** command:

```
regress dep_var list_of_independent_variables  
vce, corr
```

## R

### Dummy Coding Nominal Variables

In R, there are packages available to automate the creation of dummy variables from a nominal variable (factor/character class type). One of those packages, which we are going to illustrate, is *fastDummies*. Install and load this package using `install.packages("fastDummies")` and `library(fastDummies)`. Once you have the package installed, you can use the **dummy\_cols()** function. In our example, we are automatically creating dummy variables for the variables *race\_ethc* and *job* in the *df* dataset. In doing so, we are creating a new dataset named *new\_df* (contains all variables in *df* plus the new dummy code variables in the far-right columns).

```
new_df<-dummy_cols(df,  
select_columns = c("race_eth", "job"))
```

### Computing Nonlinear and Interaction Terms

To compute a nonlinear or interaction term, we are going to rely on the assignment operator `<-` as follows:

```
dataset_name$new_var_name <- calculation
```

As in SPSS and Stata, the calculation can be a function of one or more variables, which we illustrate below.

### Nonlinear Terms

To take a variable to a higher power, we are able to use a few different operators such as the upward pointing arrow (`^`) or two asterisks (`**`). In the two examples below, we compute a squared term for the variable *AGE* and assign it as a new variable named *AGE\_SQ*:

```
dataset_name$AGE_SQ <- dataset_name$AGE^2  
dataset_name$AGE_SQ <- dataset_name$AGE **2
```

Or more simply, just use the multiplication operator (`*`) to multiply the variable *AGE* by itself.

```
dataset_name$AGE_SQ <-  
dataset_name$AGE * dataset_name$AGE
```

### *Interaction Terms*

To create an interaction term between two variables, we are going to rely again on the <- assignment (<-) and multiplication (\*) operators, but we are going to be multiplying to different variables together. For example, suppose that we want to create an interaction term named *EDUC\_INC* for two variables that we have named *EDUCATION* and *INCOME*.

```
dataset_name$EDUC_INC <-
  dataset_name$EDUCATION * dataset_name$INCOME
```

### *Estimating the Regression Model*

After computing the nonlinear or the interaction term, we then simply treat the created variable as an additional variable added to our multiple regression model—identical to how we presented these terms in the discussion in this chapter. For situations where we are using nonlinear terms, we may need to drop the original variable. Prior research and theory indicating that a nonlinear term was appropriate will often be the best guide on what the regression equation should look like. In the case of an interaction term, keep in mind that we must include both of the original variables and the interaction term; otherwise, it will be nearly impossible to interpret the coefficients that we do obtain from a regression analysis.

### *Collinearity Diagnostics*

The **ols\_vif\_tol()** function in the *olsrr* package provides the tolerance and variance inflation factor for a regression model conducted using the **lm()** function. This is done below by using the <- assignment operator. Then, immediately after running the **lm()** function, run the **ols\_vif\_tol()** function to obtain collinearity diagnostics as follows:

```
model <- lm(dep_var ~ ind_var1 + ind_var2,
              data = dataset_name)
ols_vif_tol(model)
```

But remember to install and load the *olsrr* package before using the **ols\_vif\_tol()** function.

### **Problems**

Open the NYS data file (*nys\_1.sav*, *nys\_1\_student.sav*, or *nys\_1.dta*) to complete Exercises 1 through 4.

1. Using one of the line graph commands, generate a plot for the mean of a delinquency measure by age.

In SPSS, this should look something like as follows:

```
GRAPH /LINE(SIMPLE) = MEAN(delinquency_variable)
BY age.
```

In Stata, the corresponding syntax is a bit more complicated, but use the following lines to accomplish a similar graph as in SPSS:

```
egen mean_delinq1 = mean(delinquency_variable),
by(age) sort age
twoway (line mn_delinq1 age, connect(ascending))
```

In R, you can use the *dplyr* package first to summarize the mean delinquency by age, and then use the *ggplot2* package to make a line graph:

```
Mean_df<- df %>%
  group_by(age) %>%
  summarize(mean_del = mean(delinquency_variable,
    na.rm = TRUE))
ggplot(data = Mean_df,
  aes(x = age, y = mean_del)) + geom_line()
```

The resulting line graphs from the syntax given above will plot the mean level of delinquency—for the variable you picked—for each age recorded in the NYS sample. Try this command with other measures of delinquency, and see if you notice any variations in the shapes of the lines. (NOTE: If you are using Stata, you will need to change the output variable name in the *egen* command line, perhaps most simply by changing the number at the end.)

Do they imply a linear relationship between age and delinquency? A nonlinear relationship? If nonlinear, what kind of nonlinear relationship? (You may want to refer back to Fig. 3.2 for some approximations of different curves.)

2. Compute a squared term for the age variable as described above. Estimate a multiple regression model using one of the measures of delinquency as the dependent variable. Use age and age squared as the independent variables. As noted earlier in the chapter, this will provide a quadratic equation that will test for a nonlinear relationship between age and the measure of delinquency that you have selected.  
Report the values of the regression coefficients for age and age squared and whether or not the coefficients are statistically significant (with  $\alpha = 0.05$ ). Interpret the effect of age on this measure of delinquency.
3. Compute a multiple regression model using number of times drunk as the dependent variable. Include the measures for age, sex, race (recoded as a dummy variable), and at least two other variables that you think are related to the number of times drunk. This will be the *baseline model* in the following questions.
  - (a) Compute the tolerance statistic for each of the independent variables in the baseline model. Does it appear that there is a problem with collinearity in the regression model? Explain why.

- (b) Compute an interaction term for sex with age. Add this term to the baseline model, and rerun the regression command. Is the effect of age on number of times drunk significantly different for males and females (i.e., is the interaction effect statistically significant)? If so, interpret the effect of age on the number of times drunk for males and females.
- Compute the tolerance statistic for each of the independent variables in this model. Does it appear that there is a problem with collinearity in this model? Explain why.
- (c) Compute an interaction term for race (which should be coded as a dummy variable) and age. Add this term to the baseline model, and rerun the regression command. (The interaction effect from part (a) should no longer be included in the analysis.) Is the effect of age on number of times drunk significantly different for these two race groups? If so, interpret the effect of age on the number of times drunk for each race group.
- Compute the tolerance statistic for each of the independent variables in this model. Does it appear that there is a problem with collinearity in this model? Explain why.
- (d) If one of the additional variables that you have added to your regression model is measured at the ratio level of measurement, compute an interaction term between this variable and either the sex or the race variable. Add this term to the baseline model (there should be no other interaction terms included in this analysis), and rerun the regression command. Is the effect of this variable on number of times drunk significantly different for the two groups? If so, interpret the effect of this variable for each group.
- (e) Compute the tolerance statistic for each of the independent variables in this model. Does it appear that there is a problem with collinearity in this model? Explain why.
4. Compute a multiple regression model using number of times cheated at school as the dependent variable. Include the measures for age, sex, race (recoded as a dummy variable), grade point average, and amount of time spent studying as the independent variables. This will be the baseline model in the following questions.
- (a) Compute an interaction term for grade point average and time spent studying, and add this term to the regression model. Prior to rerunning the regression command, check the item for descriptive statistics available through the regression command window. Report the coefficient values and whether or not the coefficients are statistically significant (with  $\alpha = 0.05$ ).

- (b) Compute the tolerance statistic for each of the independent variables in this model. Does it appear that there is a problem with collinearity in this model? Explain why.
- (c) What is the effect of grade point average on number of times cheated at the mean level of time spent studying? (You will need to use the mean reported in the results for part (a).) Interpret this effect.
- How does the effect of grade point average change as the value for time spent studying increases or decreases?
- (d) What is the effect of time spent studying on number of times cheated at the mean grade point average? Interpret this effect.
- How does the effect of time spent studying change as the value for grade point average increases or decreases?

## References

---

- Agnew, R. (1992). Foundation for a general strain theory of crime and delinquency. *Criminology*, 30(1), 47–88.
- Black, D. J. (1976). *The behavior or law*. New York: Academic Press.
- Bulmer, M. G. (1979). *Principles of statistics*. New York: Dover.
- Gelman, A., & Hill, J. (2016). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.
- Steffensmeier, D., Kramer, J., & Ulmer, J. (1995). Age differences in sentencing. *Justice Quarterly*, 12(3), 583–602.
- Tonry, M. H. (1997). *Sentencing matters*. Oxford: Oxford University Press.

## Chapter four

---

# Logistic Regression

### **Logistic Regression as a Tool for Examining a Dichotomous Dependent Variable**

---

Why is It Inappropriate to Use Ordinary Least Squares Regression for a Dichotomous Dependent Variable?

What Shape Does the Logistic Model Curve Take?

How is the Outcome Altered in a Logistic Regression Model?

### **Interpreting Logistic Regression Coefficients**

---

Why is It Difficult to Interpret the Logistic Regression Coefficient?

What is an Odds Ratio and How is It Interpreted?

What is the Derivative at Mean and How is It Interpreted?

### **Comparing Logistic Regression Coefficients Within a Single Model**

---

How Can Probability Estimates be Used to Compare the Strength of Logistic Regression Coefficients?

What is the Standardized Logistic Regression Coefficient and How is It Interpreted?

## **Evaluating the Logistic Regression Model**

---

How is the Percent of Correct Predictions Interpreted?

What is Pseudo- $R^2$  and How is It Interpreted?

## **Testing for Statistical Significance**

---

What is the Test of Statistical Significance for the Overall Logistic Regression Model?

What is the Test of Statistical Significance for the Logistic Regression Coefficient?

What is the Test of Statistical Significance for a Multicategory Nominal Variable?

# O

RDINARY LEAST SQUARES (OLS) REGRESSION is a very useful tool for identifying how one or a series of independent variables affects a continuously scaled dependent variable at the interval or ratio level of measurement. As noted in Chap. 2, this method may also be used—though with caution—to explain discrete dependent variables that are measured at an ordinal level. But what should the researcher do when faced with a binary or dichotomous dependent variable? Such situations are common in criminology and criminal justice. For example, in examining sentencing practices, the researcher may want to explain why certain defendants get a prison sentence while others do not. In assessing the success of a drug treatment program, the researcher may be interested in whether offenders failed a drug test or whether they returned to prison within a fixed follow-up period. In each of these examples, the variable that the researcher seeks to explain is a simple binary outcome. It is not appropriate to examine binary dependent variables using the regression methods that we have reviewed thus far.

This chapter introduces a type of regression analysis that allows us to examine a dichotomous dependent variable. Called logistic regression analysis, it has become one of the analysis tools most frequently used in crime and justice research. We begin the chapter by explaining why the OLS regression approach described in Chap. 2 is not appropriate when the dependent variable is binary. We then describe the logistic regression approach and the logic underlying it. Finally, we illustrate the interpretation of logistic regression statistics in the context of a substantive criminal justice research example. In this chapter, as in Chap. 2, our focus will be more on explaining how logistic regression can be used in research than on describing the mathematical properties that underlie the computations used to develop logistic regression statistics.

## Why Is It Inappropriate to Use OLS Regression for a Dichotomous Dependent Variable?

---

In Chap. 2, you saw that we could use not only interval-/ratio-level variables, but also ordinal- and nominal-level variables, as independent measures in a multiple ordinary least squares regression. While OLS regression assumes that the dependent variable is measured at an interval- or ratio-level of measurement and represents a continuous underlying construct, a researcher may sometimes reasonably decide to use an ordinal-level dependent variable. But applying the OLS regression approach is inappropriate when the dependent variable is dichotomous, as is the case with a binary (two-category) dependent variable.

Why do we state this rule so unequivocally? One reason is that the logic underlying our explanation of a dichotomous dependent variable is at odds with the models that we build using the OLS regression approach. In order to expand on this idea, we need to return to how predictions are developed using OLS regression. In the simple linear model—the OLS model—we predict the value of  $y$  based on an equation that takes into account the values of a  $y$ -intercept ( $b_0$ ) and one or a series of independent variables (e.g.,  $b_1x_1$  and  $b_2x_2$ ). This model is represented below for a simple (one independent variable) regression example in which we seek to explain the yearly budget of police departments based on the number of officers employed. The equation is:

$$y = b_0 + b_1x_1$$

where  $y$  is the yearly police department budget in dollars and  $x_1$  is the number of sworn officers.

This is an additive model in which we predict the value of  $y$ —in this case, the yearly police department budget in dollars—by adding the value of the  $y$ -intercept to the value of the regression coefficient times the value of  $x_1$  (the number of sworn officers in a department).

Let us say that a representative sample of police agencies was surveyed and analysis of the responses yielded the following regression equation:

$$y = 100,000 + 100,000x_1$$

This equation suggests that for each additional officer employed, the department budget is expected to increase by \$100,000. For a police agency with 100 officers, we would expect a budget of about \$10,100,000:

### Working It Out

$$\begin{aligned}y &= 100,000 + 100,000x_1 \\&= 100,000 + 100,000(100) \\&= 100,000 + 10,000,000 \\&= 10,100,000\end{aligned}$$

For a police agency with 1000 officers, we would expect a budget of about \$100,100,000:

### Working It Out

$$\begin{aligned}y &= 100,000 + 100,000x_1 \\&= 100,000 + 100,000(1,000) \\&= 100,000 + 100,000,000 \\&= 100,100,000\end{aligned}$$

This model, like other OLS regression models, assumes that there is no real limit to the value that the dependent variable can attain. With each additional officer comes an expected increase in the departmental budget. Our model suggests that the increase is about \$100,000 for each additional officer. While this logic makes very good sense when we are speaking about interval- and ratio-scale measures, such as the budget of a police agency, does it make sense when we are dealing with a dichotomous dependent variable, such as whether a parolee has failed a drug test?

**Table 4.1**

Drug testing results and prior drug arrests for 30 parolees

<b>DRUG TEST RESULTS</b>	<b>DRUG TEST SCORE</b>	<b>NUMBER OF DRUG ARRESTS</b>
Pass	0	0
Fail	1	0
Pass	0	1
Fail	1	2
Fail	1	3
Fail	1	3
Pass	0	3
Fail	1	4
Fail	1	4
Fail	1	4
Fail	1	5
Fail	1	5
Fail	1	5
Fail	1	6
Fail	1	6
Fail	1	7
Fail	1	8

**Table 4.2**

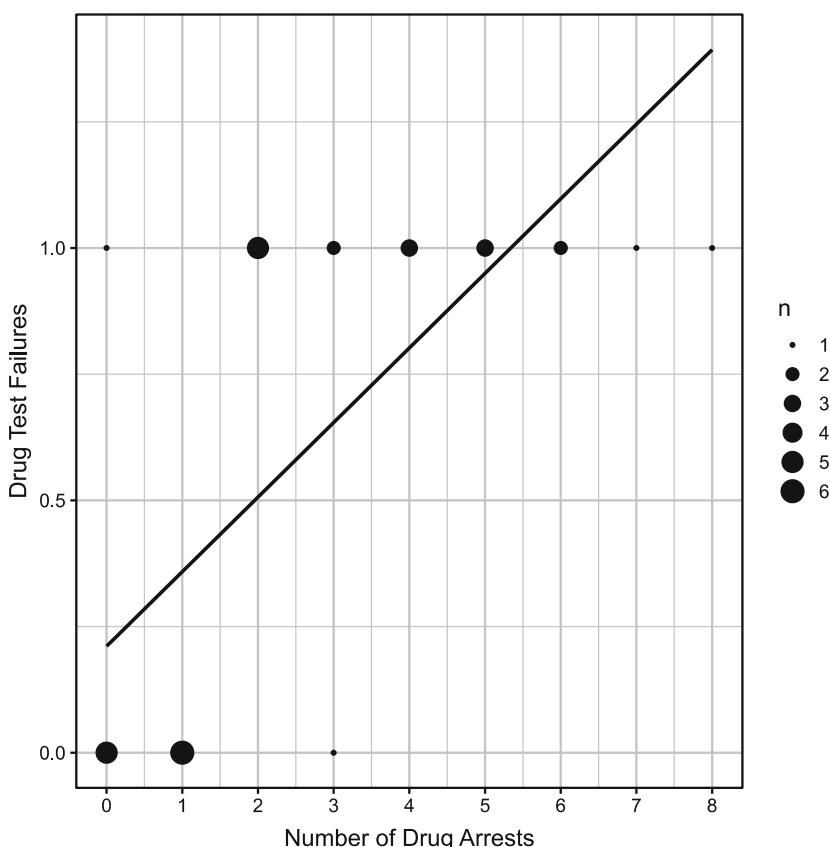
OLS regression for drug testing as the dependent variable and prior drug arrests as the independent variable

	<b>UNSTANDARDIZED COEFFICIENTS (b)</b>	<b>SE</b>	<b>STANDARDIZED COEFFICIENTS (BETA)</b>	<b>t</b>	<b>p</b>
Intercept	0.211	0.104		2.025	.052
Drug arrests	0.148	0.030	0.681	4.921	.000

Let us say that we surveyed 30 parolees who had been tested for drug use. Our independent variable is prior drug arrests. The data in Table 4.1 report whether a parolee failed the drug test and also give the number of prior drug arrests for each parolee. We have coded a failed drug test as 1 and a successful drug test as 0. Table 4.2 provides the OLS regression

**Figure 4.1**

*Scatterplot of example of drug testing with the predicted regression line*



results for our example. The OLS regression suggests a very strong relationship between prior drug arrests and failing the drug test. But if we look more closely at the regression model, we can see that this approach may lead to outcomes that are not consistent with the processes we seek to understand.

Figure 4.1 shows the data points for our example in a scatterplot, as well as the regression line drawn from the results in Table 4.2. It is clear that the OLS regression approach leads to predicted outcomes that are

not possible, given our dependent variable. For example, for a parolee with six drug arrests, our model predicts that  $y$  will have a value of 1.099:

**Working It Out**

$$\begin{aligned}y &= 0.211 + 0.148x_1 \\&= 0.211 + 0.148(6) \\&= 0.211 + 0.888 \\&= 1.099\end{aligned}$$

For a parolee with eight drug arrests, our model predicts that  $y$  will have a value of 1.395:

**Working It Out**

$$\begin{aligned}y &= 0.211 + 0.148x_1 \\&= 0.211 + 0.148(8) \\&= 0.211 + 1.184 \\&= 1.395\end{aligned}$$

But in our example, the predicted value of  $y$  should logically be no greater than 1 or no less than 0. A value of 1 means that the parolee failed the drug test, and a value of 0 means that the parolee passed the drug test. Predicting values greater than 1 or less than 0 just does not make sense given the possible outcomes of a binary dependent variable.

This example, then, illustrates a logical problem in using OLS methods to gain estimates for cases where the dependent variable is dichotomous. The OLS approach assumes that there is no limit to the predicted value of the dependent variable. But in the case of a dichotomous dependent

variable, there are limits—the values 0 and 1. While this assumption of predictions within the limits of the possible outcomes of the dependent variable is also violated when OLS regression is used for an ordinal-level dependent variable and sometimes when it is applied to specific interval-and ratio-level measures (e.g., positively values scales), the violation is most extreme in the case of a binary dependent variable, such as drug testing failures. It does not make sense to analyze such situations with a model that allows the value of  $y$  to increase at a constant rate for each change in the value of  $x$ . For our analysis to be consistent with the problem we are examining, it must provide predictions that are constrained to values between 0 and 1.

Figure 4.1 illustrates additional problems that arise in using the OLS method in a case where the dependent variable is dichotomous. In our discussion of excluded variables in Chap. 2, we noted that a central assumption of the regression approach is that there is no systematic relationship between the error term and the independent variables included in the regression. When a systematic relationship exists, estimates of the regression coefficients are likely to be biased. But if you look at parolees for whom the value of prior drug arrests is greater than 5 (see Fig. 4.1), you can see that there is a consistent relationship between the regression error and the independent variable. Because the actual value of  $y$  cannot be greater than 1, and the predicted values continue to increase in a linear fashion (as evidenced by the regression line), the regression error increases in the negative direction as the number of prior drug arrests increases. This means that as the number of prior drug arrests gets larger and larger, we will make larger and larger negative errors in prediction. When OLS regression is used to examine a binary dependent variable, we are very likely to have a systematic relationship between the independent variable and the errors we make in predicting  $y$ , because  $y$ -values are constrained to 0 and 1 and predicted values of  $y$  have no limit.

Figure 4.1 also illustrates why when we examine a dichotomous dependent variable, we violate assumptions important to making statistical inferences with OLS regression. We noted in Chap. 2 that two assumptions of our test of statistical significance in regression are that the prediction errors are normally distributed around the regression line and that they meet an assumption of homoscedasticity (equal variances around the regression line). The normality assumption is clearly violated when we have a dichotomous dependent variable. As Fig. 4.1 shows, the shape of the distribution of  $x$  around  $y$  (the focus is on the vertical distance of each  $x$  from the regression line as these represent the prediction errors) will be bimodal

because the observed values of our dependent variable are constrained in practice to 0 and 1. However, if our sample is large enough, we generally allow violations of this assumption given the central limit theorem. Our problem in regard to homoscedasticity is more serious. We noted in Chap. 2 that violations of the assumption of homoscedasticity must be large before they become a concern. In the case of a binary dependent variable, heteroscedasticity (violation of the homoscedasticity assumption) is likely to be large. As shown in Fig. 4.1 for our distribution of drug testing failures, the distribution of the scores of  $x$  around the regression line is likely to vary widely in form, depending on the scores of the independent variable.

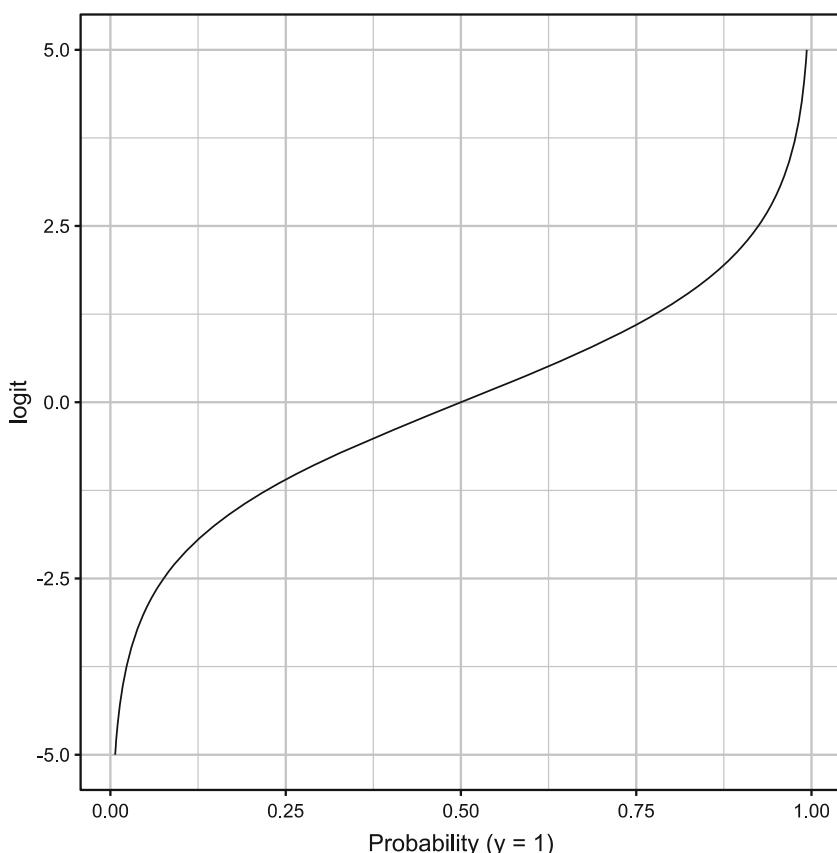
## Logistic Regression

---

While the application of OLS regression methods to a dichotomous dependent variable raises a number of substantive and statistical concerns, there are many advantages to the basic form of the regression approach introduced in previous chapters. For example, the effect of each  $b$  was constant. That is, we could define a single constant effect for each variable in the regression model. That effect took into account the other variables in the model. And we could add all of these effects and the  $y$ -intercept to get a predicted value for  $y$ . Because of the utility of the regression approach, statisticians have developed alternative methods for conducting regression analysis with dichotomous dependent variables that do not violate basic assumptions but allow us to continue to use the overall regression logic. These fall in the family of generalized linear models we described in Chap. 1. Perhaps, the most widely used of these methods is **logistic regression analysis**.<sup>1</sup> Logistic regression analysis is based on a transformation of the regression model that allows the outcomes of the regression equation to vary without limit, but constrains the predictions of the dependent variable to values between 0 and 1. At the same time, the inferential statistics used in logistic

---

<sup>1</sup>A method called generalized least squares might also be used to deal with violations of our assumptions, though logistic regression analysis is generally the preferred method. See Hanushek and Jackson (1977) for a comparison of these approaches. See also Hosmer and Lemeshow (2000). Another method, probit regression analysis, is very similar to that presented here, though it is based on the standard normal distribution rather than the logistic model curve. The estimates gained from probit regression are likely to be very similar to those gained from logistic regression. Because logistic regression analysis has become much more widely used and is available in most statistical software packages, we focus on logistic regression in this chapter.

**Figure 4.2***The logistic model curve*

regression do not rely on unrealistic assumptions regarding the population distribution of scores.

In fitting the data that are analyzed, logistic regression analysis uses the logic of a curve rather than that of a straight line. Figure 4.2 shows the **logistic model curve** for the probability that  $y = 1$ . More specifically, it is showing how any given probability between 0 and 1 maps onto a logit or specific value on the logistic curve. While the logistic regression curve follows the linear model in the middle of its distribution, it does not allow values below 0 or above 1. Indeed, as the logistic

curve approaches 0 or 1, it begins to stretch out, so it keeps coming closer to—but never actually reaches—either of these two values. Logistic regression is nonlinear; small changes in the probability near 0 and 1 are associated with large changes in the logit, whereas small changes in the probability near 0.5 are associated with relatively smaller changes in the logit. In short, the logistic curve satisfies our primary objection to the linear OLS regression method. That is, it does not allow predictions greater than or less than the actual values of the distribution of scores that we are trying to predict.

The logistic model curve provides a solution to the problem of predictions beyond the observed distribution. However, in order to gain the desired property of outcomes between 0 and 1, we have to alter the form of our regression equation. Using OLS regression, we represent our equation for the prediction of  $y$  with one independent variable as follows:

$$y = b_0 + b_1x_1$$

As noted earlier, this approach may yield values that are greater than 1 or less than 0, as was the case in our drug testing example.

In logistic regression, we alter the form of what we are trying to predict. Rather than predicting  $y$ , as in OLS regression, we now predict the natural logarithm ( $\ln$ ) of the odds of getting a 1 on the dependent variable. Although this sounds very imposing, it is simply a transformation of the equation presented above. Equation (4.1) represents the prediction equation for a simple logistic regression:

$$\ln \left( \frac{P(Y = 1)}{1 - P(Y = 1)} \right) = \ln \left( \frac{P(Y = 1)}{P(Y = 0)} \right) = b_0 + b_1x_1 \quad \text{Equation 4.1}$$

There is no change on the right side of this equation. We have the intercept (constant)  $b_0$  and the regression coefficient  $b_1$  that reflects the effect for the independent variable examined. Moreover, the outcome of this formula has no limit. But on the left side of the equation, we now have the **natural logarithm of the odds of  $y$** , or what statisticians call the **logit of  $y$** . A **logarithm**, or  $\log$ , is the exponent of the power to which a fixed number (called a base) must be raised to produce another number. So, for

example, if the base is 10, the logarithm of 100 is 2. That is, if we take 10 to the 2nd power ( $10^2$ ), we get 100. In logistic regression, we do not use a base of 10, rather we use what is called the common  $e$ , or Euler's number that is an irrational constant equal to 2.71828 (the decimal values continue indefinitely). What this means is that  $\ln(x)$  is the power to which  $e$  must be raised to get  $x$ . Mathematically, if  $a = \ln(x)$ , then  $e^a = x$ .<sup>2</sup>

What about the notation  $P(Y = 1)/[1 - P(Y = 1)]$  in Eq. (4.1)? This represents the odds of getting an outcome of 1, rather than 0, on the dependent variable. The odds are determined by dividing the probability of getting a 1 [ $P(Y = 1)$ ] by the probability of not getting a 1 [ $1 - P(Y = 1)$ ]. In our drug testing example, this would be the odds of failing a drug test divided by those of not failing the test. If an individual had an 80% predicted likelihood of failing the drug test, then the odds would be 0.80/0.20, or 4 to 1. The logit or logged odds is simply the natural log of this value.

### Working It Out

$$\text{odds} = \frac{P(Y = 1)}{1 - P(Y = 1)}$$

$$= \frac{0.80}{0.20}$$

$$= 4.0$$

$$\text{logit} = \ln(\text{odds})$$

$$= \ln(4.0)$$

$$= 1.39$$

---

<sup>2</sup>Your calculator likely has a button labeled  $e^x$ , which performs this operation. If there is no  $e^x$  button, then you should be able to locate a button labeled *INV* and another for the natural logarithm, *ln*. By pushing *INV* and then *ln* (the inverse or antilog), you will be able to perform this operation.

## ***Derivation of $P(Y = 1)$ from the Cumulative Logistic Probability Function***

---

We begin with the specification of the logistic regression model  $P(Y = 1)$ :

$$\ln \left( \frac{P(Y = 1)}{1 - P(Y = 1)} \right) = b_0 + b_1 x_1$$

To simplify, we let

$$Xb = b_0 + b_1 X_1$$

and

$$P = P(Y = 1) \Rightarrow 1 - P(Y = 1) = 1 - P$$

Using these simplifications, we can rewrite the logistic regression equation as

$$\ln \left( \frac{P}{1 - P} \right) = Xb$$

If we exponentiate both sides of the equation (i.e., take the value of e to the power of both sides of the equation), we obtain

$$e^{\ln [P/(1-P)]} = e^{Xb}$$

Then,  $\ln \left( \frac{P}{1 - P} \right)$  is the power to which we must raise e to get  $\frac{P}{1 - P}$ ; that is,

$$e^{\ln [P/(1-P)]} = \frac{P}{1 - P}$$

This leads to rewriting the logistic regression equation as

$$\frac{P}{1-P} = e^{Xb}$$

We multiply both sides of the equation by  $(1 - P)$ :

$$P = e^{Xb}(1 - P) = e^{Xb} - Pe^{Xb}$$

Then, we add  $Pe^{Xb}$  to both sides of the equation:

$$P + Pe^{Xb} = e^{Xb}$$

Next, we rewrite the equation to pull out the common factor,  $P$ :

$$P(1 + e^{Xb}) = e^{Xb}$$

Now, we divide both sides of the equation by  $(1 + e^{Xb})$  to solve for  $P$ :

$$\begin{aligned} P &= \frac{e^{Xb}}{1 + e^{Xb}} = \frac{1}{\left(\frac{1 + e^{Xb}}{e^{Xb}}\right)} = \frac{1}{\left(\frac{1}{e^{Xb}} + 1\right)} \\ &= \frac{1}{1 + e^{-Xb}} \end{aligned}$$

Since, as noted above,  $P = P(Y = 1)$ :

$$P(Y = 1) = \frac{1}{1 + e^{-Xb}}$$

If we transform this equation further, we see that it gives us the property we are looking for. That is, the predicted values of  $y$  produced by our regression equation will vary between 0 and 1, despite the fact that the outcomes in our regression equation can reach any value between plus and minus infinity. In the box above, we show how to transform the equation so that the outcome is the probability that  $y$  will be 1. The end result is a simple equation that can be calculated on a hand calculator with a natural log function. This equation is often called the **cumulative logistic probability function**.

$$P(Y = 1) = \frac{1}{1 + e^{-(b_0 + b_1 x_1)}} \quad \text{Equation 4.2}$$

By using the term  $Xb$  to represent the right side of the regression equation (i.e., to stand-in for the full regression equation),<sup>3</sup> we may write Eq. (4.2) more generally to take into account any number of independent variables:

$$P(Y = 1) = \frac{1}{1 + e^{-Xb}} \quad \text{Equation 4.3}$$

An equivalent form of Eq. (4.3) often seen in the literature on logistic regression is shown below. This equation avoids the awkwardness of the negative exponent in the denominator. However, these can be used interchangeable.

$$P(Y = 1) = \frac{e^{Xb}}{1 + e^{Xb}}$$

The  $Xb$  may seem a bit cryptic, but it greatly simplifies regression notation and comes from matrix algebra. The capital letter  $X$  represents a matrix of our independent variables where each column is a variable and each row is the data for one unit of observation, such as a person. The lower case  $b$  is a vector of regression coefficients. Because our regression model has an intercept, we also need a column of 1s in the  $X$  matrix so that the math works out. For example, a dataset with two independent variables would have an  $X$  matrix and  $b$  vector of the following form:

---

<sup>3</sup>This is the matrix notation for the regression equation.

$$X = \begin{bmatrix} x_0 & x_1 & x_2 \\ x_0 & x_1 & x_2 \\ x_0 & x_1 & x_2 \\ \dots & \dots & \dots \\ x_0 & x_1 & x_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 23.3 \\ 1 & 2 & 41.4 \\ 1 & 1 & 31.9 \\ \dots & \dots & \dots \\ 1 & 0 & 15.8 \end{bmatrix} \quad b = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

where  $x_0$  is the intercept,  $x_1$  is the first independent variable and  $x_2$  is the second independent variable. The  $Xb$  notation accommodates any number of independent variables, allowing for notation that is general and not specific to a particular model. Thus, the following are equivalent notations:

$$Xb = b_0x_0 + b_1x_1 + b_2x_2 + b_3x_3 = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

Notice that the  $x_0$  has dropped out of the right most equation given that it is simply multiplication by 1 for all cases.

Returning to Eq. (4.3), this equation divides 1 by the sum of 1 and  $e$  (the value 2.71828) taken to the  $-Xb$  power. The process of taking a number to some power is referred to as exponentiation. Here, we exponentiate  $e$  to the power  $-Xb$ . Exponentiation may also be familiar to you as the antilog or inverse log. We can also express this equation as a function of the logit. This is shown below as Eq. (4.4).

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = b_0 + b_1x_1 \quad \text{Equation 4.4}$$

where  $p$  is the probability that  $Y = 1$  or  $P(Y = 1)$ . Notice that the right-hand side of this equation looks same as OLS regression models; we have simply replaced the left-hand side with the logit of  $p$ . This can easily be extended to multiple independent variables.

Importantly, whatever the value associated with the logit, the value of  $P(Y = 1)$  will always be between 0 and 1. The value of  $P(Y = 1)$  can get closer and closer to 1 or to 0, but it will never reach or exceed that number. This is illustrated in Table 4.3, where we take very large negative and positive values of the logit. As the values get very large, the gain for each increase in  $Xb$  becomes smaller and smaller. Logistic regression, then, allows us to use the traditional regression format in which the outcome can be any real number, positive or negative. However, since we have converted what we are predicting to the logit of  $y$ , our predictions of  $y$  are bounded by 0 and 1.

Use of the natural logarithm of the odds of  $y$ , or the logit of  $y$ , has allowed us to develop a regression model in which the predicted outcomes

**Table 4.3**

Illustration of the fact that  $P(Y = 1)$  will not exceed 1 or be less than 0 for any value of the logit

LOGIT	$P(Y = 1)$
-25	0.000000000014
-20	0.000000002061
-15	0.000000305902
-10	0.000045397669
-5	0.006692850924
-4	0.017986209962
-3	0.047425873178
-2	0.119202922022
-1	0.268941421370
0	0.500000000000
1	0.731058578630
2	0.880797077978
3	0.952574126822
4	0.982013790038
5	0.993307149076
10	0.999954602131
15	0.999999694098
20	0.999999997939
25	0.99999999986

of  $y$  are constrained between 0 and 1. But what does a prediction between the values 0 and 1 mean? As we have already noted, the observed outcomes for a dichotomous dependent variable have a score of either 0 or 1. For example, in our drug testing example, either parolees failed the drug test (coded as 1) or they passed the test (coded as 0). When we examined the regression approach in previous chapters, the predicted value of  $y$  was simply one of the possible values in the distribution of scores on our interval- or ratio-level dependent measure. With a dichotomous dependent variable, our interpretation must be different. The predicted value in this case is the predicted *probability* of getting an outcome of 1. So, for example, a value of 0.50 in our drug testing example would mean that, according to our model, an individual was predicted to have about an equal chance of failing and not failing drug testing. A value of 0.90 would suggest that an individual was highly likely to have a drug testing failure.

Notice that the logistic regression model is not modeling the 0s and 1s directly but the proportion of 1s associated with each combination of the independent variables. With a single discrete independent variables such as in this example, we can illustrate this fairly easily. Table 4.4 shows the proportion of failed drug tests by the number of drug arrests. It also shows the logit associated with each of these proportions. Because these proportions are all less than 0.50, the logits are all negative. However, the logit for a proportion of 0 is listed as *-Inf* or minus infinity. When the proportion is 0 or

**Table 4.4**

Proportion and logit of failed drug testing results by prior drug arrests for 30 parolees

NUMBER OF DRUG ARRESTS	PROPORTION	LOGIT
0	0.0333	-3.367
1	0.0000	-Inf
2	0.1667	-1.609
3	0.0667	-2.639
4	0.1000	-2.197
5	0.1000	-2.197
6	0.0667	-2.639
7	0.0333	-3.367
8	0.0333	-3.367

1, the logit is undefined. If we assume for a moment that there are valid logit values for all number of drug arrests values, we could estimate an OLS regression model with the logit as the dependent variable, as implied by Eq. (4.4). This would approximate a logistic regression model. Furthermore, this shows that the regression coefficients reflect predicted change in the logits and not the original 0s and 1s. This method would be tedious with a large number of independent variables. More problematic, as shown in Table 4.4, it cannot handle the combinations of the independent variables where the proportion of the outcome is 0 or 1. The estimation method for logistic regression solves this problem. This is accomplished by using Eq. (4.3) that links the linear regression model ( $Xb$ ) to the dependent variable.

To estimate the coefficients of a logistic regression, we use a much more complex mathematical process than was used in OLS regression. It is based on **maximum likelihood estimation (MLE)** techniques. Using these techniques, we try to maximize the probability that our regression estimates will produce a distribution similar to that of the observed data. With this approach, we do not simply derive a single mathematical solution for obtaining the regression estimates.<sup>4</sup> Rather, we begin by identifying a tentative solution, which we then try to improve upon. Our criterion for improvement is termed a likelihood function. A likelihood function measures the probability of observing the results in the sample (i.e., our data), given the coefficient estimates in our model. By convention in logistic regression, we use -2 times the natural logarithm of the likelihood function (or **-2LL**), which is defined as the **log-likelihood function**. We repeat this process again and again, until the change in the likelihood function is considered negligible. We repeat the process and re-estimate our coefficients. Each repetition is called an **iteration**. Logistic regression is said to

---

<sup>4</sup>It should be noted, however, that maximum likelihood techniques do not always require an iterative process.

be an iterative procedure, because it tries a number of solutions before arriving at a final result—or, in statistical terms, converging.

Most packaged statistical programs set a default limit on the number of iterations that can be tried. Most models converge quickly and statistical software programs generally limit the number of interactions to no more than 20. **Lack of convergence** in a standard number of iterations may indicate some type of problem in the regression model. Often, this occurs when the number of variables examined is large relative to the number of cases in the study. John Tukey, a noted statistician who taught at Princeton University, suggested a rule for logistic regression: that there be at least five cases and preferably at least ten in the smaller category of the dependent variable for each independent variable examined (Tukey 1997). An implication of this is that your sample size requirements increase the rarer your outcome (i.e., when you have mostly 0s or mostly 1s for the dependent variable). Whatever the cause, if you receive a message from a statistical analysis program that your regression has failed to converge, you should look carefully at your model and the distributions of your dependent measure across combinations of your independent variables.

We have now looked at the basic logic of the logistic regression model. While the logistic regression model differs from the OLS regression model in the outcome predicted, the basic form of the additive linear model has been maintained. The right side of the equation remains an additive linear function of the  $\gamma$ -intercept and the independent variables (multiplied by their associated regression coefficients). The effect of each independent variable remains its independent effect, with the other variables in the model controlled. We also continue to be constrained by the same regression assumptions regarding correct model specification. The models are also sensitive to problems of multicollinearity. However, we do not need to concern ourselves with the assumptions of linearity or homoscedasticity. The assumption of independence is the same as with OLS regression. These concepts were discussed in Chap. 2.

## A Substantive Example: Adoption of Compstat in U.S. Police Agencies

---

Application of logistic regression to a substantive problem will help you to understand the use of logistic regression, as well as the different coefficients associated with the technique. The example we use is drawn from a Police Foundation survey of U.S. police agencies, begun in 1999 and completed in the year 2000.<sup>5</sup> The Police Foundation surveyed all police agencies with

---

<sup>5</sup>For a description of this study, see Weisburd et al. (2001).

more than 100 sworn officers ( $n = 515$ ) and got a response rate of 86%. A main concern of the survey was whether Compstat, a management system first developed in New York City in order to reduce crime and improve quality of life, had been widely adopted in some form by other U.S. police agencies. It was theorized that Compstat would be much more likely to be adopted by larger police agencies. Using logistic regression analysis, we will examine whether this hypothesis is supported by the Police Foundation data.

The dependent variable in our analysis is dichotomous, measuring whether the department claimed to have adopted a Compstat-like program. The main independent variable is the number of sworn officers serving in the department at the time of survey.<sup>6</sup> We also include, as a second independent variable, region, which divides the country into four regions: South, West, North Central, and Northeast. For this multicategory nominal variable, we use three dummy variables to represent region and define the North Central region as the reference, or excluded, category. Our regression model is represented in Eq. (4.5):

$$\text{logit}(p) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 \quad \text{Equation 4.5}$$

where  $x_1$  is the number of sworn officers,  $x_2$  is the dummy variable for the Northeast region,  $x_3$  is the dummy variable for the South region, and  $x_4$  is the dummy variable for the West region.

Table 4.5 shows the iteration history for estimating this regression. As you can see, it took only three iterations to achieve convergence. The convergence criterion used in this run was that the log-likelihood function declined by less than 0.010%. As noted earlier, we use  $-2$  times the natural logarithm of the likelihood function ( $-2\text{LL}$ ) to define the log-likelihood function. The final coefficients listed in this table are the same as the regression coefficients ( $b$ ) reported in the summary of the regression results provided in Table 4.6.

---

<sup>6</sup>Departments with 1300 or more officers were coded in our example as 1300 officers. This transformation was used in order to take into account the fact that only 5% of the departments surveyed had more than this number of officers and their totals varied very widely relative to the overall distribution. Another solution that could be used to address the problem of outliers is to define the measure as the logarithm of the number of sworn officers, rather than the raw scores. We relied on the former solution for our example because interpretation of the coefficients is more straightforward. In an analysis of this problem, a researcher would ordinarily want to compare different transformations of the dependent variable in order to define the one that best fits the data being examined.

**Table 4.5**

Iteration history of logistic regression estimation

ITERATION	-2 LOG-LIKELIHOOD	COEFFICIENTS					NUMBER SWORN
		INTERCEPT	NORTHEAST	SOUTH	WEST		
1	493.418	−1.555	.258	.629	.308	.001	
2	492.515	−1.783	.351	.795	.419	.002	
3	492.513	−1.795	.359	.805	.428	.002	

Note: Initial -2 Log-Likelihood: 528.171

**Table 4.6**

Summary of the logistic regression coefficients

VARIABLE	b	SE	WALD	df	p	EXP(b)
Northeast	.359	.372	.931	1	.335	1.432
South	.805	.332	5.883	1	.015	2.237
West	.428	.367	1.360	1	.244	1.534
Number Sworn	.002	.000	24.842	1	.000	1.002
Intercept	−1.795	.311	33.378	1	.000	.166

We can now express our regression equation in terms of the outcomes of our analysis. Inserting the values from our regression analysis, we can express the equation as follows:

$$\text{logit}(p) = -1.795 + 0.002x_1 + 0.359x_2 + 0.805x_3 + 0.428x_4$$

where  $x_1$  is the number of sworn officers,  $x_2$  is the dummy variable for the Northeast region,  $x_3$  is the dummy variable for the South, and  $x_4$  is the dummy code for the West.

We can also develop predictions of  $y$  from this model, as in the case of the OLS model. However, as explained above, our predictions of  $y$  are not the direct outcome of our additive regression model. Rather, they are the logit of the probability of  $y$  occurring and we can convert these logits into probabilities using Eq. (4.6), as shown here.

$$p = \frac{e^{\text{logit}}}{1 + e^{\text{logit}}} \quad \text{Equation 4.6}$$

For example, let us say that we want to predict the probability of a Compstat-like program in a department with 1000 officers in the North

Central region. Our first task is to determine the logit for this combination of the independent variables. We do that by applying coefficients gained in our logistic regression analysis. Because North Central is the reference category, the equation contains only the  $y$ -intercept and the effect of the number of sworn officers. The associated logit is 0.205.

### Working It Out

$$\begin{aligned}\text{logit}(p) &= b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 \\ &= -1.795 + 0.002(1000) + 0.359(0) + 0.805(0) + 0.428(0) \\ &= -1.795 + 2 \\ &= 0.205\end{aligned}$$

We can convert this logit using Eq. (4.6). By applying this equation to our result, we see that the probability of having a Compstat-like program in such a department is about 55%, at least according to this regression model.

### Working It Out

$$\begin{aligned}p &= \frac{e^{\text{logit}}}{1 + e^{\text{logit}}} \\ &= \frac{e^{0.205}}{1 + e^{0.205}} \\ &= 0.55\end{aligned}$$

For a police department in the South with 1000 officers, the predicted probability of having a Compstat-like program is fully 73%.

### Working It Out

$$\begin{aligned}
 \text{logit}(p) &= b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 \\
 &= -1.795 + 0.002(1000) + 0.359(0) + 0.805(1) + 0.428(0) \\
 &= -1.795 + 0.002(1000) + 0.805(1) \\
 &= -1.795 + 2 + 0.805 \\
 &= 1.01 \\
 p &= \frac{e^{\text{logit}}}{1 + e^{\text{logit}}} \\
 &= \frac{e^{1.01}}{1 + e^{1.01}} \\
 &= 0.733
 \end{aligned}$$

If we apply our prediction model to smaller departments, we see that they are less likely, according to our estimates, to have a Compstat-like program. For example, our model suggests that a police agency with only 100 officers from the South would have a probability of only 31% of having a Compstat-like program.

### Working It Out

$$\begin{aligned}
 \text{logit}(p) &= b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 \\
 &= -1.795 + 0.002(100) + 0.359(0) + 0.805(1) + 0.428(0) \\
 &= -1.795 + 0.002(100) + 0.805(1) \\
 &= -1.795 + 0.2 + 0.805 \\
 &= -0.79 \\
 p &= \frac{e^{\text{logit}}}{1 + e^{\text{logit}}} \\
 &= \frac{e^{-0.79}}{1 + e^{-0.79}} \\
 &= 0.312
 \end{aligned}$$

## Interpreting Logistic Regression Coefficients

---

Using Table 4.6, we can also define the specific effects of the variables in our model. In this case, the logistic regression coefficients are listed in the column labeled  $b$ . As expected, the coefficient for the number of sworn officers is positive; that is, as the number of sworn officers increases, the likelihood of having a Compstat-like program also increases. The three dummy variables included for region also have a positive impact relative to the North Central region, which is the excluded category. This means that in the Police Foundation sample, police departments in the Northeast, West, and South regions were more likely to claim to have a Compstat-like program than those in the North Central region (the reference category for this dummy coding).

What about a numeric interpretation of the logistic regression coefficient? Here, we can see the price we pay for developing a regression model in which the predictions of the probability of  $y$  are constrained between 0 and 1. In the OLS regression case, the interpretation of  $b$  is in reference to units of measurement of  $y$ . In the multiple predictor variable case,  $b$  represents the estimated change in  $y$  associated with a unit change in  $x$ , when all other independent variables in the model are held constant. So, a  $b$  of 2 in an OLS regression suggests that a unit increase in  $x$  is associated with a two-unit increase in  $y$  (all other included independent variables held constant).

The interpretation of the **logistic regression coefficient** is not as straightforward. Our regression equation is predicting not  $y$ , but the logit or logarithm of the odds of getting a 1—or, in our example, the log of the odds of having a Compstat-like program. In a multiple logistic regression,  $b$  represents the estimated change in the log of the odds of  $y$  occurring when all other independent variables are held constant. The coefficient for number of sworn officers is 0.002, meaning that each additional officer increases by 0.002 the log of the odds of having a Compstat-like program. While some researchers may have an intuitive understanding of the change in the log of the odds, the transformation of the outcome measure in the logistic regression model has made the regression coefficients very difficult to explain or interpret in a way that nonstatisticians will understand.

### The Odds Ratio

To make results easier to understand, statisticians have developed other methods of interpreting logistic regression coefficients. An approach commonly used is to report the regression coefficient in terms of its odds ratio. The **odds ratio**, sometimes called the exponent of  $b$ , is reported as  $Exp(B)$  in SPSS output, as *Odds Ratio* in Stata output, and in R, it depends on which function you use to get this output. The odds ratio represents the impact of

a one-unit change in  $x$  on the ratio of the probability of an event occurring to the probability of the event not occurring. Equation (4.7) defines the odds, and Eq. (4.8) defines the odds ratio. Note that in Eq. (4.8), the subscripts  $x_i$  simply means some value of  $x$  and  $x_1 + 1$  is that value of  $x$  plus 1. Thus, we are defining the odds ratio in terms of the calculation of the odds for two events separated by a change of one unit in  $x$ :

$$\text{odds} = \frac{p}{1-p} \quad \text{Equation 4.7}$$

$$\text{odds ratio} = \frac{\text{odds}_{x_i+1}}{\text{odds}_{x_i}} \quad \text{Equation 4.8}$$

An odds ratio greater than 1 indicates that the odds of getting a 1 on the dependent variable *increase* when the independent variable *increases*. An odds ratio less than 1 indicates that the odds of getting a 1 on the dependent variable *decreases* when the independent variable *increases*. For our example, an odds ratio greater than 1 indicates that as the independent variable increases, the odds of having a Compstat-like program also increase. If the odds ratio were 3, for example, then a one-unit change in  $x$  would make the event  $y$  about three times as likely to occur. An odds ratio less than 1 would suggest that the likelihood of having a Compstat-like program decreased as the independent variable increased. For example, an odds ratio of .5 would suggest that the likelihood of having a Compstat-like program decreased by  $\frac{1}{2}$  as the independent variable increased by 1.

In Table 4.6, we see that  $b = 0.002$  and  $e^b = 1.002$  [equivalently written as:  $\exp(b)$ ] for number of sworn officers. As noted earlier in the chapter, when we exponentiate the value of the coefficient  $b$ , we take  $e$ —the value 2.71828—to the power of the coefficient  $b$ . For number of sworn officers, it is  $e^{0.002} = 1.002$ . What this means is that for any logistic regression analysis, all we need to do is exponentiate the logistic regression coefficient to calculate the odds ratio. Most statistical software will automatically report the odds ratios for each of the independent variables included in the analysis, although for some programs you must specify this as optional output.

It is important to keep in mind that the odds ratio provides an estimate for only a *single* one-unit increase in the independent variable. When the independent variable is binary, as with a dummy variable, this makes the interpretation easy. The odds ratio reflects the differential odds of the outcome between the two categories of the dummy variable, controlling for other variables in the model. For ordinal and continuous independent variables, the interpretation is more nuanced. The odds ratio is *not* a linear function of the coefficients; thus, we cannot say that for each one-unit

increase in the independent variable, the odds increase by some amount. If we are interested in a change of more than one unit in our independent variable—say 2, 5, 10, or 100 units—we multiply that number by our coefficient  $b$  and then exponentiate that value. For example, returning to the number of sworn officers, suppose we are interested in the odds of adopting a Compstat-like program for a department that added 100 officers. We multiply our coefficient of 0.002 by 100, getting a value of 0.2, and then take  $e$  to the power of 0.2, which gives us a value of 1.2214.

### Working It Out

$$\text{odds ratio} = e^{0.002(100)} = e^{0.2} = 1.2214$$

This odds ratio tells us that the odds of adopting a Compstat-like program increase by a factor of 1.22 for a department with 100 additional officers. As an exercise, take the odds of adopting a Compstat-like program for a department with 100 officers in the North Central region (0.2029; see page 155) and calculate the odds for a department with 200 officers. Then, take the ratio of these two odds—it will equal 1.2214.

Our focus on the number of sworn officers illustrates another feature of logistic regression coefficients that is easily overlooked. There are times—usually for an interval-ratio-level independent variable—when the logistic regression coefficient will appear to have a small value. Yet, when we begin to account for the range of the independent variable and start to look at increases of 10, 100, or even 1000 in the independent variable, we may find that the odds increase by a substantial amount.

For our regional dummy variables, it should be remembered that the three measures are compared to the reference category, the North Central region. Because working out the odds ratio is tedious, we will carry out the calculations only for the South. According to the results presented in Table 4.6, the South has an associated odds ratio of 2.237, meaning that being in the South region of the country, as opposed to the North Central region, more than doubles the odds of having a Compstat-like program. As with our number of sworn officers' coefficient, we get a value of 2.2367 by taking  $e$  to the power of 0.805, which is the logistic regression coefficient for the South region.

### Working It Out

$$\text{odds ratio} = e^{0.805} = 2.2367$$

Alternatively, we can work through the calculation of the odds ratio to arrive at the same conclusion. Setting the number of sworn officers at 100, we will calculate the odds ratio of a Compstat-like program for the case where a department is in the South versus the case where it is in the North Central region.

### Working It Out: Departments in the South Region

**Step 1:** Determine the logit of the probability of  $y = 1$ .

$$\begin{aligned}\text{logit}(p) &= b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 \\ &= -1.795 + 0.002(100) + 0.359(0) + 0.805(1) + 0.428(0) \\ &= -1.795 + 0.002(100) + 0.805(1) \\ &= -1.795 + 0.2 + 0.805 \\ &= -0.790\end{aligned}$$

**Step 2:** Determine the probability of  $y = 1$ .

$$\begin{aligned}p &= \frac{e^{\text{logit}}}{1 + e^{\text{logit}}} \\ &= \frac{e^{-0.790}}{1 + e^{-0.790}} \\ &= 0.3122\end{aligned}$$

**Step 3:** Determine the odds.

$$\text{odds} = \frac{0.3122}{0.6878} = 0.4539$$

**W**orking It Out: Departments in the North Central Region

**Step 1:** Determine the logit of the probability of  $y = 1$ .

$$\begin{aligned}\text{logit}(p) &= b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 \\ &= -1.795 + 0.002(100) + 0.359(0) + 0.805(1) + 0.428(0) \\ &= -1.795 + 0.002(100) \\ &= -1.795 + 0.2 \\ &= -1.595\end{aligned}$$

**Step 2:** Determine the probability of  $y = 1$ .

$$\begin{aligned}p &= \frac{e^{\text{logit}}}{1 + e^{\text{logit}}} \\ &= \frac{e^{-1.595}}{1 + e^{-1.595}} \\ &= 0.1687\end{aligned}$$

**Step 3:** Determine the odds.

$$\text{odds} = \frac{0.1687}{0.8313} = 0.2029$$

**W**orking It Out

$$\begin{aligned}\text{odds ratio} &= \frac{\text{odds}_{x_i+1}}{\text{odds}_{x_i}} \\ &= \frac{0.4539}{0.2029} \\ &= 2.2367\end{aligned}$$

Turning to the odds comparing the West and Northeast regions with the North Central region, we can see that the differences are smaller (see Table 4.6). Like the South region statistic, these coefficients are positive, meaning that departments in these regions have a higher likelihood of reporting a Compstat-like program. The odds ratios for both regions are about 1.5. A police department in these regions is about 1.5 times as likely to have a Compstat-like program as a department in the North Central region.

### The Derivative at Mean

Another measure that sometimes makes it easier to understand the logistic regression coefficient (but that is not reported in many statistical software programs) is the **derivative at mean (DM)**. The derivative at mean converts the nonlinear logistic regression coefficient into a simple linear regression coefficient. Accordingly, it has the advantage of having the same interpretation as the result would have had if OLS regression had been appropriate to the problem. The disadvantage of the derivative at mean is that it calculates the regression coefficient as if it had a constant effect over the entire distribution of predicted values of  $y$ , based on the change observed when the predicted value of  $y$  is at its mean. In fact, the logistic curve in Fig. 4.2 shows that the impact of the parameters will change in absolute terms, depending on where in the distribution they are calculated. The derivative at mean will be largest when the mean of the dependent variable is close to the middle of the logistic curve. As the mean of the distribution moves closer to the tails of the logistic curve, the derivative will be smaller.

The interpretation of the derivative at mean is similar to that of the OLS regression coefficient. The derivative at mean may be defined as the change in  $y$  associated with a unit change in  $x$  at the mean value of the dependent variable. The derivative at mean is defined mathematically in Eq. (4.9):

$$\text{DM}_i = \bar{y}(1 - \bar{y})b_i$$

**Equation 4.9**

where  $\bar{y}$  is the mean of the dependent variable (i.e., the proportion of cases having a value of 1 for the dependent variable).

Table 4.7 provides the derivative at mean for each of the coefficients in our regression model. Since about 33% of the sample claimed to have implemented a Compstat-like program, the derivative at mean is calculated for a mean of  $y$  of 0.33. If we look at the derivative at mean for the dummy variables associated with region, we can see the advantage of this approach. Taking the South region, for which the difference from the excluded category is largest, we calculate the derivative at mean below:

**Table 4.7**

Derivative at mean for each of the regression coefficients in the Compstat example

VARIABLE	<i>b</i>	$DM = \bar{y}(1 - \bar{y})b_i$
Northeast	0.359	$(0.33)(0.67)(0.359) = 0.0794$
South	0.805	$(0.33)(0.67)(0.805) = 0.1778$
West	0.428	$(0.33)(0.67)(0.428) = 0.0946$
Number of sworn officers	0.002	$(0.33)(0.67)(0.002) = 0.0004$

### Working It Out

$$\begin{aligned}
 DM &= \bar{y}(1 - \bar{y})b_i \\
 &= (0.33)(1 - 0.33)(0.805) \\
 &= (0.33)(1 - 0.67)(0.805) \\
 &= 0.1778
 \end{aligned}$$

We can interpret this coefficient much as we interpreted the dummy variable regression coefficients in Chap. 2. If a police department is located in the South as opposed to the North Central region, its outcome on the dependent variable is about 0.1778 higher. Since the dependent variable has values ranging between 0 and 1, we can interpret this coefficient in terms of percentages. Departments in the South have, on average, about an 18-percentage-point higher chance of claiming to have a Compstat-like program when  $y$  is at its mean.

The derivative at mean for number of sworn officers is about 0.0004. This suggests that for each additional officer, there is a 0.0004 increase in the value of  $Y$ . According to the derivative at mean, an increase in 100 officers would lead to a 4-percentage-point increase in the likelihood of having a Compstat-like program. An increase of 1000 officers would lead to a 40-percentage-point increase.

### Working It Out

$$\begin{aligned}
 DM &= \bar{y}(1 - \bar{y})b_i \\
 &= (0.33)(1 - 0.33)(0.002) \\
 &= (0.33)(1 - 0.67)(0.002) \\
 &= 0.0004
 \end{aligned}$$

## Comparing Logistic Regression Coefficients

In Chap. 2, you saw how standardized regression coefficients could be used to compare the magnitude of regression coefficients measured on different scales. There is no widely accepted method for comparing the magnitude of the coefficients in logistic regression. When variables are measured on the same scale, we can rely on comparisons of the statistics we have reviewed so far. For example, if our model includes two binary dummy variables, we can easily gain a sense of the impact of each variable by comparing the size of each odds ratio (or the  $b$  coefficients).

Let us say that we are interested in predicting the likelihood of getting a prison sentence for a sample of convicted burglars. We include two binary dummy variables in our analysis. The odds ratio for the first variable, gender (0 = female; 1 = male), is 1.5. The odds ratio for the second, whether a gun was used in the burglary (0 = no; 1 = yes), is 2.0. In this case, we could say that use of a weapon has a larger effect on the likelihood of getting a prison sentence than does gender. In the case of gender, being a male as opposed to a female increases the odds of getting a prison sentence by about 50%. However, according to these estimates, using a gun in the burglary doubles the odds of getting a prison sentence.

### Using Probability Estimates to Compare Coefficients

If variables are measured on very different scales, comparing the magnitude of effects from one variable to another is often difficult. One easily understood and transparent method for doing this is to rely on the predicted probabilities of  $y$ . In a study using logistic regression, Wheeler, Weisburd, and Bode were confronted with a large number of statistically significant independent variables measured on very different scales (Wheeler et al. 1982). They decided to calculate probability estimates for measures at selected intervals when the scores of all other predictors were held at their mean. They also calculated a range of predictions computed from the 5th to 95th percentile scores for the measure of interest. The table they developed is reproduced in Table 4.8.

**Table 4.8**

Selected probability estimates and calculated range for significant variables in Wheeler, Weisburd, and Bode's Study of White-Collar Crime Sentencing

INDEPENDENT VARIABLES	PROBABILITY ESTIMATES <sup>a</sup>	RANGE <sup>b</sup>
1. Act-Related Variables		
(a) Maximum Exposure to Prison		44
• 1 day–1 year	32	
• 1 year and 1 day–2 years	35	
• 4 years and 1 day–5 years	45	
• 14 years and 1 day–15 years	76	
(b) Dollar Victimization		41
• \$101–\$500	27	
• \$2,501–\$5,000	38	
• \$10,001–\$25,000	47	
• \$25,001–\$100,000	51	
• Over \$2,500,000	68	
(c) Complexity/Sophistication		27
• 4	32	
• 6	38	
• 8	45	
• 10	52	
• 12	59	
(d) Spread of Illegality		21
• Individual	40	
• Local	47	
• Regional	54	
• National/International	61	
2. Actor-Related Variables		
(e) Social Background: Duncan S.E.I.		29
• 15.1	28	
• 49.4	41	
• 62.0	47	
• 66.1	49	
• 84.0	57	
(f) Social Background: Impeccability		17
• 7	54	
• 11	49	
• 14	45	
• 17	42	
• 21	37	
(g) Criminal Background: Number of Arrests		22
• 0	37	
• 1	43	
• 2	45	
• 5	51	
• 9	59	
(h) Criminal Background: Most Serious Prior Conviction		20
• None	37	
• Minor Offense	46	
• Low Felony	52	
• Moderate Felony	57	
(i) Role in Offense		24
• Minor	26	
• Missing	33	
• Single/Primary	50	

(continued)

**Table 4.8**

(continued)

INDEPENDENT VARIABLES	PROBABILITY ESTIMATES <sup>a</sup>	RANGE <sup>b</sup>
3. Legal Process Variables		
(j) Statutory Offense Category		39
• Antitrust Violations	28	
• Bribery	30	
• Bank Embezzlement	36	
• False Claims	36	
• Postal Fraud	38	
• Lending/Credit Fraud	45	
• SEC Violations	65	
• Tax Violations	69	
4. Other Variables		30
(k) Sex		
• Male	50	
• Female	20	
(l) Age		c
• 22	42	
• 30	48	
• 39	50	
• 48	46	
• 61	32	
(m) District		28
• Northern Georgia	34	
• Southern New York	34	
• Central California	43	
• Western Washington	43	
• Maryland	50	
• Northern Illinois	53	
• Northern Texas	62	

<sup>a</sup>Estimated likelihood of imprisonment when scores on all other variables are held at their mean<sup>b</sup>Range computed from 5th to 95th percentile score<sup>c</sup>Because of the curvilinear effect measured here, the range is not relevant

The study examined factors that explained whether or not white-collar offenders convicted in federal courts were sentenced to prison. The table gives the reader a sense of how changes in the independent variable affect changes in the dependent variable, as well as a general idea (using the range) of the overall influence of the measure examined. For example, the *amount of dollar victimization* in an offense (variable 2) and *role in offense* (variable 13) are both ordinal-level variables but are measured with a different number of categories. Looking at the table, we can see that a person playing a minor role in an offense had a predicted probability of imprisonment of about 26%, while someone playing a primary role had a 50% likelihood of imprisonment, according to the model estimated (and holding all other independent variables constant at their mean). A crime

**Table 4.9**

Table of selected probability estimates and range for the Compstat model

INDEPENDENT VARIABLES	PROBABILITY ESTIMATE	RANGE
<i>Number of sworn officers:</i>		
100 (5th percentile)	0.2468	0.53
500	0.4217	—
1300 (95th percentile)	0.7831	—
Northeast	0.4090	—
South	0.4647	—
West	0.4216	—

involving less than \$500 in victimization led to an estimated likelihood of imprisonment of 27%. A crime netting over \$2,500,000 led to an estimated likelihood of imprisonment of 68%. If we compare the range of predicted values between the 5th and 95th percentile scores for each variable, our calculation suggests that dollar victimization (with a range of 41%) has a much larger impact than role in an offense (with a range of 24%). Of course, the choice of the 5th and 95th percentiles is arbitrary. And this method also arbitrarily holds every other independent variable to its mean. Nonetheless, the advantage of this approach is that it provides a method of comparison that is straightforward and easy for the nonstatistician to understand.

To apply this method to our data, we need information on the mean for each independent variable. For our data, the means are

- Northeast: 0.225
- South: 0.373
- West: 0.229
- Number of sworn officers: 334.784

In the box on pages [“Calculating Selected Probability Estimates . . .”], the calculations are carried out according to the method employed by Wheeler, Weisburd, and Bode. Table 4.9 describes the results. Using this table, we can see that there are very large differences in the predicted probabilities of a Compstat-like program for departments of varying size. This illustrates a point made earlier, when we noted that the odds ratio for each change in number of sworn officers was small. Though the change per unit change in  $x$  is small in this case (because departments differ widely in size), the predicted change can be very large. Under this approach, the range variable suggests a larger impact for number of sworn officers than for region of country.

## **Calculating Selected Probability Estimates and Range for the Compstat Model**

For all of the following calculations,  $x_1$  = number of sworn officers,  $x_2$  = Northeast,  $x_3$  = South, and  $x_4$  = West.

Logit estimate for number of sworn officers:

$$\begin{aligned}\text{logit}(p) &= b_0 + b_1x_1 + b_2\bar{x}_2 + b_3\bar{x}_3 + b_4\bar{x}_4 \\ &= -1.795 + 0.002x_i + 0.359(0.225) + 0.805(0.373) + 0.428(0.229) \\ &= -1.795 + 0.002x_i + 0.0808 + 0.3003 + 0.0980 \\ &= -1.3159 + 0.002x_i\end{aligned}$$

$P(Y = 1)$  for 100 officers:

$$p = \frac{e^{-1.3159+0.002x_1}}{1 + e^{-1.3159+0.002x_1}} = \frac{e^{-1.3159+0.002(100)}}{1 + e^{-1.3159+0.002(100)}} = 0.2468$$

$P(Y = 1)$  for 500 officers:

$$p = \frac{e^{-1.3159+0.002x_1}}{1 + e^{-1.3159+0.002x_1}} = \frac{e^{-1.3159+0.002(500)}}{1 + e^{-1.3159+0.002(500)}} = 0.4217$$

$P(Y = 1)$  for 1300 officers:

$$p = \frac{e^{-1.3159+0.002x_1}}{1 + e^{-1.3159+0.002x_1}} = \frac{e^{-1.3159+0.002(1300)}}{1 + e^{-1.3159+0.002(1300)}} = 0.7831$$

Probability estimate for Northeast:

$$\begin{aligned}\text{logit}(p) &= b_0 + b_1\bar{x}_1 + b_2x_2 + b_3\bar{x}_3 + b_4\bar{x}_4 \\ &= -1.795 + 0.002(334.784) + 0.359x_2 + 0.805(0.373) + 0.428(0.229) \\ &= -1.795 + 0.6696 + 0.359x_2 + 0.3003 + 0.0980 \\ &= -0.7271 + 0.359x_2\end{aligned}$$

$P(Y = 1)$  for Northeast:

$$p = \frac{e^{-0.7271+0.359x_2}}{1 + e^{-0.7271+0.359x_2}} = \frac{e^{-0.7271+0.359}}{1 + e^{-0.7271+0.359}} = 0.4090$$

Probability estimate for South:

$$\begin{aligned}\text{logit}(p) &= b_0 + b_1\bar{x}_1 + b_2\bar{x}_2 + b_3x_3 + b_4\bar{x}_4 \\ &= -1.795 + 0.002(334.784) + 0.359(0.225) + 0.805x_3 + 0.428(0.229) \\ &= -1.795 + 0.6696 + 0.359(0.225) + 0.805x_3 + 0.428(0.229) \\ &= -0.9466 + 0.805x_3\end{aligned}$$

$P(Y = 1)$  for South:

$$p = \frac{e^{-0.9466+0.805x_3}}{1 + e^{-0.9466+0.805x_3}} = \frac{e^{-0.9466+0.805}}{1 + e^{-0.9466+0.805}} = 0.4647$$

Probability estimate for West:

$$\begin{aligned}\text{logit}(p) &= b_0 + b_1\bar{x}_1 + b_2\bar{x}_2 + b_3\bar{x}_3 + b_4x_4 \\ &= -1.795 + 0.002(334.784) + 0.359(0.225) + 0.805(0.373) + 0.428x_4 \\ &= -1.795 + 0.6696 + 0.0808 + 0.30003 + 0.428x_4 \\ &= -0.7443 + 0.428x_4\end{aligned}$$

$P(Y = 1)$  for West:

$$p = \frac{e^{-0.7443+0.428x_4}}{1 + e^{-0.7443+0.428x_4}} = \frac{e^{-0.7443+0.428}}{1 + e^{-0.7443+0.428}} = 0.4216$$

### Standardized Logistic Regression Coefficients

Some statistical software programs list the **standardized logistic regression coefficient**, Beta, which is analogous to the standardized regression coefficient. Like the standardized regression coefficient, the standardized logistic regression coefficient can be interpreted relative to changes (measured in standard deviation units) in the independent variable. The magnitude of the standardized logistic regression coefficient allows us to compare the relative influence of the independent variables, since a larger value for the standardized coefficient means that a greater change in the log of the odds is expected. In contrast to the standardized regression coefficients for linear regression models, the Beta calculated for logistic regression models does not fall between 0 and 1, but can take on any value.<sup>7</sup> Some statisticians warn that such coefficients should be interpreted with caution.<sup>8</sup> Nonetheless, they can provide a method for gaining a general sense of the strength of coefficients in logistic regression. The standardized logistic regression coefficient is calculated using Eq. (4.10):

$$\text{Beta} = b_i s_i$$

**Equation 4.10**

where  $b_i$  is the unstandardized coefficient for variable  $i$  from the original logistic regression model and  $s_i$  is the standard deviation for variable  $i$ . We interpret Beta as the change in the log of the odds of  $P(Y = 1)$  relative to a one standard deviation change in the independent variable. For example, a Beta of 0.4 implies that for a one standard deviation change in the independent variable, the log of the odds is expected to increase by 0.4. Alternatively, if Beta = -0.9, a one standard deviation change in the independent variable is expected to result in a decrease of 0.9 in the log of the odds that  $P(Y = 1)$ .

Returning to our example using the Compstat data, we find the standardized coefficient for number of sworn officers to be 0.6616.

### Working It Out

$$\begin{aligned}\text{Beta} &= b_i s_i \\ &= (0.002)(330.797) \\ &= 0.6616\end{aligned}$$

---

<sup>7</sup>Some researchers have proposed alternative ways of calculating standardized logistic regression coefficients that allow for interpretations related to changes in probabilities. See, for example, Kaufman (1996).

<sup>8</sup>For example, see Andy Field (2000).

**Table 4.10**

Beta and associated logistic regression coefficients for the Compstat model

VARIABLE	STANDARD DEVIATION	<i>b</i>	BETA	GELMAN BETA
Number of sworn officers	330.797	0.002	$(330.797)(0.002) = 0.6616$	1.323
Northeast	0.416	0.359	$(0.416)(0.359) = 0.1493$	0.416
South	0.484	0.805	$(0.484)(0.805) = 0.3896$	0.484
West	0.421	0.428	$(0.421)(0.428) = 0.1802$	0.421

In Table 4.10, we calculate Beta for all four of the coefficients in the model and present the accompanying unstandardized logistic regression coefficients. Though the unstandardized logistic regression coefficient for the South (0.805) seems very large relative to that for number of sworn officers (0.002), the reverse relationship is found when we look at the standardized coefficients. The standardized coefficient for number of sworn officers is 0.662, while that for the South is 0.390. This is consistent with our analysis of the probability estimates. Both of these methods take into account the fact that the scales of measurement for these measures differ widely. While you should use caution in relying on standardized logistic regression coefficients, here, as in other cases, they can provide a general yardstick for comparing the relative strength of coefficients within regression models.<sup>9</sup>

A complication with standardized regression coefficients, whether in a logistic regression model or a standard OLS model, is dummy variables. In their unstandardized form, regression coefficients for dummy variables are directly comparable to one another as each has the same scale of 0 and 1. A dummy variable with a 50/50 split will have a standard deviation of 0.5. However, as the split for the dummy variable becomes more extreme, the standard deviation becomes smaller. For example, for a dummy variable with a 10/90 split, the standard deviation is 0.22. Thus, standardization improves comparability for continuous independent variables but reduces it for dummy variables. Gelman and Hill (2007) propose a solution to this issue by leaving the regression coefficients for any binary variable, such as dummy variables, unchanged but standardizing the regression coefficient by two times the standard deviation of the independent variable. Thus, Eq. (4.10) becomes:

---

<sup>9</sup>As with standardized regression coefficients in OLS regression, you should not compare standardized logistic regression coefficients across models. Moreover, while we report standardized regression coefficients for the dummy variables included in the model, you should use caution in interpreting standardized coefficients for dummy variables.

$$\text{Beta} = b_i(2s_i)$$

**Equation 4.11**

Essentially, the Gelman and Hill method is setting the standard deviation for all dummy variables at 0.5, irrespective of the actually split between the 0s and 1s. The results for the Gelman method are shown in the far-right column of Table 4.10 and reinforce the prior conclusion that the effect of sworn officers is larger than the effect of the regional differences. However, with this standardization, we see that each of these three regions differs by about the same amount from the North Central region.

## Evaluating the Logistic Regression Model

---

In OLS regression, to assess how well our model explains the data, we use a straightforward measure of the percent of variance explained beyond the mean ( $R^2$ ). There is no equivalent statistic in logistic regression. Nonetheless, a number of measures have been proposed for assessing how well a model predicts the data.

### Percent of Correct Predictions

One widely accepted method for assessing how well a logistic regression model predicts the dependent variable is to compare the values of  $y$  predicted by the model to those that would be obtained simply by taking the observed distribution of the dependent variable. This statistic is commonly described as the **percent of correct predictions**. Table 4.11 shows the percent of correct predictions for our Compstat example. The formula for percent of correct predictions is presented in Eq. (4.12).

$$\text{Percent of correction predictions} = \left( \frac{n_{\text{correct predictions}}}{n_{\text{total}}} \right) \times 100$$

**Equation 4.12**

**Table 4.11**

Percent of correct predictions for the logistic regression of the Compstat data

OBSERVED	PREDICTED		PERCENTAGE CORRECT
	NO (0)	YES (1)	
No (0)	264	19	93.3
Yes (1)	106	30	22.1
Overall			70.2

Note: The cut value is 0.500

The observed predictions in our example represent the observed proportion of departments that report having a Compstat-like program. As we noted before, this number is about 0.33 (or 33%). We add the 106 and 30 cases in Table 4.7 where the observed value was 1, and then divide this number by the total number of cases in the analysis ( $n = 419$ ). The predicted values are drawn from Eq. (4.3). But, importantly, in order to compare these predicted values with the observed values, we must assign each case a 0 or a 1. The decision as to whether to define the predicted value as a 1 or a 0 is based on a set cutoff point. Herein lies the main drawback of this approach: The point at which you determine that the prediction is a 1 is arbitrary. In SPSS, as in other standard software packages, 0.50 is used as a natural cutoff point. That is, if we get a predicted probability of 0.50 or greater for a case in our study, it will be counted as a prediction of 1. Remember that a 1 in our case means that the department has a Compstat-like program. In this analysis, if the prediction is 0.495, the case is given a 0. Clearly, by using a single and arbitrary cutoff point, we are losing a good deal of information about how well the model fits the data.

The proportion of correct predictions is worked out below, using Eq. (4.12). The  $n$  of correct predictions is found by taking each case for which the actual and predicted values are the same. In 264 cases, the actual and predicted values are 0. In only 30 cases are the actual and predicted values equal to 1. These are the correct predictions in our analysis, so the total number of correct predictions is  $264 + 30 = 294$ . The percent of correct predictions is 70.2. This seems like a very high level of prediction. However, to interpret this statistic, we must compare it with the level we would have reached if we had not used our regression model. In that case, we would have had information only on the split in the dependent variable. As noted earlier, 33% of the departments claim to have implemented a Compstat-like program. Knowing only this, our best bet would have been to predict that every department did not have a Compstat-like program. If we did this, we would be correct about 67% of the time. Thus, our model did not improve our prediction very much over what we would have predicted with knowledge of only the outcomes of the dependent variable.

### Working It Out

$$\begin{aligned}\text{Percent of correct predictions} &= \left( \frac{n_{\text{correct predictions}}}{n_{\text{total}}} \right) \times 100 \\ &= \frac{294}{419} \times 100 \\ &= (0.7017) \times 100 \\ &= 70.17\end{aligned}$$

**Table 4.12**–2 Log-likelihood and pseudo- $R^2$  statistics

METHOD	VALUE
–2 Log-likelihood	492.513
Cox & Snell R-square	0.082
Nagelkerke R-square	0.114

**Pseudo- $R^2$** 

While there is no direct  $R^2$  measure for logistic regression, a number of what may be termed **pseudo- $R^2$**  measures have been proposed. Like standardized logistic regression coefficients, these measures are not well accepted and must be used with caution. Nonetheless, by providing a general sense of the prediction level of a model, they can add information to other statistics, such as the percent of correct predictions. A commonly used pseudo- $R^2$  measure is **Cox and Snell's  $R^2$**  (Cox and Snell 1989). As with other pseudo- $R^2$  statistics, a main component of this measure is the log-likelihood function ( $-2LL$ ). It makes good sense to rely on the log-likelihood function, since it measures the degree to which a proposed model predicts the data examined. In this case, we compare the difference between the  $-2LL$  estimate obtained when no independent variables are included (the null model) and the  $-2LL$  estimate obtained when all the independent variables are included (the full model). The  $-2LL$  value for the null model (528.171) is given in the footnote in Table 4.5. The  $-2LL$  value for the full model (492.513) is given in the model summary statistics provided in Table 4.12. Equation (4.13) provides the method of calculation for Cox and Snell's  $R^2$ .

$$\text{Cox and Snell's } R^2 = 1 - e^{-[(-2LL_{\text{null}}) - (-2LL_{\text{full}})]/n}$$

**Equation 4.13**

While this equation looks intimidating, it can be solved in two easy steps. First, we calculate the number that appears above  $e$  or the exponent of the natural log:

**Working It Out: Step 1**

$$\begin{aligned} & -[(-2LL_{\text{null}}) - (-2LL_{\text{full}})]/n \\ &= -[(528.171) - (492.513)]/419 \\ &= -35.658/419 \\ &= -0.085 \end{aligned}$$

We then take  $e$  to the power of  $-0.085$ , which, as we noted earlier, can be done simply on a basic scientific calculator. We next subtract this number from 1:

### Working It Out: Step 2

$$\begin{aligned}\text{Cox and Snell's } R^2 &= 1 - e^{-0.085} \\ &= 1 - 0.9185 \\ &= 0.0816\end{aligned}$$

Rounding 0.0816 to three decimal places gives a result of 0.082, which is the value shown in Table 4.11.

Like the percent of correct predictions, Cox and Snell's  $R^2$  suggests that our model does not provide for a very strong level of prediction. Statistical software programs may also produce another  $R^2$  statistic: the **Nagelkerke  $R^2$** . This statistic corrects for the fact that Cox and Snell's estimate and many other pseudo- $R^2$  statistics often have a maximum value of less than 1 (which would indicate that all of the variance in the dependent variable was explained by the independent variables included in the model). Nagelkerke's  $R^2$  is thus generally larger than Cox and Snell's  $R^2$ , which—especially with large values—will be too conservative (Nagelkerke 1991). Other pseudo- $R^2$  statistics will give estimates similar to those produced here. None of these values should be seen as an exact representation of the percent of variance explained in your model. But they can give you a rough sense of how well your model predicts the outcome measure.

## Statistical Significance in Logistic Regression

Statistical significance for a logistic regression can be interpreted in much the same way as it was for the regression models discussed in Chap. 2. However, a chi-square distribution is used, and thus, we do not have to be concerned with assumptions regarding the population distribution in our tests. For the overall model, there is a general test based on the difference between the  $-2LL$  statistics for the full and null models. The null model is the model with no independent variables, that is, an intercept-only model. The full model is the model with the intercept and the independent

variables. The **model chi-square** ( $\chi^2$ ) formula in logistic regression is represented in Eq. (4.14).

$$\text{Model } \chi^2 = (-2LL_{\text{null}}) - (-2LL_{\text{full}})$$

**Equation 4.14**

For our example, the model  $\chi^2$  is 35.658 (see working it out). The number of degrees of freedom is determined by the number of independent variables included in the model estimated. In our case, there are three regression coefficients for the variable region and the measure number of sworn officers. The number of degrees of freedom thus equals 4. Using a distribution function of any statistics program, we find that a  $\chi^2$  statistic of greater than 18.465 is needed for a statistically significant result at the 0.001 level. Because our chi-square statistic is much larger than this, our observed significance level is less than 0.001. Using conventional significance criteria, we would reject the null hypothesis and conclude that the model estimated provides significant improvement over that without any independent variables.

### Working It Out

$$\begin{aligned}\text{Model } \chi^2 &= (-2LL_{\text{null}}) - (-2LL_{\text{full}}) \\ &= 528.171 - 492.513 \\ &= 35.658\end{aligned}$$

In testing the statistical significance of individual parameters, statistical software packages ordinarily provide the **Wald statistic**.<sup>10</sup> This statistic also has a chi-square distribution, and so the statistical significance of a result may be checked in a chi-square table. The Wald statistic takes the ratio of the logistic regression coefficient to its standard error (see Eq. (4.15)). The standard error of the logistic regression coefficient is provided in the output (see Table 4.13). For the comparison of the South

---

<sup>10</sup>We discuss the Wald statistic in detail here because it is the most common test of statistical significance reported in many statistical software applications. It should be noted, however, that some researchers have noted that the Wald statistic is sensitive to small sample sizes (e.g., less than 100 cases). The likelihood ratio test discussed later in this chapter and in more detail in Chap. 5 offers an alternative test for statistical significance that is appropriate to both small and large samples (see Scott Long 1997).

**Table 4.13**

Summary of the logistic regression coefficients

VARIABLE	<i>b</i>	SE	WALD	df	<i>p</i>	EXP( <i>b</i> )
Northeast	0.359	0.372	0.931	1	.335	1.432
South	0.805	0.332	5.883	1	.015	2.237
West	0.428	0.367	1.360	1	.244	1.534
Number sworn	0.002	0.000	24.842	1	.000	1.002
Intercept	-1.795	0.311	33.378	1	.000	0.166

and North Central regions (the latter being the excluded category), we take the logistic regression coefficient of 0.805 and divide it by the reported standard error of 0.332. To get the Wald statistic, we square this number. The result is 5.879.<sup>11</sup>

$$W^2 = \left( \frac{b}{se_b} \right)^2 \quad \text{Equation 4.15}$$

Note that in the restricted case of a single regression coefficient, the Wald statistic is related to the *z*-statistic, which can also be used to test for the significance of individual coefficients, as shown in Eq. (4.16). An advantage of the Wald statistic, however, is that it can be used to test the joint significance of multiple coefficients, whereas the *z*-statistic cannot.

$$z = \frac{b}{se_b} \quad \text{Equation 4.16}$$

Confidence intervals are computed much the same ways as with OLS regression with the exception of using the critical value for the normal distribution, or *z*, rather than the *t*-distribution, as shown in Eq. (4.17).

$$\begin{aligned} \text{Lower 95\% } b &= b - se(z) = b - se(1.96) \\ \text{Upper 95\% } b &= b + se(z) = b + se(1.96) \end{aligned} \quad \text{Equation 4.17}$$

---

<sup>11</sup>The difference between our result and that shown in Table 4.12 is due to rounding error.

### Working It Out: South Region

$$\begin{aligned} W^2 &= \left( \frac{b}{se_b} \right)^2 \\ &= \left( \frac{0.805}{0.332} \right)^2 \\ &= 5.879 \end{aligned}$$

To determine whether this coefficient is statistically significant, we can refer to the  $\chi^2$  table for 1 degree of freedom. The number of degrees of freedom for an individual variable in a logistic regression will always be 1. Looking at Appendix 2, we see that a  $\chi^2$  of 10.827 is required for a result to be statistically significant at the 0.001 level. A  $\chi^2$  of 6.635 is required at the 0.01 level, and a  $\chi^2$  of 5.412 at the 0.02 level. Our observed significance level can therefore be defined as falling between 0.01 and 0.02. The output gives the exact observed significance level as 0.015. Using conventional levels of statistical significance, we would conclude that we can reject the null hypothesis that there is no difference in the reported implementation of Compstat-like programs in the South versus the North Central region.

Looking at the significance statistics column in Table 4.13, we can see that the number of sworn officers is also statistically significant—in this case, at greater than the 0.001 level. It is important to note that the statistics reported in this table, as well as in most statistical software, are for two-tailed significance tests. We mentioned at the outset that there was a strong hypothesis that larger departments are more likely to report a Compstat-like program. If we wanted to use a directional test of statistical significance, we would simply divide the observed significance level in our test by 2.

Looking at the other region dummy variables, we can see that there is not a statistically significant difference between the Northeast and North Central regions or between the West and North Central regions. But, as noted in Chap. 2, it is important to ask whether the variable region overall contributes significantly to the regression. To test this hypothesis, we can conduct a **likelihood ratio chi-square test**, which compares the log-likelihood function of the model with the multicategory nominal variable (the full model) with the log-likelihood function of the model without the multicategory nominal variable (the reduced model). Equation (4.18) details the likelihood ratio chi-square test. The number of degrees of freedom is defined as the number of dummy variables added by the

multicategory nominal variable. In our case, it is 3 for the three included regions.

$$\text{Likelihood ratio } \chi^2 = (-2LL_{\text{reduced model}}) - (-2LL_{\text{full}})$$

**Equation 4.18**

We can get the statistics for the test by running two separate regressions. The reduced model regression excludes the dummy variables associated with region. The  $-2LL$  for this model is 499.447 compared to the value of 492.513 for the full model that we have used previously (see Table 4.12). The latter is based on the model we have been using throughout the chapter, with the region dummy variables included, whereas the former excludes these dummy variables.

Below, we work out the likelihood ratio  $\chi^2$  using these two estimates. The likelihood ratio  $\chi^2$  for the region variable is 6.934, with 3 degrees of freedom (the number of dummy variable measures included in the model). Looking at Appendix 2, we can see that with 3 degrees of freedom, a  $\chi^2$  of 7.815 would be needed to reject the null hypothesis of no relationship between region and a reported Compstat-like program at the 0.05 significance threshold. Because our  $\chi^2$  statistic is smaller than this number, we cannot conclude that there is overall a statistically significant relationship between region and claimed development of a Compstat-like program.

### Working It Out

$$\begin{aligned}\text{Likelihood ratio } \chi^2 &= (-2LL_{\text{reduced model}}) - (-2LL_{\text{full}}) \\ &= 499.447 - 492.513 \\ &= 6.934\end{aligned}$$

## Chapter Summary

Ordinary least squares regression is not an appropriate tool for analyzing a problem in which the dependent variable is dichotomous. In such cases, OLS regression is likely to predict values that are greater than 1 and less than 0 and thus outside the observed distribution of  $y$ . Using the OLS approach in this case will also lead to violations of parametric assumptions required for associated statistical tests. **Logistic regression analysis** uses a **logistic model curve**, rather than a straight line, to predict outcomes for

$y$  in the case of a dichotomous dependent variable. This constrains predictions to between 0 and 1.

While the logistic model curve provides a solution to predictions beyond the observed distribution, the outcome variable is transformed into the **natural logarithm of the odds of  $y$** , or the **logit of  $y$** . Through use of the **cumulative logistic probability function**, the logistic regression equation may be used to predict the likelihood of  $y$  occurring. **Maximum likelihood estimation techniques** are used to estimate the coefficients in a logistic regression analysis. In this approach, we begin by identifying a tentative solution, which we then try to improve upon. Our criterion for improvement is termed the **log-likelihood function ( $-2LL$ )**. We repeat this process again and again until the change in the likelihood function is considered negligible. Each time we repeat the process and re-estimate our coefficients this is called an **iteration**. **Lack of convergence** in a standard number of iterations indicates some type of problem in the regression model that is being estimated.

The multiple **logistic regression coefficient**,  $b$ , may be interpreted as the increase in the log of the odds of  $y$  associated with a one-unit increase in  $x$  (with all other independent variables in the model held constant). The **odds ratio**, or **Exp(B)**, and the **derivative at mean, DM**, provide more easily interpreted representations of the logistic regression coefficient. The odds ratio represents the impact of a one-unit change in  $x$  on the ratio of the probability of  $y$ . Like an ordinary regression coefficient, the derivative at mean may be interpreted as the change in  $y$  associated with a unit change in  $x$ . The DM will change depending on the mean value of  $y$  in the problem examined.

There is no widely accepted method for comparing logistic regression coefficients measured on different scales. One method is to compare probability estimates at selected intervals. **Standardized logistic regression coefficients** have been suggested for logistic regression, though they should be interpreted with caution. There is no single widely accepted statistic for assessing how well the logistic regression model predicts the observed data. An approach commonly used is to calculate the **percent of correct predictions**. This method establishes an arbitrary decision point (usually 0.50) for deciding when a predicted value should be set at 1. These predictions are then compared to the observed data. **Pseudo- $R^2$**  statistics have also been developed, though they remain a subject of debate.

Statistical significance for the overall logistic regression model is assessed through computation of the **model chi-square**. Statistical significance for individual regression coefficients is evaluated with the **Wald statistic**. A **likelihood ratio chi-square test** can be used to calculate the statistical significance of a multicategory nominal variable.

## Key Terms

**Cox and Snell's  $R^2$**  A commonly used pseudo- $R^2$  measure whose main component, as in other pseudo- $R^2$  statistics, is the log-likelihood function ( $-2LL$ ).

**Cumulative logistic probability function** A transformation of the logistic probability function that allows computation of the probability that  $y$  will occur, given a certain combination of characteristics of the independent variables.

**Derivative at mean (DM)** A measure that converts the nonlinear logistic regression coefficient to a simple linear regression coefficient, which may be interpreted as the change in  $y$  associated with a unit change in  $x$ .

**Iteration** Each time we identify another tentative solution and re-estimate our logistic regression coefficients.

**Lack of convergence** Failure of a logistic regression analysis to reach a result that meets the criterion of reduction in the log-likelihood function.

**Likelihood ratio chi-square test** A test for statistical significance that allows the researcher to examine whether a subset of independent variables in a logistic regression is statistically significant. It compares  $-2LL$  for a full model to  $-2LL$  for a reduced model.

**Log-likelihood function ( $-2LL$ )** A measure of the probability of observing the results in the sample, given the coefficient estimates in the model. In logistic regression, the log-likelihood function ( $-2LL$ ) is defined as  $-2$  times the natural logarithm of the likelihood function.

**Logarithm** The power to which a fixed number (the base) must be raised to produce another number.

**Logistic model curve** The form of the predicted outcomes of a logistic regression analysis. Shaped like an S, the logistic curve begins to flatten as it approaches 0 or 1, so it keeps coming closer to—but never actually reaches—either of these two values.

**Logistic regression analysis** A type of regression analysis that allows the researcher to make predictions about dichotomous dependent variables in terms of the log of the odds of  $y$ .

**Logistic regression coefficient** The coefficient  $b$  produced in a logistic regression analysis. It may be interpreted as the change in the log of the odds of  $y$  associated with a one-unit increase in  $x$ .

**Maximum likelihood estimation** A technique for estimating the parameters or coefficients of a model that maximizes the probability that the estimates obtained will produce a distribution similar to that of the observed data.

**Model chi-square ( $\chi^2$ )** The statistical test used to assess the statistical significance of the overall logistic regression model. It compares the  $-2LL$  for the full model with the  $-2LL$  calculated without any independent variables included.

**Nagelkerke  $R^2$**  A pseudo- $R^2$  statistic that corrects for the fact that Cox and Snell's estimates, as well as many other pseudo- $R^2$  statistics, often have a maximum value of less than 1.

**Natural logarithm of the odds of  $y$  (logit of  $y$ )** The outcome predicted in a logistic regression analysis.

**Odds ratio [Exp(B)]** A statistic used to interpret the logistic regression coefficient. It

represents the impact of a one-unit change in  $x$  on the ratio of the probability of  $y$ .

**Percent of correct predictions** A statistic used to assess how well a logistic regression model explains the observed data. An arbitrary decision point (usually 0.50) is established for deciding when a predicted value should be set at 1, and then the predictions are compared to the observed data.

**Pseudo- $R^2$**  The term generally used for a group of measures used in logistic regression to create an approximation of the OLS

regression  $R^2$ . They are generally based on comparisons of  $-2\text{LL}$  for a full model and a null model (without any independent variables).

**Standardized logistic regression coefficient** A statistic used to compare logistic regression coefficients that use different scales of measurement. It is meant to approximate Beta, the standardized regression coefficient in OLS regression.

**Wald statistic** A statistic used to assess the statistical significance of coefficients in a logistic regression model.

## Symbols and Formulas

---

$e$  Base of the natural logarithm

$\ln$  Natural logarithm

$W^2$  Wald statistic

$-2\text{LL}$   $-2$  times the log-likelihood function

The logit for  $p$ , or  $P(Y = 1)$ :

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = b_0 + b_1x_1$$

To calculate the probability that  $y = 1$ :

$$P(Y = 1) = \frac{1}{1 + e^{-xb}}$$

To calculate the odds for  $P(Y = 1)$  at a given level in the independent variable  $x$ :

$$\text{odds} = \frac{p}{1-p}$$

To calculate the odds ratio for  $P(Y = 1)$ , given a one-unit change in the independent variable  $x$ :

$$\text{odds ratio} = \frac{\text{odds}_{x_i+1}}{\text{odds}_{x_i}}$$

To calculate the derivative at mean:

$$\text{DM}_i = \bar{y}(1 - \bar{y})b_i$$

To calculate the standardized logistic regression coefficient:

$$\text{Beta} = b_i s_i$$

To calculate the percent of correct predictions:

$$\text{Percent of correction predictions} = \left( \frac{n_{\text{correct predictions}}}{n_{\text{total}}} \right) \times 100$$

To calculate Cox and Snell's  $R^2$ :

$$\text{Cox and Snell's } R^2 = 1 - e^{-[(-2LL_{\text{null}}) - (-2LL_{\text{full}})]/n}$$

To calculate the model chi-square:

$$\text{Model } \chi^2 = (-2LL_{\text{null}}) - (-2LL_{\text{full}})$$

To calculate the Wald statistic:

$$W^2 = \left( \frac{b}{se_b} \right)^2$$

To calculate the likelihood ratio chi-square statistic for a subset of independent variables:

$$\text{Likelihood ratio } \chi^2 = (-2LL_{\text{reduced model}}) - (-2LL_{\text{full}})$$

## Exercises

---

- 4.1. As part of a research project for a class, a student analyzed data on a sample of adults who had been asked about their decision to report being assaulted to the police. Their decision was coded as 1 = assault reported to the police and 0 = assault not reported to the police. The student used ordinary least squares regression to estimate the effects of age (in years), sex (0 = male, 1 = female), and race (0 = white, 1 = nonwhite). The student reported the regression results as follows:

VARIABLE	<i>b</i>
Age	0.01
Sex	0.5
Race	-0.2
Constant	-0.1

- (a) Calculate the predicted values for each of the following persons:
- A 65-year-old white female
  - A 25-year-old nonwhite male
  - A 40-year-old white male
  - A 30-year-old nonwhite female
- (b) Should any of the student's predicted values lead the student to question the use of ordinary least squares regression? Explain why.
- 4.2. A research institute concerned with raising public attention about the use of force by school children calculates the following effects on the likelihood of hitting another child at school using logistic regression analysis:

VARIABLE	<i>b</i>
Sex (0 = girl, 1 = boy)	0.7
Grade in school	-0.1
Constant	-0.4

Hitting another child was coded as 1; no hitting was coded as 0.

- (a) Interpret the effects of sex and grade in school on the log of the odds that  $P(Y = 1)$ .
- (b) Calculate and interpret the odds ratios for the effects of sex and grade in school on use of force.

- 4.3. Supervision of defendants on pretrial release is thought to reduce the chance that defendants will flee the community. A government agency funds a small study to examine whether supervision affects pretrial flight (flight = 1, no flight = 0) and reports the following logistic regression results:

VARIABLE	<i>b</i>	SE
Age (years)	-0.01	0.02
Sex (1 = male, 0 = female)	0.67	0.25
Severity of offense scale (0–10)	0.21	0.03
Number of prior felony convictions	0.35	0.09
Number of contacts with supervision caseworker	-0.13	0.03
Constant	-0.52	

- (a) Calculate and interpret the odds ratio for each of the independent variables.
- (b) Can the government agency conclude that supervision in the form of contact with a caseworker has a statistically significant effect on pretrial flight? Explain why.
- (c) If the agency reports that the  $-2\text{LL}_{\text{null}}$  is 653.2 and the  $-2\text{LL}_{\text{full}}$  is 597.6, can it conclude that the model is statistically significant? Explain why.
- 4.4. A survey of adolescents indicated that 17% had used marijuana in the last year. In addition to standard demographic predictors of drug use, a researcher expects that school performance also affects the likelihood of marijuana use. The researcher's table of results follows:

VARIABLE	MEAN	STANDARD DEVIATION	<i>b</i>	SE
Age (years)	14.6	3.1	0.07	0.03
Sex (1 = male, 0 = female)	0.55	0.50	0.36	0.15
Race (1 = white, 0 = nonwhite)	0.75	0.43	-0.42	0.30
Grade point average	2.76	1.98	-0.89	0.24
Think of self as a good student (1 = yes, 0 = no)	0.59	0.49	-0.65	0.33
Constant			-0.87	

- (a) Calculate the predicted probability of marijuana use for each of the following persons:
- A 14-year-old white male who does not think of himself as a good student and has a GPA of 3.07.
  - A 17-year-old nonwhite female who thinks of herself as a good student and has a GPA of 3.22.
  - A 15-year-old white female who thinks of herself as a good student and has a GPA of 2.53.

- (b) Calculate the standardized logistic regression coefficient for each of the independent variables in the model. Which variable appears to have the largest effect on marijuana use?
- (c) Calculate the derivative at mean for each of the independent variables in the model. Which variable appears to have the largest effect on marijuana use?
- (d) Compare your answers for parts (b) and (c). How do you explain this pattern?
- 4.5. After losing a court battle over a requirement that it reduce its jail population, a county conducted an analysis to predict which offenders would pose the greatest threat of committing a violent offense if released early. A random sample of 500 inmates released from the jail in the last 3 years was analyzed to see what factors predicted arrest for a violent crime in the 12 months after release. For the final model, which included five predictors of violent arrest, the county reported the following statistics:
- $$-2LL_{\text{null}} = 876.5$$
- $$-2LL_{\text{full}} = 861.3$$
- |                            | PREDICTED NO VIOLENT ARREST | PREDICTED VIOLENT ARREST |
|----------------------------|-----------------------------|--------------------------|
| Observed no violent arrest | 439                         | 19                       |
| Observed violent arrest    | 27                          | 15                       |
- (a) Calculate the percent correctly predicted for this model. What does this statistic indicate about the county's prediction model?
- (b) Calculate the model chi-square for this model. Interpret this statistic.
- (c) Calculate Cox and Snell's  $R^2$  for this model. Interpret this statistic.
- (d) How do you explain the difference in the results for parts (a) through (c)?
- 4.6. Hoping that media attention to wrongful convictions has increased public opinion in favor of abolishing the death penalty, an abolitionist organization conducts a study to assess public support for abolishing the death penalty. Overall, the organization finds that 35% would support abolishing the death penalty if offenders could be sentenced to life without the option of parole (coded as 1 = abolish the death penalty, 0 = do not abolish the death penalty). In a logistic regression model examining the effects of respondent characteristics on support, the organization finds the following:

VARIABLE	MEAN	STANDARD DEVIATION	b	SE
Age (years)	41.2	15.4	-0.01	0.01
Sex (1 = male, 0 = female)	0.44	0.50	-0.42	0.19
Race (1 = white, 0 = nonwhite)	0.76	0.43	-0.24	0.09
Political conservative (1 = yes, 0 = no)	0.33	0.47	-1.12	0.22
Region of Country:				
South	0.23	0.42	-0.19	0.11
West	0.31	0.46	-0.09	0.04
North (omitted = Central)	0.27	0.44	0.27	0.12
Constant				0.11

- (a) Which variable has a relatively greater impact on support for abolishing the death penalty? Explain why.
- (b) If  $-2\text{LL}_{\text{reduced}} = 376.19$  and  $-2\text{LL}_{\text{full}} = 364.72$  when the region variables are omitted from the analysis, do the region variables have a statistically significant effect on support for abolishing the death penalty?

## Computer Exercises

In SPSS, Stata, and R, the general format for many of the multivariate statistical models is much the same. So, while the models may become increasingly complex, the syntax necessary to run these models is oftentimes very similar. We have included examples below and provided syntax files for Stata (Chapter\_4.sps) and SPSS (Chapter\_4.do).

### SPSS

Logistic regression analyses are performed in SPSS with the LOGISTIC REGRESSION command:

```
LOGISTIC REGRESSION VARIABLES dep_var_name
/METHOD = ENTER list_of_indep_vars.
```

The structure to this command is identical to the REGRESSION command discussed in previous chapters. Much of the output from running this command has been discussed in this chapter. The difference between the linear regression command and the logistic regression command is that the output from a logistic regression presents information for the model that includes only the intercept and is labeled *Block 0* in the SPSS output. The next section of output is labeled *Block 1*, and it contains the results discussed above: Omnibus Tests of Model Coefficients, Model Summary, Classification Table, and Variables in the Equation.

It is possible to have SPSS calculate the predicted probability and residual for each observation in the data file. To obtain one or both of these values, insert the / SAVE PRED RESID line, just as in the linear regression command—the naming

convention in the LOGISTIC REGRESSION is the same as that used in the REGRESSION command.

### Stata

Logistic regression analyses are performed in Stata with the **logit** command:

```
logit dev_var list_of_indep_vars
```

The output from running the **logit** command parallels that in Stata's **regress** command. The model summary results appear at the top of the output, followed by a table that presents the coefficients, their standard errors,  $\chi^2$ -scores, and confidence intervals (we discuss these in Chap. 5). Note that rather than the Wald statistic for each variable, Stata computes the  $\chi^2$ -score. To obtain the Wald statistic, simply square the  $\chi^2$ -score:

$$\text{Wald} = \chi^2$$

which will result in the same substantive conclusion (i.e., a significant Wald statistic will also be a significant  $\chi^2$ -score).

To obtain the odds ratios, we need to add the option **or** to the **logit** command:

```
logit dev_var list_of_indep_vars, or
```

Note that the table of coefficients will report the odds ratios rather than the coefficients interpretable as the log of the odds. To specify categorical variables, add **i.** in front of the variable name and add two **##** symbol between variable names to specify interactions:

```
logit dev_var indep_var1 i.indep_var2  
indep_var1##i.indep_var2, or
```

In the model summary statistics, Stata reports the test for the overall model as  $LR\ chisq(2)$  and what is labeled as *Pseudo-R<sup>2</sup>*. Recall that in a  $\chi^2$  test in Stata, the degrees of freedom associated with a test is included within the parentheses following *chi2*. The *Pseudo-R<sup>2</sup>* value is the Cox and Snell's  $R^2$ .

Saving predicted values and residuals following the **logit** command is identical to saving these same values after running the **regress** command:

```
predict PRE_1  
predict RES_1, r
```

which will then add two additional variables to the working data file.

The **margins** command allows for the calculation of marginal effects (or partial effects) in Stata. The command is for post-estimation such as after a logistic regression model has been fit. The **dydx(\*)** option can be added for discrete variables. This can be done as follows:

quietly logit dev\_var list\_of\_indep\_vars, **or**  
**margins, dydx(\*)**

Predictive margins can be obtained for specific variable(s) as well:

**margins, i.indep\_var2**  
**margins, indep\_var1##i.indep\_var2**  
**margins, i.indep\_var2, over(indep\_var1)**

And these can be plotted with the **marginsplot** command:

**marginsplot**

## R

Logistic regression analyses can be performed in R with the **glm()** function and setting the family argument to binomial:

```
glm(dev_var ~ ind_var1 + ind_var2,
    data = dataset_name,
    family = binomial)
```

The model provides a table that presents the coefficients, their standard errors, z-scores, and probability values. You can obtain more information about the model fit by assigning the regression model to an object and then using the **summary()** function as follows:

```
model <- glm(dev_var ~ ind_var1 + ind_var2,
            data = dataset_name,
            family = binomial)
summary(model)
```

You can also use the **coefficient()** function to have only the model coefficients be printed. To obtain the odds ratios for the coefficients, we nest the **coefficients()** function within the **exp()** function. This function tells R to exponentiate the coefficients.

**exp(coefficients(model))**

The predicted values and residuals for your model can be easily saved respectively using the **predict()** and **residuals()** functions.

**predict(model)**  
**residuals(model)**

Another means of computing logistic regression in R is to use the **lmr()** function within the *rms* package. Note that this function also prints the Nagelkerke  $R^2$  for the model.

```
lmr(dev_var ~ ind_var1 + ind_var2,
    data = dataset_name)
```

Similar to Stata, the **margins()** function allows for the calculation of marginal effects for post-estimation, such as after a logistic regression model has been fit. It requires the installation of the *margins* package. Once the package is installed, it can be used along with the **summary()** function as follows:

```
model<-glm(dev_var ~ ind_var1 + ind_var2, data =
            dataset_name, family=binomial)
summary(margins(model))
```

Predictive margins can be obtained for specific variable(s) as well with the **variables=** argument:

```
summary(margins(model, variables="ind_var1"))
summary(margins(model,
variables=c("ind_var1", "ind_var2"))
```

And these can be plotted with the **plot()** function:

```
plot(margins(model))
```

### Problems

1. Open the Compstat data file (Compstat.sav or Compstat.dta). These are the data analyzed in this chapter. Use one of the binary logistic regression commands with Compstat as the dependent variable and number of sworn officers, Northeast, South, and West as the independent variables. Note that the values reported in the software output match those reported in the text in Tables 4.5, 4.6, 4.11, and 4.12.
2. Open the Pennsylvania Sentencing data file (pcs\_98.sav or pcs\_98.dta). Use one of the binary logistic regression commands with incarceration as the dependent variable and age, race, sex, offense severity score, and prior criminal history score as the independent variables.
  - (a) Explain the logistic regression coefficients in plain English.
  - (b) Explain the odds ratios for each independent variable in plain English.
  - (c) Interpret the results of the Wald statistic for each of the logistic regression coefficients.
  - (d) Interpret the value of Cox and Snell's  $R^2$ .
  - (e) Perform a chi-square test for the overall regression model.
3. Open the NYS data file (nys\_1.sav, nys\_1\_student.sav, or nys\_1.dta). Using one of the binary logistic regression commands, run an analysis for each of the measures of delinquency below. As in the Computer Exercises in Chaps. 2 and 3, you will need to select a set of independent variables that you think are related to the dependent variable. Note that each of the delinquency items will

need to be recoded as 0 or 1 to represent whether or not the act was committed (see Chapter\_4.sps and Chapter\_4.do for examples). Do the following for each analysis:

- (a) Explain the logistic regression coefficients in plain English.
- (b) Explain the odds ratios in plain English.
- (c) Interpret the results of the Wald statistic for each of the logistic regression coefficients.
- (d) Interpret the value of Cox and Snell's  $R^2$ .
- (e) Perform a chi-square test for the overall regression model.
  - Number of thefts valued at less than \$5 in the last year; convert to any thefts in the last year.
  - Number of times drunk in the last year; convert to any times drunk in the last year.
  - Number of times the youth has hit other students in the last year; convert to any times the youth has hit other students in the last year.
  - Number of times the youth has hit a parent in the last year; convert to any times the youth has hit a parent in the last year.

## References

---

- Cox, D. R., & Snell, E. J. (1989). *The Analysis of Binary Data* (2nd ed.). London: Chapman and Hall.
- Field, A. (2000). *Discovering statistics using SPSS for windows*. London: Sage.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models* (pp. 54–57). Cambridge: Cambridge University Press.
- Hanushek, E. A., & Jackson, J. E. (1977). *Statistical methods for social scientists*. New York: Academic Press.
- Hosmer, D. W., & Lemeshow, S. (2000). *Applied regression analysis* (2nd ed.). New York: Wiley.
- Kaufman, R. L. (1996). Comparing effects in dichotomous logistic regression: A variety of standardized coefficients. *Social Science Quarterly*, 77, 90–109.
- Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78, 691–692.
- Scott Long, J. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks: Sage.
- Tukey, J. (1997). *Report to the Special Master* (p. 5); *Report to the New Jersey Supreme Court 27*.
- Weisburd, D., Mastrofski, S., McNally, A. M., & Greenspan, R. (2001). *Compstat and organizational change*. Washington, DC: The Police Foundation.
- Wheeler, S., Weisburd, D., & Bode, N. (1982). Sentencing the white collar offender: Rhetoric and reality. *American Sociological Review*, 47, 641–659.

## Chapter five

---

# Multiple Regression with Multiple Category Nominal or Ordinal Measures

## Multinomial Logistic Regression

---

How do we Analyze a Nominal Dependent Variable with More than Two Categories?

How Do We Interpret the Multinomial Logistic Regression Model?

Does the Reference Category Make a Difference?

What is the Test of Statistical Significance for Single Coefficients?

What is the Test of Statistical Significance for Multiple Coefficients?

What are the Practical Limits of the Multinomial Logistic Regression Model?

## Ordinal Logistic Regression

---

How Do We Interpret the Ordinal Logistic Regression Model?

How Do We Interpret Cumulative Probabilities?

How are Cumulative Probabilities Related to Odds Ratios?

How Do We Interpret Ordinal Logistic Coefficients?

What is the Test of Statistical Significance for Coefficients?

What are the Parallel Slopes Tests?

How is the Score Test Computed?

How is the Brant Test Computed?

What is the Partial Proportional Odds Model?

How Do We Interpret the Results from the Partial Proportional Odds Model?

**I**N THE PREVIOUS CHAPTER, we examined how to analyze data using a binary logistic regression model that included a dependent variable with two categories. This allowed us to overcome problems associated with using ordinary least squares (OLS) regression in cases where the variable that is being explained is measured as a simple dichotomy. Accordingly, we have now described tools that allow the researcher to develop explanatory models with either an interval (or ratio) dependent variable or a dichotomous dependent variable.

But there are many situations in which researchers are faced with analyzing dependent variables that include more than two categories or that are measured on an ordinal scale. For example, we may want to identify the factors that lead to a dismissal, a guilty plea conviction, or a trial conviction in court. The methods we have covered so far do not allow us to examine this problem in a single statistical model with multiple independent variables. We have also not discussed how a researcher should deal with dependent variables such as fear of crime, which are measured on an ordinal scale. As we noted in Chap. 2, the assumptions of ordinary least squares regression will often not be met when using ordinal scale dependent variables.

Fortunately, we can extend our discussion of the logistic regression model to consider such dependent variables. However, such logistic regression models need to be modified to take into account these new situations. In this chapter, we provide an introduction to multinomial and ordinal logistic regression. We think that the problems that these approaches address are common in criminal justice research and while simpler models can be used, with caveats, such as running a series of logistic regression models or using OLS regression, there are advantages to the methods discussed in this chapter. As such, it is important to understand and to be able to apply these methods.

## Multinomial Logistic Regression

---

**Multinomial logistic regression** is used to examine problems where there are more than two nominal categories in the dependent variable. We have already mentioned the case where a researcher wants to explain why convicted offenders are sentenced to prison, probation, or receive fines, but there are many situations in criminal justice in which dependent variables include multiple nominal categories. For example, a researcher may want to explain why certain offenders tend to specialize in either violent crime, property crime, or white-collar crime. Multinomial regression is particularly useful when researchers create categorizations for groups of offenders and then want to explain why certain people fall into those groups. This is common, for example, in studies in developmental criminology where offenders are placed into a small number of groups that evidence different crime trajectories.<sup>1</sup> It is then natural to ask why offenders fall into those groups. Multinomial regression provides a very useful tool for conducting multivariate analyses in such situations.

Multinomial regression is conceptually a straightforward extension of the binary logistic regression model that we discussed in the previous chapter. Recall that in the binary logistic regression model, we designated one of the two outcome categories as the presence of a given trait and the second as the absence of that trait. For example, we compared police departments that had adopted Compstat ( $y = 1$ ) versus those that did not adopt the Compstat program ( $y = 0$ ). This is also often conceptualized as success versus failure. In logistic regression, the left side of the regression equation is the natural logarithm (ln) of the odds of having a 1 on the dependent variable ( $y = 1$ ) as opposed to having a 0 ( $y = 0$ ). This transformation allowed us to develop a prediction model in which the predictions of the regression equation are constrained to fall between 0 and 1. We call this transformation the logit, as shown in Eq. (5.1).

$$\begin{aligned} \text{logit}(y = 1|x) &= \ln \left[ \frac{P(y = 1|x)}{P(y = 0|x)} \right] \\ &= b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k \end{aligned} \quad \text{Equation 5.1}$$

This equation is stating that the logit for  $y = 1$  is conditional on the independent variables ( $x$ ); that is, it is a function of the linear regression equation on the right side of the equation, where  $k$  is the number of independent variables. If we simplify this to a single binary independent

---

<sup>1</sup>See, for example, Nagin (2005).

variable with two conditions, such as treatment and control, this equation would state that the logit that  $y = 1$  is conditional on whether someone is in the treatment or control condition and  $b_1$  would reflect the difference in the logits between these two groups.

What happens when the outcome variable has more than two categories? The problem here is that we do not have a simple change in the odds for one outcome, as we did with the example of Compstat in the prior chapter. Here, we have to take into account changes in the odds in relation to more than two categories. The logit of  $y$  in Eq. (5.1) requires that there be only the absence ( $y = 0$ ) or the presence ( $y = 1$ ) of a trait. This situation is not appropriate when you have the possibility of the presence of more than one positive outcome (e.g., dismissal, a guilty plea conviction, or a trial conviction) and you want to distinguish among them. Of course, you could simply argue, for example, that you are only interested in whether individuals received a dismissal. However, you would not have the possibility in the simple logistic model to predict why they received alternatively a guilty plea conviction or a trial conviction. This is the problem that multinomial regression seeks to solve.

Suppose that our outcome variable has three categories (1, 2, and 3) with the number of observations in each being represented by  $n_1$ ,  $n_2$ , and  $n_3$ . We could begin by estimating three binary logistic regression models that would allow for all possible comparisons of the outcome categories—the logits for categories 1 versus 2, 2 versus 3, and 1 versus 3. The logit of  $y$  for each regression could be written simply as:

$$\ln \left[ \frac{P(y = 1|x)}{P(y = 2|x)} \right],$$

$$\ln \left[ \frac{P(y = 2|x)}{P(y = 3|x)} \right],$$

and

$$\ln \left[ \frac{P(y = 1|x)}{P(y = 3|x)} \right]$$

for each comparison, respectively.

Interestingly, these three logits can be linked in what can be defined as an identity equation that illustrates how knowledge of any two logits can produce the values of the third. The identity equation<sup>2</sup> can be stated as:

$$\ln \left[ \frac{P(y = 1|x)}{P(y = 2|x)} \right] + \ln \left[ \frac{P(y = 2|x)}{P(y = 3|x)} \right] = \ln \left[ \frac{P(y = 1|x)}{P(y = 3|x)} \right].$$

If we were to estimate these three separate logistic regression models, the coefficients would be interpreted in the same way as we described in Chap. 4. While this approach would allow us to make comparisons of the likelihood of subjects falling in each of the three categories examined as compared to each other, it would require us to run three separate regressions. Moreover, and more importantly from a statistical point of view, we would likely be working with three completely different samples in each of the three analyses:  $n_1 + n_2$ ,  $n_2 + n_3$ ,  $n_1 + n_3$ . This is because the cases on the dependent variable are unlikely to be distributed evenly. For example, we would not expect sentences for 300 offenders to be distributed with exactly 100 in each group (e.g., dismissal, a guilty plea conviction, or a trial conviction). Given this, each of our comparisons would be based on different samples. In comparing  $n_1$  to  $n_2$ , we would have only defendants who were in outcome categories 1 and 2. Defendants that had outcome category 3 would not be included in that comparison. But what we are really interested in is the choice among the three outcomes and how this choice is distributed in our entire sample. The statistical problem here is that the varying sample sizes would then result in incorrect standard errors for the coefficients, leading to inaccurate tests of statistical significance. The multinomial logistic regression model simultaneously accounts for these

---

<sup>2</sup>You can verify this identity by using the fact that the logarithm of a fraction is equal to the logarithm of the numerator minus the logarithm of the denominator:

$\ln(x/y) = \ln(x) - \ln(y)$ . Specifically, for this equality, we note that

$$\ln \left[ \frac{P(y=1)}{P(y=2)} \right] = \ln [P(y=1)] - \ln [P(y=2)]$$

and

$$\ln \left[ \frac{P(y=2)}{P(y=3)} \right] = \ln [P(y=2)] - \ln [P(y=3)]$$

When we put these two pieces together in a single equation, we have

$$[\ln [P(y=1)] - \ln [P(y=2)]] + [\ln [P(y=2)] - \ln [P(y=3)]]$$

$$\ln [P(y=1)] - \ln [P(y=2)] + \ln [P(y=2)] - \ln [P(y=3)]$$

$$\ln [P(y=1)] - \ln [P(y=3)]$$

$$\ln \left[ \frac{P(y=1)}{P(y=3)} \right]$$

which establishes the equality. We explain the practical implication of this equality in our discussion of the interpretation of the coefficients from a multinomial logistic regression model.

different sample sizes, ensuring a more valid estimate of significance levels as well as having more statistical power. It also has the benefit of allowing us to conduct our analysis using only one regression model.

### A Substantive Example: Case Dispositions in California

The State Court Processing Statistics database includes information on random samples of individuals arrested for felony offenses in the largest court districts in the USA. To illustrate the application of the multinomial logistic regression model, we focus on a random sample of 10,230 felony arrestees in California in the 1990s. A question of both policy and theoretical relevance is the study of the factors that affect the type of case disposition (outcome)—whether a dismissal, a guilty plea conviction, or a trial conviction.<sup>3</sup>

A first step in a multinomial regression is to define a *reference category*. This is necessary because we need to decide which category we want to use as a baseline. It is an arbitrary decision about which category is designated the reference category, but to the extent that we can make a choice that has some theoretical relevance, it makes the interpretation of the results simpler. For case disposition, suppose that we choose dismissal as the reference category, which then allows us to make two comparisons between a type of conviction. Specifically, our multinomial logistic regression results will indicate (1) the relative likelihood of a guilty plea conviction compared to a dismissal and (2) the relative likelihood of a trial conviction compared to a dismissal. The one comparison that is omitted is the relative likelihood of a guilty plea conviction compared to a trial conviction. In the multinomial logistic regression model, this comparison is not directly estimated, but as we illustrate shortly, the results can be obtained very simply from the results for the comparison of each conviction type to a dismissal. That is, this information is contained in the model just not in an obvious way.

We will use the subscript  $m$  to represent the categories and  $j$  to equal the number of categories. In our example,  $m = 1, 2, \text{ or } 3$  and we will designate category 1 as the reference category. Using these subscripts, we can write the multinomial regression equation for each of the three categories with five independent variables as follows:

---

<sup>3</sup>While it may appear odd at first glance that we have not included those individuals who were acquitted at a trial, there were very few individuals who fell into this category. Like most jurisdictions, courts in California acquit relatively few individuals through a trial—it was about 1% of all cases in the 1990s. What this means is that once the prosecutor has filed charges against a defendant, rather than dismiss the case, it will likely result in the conviction of the defendant through either a guilty plea or a trial conviction. This also implies that a dismissal of the case functions much like an acquittal, but one made by the prosecuting attorney rather than a judge or jury.

$$\begin{aligned}
 \text{logit}(y = m|x) &= \ln \left[ \frac{P(y = m|x)}{P(y = 1|x)} \right] \\
 &= b_{0,m} + b_{1,m}x_1 + b_{2,m}x_2 + b_{3,m}x_3 + b_{4,m}x_4 \\
 &\quad + b_{5,m}x_5
 \end{aligned} \tag{Equation 5.2}$$

We have  $y = 1$  in the denominator of the logit because category 1 is the reference category. If category 3 was the reference category, then this would read  $y = 3$ . If we write out Eq. (5.2) for each category separately, we get as follows:

$$\begin{aligned}
 \text{logit}(y = 1|x) &= \ln \left[ \frac{P(y = 1|x)}{P(y = 1|x)} \right] = 0 \\
 \text{logit}(y = 2|x) &= \ln \left[ \frac{P(y = 2|x)}{P(y = 1|x)} \right] \\
 &= b_{0,2} + b_{1,2}x_1 + b_{2,2}x_2 + b_{3,2}x_3 + b_{4,2}x_4 + b_{5,2}x_5 \\
 \text{logit}(y = 3|x) &= \ln \left[ \frac{P(y = 3|x)}{P(y = 1|x)} \right] \\
 &= b_{0,3} + b_{1,3}x_1 + b_{2,3}x_2 + b_{3,3}x_3 + b_{4,3}x_4 + b_{5,3}x_5
 \end{aligned}$$

Why is the right side of the first of these equal to 0? Notice that the numerator and denominator for the logit are the same, that is, both are the probability of category 1 given the data, producing a fraction that equals 1. The natural log of 1 is 0. Conceptually, this should make sense as this model is comparing category 1 to itself. This illustrates that the number of models will also be one less than the number of categories. Also notice that we have two sets of regression coefficients, one for the model between categories 2 and 1, and another for the model between 3 and 1. Equation (5.2) shows the connection of multinomial logistic regression to both ordinary linear regression models and logistic regression models. The right side is a linear additive function of the independent variables. The difference is in the connection of that model to the dependent variable.

Equation (5.2) can also be expressed as a probability equation and as a conditional odds ratio equation. To simplify the notation, we will use  $xb_m$  to represent the right side of Eq. (5.2). The probability equation expression of Eq. (5.2) is shown in Eq. (5.3) below.

$$\text{Probability Equation: } P(y = m) = \frac{e^{xb_m}}{\sum_{m=1}^j e^{xb_m}} \quad \text{Equation 5.3}$$

The numerator of Eq. (5.3) tells us to exponentiate the value of  $xb$  for category  $m$ . The denominator, in turn, tells us that we need to exponentiate the value of  $xb$  for all categories and then sum these values together. As we discovered above, the value of  $xb$  for the reference category is 0. The exponent of 0 is 1 (i.e.,  $e^0 = 1$  since  $\ln(1) = 0$ ). Below, we write out the probability equations for each of the three categories of our outcome.

$$\begin{aligned} P(y = 1) &= \frac{e^{xb_1}}{e^{xb_1} + e^{xb_2} + e^{xb_3}} = \frac{1}{1 + e^{xb_2} + e^{xb_3}} \\ P(y = 2) &= \frac{e^{xb_2}}{e^{xb_1} + e^{xb_2} + e^{xb_3}} = \frac{e^{xb_2}}{1 + e^{xb_2} + e^{xb_3}} \\ P(y = 3) &= \frac{e^{xb_3}}{e^{xb_1} + e^{xb_2} + e^{xb_3}} = \frac{e^{xb_3}}{1 + e^{xb_2} + e^{xb_3}} \end{aligned}$$

We provide the conditional odds ratio expression of Eq. (5.2) in Eq. (5.4). Note that this equation is also referred to as a relative risk ratio (Gould 2000). The conditional part of the name refers to the fact that the odds ratios are conditional on the reference category and are not the typical odds of  $y$  divided by the odds of not  $y$ . Although this equation may look more complicated, it uses the information in Eq. (5.3) for the probabilities of each category. The equation reduces to something less complex, because the denominators in the fraction in the middle of the equation cancel each other out. The only notable difference here is the notation of a second category by the subscript  $n$ . Thus, for any odds ratio that we may be interested in between categories  $m$  and  $n$ , Eq. (5.4) illustrates that it can be obtained from the respective probabilities.

$$\text{Conditional OR}_{m/n} = \frac{P(y = m)}{P(y = n)} = \frac{e^{xb_m}/\sum_{m=1}^j e^{xb_m}}{e^{xb_n}/\sum_{m=1}^j e^{xb_m}} = \frac{e^{xb_m}}{e^{xb_n}} \quad \text{Equation 5.4}$$

If we are interested in computing the odds ratio for a comparison between any category and the reference category ( $m = 1$  in our example), we obtain.

**Table 5.1**

Multinomial logistic regression coefficients

INDEPENDENT VARIABLE	TRIAL CONVICTION VS. DISMISSAL	GUILTY PLEA CONVICTION VS. DISMISSAL
Age (in years)	0.016	0.004
Sex (1 = male, 0 = female)	1.123	-0.013
Race (1 = nonwhite, 0 = white)	0.043	-0.266
Violent offense	0.657	-0.525
Number of charges	0.325	0.192
Intercept	-4.767	1.318

$$\text{Conditional OR}_{m/1} = \frac{P(y = m)}{P(y = 1)} = \frac{e^{xb_m}}{e^{xb_1}} = \frac{e^{xb_m}}{e^0} = \frac{e^{xb_m}}{1} = e^{xb_m}$$

This last result confirms how we are then to interpret the coefficients from the multinomial logistic regression model. Since the coefficients for the reference category have been fixed at 0, the coefficients for each of the remaining outcome categories will compare the relative likelihood of that category relative to the reference category.<sup>4</sup>

In practice, what these equations tell us is that we will have  $j - 1$  sets of coefficients from a multinomial logistic regression model that can be interpreted in the same way as binary logistic coefficients, where we compare each outcome ( $m$ ) to the reference category ( $m = 1$ ) for the outcome variable. In our example for case disposition, where we have designated dismissal as the reference category, one set of coefficients will give us the log of the odds or the conditional odds ratios comparing the likelihood of a guilty plea conviction relative to a dismissal, while the second set of coefficients will give us the log of the odds or the conditional odds ratios comparing the likelihood of a trial conviction relative to a dismissal.

Table 5.1 presents the results from our application of the multinomial logistic regression model. We have kept the multivariate model simple and used age, sex (male = 1, female = 0), race (nonwhite = 1, white = 0), type of crime (violent = 1, nonviolent = 0), and total number of charges as predictors of the type of case disposition for the sample of 10,230 arrestees in California in the 1990s. The first column lists all the independent variables, while the second and third columns present the coefficients for each

---

<sup>4</sup>It is worth pointing out that the binary logistic regression model presented in Chap. 4 is a special case of the multinomial logistic regression model, where  $m = 2$ . If you work through both Eqs. (5.3) and (5.4) assuming that  $m = 2$ , you will be able to replicate the equations in the previous chapter.

of the two comparisons: column 2 presents the comparison of trial conviction to dismissal, while column 3 presents the comparison of guilty plea conviction to dismissal.

The results in column 2 show that as age and the number of charges increase, the likelihood of a trial conviction increases relative to a dismissal. Similarly, defendants who are male, nonwhite, and charged with a violent offense are also more likely to be convicted at trial than to have their case dismissed. As in the previous chapter, we can also interpret each of these coefficients more directly as odds ratios. (Recall from the previous chapter that the exponentiation of the coefficient provides us with the odds ratio given a one-unit change in the independent variable.)

- If age is increased by 1 year, the odds of a trial conviction versus a dismissal increase by a factor of  $\exp(.016) = 1.016$ , controlling for all other variables in the model.
- The odds of a trial conviction versus a dismissal are  $\exp(1.123) = 3.074$  times higher for male than for female defendants, controlling for all other variables in the model.
- The odds of a trial conviction versus a dismissal are  $\exp(.043) = 1.044$  times higher for nonwhite than white defendants, controlling for all other variables in the model.
- The odds of a trial conviction versus a dismissal are  $\exp(.657) = 1.929$  times higher for defendants charged with a violent rather than a nonviolent offense, controlling for all other variables in the model.
- If the number of charges is increased by one, the odds of a trial conviction versus a dismissal increase by a factor of  $\exp(.325) = 1.384$ , controlling for all other variables in the model.

We can similarly interpret the results in column 3, which show that as age and number of charges increase, the likelihood of a guilty plea conviction relative to a dismissal increases. We also see from these results that defendants who are male, nonwhite, and charged with a violent offense will be less likely to be convicted with a guilty plea than to have their cases dismissed. Again, the direct interpretations of the coefficients would be the following:

- If age is increased by 1 year, the odds of a guilty plea conviction versus a dismissal increase by a factor of  $\exp(.004) = 1.004$ , controlling for all other variables in the model.
- The odds of a guilty plea conviction versus a dismissal are  $\exp(-.013) = .987$  times smaller for male than for female defendants, controlling for all other variables in the model.
- The odds of a guilty plea conviction versus a dismissal are  $\exp(-.266) = .766$  times smaller for nonwhite than white defendants, controlling for all other variables in the model.

- The odds of a guilty plea conviction versus a dismissal are  $\exp(-.525) = .592$  times smaller for defendants charged with a violent rather than a nonviolent offense, controlling for all other variables in the model.
- If the number of charges is increased by one, the odds of a guilty plea conviction versus a dismissal increase by a factor of  $\exp(.192) = 1.212$ , controlling for all other variables in the model.

### The Missing Set of Coefficients

As we noted earlier, when we estimate a multinomial logistic regression model, we obtain coefficients for all contrasts but one. In the example of case disposition, we are missing the contrast between guilty plea conviction and trial conviction. Based on the identity relationship of multiple logits that we described earlier in the chapter (see, also, Footnote 2), for all possible comparisons of the outcome categories, the most direct way of obtaining the missing coefficients is to simply subtract one set of coefficients from another set of coefficients. In Table 5.1, the results in column 2 represent the logit for *Trial Conviction vs. Dismissal*, while those in column 3 represent the logit for *Guilty Plea Conviction vs. Dismissal*.

Since the logarithm of a fraction can be rewritten as the subtraction of the logarithm of the denominator from the logarithm of the numerator, the logits can be rewritten as:

$$\ln \left( \frac{P(Y = \text{Trial Conviction})}{P(Y = \text{Dismissal})} \right) = \ln [P(Y = \text{Trial Conviction})] - \ln [P(Y = \text{Dismissal})]$$

and

$$\ln \left( \frac{P(Y = \text{Guilty Plea Conviction})}{P(Y = \text{Dismissal})} \right) = \ln [P(Y = \text{Guilty Plea Conviction})] - \ln [P(Y = \text{Dismissal})]$$

By performing simple subtractions of the logits, we can generate additional contrasts between the outcome categories. To obtain the missing coefficients for the comparison of Guilty Plea Conviction to Trial Conviction, we subtract the logit for Trial Conviction and Dismissal from the logit for Guilty Plea Conviction and Dismissal:

$$\begin{aligned}
 & \ln \left( \frac{P(Y = \text{Guilty Plea Conviction})}{P(Y = \text{Trial Conviction})} \right) \\
 &= \ln \left( \frac{P(Y = \text{Guilty Plea Conviction})}{P(Y = \text{Dismissal})} \right) - \ln \left( \frac{P(Y = \text{Trial Conviction})}{P(Y = \text{Dismissal})} \right) \\
 &= [\ln(P(Y = \text{Guilty Plea Conviction})) - \ln(P(Y = \text{Dismissal}))] \\
 &\quad - [\ln(P(Y = \text{Trial Conviction})) - \ln(P(Y = \text{Dismissal}))] \\
 &= [\ln(P(Y = \text{Guilty Plea Conviction})) - \ln(P(Y = \text{Trial Conviction}))]
 \end{aligned}$$

What this algebraic manipulation of logits shows us is that we can obtain the coefficients for the omitted contrast simply by subtracting one set of coefficients from another set of coefficients.

When applied to our case disposition coefficients, we obtain the results presented in Table 5.2. Here, we see that as age and the number of charges increase, the likelihood of a guilty plea conviction relative to a trial conviction decreases. Similarly, defendants who are male, nonwhite, and charged with a violent offense will be less likely to be convicted through a guilty plea than through a trial.

In regard to the direct interpretation of the coefficients, we have the following:

- If age is increased by one year, the odds of a guilty plea conviction versus a trial conviction decrease by a factor of  $\exp(-.012) = .988$ , controlling for all other variables in the model.
- The odds of a guilty plea conviction versus a trial conviction are  $\exp(-1.136) = .321$  times lower for male than for female defendants, controlling for all other variables in the model.
- The odds of a guilty plea conviction versus a trial conviction are  $\exp(-.309) = .734$  times lower for nonwhite than white defendants, controlling for all other variables in the model.

**Table 5.2**

Coefficients for the omitted contrast of guilty plea conviction vs. trial conviction through subtraction of column 3 from column 2

INDEPENDENT VARIABLE	GUILTY PLEA CONVICTION VS. DISMISSAL	TRIAL CONVICTION VS. DISMISSAL	GUILTY PLEA CONVICTION VS. TRIAL CONVICTION
Age (in years)	0.004	0.016	-0.012
Sex (1 = male, 0 = female)	-0.013	1.123	-1.136
Race (1 = nonwhite, 0 = white)	-0.266	0.043	-0.309
Violent offense	-0.525	0.657	-1.182
Number of charges	0.192	0.325	-0.133
Intercept	1.318	-4.767	6.085

**Table 5.3**

Multinomial logistic regression coefficients using trial conviction as the reference category (re-estimated model)

INDEPENDENT VARIABLE	DISMISSAL VS. TRIAL CONVICTION	GUILTY PLEA CONVICTION VS. TRIAL CONVICTION
Age (in years)	-0.016	-0.012
Sex (1 = male, 0 = female)	-1.123	-1.136
Race (1 = nonwhite, 0 = white)	-0.043	-0.309
Violent offense	-0.657	-1.182
Number of charges	-0.325	-0.133
Intercept	4.767	6.085

- The odds of a guilty plea conviction versus a trial conviction are  $\exp(-1.182) = .307$  times lower for defendants charged with a violent rather than a nonviolent offense, controlling for all other variables in the model.
- If the number of charges is increased by one, the odds of a guilty plea conviction versus a trial conviction decrease by a factor of  $\exp(-.133) = .875$ , controlling for all other variables in the model.

A second way to obtain the coefficients for the comparison of guilty plea conviction to trial conviction would be to simply redefine our statistical model so that trial conviction was chosen as the reference category and re-estimate our multinomial model. Upon rerunning the multinomial logistic regression model, we obtain the results presented in Table 5.3.

Note that column 2 contains the coefficients for the contrast between dismissal vs. trial conviction, which substantively gets at the same comparison that appears in Table 5.1, column 2, except that the direction of the effects are reversed, reflecting that the order of comparison was reversed. Again, what this indicates to us is that the selection of reference categories is arbitrary and that we will obtain the same substantive results, regardless of which category is selected. At the same time, we need to be aware of the selection of categories so that we correctly interpret our results.

Column 3 presents the contrast for guilty plea conviction relative to trial conviction. The results in this column are identical to those appearing in column 4 of Table 5.2, which were based on simply subtracting one set of coefficients from another.

Had we been interested in the contrast of *Trial Conviction* relative to *Guilty Plea Conviction*, we would have reversed the original order of subtraction (i.e., the coefficients in column 3 would have been subtracted from the coefficients in column 2). Then, the only difference that we would have seen in Table 5.2 is that the signs of the coefficients would have been reversed. Otherwise, the substantive interpretation of the results would be identical.

## Statistical Inference

### Single Coefficients

The results from a multinomial logistic regression analysis slightly complicate tests of statistical significance. Since we now have multiple coefficients for each independent variable, there are questions about how to discern whether an independent variable has an effect on the dependent variable as a whole. Specifically, there are two issues of statistical inference that are important for interpreting the results from a multinomial logistic regression analysis. For each coefficient, we can estimate the statistical significance of each category compared to the reference category. But we also can estimate the overall significance of the independent variable in predicting the multicategory dependent variable.

To test the effect of each individual coefficient in comparison to the reference category, we would again use the Wald statistic described in Chap. 4, with degrees of freedom equal to 1. As noted in Chap. 4, the Wald statistic has a chi-square distribution, and so the statistical significance of a result may be checked in the chi-square table. Table 5.4 presents the multinomial logistic regression coefficients from the original model along with the standard errors (se) of the coefficients and value of the Wald statistic ( $W$ ). Alternative, we can use the  $z$  statistic which is simply the square root of  $W$ , or  $z = \sqrt{W}$ . Some statistical software programs report  $W$ , whereas others report  $z$ .

If we set the significance level at 5%, the critical value of the Wald statistic with  $df = 1$  is 3.841 and for  $z$  the critical value is 1.96. We can see that the violent offense charge and number of charges have statistically significant effects for both pairs of outcomes. The demographic characteristics have varying effects, where age and sex have statistically significant effects on

**Table 5.4**

Multinomial logistic regression coefficients, standard errors, and Wald test results

MULTINOMIAL CONTRAST AND INDEPENDENT VARIABLES	<i>b</i>	SE	WALD	-95%	+95%
<i>Trial conviction vs. dismissal</i>					
Age	0.016	0.008	4.181	0.000	0.032
Male	1.123	0.373	11.350	0.392	1.854
Nonwhite	0.043	0.158	0.072	-0.267	0.353
Violent offense	0.657	0.160	16.834	0.343	0.971
Number of charges	0.325	0.043	58.524	0.241	0.409
Intercept	-4.767	0.438	118.726	-5.625	-3.909
<i>Trial conviction vs. dismissal</i>					
Age	0.004	0.003	2.110	-0.002	0.010
Male	-0.013	0.071	0.032	-0.152	0.126
Nonwhite	-0.266	0.053	24.827	-0.370	-0.162
Violent offense	-0.525	0.060	76.287	-0.643	-0.407
Number of charges	0.192	0.021	82.642	0.151	0.233
Intercept	1.318	0.124	112.754	1.075	1.561

the likelihood of a trial conviction compared to a dismissal, but race has a statistically significant effect on the likelihood of a guilty plea conviction compared to a dismissal.

We can also compute confidence intervals around each individual coefficient in the usual fashion using the standard error and the  $z$  critical value. For a 95% confidence interval, this would be a  $z$  of 1.96, as shown above, and the computation for the lower and upper limits are shown in Eq. (5.5).

$$\text{Lower 95\% } b = b - 1.96se$$

$$\text{Upper 95\% } b = b + 1.96se$$

**Equation 5.5**

### *Multiple Coefficients*

In Table 5.4, there are two coefficients for each independent variable. As we noted above, the number of coefficients from a multinomial logistic regression model for each independent variable will be one less than the number of categories on the dependent variable (i.e.,  $j - 1$ ). How do we assess the overall effect of each independent variable on the dependent variable? There are two key ways of doing this—one is a likelihood ratio test similar to the test we discussed in the previous chapter on binary logistic regression. The other test is an extension of the Wald test we have also already used. Regardless of the statistical software package one uses to estimate a multinomial logistic regression model, one of these two methods will be reported to test the overall effect of each independent variable.

The likelihood ratio (LR) test involves estimating the full multinomial logistic regression equation with all variables and then estimating reduced models that eliminate one independent variable from each analysis. The difference in the  $-2$  log-likelihood function for each equation will then allow for the test of each independent variable.

For example, in the case dismissal analysis, the value of the  $-2$  log-likelihood for the full model is 3625.670. When we estimate the same model, but eliminate the variable nonwhite from the analysis, the value of the  $-2$  log-likelihood is 3653.501. The difference of the two log-likelihood functions is  $3653.501 - 3625.670 = 27.831$ . By eliminating the variable for nonwhite, we have removed two coefficients from the analysis. If you refer again to Table 5.4, you will see that each independent variable appears twice to indicate its overall effect on case disposition—once to represent the effect on trial conviction versus dismissal and once more to represent the effect on guilty plea conviction versus dismissal. The degrees of freedom for the test will be  $df = 2$  to reflect the removal of the two coefficients. At a significance level of 5%, the critical value of the chi-square is 5.991. This means that we would conclude that the race of the defendant has a statistically significant effect on the type of case disposition.

**Table 5.5**

Likelihood ratio and Wald statistic results for the overall effect of each independent variable

INDEPENDENT VARIABLES	df	-2 LOG-LIKELIHOOD FOR REDUCED MODEL	LR TEST STATISTIC	WALD
Age	2	3630.528	4.859	4.94
Male	2	3642.342	16.672	11.98
Nonwhite	2	3653.501	27.831	27.77
Violent offense	2	3747.168	121.498	125.95
Number of charges	2	3737.877	112.207	99.08

Note:  $-2 \log\text{-likelihood}$  for the full model = 3625.670

Table 5.5 presents the values of the  $-2 \log\text{-likelihood}$  function for each of the reduced models and the value of the likelihood ratio test for each independent variable. Based on the critical value of the chi-square of 5.991, we see that race and sex of defendant, violent offense charge, and number of charges all have statistically significant effects on type of case disposition, while age of defendant does not have a statistically significant effect.

An alternative test of each independent variable is to use the Wald statistic. Up to this point, we have used the Wald statistic to test the statistical significance of a single coefficient, but it can also be used to test the group of coefficients representing the effect of any given independent variable. Recall that the Wald test statistic for a single coefficient is computed by dividing the coefficient by its standard error and then squaring this value. The Wald statistic for a group of coefficients involves an analogous calculation, but requires the use of matrix algebra—a topic beyond the scope of our text. That said, many statistical software packages will generate the results for the Wald test as part of the standard output, and our attention here is focused more on illustrating the interpretation of the results, rather than the actual calculation. In most applications, the value of the Wald statistic will be very similar to the value of the LR test.<sup>5</sup>

To test the overall effect of an independent variable with the Wald statistic, we continue to use a chi-square distribution with degrees of freedom equals the number of coefficients being tested—the number of outcome categories minus 1 (i.e.,  $df = j - 1$ ).

In our case disposition example, we have three outcome categories ( $j = 3$ ), so the degrees of freedom will be  $3 - 1 = 2$ , which again corresponds to the number of sets of coefficients that have been estimated. The values of the Wald test for each of the independent variables included in our analysis are presented in Table 5.5.

---

<sup>5</sup>Recall from Footnote 10 in Chap. 4 that the Wald statistic is sensitive to small samples (e.g., less than 100), while the LR test is not.

Using a significance level of 5%, the critical  $\chi^2$  statistic has a value of 5.991, as determined by the distribution function of any statistics program. Based on this value, we see that all the independent variables, except for age of defendant, have statistically significant effects on type of case disposition. The substance of these results is identical to that using the LR test.

How should we address mixed results? For example, it is not uncommon for a researcher to find that the overall effect of an independent variable is not statistically significant, but one of the individual coefficients does have a significant effect on a comparison of two outcome categories. Alternatively, the Wald test for the overall effect of an independent variable may show it to have a statistically significant effect, but there may be individual coefficients representing the effect of that independent variable on a specific comparison that are not statistically significant.

This kind of difficulty is illustrated in the results presented in Tables 5.4 and 5.5. Age of defendant does not have a statistically significant effect on case disposition overall, yet age does have a statistically significant effect on the comparison of trial conviction to dismissal. In such a case, the researcher should carefully note the pattern of results, explaining that overall age does not affect case disposition but that there appears to be a statistically significant impact of age on gaining a trial conviction as compared to a dismissal.

Alternatively, we also see that the overall effect of race is statistically significant, but the individual coefficient for race on the comparison between trial conviction and dismissal is not statistically significant. The safest approach for the researcher in this type of situation is to note the significance of the overall effect of the independent variable, but again to clearly explain the pattern of results for the individual coefficients. In this case, our model suggests that race has an overall impact on case disposition but our data do not allow us to conclude, despite this, that race has a significant effect on gaining a trial conviction as opposed to a dismissal.

Our suggestion is to use caution in interpreting the results and to be as clear as possible in explaining the nature and type of effect that is statistically significant. In multinomial regression, a number of statistical outcomes are included and the researcher should be careful not to draw selectively from the results gained.

### *Overall Model*

In addition to testing the statistical significance of the individual coefficients, we are also often interested in assessing the statistical significance of the overall model. To assess the statistical significance of the full multinomial regression model, we compute a model chi-square statistic that is identical in form to that used for the binary logistic regression model discussed in Chap. 4. Recall that the model chi-square is computed as:

$$\text{Model } \chi^2 = (-2LL_{\text{null model}}) - (-2LL_{\text{full model}})$$

For our case disposition analysis, the  $-2LL_{\text{null model}} = 3923.540$  and the  $-2LL_{\text{full model}} = 3625.670$ , resulting in a model chi-square of  $3923.540 - 3625.670 = 297.870$ . We have already noted that a total of 10 coefficients have been estimated (2 for each of the 5 independent variables), which gives us a degrees of freedom value for this test equal to 10. At a significance level of 5%, we see that a chi-square statistic greater than 18.307 is needed to reject the null hypothesis that the model has no statistically significant effect on case disposition. Since our model chi-square is larger than the critical value of the chi-square, we conclude that the overall model has a statistically significant effect on case disposition.

### A Concluding Observation About Multinomial Logistic Regression Models

In our substantive example, we selected a dependent variable with only three categories. Realistic applications of multinomial logistic regression models with more than three categories can quickly become unwieldy in regard to the number of contrasts that are being analyzed. For example, if we had a dependent variable with four categories, we would have three sets of coefficients to represent a total of six different contrasts ( $C_1$  and  $C_2$ ,  $C_1$  and  $C_3$ ,  $C_1$  and  $C_4$ ,  $C_2$  and  $C_3$ ,  $C_2$  and  $C_4$ , and  $C_3$  and  $C_4$ ). If we increased the number of outcome categories to five, we would have four sets of coefficients to represent a total of ten different contrasts, at which point the results from a multinomial logistic regression analysis likely become too difficult for most researchers to summarize in a coherent and concise way.

## Ordinal Logistic Regression

---

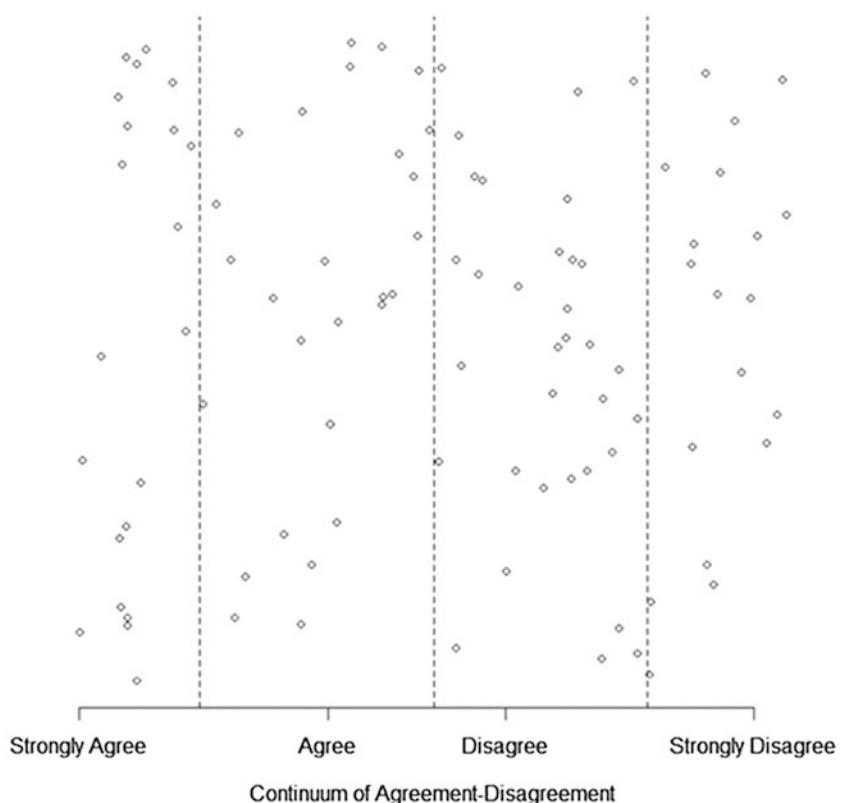
Multinomial regression provides a solution to the important problem of predicting multiple nominal category dependent variables. But in our discussions so far, we have not examined how to analyze ordinal-level dependent variables. For many years, researchers simply used OLS regression to deal with this type of analysis. There are times when this approach makes sense and is unlikely to lead to estimation problems that are important. For example, if a researcher is examining an ordinal-level variable that is measured as ten ascending categories and can assume that the interval between each of the levels of the scale is similar in meaning, then OLS estimates are likely to be satisfactory from a statistical perspective. An OLS regression model with a 5-point Likert-type scale as the dependent variable

and a sufficiently large sample size will generally lead to the same substantive conclusions as a more statistically justifiable model designed for an ordinal dependent variable. However, in such a case, the ordinal-level dependent variable is assumed to have characteristics close to that of an interval- or ratio-level measure.

Until recently, other estimation approaches for ordinal dependent variables were not easily accessible. As such, researchers often made assumptions regarding an ordinal dependent variable that were potentially not appropriate. For example, when we examine fear of crime measured as a series of categories from *very fearful* to *not fearful at all*, it is questionable that there are equal intervals between these qualitative responses. All modern general-purpose statistical software packages have the ability to estimate ordinal regression models, and as such, researchers should be cautious in applying OLS regression to ordinal-level measures. Furthermore, if they do so, they should ensure that the results are not misleading by comparing them to a model for ordinal dependent variables. In this section, we present the **ordinal logistic regression** model that explicitly takes into account an ordered categorical (discrete) dependent variable. Ordered probit regression is also appropriate for an ordered dependent. These two regression methods are highly similar and will almost always produce comparable results.

In order to set up the application and interpretation of the ordinal logistic model, we need to reconsider what a variable measured at the ordinal level tells us. Recall from introductory statistics that an ordinal variable has ranked categories that are assumed to represent an underlying continuum. For example, when respondents to a survey are presented with a statement that has as response choices Strongly Agree, Agree, Disagree, and Strongly Disagree, the variable is assumed to represent an underlying continuum of agreement–disagreement with some issue. Yet, we know that however an individual responds to the question, any two individuals falling in the same category may not mean exactly the same thing. For example, if we randomly selected two individuals who had responded Strongly Disagree with a policy statement and we were able to ask more in-depth follow-up questions, we would likely discover that there were degrees of how strongly each disagreed.

If we assume that an ordinal variable's categories represent an underlying continuum, we can think of **thresholds** as those points where an individual may move from one ordinal category to another (adjacent) category. In the example above, we could make note of the thresholds between Strongly Agree and Agree, Agree and Disagree, and Disagree and Strongly Disagree. Figure 5.1 illustrates the link between the underlying continuum and the variable measured at the ordinal level. In Fig. 5.1, each dot represents the true value for an individual's attitudes about a given issue—but this true value cannot be measured directly, and we are left with

**Figure 5.1***Hypothetical ordinal variable and underlying continuum*

the four response choices indicating degree of agreement or disagreement. Each of the vertical lines marks the point between one of the possible response choices and indicates the threshold for each response category.

The ordinal logistic regression model represents something of a hybrid of the binary logistic and multinomial logistic regression models. Similar to the multinomial logistic regression model's estimation of multiple model intercepts, the ordinal logistic model estimates multiple intercepts that represent the values of the thresholds. Comparable to the binary logistic model, the ordinal logistic model estimates one coefficient for the effect of each independent variable on the dependent variable. In part, this is due to the added information contained in an ordinal variable, rather than a multicategory nominal variable. The interpretation of the results from the ordinal logistic model is also potentially much simpler than the results from the multinomial logistic model.

One of the key differences between the ordinal logistic model and other logistic models is that rather than estimating the probability of a single category as in the binary and multinomial logistic models, the ordinal logistic model estimates a cumulative probability—the probability that the outcome is equal to or less than the category of interest, as shown here:

$$P(y \leq m) = \sum_{m=1}^{j-1} P(y = m) \quad \text{Equation 5.6}$$

where  $m$  is the category of interest and can take on values ranging from 1 to  $j - 1$ . The summation sign tells us that we are to add the probabilities for each individual outcome from the first category (i.e.,  $y = 1$ ) to the last category of interest (i.e.,  $y = j - 1$ ). For example, using the four response categories above would mean that  $j = 4$ , and we could compute a total of  $j - 1 = 4 - 1 = 3$  cumulative probabilities. If we define a Strongly Agree response as 1 and Strongly Disagree response as 4, we could then compute probabilities for  $P(y \leq 1)$ ,  $P(y \leq 2)$ , and  $P(y \leq 3)$ , representing  $P(y \leq \text{Strongly Agree})$ ,  $P(y \leq \text{Agree})$ , and  $P(y \leq \text{Disagree})$ , respectively. We would not include the final category (i.e.,  $P(y \leq 4)$  or  $P(y \leq \text{Strongly Disagree})$ ), since it would have to be equal to 1 (or 100%)—all possible values have to fall in one of the four response categories.

As with all logistic regression-type models, the analysis is based on logits or the logged odds of moving from one level to the next in the ordinal scale. In the context of ordinal logistic regression, the logits are defined as shown in Eq. (5.7).

$$\text{logit}[P(y \leq m)] = \ln \left[ \frac{P(y \leq m)}{P(y > m)} \right] \quad \text{Equation 5.7}$$

To illustrate this, suppose that the 4-point scale above had the following proportion of cases at each level: Strongly Agree = .15, Agree = .30, Disagree = .35, and Strongly Disagree = .20. The corresponding logits would be as follows:  $-1.735$ ,  $-0.201$ , and  $1.386$ , for the first three of these levels, respectively. This is a model without any independent variables, or what is called an intercept-only model, and these three values represent the cut-points between the four levels, hence there are three of them just as there are three vertical cut-points in Fig. 5.1.

Ordinal logistic regression models set these cut-points as a function of the independent variables. In the literature and across statistical software programs, there are two ways to write the regression function with

implications for how the coefficients are interpreted. These are shown in Eqs. (5.8) and (5.9).

$$\begin{aligned}\text{logit}(y \leq m|x) &= \ln \left[ \frac{P(y \leq m|x)}{P(y > m|x)} \right] \\ &= \tau_{0,m} + b_1x_1 + b_2x_2 + \cdots + b_kx_k\end{aligned}\quad \text{Equation 5.8}$$

$$\begin{aligned}\text{logit}(y \leq m|x) &= \ln \left[ \frac{P(y \leq m|x)}{P(y > m|x)} \right] \\ &= \tau_{0,m} - b_1x_1 - b_2x_2 - \cdots - b_kx_k\end{aligned}\quad \text{Equation 5.9}$$

The difference between Eqs. (5.8) and (5.9) is whether the coefficients for the independent variables are added to or subtracted from the intercept. The effect of this is to change the direction of the coefficients in the estimated models. It is critical to know which equation is being used by the software program you are using to ensure that you are interpreting the direction of the effects correctly. The descriptive statistics and predicted margin effects are useful tools in this regard (see Computer Exercise at the end of the chapter for information on how to do the latter).

The general form for this equation is very similar to that for the binary and multinomial logistic model except that we have introduced a new term ( $\tau_m$ ). This term represents the threshold parameters, which function as intercepts in the model and will take on values for  $m = 1$  to  $j - 1$ .

Equations (5.8) and (5.9) form the basis for estimating ordinal logistic regression models. That is, the model is regressing the logits (logged odds) as a linear function of the independent variables. Stated differently, it is determining the relationship between the independent variables and the logged odds of each level of our ordinal variable.

### Interpretation of Ordinal Logistic Regression Coefficients

In our discussion of the binary logistic regression model, we illustrated how a one-unit increase in the independent variable would modify the odds of success versus failure on an outcome by a factor of  $\exp(b)$  (equivalently  $e^b$ ). These exponentiated coefficients are odds ratios. Similarly, with ordinal logistic regression, we can compute odds ratios, but the interpretation is different. These reflect how a one-unit increase in the independent variable modifies the odds of moving up one level on the ordinal outcome variable. The computation of the odds ratio depends on whether the estimation of the model is based on Eq. (5.8) or (5.9), with the computation being  $\exp(b)$  and  $\exp(-b)$ , respectively.

To illustrate, recall that the regression coefficient  $b$  for a given independent variable  $x$  reflects the predicted change in the logit for a one-unit increase in  $x$  and that a logit is a logged odds. Also, recall that  $\ln(a) - \ln(b) = \ln(a/b)$ . Thus, the difference or change in logits equals the ratio of the odds. Suppose we have two values of one of our independent variables:  $x$  and  $x + 1$  and that our model was estimated using Eq. (5.9). The odds ratio is the odds for  $x + 1$  divided by the odds for  $x$  and can be expressed mathematically as:

$$\begin{aligned} \text{OR}_{(x+1)/x} &= \frac{\text{odds}_{m_{x+1}}}{\text{odds}_{m_x}} = \frac{\exp(\tau_m - (x+1)b)}{\exp(\tau_m - xb)} = \exp([x - (x+1)]b) \\ &= \exp(-b) \end{aligned}$$

Thus, to interpret the effect of a one-unit change in the independent variable in an ordinal logistic regression model, we will need to exponentiate the *negative* value of the estimated coefficient (which if it is negative, becomes positive!) if our model was estimated using Eq. (5.9), but we use the coefficient as is if using Eq. (5.8). We can then interpret the coefficient as indicating the odds of an outcome less than or equal to category  $m$  versus the odds of a category greater than  $m$ .

### **Substantive Example: Severity of Punishment Decisions**

Using the State Court Processing data for California in the 1990s resulted in a sample of 8197 individuals being convicted for some type of crime. The primary punishment outcomes—community-based, jail, and prison—represent a continuum of punishment severity with prison the most severe sentence. Again, if we keep our set of independent variables limited by using the same variables as in our case disposition example above, we have measures of age, sex (male = 1, female = 0), race (nonwhite = 1, white = 0), type of crime (violent = 1, nonviolent = 0), and total number of charges as predictors of severity of punishment. Table 5.6 presents the results of our ordinal logistic regression model. You will notice that ordinal regression, like multinomial regression uses the Wald statistic to assess the statistical significance of individual parameters.

### **Interpreting the Coefficients**

While ordinal regression accounts for the fact that the categories in the dependent variable are ranked, for example, in our case from less to more severe sanctions, the interpretation of the coefficients is similar to that used

**Table 5.6**

Ordinal logistic regression results for severity of punishment

INDEPENDENT VARIABLES	<i>b</i>	SE	-95%	+95%	WALD
Age	0.004	0.002	0.000	0.008	4.000
Male	0.821	0.058	0.707	0.935	202.991
Nonwhite	0.166	0.043	0.082	0.250	14.862
Violent offense	0.328	0.053	0.224	0.432	38.312
Number of charges	0.014	0.014	-0.013	0.041	1.000
$\tau_1$ (Intercept 1)	-0.881	0.099	-1.075	-0.687	79.175
$\tau_2$ (Intercept 2)	1.720	0.101	1.522	1.918	291.980

in multinomial regression. In this case, we can compare lower categories to the categories ranked above them. For example, in the case of sentences, we can compare either community-based punishment to jail and prison sentences or community-based punishment and jail sentences to prison sentences. In both these cases, the exponent of the negative of the coefficient provides the odds ratio for change. Since age, for example, is measured at the ratio level of measurement, we would note that for a 1-year increase in age, the odds ratio changes by a factor of  $\exp(-.004) = .996$ , controlling for the other variables in the model. We can write out the interpretations as follows:

- The odds of receiving a community-based punishment versus a jail and prison punishment decrease by a factor of .996 for a 1-year increase in age, controlling for all other variables in the model.
- The odds of receiving a community-based and a jail punishment versus a prison punishment decrease by a factor of .996 for a 1-year increase in age, controlling for all other variables in the model.

The coefficient for the male dummy variable is .821. By exponentiating the negative of .821 ( $\exp(-.821) = .440$ ), we see that males are likely to receive more severe punishments than females, controlling for the other independent variables in the model. More concretely, we can state the following about the punishment of male offenders:

- The odds of receiving a community-based punishment versus a jail and prison punishment are .440 times smaller for males than for females, controlling for all other variables in the model.
- The odds of receiving a community-based and jail punishment versus a prison punishment are .440 times smaller for males than for females, controlling for all other variables in the model.

The effect of race on punishment severity is  $\exp(-.166) = .847$ . Writing out direct interpretations of this coefficient leads to the following statements:

- The odds of receiving a community-based punishment versus a jail and prison punishment are .847 times smaller for nonwhites than for whites, controlling for all other variables in the model.
- The odds of receiving a community-based and jail punishment versus a prison punishment are .847 times smaller for nonwhites than for whites, controlling for all other variables in the model.

The effect of a violent offense charge is  $\exp(-.328) = .720$ , indicating that a violent offense is likely to result in more severe forms of punishment (as we would expect):

- The odds of receiving a community-based punishment versus a jail and prison punishment are .720 times smaller for individuals charged with a violent offense rather than a nonviolent offense, controlling for all other variables in the model.
- The odds of receiving a community-based and jail punishment versus a prison punishment are .720 times smaller for individuals charged with a violent offense rather than a nonviolent offense, controlling for all other variables in the model.

Finally, the effect of a one-unit increase in the number of charges is  $\exp(-.014) = .986$ . In practice, we would not spend much time interpreting this coefficient, since the Wald statistic indicates it is not significantly different from 0.<sup>6</sup> However, as another illustration for how to interpret coefficients from an ordinal logistic regression model, it is useful to write out the interpretations of this coefficient:

- The odds of receiving a community-based punishment versus a jail and prison punishment decrease by a factor of .986 for a one-unit increase in the number of charges, controlling for all other variables in the model.
- The odds of receiving a community-based and a jail punishment versus a prison punishment decrease by a factor of .986 for a one-unit increase in the number of charges, controlling for all other variables in the model.

Note, too, that there are two threshold parameters representing the threshold points between each of the ordered categories (i.e., community-based and jail punishments and then between jail and prison punishments).

### Statistical Significance

As we noted above, the test of statistical significance for each individual coefficient is a Wald statistic that is computed and is interpreted in exactly the same way as the Wald statistic for binary logistic and multinomial

---

<sup>6</sup>You should verify the statistical significance of each coefficient presented in Table 5.6 using a Wald test statistic with  $df = 1$ .

logistic regression models. Alternatively, we can use a  $z$  statistic which is developed by taking the coefficient ( $b$ ) divided by its standard error. The Wald statistic simply equals the square of this value ( $z^2$ ) for a single coefficient. Table 5.6 reports the values of the Wald statistic for each independent variable. The statistical significance of the overall model is based on a model  $\chi^2$  statistic that is also computed and interpreted in exactly the same way as for the binary logistic and multinomial logistic regression models. In our punishment example, the  $-2LL_{\text{null}} = 5,883.113$  and the  $-2LL_{\text{full}} = 5,601.386$ , resulting in a model  $\chi^2$  of  $5,883.113 - 5,601.386 = 281.727$ . Since a total of five coefficients have been estimated (one for each independent variable), the degrees of freedom value for this test is equal to 5. The 5% critical value for a  $\chi^2$  with 5 degrees of freedom is 11.070. Thus, the model  $\chi^2$  needs to equal or exceed this value for us to reject the null hypothesis that the model had no effect on punishment severity. Since our model  $\chi^2$  is larger than the critical value, we conclude that the overall model had a statistically significant effect on punishment severity.

Table 5.6 also shows the lower and upper bounds of the 95% confidence intervals (−95% and +95%, respectively). For number of charges, the confidence interval includes zero, indicating that we cannot reject zero as a plausible value. Age, while statistically significant, has a confidence interval that touches zero, although it is positive out to several decimal places. This suggests that it is plausible that the effect of age is quite small.

### Parallel Slopes Tests

As we noted earlier, the proportional odds model assumes that the effects of the independent variables are constant across all categories of the dependent variable, which is analogous to our interpretation of coefficients in a multiple linear regression model. Regardless of the level (or category) of the dependent variable, we expect the independent variable to exert a constant (i.e., proportional) effect on the dependent variable. The constant effect of each independent variable should have also been clear in the direct interpretations of the coefficients noted in the previous section. This is known more generally as the **parallel slopes assumption**. Most statistical packages include a score test of this assumption that informs the user of the appropriateness of the ordinal logistic model. Somewhat less common is the Brant test, which tests for parallel slopes for the overall model and for each independent variable.

### Score Test

Conceptually, the parallel slopes score test is based on the idea that we could estimate a series of  $j - 1$  binary logistic regression models (i.e., one model less than the number of ordered categories in the dependent variable) of the form  $P(Y \leq m)$  that allowed the effects for all  $k$  independent

variables to vary by outcome category on the dependent variable. In other words, we could perform a multinomial regression model as discussed earlier in this chapter. The test would then focus on whether a single coefficient or multiple coefficients best represented the effect of the independent variables on the dependent variable. Technically, the score test uses information about the log-likelihood for the ordinal logistic regression model and assesses how much it would change by allowing the coefficients for all the independent variables to vary by the outcome category on the dependent variable. The degree of change in the likelihood function then indicates whether the parallel slopes assumption is met. The null hypothesis of the score test is parallel (equal) slopes. The research hypothesis is that the slopes are not parallel. The value of the score test (reported by most statistical software) is distributed as a chi-square with  $k(j - 2)$  degrees of freedom.

For our severity of punishment example, we have  $k = 5$  (i.e., five independent variables) and  $j = 3$  (i.e., three outcome categories on the dependent variable). The corresponding degrees of freedom for our score test is equal to  $5(3 - 2) = 5$ . Based on the critical values for the  $\chi^2$ , the critical  $\chi^2$  for a significance level of 5% is 11.070. The value of the score test for our model is 57.890, which indicates that we should reject our null hypothesis of parallel slopes and conclude that our model does not meet the parallel slopes assumption.

### Brant Test

Similar to the score test, the Brant test (Brant 1990) is a Wald test that assesses whether all the coefficients in a proportional odds model satisfy the parallel slopes assumption. The computation of the Brant test is based on the values of the coefficients and their respective variances. In addition to providing an overall test for the parallel slopes' assumption, the Brant test can be decomposed into values for each of the independent variables in the ordinal logistic regression model to test whether each independent variable meets the parallel slopes assumption.

The Brant test for the overall model is distributed as a chi-square with  $k(j - 2)$  degrees of freedom (same as in the score test). Each independent variable's test statistic is distributed as a  $\chi^2$  with  $j - 2$  degrees of freedom.

The results of the Brant test for our severity of punishment example appear in Table 5.7. As expected, the overall test again indicates that the parallel slopes assumption is violated for our model. The  $\chi^2$  is computed as 65.34, and with  $df = 5$  and a critical chi-square of 11.070, we reject the null hypothesis of parallel slopes for the full model. For each independent variable, the critical  $\chi^2$  based on  $3 - 2 = 1$  degree of freedom at the 5% level of significance is 3.841. We see that age ( $\chi^2 = 40.25$ ), male ( $\chi^2 = 4.72$ ), and violent offense charge ( $\chi^2 = 15.82$ ) would lead us to reject the null hypothesis of parallel slopes for each of these variables, since all have

**Table 5.7**

Brant test results for severity of punishment

INDEPENDENT VARIABLES	$\chi^2$	df	p
Overall	65.34	5	0.000
Age	40.25	1	0.000
Male	4.72	1	0.030
Nonwhite	2.28	1	0.131
Violent	15.82	1	0.000
Total number of charges	2.14	1	0.143

chi-square values greater than 3.841. This means that the effects of these three independent variables are not proportional (constant) across the levels of severity of punishment. In contrast, the effects of nonwhite ( $\chi^2 = 2.28$ ) and total number of charges ( $\chi^2 = 2.14$ ) have chi-square values less than 3.841, leading us to fail to reject the parallel slopes assumption and conclude that the effects of these two independent variables are proportional across the levels of severity of punishment.

### Partial Proportional Odds

In much of the research in criminology and criminal justice, it is quite common for the parallel slopes assumption not to be met in practice. Historically, when researchers have been confronted with results from the score test indicating that the model failed to satisfy the parallel slopes assumption, they were left with a choice of fitting the proportional odds model and violating a key assumption of the model or of fitting a multinomial logistic regression model and ignoring the ordinal nature of the dependent variable, complicating the interpretation of the results through the increased number of coefficients. A class of models referred to as partial proportional odds or generalized ordinal logistic regression models has received increasing attention as a method of dealing with these issues.<sup>7</sup> The logic to the partial proportional odds model is to allow some or all of the coefficients of the independent variables to vary by the level of the dependent variable, much like we see in the application of multinomial logistic regression, but to constrain other coefficients to have a single value, as in the proportional odds model.

We obtain the partial proportional odds model by generalizing the proportional odds to allow the coefficients (the  $b_1$  to  $b_k$  where  $k$  is the number of independent variables) to vary by level of the dependent variable ( $m$ ):

---

<sup>7</sup>There are a number of sources interested readers can consult, although most of these are much more technical than the material presented in this text. See, for example, Fu (1998), Lall et al. (2002), O'Connell (2006), Peterson and Harrell (1990), and Williams (2006).

$$\begin{aligned}\text{logit}(y \leq m|x) &= \ln \left[ \frac{P(y \leq m|x)}{P(y > m|x)} \right] \\ &= \tau_{0,m} + b_{1,m}x_1 + b_{2,m}x_2 + \dots + b_{k,m}x_k\end{aligned}$$

Notice that there is now a regression coefficient for each independent variable and each cut-point. Without any further constraints on the coefficients, the total number of coefficients estimated will be identical to that obtained from a multinomial logistic regression analysis. It is important to note, however, that the coefficients do not mean the same thing. Recall from our discussion above that multinomial logistic regression coefficients refer to comparisons between a given category and the reference category. As noted in the equation above, the logit in the partial proportional odds model is identical to that in the proportional odds model and refers to the odds of a category less than or equal to  $m$  versus a category greater than  $m$ .

Due to the potentially large number of coefficients in a fully generalized ordinal logit model, most researchers will want to limit the number of variables with nonconstant effects. The results from the Brant test are useful for determining which independent variables, if any, appear to have varying effects on the different categories of the dependent variable (i.e., the slopes are not parallel). If the overall Brant test result is not statistically significant, it implies that the parallel slopes assumption is met for the full model. In this case, there is likely little to be gained by relaxing the parallel slopes assumption for a single variable—the results become unnecessarily complicated and will not add much statistically to the model.

In those cases where the overall Brant test result is statistically significant, then the Brant test results for individual variables will point to those variables with the greatest divergence from the parallel slopes assumption and the best candidates for allowing the effects to vary across the different ordinal logits.

All other features of the partial proportional odds model—tests for statistical significance, interpretation of the coefficients, and the like—are the same as found in the proportional odds model.

### **Severity of Punishment Example**

In our application of the proportional odds model to the severity of punishment data from California, we also noted that the parallel slopes assumption was not satisfied for the overall model. In particular, the effects of age, male, and violent offense charge violated the parallel slopes assumption, while those of nonwhite and total number of charges did not (see Table 5.7).

To illustrate the application and interpretation of the partial proportional odds model, we begin by allowing all five independent variables to have different effects on the two ordinal logits:

$$\begin{aligned}\text{logit}_1 &= \ln \left( \frac{P(y \leq 1)}{P(y > 1)} \right) \\ &= \ln \left( \frac{P(y = \text{Probation})}{P(y = \text{Jail or Prison})} \right) \\ &= \tau_1 - b_{1,1}x_1 + b_{2,1}x_2 + b_{3,1}x_3 + b_{4,1}x_4 + b_{5,1}x_5\end{aligned}$$

and

$$\begin{aligned}\text{logit}_2 &= \ln \left( \frac{P(y \leq 2)}{P(y > 2)} \right) \\ &= \ln \left( \frac{P(y = \text{Probation or Jail})}{P(y = \text{Prison})} \right) \\ &= \tau_2 - b_{1,2}x_1 + b_{2,2}x_2 + b_{3,2}x_3 + b_{4,2}x_4 + b_{5,2}x_5\end{aligned}$$

where  $b_{1,1}$ ,  $b_{2,1}$ , etc., are the coefficients for the probation vs. jail and prison categories, and  $b_{1,2}$ ,  $b_{2,2}$ , etc., are the coefficients for the probation and jail vs. prison category.

Since there are two different ordinal logits being estimated, there are two full sets of unique coefficients to interpret that illustrate the different effects the independent variables have on the two different ordered logits. These coefficients are presented in Table 5.8.

Some highlights found in the results presented in Table 5.8:

**Table 5.8**

Partial proportional odds model for severity of punishment—all coefficients allowed to vary

INDEPENDENT VARIABLE	<i>b</i>		<i>SE</i>	
	PROBATION VS. JAIL/PRISON	PROBATION/JAIL VS. PRISON	PROBATION VS. JAIL/PRISON	PROBATION/JAIL VS. PRISON
Age	-0.010	0.011	0.003	0.003
Male	0.728	0.916	0.073	0.074
Nonwhite	0.103	0.190	0.064	0.048
Violent offense	0.101	0.410	0.081	0.057
Total number of charges	-0.006	0.021	0.020	0.015
Constant ( $\tau$ )	-1.496	2.073	0.133	0.133

- Age:

- The odds of probation versus jail or prison increase by a factor of  $\exp(-(-0.010)) = 1.010$ , controlling for other variables in the model. For a 10-year increase in age, the odds of probation versus a jail or a prison sentence increase by a factor of  $\exp(-10 \times -0.010) = 1.105$ , controlling for all other variables in the model.
- The odds of probation or jail versus prison decrease by a factor of  $\exp(-(0.011)) = 0.989$  for a one-unit increase in age, controlling for all other variables in the model. For a 10-year increase in age, the odds of probation or jail versus a prison sentence decrease by a factor of  $\exp(-(10 \times 0.011)) = 0.896$ , controlling for all other variables in the model.

- Violent offense charge:

- The odds of probation versus jail or prison are  $\exp(-(0.101)) = 0.904$  times smaller for offenders charged with a violent offense than for offenders charged with a nonviolent offense, controlling for all other variables in the model.
- The odds of probation or jail versus prison are  $\exp(-(0.410)) = 0.664$  times smaller for offenders charged with a violent offense than for offenders charged with a miscellaneous offense, controlling for all other variables in the model.

Substantively, these results indicate that offenders charged with a violent offense are less likely to be treated leniently in the form of receiving either a probation or a jail sentence and are more likely to receive a prison sentence. The results for age suggest that while older offenders are more likely to receive a probation sentence rather than a jail or a prison sentence, they are less likely to receive a probation or a jail sentence rather than a prison sentence. These results may seem contradictory, but one way of interpreting the pattern is that older offenders are more likely to be sentenced to probation or prison, depending on crime, criminal history, and so on, but less likely to receive jail sentences. This kind of variability in the effect of age on sentencing has also been found in prior research (Steffensmeier et al. 1998).

Since the results of the Brant test indicated that there were only three of the five independent variables that did not satisfy the parallel slopes assumption, we have rerun the partial proportional odds model allowing only the effects for age, male, and violent offense charge to vary. Table 5.9 presents these results. Since the results presented here are nearly identical to those presented in Table 5.8 for the coefficients allowed to vary, we limit

**Table 5.9**

Partial proportional odds model for severity of punishment—some coefficients constrained

VARIABLE	CONSTRAINED	<i>b</i>		<i>SE</i>		PROBATION/ JAIL VS. PRISON
		PROBATION VS. JAIL/PRISON		PROBATION/ JAIL VS. PRISON	CONSTRAINED	
		PROBATION VS. JAIL/PRISON	PROBATION/ JAIL VS. PRISON	PROBATION VS. JAIL/PRISON	PROBATION/ JAIL VS. PRISON	
Age		-0.010	0.011	0.003	0.003	
Male		0.733	0.915	0.073	0.074	
Nonwhite	0.164			0.043		
Violent offense		0.094	0.413	0.081	0.057	
Num. of charges	0.013			0.014		
Constant ( $\tau_0$ )		-1.428	2.042	0.128	0.115	

our discussion to the effects of male (variable) and nonwhite (constrained or equal):

- Male:
  - The odds of probation versus jail or prison are  $\exp(-(0.733)) = 0.480$  times smaller for male offenders than for female offenders, controlling for all other variables in the model.
  - The odds of probation or jail versus prison are  $\exp(-(0.915)) = 0.401$  times smaller for male offenders than for female offenders, controlling for all other variables in the model.
- Nonwhite:
  - The odds of probation versus jail or prison are  $\exp(-(0.164)) = 0.849$  times smaller for nonwhite offenders than for white offenders, controlling for all other variables in the model.
  - The odds of probation or jail versus prison are  $\exp(-(0.164)) = 0.849$  times smaller for nonwhite offenders than for white offenders, controlling for all other variables in the model.

Substantively, these results indicate that male and nonwhite offenders are less likely to receive more lenient punishments (either probation or jail) and more likely to receive a prison sentence. As noted above, this kind of pattern is consistent with much of the prior research on punishment severity.

## Chapter Summary

In this chapter, we have examined two different multiple regression-based statistical models to be used when we are confronted with a categorical dependent variable that has more than two categories. When the

dependent variable has three or more categories, we can use the **multinomial logistic regression model**. The multinomial logistic regression model allows for the computation of odds ratios that indicate the effects of the independent variables on the relative likelihood of the different outcome categories.

Since the multinomial logistic regression model estimates a set of coefficients for each independent variable, we have two issues of statistical significance to assess: the individual coefficients and the overall effect of the independent variable on the dependent variable. For the individual coefficients we use the Wald or  $z$ -statistic. For the overall effect of the independent variable on the dependent variable, where we are testing multiple coefficients, we can use the likelihood ratio (LR) test or the Wald statistic. Both test statistics are distributed as a  $\chi^2$  with  $j - 1$  degrees of freedom.

When the dependent variable is measured at the ordinal level of measurement, we can use the **ordinal logistic regression model** (or **proportional odds model**). The ordinal logistic regression model also allows for the computation of odds ratios, but the focus is on the likelihood of increasing or decreasing categories on the ordered dependent variable. The ordinal logistic model assumes that the effects of the independent variables are constant across the categories of the dependent variable (**parallel slopes assumption**), which can be tested with the parallel slopes test that is commonly reported in the output of most statistical programs. The parallel slopes test statistic is distributed as a chi-square with  $k(j - 2)$  degree of freedom. The null hypothesis in such a test is that the slopes are parallel, while the research hypothesis is that the slopes are not parallel.

When there is evidence that the parallel slopes assumption is not satisfied, we can use the **partial proportional odds model** that allows one or more of the effects of the independent variable to vary across the levels of the ordinal dependent variable. The interpretation of the results and the tests for statistical significance work the same way in the partial proportional odds model as they do in the ordinal logistic regression model.

## Key Terms

---

**Multinomial logistic regression** A statistical technique to predict the value of a dependent variable with three or more categories measured at the nominal level of measurement.

**Ordinal logistic regression (proportional odds model)** A statistical technique to predict the value of a dependent variable with three or more categories measured at the ordinal level of measurement.

**Parallel slopes assumption** In an ordinal logistic regression model, the effect of each independent variable is assumed to be constant across all categories of the dependent variable.

**Partial proportional odds model** An ordinal logistic regression model that allows

the effects of one or more of the independent variables to vary across the levels of the ordinal dependent variable. Useful when the parallel slopes assumption is violated.

**Thresholds** Points that mark the limits of the underlying continuum measured by an ordinal variable.

## Formulas

---

To calculate the probability that  $Y = m$ :

$$\text{Probability Equation: } P(Y = m) = \frac{\exp(Xb_m)}{\sum_{j=1}^J \exp(Xb_j)}$$

To calculate the odds ratio in a multinomial logistic regression model for  $P(Y = m)$  relative to  $P(Y = n)$ , given a one-unit change in an independent variable:

$$\text{Conditional } OR_{m|n} = \frac{P(Y = m)}{P(Y = n)} = \frac{\frac{\exp(Xb_m)}{\sum_{j=1}^J \exp(Xb_j)}}{\frac{\exp(Xb_n)}{\sum_{j=1}^J \exp(Xb_j)}} = \frac{\exp(Xb_m)}{\exp(Xb_n)}$$

To calculate the odds ratio in an ordinal logistic regression model using cumulative probabilities:

$$OR_m = \frac{P(Y \leq m)}{1 - P(Y \leq m)} = \frac{P(Y \leq m)}{P(Y > m)}$$

Ordinal logit equation:

$$\ln \left( \frac{P(Y \leq m)}{P(Y > m)} \right) = \ln [\exp(\tau_m - Xb)] = \tau_m - Xb$$

## Exercises

---

- 5.1. A large survey of adults asked about violent victimization experiences. A question of particular interest to one researcher was the location of the victimization event—home, work, or elsewhere. She computed a multinomial logistic regression model that produced the following results:

INDEPENDENT VARIABLE	HOME VS. WORK	ELSEWHERE VS. WORK
Age (years)	0.01	0.05
Sex (1 = male, 0 = female)	-0.19	0.22
Married (1 = yes, 0 = no)	0.37	-0.13
Number of nights out per week for leisure	0.07	0.16

- (a) Calculate the odds ratio for each coefficient, and explain what each odds ratio means.
- (b) Calculate the coefficients and the odds ratios for the omitted comparison, and explain what each odds ratio means.
- 5.2. In an attempt to better understand how non-incarcerative punishments were being used by judges, Blue State funded an evaluation study of misdemeanor punishment decisions. The evaluators classified non-incarcerative sentences in the following four categories: fine, restitution, community service, and electronic monitoring. The researchers' final analysis produced the following results:

INDEPENDENT VARIABLE	FINE V. ELECTRONIC MONITORING	RESTITUTION V. ELECTRONIC MONITORING	COMMUNITY SERVICE V. ELECTRONIC MONITORING
Any prior criminal record (1 = yes, 0 = no)	-0.06	-0.07	-0.10
Severity of offense	-0.10	-0.12	-0.14
Employed (1 = yes, 0 = no)	0.25	0.23	0.36

- (a) Calculate the odds ratios for the effect of any prior record, and explain what each odds ratio means.
- (b) Calculate the odds ratios for the effect of severity of offense, and explain what each odds ratio means.
- (c) Calculate the odds ratios for the effect of employed, and explain what each odds ratio means.
- (d) Calculate the coefficients and the odds ratios for the comparison between *fine* and *community service*. Explain what each odds ratio means.

- 5.3. Criminological theory has attempted to explain both the frequency of delinquency and the type of delinquency an individual is likely to commit. A longitudinal study of adolescents tested for the effects of several background characteristics on the likelihood an individual would commit a drug, property, violent, or public order offense. The researchers used a multinomial logistic regression model and found the value of the  $-2 \log\text{-likelihood}$  for the full model to be 5263.1. The values for the  $-2 \log\text{-likelihood}$  for each of the independent variables were reported as:

INDEPENDENT VARIABLE	$-2 \log\text{-LIKELIHOOD}$
Age	5264.7
Sex	5322.5
Race	5271.1
Grade point average	5267.9
Employment status	5414.6
Parental supervision	5272.3
Number of friends who had been arrested	5459.4

Calculate the LR Test statistic for each independent variable, and state whether this variable has a statistically significant effect on type of crime (assume  $\alpha = 0.05$ ).

- 5.4. In response to public perceptions that the police in Riverside City were too prone to use physical force on suspects, a study was commissioned to examine the factors related to when police did use physical force. After a year of data collection, the researchers classified police use of force into the following three categories: *None*, *Mild restraint*, and *Complete restraint*. The ordinal logistic regression model of only demographic characteristics produced the following results:

INDEPENDENT VARIABLE	<i>b</i>
Age of officer (years)	-0.02
Sex of officer (1 = male, 0 = female)	0.18
Race of officer (1 = white, 0 = nonwhite)	0.13
Age of suspect (years)	-0.03
Sex of suspect (1 = male, 0 = female)	0.33
Race of suspect (1 = white, 0 = nonwhite)	-0.11

Calculate the odds ratio for each coefficient, and explain what each odds ratio means.

- 5.5. A survey of adults in the USA asked a series of questions about support for various policies related to the treatment of criminal offenders. One question focused on the level of support for the use of the death penalty—whether the respondent was opposed to its use, neutral, or favored its use. An ordinal logistic regression model that included age (years), sex (1 = male, 0 = female), race (1 = African American, 0 = white), education (number of years completed), and degree of political liberalism (1 = low, 10 = high) produced the following results:

INDEPENDENT VARIABLE	<i>b</i>
Age	0.03
Sex	0.41
Race	-0.65
Education	-0.18
Liberalism	-0.21

Calculate the odds ratio for each coefficient, and explain what each odds ratio means.

- 5.6. In a study of community perceptions of the local police department, individuals were asked a series of questions about their perceptions of police behavior when interacting with local residents. Of particular interest to the researchers was a question about trust that residents had in the police: *Would you say that your level of trust in the police is . . .* The responses were limited to Very Low, Low, Moderate, High, and Very High. The researchers estimated an ordinal logistic regression model and Brant test, and found the following:

INDEPENDENT VARIABLE	<i>b</i>	BRANT TEST
Age (in years)	0.02	4.372
Sex (1 = male, 0 = female)	-0.38	8.914
Race (1 = nonwhite, 0 = white)	-0.42	12.695
Ever arrested? (1 = yes, 0 = no)	-0.67	2.720
Ever reported a crime to the police? (1 = yes, 0 = no)	-0.26	5.661
Total		34.362

- (a) Calculate the odds ratios for each coefficient, and explain what it means.
- (b) Test the parallel slopes assumption for the full model and each coefficient. What can you conclude about this model? What would your recommendation be to the researchers about the use of the ordinal logistic regression model? Explain why.
- 5.7. A longitudinal study of delinquent and criminal behavior classified a cohort of males (all the same race–ethnicity) into one of the three categories based on patterns of illegal behavior throughout adolescence: Nondelinquent, Low-rate delinquent, and High-rate delinquent. On the basis of Brant test results, the researchers estimated a partial proportional odds model using a small subset of background characteristics to predict delinquent group:

INDEPENDENT VARIABLE	NONDELINQUENT VS. LOW AND HIGH RATE <i>b</i>	NONDELINQUENT AND/OR LOW RATE VS. HIGH RATE <i>b</i>
Academic performance (1 = low to 10 = high)	-0.12	-0.08
Risk scale (1 = low to 20 = high)	0.23	0.33

(continued)

INDEPENDENT VARIABLE	NONDELINQUENT VS. LOW AND HIGH RATE <i>b</i>	NONDELINQUENT AND/OR LOW RATE VS. HIGH RATE <i>b</i>
Parent arrested? (1 = yes, 0 = no)	0.47	0.85
Number of friends arrested	0.17	0.29
Level of parental supervision (1 = low, 10 = high)	-0.09	-0.11

Interpret and explain these results.

## Computer Exercises

The data file used to illustrate the application of the multinomial and ordinal models in this chapter can be found in either SPSS (ca\_scps9098.sav) or Stata (ca\_scps9098.dta) format. The illustration of the commands below assumes that you have opened one of these files into SPSS, Stata, or R, and can also be found in the sample syntax files in both SPSS (Chapter\_5.sps) and Stata (Chapter\_5.do) format.

### SPSS

#### *Multinomial Logistic Regression*

To estimate a multinomial logistic regression model, you will need to use the NOMREG command:

```
NOMREG depvar (BASE = #) WITH indepvars  
/PRINT = PARAMETER SUMMARY LRT CPS MFI.
```

As in previous illustrations of SPSS commands, everything can be issued in upper or lowercase, but we have used uppercase lettering to highlight the key components of the command. The /PRINT = option forces SPSS to print all of the model and individual coefficient results. Also, take note that the default reference category in SPSS is the category with the highest number (i.e., category value). To force a particular category as the reference, use the base option in parentheses.

To reproduce our results in Table 5.1, enter the following command:

```
NOMREG casedisp (BASE = 1)  
WITH age male nonwhite violent total_charges  
/PRINT = PARAMETER SUMMARY LRT CPS MFI.
```

Similarly, to reproduce the results in Table 5.3, where *Trial* was used as the reference category, use the following command:

```
NOMREG casedisp (BASE = 3)  
WITH age male nonwhite violent total_charges  
/PRINT = PARAMETER SUMMARY LRT CPS MFI.
```

Note that the only difference between these two commands is changing the base from 1 to 3.

### *Ordinal Logistic Regression*

To estimate an ordinal logistic regression model in SPSS, use the PLUM command:

```
PLUM depvar WITH indepvars
/LINK = LOGIT
/PRINT = FIT PARAMETER SUMMARY TPARALLEL.
```

Since there are other types of models for ordinal regression, the /LINK = option forces SPSS to estimate an ordinal logistic regression model. The /PRINT = option forces SPSS to generate a full set of output that is consistent with the items we have discussed in this chapter.

To reproduce the results in Table 5.6, enter the following command:

```
PLUM ord_punishment
WITH age total_charges nonwhite male violent
/LINK = LOGIT/PRINT = FIT PARAMETER SUMMARY TPARALLEL.
```

At the time of this writing, SPSS does not have the option of computing the Brant test or estimating partial proportional odds models.

### **Stata**

#### *Multinomial Logistic Regression*

To estimate a multinomial logistic regression model in Stata, we use the **mlogit** command:

```
mlogit depvar indepvars, baseoutcome(#)
```

where *baseoutcome* refers to the category that should be used as the reference category. The default in Stata is to use the category with the greatest number of cases.

To reproduce our results in Table 5.1, enter the following command:

```
mlogit casedisp age male nonwhite violent total_charges,
baseoutcome(1)
```

Note that the **baseoutcome(#)** option was used to force Stata into using *dismissal* as the reference category. If this option had been omitted, *plea* would have been used as the reference category.

Similarly, to reproduce the results in Table 5.3, the command would be

```
mlogit casedisp age male nonwhite violent total_charges,
baseoutcome(3)
```

Note, too, that the output provided in Stata in regard to statistical significance is a *z*-score rather than a Wald statistic. This is not problematic, since there is a simple relationship between the Wald and *z*-score: Wald =  $z^2$ .

Consequently, if you square the values of the *z*-scores in the Stata output, it will reproduce the Wald statistics reported in the text (with some rounding error).

### *Ordinal Logistic Regression*

To estimate an ordinal logistic regression model in Stata, use the **ologit** command:

```
ologit depvar indepvars
```

To reproduce the results in Table 5.6, enter the following command:

```
ologit ord_punishment age male nonwhite violent total_charges
```

The Brant test requires downloading and installing a set of commands written by Long and Freese (2006). To install this package of user-written commands, use the **findit** command (one time only):

```
findit spost13_ado
```

After this command has been run, a pop-up window will open and allow you to install the package. Then, the Brant test should be available and only requires entering the command **brant** in the line after running the **ologit** command. If you run into problems with *spost13\_ado*, use the command **ssc install oparallel** as the *oparallel* test has a brant option. The following commands will reproduce the results for Table 5.6 and then Table 5.7:

```
ologit ord_punishment age male nonwhite violent total_charges  
brant
```

### *Partial Proportional Odds*

Before running a partial proportional odds model, we again need to download and install a user-written procedure called **gologit2**. The process here is the same as before. Enter the following command (one time only):

```
ssc install gologit2
```

This command will install gologit2 from an archive of procedures housed and maintained by Stata. Once this command has been run, the basic structure of the **gologit2** command is

```
gologit2 depvar indepvars
```

This will estimate a fully unconstrained model, where all of the independent variables are allowed to have variable effects across the levels of the ordinal dependent variable.

To estimate a partial proportional odds model that constrains some independent variables to have the same effect and allows others to have variable effects, use the **pl**(constrained\_indepvars) option:

```
gologit2 depvar indepvars, pl(constrained_indepvars)
```

The variable names listed inside the parentheses with the **pl** option will be constrained. Any other independent variable listed in the **gologit2** command line will be allowed to have variable effects.

To reproduce the results in Table 5.8, enter the following command:

```
gologit2 ord_punishment age male nonwhite  
violent total_charges
```

To reproduce the results in Table 5.9, which constrains the effects of nonwhite and number of charges, enter the following command:

```
gologit2 ord_punishment age male nonwhite  
violent total_charges, pI(nonwhite total_charges)
```

## R

### Multinomial Logistic Regression

We estimate a multinomial logistic regression model in R using the **multinom()** function. This function is within the *mlogit* package, so you must install and load the package, respectively, with `install.packages("nnet")` and `library(nnet)`. The basic structure of *mlogit* is as follows:

```
multinom(depvar ~ indepvar1 + indepvar2,  
data = dataset_name)
```

To reproduce our results in Table 5.1, we must first change the class of the variable `total_charges` using the **as.numeric()** function because in the prior examples, it is treated as a ratio variable. If we do not, R will treat the variable as if it were a nominal. We are going to store the results of our model in an object named `test`. Also, remember to specify the reference category as *Dismissal*. Then, use the **summary()** function to obtain information a summary of the regression model.

```
df$total_charges <- as.numeric(df$total_charges)  
test <- multinom(casedisp ~ age + male + nonwhite +  
violent + total_charges,  
data = df)  
summary(test)
```

To reproduce our results in Table 5.3, we must change the reference category. This is done using the **relevel()** function. Note that we are referencing the value on the `casedisp` variable, but we could also just specify `ref = "Trial"`.

```
df$casedisp2 <- relevel(df$casedisp, ref = 3)  
test2 <- multinom(casedisp2 ~ age + male + nonwhite +  
violent + total_charges,  
data = df)  
summary(test2)
```

To get the *p*- and *z*-values for the model output, we are going to use the **coeftest()** function within the *AER* package. Once the package is installed, this can be done with the following command:

```
coeftest(test2)
```

Simply just square  $z$ -values to obtain the Wald statistics.

### *Ordinal Logistic Regression*

To estimate an ordinal logistic regression model in R, use the **polr()** function in the *MASS* package:

```
polr(depvar ~ indepvar1 + indepvar2,  
      data = dataset_name, Hess = FALSE)
```

If you want just the coefficients provided, you specify `Hess = FALSE`. We are going to set it to `TRUE` and then are going to use the `summary()` function to return the results. Additionally, we need to tell R that we want our dependent variable to be treated as a factor, which we are going to be doing with the **factor()** function, and we want treat the *total\_charges* variable as ratio using **as.numeric()**.

```
df$ord_punishment2 <- factor(df$ord_punishment)  
df$total<-as.numeric(df$total_charges)
```

Then, to reproduce the results in Table 5.6, enter the following command:

```
model <- polr(ord_punishment2 ~ age + male + nonwhite +  
           violent + total, data=df, Hess = TRUE)  
summary(model)
```

The Brant test requires installing loading the *brant* package. After this package has been installed, the Brant test can be conducted using the **brant()** function. If you already have it installed in Stata but received an error, uninstall the *brant* package. The following commands will reproduce the results for Tables 5.7:

```
brant(model)
```

### *Partial Proportional Odds*

Before running a partial proportional odds model in R by using the **clm()** function, you will want to install the *ordinal* package. The basic structure of conducting a partial proportional odds model using the **clm()** function is as follows:

```
clm(depvar ~ indepvar1 + indepvar2, data=dataset_name)
```

If we want to allow our regression parameters to vary with the responses on the dependent variable, we are going to introduce nominal effects using the **nominal=** argument, which allow variables specified using this argument to vary and those independent variables specified to the left of the argument not allowing to vary. This is done as follows:

```
clm(depvar ~ 1, nominal= ~ indepvar1 + indepvar2,  
      data=dataset_name)
```

To reproduce the results in Table 5.8, you may need to specify that the *total\_charges* variable be specified as numeric if it is not already and then you can run your model using the **clm()** function:

```
df$total<-as.numeric(df$total_charges)
```

```
clm(ord_punishment ~ 1, nominal = ~ nonwhite +
    total + age + male + violent, data = df)
```

Recall that the coefficients will have a slightly different interpretation to Table 5.8 (see discussion of Eqs. (5.8) and (5.9)).

To reproduce the results in Table 5.9, which constrains the effects of nonwhite and number of charges, enter the following command:

```
clm(ord_punishment ~ nonwhite + total, nominal = ~
    age + male + violent, data = df)
```

### Problems

- As a first step in working with these two commands, open either the SPSS (Chapter\_5.sps) or the Stata (Chapter\_5.do) files that will run all of the commands described above. If you prefer to work directly with the California data file, then open either the SPSS (ca\_scps9098.sav) or the Stata version (ca\_scps9098.dta). The syntax and data files contain the felony arrest cases used in the examples in this chapter and will allow you to reproduce the results. Follow the commands as described above or run the appropriate syntax file.

Questions 2 through 4 use the NYS data (nys\_1.sav or nys\_4\_student.sav for SPSS; nys\_1.dta for Stata; or use any dataset in R using `read_sav()` or `read_dta()` from the *haven* package).

- Compute a multinomial logistic regression model using employment status (full-time, part-time, and not employed) as the dependent variable. From the remaining variables included in the NYS data file, select at least five variables that you think might have some relationship to an adolescent's employment status. Calculate the odds ratio for each coefficient, and explain what each odds ratio means in plain English.
- Compute an ordinal logistic regression model using the same set of dependent and independent variables that you used in Question 2.
  - Calculate the odds ratio for each coefficient, and explain what each odds ratio means in plain English.
  - Test the parallel slopes assumption. If you have access to Stata, use the Brant test. What do these results mean? Should you use the multinomial results? The ordinal model results? Or estimate a different model?

- (c) If you estimate a different model, report the results, and explain what each coefficient or odds ratio means in plain English.
4. Select one of the measures of delinquency, and recode it into three categories representing no delinquency (a score of 0), one delinquent act, and two or more delinquent acts. Compute an ordinal logistic regression model using this recoded measure of delinquency as the dependent variable. From the remaining variables in the data file, select at least five variables that you think might have some relationship with this measure of delinquency.
- Calculate the odds ratio for each coefficient, and explain what each odds ratio means in plain English.
  - Test the parallel slopes assumption. What does this result mean? (Use the Brant test, if available.)
  - If you have access to Stata or use R, estimate a partial proportional odds model using the same dependent and independent variables that takes into account your results from part (b). If you only have access to SPSS, estimate a multinomial logistic regression model using the same dependent and independent variables. Calculate the odds ratio for each coefficient, and explain what each odds ratio means in plain English.
  - How does the substance of the results from the initial ordinal logistic regression model compare to the substance of the results from either the partial proportional odds model or the multinomial logistic regression model? Which gives a more accurate representation of the relationships in the data? Explain why.

## References

---

- Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, 46, 1171–1178.
- Fu, V. (1998). Estimating generalized ordered logit models. *Stata Technical Bulletin*, 8, 160–164.
- Gould, W. (2000). Stata January 2000 (Technical STB-53).
- Lall, R., Walters, S. J., Morgan, K., & MRC CFAS Co-operative Institute of Public Health. (2002). A review of ordinal regression models applied on health-related quality of life assessments. *Statistical Methods in Medical Research*, 11, 49–67.
- Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using Stata* (2nd ed.). College Station: Stata Press.
- Nagin, D. (2005). *Group-based modeling of development*. Cambridge: Harvard University Press.
- O’Connell, A. A. (2006). *Logistic regression models for ordinal response variables*. Thousand Oaks: Sage.

- Peterson, B., & Harrell, F. E., Jr. (1990). Partial proportional odds models for ordinal response variables. *Applied Statistics*, 39, 205–217.
- Steffensmeier, D., Ulmer, J., & Kramer, J. (1998). The interaction of race, gender, and age in criminal sentencing: The punishment cost of being young, black, and male. *Criminology*, 36, 763–798.
- Williams, R. (2006). Generalized ordered logit/partial proportional odds models for ordinal dependent variables. *The Stata Journal*, 6(1), 58–82.

## Chapter six

---

# Count-Based Regression Models

### **C o u n t - B a s e d D e p e n d e n t V a r i a b l e s**

---

Why is it Inappropriate to Use Ordinary Least Squares (OLS) regression for Count-Based Dependent Variables?

What are the Characteristics of the Poisson Distribution?

### **P o i s s o n R e g r e s s i o n**

---

How does Poisson Regression Differ from OLS Regression?

How are Poisson Regression Coefficients Interpreted?

What is the Incident Rate Ratio and How is it Interpreted?

What is Over-Dispersion and What Produces It?

### **Q u a s i - P o i s s o n a n d N e g a t i v e B i n o m i a l R e g r e s s i o n**

---

How do Quasi-Poisson and Negative Binomial Regression Compare to Poisson Regression?

How Do you Test for Over-Dispersion?

Which Model Should You Use?

## **Z e r o - I n f l a t e d R e g r e s s i o n**

---

What is a Zero-Inflated Count Distribution?

How do Zero-Inflated Count Models Account for Too Many Zeros?

How Do you Interpret the Results from a Zero-Inflated Poisson or Zero-Inflated Negative Binomial Model?

**I**T IS COMMON IN CRIMINAL JUSTICE RESEARCH for the dependent variable of interest to be a count of some kind, such as the number of crimes occurring on a street over a one-month period or the number of arrests for adjudicated delinquents over one year. **Counts** are ratio-level measures. Each unit change in the variable of interest is of equal intervals, and zero means that the variable has no events—for example, a hot spot has no crimes during the period observed. Accordingly, it would seem that we should be able to use ordinary least squares (OLS) regression when we have a dependent variable that is measured as a count because OLS regression assumes that the dependent variable is an interval- or ratio-level measure. However, recall that OLS regression also assumes that the errors for the population model are normally distributed. In contrast, count data are typically positively skewed. The count of crimes on streets or communities, for example, will generally include many values close to zero, but at the same time, a small number of streets will have very large numbers of crimes. The same is true for records of offending among people. Thus, the normality assumption is likely not met. Additionally, OLS regression assumes homoscedasticity of variance. With count data, the variance generally grows as the counts get larger, violating this assumption as well. Another problem in using OLS regression with a count dependent variable is the possibility that the regression model will predict negative counts. Finally, OLS regression assumes that the dependent variable reflects a continuous latent construct even if measured on a discrete scale. Counts by nature reflect discrete or countable events or entities.

Because of these possible violations of assumptions, an OLS regression model may be biased when applied to a count-based dependent variable and, more importantly, may do a poor job of modeling the phenomenon of interest. With an OLS regression model, we assume that the phenomenon of interest (i.e., the dependent variable) was generated through an additive process. As was discussed in the section on transformations in Chap. 3, counts often arise out of multiplicative processes. A model based on the former will do a poor job of explaining data generated from the latter. The

solution to these problems is to use a regression modeling method designed for counts. These include Poisson regression, quasi-Poisson regression, negative binomial regression, zero-inflated Poisson regression, and zero-inflated negative binomial regression. These regression methods are the focus of this chapter.

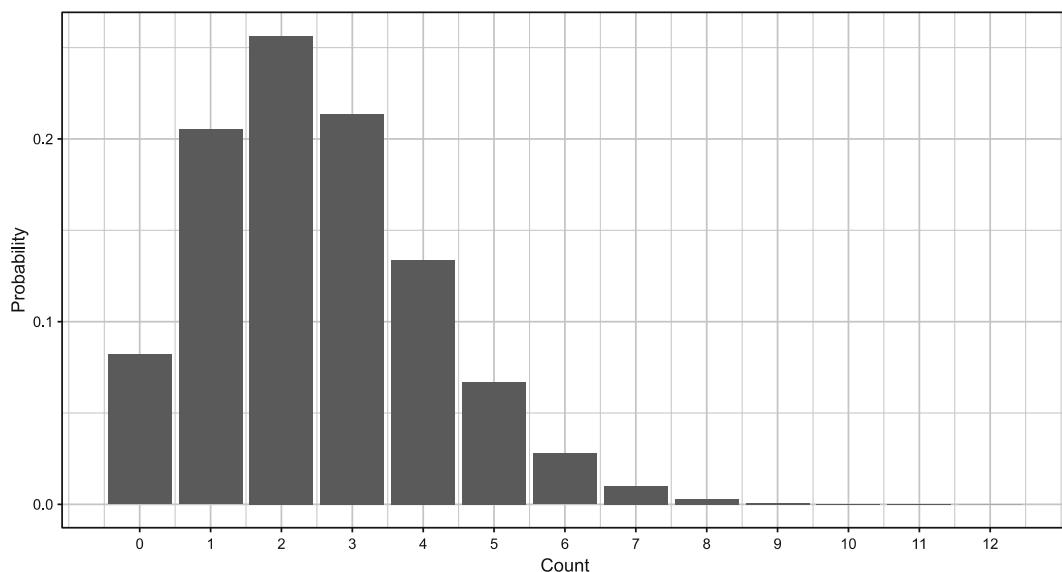
## The Poisson Distribution

In contrast to OLS regression that assumes the errors are normally distributed, **Poisson regression** assumes that the errors are Poisson distributed. The Poisson probability distribution is a family of distributions, each with a different mean ( $\mu$ ). An example of a Poisson distribution with a mean of 2.5 is shown in Fig. 6.1. This shows that a count of 0 will occur roughly 8% of the time (a probability of roughly 0.08), a count of 1 will occur a little over 20% of the time, and counts 12 and above will be very rare. Also, notice that the distribution is positively skewed; it is discrete and positively valued (i.e., it only includes whole numbers).

Figure 6.2 shows the Poisson probability distribution with a mean of 25. Notice that this distribution looks nearly normally distributed with only a slight positive skew. The larger the mean of the Poisson distribution, the

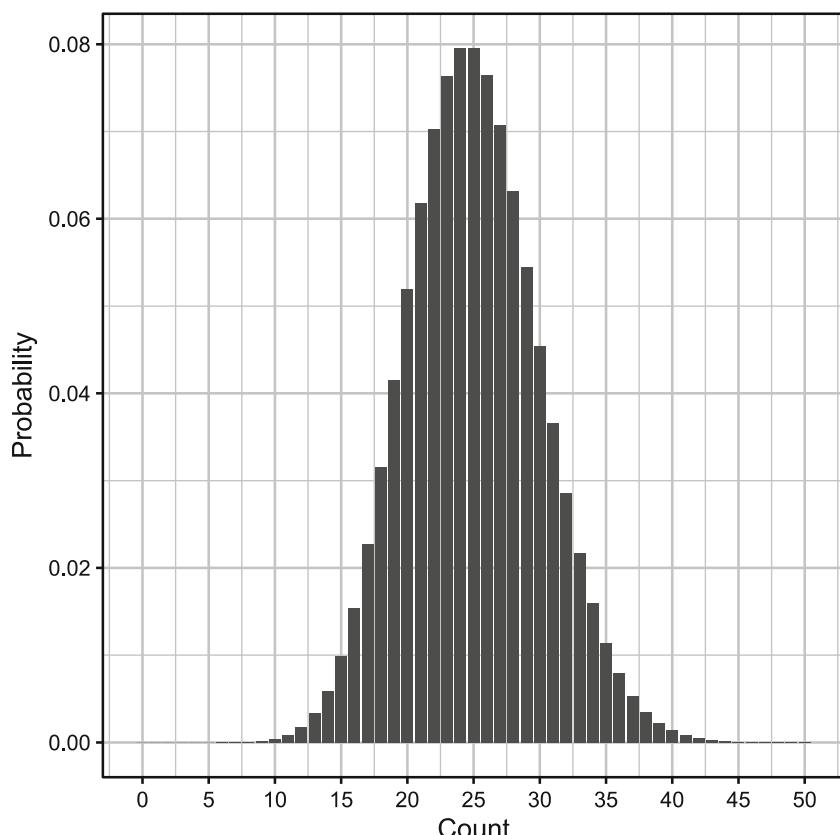
**Figure 6.1**

*Poisson probability distribution with a mean of 2.5*



**Figure 6.2**

*Poisson probability distribution with a mean of 25*



An important characteristic of the Poisson probability distribution is that the variance equals the mean. It is this relationship that produces heteroscedasticity. If we imagine a positive relationship between an independent variable and a count-based dependent variable, then the predicted mean count increases as the independent variable increases. As such, the variance in the dependent variable also increases as the independent variable increases. In an OLS model, this would produce a fan-shaped scatter plot of residuals against predicted values. If our only goal was to deal with heteroscedasticity, then we can simply square-root transform the counts. This is a variance stabilizing transformation. However, interpreting

regression coefficients from a model with a square-root transformed dependent variable is difficult, particularly in the case of counts. This relationship between the mean and variance, however, will be useful when we assess for over-dispersion, that is, excess variance, in count data, a topic that will be addressed later in the chapter.

The Poisson distribution makes strong assumptions about the data that may or may not be true. To understand these assumptions, we need to first explore the nature of count data. The first thing to realize is that if we are counting something, we are counting across time or space, sometimes both. For example, if we are counting the number of arrests for a group of youth who have previously been adjudicated for delinquent acts, we would need to do so for a period of time, such as 1 year. Alternatively, we could count the number of liquor stores within geographic divisions of an urban area at a specific point in time. Crime counts generally reflect both a geographic area and a unit of time, such as the number of crimes per hot spot over 6 months.

We may also think of binomial data as a count, but there is a critical difference between it and Poisson type count data. Under the binomial distribution, the count is the result of a fixed number of trials, such as the number of heads when a coin is flipped 100 times. In this example, the sample size is 100, and the outcome for each trial (coin flip) is recorded and has only one of two possibilities (heads or tails). With a Poisson distribution, we can also think of the event being counted as either occurring or not occurring for some small unit of time or space, but we only count the occurrences. Framed in this way, there is no meaningful upper limit to the number of trials (i.e., the sample size). Using the adjudicated youth example, we could divide the year into seconds and determine if each youth had or had not committed a crime during any given second. There are roughly  $3.154e + 7$  seconds per year (a huge number). But why seconds and not milliseconds? Furthermore, this would produce a dataset that was extremely large and contained mostly zeros (zero represents no crime during that unit of time). Thus, the Poisson distribution reflects the number of events that occurred over time or space rather than a binary outcome over a fixed number of trials.

An implication of this connection between the binomial and Poisson distributions is that the latter assumes that the events being counted are independent and that only one event can occur at a time or in a given location at a given time (Agresti 2003). This may seem overly restrictive, but it is not. In most situations, if you divide time or space into small enough units, then at a small enough unit you will have only one event per time/space. The exception would be two or more events that are not independent and co-occur. For example, if we consider crime counts over a year for a city, there might well be two or more crimes that occur during the same minute but at different locations (so independent). However, if two or more

crimes occur at the very same place at the same small unit of time, they are likely to be a joint event, such as multiple crimes committed by the same person at the same time and place. A lack of independence in count data creates over-dispersion (increased variability in the counts) and violates an assumption of Poisson models. Quasi-Poisson and negative binomial regression models correct for over-dispersion, addressing this problem.

## Poisson Regression

---

Poisson regression models a count dependent variable as a function of a set of independent variables. As a specific form of the general linear model, Poisson regression uses a function (called a link function) to connect the linear regression model that includes the independent variables and associated regression coefficients to the dependent variable. In the case of Poisson regression, this is the natural log-link [written as  $\ln(a)$ , the natural logarithm of  $a$ ]. A Poisson regression model with a single independent variable can be written as

$$\ln(y) = b_0 + b_1x_1, \quad \text{Equation 6.1}$$

when expressed in terms of the population parameters or

$$\ln(Y) = \beta_0 + \beta_1x_1, \quad \text{Equation 6.2}$$

when expressed in terms of the sample statistics, where  $y$  is a count-based dependent variable,  $b_0$  is the intercept, and  $b_1$  is the regression coefficient or slope associated with the independent variable,  $x_1$ . Conceptually, this is fitting a regression model to a log-transformed dependent variable, as we did with the sentence length in months in data in Chap. 3, as shown in Fig. 3.4. However, when  $y = 0$ , the log transformation is undefined, creating a problem with simply transforming  $y$  and estimating an OLS regression model, ignoring for the moment problems with other assumptions of OLS regression. Poisson regression's solution is to exponentiate each side of the equation, producing

$$y = e^{b_0 + b_1x_1}. \quad \text{Equation 6.3}$$

In this way, Poisson regression can model counts equal to zero. As with logistic regression, we estimate this model using iterative maximum

likelihood methods. Additional independent variables can be included, of course, and the model-building issues relevant to OLS regression remain the same. The differences are the log-link between the dependent variable and the linear regression function and the assumption that the errors are Poisson distributed.

We illustrate a simple Poisson regression model with fictitious data on the relationship between the number of rearrests over one year for a sample of 100 offenders and their risk score, measured on a four-point scale. The mean and variance of the number of rearrests are shown in Table 6.1 for each of the four risk levels.

The Poisson regression model to be tested is

$$\ln(y_i) = b_0 + b_1x_i + b_2x_2 + e_i,$$

Equation 6.4

where  $y_i$  is the number of rearrests per offender,  $i$ , and  $x_i$  is the risk score for each offender. Table 6.2 presents an OLS regression model on the raw untransformed counts and the results of the Poisson regression model. The coefficient for risk score is highly statistically significant in both models.

The difference between the OLS and Poisson model can be seen in Fig. 6.3 that shows a scatterplot of the raw data and the two regression lines. In this example, the difference is not large and would likely lead to a similar substantive conclusion regarding the relationship between risk score and the number of rearrests. However, the Poisson model has a slightly better fit and, more importantly, makes assumptions more consistent with the nature of the data and the underlying generating mechanisms.

**Table 6.1**

Number of rearrests over a 1-year period by risk level for 100 offenders

RISK LEVEL	MEAN	VARIANCE	MINIMUM	MAXIMUM	N
1	0.44	0.6733	0	3	25
2	0.68	0.5600	0	2	25
3	2.20	2.2500	0	5	25
4	3.68	4.0600	0	8	25
Total	1.75	3.5429	0	8	100

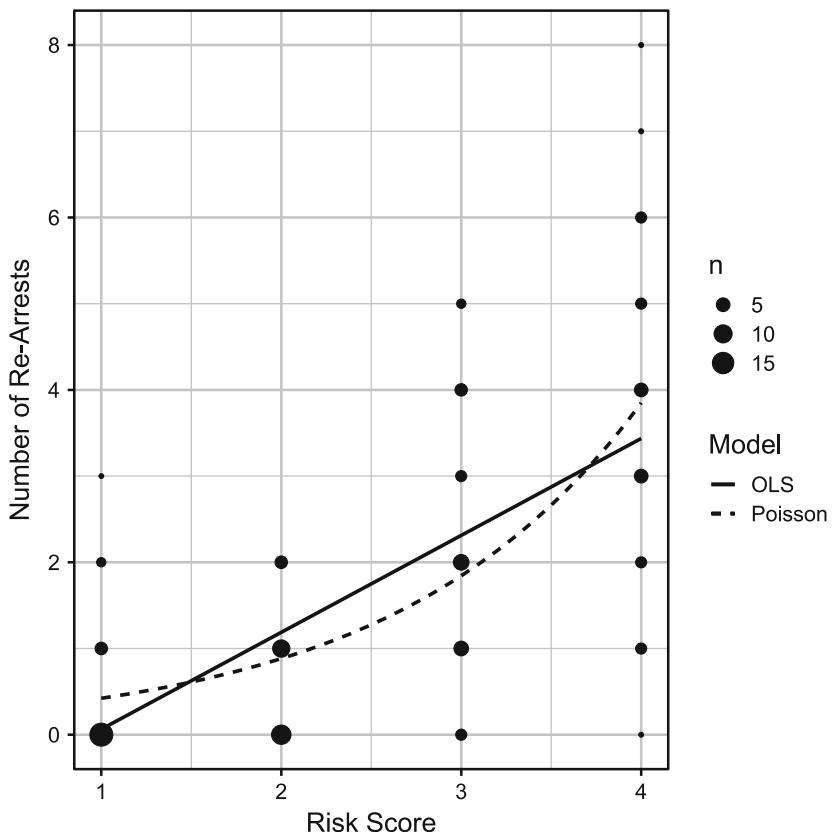
**Table 6.2**

Comparison of OLS and Poisson regression predicting rearrests from risk score

VARIABLE	OLS		POISSON	
	b	SE	b	SE
Intercept	-1.06	0.344	-1.60	0.285
Risk score	1.12	0.126	0.74	0.083

**Figure 6.3**

*OLS and Poisson regression lines and raw data for number of rearrests and risk score*



### Incident Rate Ratios (IRRs)

The interpretation of the coefficients from a Poisson model differs from an OLS model. In the OLS model in Table 6.1, the regression coefficient indicates that as the risk score increases by 1, the number of rearrests is predicted to increase by roughly 1 (1.12 to be exact). In the Poisson model, the coefficient of 0.74 indicates that as the risk score increases by 1, the logged number of predicted rearrests increases by 0.74. It is not easy to think in terms of a natural log scale. However, we can convert Poisson regression coefficients into an effect size that is easy to interpret.

Recall that in logistic regression, the exponent of the regression coefficient is an odds ratio and that an odds ratio provides a useful way to interpret logistic regression coefficients. Similarly, in Poisson and other count-based regression models, the exponent of the regression coefficient

facilitates interpretation. In this case, this value is called an **incidence rate ratio (IRR)** and reflects the rate at which the dependent variable grows relative to a one-unit change in the independent variable. A Poisson coefficient of zero indicates no relationship between the independent variable and the dependent variable, accounting for other variables in the model. The exponent of 0 is 1. An IRR of 1 indicates that the rate of growth is flat, as anything times 1 equals itself. Values greater than 1 indicate a positive relationship, and values less than 1 indicate a negative relationship. An IRR value of 2 indicates that the count doubles for every 1 unit change in the independent variable.

In our example, the Poisson coefficient for the risk score is 0.74. The exponent of this is 2.09 ( $\exp(0.736) = e^{0.736} = 2.088$ ). If we take a number and multiply it by 2.09, it roughly doubles or increases by 109%. If the coefficient had been  $-0.736$ , then the IRR would equal 0.48. Multiplying a value by 0.48 decreases it by roughly 1/2 or 52%. Thus, we can either interpret the IRRs directly or convert them to a percent change. This is accomplished with the two equations below, depending on whether the coefficient is positive or negative.

$$\% \text{ change} = \begin{cases} (1 - e^b) \times 100 = (1 - \text{IRR}) \times 100 & \text{if } b < 0 \text{ or } \text{IRR} < 1 \\ (e^b - 1) \times 100 = (\text{IRR} - 1) \times 100 & \text{if } b > 0 \text{ or } \text{IRR} > 1 \end{cases}$$

Because this is a simple Poisson model with a single independent variable, we can compute the incident rate ratios between adjacent risk scores from the raw means. The mean number of rearrests for risk scores of 1 and 2 are 0.44 and 0.68, respectively. Thus, the risk ratio is  $0.68/0.44$ , or 1.54. Continuing with other adjacent categories, we get 3.24 between risk scores of 2 and 3 ( $2.20/0.68$ ) and 1.67 between risk scores of 3 and 4 ( $3.68/2.20$ ). The average of these is 2.15, which is roughly similar to the results from the Poisson model. This illustrates the connection between the regression coefficient that reflects differences on a log scale to IRRs that reflect ratios or rates of change.

The fictitious Poisson model suggests that the number of rearrests increases by 109% (slightly more than doubling) as the risk score increases by 1. Notice that unlike OLS regression that is linear and would predict the same increase in rearrests throughout the full range of rearrest values, the Poisson model is linear on the log scale but increases multiplicatively on the raw scale, that is, by a proportion of the starting value. This is consistent with the assumption made by a Poisson model that the underlying process reflects multiplicative growth, producing a curved regression line.

### Significance Testing

Statistical significance for a Poisson regression model depends on the  $-2LL$  statistics (the  $-2$  log-likelihood) as it does with the logistic regression and other maximum likelihood-based regression models. As has been discussed in prior chapters, the difference between two  $-2LLs$  (or equivalently  $-2$  times the difference between two negative log-likelihoods) is distributed as a  $\chi^2$  with the degrees of freedom equal to the difference in the number of parameters in each model. The test of the statistical significance for the regression model, that is, whether the model significantly accounts for variability in the dependent variable, is the difference between the  $-2LL$  for the null model and the full model. The former is the model with no predictors (an intercept-only model), and the full model is the model with all of the independent variables of interest. This is shown below:

$$\text{Model } \chi^2 = (-2LL_{null}) - (-2LL_{full}). \quad \text{Equation 6.5}$$

For our example above, the null model has a  $-2LL$  of  $-195.691$ , and the full model has a  $-2LL$  of  $-147.525$ . Working this out produces a  $\chi^2 = 96.33$ . This has 1 degree of freedom and is highly significant, given that the critical value for a  $\chi^2$  with 1 degree of freedom and significance level of 0.05 is 3.84.

$$\begin{aligned} \text{Model } \chi^2 &= (-2LL_{null}) - (-2LL_{full}) \\ &= (-195.691) - (-147.525) \\ &= 96.33 \end{aligned} \quad \text{Equation 6.6}$$

This is interpreted much as an overall  $F$ -test is in OLS regression. It indicates that the model overall is statistically significant. In our example with a single predictor, it simply tests that this single coefficient is statistically significant. However, with a more complex model, it is testing whether the model as a whole is statistically significant. This is typically a low bar and should not be given a lot of weight in the interpretation. It does not tell you whether the model is the right model or a correctly specified model, only that it is better than no model.

The statistical significance of individual coefficients is also determined in the same fashion as with other regression models, such as logistic regression. Dividing the coefficient,  $b$ , by its standard error produces a  $z$ -test, and values greater than 1.96 are statistically significant at the .05 significance level. This is shown below:

$$z = \frac{b}{se_b}$$

The square of this may also be reported by statistical software programs as a Wald statistic.

### Exposure and Offsets

The discussion thus far has assumed that the counts across observational units are directly comparable to one another. For example, the number of rearrests per offender represented the number of rearrests for a fixed length of time, such as one year. Count data are often not directly comparable across the observational units. If you were studying crime counts across a sample of geographic areas, these areas may differ in terms of the population size. The crime rates per some unit of the population, such as crimes per 100,000 persons, may be more meaningful. Similarly, if the amount of time differs across units of observation, it may be necessary to standardize the counts by a common unit of time, such as the crimes per month or per week. It may also be meaningful to express a count relative to the size of an area. In each of these cases, we are converting a raw count into a rate to ensure comparability across observational units.

In count-based models, the denominator used to convert counts to rates is called the **exposure**. This word makes intuitive sense when we are dealing with different lengths of time, as it is easy to think of time as exposure. However, the concept is the same for other denominators, such as area size or population size or a combination of these.

Rather than calculating the rate ourselves and using that as the dependent variable, Poisson regression incorporates the exposure variable into the model using what is called an **offset**. The offset is the log of the exposure variable and is included on the right side of the regression equation but does not have an associated regression coefficient (or you could think of the coefficient as fixed to equal 1). For a simple model with a single independent variable, the regression equation would be written as

$$\ln(y) = b_0 + b_1x_1 + \text{offset}(x_2)$$

**Equation 6.7**

where  $y$  is our count variable,  $x_1$  is our independent variable, and  $x_2$  is the natural log of our exposure variable. To understand how this works, we will rewrite this equation, moving the offset to the left side of the equation by subtracting it from each side, producing:

$$\ln(y) - \text{offset}(x_2) = b_0 + b_1 x_1.$$

Substituting  $\ln(x_2)$  for the offset term, as they are equivalent, produces the following expression:

$$\ln(y) - \ln(x_2) = b_0 + b_1 x_1.$$

Recall that the difference between two logged values (i.e.,  $\ln(a) - \ln(b)$ ) is the same as the log of the ratio of the two values (i.e.,  $\ln(a/b)$ ). Thus, we can further manipulate the expression as follows:

$$\ln(y/x_2) = b_0 + b_1 x_1.$$

This shows that using an offset that is the log of an exposure variable converts the counts into rates. As such, we can think of Poisson regression as predicting rates, even though the dependent variables are counts.

In our running example, the length of time was equal for all offenders. As such, there is no need to use an offset, as the counts are directly comparable without converting them to rates. However, what would happen to this model if we used an offset? Notice that if we used an exposure of 1 to reflect 1 year of time, the log of 1 is 0. This would result in a left side of the equation equaling:

$$\ln(y) - \ln(1) = \ln(y),$$

or equivalently,

$$\ln(y/1) = \ln(y).$$

Thus, using an exposure of 1 would have no effect on the model. However, how would things change if we used an exposure of 12, thus converting the counts to the monthly rate of rearrests? The left side would become

$$\ln(y) - \ln(12).$$

If we move the offset back to the right side, the above equation becomes as follows:

$$\ln(y) = b_0 + b_1x_1 + \text{offset}(\ln(12)).$$

The log of 12 equals 2.48. Thus, we are subtracting a constant from the dependent variable or equivalently, adding a constant to the regression equation. The only effect this has on the model is to shift the intercept. Coefficients for independent variables remain unaffected. When we do this with the example above, the intercept changes from  $-1.60$  to  $-4.08$ , which happens to be a difference of 2.48. The coefficient for risk score remains the same, or 0.74, and thus, the IRR is still 2.09. Thus, a one-unit increase in the risk score increases the incident rate for rearrests by 109%. In situations where the exposure is not equal across observations, a model with an offset will clearly produce different, and more valid, results than a model without an offset. However, with an equal exposure across observational units, an offset is not needed, but harmless if included.

Why not just compute the rate and use that as the dependent variable? If you were using OLS regression, this would be a reasonable approach. However, with Poisson regression, the standard errors are based on the Poisson probability distribution that assumes each value of the dependent variable is a count, not a rate. Thus, it is critical that you use the raw counts as the dependent variable in count-based regression models. If these counts are only comparable across observations if converted to a rate, then use an offset to accomplish this. Some software programs allow you to specify either an exposure variable or an offset variable. The difference is that the latter must be the natural log of the former.

### An Example: California 1999 Uniform Crime Report Data

To illustrate a Poisson model with differential exposure across observational units, we will use the 1999 FBI Uniform Crime Report data for the state of California. These data include 58 counties, each county's population in 1999, and county-level crime counts. The data also include the unemployment rate for each county in 1999. We will explore whether the county's burglary rate is related to the unemployment rate.

The descriptive statistics for the variables that will be used in the model are shown in Table 6.3. As expected with a count variable, the number of burglaries per county is seriously positively skewed with a mean of a bit less than 4000 (3857) but a median of only 1271. The range is from a low of 24 burglaries to a high of 57,051 burglaries. The population across counties also differs substantially. The smallest county has a population of just over 1000 (1227), and the largest county has a population of over nine million (9,348,390). Thus, the raw counts of burglaries per county are not directly comparable. Rather, it is the rates we want to be comparing, in this case, the rates per 100,000 persons. These burglary rates ranged from 325 to 2037 with a mean and median that are roughly the same at just under 800 per

**Table 6.3**

California 1999 uniform crime report data for all 58 counties

VARIABLE	MEAN	MEDIAN	SD	MIN	MAX
Burglary count	3857.0	1271.0	8201.0	24.0	57051.0
Burglary rate per 100,000	790.9	796.2	315.0	325.0	2037.0
Unemployment rate	7.4	6.7	4.2	1.9	23.4
Population	571,283.2	158,463.5	1,327,538.3	1227.0	9,348,390.0

100,000. The unemployment rate also varies substantially across these counties from a low of 1.9% to a high of 23.4%.

We ran two regression models on these data. The first was simply an OLS model predicting the burglary rate from the unemployment rate. This model is shown here:

$$y_i = b_0 + b_1 x_{1i} + e_i$$

where  $y_i$  is the burglary rate per 100,000,  $x_1$  is the unemployment rate,  $e_i$  is the residual, and the subscript  $i$  is for each of the 58 counties, that is,  $i = 1 \dots 58$ . The second model was a Poisson regression model predicting the number of burglaries (i.e., the burglary count). This model included an offset that was the natural log of the population for each county. This model is shown here:

$$\ln(y_i) = b_0 + b_1 x_{1i} + \text{offset}(\ln(x_{2i})) + e_i$$

where  $y_i$  is the burglary count and  $x_2$  is the population for each county. Because of the offset, both models are modeling the burglary rate.

The results of these two regression models are shown in Table 6.4. The coefficient for the unemployment rate is significant in both models. For the OLS model, the coefficient is 36.75 and indicates that the burglary rate increases by roughly 37 crimes per 100,000 as the unemployment rate increases by 1-percentage point. For the Poisson model, the coefficient is 0.051. Exponentiating this produces an IRR of 1.05, indicating that the burglary rate increases by 5% for each 1-percentage-point increase in the unemployment rate. Thus, it appears that the burglary rate is affected by the unemployment rate, as we might expect. The scatterplot for these data and the regression line for the Poisson model are shown in Fig. 6.4.

An important difference between the OLS and Poisson model is in the significance test for the regression coefficient for the independent variable. The  $t$ -test for this coefficient in the OLS model is 4.218, whereas the  $z$ -test for the Poisson model is 89.5 (note that both the  $t$  and  $z$  reflect the ratio of the regression coefficient to its standard error and as such are directly

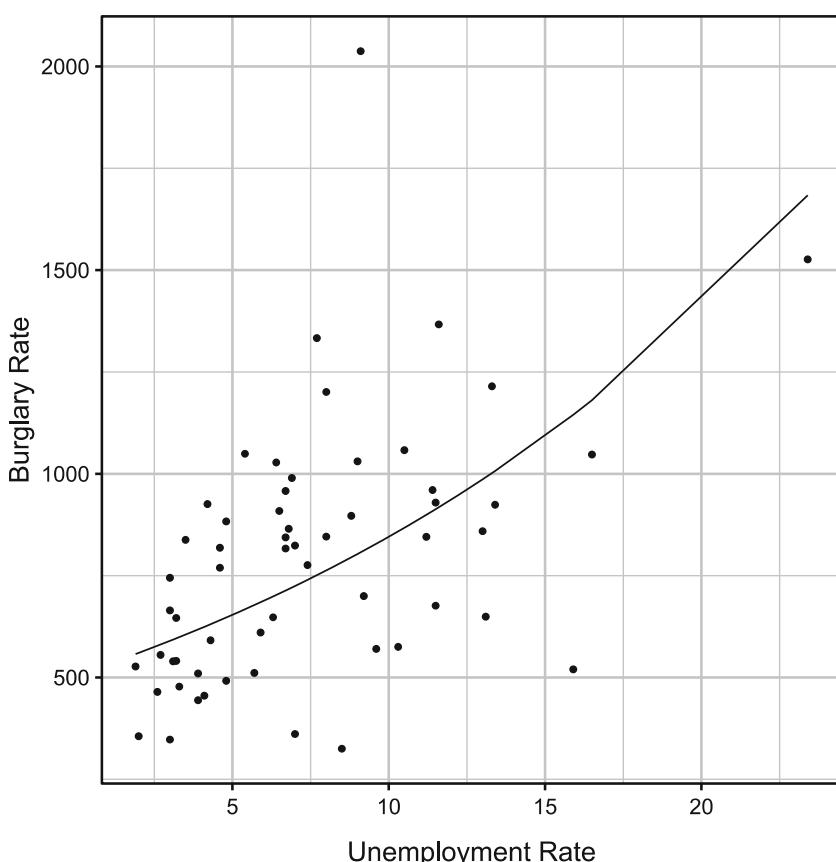
**Table 6.4**

Comparison of OLS and Poisson regression predicting burglary rate per 100,000 based on the unemployment rate

VARIABLE	OLS			POISSON		
	b	SE	t	b	SE	z
Intercept	518.70	74.074	7.002	-5.287	0.004	-1318.0
Unemployment Rate	36.75	8.713	4.218	0.051	0.0006	89.5

**Figure 6.4**

Scatter plot and Poisson regression line for Burglary rate per 100,000 by unemployment rate



comparable). The latter results in a much smaller  $p$ -value than the former. While this may seem like a reason to prefer the Poisson model over the OLS model, in this case, it actually reflects a problem with the Poisson model. Specifically, the burglary count data are over-dispersed resulting in an underestimate of the standard errors.

## Over-Dispersion in Count Data

---

Unlike OLS regression that computes the standard errors based on the variability in the observed data, after accounting for the independent variables, Poisson regression computes the standard errors based on the conditional mean of the model, ignoring the observed variability in the counts. That is, a Poisson regression model assumes that the conditional variance equals the conditional mean. The term conditional simply indicates that it is the mean and variance of the dependent variable after accounting for the independent variables. Thus, an important assumption of Poisson regression is that the conditional variance equals the conditional mean. If this is not the case (referred to as **over-dispersion**), then the model will be biased and the size of the bias may be substantial. This is often viewed as a major weakness of Poisson models.

Our example with rearrest counts for a sample of 100 offenders by four levels of risk provides a useful illustration of the relationship between the mean and variance. The means and variances for the number of rearrests for each risk score are shown in Table 6.1. These data were generated as random draws from four different Poisson distributions, each with a different mean. The population means for these four samples were 0.37, 0.78, 1.68, and 3.49, respectively. We see that the sample means differ slightly from the population means due to sampling error. More importantly, we see that the variances are roughly similar to the means within each risk score, again differing only due to sampling error. In the population, each of these four samples would have a mean and variance that were equal. Note, however, that the mean for the total sample is substantially smaller than the total variance. The overall distribution of the number of rearrests is over-dispersed given that it is a composite of four different Poisson distributions. Once these differences have been modeled by including risk score as an independent variable in a Poisson model, the conditional mean and variance are roughly equal and this assumption of Poisson regression is satisfied.

This illustrates that a potential source of over-dispersion in count data is differences in the underlying Poisson rate across observational units, such as individuals or counties. If our model fully accounts for these differences, the errors will be Poisson distributed. If the model does not fully account for these differences, then the errors will remain over-dispersed.

In most research contexts in criminology and criminal justice, we would assume that there are meaningful differences in the underlying Poisson rate for our dependent variable across the units of observation. More importantly, it would be rare for our regression model to fully account for these differences. Count data in criminology, such as crime counts, are typically over-dispersed, and our regression models rarely fully account for sources

of dispersion (Berk and MacDonald 2008). As will be shown below, this is the case with our California UCR burglary example.

Another potential source of over-dispersion in count data is a lack of independence in the counts. Recall that the Poisson distribution assumes that each event being counted is independent of every other event being counted. A dependency in counts may exist if an event increases the likelihood of another event occurring (Osgood 2000). In crime count data, such dependencies are likely, as crime tends to beget crime. Furthermore, multiple offenders may be arrested for the same crime or a single offender may commit multiple crimes over a short period of time. These are all potential sources of dependence in crime count data that can lead to over-dispersion.

A final potential source of over-dispersion is an excess of zeros in the count data. This can occur when there are two generating mechanisms underlying the counts, one that models whether a count is zero or nonzero and another that models the nonzero counts (see Greene 2018). For example, the number of times a youth has used marijuana within the past 30 days is likely to be zero-inflated. There is likely to be a binomial process that reflects a youth's decision to abstain from marijuana or use marijuana. There is likely a separate Poisson process that reflects the number of times those youth who use marijuana do so over a 30-day period. These two processes generate a zero-inflated count distribution and result in over-dispersion.

The first two sources of over-dispersion can be handled statistically by using either a **quasi-Poisson** or **negative binomial regression** model. Both of these regression methods take into account the conditional variability in the data. Zero-inflated over-dispersion can be handled through the use of a **zero-inflated Poisson (ZIP) regression** model or a **zero-inflated negative binomial (ZINB) regression** model. These zero-inflated models generate two sub-models, one for modeling the zero/nonzero process and one for modeling the nonzero count process. The ZINB version of these models can address over-dispersion in the nonzero counts.

A clue to whether a count-based dependent variable is over-dispersed is simply the ratio of the mean to the variance. If the variance is substantially larger than the mean, then the data may be over-dispersed. However, this is rather crude and does not take into account the extent to which the independent variables account for this over-dispersion. There are several tests for over-dispersion. The most straightforward is to re-estimate a Poisson regression model as a negative binomial regression and rely on the likelihood ratio test that compares the two. If it is significant, it reflects that the data are over-dispersed. We caution against relying entirely on such a test as there is little harm in using the negative binomial or quasi-Poisson model even if the test for over-dispersion is nonsignificant. The quasi-

Poisson and negative binomial models converge with the Poisson model as the observed over-dispersion decreases. With no over-dispersion, the quasi-Poisson and negative binomial models produce the same results of a Poisson model.

## Quasi-Poisson and Negative Binomial Regression

---

Both the quasi-Poisson and negative binomial models adjust for over-dispersion. Each method estimates an over-dispersion parameter; however, they differ in how this is estimated and in how this parameter value is incorporated into the model. In the quasi-Poisson model, the over-dispersion parameter is simply the ratio of the conditional variance to the conditional mean. A value of 1 indicates no over-dispersion. Values greater than 1 reflect over-dispersion, and values less than 1 reflect under-dispersion. This over-dispersion parameter,  $\theta$ , is used to adjust the standard errors of the Poisson regression model. Since  $\theta$  is scaled as a variance, the standard errors are inflated by the square root of  $\theta$ , as shown here:

$$se_{\text{quasi-Poisson}} = se_{\text{Poisson}} \sqrt{\theta}$$

where  $se_{\text{Poisson}}$  is a standard error for a regression coefficient from a Poisson regression model and  $se_{\text{quasi-Poisson}}$  is the over-dispersion adjusted standard error from a quasi-Poisson regression model. Essentially, the errors in a quasi-Poisson model are assumed to be Poisson distributed with a variance equal to  $\mu\theta$  rather than simply  $\mu$ . Importantly, the over-dispersion parameter is estimated from the results of a Poisson model, using the predicted and residual values from the model. Specifically,  $\theta$  is estimated as:

$$\theta = \frac{1}{n - k - 1} \sum \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i}$$

where  $y_i$  are the counts for each observation,  $i$ ,  $n$  is the number of observations,  $k$  is the number of independent variables in the model, and  $\hat{y}_i$  is the predicted count for each observation. Note that in the context of Poisson models, the  $\hat{y}$  is the predicted population mean ( $\mu$ ) of the Poisson distribution for each observation. Thus,  $\theta$  is not estimated directly as part of the model but as a post-estimate value. The *quasi* in the name indicates that the model is based on a Poisson probability distribution that has been altered (the variance has been inflated or deflated), making it a quasi-Poisson probability distribution. An implication of this is that, unlike the over-

dispersion parameter in a negative binomial model, discussed below, there is no statistical significance test for  $\theta$ . However, if  $\theta$  is large enough to change which regression coefficients are significant, then it is the more credible model.

If we apply this to our example where we are predicting the number of rearrests from an offender's risk score, we get an over-dispersion parameter of 1.13. This value is close to 1 and reflects only a small amount of over-dispersion in the observed data relative to the expectation under a Poisson model. We should not be surprised by this because these data were drawn randomly from a Poisson probability distribution. Fitting a quasi-Poisson regression model to these data results in coefficients that remain the same. This is an important feature of quasi-Poisson regression: The coefficients will be identical to the coefficients from a Poisson regression. The only difference will be in the standard errors. In Table 6.2, the standard error for the risk score coefficient was 0.83. Adjusting this for the observed over-dispersion produces a quasi-Poisson standard error of 0.088, as shown below.

$$se_{\text{quasi-Poisson}} = 0.083\sqrt{1.13} = 0.088$$

The associated  $z$ -test changes from 8.85 ( $0.736/0.083$ ) to 8.31 ( $0.736/0.088$ ), a trivial change that does not alter the substantive interpretation regarding the relationship between risk score and the number of rearrests.

In the case of negative binomial regression, we estimate the over-dispersion parameter as part of the model itself and the distribution of the errors is assumed to conform to a negative binomial probability distribution. In contrast to the Poisson distribution where the variance equals the mean (i.e.,  $Var(Y) = \mu$ ), the variance of a negative binomial distribution is a function of the mean and an over-dispersion parameter. This function is shown below:

$$Var(Y) = \mu + \mu^2\alpha \quad \text{Equation 6.8}$$

where  $\mu$  is the expected mean and  $\alpha$  is the over-dispersion parameter (Ver Hoef and Boveng 2007). This parameter is estimated via maximum likelihood as part of the model fitting process. As such, it is a parameter of the regression model and we can use the likelihood ratio (LR) test to determine if the estimated over-dispersion parameter significantly differs from zero in the population. That is, do we have sufficient evidence to conclude that the over-dispersion is not just sampling error? An important difference between  $\theta$  and  $\alpha$  is that the former is 1 when there is no over-dispersion, whereas  $\alpha$  is

0 when there is no over-dispersion. Also,  $\theta$  can be less than 1, reflecting under-dispersion, but  $\alpha$  has 0 as a minimum value.

Reanalyzing the number of rearrests data shown in Table 6.2 using a negative binomial regression model alters both the regression coefficients and the standard errors, although in this example the differences are slight given the trivial level of over-dispersion. That is, in this case, the quasi-Poisson and negative binomial models are the same to 3 digits. The effect of model choice on the coefficients will be explored when we reanalyze the California UCR data, where the over-dispersion is more significant. However, we can now test whether the level of over-dispersion in these data is statistically significant. This is done with the LR test that we have already used to compare nested models. In this case, we compare the Poisson model to the negative binomial model. The difference between these two is that the latter includes an additional parameter,  $\alpha$ . The LR test is as follows:

$$\text{LR test} = -2[(\text{LL}_{\text{Poisson}}) - (\text{LL}_{\text{negative binomial}})].$$

In our example, the negative log-likelihood for the Poisson model is  $-147.5251$ , and for the negative binomial model, it is  $-147.4533$ . This produces a LR test of 0.14, as shown below.

$$\text{LR test} = -2[(-147.5251) - (-147.4533)] = 0.14.$$

This is distributed as a  $\chi^2$  with one degree of freedom. The significance level for this  $\chi^2$  is 0.70. Thus, this over-dispersion parameter is not statistically significant, consistent with expectation given that we know that the conditional distributions (i.e., the distribution within each risk score) were drawn from true Poisson probability distributions.

### An Example: Reanalysis of the California 1999 Uniform Crime Report Data

A more realistic comparison of the Poisson, quasi-Poisson, and negative binomial regression can be carried out by reanalyzing the California 1999 Uniform Crime Report (UCR) data. We show the results of these three models in Table 6.5. The over-dispersion parameter for the quasi-Poisson model was 258.44. Thus, the observed conditional variance was over 250 times larger than expected under a true Poisson distribution, indicating a high degree of over-dispersion. The over-dispersion parameter for the negative binomial model is 0.09. Recall that these two over-dispersion parameters are scaled differently and as such are not directly comparable. Any value greater than 0 reflects at least some observed over-dispersion.

**Table 6.5**

Comparison of Poisson, quasi-Poisson, and negative binomial regression models predicting county-level burglary rates per 100,000 based on county unemployment rates in California in 1999

VARIABLE	POISSON			QUASI-POISSON			NEGATIVE BINOMIAL		
	b	SE	z	b	SE	z	b	SE	z
Intercept	-5.287	0.0040	-1318.0	-5.287	0.0645	-81.98	-5.202	0.0809	-64.33
Unemployment rate	0.051	0.0006	89.5	0.051	0.0092	5.57	0.045	0.0095	4.722

We can use the LR test to determine if we can reject the null hypothesis that  $\alpha$  is zero in the population. The negative LL for the Poisson model is  $-7215.212$ , and for the negative binomial model, it is  $-422.011$ . The LR test value is  $13,586.4$ , as shown below.

$$\text{LR test} = -2[(7215.212) - (-422.011)] = 13584.4$$

The critical value for a  $\chi^2$  with one degree of freedom is 3.84. This LR test is highly statistically significant, allowing us to reject the null that  $\alpha$  is zero and conclude that these data are over-dispersed. Note that this supports using either a quasi-Poisson or negative binomial model.

How do the results across these three models compare? First, notice that the regression coefficients for the Poisson and quasi-Poisson model are identical, reflecting that the quasi-Poisson model is simply a Poisson model with standard errors inflated (or deflated) based on observed over- or under-dispersion. Because of this, we cannot estimate the  $-\text{LL}$  for a quasi-Poisson model. The regression coefficients for the negative binomial model differ slightly, however, from the two Poisson models. The difference is not large. Converting the regression coefficients for the unemployment rate to IRRs suggests that the burglary rate increases either by about 5.2% or 4.6% for every one-percentage-point increase in the unemployment rate. The reason for the difference is in the weighting each model gives to each count. In the Poisson and quasi-Poisson models, each count is weighted directly relative to its size. That is, the weights are proportional to the count or conditional mean. Thus, there is a linear relationship between the weight given each observation and its count value. In contrast, the negative binomial model weights the counts differently because of the different functional form of the relationship between the expected counts (expected means) and the variance. The expected variance has a squared term in it (Eq. 6.8), and this produces a quadratic, rather than a linear,

relationship between the expected variances and means such that the variance increases quadratically as the expected mean increases. These count-based models weight observations by the inverse of their expected variance. Thus, the negative binomial model will give less weight to the larger counts relative to a Poisson/quasi-Poisson model, producing slightly different regression coefficients (see Ver Hoef and Boveng 2007, for detailed exploration of this issue).

Another difference in Table 6.5 is the change in the standard errors. The standard error for the unemployment rate coefficient changes from 0.0006 in the Poisson model to 0.0092 and 0.0095 in the quasi-Poisson and negative binomial models, respectively. This is roughly a 15-fold increase, suggesting that the Poisson standard errors were substantially downwardly biased. Adjusting for over-dispersion results in much more modest  $z$ -tests of 5.57 and 4.722 compared to 89.5 for the Poisson model.

When faced with over-dispersed counts, should we use the quasi-Poisson or negative binomial approach? First, in most situations, it should not matter much, producing results that would lead to comparable substantive conclusions. Second, there is no clear consensus in the literature. Both Gardner et al. (1995) and Ver Hoef and Boveng (2007) argue in favor of the quasi-Poisson model in most situations unless the focus is on generating a population estimate of over-dispersion, in which case the negative binomial model would be preferred. As of this writing, the negative binomial model is far more common in the criminology research literature.

## Zero-Inflated Poisson and Negative Binomial Regression

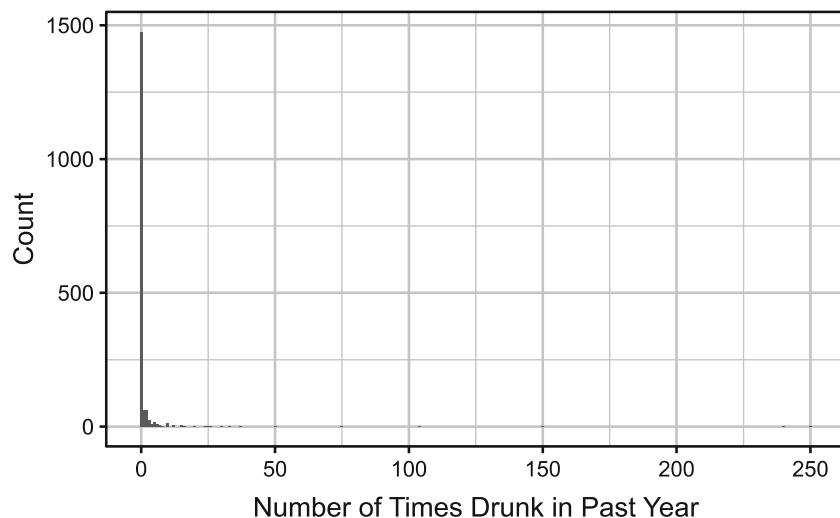
---

Count data may have too many zeros. What we mean by this is that there are more zeros than would be expected under a Poisson distribution. We describe such distributions as being zero-inflated, leading to over-dispersion. Unfortunately, fitting a quasi-Poisson or negative binomial model does not effectively model zero-inflated data. As the names imply, the zero-inflated Poisson and zero-inflated negative binomial regression models do appropriately handle data of this type.

A zero-inflated regression model is essentially two regression models in one. The first is a logistic regression model predicting whether the count is zero or nonzero. The second model is a Poisson or negative binomial model predicting the nonzero counts. This reflects that a zero-inflated count distribution likely has two underlying data-generating mechanisms, one for the zero/nonzero distinction and the other for the count portion.

**Figure 6.5**

*Histogram of the number of times drunk in past year for a sample of 1710 youth from the National Youth Survey*



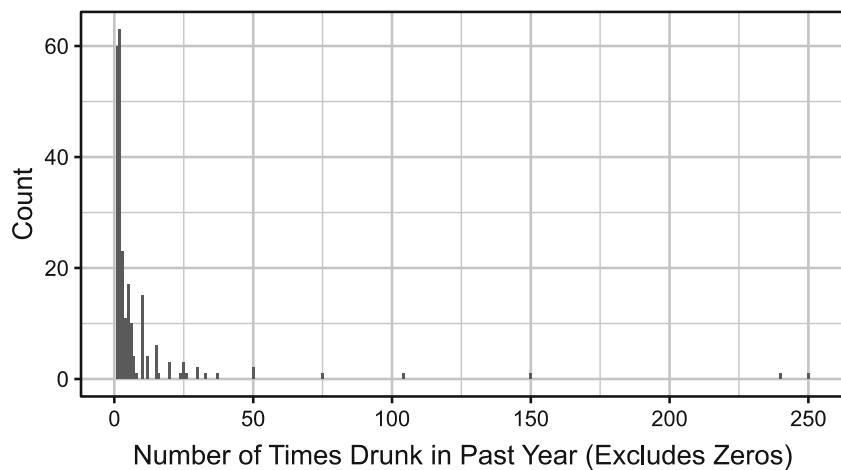
We use data from the National Youth Survey to illustrate this method. This sample includes a subset of 1710 youth aged 11 through 17.<sup>1</sup> A question on the survey was, *How many times in the last year have you been drunk?* The responses to this question are zero-inflated with 1476 (86%) of the youth responding zero times and 234 of the youth responding one or more times. The range was from 0 to 250. A histogram of these data is shown in Fig. 6.5 with a very large spike at zero. There are single cases with responses up to 250 that cannot be seen on this graph. To better visualize the nonzero data, we show a histogram in Fig. 6.6 of only the 234 cases that responded that they had been drunk one or more times in the past year.

---

<sup>1</sup>The data file we will use first represents a subset of the data from the National Youth Survey, Wave 1. The sample of 1,725 youth is representative of persons aged 11–17 years in the USA in 1976, when the first wave of data was collected. While these data may seem old, researchers continue to publish reports based on new findings and interpretations of these data. One of the apparent strengths of this study was its design; the youth were interviewed annually for 5 years from 1976 to 1980 and then were interviewed again in 1983 and 1987. The data file on our Website was constructed from the full data source available at the Inter-University Consortium of Political and Social Research, which is a national data archive. Data from studies funded by the National Institute of Justice (NIJ) are freely available to anyone with an Internet connection; go to <http://www.ipusr.umich.edu/NACJD>. All seven waves of data from the National Youth Survey are available, for example.

**Figure 6.6**

*Histogram of the number of times drunk in past year, excluding zeros, for a sample of 234 youth from the National Youth Survey*

**Table 6.6**

Means as associated statistics for 1710 youth from the National Youth Survey

VARIABLE	MEAN	SD	MIN.	MAX.
Sex (1 = female; 0 = male)	0.468	0.499	0	1
Age	13.870	1.940	11	17
Importance of college	3.927	1.441	1	5
GPA	2.729	0.822	0	4

We theorize that the following variables may influence a youth's drinking behavior: age, sex, school GPA, and the importance they place on going to college. The latter is a simple 3-point ordinal scale coded as 1, 3, and 5. The descriptive statistics for these independent variables are shown in Table 6.6.

We present the zero-inflated negative binomial model in Table 6.7. Notice that there are two models reported. The first is the negative binomial model, and the second is the logistic model. It is important to note that the logistic regression model is predicting whether a youth responded that they had not been drunk in the past year. Thus, the coefficients are flipped from what you might expect. For example, the positive coefficient for sex, coded as 1 for females and 0 for males, indicates that females are more likely to have zeros relative to males. Conversely, they are less likely to have been drunk at any point in the last year. However, this effect is not statistically significant. The count-based model is interpreted as expected with positive

**Table 6.7**

Zero-inflated negative binomial for the number of times drunk in the past year using National Youth Survey Data

SUB-MODEL AND VARIABLE	<i>b</i>	SE	<i>z</i>	<i>p</i>
Count				
Sex (1 = female; 0 = male)	-0.457	0.256	-1.79	0.074
Age	0.397	0.125	3.18	0.001
Importance of college	-0.470	0.187	-2.51	0.012
GPA	0.064	0.091	0.70	0.484
Intercept	-4.122	1.874	-2.20	0.028
Zero/nonzero (predicting 0)				
Sex (1 = female; 0 = male)	0.577	0.457	1.26	0.207
Age	-1.105	0.186	-5.93	0.000
Importance of college	0.401	0.294	1.36	0.173
GPA	0.589	0.225	2.61	0.009
Intercept	11.881	2.263	5.25	0.000

coefficients indicating a positive relationship between the independent variable and the count.

The Vuong test, related to a likelihood ratio test, compares the zero-inflated model (Poisson or negative binomial) to a nonzero-inflated model (a corresponding Poisson or negative binomial). For our example, this test is statistically significant ( $p < .0005$ ), supporting the inference that the data are zero-inflated. Stated more simply, the zero-inflated model fits the data better than a nonzero-inflated negative binomial model. The test for the over-dispersion parameter,  $\alpha$ , is also significant, indicating that there is significant over-dispersion in the count-based portion of this model after accounting for the excess zeros.

Focusing first on the logistic sub-model, we see that age is related to whether or not a youth has been drunk in the past year. The negative coefficient means that younger youth are more likely not to have been drunk. Flipping this around, older youth are more likely to have been drunk, as we would expect. Also, youth with higher GPAs are less likely to have been drunk than youth with lower GPAs. Notice, however, that while age has a similar effect on the frequency of being drunk with the frequency increasing with age, GPA is unrelated to the frequency of being drunk. Thus, it appears that GPA is related to the decision to drink to the point of being drunk but not related to how often a youth gets drunk, illustrating the ability of zero-inflated models to differentiate these two processes.

The importance a youth places on going to college also has a differential effect on these two processes. In contrast to GPA, the importance a youth places on going to college does not appear to relate to whether that youth has been drunk in the past year but does appear to relate to his or her frequency of being drunk. Youth who place greater importance on going to college report having been drunk fewer times in the past year.

These effects can be converted to more interpretable effect sizes. As with a regular logistic regression model, the coefficients for the zero/nonzero sub-model when exponentiated become odds ratios. Similarly, the exponentiated coefficients for the count sub-model become IRRs. Flipping the sign of the coefficients for the zero/nonzero model to ease interpretation, the odds ratio for age is 3.02. As age increases by 1 year, the odds that a youth has been drunk at any point in the last year increases threefold. The IRR for age from the count sub-model is 1.49. Subtracting 1 from this and multiplying by 100 produce 49%. For those youth who have been drunk at some point in the past year, the frequency of being drunk increases by roughly 50% for each 1-year increase in age. Focusing on the importance of college, the odds ratio for the zero/nonzero sub-model, again flipping the sign of the coefficient, is 0.45, suggesting that the likelihood of being drunk at some point in the past year decreases by roughly a half as the importance of going to college increases by 1. The corresponding IRR for the count sub-model is 0.63. Subtracting this from 1 and multiplying by 100 produce 37%. Thus, the frequency of being drunk in the past year decreases by 37% for each 1-point increase in the importance of college.

We included the same independent variables in both sub-models of this analysis. This is not necessary. The count sub-model and the zero/nonzero sub-model can have different sets of independent variables. These can be overlapping or completely distinct. It is possible for the generating mechanism for each model to be different. As such, your theoretical propositions regarding these processes may specify different variables and you can incorporate this into the specification of the model tested.

## Chapter Summary

---

Count-based data are common in criminological research including outcomes such as crime **counts** for geographic areas or the number of rearrests over a given time period for a group of individuals. Counts by nature are discrete, positively valued whole numbers. When we want to model counts as a dependent variable in a regression model, OLS regression is generally a poor choice. This stems from several features of counts, including that the variance increases with the count, resulting in heteroscedasticity, a positively skewed distribution, and the impossibility of negative values. On a natural log scale, counts tend to be normally distributed, unless over-dispersed. More importantly, counts tend to grow multiplicatively rather than in an additive manner. That is, it is generally more meaningful to conceptualize change in the counts as percent change rather than as a raw change in number. The regression coefficients can also be transformed into **incident rate ratios** for ease of interpretation and

comparability. **Exposure** is the denominator in count-based regression models that is used to convert counts to rates, and with Poisson regression, this is incorporated into the model using an **offset**.

Count-based regression approaches such as **Poisson**, **quasi-Poisson**, and **negative binomial regression** models appropriately handle these characteristics of counts as a dependent variable and do so by using a log-link, thus modeling the log of the count. The difference between a Poisson model and both quasi-Poisson and negative binomial models is that the latter two adjust for over-dispersion in the count data. Count data where the variance is greater than the mean represent **over-dispersion**. Sources of over-dispersion include unmodeled differences in the count rate across observational units and dependencies in the counts. The quasi-Poisson model produces regression coefficients that are identical to a Poisson model but with standard errors that are adjusted for any observed over-dispersion. Negative binomial regression models adjust for over-dispersion differently, and the regression coefficients may differ compared to a Poisson model but usually only slightly.

Another complication with count data is the possibility that the distribution has an excess of zeros relative to what would be expected from a Poisson process. These are called zero-inflated distributions and can be modeled with zero-inflated versions of either the Poisson or negative binomial models (referred to as **ZIP** and **ZINB regression models**). Zero-inflated approaches separately model the zeros versus nonzeros and the counts, producing two sub-models. The former is a logistic regression model, and the latter is a Poisson or negative binomial model, but they are estimated together. The independent variables contributing to each model can be the same or different.

Historically, particularly in ecology where count data is also very common, using OLS regression on log-transformed counts was a common approach to handle this type of data (O'Hara and Kotze 2010). However, with the ready availability of statistical software able to perform the count-based regression models discussed in this chapter, there is little reason to not use these modeling methods. They better reflect the nature of the data and are less likely to violate key assumptions.

## Key Terms

---

**Counts** A measure that reflects the number of events or some other countable entity. The values are discrete, whole numbers and may include 0 as a valid count, i.e., no observed events for a given observational unit.

**Exposure** A variable that reflects the exposure for a count. This might be the length of time, such as a minute or year during which the counts were made, the population base for the counts, the size of a

geographic area, or some combination of these. An exposure variable is the denominator in the ratio that converts a count into a meaningful rate, such as crimes per 100,000 persons per county per year.

**Incident rate ratio (IRR)** The incident rate ratio is an effect size for interpreting Poisson and negative binomial regression coefficients. It is the exponent of the coefficient that reflects the rate at which the count is increasing or decreasing for every one-unit change in the independent variable.

**Negative binomial regression** A variant on Poisson regression that accounts for observed over-dispersion. This method estimates an over-dispersion parameter as part of the model and will produce regression coefficients that may differ from a Poisson model.

**Offset** A term in a count-based regression model used to convert the counts into meaningful rates. The offset is the natural log of exposure.

**Over-dispersion** Variability in a distribution of counts that is in excess of what would be expected for a Poisson distribution where the variance equals the mean.

**Poisson regression** A regression method for count-based dependent variables. This method makes the assumption that the counts, after accounting for variability due to the independent variables, are Poisson distributed. That is, this method does not adjust for over-dispersion.

**Quasi-Poisson regression** A variant on Poisson regression that accounts for observed over or under-dispersion. The regression coefficients will be the same as with a Poisson model, but the standard errors are adjusted for observed dispersion in the data.

**ZIP, ZINB regression** Zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models deal with count data that have an excess of zeros relative to expectation under a Poisson or negative binomial model.

## Symbols and Formulas

---

IRR      Incident rate ratio

$\theta$       Over-dispersion parameter

A single independent variable Poisson model in population form:

$$\ln(Y) = \beta_0 + \beta_1 x_1$$

A single independent variable Poisson model in sample form:

$$\ln(y) = b_0 + b_1 x_1$$

A single independent variable Poisson model in exponentiated form:

$$y = e^{b_0 + b_1 x_1}$$

The incident rate ratio (IRR) of a regression coefficient ( $b$ ):

$$\text{IRR} = e^b$$

Percent change in the dependent variable for a 1-unit change in an independent variable:

$$\% \text{ change} = \begin{cases} (1 - e^b) \times 100 = (1 - \text{IRR}) \times 100 & \text{if } b < 0 \text{ or IRR} < 1 \\ (e^b - 1) \times 100 = (\text{IRR} - 1) \times 100 & \text{if } b > 0 \text{ or IRR} > 1 \end{cases}$$

Count-based regression model with an offset where  $x_2$  is the natural log of exposure:

$$\ln(y) = b_0 + b_1 x_1 + \text{offset}(x_2)$$

Over-dispersion parameter for a quasi-Poisson model:

$$\theta = \frac{1}{n - k - 1} \sum \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i}$$

Standard errors for a quasi-Poisson model adjusted for over-dispersion:

$$se_{\text{quasi-Poisson}} = se_{\text{Poisson}} \sqrt{\theta}$$

## Exercises

---

- 6.1. Assume we have count data for two groups, shown below.

Group 1:	6	7	7	10	6	6	8	2	6	5
Group 2:	5	12	7	6	5	9	6	10	5	7

Calculate the following:

- (a) The sum of the counts for each group.
  - (b) The natural log of these sums.
  - (c) The difference between the natural log of these sums.
  - (d) The natural log of the ratio of these sums.
  - (e) The exponent of the answer to (c) above.
  - (f) The ratio of the mean for group 1 relative to the mean for group 2.
  - (g) How do (e) and (f) compare?
  - (h) The answer to (e) above is an IRR. Convert this to a percent change.  
How would you interpret this value?
- 6.2. Below are the results from a Poisson regression model.

INDEPENDENT VARIABLE	<i>b</i>	SE	<i>z</i>
Intercept	-1.25	0.15	-8.33
$x_1$	-0.13	0.06	-2.17
$x_2$	0.46	0.08	5.75
$x_3$	0.05	0.04	1.25

- (a) Assume that the quasi-Poisson over-dispersion parameter,  $\theta$ , is 2.00.  
Adjust the standard errors using this value.
- (b) Compute over-dispersion adjusted *z*-values.
- (c) Assuming that the *z* critical value for a .05 significance level is 1.96,  
how do the results change after adjusting for over-dispersion?
- (d) Compute the percent change associated with a one-unit change in  
each of the independent variables.

## Computer Exercises

The data file used to illustrate the application of the count-based regression models in this chapter can be found in either SPSS (nys\_l.sav or nys\_l\_student.sav) or Stata (nys\_l.dta) format. Alternatively, one of the files can be imported into R using the `read.sav()` or `read.dta()` functions from the *foreign* package. The illustration of the commands below assumes that you have opened one of these files into SPSS, Stata, or R, and can also be found in the sample syntax files in both SPSS (Chapter\_5.sps) and Stata (Chapter\_5.do) format.

## SPSS

### *Poisson Regression*

The GENLIN procedure is used to estimate generalized linear models in SPSS (version 15 and later), and it has the option of specifying a Poisson distribution. The structure of the syntax is as follows:

```
GENLIN Dep_var BY IV_Factors (ORDER=ASCENDING)
  WITH IV_scale
  /MODEL IV_Factors IV_scale
  /INTERCEPT=YES
  /OFFSET=offset_var
  /DISTRIBUTION=POISSON LINK=LOG
  /CRITERIA CILEVEL=95 CITYPE=WALD LIKELIHOOD=FULL
  /PRINT FIT SUMMARY SOLUTION.
```

Note that the dependent variable is specified after the GENLIN command. The BY argument is used to specify categorical independent variables, while scale independent variables come after the WITH argument. An offset variable may be added with the OFFSET= argument.

### *Quasi-Poisson Regression*

SPSS does not have the option to calculate quasi-Poisson regression.

### *Negative Binomial Regression*

The GENLIN procedure can also be used to estimate a generalized linear model where the dependent variable has a negative binomial distribution. The structure of the syntax is as follows:

```
GENLIN Dep_var BY IV_Factors (ORDER=ASCENDING)
  WITH IV_scale
  /MODEL IV_Factors IV_scale
  /INTERCEPT=YES
  /OFFSET=offset_var
  /DISTRIBUTION=NEGBIN(1) LINK=LOG
  /CRITERIA CILEVEL=95 CITYPE=WALD LIKELIHOOD=FULL
  /PRINT FIT SUMMARY SOLUTION.
```

The code is the same as a Poisson regression generalized linear model with the exception of the DISTRIBUTION= argument. Negative binomial regression is specified by NEGBIN(1) instead of POISSON.

### *Zero-Inflated Poisson/Negative Binomial Regression*

SPSS does not offer regression models that accommodate a dependent variable with a zero-inflated distribution.

## Stata

### Poisson Regression

The **glm** command for generalized linear models can be used for Poisson regression models in Stata. It is specified with the **family()** argument, and the family type for a Poisson model is **poisson** (make sure to use a lowercase *p*). You must also specify the argument **link(log)** and have the option to add an offset variable with the **offset()** argument. The basic structure of a Poisson regression model using the **glm** command is as follows:

```
glm dep_var indep_vars, family(poisson) link(log)
offset(offset_var_name)
```

Recall that you need to specify **i.** in front of any categorical variables (e.g., **i.cat\_indep\_var1**). Relative risk ratios can be obtained instead of odds ratios by adding the argument **eform** to the right of the comma, as follows:

```
glm dep_var indep_vars, family(poisson) link(log) eform
```

The **predict** command is for post-estimation and can be used to obtain predicted estimates. Simply execute your *glm* model, and then, use the **predict** command, along with the name of the new variable to be used to store the predicted estimates.

```
predict new_var_name
```

### Quasi-Poisson Regression

Stata does not have the option to calculate quasi-Poisson regression.

### Negative Binomial Regression

Negative binomial regression models are also conducted in Stata using the **glm** command, whereby the **family(nbinomial)** argument is specified. You must also specify the argument **link(log)**, and you have the option to add an offset variable with the **offset()** argument.

```
glm dep_var indep_vars, family(nbinomial) link(log)
offset(offset_var_name)
```

As with Poisson regression when using *glm*, relative risk ratios can be obtained instead of odds ratios by adding the argument **eform** to the right of the comma, as follows:

```
glm dep_var indep_vars, family(nbinomial) link(log) eform
```

The **predict** command is for post-estimation and can be used to obtain predicted estimates. Simply execute your *glm* model, and then, use the **predict** command, along with the name of the new variable to be used to store the predicted estimates.

```
predict new_var_name
```

### *Zero-Inflated Poisson/Negative Binomial Regression*

Zero-inflated Poisson and negative binomial regression models can be estimated with maximum likelihood in Stata by, respectively, using the **zip** or **zinb** commands. You must also specify the **inflate()** argument, which is where you add the variable(s) that predict excess 0s. If the **offset()** argument is used, it is specified within the **inflate()** command. The structure of both commands is as follows:

```
zip dep_var indep_vars, inflate( var_pred_0s,  
    offset(offset_var))  
zinb dep_var indep_vars, inflate( var_pred_0s,  
    offset(offset_var))
```

With both **zip** and **zinb** commands, if you want relative rate ratios reported, add the argument **irr** to the right of the comma:

```
zinb dep_var indep_vars, irr  
    inflate(var_pred_0s)
```

Additionally, if you would like to compare whether the zero-inflated Poisson is a better fit in comparison to the zero-inflated negative binomial, specify the argument **zip** to the right of the comma:

```
zinb dep_var indep_vars, zip  
    inflate(var_pred_0s)
```

As with using *glm*, the **predict** command is for post-estimation and can be used to obtain predicted estimates. Simply execute your *zip* or *zinb* model, and then, use the **predict** command, along with the name of the new variable to be used to store the predicted estimates.

```
predict new_var_name
```

## R

### *Poisson Regression*

The **glm()** function for generalized linear models, which is in the *stats* package, can be used to fit many types of count regression models. It is specified with the **family=** argument, and the family type for a Poisson model is **poisson** (make sure to use a lowercase *p*). You can view the model output using the **summary()** function. The basic structure of a Poisson regression model using the **glm()** function is as follows:

```
model <- glm(dep_var ~ indep_var1 + indep_var2,  
    data=dataset_name,  
    family="poisson")  
summary(model)
```

You may add  $\sim 1$  instead of  $\sim \text{indep\_var1} + \text{indep\_var2}$  if you wish to run an intercept-only model. And an exposure variable can be incorporated into the model using the **offset()** function. However, recall that you also need to log it, which can be done with the **log()** function as follows:

```
glm(dep_var ~ indep_var1 +
    offset(log(offset_var)),
    data=dataset_name, family="poisson")
```

The **predict()** function can be used along with your *glm*-class object to obtain predicted estimates for each case, and you can assign it as a new variable using the <- assignment operator. Then, the **aggregate()** function can be nested within the **predict()** function to get a mean of the predicted estimates. Note that the user must specify the formula within the aggregate function whereby the independent variable is the grouping variable.

```
predict(model)
df$pred_est_var<-predict(model)
aggregate(predict(model) ~x, data=df, mean)
```

The coefficients from the model can be transformed into odds ratios by exponentiating the coefficients. To do this, use both the **coef()** and **exp()** functions on the *glm*-class object that is storing the model output (model in our case).

```
exp(coef(model))
```

### *Quasi-Poisson Regression*

The specification of a quasi-Poisson regression model with the **glm()** function is similar to a standard Poisson regression model, but you will change the **family=** argument to type **quasipoisson**. As with above, remember that this will require the installation of the *stats* package. The basic structure of the syntax is as follows:

```
model <- glm(dep_var ~ indep_var1 + indep_var2,
                 data=dataset_name,
                 family="quasipoisson")
summary(model)
```

You may add  $\sim 1$  instead of  $\sim \text{indep\_var1} + \text{indep\_var2}$  if you wish to run an intercept-only model. And an exposure variable can be incorporated into the model using the **offset()** function. However, recall that you also need to log it, which can be done with the **log()** function as follows:

```
glm(dep_var ~ indep_var1 +
    offset(log(offset_var)),
    data=dataset_name,
    family="quasipoisson")
```

The **predict()** function can be used along with your *glm*-class object to obtain predicted estimates for each case, and you can assign it as a new variable using the <- assignment operator. Then, the **aggregate()** function can be nested within the **predict()** function to get a mean of the predicted estimates. Note that the user must specify the formula within the aggregate function whereby the independent variable is the grouping variable.

```
predict(model)
df$pred_est_var<-predict(model)
aggregate(predict(model) ~x, data=df, mean)
```

The coefficients from the model can be transformed into odds ratios by exponentiating the coefficients. To do this, use both the **coef()** and **exp()** functions on the *glm*-class object that is storing the model output (model in our case).

```
exp(coef(model))
```

### Negative Binomial Regression

Negative binomial regression models are set up a little differently than Poisson and quasi-Poisson as they use the function **glm.nb()**, which is from the *MASS* package. Notice that the family= argument is no longer needed.

```
model <- glm.nb(dep_var ~ indep_var1 +
    indep_var2, data=dataset_name)
summary(model)
```

You may add ~1 instead of ~indep\_var1+indep\_var2 if you wish to run an intercept-only model. You can still incorporate an exposure variable into the model using the **offset()** function, but remember that you will need to log it using the **log()** function as follows:

```
glm.nb(dep_var ~ indep_var1 +
    offset(log(offset_var)),
    data=dataset_name)
```

The **predict()** function can be used to obtain predicted estimates for each case, and you can assign it as a new variable using the <- assignment operator. Note that even though we are using the function **glm.nb()**, it still produces a *glm*-class object. Then, the **aggregate()** function can be nested within the **predict()** function to get a mean of the predicted estimates. Note that the user must specify the formula within the aggregate function whereby the independent variable is the grouping variable.

```
predict(model)
df$pred_est_var<-predict(model)
aggregate(predict(model) ~x, data=df, mean)
```

The coefficients from the model can be transformed into odds ratios by exponentiating the coefficients. To do this, use both the **coef()** and **exp()** functions on the *glm*-class object that is storing the model output (model in our case).

```
exp(coef(model))
```

### *Zero-Inflated Poisson/Negative Binomial Regression*

Zero-inflated regression models can be estimated with maximum likelihood in R by using the **zeroinfl()** function in the *pscl* package. The distribution, whether Poisson or negative binomial, is specified with the **dist=** argument. Additionally, you must specify **link="logit"** when using count data. The **zeroinfl()** function is set up for zero-inflated Poisson regression models as follows:

```
zip <- zeroinfl(dep_var ~ indep_var1 +
  indep_var2, data=dataset_name,
  dist="poisson", link="logit")
summary(zip)
```

The **zeroinfl()** function is set up for zero-inflated negative binomial regression models as follows:

```
negbi <- zeroinfl(dep_var ~ indep_var1 +
  indep_var2, data=dataset_name,
  dist="negbin", link="logit")
summary(negbi)
```

You can then assess which model is a better fit with the likelihood ratio test by using the **lrtest()** function as follows:

```
lrtest(zip, negbi)
```

As with the **glm()** function, you can incorporate an offset into the model using the **offset()** and **log()** functions:

```
zeroinfl(dep_var ~ indep_var1 +
  offset(log(offset_var)), 
  data=dataset_name,
  dist="poisson", link="logit")
```

**Problems**

1. Enter the data from Exercise 6.1 above into SPSS, Stata, or R. Analyze these data with the counts as the dependent variable and group as the independent variable. How does the regression coefficient for the values you computed as part of Exercise 6.1 compare with the results from this analysis?

Open the NYS Wave I data file (nys\_1.sav, nys\_1\_student.sav, or nys\_1.dta) to complete exercises 2–4. Make sure to select a variable from the dataset to be the exposure/offset variable when appropriate. When selecting independent variables, keep in mind the guidelines for selecting covariates that you learned in prior chapters (e.g., be mindful of multicollinearity).

2. Compute a Poisson regression model using the number of nights a week the student reported studying on average (*eve\_stdy*) as the dependent variable. Include their grade point average (*gpa*) and their perception of the importance of attending college (*imp\_coll*) as independent variables.
  - (a) Explain the results in plain English.
  - (b) Based on the results, predict the number of nights the student spent studying, and then, present the average number of nights the students spend studying for each grade point average.
3. Compute a negative binomial regression model using the self-reported frequencies that the student hit his/her parent in the previous year (*bit\_prnt*) as the dependent variable. Add the student’s perceived importance of attaining a good career (*imp\_job*) and the perceived importance of having friends (*imp\_frns*) as independent variables.
  - (a) Explain the results in plain English.
  - (b) Compare the negative binomial regression model results with a Poisson regression model, and then, explain the differences. Which appears to be a better model fit?
4. Compute both a zero-inflated Poisson and zero-inflated negative binomial model using the self-reported frequencies that the student cheated on a school exam (*cheat*) as the dependent variable. Further, select five independent variables from the dataset that you suspect may affect cheating.
  - (a) Compare the regression coefficients between the two models. Describe which model you think is a better fit.

## References

---

- Agresti, A. (2003). *Categorical data analysis* (2nd ed.). Hoboken, NJ: John Wiley & Sons.
- Berk, R., & MacDonald, J. M. (2008). Overdispersion and Poisson regression. *Journal of Quantitative Criminology*, 24(3), 269–284.
- Gardner, W., Mulvey, E. P., & Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological Bulletin*, 118(3), 392.
- Greene, W. H. (2018). *Econometric analysis* (8th ed., p. 905). Chennai: Pearson Education India.
- O'Hara, R., & Kotze, J. (2010). Do not log-transform count data. *Nature Proceedings*, 1 (2), 118–122.
- Osgood, D. W. (2000). Poisson-based regression analysis of aggregate crime rates. *Journal of Quantitative Criminology*, 16(1), 21–43.
- Ver Hoef, J. M., & Boveng, P. L. (2007). Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology*, 88(11), 2766–2772.

## Chapter seven

---

# Multilevel Regression Models

## **H o w D o W e A n a l y z e C l u s t e r e d o r M u l t i l e v e l D a t a ?**

---

What are Clustered Data or Multilevel Data?

Why Can't we Use Ordinary Least Squares Regression for Such Data?

## **V a r i a n c e C o m p o n e n t s M o d e l**

---

What are Fixed and Random Effects?

How Do We Interpret the Results from a Variance Components Model?

What is the Intraclass Correlation and How is it Interpreted?

## **R a n d o m I n t e r c e p t M o d e l**

---

How is the Random Intercept Model Different from the Variance Components Model?

How is Explained Variance Computed and Interpreted?

What are Between and Within Effects?

How Do We Test for Differences in Between and Within Effects?

## **R a n d o m C o e f f i c i e n t M o d e l**

---

How is the Random Coefficient Model Different from the Random Intercept Model?

How Do We Test for Differences Between the Random Coefficient Model and the Random Intercept Model?

## **A d d i n g C l u s t e r (L e v e l 2) C h a r a c t e r i s t i c s**

---

How are Cluster (Level 2) Characteristics Added to the Random Coefficient Model?

How Do We Interpret the Effects of Cluster (Level 2) Characteristics?

**I**N THE REGRESSION MODELS that we have presented so far, such as ordinary least squares (OLS) regression, logistic regression, and Poisson regression, we have assumed a single sample of cases. This is expressed in the assumption of independence that we make for these models. However, sources of nonindependence are common in criminology and criminal justice. For example, we might have repeated measurements on a sample of individuals. The individuals may be independent of one another, but the repeated measurements within individuals are not. Perhaps less obvious but statistically quite similar would be observations that occur within some natural clustering. A classic example in the social sciences is students nested or clustered within classrooms. If we were interested in studying the academic performance of primary school students, there may well be a statistical dependency associated with the class that each student is in. That is, academic achievement might be affected by the classroom for each student. This would result in the achievement scores of students that share a classroom to be at least slightly more similar to each other than to students as a whole. In criminology, we might be interested in studying police officer attitudes toward procedural justice and survey officers from a sample of police departments. The officers are clustered in the departments, creating a potential dependency in the data, with officers within a department sharing more similar views. Or we might sample people who live in crime hot spots, or high crime communities. People who live in the same hot spots or in the same high crime communities will likely be more alike to each other creating a degree of dependence.

How are we to analyze clustered data? Such data, also called **multilevel data** (and hence, the name of the statistical models we discuss in this chapter), are very common in criminal justice research. In thinking about clustered data as multilevel data, we would define the cluster—neighborhood, school, police department, hot spot above—as the level-2 data. The unique observations—typically, individual cases—would be defined as the level-1 data. We are also not limited to thinking of our data as having only two levels and could conceivably work with data that have three or more

levels. An example of three-level data might involve a study that begins with a sample of communities (level 3), followed by a sample of hot spots within each of those communities (level 2), and then the individual residents within each of the selected hot spots (level 1). Put in terms of clustered data, we have residents clustered within hot spots that are clustered within communities. The more general point to describing clustered data as multilevel data is that the lowest level of data will represent the total number of observations in our data—whatever these observations happen to represent. Each additional level of clustering then reflects a higher level of the data structure.

Why does the clustering of data matter? There are both statistical and theoretical reasons for why we may want to pay attention to clustered data. Statistically, observations within a cluster will tend to be more similar to each other than to observations from different clusters. For example, survey respondents within a neighborhood will tend to be more alike on key individual characteristics when compared to survey respondents from another neighborhood, regardless of whether that other neighborhood is across town or across the nation. The increased similarity of cases within a cluster has consequences for our statistical tests, making it more likely that we will find statistically significant results, since cases within a cluster will tend to exhibit a similar pattern of association and consequently smaller, and biased, standard errors.<sup>1</sup>

Theoretically, we may also have an interest in the multilevel structure of the data that points to important effects of the cluster on relationships observed at level-1. For example, how might characteristics of a neighborhood—such as poverty rate or unemployment rate—affect the relationship we might observe between a respondent's gender and fear of crime? If we find that female respondents express higher levels of fear of crime, then we could ask the question about whether this statistical relationship is the same across neighborhoods. Does the effect of gender on fear of crime change across neighborhood? If the effect is essentially the same across neighborhoods, it tells us that neighborhoods may be unimportant for understanding fear of crime. In contrast, if we find that the effect of gender does vary across neighborhood, we may then want to investigate why the effect varies. Is it due to other characteristics of the neighborhood, such as poverty, unemployment, vacant houses, and the like? Multilevel data are structured in such a way that the clustering of cases presents both a challenge and an opportunity to test for the effects of different independent variables measured on different units of analysis.

---

<sup>1</sup>If our concern is primarily in statistically accounting for clustered observations, we can use what are referred to as robust standard errors and is available as an option in most statistical packages. We do not discuss these standard errors in this chapter, but encourage curious readers to consult Angrist and Pischke (2009).

In this chapter, we provide a brief introduction to what are known as **multilevel models**<sup>2</sup> that account for the clustering of cases—the multilevel structure of the data—and can tell us interesting things about the nature of the statistical relationships we are studying. We take as a given that there is something informative or interesting about the multilevel structure of the data—that the clustering of observations is not simply a statistical nuisance to be corrected for. In the discussion that follows, we restrict our attention to the analysis of dependent variables measured at the interval or ratio level of measurement. We also limit our discussion to two-level models: We have individual-level data (level 1) nested within one set of clusters (level 2) or measurement time points (level 1) nested within individuals (level 2). We also provide a substantive example using a count regression model. At the same time, we want to note that there is a growing literature on how to apply multilevel models to other types of dependent variables within the framework of the generalized linear model. The overall logic and approach we present here remain fundamentally unchanged when conducting these types of analyses.

## A Simple Multilevel Model

---

The simplest possible multilevel regression model is an intercept-only model. Such a model would have no predictor variable or independent variables but would include a term to reflect the clustering of the data. To illustrate this, we will start with a very basic model without any multilevel elements, shown here

$$y_i = \beta_0 + \epsilon_i \quad \text{Equation 7.1}$$

where  $y_i$  is the dependent variable for each individual, with  $i = 1 \dots n$ . In this simple model, the intercept,  $\beta_0$ , will equal the mean, and the error term,  $\epsilon_i$ , will reflect the errors or deviation of each individual case from the overall mean (we will call this the grand mean to differentiate it from the group or cluster means). We assume that errors in the population are normally distributed with a mean of 0 and a variance of  $\sigma_\epsilon^2$ . Because there are no predictors in this model, the variance of the errors is simply the variance of

---

<sup>2</sup>These models are also known as mixed models, random-effects models, growth-curve models, and hierarchical linear models. Since these phrases all take on different meanings across the social and behavioral sciences, we use multilevel models, since that phrase seems to hold less potential for confusion across disciplinary boundaries.

the dependent variable (i.e., the total variance). Furthermore, the intercept is a fixed effect and the error term is a random effect. This distinction between fixed and random effects will be explored below but is worth taking note of here.

Let us assume that the data are clustered in some way and that this clustering is reflected in a nominal variable indicating cluster membership. Recall from Chap. 3 that we can include a nominal independent variable into a regression via dummy coding, that is, by creating a set of binary 0/1 variables, each indicating whether an observation is a member of a particular group, with 1 = yes and 0 = no. Also, recall that if the number of categories for our nominal variable equals  $k$ , then we only need  $k - 1$  dummy variables to fully account for the nominal variable in the regression model. This is shown in the equation below.

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{j-1} x_{j-1} + \epsilon_i \quad \text{Equation 7.2}$$

where  $x_1$  to  $x_{j-1}$  are the dummy variables for the cluster or grouping variable. All of the terms in this regression equation except for the final error term are fixed, and as such, this is considered a fixed-effects model.

A multilevel model changes how the cluster variable is entered into the regression model. Rather than being entered as a series of dummy variables, it is treated as a random effect. The model is rewritten as

$$y_{ij} = \beta_0 + u_j + \epsilon_{ij} \quad \text{Equation 7.3}$$

where  $u_j$  is the deviation of the mean for each cluster from the overall mean with  $j = 1 \dots k$ . We can think of Equation 7.3 as a series of separate regression equations, one for each cluster or group, such as:

$$\begin{aligned} y_{i1} &= \beta_{0,1} + \epsilon_{i1} \\ y_{i2} &= \beta_{0,2} + \epsilon_{i2} \\ y_{i3} &= \beta_{0,3} + \epsilon_{i3} \\ &\dots \\ y_{ik} &= \beta_{0,k} + \epsilon_{ik} \end{aligned} \quad \text{Equation 7.4}$$

Thus,  $\beta_{0j}$  is the intercept (in this case, the mean) for each cluster and  $\epsilon_{ij}$  is the error within each cluster. These are our level-1 models because we are modeling the observations within each cluster. With this approach, we do not need to specify a reference category that is omitted. That is, we have a

regression equation for each cluster. We can then write a level-2 model that treats these cluster intercepts (means) as the dependent variable, as shown here:

$$\beta_{0j} = \beta_{00} + u_j \quad \text{Equation 7.5}$$

where  $u_j$  is the random errors of the intercepts across clusters. Note that we now have double subscripts for the intercept, indicating that this is the intercept across individuals and clusters. Combining these is equivalent to Eq. (7.3). Because  $u_j$  is a random variate associated with the cluster intercepts, this is a random-intercept model. More importantly, we have decomposed the error variance in Eq. (7.1) into two components: that associated with the clusters (the variance of  $u_j$ ) and the variance of the individual observations within their respective clusters (the residual variance of  $e_{ij}$  after accounting for any clustering effect). This is why these models are also called **variance components models**.

### Fixed-Effects and Random-Effects

What is a **fixed-effect** versus a **random-effect**? The distinction is important in multilevel models, but, unfortunately, the statistics literature is filled with conflicting definitions and explanations of these concepts. Within the classical framework of analysis of variance, a fixed-effects model is one where the levels of the independent variable represent the levels of interest (i.e., *fixed* levels), whereas in a random-effects model, the levels are assumed to have been sampled from a distribution of possible levels (see for example Hays 1988, see also Gelman 2005). For example, if the independent variable is three different forms of hot spots policing and we are interested in testing the differences between these three levels, then we have a fixed-effects model. That is, we have a set of *fixed* regression coefficients comparing the difference in the means between these conditions. In contrast, if the independent variable is three different police units and our interest is not whether these particular units differ but whether there is meaningful variability in our dependent variable across units and these are just three possible units, then we have a random-effects model. Notice that the same data could be conceptualized and estimated under either a fixed- or random-effects model. However, our focus shifts from the fixed differences between the groups (clusters, in our multilevel framework) to the variability among the groups.

Most multilevel models are mixed-effects because they include a mix of random- and fixed-effects. Within the multilevel or mixed-effects regression framework, the distinction between the fixed- and random-effects is reflected in whether an independent variable is represented by a regression coefficient (a fixed-effect) or by a random variate (a random-effect).

### A Substantive Example: Bail Decision-Making Study

One of the most important decisions in the criminal process is the bail and release decision made by the judge shortly after the arrest of most individuals. Several decades of research have shown that defendants who have been released during the pretrial period—the time between arrest and disposition (conclusion) of a case—will tend to receive more lenient punishments if convicted. Those defendants who remain in jail during the pretrial period, due to the judge denying release altogether or to the judge requiring a bail amount the defendant could not pay, will typically be more likely to go to prison and to receive a slightly longer sentence if sentenced to prison. Importantly, these effects hold even after taking into account other characteristics of the defendant, such as prior record and severity of the offense.

As part of a larger project exploring judicial decision making in the bail and release decision, John Goldkamp and Michael Gottfredson conducted two studies in Philadelphia—the first a pilot study to examine the factors that influenced the level of bail judges required and the second a test of whether the use of what were called bail guidelines made the decision making more consistent and equitable across defendants (Goldkamp and Gottfredson 1985). We focus our attention on data from the first study. Goldkamp and Gottfredson selected a random sample of 20 judges to participate in the study and then selected a random sample of 240 cases per judge that required a bail and/or release decision. This resulted in a total sample of 4800 cases clustered evenly across the 20 judges. Put in the terminology of levels of data, the 4800 cases represent the level-1 data, while the 20 judges represent the level-2 data.

We can consider each judge as a separate experimental condition—cases were randomly assigned to each judge, ensuring broad similarity of the cases, and thereby creating the opportunity to assess how similarly or differently judges would process these cases. Our attention in the example that follows is the bail decision for each case that was indicated by the dollar amount the judge set for the person’s release.<sup>3</sup> Our dependent variable is the common logarithm of bail amount. Of the original 4,800 cases, bail amounts were required of 2314 defendants, which comprise the sample for the following analyses.

Table 7.1 presents the means and standard deviations for bail amount (in dollars) and logged bail amount for each of the 20 judges. The average bail amount required by each judge varies considerably. For example, the average bail amount required by Judge 9 was \$1652, while for judge 6, it

---

<sup>3</sup>There was a 10% rule in effect in Philadelphia at the time of the study, meaning that defendants would only need to post 10% of the dollar amount requested by the judge in order to ensure their freedom during the pretrial period.

**Table 7.1**

Means and Standard deviations of bail amounts by Judge in Philadelphia

JUDGE	BAIL AMOUNT (DOLLARS)		BAIL AMOUNT (LOGGED)	
	MEAN	SD	MEAN	SD
1	2076.30	4513.85	3.04	0.45
2	4784.88	8522.53	3.28	0.56
3	1901.68	3414.13	3.05	0.41
4	1830.43	2808.85	2.97	0.46
5	2204.58	3890.20	2.98	0.50
6	17,486.78	60,861.05	3.51	0.73
7	1842.76	3320.16	3.00	0.42
8	2117.76	3583.45	3.01	0.47
9	1652.86	2099.00	2.95	0.46
10	3627.04	10,960.82	2.96	0.57
11	4576.19	7170.58	3.38	0.47
12	7299.55	16,270.59	3.34	0.63
13	3385.96	6310.05	3.21	0.50
14	7945.31	24,293.52	3.32	0.65
15	4944.00	11726.94	3.28	0.51
16	8747.40	19,082.44	3.57	0.50
17	2184.62	3182.93	3.11	0.41
18	4158.82	7765.01	3.24	0.51
19	1956.30	3053.54	3.07	0.38
20	6246.90	23,285.99	3.23	0.58

was \$17,487. Note that the values for logged bail are much smaller and have a more limited range. This is due to the fact that the logarithm used here—base 10—reflects the exponent for the number of times 10 would be multiplied by itself to reproduce the bail amount (e.g.,  $\log(100) = 2$  and  $\log(1000) = 3$ ). Consequently, a difference of 1.0 on the logarithmic scale used here is equivalent to a ten-fold increase in bail amount.

Table 7.2 presents the fixed-effects regression results for this example using OLS regression and dummy variables for judges 2 through 20 with judge 1 as the reference category. Effect coding rather than simply dummy coding was used. The implication of this is that the intercept reflects the overall mean and each coefficient reflects the deviation from the grand mean for that particular judge. Thus, the overall mean is 3.04 and the mean for judge 2 is 0.24 above the overall mean, or  $3.04 + 0.24 = 3.28$ .

To help establish a baseline for the multilevel models discussed in the remainder of this chapter, it will be useful to present the results from a one-way ANOVA, where we treat each judge as a type of experimental condition and test for differences in mean bail amounts. Table 7.3<sup>4</sup> presents the results for logged bail amount (this will be the outcome measure we rely on in subsequent analyses in this chapter). The test of the null hypothesis of

<sup>4</sup>Due to rounding, some of the judge means estimated with the coefficients in Table 7.2 will differ at the second decimal when compared to the means reported in Table 7.1.

**Table 7.2**

Fixed effects OLS regression analysis of logged bail amount

JUDGE	COEFFICIENT $b_k$	STANDARD ERROR
2	0.24	0.07
3	0.01	0.06
4	-0.08	0.07
5	-0.06	0.06
6	0.47	0.06
7	-0.05	0.06
8	-0.04	0.07
9	-0.09	0.06
10	-0.08	0.07
11	0.33	0.07
12	0.30	0.07
13	0.16	0.07
14	0.28	0.06
15	0.24	0.06
16	0.53	0.07
17	0.07	0.07
18	0.19	0.07
19	0.03	0.06
20	0.18	0.06
Intercept	3.04	0.04

**Table 7.3**

ANOVA summary table for regression model from Table 7.2

SOURCE	SS	df	MS	F
Between	74.28	19	3.91	14.48
Within	619.38	2294	0.27	
Total		2213		

equality of means across judge gives us an  $F$ -test value of  $F = 14.48$  with  $df_1 = 19$ ,  $df_2 = 2294$ , and  $p < 0.001$ . We then conclude that the mean logged bail amount across this sample of 20 judges is significantly different. The summary ANOVA table for this analysis is shown in Table 7.3.

How would we change this model such that the cluster for the variable judges becomes a random effect instead of a fixed effect? In this case, we are no longer interested in the individual coefficients for each judge but rather the variability in the bail outcomes attributed to judges, not just those in this sample, but judges more generally. Stated differently, what is our estimate of variance of  $\sigma_u^2$  from Equation 7.3? Intuitively, you might think that it is simply the  $MS_{\text{between}}$  from the ANOVA model as this separated the variance into the observed variance between the groups and within the groups. However, it is a bit more complex than this. If there are no differences in the bail amounts across the judges, then the true variance associated with judges should be zero. However, under the null, our expectation for the  $F$ -statistic is that it equals one, that is that the variance

between equals the variance within. Why is this? Well, we would expect some variability in the means across judges simply due to sample error alone, even if there is no effect associated with the judge. It would be a highly improbable outcome for the means across the judges to be exactly equal (something would be amiss). But how much variability would we expect across the judges just by chance? The statistical expectation is that the means across judges will vary by the same amount as the error variance across individual defendants, assuming no judge effect. Thus, the  $MS_{\text{between}}$  includes variance associated with judges and sampling error across individual defendants. Stated mathematically, the expected value for the  $MS_{\text{between}}$  in a one-way ANOVA is

$$MS_{\text{between}} = n\sigma_u^2 + \sigma_\epsilon^2$$

where  $n$  is the sample size within each group and  $\sigma_\epsilon^2$  equals the  $MS_{\text{within}}$ . Solving for  $\sigma_u^2$  as a function of the  $MS_{\text{between}}$ , the  $MS_{\text{within}}$ , and the sample size within produces the following equation:

$$\sigma_u^2 = \frac{MS_{\text{between}} - MS_{\text{within}}}{n} \quad \text{Equation 7.6}$$

When the design has unequal sample sizes across groups or clusters, the harmonic mean is used. In the case of our bail study, the harmonic mean of the number of defendants within judges is 117.41. If we work this out for our bail example, we estimate  $\sigma_u^2$  to equal 0.03. Without context, this number is difficult to interpret. The intraclass correlation coefficient, however, converts this into a useful metric (discussed in next section).

### Working It Out

$$\begin{aligned}\sigma_u^2 &= \frac{MS_{\text{between}} - MS_{\text{within}}}{n} \\ &= \frac{3.9096 - 0.27}{117.41} \\ &= 0.03\end{aligned}$$

### Intraclass Correlation and Explained Variance

Treating the level-2 clusters as random rather than fixed shifts the focus from estimating regression coefficients between the clusters (i.e., the mean differences between clusters) to estimating the variability attributable to

clusters. In a basic two-level model, we have two random variance components, the level-1 errors ( $\sigma_e^2$ ), and the level-2 errors ( $\sigma_u^2$ ). The total of these two is the variance not accounted for by the fixed part of our equation, that is, by the fixed regression coefficients (i.e., other independent variables). Based on this, we can compute a measure of the proportion of error variance explained by our level-2 random effect:

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} \quad \text{Equation 7.7}$$

where  $\rho$  has values ranging from 0 to 1 and measures the proportion of residual variation (variation not explained by independent variables) in the dependent variable that is due to the clusters. At  $\rho = 0$ , clusters explain none of the variation in the dependent variable not explained by the independent variables, while at  $\rho = 1$ , the group explains all of the variation in the dependent variable not explained by the independent variables.

An alternative interpretation of  $\rho$  is as the **intraclass correlation**, which indicates the level of absolute agreement of values within each group. By absolute agreement, we are trying to assess the extent to which the values within a group are identical. The intraclass correlation provides a measure that can be viewed in two different ways. In part, it provides a measure of intergroup heterogeneity by measuring how much of the total variation is due to the group. At the same time, it provides a measure of within-group homogeneity by measuring how similar the values are within each group.

### Deciding Between and Fixed- and Random-Effects Model

How do we know when to choose a fixed- or random-effects model? First off, if you have nested or clustered data, then you must model the resulting dependencies using an appropriate method. Simply ignoring dependencies in your data, whether naturally occurring or a function of the research design or sampling method, will produce biased results. Second, choosing between fixed-effect adjustment for clusters (Eq. 7.3) or random-effects adjustment for clusters (Eq. 7.4) are both defensible choices although the latter is often conceptually more appropriate.

From a practical standpoint, there are no firm rules about the sample sizes needed to estimate models with fixed and random effects. The total sample size is used to estimate the fixed effects and much like estimating any linear regression model, relatively modest sample sizes (100–200 cases) are often adequate. That same guideline holds for multilevel models. Since the random effects are estimated at the level of the cluster, it is unclear just how many clusters are necessary to estimate a multilevel model, although

10–20 clusters is a recommended lower bound by some authors (Rabe-Hesketh and Skrondal 2012). However, this depends entirely on the purpose of the analysis. If you are modeling level-2 effects, then clearly you need a sufficient number of clusters for the level-2 model. In contrast, if you are simply adjusting for the cluster effect, then small numbers of clusters (even less than 5) will produce valid estimates for the level-1 model. According to Gelman, such a model converges on the classic fixed-effects model shown in Equation 7.3 (see page 247 of Gelman and Hill 2007). It should be noted that the estimate of the level-2 variance component becomes unstable when the number of clusters becomes small. In summary, if the goal is to have standard errors for the fixed regression coefficients that are robust to the cluster effect, then a small number of clusters is acceptable. However, if the goal is to model between cluster differences, then a larger number of clusters is needed.

### Statistical Significance

A natural question that arises in the application of random-effects models is whether the random effect, the estimate of variance  $\sigma_u^2$ , is statistically significant. Substantively, this is a question about whether allowing for random variation around the overall mean adds anything statistically to the model over and above a fixed-effects model. In the simple model above, it is the test of whether judges explain some of the variability in bail amounts across individual defendants.

To test the statistical significance of  $\sigma_u^2$ , we rely on a likelihood ratio test, similar to that used in previous chapters. To compute the likelihood ratio (LR) test for a variance component (or multiple components as we will see later in the chapter), we need two values of the log-likelihood: (1) log-likelihood for the null model that excludes the random effect and (2) log-likelihood for the model that includes the random effects (note: All models have the level-1 random effect associate with the lowest level of measurement). The LR test is computed as:

$$\chi^2 = -2(LL_1 - LL_2) \quad \text{Equation 7.8}$$

The likelihood ratio test statistic has a  $\chi^2$  sampling distribution with 1 degree of freedom. We then divide the observed level of statistical significance for the computed  $\chi^2$  by 2, since it is a test of variances, which can only take on positive values and effectively truncating the sampling distribution to positive values.

### Bail Decision-Making Study

We return to our example from the Bail Decision-Making Study and present the results for a variance components model in Table 7.4. The model

**Table 7.4**

Variance components results for logged bail amount

VARIABLE/EFFECT	<i>b</i>	<i>SE</i>	<i>z</i>	$\sigma^2$
Fixed effect:				
Intercept	3.17	0.04	77.73	
Random effects:				
Intercept: judges				0.031
Error: defendants				0.270
$\rho = 0.104$				

intercept is estimated to be 3.17, which is also the weighted overall mean for logged bail amount. The variance of the groups ( $\sigma_u^2$ ) is estimated to be 0.031, indicating the degree to which the group (i.e., judge) means vary around the full sample mean. The unexplained error variance ( $\sigma_e^2$ ) is quite a bit larger and is estimated to be 0.27.

To what extent does the judge making the decision about bail affect the required amount? The intraclass correlation provides an indicator of the influence of the judge and is estimated to be 0.10 for logged bail. The intraclass correlation can also be obtained from the two variance components estimates:

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} = \frac{0.031}{0.031 + 0.270} = 0.10$$

The value of the intraclass correlation suggests that the decision-making judge only accounts for about 10% of the variation in the logged bail amounts. The log-likelihood for the null model (in this case Eq. 7.1) is  $-1871.73$ , whereas the log-likelihood for the random-effects model (in this case Eq. 7.3) is  $-1777.35$ . Using these values, we can determine the LR  $\chi^2$  as:

$$\chi^2 = -2[(-1871.73) - (-1777.35)] = 188.76$$

Based on 1 degree of freedom, we find the critical  $\chi^2$ , assuming a *p*-value  $< 0.05$ , to be 3.841. Our computed  $\chi^2$  of 188.77 has a *p*-value much less than 0.05, meaning the variance components model represents a significant improvement over the null model. Substantively, these results indicate that the decision-making judge is important to understanding bail amount requested. This can be generalized to the population of judges from which these 20 judges are a representative sample.

## Random Intercept Model with Fixed Slopes

---

We can extend the basic variance components model to include independent variables. That is, we can incorporate fixed slopes into the model. As with the variance components model above, this is still a **random intercept model** because the intercepts for the clusters are random, but the slopes are fixed. In the next section we will incorporate random slopes, that is, allow the slopes to vary across clusters. Equation (7.3) is expanded as shown below to includes  $k$  independent variables

$$y_{ij} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + u_j + \epsilon_{ij} \quad \text{Equation 7.9}$$

where the terms are defined as above with the addition of  $\beta_1$  to  $\beta_k$  and  $x_1$  to  $x_k$ . As with OLS regression, these are the regression coefficients and independent variables, respectively. The  $u_j$  is the random effect of the cluster on the model intercept, as it was with Equation 7.3, and the  $\epsilon_{ij}$  represents the random error term for each individual or level-1 observation if other than individuals (e.g., repeated measurements within individuals).

In the random intercept model, the regression coefficients are interpreted in the same way as discussed previously—a one-unit change in the independent variable is expected to result in a change in the dependent variable equal to  $b_k$ . The only meaningful change between this model and a basic OLS model without the  $u_j$  is that this model accounts for the statistical dependencies associated with the cluster variable, whether that is observations within individuals or individuals within some naturally (or experimentally) occurring clusters. The standard errors and associated statistics, such as  $t$ -test and confidence intervals, will be more accurate, assuming that the model correctly identifies the source of any nonindependence in the data.

### Statistical Significance

The results from a two-level random intercept model will lead to testing the statistical significance of both the effects of the independent variables included in the model and the significance of the level-2 random-effect variance component. In regard to testing for the statistical significance of the effects of the independent variables (e.g.,  $b_1$  and  $b_2$ , in the equation above), we would use the following familiar equation:

$$z = \frac{b}{se_b} \quad \text{Equation 7.10}$$

where  $se_b$  is the standard error of the coefficient and  $z$  (the test statistic) is assumed to have a normal distribution. Some software reports this as  $t$ . The computation is the same with the difference being that the  $p$ -value is determined from the  $t$ -distribution rather than the  $z$ -distribution. The maximum likelihood estimation procedure for the random intercept model estimates standard errors that are adjusted for the clustered nature of the data. Depending on the particular data, the standard errors for the coefficients in a random intercept model will always be at least as large, but more likely larger, than those estimated in an OLS linear regression model with the same variables, since the clustered nature of the data has been taken into account during the estimation process.

The test of the random intercept (e.g., the level-2 random effects) involves the use of likelihood ratio (LR) test discussed previously as Eq. (7.8), the difference being that the first log-likelihood in the equation is for a standard OLS regression model that excludes the level-2 random intercept. The second log-likelihood is the full model shown in Eq. (7.10). The goal of this test is to assess whether allowing the model intercept to vary randomly across clusters improves the fit of the statistical model.

As before, the likelihood ratio test statistic has a  $\chi^2$  sampling distribution with 1 degree of freedom for the one random-effects variance estimate. Substantively, this test will indicate whether the addition of a random intercept to a linear regression model makes a statistically significant contribution. Some software programs report the test of the significance of a single random-effects variance component as a  $z$ . This  $z$  value is simply the square root of the  $\chi^2$ . Tests of the joint significance of two or more random effects will also be reported as a  $\chi^2$ .

It may be tempting to simplify the model to a nonmultilevel OLS regression model if the level-2 random-effects variance component is not statistically significant. This is not recommended and amounts to accepting the null that the true value for this variance equals zero. The decision to fit a random intercept multilevel model should be based on a theoretical understanding of the data and a belief that there may be a statistical dependency in the data associated with some form of known clustering. If the model produced a nonzero value for the random intercept, then some dependence was identified in the data, even if the amount of dependence was not statistically significant given the sample size. The only time where it makes sense to drop a random intercept is when the term equals zero.

### Centering Independent Variables

There are many instances where the interpretation of the model intercept is important to understanding the implications of the estimated results. When we are confronted with an independent variable that has no meaningful zero point, the meaning of the intercept is difficult to explain. A straightforward way of dealing with this issue is to center the independent

variables. What do we mean by center a variable? Centering refers to subtracting each value of a variable from its mean.

The two types of centering of independent variable that are often most useful for estimating multilevel models are (1) grand or overall mean centering and (2) cluster or group-based centering. In grand mean centering, we subtract the overall sample mean of the independent variable from each observation:

$$\text{Grand mean centering: } x_{ij} = x_{ij} - \bar{x}_{..}$$

Note that the two periods in the subscript of  $\bar{x}$  indicate that this is the mean across individuals ( $i$ ) and clusters ( $j$ ).

In cluster-based centering, we subtract the relevant cluster (group) mean from each observation:

$$\text{Cluster mean centering: } x_{ij} = x_{ij} - \bar{x}_j$$

Note that there is a period in place of the  $i$  subscript to indicate that this is the mean across individuals ( $i$ ) but not clusters, that is, there is a separate mean for each cluster ( $j$ ).

How does the inclusion of a centered variable, instead of the original independent variable, affect the interpretation of the results? The interpretation of a regression coefficient for a grand mean centered independent variable is exactly the same as with an uncentered independent variable: A one-unit change in  $x$  is expected to result in a change of  $b$  units in the dependent variable. Keep in mind that all we have done by grand mean centering a variable is shifted its location; its scale is unchanged and so the slope is unchanged.

The difference between a model with uncentered and grand mean centered variables is entirely in the intercepts. With uncentered data, the overall intercept and the cluster intercepts are the mean value of  $y$  when all of the independent variables equal 0. As discussed in Chap. 3, these intercepts are often outside the scope of the data and these values are of little interest, although for the clusters, the differences are meaningful. With grand mean centering, the intercepts are now the means. The overall intercept is the overall mean, and the cluster intercepts are the cluster means, adjusted for the independent variables. The model is otherwise unaffected.

Cluster or group mean centering is a bit more complicated in terms of its effect on the model parameters. Furthermore, you would want to include the cluster mean into the model as a level-2 predictor. By doing so, you can test the effect at level-1 of being above or below the cluster mean on the

dependent variable (the within-cluster effect of the independent variable) and at level-2 the effect of the cluster mean on the independent variable on the cluster's mean on the dependent variable (the between cluster effect of the independent variable).

The other implication for interpreting the results is focused on the variance component  $\sigma_u^2$ . When grand mean centering has been used,  $\sigma_u^2$  represents variation in the group means around the overall mean, identical to the case where no centering has been used. With cluster-based centering,  $\sigma_u^2$  is interpreted as the variation of group means around the weighted average of cluster means.

When should centering be used? In general, centering an independent variable aids in the estimation of multilevel models, particularly some of the more complex models that involve estimating interaction effects across levels of data (which we discuss below). Grand mean centering should not change the substance of the statistical results, since all that centering accomplishes is a shifting of the independent variable so that a value of 0 represents either the overall mean for the sample or the cluster mean for each group.

Dummy variables may also be centered. In this case, the overall or group means simply represent the proportion of cases in the full sample or the cluster that has the characteristic measured by the dummy variable.

### Bail Decision-Making Study

To illustrate the application of random intercept models and the use of centering, we return to the Bail Decision-Making Study. In the interest of keeping the statistical model simple, we include only one independent variable predicting bail amount. Clearly, there are numerous characteristics of defendants and their cases that affect the bail decision. Our goal here is to illustrate how one would go about estimating and interpreting a random intercept model with and without a centered independent variable. We will develop a more complex model in a later section.

In making assessments about bail, the judge is expected to consider both the chances of the defendant fleeing the community and the potential threat to public safety. One indicator of a defendant's overall risk is the number of prior drug offenses in the person's criminal history record. In general, the greater the evidence of prior drug offending, the higher the perceived risk of some kind of pretrial misconduct and consequently higher bail amounts being requested from defendants.

Table 7.5 presents results from four different models. The first column is the results of an OLS regression model, that is, a model without clustering. The next three are multilevel random intercept models with either no centering, grand mean centering, or cluster mean centering of the independent variable (number of prior drug arrests). We see from the results in

**Table 7.5**

Regression results for logged bail amount on number of prior drug offenses

VARIABLE/EFFECT	MULTILEVEL RANDOM INTERCEPT MODEL							
	OLS		NO CENTERING		GRAND MEAN CENTERING		CLUSTER MEAN CENTERING	
	b	SE	b	SE	b	SE	b	SE
Fixed effects:								
Intercept	2.69	0.019	2.71	0.040	3.17	0.037	3.17	0.041
Number of prior drug offenses	0.12	0.004	0.11	0.004	0.11	0.004	0.11	0.004
Random effects:								
Intercept: judge			$\sigma^2$		$\sigma^2$		$\sigma^2$	
Residual			0.026		0.026		0.031	
			0.192		0.192		0.192	

Column 1 that a one-unit increase in the number of prior drug offenses increases logged bail by 0.12 units. Since talking about changes in logged units likely makes little intuitive sense, a more meaningful interpretation of this coefficient is in terms of percentage change in the logged outcome variable. Our observed coefficient of 0.12 for number of prior drug offenses can alternatively be interpreted as expecting bail amount to increase by 12% for each additional prior drug offense.

Notice that the intercept remains nearly the same for the no centering model as in the OLS model (2.69 vs. 2.72, respectively) as does the effect of number of prior drug offenses (0.12 vs. 0.11, respectively). For the two models with centering, the estimates for the intercept and the effect of number of prior drug offenses are the same through two significant digits (but not beyond). Note that the estimate of the intercept is the overall mean for logged bail amount reported earlier, which is expected under the use of grand mean centering. The use of cluster mean centering estimates a model intercept that is the weighted average of the group means. The reason the estimates for the intercept are so similar in the two models using different types of centering is an artifact of the study design that used a balanced approach to select the same number of cases for each judge. The small variation in the number of cases per judge used in these analyses accounts for the minor differences that do appear in the intercept and the effect of number of prior drug offenses in decimals beyond two places.

In regard to the effect of number of prior drug offenses, please note that the effect is the same for all three multilevel models, regardless of whether no centering, grand mean centering, or cluster mean centering was used. This is to be expected, as centering an independent variable only shifts the distribution of cases and does not alter anything else about the values, meaning that the coefficient representing the effect of the independent variable should stay the same.

An important difference in the cluster mean centered model from the other two multilevel models is in the estimate of the random-effects

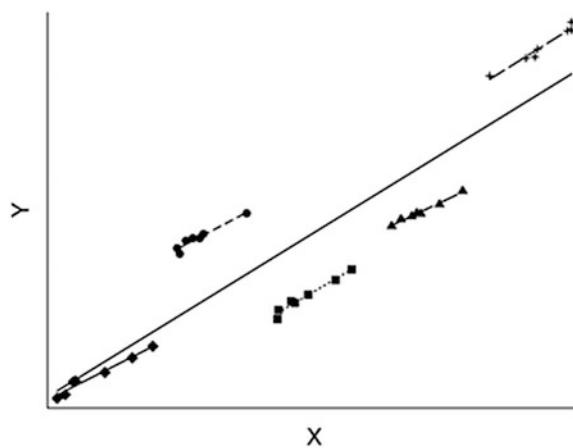
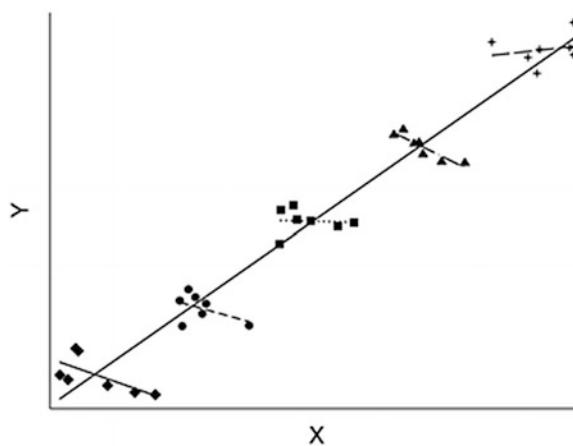
variance component for the level-2 intercepts (judges). Notice that this value is the same as for the model we started with that had no independent variable but is slightly less for the no centering and grand mean centering models. For the former, the random intercept for level-2 reflects the variability across the simple means for each cluster, whereas for the latter, it reflects the variability across the adjusted means for each cluster, adjusting for number of prior drug offenses.

### Between and Within Effects

In our discussion of centering, we noted that centering variables can assist in estimating multilevel models. What this means is that the algorithms used by various statistical packages to estimate multilevel models perform better when using centered independent variables. Although the explanation for how these algorithms work goes beyond the focus of our text, we note that statistical packages that estimate multilevel models often require multiple iterations to come to a solution—the estimates of the intercept and the other coefficients.

We are still left with the question, then, of which method of centering to use. How do we make this determination? One of the issues that naturally arises in the study of clustered or multilevel data is whether the effects of the independent variables are the same across groups as they are within groups. For example, in the analysis of judicial bail decision making, we might wonder whether the effect of number of prior drug offenses across judge—what we will call a **between effect**—is the same as the effect of number of drug offenses processed by each judge—what we will call a **within effect**. Conceptually, what we are attempting to get at is whether a regression model for each judge (the within regression for each judge) is parallel to a single regression line based on the means for each judge (the between regression for all judges). To the extent the slopes (coefficients) are parallel, there are similar between and within effects, meaning that each judge uses information on number of prior drug offenses in approximately the same way. To the extent the slopes differ, there is a difference in the between and within effects, indicating that judges weight information about number of prior drug offenses differently.

Figures 7.1 and 7.2 present a way of thinking hypothetically about similar and different effects. In each figure, the solid line represents the overall regression slope for the effect of  $x_1$  on  $y$ . The dashed lines represent the regression lines within each of the five clusters plotted. In Fig. 7.1, the between and within effects are parallel to each other. The different placement of the dashed lines represents the random effect of each cluster, where two clusters have positive random effects (and appear above the solid line), and three clusters have negative random effects (and appear below the solid line). In Fig. 7.2, the between and within effects are different—one slope is positive, one slope is essentially flat, while the

**Figure 7.1***Parallel between and within effects***Figure 7.2***Different between and within effects*

remaining three slopes are negative. Although these figures are informative in highlighting similarities and differences in the between and within effects, it is impractical to plot out regression lines for many groups and instead you should generally rely on a statistical test for differences in the between and within effects.

### Testing for Between and Within Effects

The most direct way of testing for a difference in the between and the within effects is with the addition of the cluster mean for a cluster mean centered independent variable. This is shown in the equation below for a model with a single independent variable:

$$y_{ij} = \beta_0 + \beta_1(x_{1j} - \bar{x}_{1j}) + \beta_2\bar{x}_{1j} + u_j + \epsilon_{ij} \quad \text{Equation 7.11}$$

where  $\bar{x}_{1j}$  is the cluster mean. Notice that we now have two regression coefficients,  $b_1$  and  $b_2$ . The first reflects the level-1, or within effect of  $x_1$  on the dependent variable. The second reflects the level-2, or between effect of  $x_1$  on the dependent variable. In the context of our bail data using the number of prior drug offenses, independent variable  $b_1$  indicates the effect within judges of the number of prior drug offenses; that is,  $b_1$  indicates the change in logged bail amount for a one-unit change in the prior drug offense once you remove judge effect. The  $b_2$  indicates whether the mean number of prior drug offenses within the sample of defendants seen by a judge effects the logged bail amount. That is, do judges who see defendants with higher or lower prior drug offense histories, on average, hand down larger or smaller average logged bail amounts.

### Bail Decision-Making Study

In the Bail Decision-Making Study, one of the key areas of attention was a question about whether judges weighted information about defendants in similar or different ways. A direct test of this is provided by a test for similarity of between and within effects. If we continue the example started previously using number of drug offenses as the independent variable, we estimate the model shown in Eq. (7.11). The results of this model are shown in Table 7.6.

As expected, given the previous set of results, the between effect of number of drug offenses is greater than the within effect of number of drug offenses, again confirming that these 20 judges differentially used information about drug offending when making bail decisions. It is worth noting that the effect of cluster mean centering is the same as the raw score

**Table 7.6**

Test of between and within effects similarity

VARIABLE	<i>b</i>	<i>SE</i>	<i>z</i>
Intercept ( $b_0$ )	1.29	0.523	2.47
Number of Drug Offenses (cluster mean centered) ( $b_1$ )	0.11	0.004	29.80
Cluster Mean of Drug Offenses ( $b_2$ )	0.46	0.128	3.61

estimates presented in Tables 7.5. This is to be expected, since this is just the effect for the cluster mean centered number of drug offenses.

## Random Coefficient Model

---

A straightforward extension of the random intercept model involves thinking about the effects of one or more of the independent variables in a multilevel model also varying across clusters. Put another way, we may have justification, based on prior research and theory, to expect the slope coefficients for a key variable to vary across clusters. For example, in a study of fear of crime across neighborhoods, we might expect the effect of gender to vary by neighborhood. Similarly, in our study of judicial decision making, we might expect judges to weight information about cases and defendants differently, suggesting that we will find different slopes for key predictors of the outcome variable.

The development of the **random coefficient model** begins with the random intercept model. Equation (7.12) below is simply Eq. (7.3) with a single level-1 independent variable,  $x_1$ ,

$$y_{ij} = \beta_0 + \beta_1 x_1 + u_{0j} + \epsilon_{ij} \quad \text{Equation 7.12}$$

where all terms are defined as above except that there is now a 0 included in the subscript of the random effect  $u_j$  to indicate the connection to the intercept ( $\beta_0$ ). For a random coefficient model, we add a random effect for the slope coefficient in question. In our example, to estimate a model in which  $\beta_1$  is allowed to vary across cluster, we would add an additional random effect,  $u_{1j}$ , as shown here:

$$y_{ij} = \beta_0 + \beta_1 x_1 + u_{0j} + u_{1j} + \epsilon_{ij} \quad \text{Equation 7.13}$$

We can rewrite this equation as separate level-1 and level-2 models as we did with Eq. (7.3). These are shown as Eqs. (7.14–7.16).

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + \epsilon_{ij} \quad \text{Equation 7.14}$$

$$\beta_{0j} = \beta_0 + u_{0j} \quad \text{Equation 7.15}$$

$$\beta_{1j} = \beta_1 + u_{1j}$$

**Equation 7.16**

Equation (7.14) is the level-1 model and essentially estimates a separate regression equation for each cluster ( $j$ ). We now have two level-2 models. Equation (7.15) is the level-2 model for the random intercepts, and Eq. (7.16) is the level-2 model for the random slopes. Conceptually, what the random coefficient model does is analogous to estimating a regression model for each cluster and then examining whether the intercepts and slope coefficients vary in any meaningful way across the clusters. We now turn to a more formal examination of variance estimates from the random coefficient model.

### Variance Estimates

Similar to the variance components and random intercept models, the level-1 error variance continues to be represented by  $\sigma_e^2$ . The variance of the random effects ( $u_{0j}$  and  $u_{1j}$ ) now take on multiple values:

Variance of the intercept across cluster:  $\sigma_{u_0}^2$

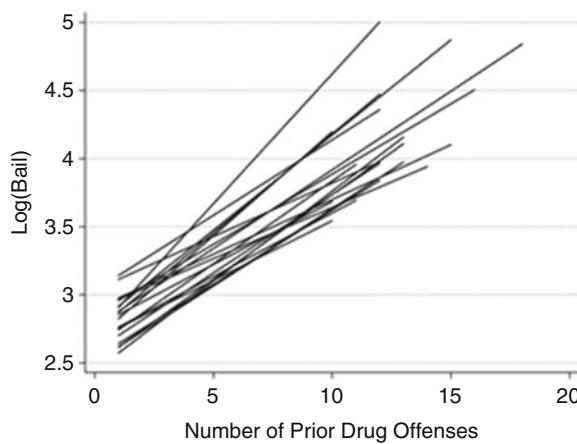
Variance of the slope coefficient across cluster:  $\sigma_{u_1}^2$

We are now confronted with another choice regarding the estimation of the variance components. In addition to the two variance estimates for the level-2 random effects, we may estimate the covariance of the random effects for the intercept and the slope, which we can label either  $\sigma_{u_{01}}^2$  or  $\sigma_{u_{10}}^2$ . What does this estimate of the covariance of  $\sigma_{u_0}^2$  and  $\sigma_{u_1}^2$  assess? In general, it will indicate whether the magnitude of the random effect for the intercept covaries with the magnitude of the random effect for the slope coefficient. More directly, a positive value of the covariance indicates that clusters with larger intercepts will tend to have larger values for the slope coefficient. Conversely, a negative covariance would suggest that smaller values of the intercept are associated with larger values of the slope coefficient.

It is important to note that we cannot make direct comparisons of the variance and covariance components, since each estimate reflects the different metrics of the variables being analyzed. Simply by changing the scale of one or another variable, we could alter the variance or covariance estimate. For example, by expanding the scale of a variable, say from (0,1) to (0,10), it would inflate the values of each variance estimate without changing the substantive interpretation of the results.

### Bail Decision-Making Study

To gain an appreciation of the random coefficient model, we begin our analysis of the bail decision-making data by estimating 20 separate

**Figure 7.3***Judge-specific regression lines*

regression equations, one for each of the 20 judges included in the sample. We continue to rely on the same simple model of logged bail as the dependent variable and number of prior drug offenses as the only independent variable. The regression equation for these models is shown here<sup>5</sup>:

$$\begin{aligned}y_{i,1} &= b_{0,1} + b_{1,1}x_{1,1} + e_{i,1} \\y_{i,2} &= b_{0,2} + b_{1,2}x_{1,2} + e_{i,2} \\y_{i,3} &= b_{0,3} + b_{1,3}x_{1,3} + e_{i,3} \\\dots \\y_{i,20} &= b_{0,20} + b_{1,20}x_{1,20} + e_{i,20}\end{aligned}$$

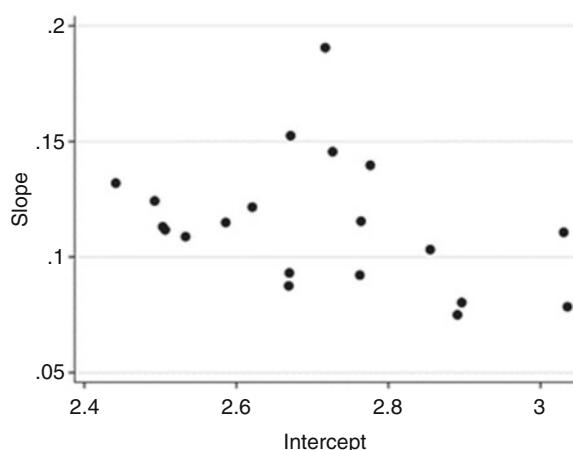
where  $y_{ij}$  is the logged bail amount for each defendant ( $i$ ) within judge ( $j$ ) and  $x_{ij}$  is the number of prior drug offenses for each defendant. Note from the subscripts to the intercept ( $b_{0,j}$ ) and the slope coefficient ( $b_{1,j}$ ) that there will be unique estimates for each value across the judges/clusters ( $j$ ). Rather than present the results from 20 regression analyses in a table, the results are presented graphically in Fig. 7.3, where each line represents the regression line (the  $b_1$ s) for a single judge. Clearly, the results show variation in the intercept across judge—note the vertical placement of each regression line that reflects larger or smaller values of the judge-specific

---

<sup>5</sup>Note that we are now expressing this in terms of the sample model, replacing the population parameters  $\beta$  and  $\epsilon$  with the sample values of  $b$  and  $e$ .

**Figure 7.4**

*Scatterplot of judge-specific intercepts and slope coefficients*



intercept ( $b_0$ s). We also note that there is variation in the regression slopes across the 20 judges—some of the slopes are steeper; some are flatter.

One way of starting to assess how similar or different the intercepts and slope coefficients are for each judge can be viewed in a scatterplot of the intercepts and the slopes. Figure 7.4 presents these results, where each point on the graph represents one of the 20 judges included in the analysis. The spread of cases across both axes confirms what we viewed in Fig. 7.3—there is variation in the intercept and the slope coefficient across judges. Figure 7.4 shows a pattern consistent with a negative association: Larger values of the intercept tend to have smaller slope coefficients, while smaller values of the intercept tend to have larger slope coefficients. The Pearson correlation for these values is  $-0.34$ . Substantively, this pattern suggests that judges with large intercepts—their cases receive relatively larger bail amounts on average—tend to place less weight on the defendant’s number of prior drug offenses. In contrast, for judges with smaller intercepts—their cases tend to receive relatively lower bail amounts on average—place greater weight on the defendant’s number of prior drug offenses.

The use of the random coefficient model offers a more efficient way of assessing the similarities and differences across the judges, but most importantly will allow us to determine whether the variations in the traditional regression model are statistically meaningful or reflect random variation in the values for each judge. Table 7.7 presents the results for the random coefficient model:

**Table 7.7**

Regression results for logged bail amount on number of prior drug offenses comparing a random intercept model to a random intercept and slope model

VARIABLE/EFFECT	RANDOM INTERCEPT <sup>a</sup>		RANDOM INTERCEPT AND SLOPE <sup>b</sup>	
	<i>b</i>	SE	<i>b</i>	SE
Fixed effects:				
Intercept	2.71	0.040	2.71	0.037
Number of prior drug offenses	0.11	0.004	0.11	0.006
Random effects:				
Intercept: judge	$\sigma^2$		$\sigma^2$	
Slope: number of prior drug offenses	0.026		0.0209	
Covariance			0.0005	
			-0.0004	

<sup>a</sup>Log-likelihood = -1400.06

<sup>b</sup>Log-likelihood = -1391.74

$$y_{ij} = b_0 + b_1 x_1 + u_{0j} + u_{1j} + e_{ij}$$

where the terms are defined as above. Recall that this model combines the 20 separate regression models from above with the 20 intercepts and 20 slopes becoming the random effects of  $u_{0j}$  and  $u_{1j}$ .

Table 7.7 provides the results for this model with the random intercept and slope to the model with only a random intercept. Note that the estimates of the overall intercept and slope coefficients are the same between these two models. Hopefully, this is intuitive as the intercept reflects the overall weighted mean and the slope is similarly the overall effect of the number of prior drug offenses. What differs in Table 7.7 is the inclusion of the random effects for the slope and the covariance of the intercept and slope random effects. As before, we can test whether the addition of the random slope represents an improvement in the statistical model.

Recall from above that the test of statistical significance for the addition of a random effect is a  $\chi^2$  test that compares the log-likelihood values from two different models. In the present case, we are comparing the log-likelihood values from the model with only the random intercept to the model with both the random intercept and random slope. This is shown below.

$$\chi^2 = -2(LL_1 - LL_2) = -2[(-1400.06) - (-1391.74)] = 16.64$$

Based on 1 degree of freedom, we find the critical  $\chi^2$ , assuming a *p*-value < 0.05, to be 3.841. Our computed  $\chi^2$  of 16.64 has a *p*-value much less than 0.05. This result indicates that the multilevel model with a random effect for

the slope coefficient (number of prior drug offenses) represents a significant improvement in the statistical model. The effect of number of prior drug offenses on logged bail amount differs across judges.

## Adding Cluster (Level 2) Characteristics

---

Thus far in our discussion of multilevel models, we have focused strictly on characteristics of the level-1 observations. In our bail example, these are characteristics of the defendants. One of the great strengths of multilevel modeling is the ability to include cluster-level characteristics that will indicate how the effects of the independent variables may vary across levels of a cluster characteristic. In the example we have used thus far regarding judges and bail decision making, we might hypothesize that characteristics of judges would affect how each would weigh information about defendants in making bail decisions. For example, gender of judge may alter the relationship that we have observed between number of prior drug offenses and bail amount. Or, years of service as a judge may affect the observed relationship between number of prior drug offenses and bail amount. These are the kinds of questions to which we now turn.

When considering adding cluster-level characteristics to a multilevel analysis, the researcher is confronted with two important questions about a cluster characteristic: (1) Is there an expectation that the cluster characteristic will directly affect the dependent variable? (2) Is there an expectation that the effect of an independent variable will vary by the level of the cluster characteristic?

Both of these questions force us to consider prior theory and research in thoughtfully developing our multilevel model. The first question is the more straightforward of the two questions and will often reflect prior research showing that the cluster characteristic is likely important to the dependent variable being analyzed. For example, there is research indicating that gender of judge affects how criminal defendants and offenders are treated. We would have justification for hypothesizing that gender of judge would affect bail amount. Similarly, if we were studying fear of crime across a large sample of different neighborhoods, we would have justification for hypothesizing that official crime rates in the neighborhoods may have a direct effect on an individual's fear of crime. In terms of the model, this question is about predicting variability in the random intercepts across clusters.

The second question requires considerable care in developing the model, especially since there is likely to be less evidence and/or theory on which to base a hypothesis of an independent variable's (level 1) effect varying by the level of the cluster characteristic. In the literature on

multilevel models, this kind of relationship is often referred to as a **cross-level interaction**, since they imply the effect of one variable (the level-1 independent variable) changes across the levels of another variable (the level-2 cluster characteristic). For example, in considering fear of crime, we may hypothesize that official crime rates may interact with the effect of age of a resident on fear of crime. Older individuals are more fearful of being crime victims in general, and we could hypothesize that as the neighborhood crime rate increased, there was a multiplier effect on the fear of crime among elderly residents. At the same time, neighborhood crime rates may not affect the level of fear of younger individuals. In terms of the model, this question is about predicting variability in the random slopes across clusters.

Although the interpretation of cluster-level characteristics in a multilevel model may become complicated, their inclusion in the statistical model is not complicated. For the situation where we expect the cluster characteristic to have a direct effect on the dependent variable, we simply include it as an additional independent variable in our random intercept or random coefficient model. To keep straight which independent variables were measured at level-1 versus level-2, we are going to use  $w$  as the notation for a level-2 independent variable instead of  $x$ . It is simplest to understand how we are changing the model when the two levels are expressed separately, as we have done previous, such as with Eqs. (7.14) through (7.16). In the form of a simple random coefficient model, we would estimate the following model:

$$\text{Level-1 model: } y_{ij} = b_{0j} + b_{1j}x_{1ij} + e_{ij}$$

$$\text{Level-2 intercept model: } b_{0j} = b_0 + b_2w_2 + u_{0j}$$

$$\text{Level-2 slope}(b_1) \text{ model: } b_{1j} = b_1 + u_{1j}$$

We can combine these into a single model shown below:

$$y_{ij} = b_0 + b_1x_1 + b_2w_2 + u_{0j} + u_{1j} + e_{ij}$$

Like any other variable we might include in a linear regression model, the cluster characteristic can be a dummy variable or a scaled variable. The interpretation of the cluster characteristic's effect ( $b_2$ ) is no different than that for other independent variables included in the model: A unit change in  $w_2$  is expected to change the value of the dependent variable by  $b_2$ .

If we expect the effect of one of our level-1 independent variables to vary by level of a cluster characteristic, we include an interaction term

between the two variables—our cross-level interaction—and add it to the model. Starting with the separate models, we would add a predictor to the level-2 slope model; that is, we are predicting variability in a particular slope across clusters. Thus, the level-2 slope model becomes:

$$b_{1j} = b_1 + b_3 w_3 + u_{1j}$$

When we combine these into a single regression model, we get the following:

$$y_{ij} = b_0 + b_1 x_1 + b_2 w_2 + b_3 x_1 w_2 + u_{0j} + u_{1j} + e_{ij}$$

where we now have an interaction term ( $b_3$ ) reflecting the interaction between the level-1  $x_1$  and level-2  $w_2$  independent variables. Note that with the exception of the level-2 random effects ( $u_{0j}$  and  $u_{1j}$ ), this model is equivalent to an interaction model between two independent variables.

How do you interpret these results and make sense of the interaction effect? In Chap. 3, we noted that interaction effects can usually be interpreted in two different ways: We fix the value of one variable and assess the effect of the second variable. In a multilevel model, we will always fix the level (value) of the cluster characteristic first and then interpret the effect of the level-1 independent variable. In the equation above, the effect of  $x_1$  can be written as:

$$b_1 x_1 + b_3 x_1 w_2$$

If  $w_2$  is a dummy variable, then for  $w_2 = 0$  we have as follows:

$$b_1 x_1 + b_3 x_1 0 = b_1 x_1$$

meaning the effect of  $x_1$  when  $w_2 = 0$  is simply  $b_1$ . In contrast, if  $w_2 = 1$ , we have as follows:

$$b_1 x_1 + b_3 x_1 1 = b_1 x_1 + b_3 x_1 = (b_1 + b_3) x_1$$

meaning the effect of  $x_1$  when  $w_2 = 0$  is  $b_1 + b_3$ . For cluster-level characteristics measured on an interval scale of measurement, we would typically pick out some meaningful values and highlight the effect of  $x_1$  at those values.

### A Substantive Example: Race and Sentencing Across Pennsylvania Counties

In an analysis of sentencing decisions in Pennsylvania in the 1990s, Britt used a multilevel model to assess the effects of various social, economic, and crime measures on punishment severity decisions for offenders (Britt 2000). Of particular interest in Britt's analysis was the effect of these kinds of community characteristics on the effect of offender's race on punishment severity. For example, were black offenders punished more severely in those counties with higher rates of crime? Alternatively, were black offenders punished less severely in those counties with proportionally larger black populations? The theoretical rationale for these different hypotheses is presented in the original paper.

In what follows, we highlight a few of his key findings as examples of the power of a multilevel model. The data in the analysis represented more than 70,000 sentence length decisions spanning four years for all 67 of Pennsylvania's counties. In the context of a multilevel model, the sentence length decisions represent the individual-level data (i.e., level 1) and the counties in which the sentences were given represent the cluster-level data (level 2). We focus our discussion on the following variables:

#### Individual-level (Level 1)

- Sentence length: Months sentenced to jail or prison.
- Black: Coded as 1 if offender was black, 0 if offender was white (all other cases were excluded from analysis).

#### County-level (Level 2)

- Percentage of population classified as black.
- Percentage of population living in an urban area.
- Trend in unemployment rate (increasing, decreasing, or flat).
- Average violent crime rate.

Since our intention here is to illustrate the use and interpretation of multilevel models with cluster-level characteristics, we report abridged results in Table 7.8, showing only those elements focused on the effect of county characteristics on overall sentence length and county characteristics interacting with race of the offender. Omitted from the table are numerous case and offender characteristics relevant to predicting punishment severity, such as offense severity, criminal history, and plea bargaining.

To begin the interpretation of the results, note that percentage of the population classified as black, difference in white and black per capita income, percentage living in urban areas, and trend in unemployment all have direct effects on sentence length. The other results in Table 7.8 show the direct effect for being a black offender and the interaction terms for black offender with percentage of the population classified as black and the

**Table 7.8**

Multilevel regression results for sentence length

VARIABLE	<i>b</i>	SE	<i>z</i>
Intercept	13.968	0.574	24.334
Level-1 effect			
Black	-2.277	0.618	-3.684
Level-2 effects			
Percentage Black	-0.161	0.021	-7.750
Percentage urban	0.026	0.007	3.910
Trend in unemployment	0.820	0.228	3.596
Cross-level interactions			
Black × percentage Black	-0.315	0.092	-3.292
Black × violent crime rate	0.009	0.003	3.000

average violent crime rate. The interpretation of the results at this point is no different than the interpretations we made in the linear regression model. For the direct effects on sentence length:

- Black defendants on average get shorter sentence lengths.
- As the percentage of blacks in a county increases, the average sentence length decreases.
- As the percentage of a county's population living in an urban area increases, the average sentence length increases.
- As the county-level unemployment rate increased over time, the average sentence length increases.

For the cross-level interaction effects of offender race with percentage of black residents in a county and average crime rate, the interpretations follow the logic of any other interaction effect. In general, what we find is that as the percentage of a county's population classified as black increases, the *effect* of being black decreases, meaning that black offenders received significantly shorter sentences than white offenders overall, but the magnitude of this difference increases as the percentage of blacks in a county increases. Conversely, in counties where the average violent crime rate was higher, the *effect* of being black *increases*, meaning that the punishments received by black offenders were more severe than those for white offenders in counties with higher violent crime rates.

The following hypotheticals will help to illustrate these patterns. Suppose that we have four different counties with the following characteristics:

- County A: Percentage Black = 10, Violent Crime Rate = 100
- County B: Percentage Black = 20, Violent Crime Rate = 100
- County C: Percentage Black = 10, Violent Crime Rate = 200
- County D: Percentage Black = 20, Violent Crime Rate = 200

The equation for the effect of being a black offender on sentence length is:

$$\begin{aligned} -2.277Black - 0.315Black \times PercentBlack + 0.009Black \\ \times ViolentCrimeRate. \end{aligned}$$

The effect of being a black offender in County A:

$$\begin{aligned} -2.277Black - 0.315 \times 10 \times Black + 0.009 \times 100 \times Black \\ = -4.527Black. \end{aligned}$$

The effect of being a black offender in County B:

$$\begin{aligned} -2.277Black - 0.315 \times 20 \times Black + 0.009 \times 100 \times Black \\ = -7.677Black. \end{aligned}$$

The effect of being a black offender in County C:

$$\begin{aligned} -2.277Black - 0.315 \times 10 \times Black + 0.009 \times 200 \times Black \\ = -3.627Black. \end{aligned}$$

The effect of being a black offender in County D:

$$\begin{aligned} -2.277Black - 0.315 \times 20 \times Black + 0.009 \times 200 \times Black \\ = -6.777Black. \end{aligned}$$

In all cases, the effect of being a black offender resulted in a shorter sentence, ranging from about 3.6 months to just under 7.7 months.<sup>6</sup> For the two pairs of counties with matching percentage black populations, the

---

<sup>6</sup>While this finding often strikes many criminal justice students as counterintuitive, it is consistent with much of the research done on the race effects on sentencing outcomes. What is not shown here, but is included in Britt's article, is the effect of race on the likelihood of being incarcerated, where black offenders were much more likely than white offenders to be sentenced to prison. The findings here just highlight that, once sentenced to incarceration, the length of time is shorter for black offenders compared to white offenders. In Pennsylvania, this has typically taken the form of more black offenders being sentenced to local jails for relatively short periods of time, while white offenders who have received incarceration sentences will be sent to state prisons for relatively longer stays.

increase in the violent crime rate resulted in a shrinking of the effect of being black and moving the coefficient closer to 0, where there is no difference between black and white offenders. For the two pairs of counties with matching violent crime rates, as the percentage of the black population increased, the sentence disparity increased further, with black offenders receiving even more lenient sentences.

### Multilevel Negative Binomial Regression

In the previous chapter, we discussed count regression models. As we noted there, count regression models are common in criminology and criminal justice because we are often looking at counts of such outcomes as arrests or crimes. We provide here a substantive example of a multilevel negative binomial regression in which hot spot and non-hot spot street segments ( $n = 449$ ) were nested within communities ( $n = 55$ ) (Weisburd et al. 2020). The key question in the study was whether collective efficacy, a measure of the extent to which people on a street segment trusted each other and were willing to intervene in problems on the street (Sampson et al. 1997), was related to emergency calls to the police on these streets. But it could be argued that if such a relationship existed, it was in fact spurious, because the key variables in understanding crime calls on the street were structural indicators of collective efficacy at the community level.

The researchers had a unique dataset in which they measured directly collective efficacy on the streets in the study. They also had access to a series of structural indicators of collective efficacy in the community drawn from census information. At level-2 in the model (the Community Statistical Areas [CSAs]), they measured concentrated disadvantage, racial diversity, and the percentage aged 19–24-year old at the CSA level. They also wanted to ensure that at level 1 of the model (the street segments), they took into account a series of possible control variables that might confound their estimate of collective efficacy at level 1. The control variables are listed in Table 7.9 (from *age* through *social service building*). The nature of the variables is not important for our purposes here but can be examined in the original article. We present the Table from the article so that you can see a typical journal reporting table of a multilevel model.

The model estimated was a random intercept, fixed slope model with no cross-level interactions (i.e., random slopes). There were also no within-level interaction terms included in the model. Thus, the regression coefficients can be interpreted as the simple marginal effect of each variable on the dependent variable, adjusting for the other independent variables in the model and any dependencies of street segments within the same community.

**Table 7.9**

Reporting of multilevel regression results

	MODEL 1 (UNCONDITIONAL)		MODEL 2 (COLLECTIVE EFFICACY ONLY)			MODEL 3 (FULL MODEL)				
	b	SE	b	IRR	SE	p-Value	b	IRR	SE	p-Value
<i>Fixed Effects</i>										
Intercept	4.256***	0.070	6.927***	—	0.413	0.000	4.941***	—	0.413	0.000
Street-level variables										
Collective efficacy	—	—	-0.730***	0.482	0.109	0.000	-0.296**	0.744	0.098	0.003
Age	—	—	—	—	—	—	0.001	1.001	0.005	0.826
Female	—	—	—	—	—	—	0.001	1.001	0.002	0.653
Socioeconomic status	—	—	—	—	—	—	-0.225***	0.798	0.041	0.000
Youth presence	—	—	—	—	—	—	-0.046	0.955	0.033	0.156
Sidewalk physical disorder	—	—	—	—	—	—	0.157***	1.170	0.035	0.000
Structural physical disorder	—	—	—	—	—	—	0.049	1.050	0.033	0.143
Urbanization	—	—	—	—	—	—	-0.111**	0.895	0.041	0.006
Business activity	—	—	—	—	—	—	0.113***	1.120	0.036	0.001
Bus stops	—	—	—	—	—	—	-0.006	0.994	0.005	0.237
Street population	—	—	—	—	—	—	0.003***	1.003	0.000	0.000
Social service building	—	—	—	—	—	—	-0.054	0.948	0.073	0.460
Community-level variables										
Concentrated Disadvantage	—	—	—	—	—	—	0.075	1.078	0.055	0.172
Racial Diversity	—	—	—	—	—	—	0.003	1.003	0.002	0.066
Percent aged 19-24	—	—	—	—	—	—	-0.013	0.987	0.007	0.053
Random effects										
$\tau_{00}$	0.157		0.101				0.006			
$\chi^2$	32.99***		23.89***				0.77			
Log Likelihood	-2368.635		-2346.584				-2248.065			

Note: Incidence rate ratio (IRR) =  $\exp(b)$ \* $p \leq .05$ , \*\* $p \leq .01$ , \*\*\* $p \leq .001$  $N = 449$  streets (level 1);  $N = 53$  CSAs (level 2).

Because the dependent variable was crime counts and crime counts are typically over-dispersed, multilevel mixed-effects negative binomial regression models were estimated. Recall that in Chap. 6, we noted the importance of identifying whether the data for the dependent variable were over-dispersed. The likelihood ratio test for over-dispersion was statistically significant ( $\chi^2 = 9502.44$ ;  $p < 0.001$ ) in this example; this test compares the fit of a negative binomial model to a Poisson model, with the former clearly providing a better fit, strongly suggesting that the negative binomial model is more appropriate. As discussed in Chap. 6, do not place too much emphasis on this significance test. So long as there is any evidence of over-dispersion, even if not statistically significant, a negative binomial or quasi-Poisson model will produce results with more accurate  $p$ -values.

First, an unconditional multilevel model was estimated to assess the degree to which street-level crime varied across communities. The reported likelihood ratio test comparing the multilevel negative binomial model to a traditional negative binomial was significant ( $\chi^2 = 32.99; p \leq 0.001$ ), indicating significant variation across communities, justifying the need for multilevel modeling. The intraclass correlation was 0.264, which indicates that 26.4% of the variability in crime counts on the street can be attributed to differences across communities.

The estimated model can be more easily conceptualized if represented as two models, one for each level, as shown below:

$$\ln(Y_{ij}) = \beta_{0j} + \beta_1 \ln(x_1) + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k + \epsilon_{ij}$$

$$\beta_{0j} = \gamma_0 + \gamma_1 w_1 + \dots + \gamma_k w_k + u_j$$

where  $\ln(Y_{ij})$  is the natural log of the crime call counts for each street segment,  $\beta_{0j}$  is the intercept for each community,  $\ln(x_1)$  is collective efficacy for each street segment,  $x_2$  through  $x_k$  are the control variables at level 1, and  $\epsilon_{ij}$  is the level-1 error term. The community-level intercepts are then modeled by the second equation, where  $\gamma_0$  is the intercept across communities,  $w_k$  are the community level independent variables, and the  $u_j$  is the random coefficient for the community-level intercept. The level-1 regression coefficients are denoted as  $\beta_k$ , and the level-2 regression coefficients are denoted as  $\gamma_k$ .

To begin, the authors estimated the unconditional model (Model 1). They then included street-level collective efficacy in Model 2. This showed that higher levels of collective efficacy on the street were significantly related to lower crime levels on the street. Finally, in the full model (Model 3), they included the control variables at the street segment level, as well as the level-2 CSA structural indicators of collective efficacy. While the importance of collective efficacy declines in this model, it remains an important influence and is statistically significant ( $p = 0.003$ ). Using the IRR as a metric, a one-unit increase on the collective efficacy scale (scale ranges between 2.69 and 4.56) would be expected to lead to about a 25% decline in the crime rate of a street.

Structural indicators of collective efficacy were not statistically significant in these analyses, suggesting that scholars may overestimate the importance of structural measures at the community-level in predicting street-level collective efficacy. An important point to note in this model is that the level-2 coefficients are explaining variability in the crime counts across communities and the level-1 coefficients are explaining variability in the crime counts across street segments, taking into account community-level effects.

## Chapter Summary

---

**Multilevel models** offer an important extension to traditional linear regression models by statistically accounting for possible clustering in a sample of data. Observations that come from the same cluster (e.g., multiple survey respondents within the same neighborhood, multiple cases processed by a judge or prosecutor, and repeated measurements on the same individuals) will tend to be more similar to each other than to observations from different clusters. This results in an increased likelihood of erroneously finding statistically significant effects, since many of the cases within a cluster will exhibit a similar pattern of association. Multilevel models account for clustering by allowing for random variation in the intercepts and possibly the coefficients of the independent variables. Models that allow for variation in the model intercept are referred to as **random intercept models**, while models that contain a random intercept and at least one random slope coefficient are referred to as **random coefficient models**.

Variation in both the intercept and the coefficient for an independent variable is measured with what are called **variance components model**—measures of how much the intercept and slope may vary across cluster. These are also called the **random effects**. We test the significance of the variance estimates with a chi-square test that compares the model with the random effects against a linear regression model without any random effects.

In estimating multilevel models, we may also center the values of the independent variables. Centering can take on two forms: grand mean or cluster mean centering. Centering has no effect on the interpretation of the slope coefficients but will alter the substantive meaning of the model intercept. In **grand mean centering**, the model intercept represents the overall sample mean for the dependent variable. In **cluster mean centering**, the model intercept represents a weighted sample mean for the dependent variable that is conditioned on the number of cases per cluster. The more balanced the size of the clusters, the more similar the two estimates of the model intercept will be. In general, centering the values of the independent variables will tend to simplify the estimation of the overall multilevel model.

Centering also allows for the testing of different **between-** and **within-effects** of the independent variables. Much of the research in criminology and criminal justice assumes that the between and within effects are the same without ever testing for similarity. By estimating models that include the cluster means as independent variables, it is possible to assess directly how or whether the between cluster effects are the same as the within-cluster effects. If the results of these tests indicate the effects are the same,

then grand mean centering is appropriate. Alternatively, if the between and within effects are different, then cluster mean centering is a more appropriate technique.

## Key Terms

---

**Between effect** Effect of an independent variable on the dependent variable using the cluster as the unit of analysis—a regression of cluster-level averages across all the clusters included in the analysis.

**Cluster mean centering** Computed difference between the observed raw score on some variable for each observation in the sample and the cluster mean for that variable.

**Cross-level interaction** An interaction effect included in a multilevel model between a level-1 independent variable and a level-2 cluster's characteristics.

**Effect coding** A method for recoding a multicategory nominal variable into multiple indicator variables (one less than the total number of categories), where the indicator category is coded as 1, the reference category is coded as -1, and all other categories are coded as 0. Effect coding ensures that the sum of all the estimated effects for the indicator variable is equal to 0.

**Fixed effects** A descriptive label for the regression coefficients ( $b_k$ ) estimated in a model with random effects. Fixed effects represent the average effects of the independent variables on the dependent variable across all individuals and clusters in a multilevel model.

**Grand mean centering** Computed difference between the observed raw score on some variable for each observation in the sample and the overall sample mean for that variable.

**Intraclass correlation** A measure of association that measures the level of absolute agreement of values within each cluster.

**Multilevel data** Sample data where individual observations (level-1 data) are clustered within a higher-level sampling unit (level-2 data).

**Multilevel model** A type of model that accounts for nested or clustered data, such as students within classrooms, repeated measures within individuals, police officers within police departments, residents within hot spots, and households within neighborhoods. These models can include multiple levels of clustering.

**Random effects** A descriptive label for the random error terms included in a multilevel model that allow for variation across cluster from the sample average estimated in the fixed effects. Random effects are assumed to be normally distributed in most multilevel models.

**Random intercept model** A linear regression model that allows the intercept to vary randomly across cluster—random effects are included for the model intercept.

**Random coefficient model** A linear regression model that allows the intercept and the effect of at least one independent variable to vary randomly across clusters—random effects are included for the model intercept and at least one independent variable.

**Regression coding** A method for recoding a multicategory nominal variable into multiple indicator dummy variables (one less than the total number of categories), where the indicator category is coded as 1 and all other categories are coded as 0. The reference category does not have an indicator variable and is coded as a 0 on all the indicator dummy variables.

**Variance components model** A one-way analysis of variance model that includes random effects for each cluster that assesses whether there is random variation in the mean of the dependent variable across the clusters included in the analysis.

**Within effect** Effect of an independent variable on the dependent variable within each cluster and then averaged across all clusters or groups included in the analysis.

## Symbols and Formulas

---

$u_{0j}$	Random effect for the model intercept $b_0$
$u_{kj}$	Random effect for the regression coefficient for the independent variable $k$ ( $b_k$ )
$\sigma^2_{u_{0j}}$	Variance of the random effect for the model intercept
$\sigma^2_\epsilon$	Error variance (unexplained variance) in a random intercept model
$\sigma^2_{u_{0j}^2}$	Variance of the random effect for the model intercept in a random coefficient model
$\sigma^2_{u_{kj}}$	Variance of the random coefficient for independent variable $k$
$\sigma_{u_{0k}}^2$	Covariance of the random effects for the model intercept and for the independent variable $k$ in a random coefficient model

General equation for the variance components model:

$$y_{ij} = b_0 + u_j + e_{ij}$$

General equation for the random intercept model with one independent variable:

$$y_{ij} = b_0 + b_1 x_1 + u_{0j} + e_{ij}$$

General equation for the random coefficient model with one independent variable:

$$y_{ij} = b_0 + b_1 x_1 + u_{0j} + u_{1j} + e_{ij}$$

Likelihood ratio (LR) test for variance components comparing two models (1 and 2):

$$\chi^2 = -2(LL_1 - LL_2)$$

Equation for computing the intraclass correlation ( $\rho$ ):

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$$

Equation for grand mean centering:

$$x_{ij} - \bar{x}_{..}$$

Equation for cluster mean centering:

$$x_{ij} - \bar{x}_{.j}$$

## Exercises

---

- 7.1. Researchers interested in the possible effects of neighborhood poverty on patterns of intimate partner violence (IPV) gathered interview data from 100 residents in each of a city's 53 neighborhoods. As a first step in establishing neighborhood variability in IPV, the researchers estimated a variance components model and obtained the following results:

$$-2LL (\text{null model}) = -2000$$

$$-2LL(\text{random-effects model}) = -1900$$

$$\sigma_u^2 = 0.10$$

$$\sigma_\epsilon^2 = 0.40$$

- (a) Test whether there is significant variation in the prevalence of IPV across these 53 neighborhoods. Explain the substantive meaning of your findings.
- (b) To what extent does neighborhood affect the prevalence of IPV? Calculate the intraclass correlation coefficient, and explain its substantive meaning.
- 7.2. In a study of 13,726 students across 491 schools, a study asked about self-reported delinquent behavior, which was measured with a scale that ranged from 0 (no delinquency) to 7 (high rate of delinquency). The researchers were particularly interested in the effects of academic performance on delinquent behavior and estimated a random intercept model, obtaining the following results:
- | VARIABLE                | COEFFICIENT |
|-------------------------|-------------|
| Intercept               | 0.50        |
| GPA                     | -0.30       |
| Educational aspirations | -0.05       |
| Father's education      | -0.20       |
| Mother's education      | -0.35       |
| $\sigma_u^2$            | 0.15        |
| $\sigma_e^2$            | 0.95        |
- (a) Calculate the intraclass correlation coefficient for the variance components model, and explain its substantive meaning.
- (b) Calculate the intraclass correlation coefficient for the random intercept model, and explain its substantive meaning.
- (c) Interpret the change in the value of the intraclass correlation coefficient between the variance components and random intercept models.
- (d) How would the intraclass correlation coefficient change if additional covariates were added to the model? Describe how the statistical significance of the added covariate affects the change in the value of the intraclass correlation coefficient in this case.
- 7.3. A study of antisocial behavior among children collected information on 674 families, each with at least two children who could participate in the study. Based on prior research, the investigators expected the within-family and between-family effects of parental attachment to be different. The investigators found the following effects:

$$b_{Attachment} = -0.23, se = 0.07$$

$$b_{ClusterMeanofAttachment} = -0.16, se = 0.03$$

- (a) Explain whether the investigators found evidence of different within-family and between-family effects of parental attachment.
- (b) Which type of centering would be most appropriate for these data if the investigators simply want to estimate a single effect for parental attachment that ignores the between and the within effects?
- 7.4. Researchers interested in studying the effects of neighborhood characteristics on individuals' perceptions of fear of crime victimization selected a random sample of 100 neighborhoods and then interviewed 50 residents within each neighborhood. Using a fear of crime scale as the dependent variable and demographic characteristics as covariates, the researchers estimated a series of regression models to test for random effects across neighborhood. The results appear in the following table (assume all fixed effects are statistically significant with  $p < 0.05$ ):

VARIABLE	OLS	RANDOM INTERCEPT MODEL	RANDOM COEFFICIENT MODEL
<i>Fixed effects</i>			
Intercept	4.50	4.40	4.50
Age	0.07	0.06	0.05
Female	0.78	0.80	0.90
Black	0.64	0.52	0.48
<i>Random effects</i>			
Intercept		0.03	0.03
Age			0.01
Female			0.12
Black			0.05
<i>Model information</i>			
– 2LL	–3176	–2273	–2095

- (a) Test whether there is significant variation in the model intercept across the 100 neighborhoods using a significance level of 5%. Interpret your result.
- (b) Test whether there is significant variation in the coefficients for age, female, and black using a significance level of 5%. Interpret your result.
- 7.5. A study of sentence length decisions began by selecting a random sample of 35 judges within a large state. For each judge, the researchers selected a random sample of 200 cases among those involving sentences to jail or prison. Sentence length was measured as the number of months sentenced to incarceration. To assess the impact of legal characteristics on sentence length decisions, the researchers developed measures of severity of the conviction crime and of criminal history. After establishing that the intercept varied across judge, the researchers investigated a series of random coefficient models that examined whether there were correlations of the random effects for the intercept and the two covariates. The following table presents their results:

VARIABLE	RANDOM INTERCEPT MODEL	RANDOM COEFFICIENT MODEL 1	RANDOM COEFFICIENT MODEL 2
<i>Fixed effects</i>			
Intercept	5.19	5.23	7.98
Severity of offense	2.72	2.65	1.95
Criminal history	5.31	5.62	4.97
<i>Random effects</i>			
Intercept	0.29	0.26	0.21
Severity of offense		0.16	0.12
Criminal history		0.21	0.18
<i>Covariances</i>			
Intercept severity			0.06
Intercept history			0.04
Severity history			0.11
<i>Model information</i>			
-2LL	-2984	-2781	-2779

- (a) Test whether there is significant variation in the coefficients for offense severity and criminal history using a significance level of 5%. Interpret your result.
- (b) Test whether the addition of the random-effect covariances is statistically significant using a significance level of 5%. Interpret your result.

## Computer Exercises

We have noted throughout this chapter that multilevel models can become complex very quickly, and so we have tried to keep our focus on the basic elements of multilevel models. The essential syntax required to estimate these models is generally straightforward and does not become complicated until we start customizing the statistical model. The examples below and the accompanying syntax files for SPSS (Chapter\_7.sps) and Stata (Chapter\_7.do) illustrate key components to the multilevel commands without getting bogged down in too many of the options and details.

### SPSS

In SPSS, random intercept models and random coefficient models are both estimated with the MIXED command:

```
MIXED dep_var WITH list_of_indep_vars
/PRINT = SOLUTION TESTCOV
/FIXED = INTERCEPT list_of_indep_vars
/RANDOM = INTERCEPT random_indep_var(s) |
SUBJECT(cluster_variable) COVTYPE(UN).
```

where the first line has the same structure as many of the other commands in SPSS. The /PRINT=SOLUTION TESTCOV option requests SPSS to print out the coefficient table (SOLUTION) and to test the random-effect variances and covariances for statistical significance (TESTCOV). The /METHOD=ML option forces SPSS to estimate the models with maximum likelihood, while the FIXED = option line should include INTERCEPT and all of the independent variables included in the model. The /RANDOM = option will then determine whether a random intercept model (RIM) or random coefficient model (RCM) will be estimated. If a RIM is to be estimated, the /RANDOM line simplifies to:

```
/RANDOM = INTERCEPT | SUBJECT(cluster_variable).
```

For an RCM that only estimates random-effects variances:

```
/RANDOM = INTERCEPT random_indep_var(s) |
SUBJECT(cluster_variable).
```

For an RCM that estimates random-effects covariances and variances, you will need to add the COVTYPE(UN) option to the /RANDOM line, since the default in SPSS is to estimate an RCM without random-effect covariances:

```
/RANDOM = INTERCEPT random_indep_var(s) |
SUBJECT(cluster_variable) COVTYPE(UN).
```

The output from running any of these commands is the same and includes tables of coefficients (the fixed effects), of covariances and variances for the random effects, and of model summary statistics.

## Stata

### *Random Intercept Models*

To estimate random coefficient models in Stata, you will have access to two different commands: **xtreg** and **xtmixed**. The structure to the **xtreg** command is as follows:

**xtreg** depvar list\_of\_indep\_vars, i(cluster\_variable) **mle var**

The basic format is similar to the regress command—the difference follows the comma, where the **i()** option indicates which variable provides information on the cluster. In this chapter, the cluster was the judge identifier. The **var** option is included to force Stata to estimate the variance of the random intercept—the default output in Stata is to report the square root of the variance (i.e., the standard

deviation of the random effect). Finally, the **mle** option forces Stata to compute the maximum likelihood estimates for the RIM.

The use of the **xtmixed** command is similar:

```
xtmixed depvar list_of_indep_vars || cluster_variable; mle var
```

where instead of a comma immediately following the list of independent variables, we have two vertical bars (||) that are followed by the cluster variable with a colon (:) appended and then the comma and request for maximum likelihood estimates. As we explain in the next section regarding the estimation of random coefficient models, the vertical bars will be useful for designating random coefficients.

### *Random Coefficient Models*

The **xtreg** command cannot be used for random coefficient models, while **xtmixed** can be used. To use **xtmixed** for an RCM, we simply add the name of the independent variable(s) after the : that we want to estimate random effects for:

```
xtmixed depvar list_of_indep_vars || cluster_variable:  
random_indep_var(s), mle var
```

The structure to the rest of the command is the same—all we do is include one or more independent variable names after the : and before the comma. Note that the default RCM in Stata is to estimate a model with no covariances of the random effects. If we want to estimate the covariances of the random effects, we add the option **cov(unstructured)** to the command line:

```
xtmixed depvar list_of_indep_vars || cluster_variable:  
random_indep_var(s), mle var cov(unstructured)
```

The output will then contain the variances for the intercept, any independent variables with effects allowed to vary across cluster, and all possible covariances of the random effects.

## R

### *Random Intercept Models*

The **lmer()** function can be used for random coefficient models, which is within the *lme4* package. This is done by specifying 1| before the cluster variable, as follows:

```
lmer(depvar~ indep_var1 + indep_vars2 + (1|  
cluster_variable), data=dataset_name, REML=FALSE)
```

We want the models to be estimated using maximum likelihood, so you will specify REML as FALSE. For instance, see the following example with the bail dataset:

```
lmer(logbail ~ (1 | judge), data=df, REML=FALSE)
```

### *Random Coefficient Models*

To estimate random coefficient models, we can use the same **lmer()** function, but we need to specify it a little differently. Here, we need to specify **0 +** before the independent variable and end it with the | symbol and clustering variable, whereby you specify the slope variable first and the predict of the slope afterwards.

```
lmer(depvar~ indep_var1 + indep_vars2 + (1|cluster_variable)
    + (0 + indep_var1|slope_predictor),
    data=dataset_name, REML=FALSE)
```

An example of this using the bail dataset is as follows:

```
lmer(logbail ~ numdoff + (1|judge) + (0 + numdoff|judge),
    data=df, REML=FALSE)
```

To allow the intercept and slope to be correlated when using **lmer()**, just replace **(0 + indep\_var1|slope\_predictor)** with **(1 + indep\_var1|slope\_predictor)**. This is the same as Stata's *cov(unstructured)*.

The **lmer()** function does not provide the user with a *p*-value to assess the statistical significance of the overall model. That can be obtained using the **anova()** function on the *lmer* object used to store the regression model results:

```
model<-lmer(logbail ~ numdoff + (1|judge) +
    (0 + numdoff|judge), data=df, REML=FALSE)
anova(model)
```

The **devfun2()** and **VarCorr()** functions from the *lme4* package can be used to get Wald standard error of variance estimates. We will also rely on the *numDeriv* package to use the **hessian()** function. To do this, run the regression model as we do above and store the results in an object (we name the object *model*). Then, once the required packages are installed, extract the parameters as standard deviations and compute the Hessian matrix, as follows:

```
dd.ML <- devfun2(model, useSc=TRUE, signames=FALSE)
my.data <- as.data.frame(VarCorr(model))
pars <- my.data[,"sdcor"]
hess <- hessian(dd.ML,pars)
my.data2 <- 2*solve(hess)
sqrt(diag(my.data2))
```

### **Problems**

1. Open the Bail Decision-Making data file (bail-data-example.sav or bail-data-example.dta). The sample syntax files for this chapter include the syntax required to reproduce most of the tables in this chapter. Work your way through one of the syntax files, and make sure you understand how it works.

2. Using the Bail Decision-Making data file, use *logbail* as the dependent variable. Select a set of 4–6 independent variables, and estimate the following models:
  - (a) Random intercept model:
    - Which of the fixed effects are statistically significant at a 5% level of significance?
    - Interpret each of the statistically significant fixed effects.
    - Test whether the RIM offers a significant improvement over a linear regression model. Use a significance level of 5%.
  - (b) Random coefficient model: Select at least one, but no more than three, of the independent variables to have a random coefficient. Do not estimate the covariances of the random effects.
    - Did anything change in regard to the effects of these independent variables? Value of the coefficient? Level of statistical significance?
    - Test whether this RCM offers a significant improvement over the RIM estimated in part (a). Use a significance level of 5%.
  - (c) Random coefficient model: Use the same RCM as in part (b), but estimate the random-effect covariances.
    - Did anything change in regard to the effects of these independent variables? Value of the coefficient? Level of statistical significance?
    - Test whether this RCM offers a significant improvement over the RCM estimated in part (b). Use a significance level of 5%.
3. Continue to use the Bail Decision-Making data file, change the dependent variable to the bail amount requested, and estimate the same set of models with the same independent variables as in Question 2.
  - (a) Explain how the results are similar or different. Focus on the values of the coefficients and the covariances and variances of the random effects.
  - (b) What might account for these differences? (Hint: You may want to generate histograms for both dependent variables as a starting point.)

## References

---

- Angrist, J.D., & Pischke, J. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.
- Britt, C. L. (2000). Social context and racial disparities in punishment decisions. *Justice Quarterly*, 17, 801–826.

- Gelman, A. (2005). Analysis of variance: Why it is more important than ever (with discussion). *Annals of Statistics*, 33, 1–53.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.
- Goldkamp, J. S., & Gottfredson, M. R. (1985). *Policy guidelines for bail: An experiment in court reform*. Philadelphia, PA: Temple University Press.
- Hays, W. L. (1988). *Statistics*. New York, NY: Holt, Rinehart, and Winston.
- Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using stata, volume I: Continuous responses* (3rd ed.). College Station, TX: Stata Press.
- Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, 277(5328), 918–924.
- Weisburd, D., White, C., & Wooditch, A. (2020). Does collective efficacy matter at the micro geographic level? Findings from a study of street segments. *The British Journal of Criminology*, 60(4), 873–889.

## Chapter eight

---

# Statistical Power

### **How Is Statistical Power Defined?**

---

What is Statistical Power?

How Do Significance Criteria Influence Statistical Power?

How Does Effect Size Influence Statistical Power?

How Does Sample Size Influence Statistical Power?

### **Estimating Statistical Power**

---

How Do We Define the Significance Criteria and Effect Size in a Statistical Power Analysis?

How Do We Determine the Sample Size Needed to Ensure a Statistically Powerful Study?

**A**S WE HAVE SEEN IN EARLIER CHAPTERS, researchers place a premium on statistical inference and its use in making decisions about population parameters from sample statistics. In assessing statistical significance, the focus is the problem of Type I, or alpha ( $\alpha$ ), error: the risk of falsely rejecting the null hypothesis. Paying attention to the statistical significance of a finding should keep researchers honest, because it provides a systematic approach for deciding when the observed statistics are convincing enough for the researcher to state that they reflect broader processes or relationships in the general population from which the sample was drawn. If the threshold of statistical significance is not met, then the researcher cannot reject the null hypothesis and cannot conclude that a relationship exists.

Another type of error that most researchers are aware of, but pay relatively little attention to, is Type II, or beta ( $\beta$ ), error: the risk of falsely failing to reject the null hypothesis. A study that has a high risk of Type II error is likely to mistakenly conclude that treatments are not worthwhile or that a relationship does not exist when in fact it does in the population. Understanding the risk of a Type II error is crucial to the development of a research design that will give the researcher a reasonable chance of finding a treatment effect or a statistical relationship, if those effects and relationships exist in the population. This is fundamentally what we mean by statistical power—given the current design of a study, does it have the ability (i.e., the power) to detect statistically significant effects and relationships?

Although researchers in criminology and criminal justice have placed much more emphasis on the statistical significance than on the statistical power of a study, research in fields such as medicine and psychology routinely reports estimates of statistical power (see, for example, Maxwell et al. 2008). Federal funding agencies (e.g., National Institutes of Health, the National Institute of Justice) typically require research proposals to estimate how powerful the proposed research design will be. The purpose of this chapter is to present an introductory discussion of the key components in

an assessment of the statistical power of a research design and to explain why it is important to have a basic understanding of the importance of statistical power in designing and evaluating research.

## Statistical Power

---

**Statistical power** measures the probability of rejecting the null hypothesis when it is false, but it cannot be measured directly. Rather, statistical power is calculated by subtracting the probability of a Type II error—the probability of falsely failing to reject the null hypothesis—from 1:

$$\text{Power} = 1 - \text{Probability}(\text{Type II error}) = 1 - \beta.$$

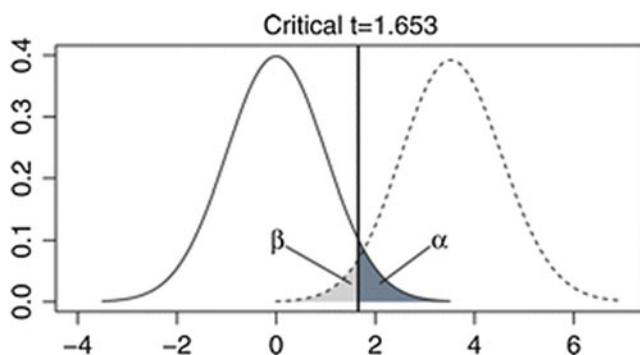
For many sample statistics, the Type II error can be estimated directly from the sampling distributions commonly assumed for most test statistics. In contrast to a traditional test of statistical significance, which identifies for the researcher the risk of stating that factors are related when they are not (i.e., the Type I error), statistical power measures how often one would fail to identify a relationship that in fact does exist in the population. For example, a study with a statistical power level of 0.90 has only a 10% probability of falsely failing to reject the null hypothesis. Alternatively, a study with a statistical power estimate of 0.40 has a 60% probability of falsely failing to reject the null hypothesis. As the statistical power of a proposed study increases, the risk of making a Type II error decreases.

Figure 8.1 presents the relationship between Type I and Type II errors graphically. Suppose that we are interested in a difference in group means, say between a control and treatment group in a criminal justice experiment, and based on prior research and theory, we expect to find a positive difference in the outcome measure. We would test for a difference in the group means by using a one-tailed  $t$ -test. If we have 100 cases in each group, then the critical  $t$ -value is 1.653 for  $\alpha = 0.05$ . The distribution on the left side of Fig. 8.1 indicated by the solid line represents the  $t$ -distribution—the sampling distribution—under the null hypothesis (i.e., the center or mean of this distribution is the hypothesized null value or 0), with the significance level ( $\alpha$ ) shaded in the right tail of the distribution—the vertical line represents the critical  $t$ -value.

The distribution on the right side of Fig. 8.1 and indicated by the dashed line represents the hypothesized sampling distribution based on prior research and theory and our expectations for the expected differences in the two group means (i.e., the center or mean of this distribution is the expected effect in the population). The hypothesized sampling distribution

**Figure 8.1**

*Graphical representation of Type I and Type II errors in a difference of means test (100 cases per sample)*



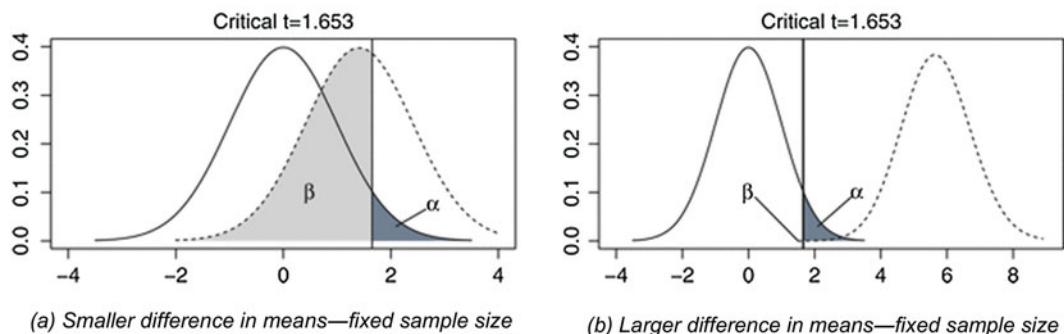
is also a  $t$ -distribution, but it is known as a noncentral  $t$ -distribution—we illustrate below how this distribution is used to compute statistical power. The probability of making a Type II error ( $\beta$ ) is denoted in the figure and is the cumulative probability in the distribution on the right up to the critical  $t$ -value (i.e.,  $t < 1.653$ ). The statistical power of this difference of means test is represented in the figure by the area under the dashed line that falls to the right of the critical value and represents  $t$ -values of 1.653 and above, which would be interpreted as evidence in favor of rejecting the null hypothesis. The area of the curve denoted by  $\beta$  represents the proportion of the distribution where we would fail to reject the null.

It is important to note that our estimate of  $\beta$  is fully dependent on our estimate of the expected magnitude of the difference between the two groups. Figure 8.2 illustrates the differences for two alternative effect sizes while assuming that the sample sizes remain fixed at 100 cases per group. For example, if we expect the difference of means to be smaller, we would shift the hypothesized sampling distribution (the dashed line) to the left, increasing our estimate of  $\beta$  (see Panel (a)). If we expect a larger difference, we would shift the hypothesized sampling distribution to the right, reducing the estimate of  $\beta$  (see Panel (b)).

If the statistical power of a research design is high and the null hypothesis is false for the population under study, then it is very likely that the researcher will reject the null hypothesis and conclude that there is a statistically significant finding. If the statistical power of a research design is low, it is unlikely to yield a statistically significant finding, even if the research hypothesis is in fact true. Studies with very low statistical power are sometimes described as being *designed for failure*, because a study that

**Figure 8.2**

*Graphical representation of Type I and Type II errors in a difference of means test—changing the difference in mean values*



is underpowered is unlikely to yield a statistically significant result, even when the outcomes observed are consistent with the research hypothesis.<sup>1</sup>

Consider the implications for theory and practice in criminology and criminal justice of a study that has low statistical power. Suppose that a promising new program has been developed for dealing with spouse assault. If that program is evaluated with a study that has low statistical power, then the research team will likely fail to reject the null hypothesis based on the sample statistics, even if the program does indeed have the potential for reducing spouse assault. Although the research team is likely to say that the program does not have a statistically significant impact on spouse assault, this is not because the program is not an effective one, but because the research team designed the study in such a way that it was unlikely to be able to identify program success. Conceptually, this same problem occurs in the analysis of other types of data when trying to establish whether a relationship exists between two theoretically important variables. The relationship may exist in the population of interest, but a study with low statistical power will be unlikely to conclude that the relationship is statistically significant.

One might assume that researchers would work hard to develop statistically powerful studies, because such studies are more likely to support the research hypothesis proposed by the investigators. Unfortunately, statistical power is sometimes ignored altogether by criminology and criminal justice researchers, which results in many studies having a low level of statistical power. Furthermore, there are often constraints on what is feasible in terms of sample size that limit the statistical power of individual studies.

---

<sup>1</sup>For an extended discussion of this, see Weisburd (1991).

### Setting the Level of Statistical Power

What is a desirable level of statistical power? There is no single correct answer to this question, since it depends on the relative importance of Type I and Type II errors for the researcher. That said, one of the more common suggestions in the statistical power literature has been that studies should attempt to achieve a power level of 0.80, meaning that the chances of a Type II error are  $\beta = 0.20$ . There are many ways in which this is an arbitrary threshold. At the same time, it implies a straightforward gauge for the relative importance of both types of error. If we use a conventional level of statistical significance ( $\alpha = 0.05$ ) and statistical power (0.80,  $\beta = 0.20$ ), it implies that the researcher is willing to accept a risk of making a Type II error that is four times greater than the risk of a Type I error:

$$\beta/\alpha = 0.20/0.05 = 4.0.$$

If the target level of statistical power is 0.90, then  $\beta = 0.10$ , and the ratio of probabilities decreases to  $0.10/0.05 = 2.0$ . What this means is that for a fixed level of statistical significance ( $\alpha$ ), increasing the level of statistical power reduces the chances of a Type II error ( $\beta$ ) at the same time that the ratio of  $\beta/\alpha$  moves closer to 1.0, where the chances of both types of error are viewed as equally important.

What happens if we reduce the desired level of statistical significance? For example, suppose we were particularly concerned about our chances of making a Type I error and reduced the significance level of a test from 0.05 to 0.01. For a statistical power level of 0.80, this would imply that we are willing to accept a probability of making a Type II error that is 20 times greater than the probability of a Type I error. If we simultaneously increase the level of statistical power to 0.90 at the same time we reduce the significance level, the  $\beta/\alpha$  ratio decreases to 10, but it still implies a much greater likelihood of a Type II error. If we wanted to keep the ratio of error probabilities at 4.0, we would need a study with a power level of 0.96 ( $=1 - 4(\alpha) = 1 - 0.04$ ). Intuitively, this makes good sense though: If we are going to make it more difficult to reject the null hypothesis by reducing  $\alpha$ , we will simultaneously increase our chances of failing to reject a false null hypothesis unless we have a more powerful study.

### Components of Statistical Power

---

The level of statistical power associated with any given test of a sample statistic is influenced by three key elements:

1. Level of statistical significance, including directional tests when appropriate
2. Sample size
3. Effect size (the true population effect)

The level of statistical significance and sample size are assumed to be within the control of the researcher, while the estimated effect size is not. The following discussion briefly describes the links between each element and the statistical power of any given test.

### **Statistical Significance and Statistical Power**

The most straightforward way to increase the statistical power of a test is to change the significance level used. As we reduce the chances of making a Type I error by reducing the level of statistical significance from 0.10 to 0.05 to 0.01, it becomes increasingly difficult to reject the null hypothesis. Simultaneously, the power of the test is reduced. A significance level of 0.05 results in a more powerful test than a significance level of 0.01, because it is easier to reject the null hypothesis using the more lenient significance criteria. Conversely, a 0.10 level of significance would make it even easier to reject the null hypothesis.

As a simple illustration, Table 8.1 presents  $z$ -scores required to reject the null hypothesis for several levels of statistical significance using a two-tailed test. It would take a  $z$ -score greater than 1.645 or less than -1.645 to reject the null hypothesis with  $\alpha = 0.10$ , a  $z$ -score greater than 1.960 or less than -1.960 with  $\alpha = 0.05$ , and a  $z$ -score greater than 2.576 or less than -2.576 for  $\alpha = 0.01$ . Clearly, it is much easier to reject the null hypothesis with a 0.10 significance threshold than with a 0.01 significance threshold.

This method for increasing statistical power is direct, but it means that any benefit we gain in reducing the risk of a Type II error is offset by an increase in the risk of a Type I error. By setting a more lenient significance threshold, we do indeed gain a more statistically powerful research study. However, the level of statistical significance of our test also declines. Since a 0.05 significance level has become the convention in much of the research

**Table 8.1**

***$z$ -scores needed to reject the null hypothesis in a two-tailed test of statistical significance by level of  $\alpha$***

$\alpha$	$z$
0.20	$\pm 1.282$
0.10	$\pm 1.645$
0.05	$\pm 1.960$
0.01	$\pm 2.576$
0.001	$\pm 3.291$

in criminology and criminal justice, it is important for authors to note why a more (or less) restrictive level of statistical significance is used.

### *Directional Hypotheses*

A related method for increasing the statistical power of a study is to limit the direction of the research hypothesis to either a positive or a negative outcome, which implies the use of a one-tailed statistical test. A one-tailed test will provide greater statistical power than a two-tailed test for the same reason that a less stringent level of statistical significance provides more power than a more stringent one. By choosing a one-tailed test, the researcher reduces the absolute value of the test statistic needed to reject the null hypothesis by placing all of the probability of making a Type I error in a single tail of the distribution.

We can see this in practice again with the  $z$ -test. Table 8.2 lists the  $z$ -scores needed to reject the null hypothesis in one- and two-tailed tests for five different levels of statistical significance. (For the sake of simplicity, we assume in the one-tailed test that the outcome will be positive.) At each level, as in other statistical tests, the test statistic required to reject the null hypothesis is smaller in the case of a one-tailed test. For example, at  $\alpha = 0.05$ , a  $z$ -score greater than 1.960 or less than -1.960 is needed to reject the null hypothesis in the two-tailed test. In the one-tailed test, the  $z$ -score needs only to be greater than +1.645. When we reduce the significance level to  $\alpha = 0.01$ , a  $z$ -score greater than +2.576 or less than -2.576 is needed to reject the null hypothesis in the two-tailed test, but in the one-tailed test, the  $z$ -score needs only to be greater than +2.326.

Although the researcher can increase the statistical power of a study by using a directional, as opposed to a nondirectional, research hypothesis, there is a price for shifting the rejection region to one side of the sampling distribution. Once a one-directional test is defined, a finding in the direction opposite to that originally predicted cannot be recognized, no matter how big. To do otherwise would bring into question the integrity of the assumptions of the statistical test used in the analysis.

**Table 8.2**

$z$ -scores needed to reject the null hypothesis in one-tailed and two-tailed tests of statistical significance

$\alpha$	<b><math>z</math>-SCORE (ONE-TAILED TEST)</b>	<b><math>z</math>-SCORE (TWO-TAIL TEST)</b>
0.20	+0.842	$\pm 1.282$
0.10	+1.282	$\pm 1.645$
0.05	+1.645	$\pm 1.960$
0.01	+2.326	$\pm 2.576$
0.001	+3.090	$\pm 3.291$

### Sample Size and Statistical Power

The method used most often to change the level of statistical power in social science research is to vary the size of the sample. Similar to specifying the level of statistical significance, sample size often can be controlled by the researcher. Modifying the size of the sample is typically a more attractive option for increasing statistical power than modifying the level of statistical significance, since the risk of a Type I error remains fixed—presumably at the conventional  $\alpha = 0.05$ .

The relationship between statistical power and sample size is straightforward. All else being equal, larger samples provide more stable estimates of the population parameters than do smaller samples. Assuming that we are analyzing data from random samples of a population, the larger sample will have smaller standard errors of the coefficients than will the smaller sample. As the number of cases in a sample increases, the standard error of the sampling distribution (for any given statistical test) decreases. For example, the standard error for a single-sample  $t$ -test is

$$\sigma_{\mu} = \frac{\sigma}{\sqrt{n-1}}.$$

As  $n$  gets larger, irrespective of the value of the standard deviation ( $\sigma$ ) itself, the standard error of the estimate ( $\sigma_{\mu}$ ) gets smaller. As the standard error of a test decreases, the likelihood of achieving statistical significance grows, because the test statistic for a test of statistical significance is calculated by taking the ratio of the difference between the observed statistic and the value proposed in the null hypothesis (typically 0) to the standard error of that difference. If the difference is held constant, then as the sample size increases, the standard error decreases, and a larger test statistic is computed, making it easier to reject the null hypothesis.

The effect of sample size on statistical power for a  $t$ -test of the difference of two independent sample means is illustrated in Table 8.3. The last column of Table 8.3 indicates the number of statistically significant outcomes expected in 100 two-sample  $t$ -tests in which a mean difference of two arrests between groups ( $\sigma = 1$ ) is examined for four different scenarios (using a 5% significance threshold and a two-tailed test). In the first scenario, the sample size for each group is only 35 cases; in the second scenario, the sample size is 100; in the third, 200; and in the fourth, fully 1000. Table 8.3 shows that the likelihood of rejecting the null hypothesis changes substantially with each increase in sample size, even though all other characteristics are held constant across the four scenarios. Under the first scenario, we would expect only about 13 statistically significant outcomes in 100 tests. In the second scenario, 29 significant outcomes would be expected and in the third, 51. In the final scenario of samples of 1000,

**Table 8.3**

Number of statistically significant outcomes expected in 100 two-sample t-tests for six different scenarios

SCENARIO	SAMPLE SIZE (PER GROUP)	$\mu_1 - \mu_2$	$\sigma$	EXPECTED SIGNIFICANT OUTCOMES
1	35	0.2	1	13
2	100	0.2	1	29
3	200	0.2	1	51
4	1,000	0.2	1	99

nearly every test (99 out of 100) would be expected to lead to a significant result.

Sample size is often a primary concern in statistical power analysis because (1) it is directly related to statistical power, (2) it is a factor usually under the control of the researcher, and (3) it can be manipulated without altering the criteria for statistical significance of a study.

In most cases, researchers maximize the statistical power of a study by increasing sample size. The concern with sample size is also reflected in the number of publications focused on advising researchers in all behavioral and social science fields on how to determine the appropriate sample size for a proposed research study.<sup>2</sup>

Although sample size should be under the control of the researcher, it is important to be aware of the unanticipated consequences of simply increasing sample size may have on other factors that influence statistical power, particularly in evaluation research (Weisburd 1991). For example, suppose a researcher has developed a complex and intensive method for intervening with high-risk youth. The impact of the treatment is dependent on each subject receiving the full dosage of the treatment for a 6-month period. If the researcher were to increase the sample size of this study, it might become more difficult to deliver the treatment in the way that was originally intended. More generally, increasing the sample size of a study can decrease the integrity or the dosage of the interventions that are applied and result in the study showing a small or null effect of the treatment. Also, studies are likely to include more heterogeneous groups of subjects as sample size increases. For example, in a study of intensive probation, eligibility requirements were continually relaxed in order to meet project goals regarding the number of participants (Petersilia 1989). As noted below, these are not only threats to the integrity of a study, but also are likely to reduce the statistical power of a study.

---

<sup>2</sup>For a range of examples, see Dattalo (2008), Kraemer and Thiemann (1987), and Murphy and Myors (2003).

### Effect Size and Statistical Power

An **effect size (ES)** is simply an index of the size of the effect of interest. The specific definition will change depending on the type of data and statistical model of interest. In the examples we have discussed so far in this chapter, the effect size is the difference between two means. An effect size might also be a correlation coefficient, a simple proportion, a regression coefficient, or an index of the variability across 3 or more means, among others. In using the term *effect*, we are not implying causation, as in *cause-and-effect* but simply the size of the statistical parameter of interest in our particular research context.

The same effect size in one study might be statistically significant but not in another study given differences in sample sizes. This was illustrated in Table 8.3. Each row of this table has the same effect size (the same difference between the means), but the rows with larger samples sizes have greater power, that is, be more likely to correctly reject the null. Stated in another way, the effect size should reflect the effect of interest and not the combination of the effect of interest and the sample size.

The noncentrality parameter ( $\delta$ ) in the context of power analysis is the difference between the location of the sampling distribution in the population relative to the null sampling distribution. Because sampling distributions are based on our test statistics, the noncentrality parameter is the difference between the mean of the true sampling distribution and the null sampling distribution in units of the test statistics. This is expressed in Eq. (8.1).

$$\delta = \text{Mean test statistic}_{\text{population}} - \text{Mean test statistic}_{\text{null}}$$

**Equation 8.1**

This equation is finding the location of the presumed population sampling distribution shown as the distribution on the left in Figs. 8.1 and 8.2, relative to the null distribution.

In most situations, the null hypothesis is zero, such as no difference between two means or a correlation equal to zero. Thus, the mean of the test statistic under the assumptions of a null hypothesis equal to zero is also zero. In these cases, the noncentrality parameter,  $\delta$ , simplifies to equal the mean of the test statistic for the true value in the population (in statistical power analysis, we do not know this value, so it is the expected value in the population). However, there are situations where the null is not zero. For example, if you were testing whether a coin is unbiased, the null hypothesis is that the population value for the proportion of heads versus tails is 0.50. In such a case, the noncentrality parameter is the difference between the true proportion (i.e., based on an infinite number of coin flips) and 0.50. In power analysis, you would most likely set this value to be the minimum amount of bias that you think is substantively meaningful in your context.

In estimating statistical power, it is useful to use standardized effect sizes, particularly in the context of mean differences. A mean difference of 24.61 is meaningless without knowledge of the nature of the scale and more importantly the variability across individuals (or other units) on this scale. This mean difference would be very large if the scale ranged from 0 to 50 with a standard deviation of 10 but would be very small if the scale ranged from 0 to 5,000 with a standard deviation of 1,000. Standardizing on the variability of a distribution of scores allows for the comparison of effects across studies that may have used different scales or slightly different types of measures. It also facilitates the estimation of statistical power across a wide range of studies and types of analyses. The correlation coefficient is a standardized effect size that you are already familiar with. Correlation coefficients are scaled between  $-1.0$  and  $+1.0$  independently from the scaling of the two measures being correlated.

The relationship between effect size and statistical power should be clear. When the standardized population parameter differs substantially from that proposed under the null hypothesis, the researcher should be more likely to observe a significant difference or effect in a particular sample. The larger the effect, the easier it is to find given a fixed sample size.

A difference of means test for two independent samples provides a simple illustration for these relationships. The standardized effect size in this case is Cohen's  $d$  (Cohen 1988). This effect size standardizes the mean difference relative to the pooled standard deviation within the groups. Conceptually, this expresses the mean difference as a  $z$ -score except that the standard deviation is the natural variability in the dependent variable without the effect of the independent variable, if there is one (e.g., without any treatment effect variability). Thus, the ES for a difference of means test may be defined simply as the raw difference between the two population parameters, divided by their common (within groups) standard deviation, as shown in Eq. (8.2):

$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

Equation 8.2

To reiterate an earlier comment, when the difference between the population means is greater, the ES for the difference of means will be larger. Also, as the variability of the scores grows, as represented by the standard deviation of the estimates, the ES will get smaller.

Table 8.4 presents a simple illustration of the relationship between effect size and statistical power in practice. The last column of Table 8.4 presents the number of statistically significant outcomes expected in 100  $t$ -tests (using a 0.05 significance threshold and a nondirectional research

**Table 8.4**

Number of statistically significant outcomes expected in 100 two-sample t-tests for six different scenarios (100 cases in each sample)

SCENARIO	$\mu_1$	$\mu_2$	$\sigma$	$d$	EXPECTED SIGNIFICANT OUTCOMES
(a) Means differ; standard deviations constant					
1	0.3	0.5	2	-0.10	10
2	0.3	0.9	2	-0.30	56
3	0.3	1.3	2	-0.50	94
(b) Means constant; standard deviations differ					
4	0.3	0.5	0.5	-0.40	80
5	0.3	0.5	1	-0.20	29
6	0.3	0.5	2	-0.10	10

hypothesis—resulting in a two-tailed test), each with 100 cases per sample, and illustrates six different scenarios. In the first three scenarios, the mean differences between the two populations are varied and the standard deviations for the populations are held constant. In the last three scenarios, the mean differences are held constant and the standard deviations differ.

As Table 8.4 shows, the largest number of statistically significant outcomes is expected in either the comparisons with the largest differences between mean scores or the comparisons with the smallest standard deviations. As the differences between the population means grow (scenarios 1, 2, and 3), so too does the likelihood of obtaining a statistically significant result. Conversely, as the population standard deviations of the comparisons get larger (scenarios 4, 5, and 6), the expected number of significant outcomes decreases.

As this exercise illustrates, there is a direct relationship between the two components of effect size and statistical power. Studies that examine populations in which there is a larger effect size will, all else being equal, have a higher level of statistical power. Importantly, the relationship between effect size and statistical power is unrelated to the significance criteria we use in a test. In this sense, effect size allows for increasing the statistical power of a study (and thus reducing the risk of Type II error) while minimizing the risk of Type I error (through the establishment of rigorous levels of statistical significance).

Although effect size is often considered the most important component of statistical power, it is generally very difficult for the researcher to manipulate in a specific study (Lipsey 1990). Ordinarily, a study is initiated in order to determine the type and magnitude of a relationship that exists in a population. In many cases, the researcher has no influence at all over the raw differences or the variability of the scores on the measures examined. For example, a researcher who is interested in identifying whether male and female police officers have different attitudes toward corruption may

have no idea prior to the execution of a study the nature of these attitudes or their variability. It is then not possible for the researcher to estimate the nature of the effect size prior to collecting and analyzing data—the effect size may be large or small, but it is not a factor that the researcher is able to influence.

In contrast, in evaluation research—in which a study attempts to assess a specific program or intervention—the researcher may have the ability to influence the effect size of a study and thus minimize the risk of making a Type II error. There is recognition, for example, of the importance of ensuring the strength and integrity of criminal justice interventions (Petersilia 1989). Moreover, many criminal justice evaluations fail to show a statistically significant result simply because the interventions are too weak to have the desired impact or the outcomes are too variable to allow a statistically significant finding (Weisburd 1991; Weisburd et al. 2003).

Statistical power suggests that researchers should be concerned with the effect size of their evaluation studies if they want to develop a fair test of the research hypothesis. First, the interventions should be strong enough to lead to the expected differences in the populations under study. Of course, the larger the differences expected, the greater the statistical power of an investigation. Second, interventions should be administered in ways that maximize the homogeneity of outcomes. For example, interventions applied differently to each subject will likely increase the variability of outcomes and thus the standard deviation of those scores. Finally, researchers should recognize that the heterogeneity of the subjects studied (and thus the heterogeneity of the populations to which they infer) will often influence the statistical power of their tests. Different types of people are likely to respond in different ways to treatment or interventions. If they do respond differently, the variability of outcomes will be larger, and thus, the likelihood of making a Type II error will increase.

As a caution, we note that a wide range of research in criminology and criminal justice has increasingly made use of archival datasets that result in researchers analyzing populations rather than samples. Examples of this would include studies that rely on archival data on all punishment decisions made in the U.S. Federal District Courts or census data on all prisoners in a state on a specific date. A number of years ago, Maltz (1994) noted that the increased frequency with which populations are analyzed calls into question many of the assumptions about performing tests for statistical significance. Put simply, the analysis of population data implies no need for statistical significance testing, since the researcher is not trying to generalize from a sample to a population. Clearly, issues of statistical power are not relevant when we analyze a population: Setting a significance level makes

little sense, the number of cases in the dataset is as large as it possibly can be, and the effect size is simply what is observed (excluding measurement error).

## **Estimating Statistical Power and Sample Size for a Statistically Powerful Study**

---

A number of texts have been written that provide detailed tables for defining the statistical power of a study.<sup>3</sup> All of these texts also provide a means for computing the size of the sample needed to achieve a given level of statistical power. In both cases—the estimation of statistical power or the estimation of necessary sample size—assumptions will need to be made about effect size and level of statistical significance desired. There are also a large number of computer programs that allow you to develop estimates of statistical power. The following discussion provides a basic illustration for how to compute estimates of statistical power. (The computations reported in the following discussion have been performed with a variety of statistical software tools, several of which are freely available. More detail on several easily accessible resources to compute power estimates is provided in the computer problems section at the end of this chapter.)

The most common application of statistical power analysis in criminology and criminal justice research has been to compute the sample size needed to achieve a statistically powerful study (generally at or above 80%). As noted above, we need to be cautious about simply increasing the size of the sample, since a larger sample can affect other important features of statistical power. Thus, in using increased sample size to minimize Type II error, we must consider the potential consequences that larger samples might have on the nature of interventions or subjects studied, particularly in evaluation research. Nonetheless, sample size remains the tool most frequently used for adjusting the power of studies, because it can be manipulated by the researcher and does not require changes in the significance criteria of a test.

To define how many cases should be included in a study, we must conduct power analyses before the study is begun, generally referred to as prospective or *a priori* power analysis, and where our attention has been focused thus far in this chapter. Some authors have advocated the use of power analysis to evaluate whether studies already conducted have acceptable levels of statistical power, based on the sample statistics, referred to as retrospective or *post-hoc* power analysis. Although there is much

---

<sup>3</sup>Among some of the more widely used examples are Cohen (1988), Kraemer and Thiemann (1987), Lipsey (1990), and Murphy and Myors (2003).

agreement about the utility of prospective power analysis, there is little consensus about the appropriateness of retrospective power analysis.<sup>4</sup> The widespread use of secondary data sources in the study of crime and criminal justice further complicates the interpretation of results from a statistical power analysis. Since it is not possible for researchers to augment the original study's sample, results from a power analysis will still be informative in the sense that the results will indicate to the researchers using these data sources what the archived data set can and cannot tell them about the statistical relationships they may be most interested in.

To define the sample size needed for a powerful study, we must first clearly define each of the components of statistical power other than sample size. These include as follows:

1. The statistical test
2. The significance level
3. The research hypothesis (whether directional or nondirectional)
4. The effect size

The first three of these elements should be familiar, since they are based on common assumptions made in developing any statistical test. The statistical test is chosen based on the type of measurement and the extent to which the study can meet certain assumptions. For example, if we want to compare three sample means, we will likely use analysis of variance as our test. If we are comparing means from two samples, we will likely use a two-sample *t*-test. If we are interested in the unique effects of a number of independent variables on a single interval-level dependent variable, we will likely use OLS regression and rely on *t*-tests for the individual coefficients and *F*-tests for either the full regression model or a subset of variables from the full model.

To calculate statistical power, we must also define the significance level of a test and its research hypothesis. By convention, we generally use a 0.05 significance threshold, and thus, we are likely to compute statistical power estimates based on this criterion. The research hypothesis defines whether a test is directional or nondirectional. When the statistical test allows for it, we will typically choose a nondirectional test to take into account the different types of outcomes that can be found in a study (Cohen 1988). If we were evaluating an existing study, we would use the decisions as stated by the authors in assessing that study's level of statistical power.

The fourth element, defining effect size, is perhaps the most difficult component. If we are trying to estimate the magnitude of a relationship in the population that has not been well examined in the past, how can we

---

<sup>4</sup>For an example, see the exchange between Hayes and Steidl (1997) and Thomas (1997).

estimate the effect size in the population? It may be useful to reframe this criterion. The purpose of a power analysis is to see whether our study is likely to detect an effect of a certain size. Usually, we define that effect in terms of what is a meaningful outcome in a study. A power analysis, then, tells us whether our study is designed in a way that is likely to detect that outcome (i.e., reject the null hypothesis on the basis of our sample statistics). This is one of the reasons why statistical power is sometimes defined as **design sensitivity** (Lipsey 1990). It assesses whether our study is designed with enough sensitivity to be likely to reject the null hypothesis if an effect of a certain size exists in the population under study.

The task of defining the effect size has been made easier by identifying broad categories of effect size that can be compared across studies. Cohen's suggestions have been the most widely adopted by other researchers and simply refer to classifying effect sizes as small, medium, and large (Cohen 1988). The numeric value associated with an effect size classified as small, medium, or large is contingent on the specific statistical test being considered. For example, if our focus is on a difference of means test for two independent samples, then Cohen's  $d$  is the effect size of choice and is considered to be a small effect if it is 0.20, a medium effect if it is 0.50, and a large effect if it is 0.80. In contrast, if we are considering the statistical power of an OLS regression model, the standardized effect size estimate is known as  $f^2$  and is considered to be a small effect (if it is 0.02), a medium effect (if it is 0.15), and a large effect (if it is 0.35). Other authors have followed suit and attempted to define similar types of standardized effects for more complex statistical models not addressed in Cohen's work or this book.

The following illustrations turn to a discussion of the computation of statistical power for several common situations in criminology and criminal justice research: difference between two means, one-way ANOVA (i.e., difference between three or more means), correlation coefficient, and ordinary least squares regression.

The computation of statistical power estimates requires the comparison of a sampling distribution under the null hypothesis with a sampling distribution under the alternative or the research hypothesis. The sampling distribution under the research hypothesis is referred to as a noncentral distribution, and the noncentrality parameter shown in Eq. (8.1) determines where in this distribution the null hypothesis can be rejected—that is where in this distribution we would decide that an effect is statistically significant.

For each of the statistical tests discussed below, we describe both the standardized effect and the noncentrality parameter and explain how to use these values to estimate the statistical power of a sample as well as the size of sample needed to meet a target level of statistical power.

### Difference of Means Test

As discussed above, we have pointed to the difference of means test as an example for many of the points we wanted to make about statistical power. This estimation depends on Cohen's  $d$  effect size, also called the standardized mean difference, shown in Eq. (8.2).

The noncentrality parameter,  $\delta$ , is the location of the mean for the noncentral  $t$ -distribution associated with a specific value of  $d$ , the effect size, computed as:

$$\delta = d \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

where  $n_1$  and  $n_2$  are the sample sizes associated with each of the two means.

Referring back to Figs. 8.1 and 8.2, the noncentrality parameter tells us the difference between the two distributions, the sampling distribution under the null and the sampling distribution for our presumed population effect size. To determine  $\beta$  and by extension statistical power, or  $1 - \beta$ , we need to determine the location of the vertical line in these figures and then the portion of the population distribution on the right of this vertical line. In all of these examples, we are assuming a positive value for the effect size. If you are working with negative effects, everything would be flipped but otherwise the same. The location of the line can be found with the following equation:

$$t_\beta = \delta - t_{CV}$$

**Equation 8.3**

where  $t_{CV}$  is the critical value for the  $t$  based on the  $\alpha$  level and whether the test is two-tailed or one-tailed. Statistical power is then determined by finding the area under the curve of the resulting  $t$ -value with the appropriate degrees of freedom (the same as used for the  $t$ -critical value) to the right of this value, as shown in Figs. 8.1 and 8.2.

To illustrate the computation of a statistical power estimate, suppose that we want to assess the effectiveness of a treatment program for drug offenders. Our design calls for random assignment of 100 cases to each group. We expect the program to be effective at reducing recidivism in the treatment group and so can assume a one-tailed  $t$ -test with a significance level of 5%. What is the statistical power of our design for detecting standardized effects at the small ( $d = 0.2$ ), medium ( $d = 0.5$ ), and large ( $d = 0.8$ ) levels?

For all three scenarios, the critical  $t$ -value will be 1.653, based on a one-tailed test with a significance level of 0.05 and  $df = N - 2 = 198$ . For a small effect, the non-centrality parameter  $\delta$  is 1.414, as shown below:

$$\delta = 0.2 \sqrt{\frac{100 \times 100}{100 + 100}} = 1.414.$$

The  $t_\beta$  is the difference between  $\delta$  and our critical  $t$ -value, or  $-0.239$ , as shown below:

$$t_\beta = 1.414 - 1.653 = -0.239.$$

The one-tailed significance for this  $t$  is 0.406. This is the area in the left tail of this distribution, or  $\beta$ . This provides us with an estimate for risk of making a Type II error of  $\beta = 0.594$ , suggesting that we have a probability of 59.4% of making a Type II error and fail to reject the null hypothesis when it is false. (We will illustrate how to obtain estimates of the noncentral  $t$  from statistical packages in the exercises). The corresponding estimate of statistical power is  $1 - 0.594 = 0.406$ . Note that we always use the one-tailed  $p$ -value for  $t_\beta$  even if the critical value  $t$  was based on a two-tailed test. Whether the power analysis is one-tailed or two-tailed is based on the latter and not the former. Substantively, this result suggests that if we have only 100 cases in each group, our probability of rejecting the null hypothesis when it is false is only about 40.6%. Using the same approach, we can estimate power for other effect sizes. In regard to a medium effect size,  $\delta = 3.536$ ,  $\beta = 0.030$ , and power = 0.970. For a large effect size,  $\delta = 5.657$ ,  $\beta < 0.0001$ , and power  $> 0.9999$ . Putting these results together indicates that our design with 100 cases assigned to each group provides a high level of statistical power for detecting medium effects and larger but an inadequate level of power for detecting small effects.

Alternatively, we may be interested in determining the sample size needed to provide us with a statistical power estimate of 80% for each of the three effect sizes: small, medium, and large. This can be done either through following through on calculations as we have done here to identify when a power level is achieved, or using computer programs that simply provide such estimates. In the case of a small effect, we find that we need a total sample of 620 cases—310 in each group—to assure us that we will be able to reject the null hypothesis when it is false about 80% of the time. To achieve a power estimate of 80% for a medium effect, we only need 102 cases (51 in each group). For a large effect, the sample size drops to 40 (20 in each group).

## ANOVA

For a simple one-way ANOVA, where we are looking only at the differences across three or more means with an equal sample size in each group, the standardized effect size  $f$  is defined as

$$f = \frac{\sigma_m}{\sigma_e}, \quad \text{Equation 8.4}$$

where  $\sigma_m$  is the standard deviation of the means and  $\sigma_e$  is the standard deviation within groups, based on the error or residual variance. The former is computed as

$$\sigma_m = \sqrt{\sum_{t=1}^k \frac{(m_t - m)^2}{k}},$$

where  $k$  is the number of groups,  $m$  is the grand mean, and  $m_t$  represents each of the group means with  $n_1 = n_2 = \dots = n_k$ .

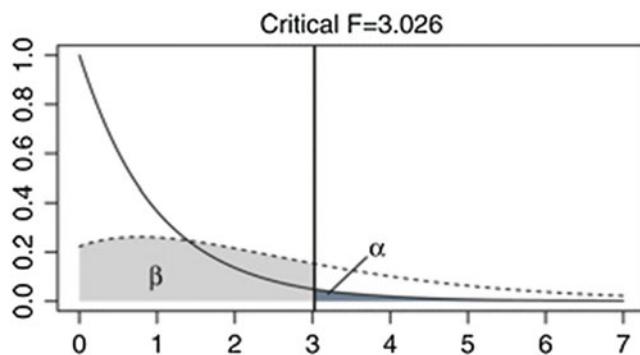
The noncentrality parameter  $\lambda$  for the  $F$ -distribution is

$$\lambda = n \times f^2 \quad \text{Equation 8.5}$$

where  $f^2$  refers to the square of the standardized effect size ( $f$ ) and  $n$  refers to the total sample size.

As an illustration of the calculation of statistical power estimates for a fixed-effects ANOVA model, assume that we have three groups, each with 100 cases participating in an experiment aimed at reducing recidivism among violent offenders: a control group and two different kinds of treatment groups. Assume that the significance level has been set at 5%. What is the level of statistical power of our design for detecting standardized effects at the small ( $f = 0.1$ ), medium ( $f = 0.25$ ), and large ( $f = 0.4$ ) levels?

For each of the three scenarios, the critical value of the  $F$ -statistic is 3.026 ( $df_1 = 2$ ,  $df_2 = 297$ ). For a small effect, the noncentrality parameter  $\lambda$  is 3 ( $0.1^2 \times 300 = 3$ ). Unlike the  $t$ -test where we could compute a  $t_\beta$  and determine power from that, the noncentral  $F$ -distribution is more complicated. However, most of all the implementations of this distribution in software programs include the ability to obtain the  $p$ -value for the upper tail based on a given  $F$  critical value, a noncentrality parameter, and the numerator and denominator degrees of freedom. For our illustration, doing so provides us with an estimate for risk of making a Type II error of  $\beta = 0.681$  (the lower tail), suggesting that we have a probability of 68.1%

**Figure 8.3***Graphical representation for power analysis in a one-way ANOVA*

of making a Type II error and fail to reject the null hypothesis when it is false. The corresponding estimate of statistical power is  $1 - 0.681 = 0.319$  (the upper tail), meaning that we have only a 31.9% chance of rejecting the null hypothesis when it is false. This result is presented graphically in Fig. 8.3. Similar to the layout in Fig. 8.1, the vertical line indicates the critical  $F$  of 3.026, the solid line the  $F$ -distribution, and the dashed line the noncentral  $F$ -distribution. Below the two curves, represented by two different shades of gray, alpha is indicated by the darker shading in the right tail of the  $F$ -distribution beyond the critical value, and beta is represented by the lighter shaded area to the left of the critical value and under the noncentral  $F$ -distribution.

For the medium and large effect size analyses, the  $F$ -distribution remains the same, but the noncentral  $F$ -distribution is shifted further to the right. For the medium effect size,  $\lambda = 18.75$ ,  $\beta = 0.022$ , and power = 0.978. The large effect size has  $\lambda = 48$ ,  $\beta < 0.0001$ , and power > 0.9999. Similar to the previous analysis comparing the means for only two groups, our research design with 100 cases assigned to each of the three groups provides a high level of statistical power for detecting medium and large effects but an inadequate level of power for detecting small effects.

If our concern is focused on the size of the sample needed for a power level of 80% for each of the three effect sizes—small, medium, and large—then we would again proceed in the same way as in the two-sample  $t$ -test. To have an 80% chance of detecting a small effect ( $f = 0.10$ ), we would need a sample of 969 cases (323 in each group). For the medium effect, we would need only 159 cases (53 in each group), and for the large effect, only 66 cases (22 in each group).

### Correlation

To test the statistical power of a correlation coefficient, we use the correlation coefficient ( $r$ ) as the standardized effect size. As with Cohen's  $d$ , power for a correlation is based on the  $t$  probability distribution. Thus, the noncentrality parameter,  $\delta$ , is the  $t$  associated with the population correlation coefficient. The computation of the noncentrality parameter  $\delta$  for the correlation coefficient is

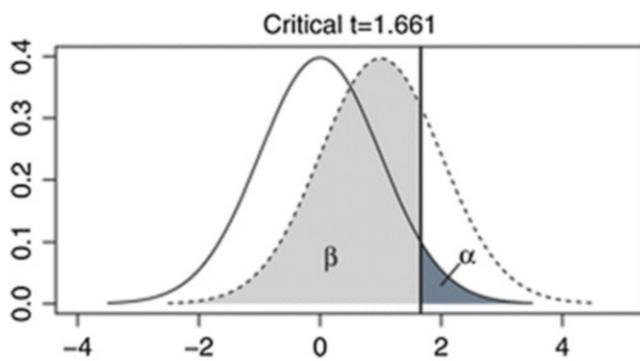
$$\delta = \frac{r\sqrt{(n-2)}}{\sqrt{1-r^2}} \quad \text{Equation 8.6}$$

where  $r$  is the presumed population correlation and  $n$  is the sample size.

We can again illustrate the calculation of statistical power for correlations by assuming that we have 100 observations that would allow us to compute a correlation between two variables. For example, suppose we interview a random sample of police officers and are interested in the correlation between the number of years on the police force and a scale that measured hostility toward judges. We might expect that more years on the police force will have a positive correlation with hostility toward judges, implying that we can conduct a one-tailed  $t$ -test of statistical significance. As with the preceding examples, assume that the level of statistical significance is 5%. What is the level of statistical power of our design for detecting standardized effects at the small ( $r = 0.1$ ), medium ( $r = 0.3$ ), and large ( $r = 0.5$ ) levels?

The one-tailed critical  $t$ -value for all three scenarios is 1.661, based on  $df = n - 2 = 98$ . For a small effect size ( $r = 0.1$ ), the noncentrality parameter is  $\delta = 0.995$ . The difference between  $\delta$  and the  $t$  critical value is  $-0.666$ . The area under the curve to the left of this  $t$ -value provides us with an estimate for risk of making a Type II error of  $\beta = 0.746$ , suggesting that we have a probability of 74.6% of making a Type II error and would fail to reject the null hypothesis when it is false. The corresponding estimate of statistical power is 0.254, indicating that we would only reject the null hypothesis when it was false about 25% of the time. Figure 8.4 presents these results graphically. The statistical power analysis of the medium effect indicates that  $\delta = 3.145$ ,  $\beta = 0.070$ , and power = 0.930. The large effect shows an even greater level of statistical power, where  $\delta = 5.774$ ,  $\beta < 0.0001$ , and power = 0.9999.

The sample size required to detect each of the three effect sizes—small, medium, and large—with a statistical power of 80% again requires the use of the  $t$ -distribution. To achieve a power level of 80% for a small effect ( $r = 0.1$ ), a sample of 614 cases would be needed. For the medium effect ( $r = 0.3$ ), the required number of cases drops to 64, while for the large effect

**Figure 8.4***Graphical representation for power analysis of a correlation*

( $r = 0.5$ ), only 21 cases are required to have an 80% chance of rejecting the null hypothesis when it is false.

### **Least Squares Regression**

The statistical power analysis of least squares regression can take two different, but related, forms. One question asks about the ability to detect whether a regression model—a single dependent variable and two or more independent variables—has a statistically significant effect on the dependent variable. This means that the null hypothesis is focused on whether the regression model in its entirety has an effect on the dependent variable. A second question asks about the ability to detect the effect of a single variable or a subset of variables added to a regression model. This addresses the more common substantive question in much of the published research: Once the other relevant independent and control variables have been taken into account statistically, does variable  $X$  add anything to the overall model?

Whether we are analyzing the full model or a subset of the full model, the standardized effect size (denoted as  $f^2$ ) is based on either the  $R^2$  for the full model or the partial  $R^2$  for the subset of variables we are interested in analyzing. Specifically,

$$R^2 = \frac{f^2}{1 + f^2} \quad \text{Equation 8.7}$$

As indicated above, Cohen's recommendations for small, medium, and large standardized effect sizes in a regression model are 0.02, 0.15, and 0.35, respectively. To provide some context to these values, an  $f^2$  value of 0.02

corresponds to an  $R^2$  of 0.0196, while  $f^2 = 0.15$  implies that  $R^2 = 0.13$ , and  $f^2 = 0.35$  implies that  $R^2 = 0.26$ . Statistical power analysis for least squares regression uses the  $F$ -distribution.

As noted in the discussion of statistical power analysis for ANOVA models, the noncentrality parameter  $\lambda$  for the  $F$ -distribution is

$$\lambda = n \times f^2$$

Determining statistical power proceeds in the same fashion as with the one-way ANOVA. The area to the left of the  $F$  critical value from an  $F$ -distribution with the above noncentrality parameter provides the estimate of  $\beta$ .

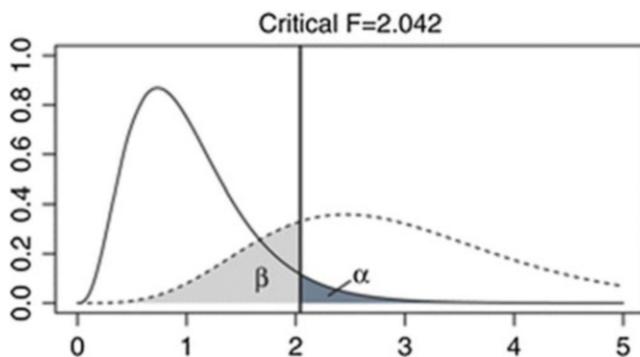
To assess the statistical power for the full regression model, consider the following simple example. Suppose that we are interested in the effects of various case and defendant characteristics on the amount of bail required by a court. Typical analyses of bail decisions would consider some of the following characteristics (as well as others not listed): (1) severity of the prior record, (2) severity of the current offense, (3) number of counts of the current offense, (4) type of attorney, (5) whether the defendant was under criminal justice supervision at the time of the current offense, (6) age of the defendant, (7) race of the defendant, and (8) gender of the defendant. This provides us with a regression model with eight independent variables.

As a point of illustration, we may want to estimate the statistical power of the regression model assuming that we have a sample of only 100 cases and have set a significance level of 5%, giving us a crucial  $F$ -value of 2.042. For the small effect size ( $f^2 = 0.02$ ), we have noncentrality parameter  $\lambda = 2.0 (= 0.02 \times 100)$ . We then find  $\beta = 0.876$ , meaning that with only 100 cases, we have a probability of making a Type II error of just under 88%. Alternatively, the estimate of statistical power is 0.124, meaning that we have a probability of only 12.4% of rejecting the null hypothesis when it is false. The results for the medium effect ( $f^2 = 0.15$ ) appear in Fig. 8.5, which is based on  $\lambda = 15.0$ ,  $\beta = 0.242$ , and power = 0.758. This is still an inadequate level of power but is much closer to the target of 80%. For the large effect ( $f^2 = 0.35$ ),  $\lambda = 35.0$ ,  $\beta = 0.007$ , and power = 0.993, which is well beyond the desired level of 80%.

For a regression model with eight independent variables, what sample size is required to achieve a statistical power level of 80% for detecting effects at the small ( $f^2 = 0.02$ ), medium ( $f^2 = 0.15$ ), and large ( $f^2 = 0.35$ ) levels? For the small effect, we would require a sample of 759 cases to achieve a power level of 80%. For the medium and large effects, we would require samples of 109 and 52 cases, respectively. The number of cases required to detect a statistically significant effect at either the medium or the large effect level may strike many readers as small. It is important to keep in mind that we have only been assessing the full model—the number of cases

**Figure 8.5**

*Graphical representation for power analysis of a regression model (with eight independent variables)*



required for detecting individual effects will tend to be different than the number of cases required for detecting whether the full model is significant.

The assessment of statistical power for a single independent variable or a small subset of independent variables proceeds in much the same way as the analysis for the full model. The key difference is in the degrees of freedom required for the  $F$ -distribution. In the case of a single independent variable, the numerator  $df = 1$ , while the denominator  $df$  remains the same as in the full model. For a subset of independent variables, the numerator  $df =$  the number of variables in the subset (the denominator  $df$  remains the same).

If we return to the bail example above, the analysis of statistical power for any one of the independent variables will be identical. We continue to keep the sample size at 100 cases, the level of statistical significance at 5%, and the definition of small, medium, and large effects the same as before. For the small effect ( $f^2 = 0.02$ ),  $\lambda = 2.0$ ,  $\beta = 0.712$ , and power = 0.288, meaning that we would only be able to reject the null hypothesis of no relationship between the independent and dependent variables about 28.8% of the time. For the medium effect ( $f^2 = 0.15$ ),  $\lambda = 15.0$ ,  $\beta = 0.031$ , and power = 0.969, while for the large effect ( $f^2 = 0.35$ ),  $\lambda = 35.0$ ,  $\beta < 0.0001$ , and power > 0.9999.

Similarly, we may be interested in assessing the statistical power of a subset of variables. For example, in the bail example, the subset of demographic characteristics (age, race, and gender) may be important in testing some aspect of a theory predicting differential treatment of defendants within the courts. We find a similar pattern to the results. For the small effect ( $f^2 = 0.02$ ),  $\lambda = 2.0$ ,  $\beta = 0.814$ , and power = 0.186, again indicating a low level of statistical power for detecting a statistically significant

relationship between demographic characteristics and bail amount. For the medium effect ( $f^2 = 0.15$ ),  $\lambda = 15.0$ ,  $\beta = 0.095$ , and power = 0.905, while for the large effect ( $f^2 = 0.35$ ),  $\lambda = 35.0$ ,  $\beta = 0.001$ , and power = 0.999.

Sample size calculations work in the same way as for the full model. If we hope to achieve a power level of 80%, what size sample is necessary to detect small, medium, and large effects for either single variables or subsets of variables? Continuing the bail example, we assume that there are eight independent variables. For the single variable, the number of cases required to detect a small effect with a probability of 80% is 395. A medium effect requires only 55 cases, while a large effect requires only 26 cases. It is worth noting that sample size calculations for single variable effects are not affected by the number of variables included in the full regression model.

In practice, many of the individual effects that researchers are trying to assess in their multivariate models will tend toward the small effect size. For example, much survey research aimed at trying to explain attitudes toward a particular topic will often incorporate 10–20 independent variables and have a full model  $R^2$  typically between 0.15 and 0.20. This implies that many of the effects of individual variables will tend to be quite small in magnitude. In order for an analysis to detect a statistically significant relationship, a large sample becomes necessary.

## Summing Up: Avoiding Studies Designed for Failure

---

The statistical power of a test can be compared to the sensitivity of a radiation meter. A very sensitive meter will be able to identify even the smallest deposits of radioactivity. A meter that is not very sensitive will often miss such small deposits, although it likely will detect very large radiation signals from areas rich in radioactivity. Similarly, a statistically sensitive study will be able to identify even small effects. This is usually because the researcher has increased the sample size of the study to make it more statistically powerful. Conversely, a study that has little sensitivity is unlikely to yield a statistically significant result even when relatively large differences or program impacts are observed. Such studies may be seen as *designed for failure*, not because of inadequacies in the theories or the programs evaluated, but because the investigator failed to consider statistical power at the outset of the study.

You might question why we would even bother to define the size of the sample needed for statistically powerful studies. Why not just collect 1000 or more cases in every study and be almost assured of a statistically powerful result? The simple answer is that although you should try to sample as many cases as you can in a study, there are generally constraints in developing samples. These constraints may be monetary, related to time, or associated with access to subjects. It is often important to know the

minimum number of cases needed to achieve a certain threshold of statistical power so that you can try, within the constraints of the research setting, to reach an adequate level of statistical power in your study. It is also important to be able to assess whether studies that you read or evaluate were designed in such a way that they are reasonable tests of the hypotheses presented. If such studies are strongly underpowered, then you should have much less confidence in findings that do not support the research hypothesis.

## Chapter Summary

---

A statistically powerful test is one for which there is a low risk of making a Type II error. **Statistical power** can be defined as 1 minus the probability of falsely accepting the null hypothesis. A test with a statistical power of 0.90 is one for which there is only a 10% probability of making a Type II error. If the power of a test is 0.10, the probability of a Type II error is 90%. A minimum statistical power level of at least 0.50 is recommended. However, it is generally accepted that in better studies, the level of statistical power will be at least 0.80. A study with a low level of statistical power can be described as designed for failure, as it is unlikely to produce a statistically significant result even if the expected effect exists in the population under study. The statistical power of a study is also defined as its **design sensitivity**.

There are several ways in which statistical power can be maximized. First, we may raise the significance threshold. Doing so, however, also increases the risk of a Type I error. Second, we may limit the direction of the research hypothesis and conduct a one-tailed test. Doing so, though, will necessarily ignore outcomes in the opposite direction. Third, we may try to maximize the **effect size**. The greater the differences between the populations and the smaller the variability of those differences, the larger the population effect size will be. Effect size, however, is usually beyond the control of the researcher. Fourth, we may increase the sample size. A larger sample produces a smaller standard error for the sampling distribution and a larger test statistic. The larger the sample, all else being equal, the greater the chance of rejecting the null hypothesis.

Sample size is generally the most useful tool for maximizing statistical power. A power analysis before a study is begun will define the number of cases needed to identify a particular size effect—small, medium, or large. A power analysis of an existing study will help to identify whether it was well designed to assess the questions that were examined. To identify a small effect size, the overall sample must be very large. For a large effect size, a much smaller sample will suffice.

Statistical power can be estimated by hand, though there are many computer programs that do this for the researcher. Ordinarily, the researcher will try to determine the size of a sample needed to gain a particular level of statistical power. To do this, the statistical test used, significance level, directionality of the research hypothesis, and hypothesized effect size must be defined at the outset.

## Key Terms

---

**Design sensitivity** The statistical power of a research study. In a sensitive study design, statistical power will be maximized, and the statistical test employed will be more capable of identifying an effect.

**Effect size (ES)** A general term of a statistical index of the size of an effect, such as a mean difference, correlation coefficient, or regression coefficient. Typically, these are

standardized, so they no longer depend on the raw units of the outcome measure examined. Effect sizes are a critical component of statistical power analysis.

**Statistical power** One minus the probability of a Type II error. The greater the statistical power of a test, the less chance there is that a researcher will mistakenly fail to reject the null hypothesis.

## Symbols and Formulas

---

ES Effect size

$d$  Cohen's standardized mean difference effect size

$f$  Standardized effect size for one-way ANOVA and OLS regression models

$n$  Sample size

$n_i$  Sample size in group  $i$

$\delta$  Noncentrality parameter for the  $t$ -distribution

$\lambda$  Noncentrality parameter for the  $F$ -distribution

Cohen's  $d$  effect size for the difference between two population means:

$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

The difference between the non-centrality parameter and the critical value of  $t$ :

$$t_\beta = \delta - t_{\text{cv}}$$

Noncentrality parameter for the  $t$ -distribution for a Cohen's  $d$ :

$$\delta = d \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

Cohen's  $f$  effect size for a one-way ANOVA with three or more means:

$$f = \frac{\sigma_m}{\sigma_e}$$

Noncentrality parameter for the  $F$ -distribution for a Cohen's  $f$ :

$$\lambda = n \times f^2$$

Noncentrality parameter for the  $t$ -distribution for a correlation coefficient:

$$\delta = \frac{r \sqrt{(n - 2)}}{\sqrt{1 - r^2}}$$

The proportion of variation explained:

$$R^2 = \frac{f^2}{1 + f^2}$$

## Computer Exercises

In contrast to many of the other computer exercises in this text, the computation of statistical power estimates is not easily performed in any of the large stand-alone statistical packages. There are a variety of software packages available for computing statistical power as well as a number of websites that host power calculators for a wide range of statistical tests. All of the analyses presented in this chapter were performed with G\*Power (version 3.1.7). G\*Power 3 is freely available to download from the *Institut für Experimentelle Psychologie at Universität Düsseldorf* (<http://www.psych.uni-duesseldorf.de/abteilungen/aap/gpower3/>). G\*Power

3 is a specialized package devoted to statistical power estimation and offers a wide range of tests beyond those discussed here. G\*Power 3 also features the simple creation of powerful graphs that will plot power estimates across a range of sample sizes, effect sizes, and statistical significance levels. The figures presented in this chapter are similar to what are produced with G\*Power 3. Faul et al. (2007) provide a useful overview of the capabilities of G\*Power 3.

Power and Precision v. 2.0 is a commercially available software package designed to compute power estimates for a wide range of statistical models in a user-friendly environment (Borenstein et al. 2001). As a commercial software package, its range of capabilities is significantly greater than G\*Power 3. A particularly useful feature is that all of the output—text and graphs—can be easily exported to other programs.

In the case that one simply wants to compute a small number of power estimates without bothering to learn a new software package, a reasonably comprehensive list of Web-based power calculators can be found at <http://statpages.org/#Power>. The list of websites hosting power calculators is categorized by the type of statistical test that the user is searching for—one-sample *t*-test, two-sample *t*-test, correlation, regression, and so on.

On a technical note, it is worth highlighting that there will be slight differences across statistical software packages and power calculators in the estimated sample sizes needed to achieve a given level of statistical power. The primary reason for this appears to be focused on rounding the estimated sample size to an integer, since we cannot sample a fraction of a case in any research study. Some packages round up so that the estimated statistical power is always at least as great as the target entered into the computation. Other packages and calculators will round to the closest integer (regardless of whether it is larger or smaller), so the overall estimate of statistical power may be slightly less than the initial target.

## Stata

### *Two-Sample Difference of Means Test*

In Stata, one- and two-sample difference of means tests are performed with the **sampsi** command:

```
sampsi Mean1 Mean2, sd1(#) sd2(#) n1(#) n2(#) ///
power(#) onesided
```

where Mean1 and Mean2 refer to the expected population means for the two samples being compared, **sd1(#)** and **sd2(#)** refer to the expected standard deviations for each sample (values inserted in the parentheses), **n1(#)** and **n2(#)** refer to the expected number of cases (values inserted in the parentheses) in each sample, **power(#)** is a designated level of power for sample size estimation (the default is a power level of 0.9), and **onesided** indicates that a one-tail test should be used (a two-tail test is the default). Note that the *///* symbol in the command allows commands to span multiple lines in a command “do” file. In the situation where we

are trying to estimate power and assume constant standard deviations and sample sizes across the two samples, this can be simplified to

```
samps1 Mean1 Mean2, sd(#) n(#)
```

Upon entering the command, the output will list all of the assumptions (alpha level, etc.) and then compute the power. For example, the first line of Table 8.3 indicates that the expected number of significant results (out of 100 tests) is 13, meaning that the estimate of power for a situation involving two samples of 35 cases each, a difference of population means of 0.2, a common standard deviation of 1.0 is 0.13 (with rounding). Using the Stata **samps1** command, we would enter the following command:

```
samps1 0 0.2, sd(1) n(35)
```

The use of 0 and 0.2 for the two-sample means is a convenient way to represent the difference. It would make no difference what two numbers we inserted here so long as the difference was 0.2 (try it).

Stata produces the following output:

```
. samps1 0 0.2, sd(1) n(35)
```

Estimated power for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1

and m2 is the mean in population 2

Assumptions:

alpha =	<b>0.0500</b>	(two-sided)
m1 =	<b>0</b>	
m2 =	<b>.2</b>	
sd1 =	<b>1</b>	
sd2 =	<b>1</b>	
sample size n1 =	<b>35</b>	
n2 =	<b>35</b>	
n2/n1 =	<b>1.00</b>	

Estimated power:

```
power = 0.1332
```

The power estimate of 0.1332 would lead us to expect 13 significant results in 100 tests ( $=0.1332 \times 100$ ). The remainder of Table 8.3 can be reproduced simply by changing the expected sample size in the command above from 35 to 100, 200, and 1,000, respectively.

If our interest is in estimating the sample size required to achieve a given level of statistical power, we can use Stata's built-in **power** command. We no longer specify

an argument for sample size, and we add an argument for power. To compute the sample size needed to detect our effect with a power of 0.8 (one-tailed test), we would enter the following command:

```
power twomeans 0 0.2, sd(1) power(.8) onesided
```

### ANOVA

Power and sample size analysis was extended to include methods for analyzing ANOVA models in Stata 13.1. To compute power for a simple one-way ANOVA, the command **power oneway** can be used. The basic components to the *power oneway* command to obtain statistical power are the following:

```
power oneway, n() delta() ngroups() power()
```

where **n(#)** represents the total number of observations, **delta(#)** represents the standardized effect size (the default is  $f = 0.1$ ), **ngroups(#)** is the number of groups being compared, and **power(#)** is the designated power used to calculate sample size(s). For our example above, we computed the power of a one-way ANOVA design with three groups ( $k$ ) and 100 cases in each group for three different effect sizes  $f(0.1, 0.25, \text{ and } 0.4)$ .

The *power oneway* command to compute the power estimate for the small effect (i.e.,  $f = 0.10$ ) is as follows:

```
power oneway, n(300) delta(.1) ngroups(3)
```

Since we specified 100 cases per group (for a total sample size of 300), you will obtain the following output:

```
Estimated power for one-way ANOVA
F test for group effect
Ho: delta = 0 versus Ha: delta != 0
```

Study parameters:

```
alpha = 0.0500
N = 300
N per group = 100
delta = 0.1000
N_g = 3
Var_m = 0.0100
Var_e = 1.0000
```

Estimated power:

```
power = 0.3186
```

If you rerun this command but change the value for  $f$  to include the medium and strong effect sizes, the power estimates reported above will also be reproduced. This can be done by changing **delta(.1)** to **delta(.1 .25 .4)**.

The **power oneway** command can also be used to estimate power for a given effect size over a range of group sizes by modifying the **n(#)** argument to include the minimum sample size to be tested (**min**), maximum sample size to be tested (**max**), and the increments to be tested (**by**): **n(min (by) max)**. For example, to determine the power to detect a small effect ( $f = 0.1$ ) in a study with three groups starting at 100 observations per group to 500 observations per group (in increments of ten observations per group), it would be specified as follows:

```
. power oneway, n(300 (30) 1500) delta(.1) ngroups(3)
```

Estimated power for one-way ANOVA

F test for group effect

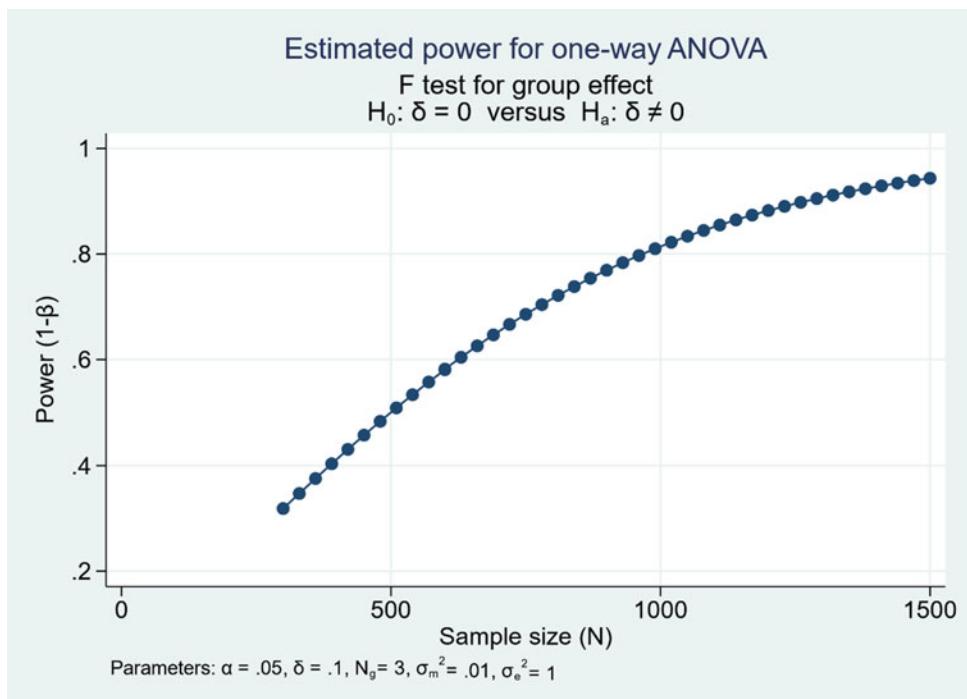
Ho: delta = 0 versus Ha: delta != 0

alpha	power	N	N_per_group	delta	N_g	Var_m	Var_e
.05	.3186	300	100	.1	3	.01	1
.05	.3473	330	110	.1	3	.01	1
.05	.3756	360	120	.1	3	.01	1
.05	.4035	390	130	.1	3	.01	1
.05	.4309	420	140	.1	3	.01	1
.05	.4578	450	150	.1	3	.01	1
.05	.484	480	160	.1	3	.01	1
.05	.5096	510	170	.1	3	.01	1
.05	.5344	540	180	.1	3	.01	1
.05	.5585	570	190	.1	3	.01	1
.05	.5818	600	200	.1	3	.01	1
.05	.6042	630	210	.1	3	.01	1
.05	.6259	660	220	.1	3	.01	1
.05	.6467	690	230	.1	3	.01	1
.05	.6667	720	240	.1	3	.01	1
.05	.6858	750	250	.1	3	.01	1
.05	.7041	780	260	.1	3	.01	1
.05	.7216	810	270	.1	3	.01	1
.05	.7382	840	280	.1	3	.01	1
.05	.7541	870	290	.1	3	.01	1
.05	.7692	900	300	.1	3	.01	1
.05	.7836	930	310	.1	3	.01	1
.05	.7972	960	320	.1	3	.01	1
.05	.8101	990	330	.1	3	.01	1
.05	.8223	1,020	340	.1	3	.01	1

.05	.8338	1,050	350	.1	3	.01	1
.05	.8447	1,080	360	.1	3	.01	1
.05	.855	1,110	370	.1	3	.01	1
.05	.8646	1,140	380	.1	3	.01	1
.05	.8738	1,170	390	.1	3	.01	1
.05	.8823	1,200	400	.1	3	.01	1
.05	.8904	1,230	410	.1	3	.01	1
.05	.898	1,260	420	.1	3	.01	1
.05	.9051	1,290	430	.1	3	.01	1
.05	.9117	1,320	440	.1	3	.01	1
.05	.918	1,350	450	.1	3	.01	1
.05	.9238	1,380	460	.1	3	.01	1
.05	.9293	1,410	470	.1	3	.01	1
.05	.9344	1,440	480	.1	3	.01	1
.05	.9391	1,470	490	.1	3	.01	1
.05	.9436	1,500	500	.1	3	.01	1

If you would prefer that your output be presented in a figure, you can add the **graph** argument:

```
power oneway, n(300 (30) 1500) delta(.1) ngroups(3) graph
```



You can also directly estimate the sample size required for a designated level of statistical power for a one-way ANOVA. This can be done by replacing **n(#)** with the **power(#)** argument, as follows:

```
. power oneway, power(.8) delta(.1) ngroups(3)
```

Estimated sample size for one-way ANOVA

F test for group effect

Ho: delta = 0 versus Ha: delta != 0

Study parameters:

alpha =	<b>0.0500</b>
power =	<b>0.8000</b>
delta =	<b>0.1000</b>
N_g =	<b>3</b>
Var_m =	<b>0.0100</b>
Var_e =	<b>1.0000</b>

Estimated sample sizes:

N =	<b>969</b>
N per group =	<b>323</b>

### *Correlation*

There is one user-written procedure that we are aware of for computing power estimates of correlation coefficients in Stata. The command is **sampsrho**, which bases power calculations on converting the correlation coefficient with the Fisher  $z$  formula and then using the normal distribution (instead of a  $t$ -distribution for the untransformed correlation coefficient). This command can be installed with the following command:

```
ssc install sampsrho
```

The basic structure of the **sampsrho** command is

```
sampsrho, null(#) alt(#) n(#) power(#) ///
solve() alpha(#) onesided
```

where **null(#)** specifies the value of the correlation for the null hypothesis (default is 0), **alt(#)** specifies the alternative hypothesis value of the correlation (default is 0.5), **n(#)** specifies the sample size (default is 100), **power(#)** indicates the desired level of power (default is 0.9), **solve()** notes whether to solve for sample size (*n*, which is the default) or power, **alpha(#)** specifies the alpha level if different than

0.05 (the default), and **onesided** indicates that a one-tail test is to be performed (a two-tailed test is the default).

To replicate the values above in our analysis of power estimates for correlation coefficients for a sample size of 100, we would enter the following command in Stata to estimate the power to detect a weak correlation:

```
sampsrho, solve(power) n(100) alt(0.1) onesided
```

The estimated power is 0.260, very nearly the same as the estimate produced using the correlation coefficient and the *t*-distribution. If you were interested in reproducing the power estimates for the medium and strong effects, you would just need to change the value of **alt(#)** to **alt(0.3)** and **alt(0.5)**, respectively.

In a similar way, we can estimate the sample size needed to achieve a designated level of statistical power for a hypothesized effect size by making just a few changes to the **sampsrho** command. For example, if we wanted to estimate the sample size needed to detect a medium correlation (i.e.,  $r = 0.3$ ) with a power level of 0.80, we would omit the sample size and solve options but insert **power(0.8)** and enter the following command:

```
sampsrho, alt(0.3) power(0.8) onesided
```

We find that the estimated sample size is 67.53. Since we cannot find a fraction of a case, we would typically round up to 68 in this case. The rationale, as we noted above, in rounding up is to ensure that a power level of no less than our target (e.g., 0.80) is achieved. Note, too, that the sample size estimated here (68) is slightly larger than that estimated above in the text (64). The difference is entirely due to the use of the Fisher-transformed value of the correlation and use of the normal distribution and is to be expected.

### *OLS Regression*

Power and sample size analysis can be performed for an  $R^2$  test in a multiple linear regression by using the **power rsquared** command (Stata 15 or later). The basic structure to the **power rsquared** command is as follows:

```
power rsquared r2r r2f, ntested(#) ncontrol(#) n(#) power(#)
```

where **r2r** is the  $R^2$  for the reduced (null) model, **r2f** is the hypothesized value of  $R^2$  expected, **ntest(#)** refers to the number of independent variables being tested, **ncontrol(#)** refers to the total number of control variables included in the regression model, **n(#)** refers to the number of observations, and **power(#)** is the desired statistical power. Alpha is assumed to be 0.05.

To reproduce the results reported above for power in OLS regression for a weak effect (i.e.,  $R^2 = 0.02$ ), we would use the following syntax:

```
power rsquared 0 .02, ntested(8) ncontrol(8) n(100)
```

Note that the value for **r2r** is entered as 0—this is the expected value of  $R^2$  without any of the independent variables included in the analysis. The power estimate reported by Stata is 0.1247, very close to the result of 0.124 reported above. Results for the moderate ( $R^2 = 0.13$ ) and strong ( $R^2 = 0.26$ ) effect sizes are obtained by simply altering the value of **r2f** in the **power rsquared** command.

To compute the sample size needed to achieve a designated level of statistical power, we would omit the **n(#)** option but insert an option for **power(#)**:

```
power rsquared 0.02, ntested(8) ncontrol(8) power(.8)
```

We find that the estimated sample size needed to detect a weak effect ( $R^2 = 0.02$ ) is 744, which is different from the value of 759 reported above (you will get a slightly different results in R as well). This is due to the calculation of the standardized effect ( $f^2$ ) and rounding error for the weak effect size.<sup>5</sup>

## R

### *Two-Sample Difference of Means Test*

In R, one- and two-sample difference of means tests are performed with the **pwr.t.test()** function from the *pwr* package:

```
pwr.t.test(n = NULL, d = NULL, sig.level = 0.05,
           power = NULL,
           type = c("two.sample", "one.sample", "paired"),
           alternative = c("two.sided", "less", "greater"))
```

where **n** refers to the number of observations per group being compared, **d** refers to the effect size, **sig.level** refers to the significance level (alpha level), **power** refers to the power level, and **type** is the type of *t*-test being conducted (i.e., a one-, two-, or dependent samples *t*-test). The **alternative** option refers to the alternative hypothesis, where it can take on the values of "two.sided", which is the default for this function, or "greater" or "less", which are both one-sided hypotheses.

For estimating the power for a test involving two groups of 35 observations each, a difference in population means of 0.2, and a standard deviation of 1, we would input the following arguments:

```
pwr.t.test(n=35, d= (0.2-0)/1, sig.level=0.05,
           power= NULL, type="two.sample")
```

Note that **d** is simply equal to the difference in means divided by the pooled standard deviation, so we can enter this calculation directly in the function. We have

---

<sup>5</sup>If the value of **r2r(#)** in the command is changed to **r2r(0.0196)**, the resulting estimate for required sample size is 760—a value that differs by 1 from the 759 cases reported above.

left the power argument empty, which tells R that this is what we want to estimate. R produces the following output:

```
pwr.t.test(n=35, d= (0.2-0)/1, sig.level=0.05,
           power= NULL, type="two.sample")
```

Two-sample t test power calculation

```
n = 35
d = 0.2
sig.level = 0.05
power = 0.1308497
alternative = two.sided
```

NOTE: n is number in \*each\* group

The power estimate of 0.13 would lead us to expect 13 significant results in 100 tests ( $0.13 \times 100$ ). Changing the expected sample size will result in different power estimates. Just remember that the value of **n** that you include should be the number of observations in each group.

If we are interested in estimating the sample size needed to attain a certain level of power, we would leave the **n** argument empty, and instead add an argument for **power**. To compute the sample size needed to detect an effect with a power of 0.8, we would enter the following command:

```
pwr.t.test(n=NULL, d=(0.2-0)/1, sig.level=0.05,
            power= 0.8, type="two.sample")
```

### **ANOVA**

R also allows for conducting power and sample size analysis for ANOVA models. This can be accomplished using the **pwr.anova.test()** function from the *pwr* package:

```
pwr.anova.test(k = NULL, n = NULL, f = NULL,
                sig.level = 0.05, power = NULL)
```

where **k** is the number of groups in the analysis, **n** is the number of observations in each group, **f** is the effect size, **sig.level** is the significance level, and **power** is the power of the test. Like the previous function, R knows which value to compute based on which argument is left empty or null. For example, to compute the power needed of a test, you would leave the **power** argument empty. We could compute the power of a one-way ANOVA design with three groups (**k**) and 100 cases in each

group for three different effect sizes  $f(0.1, 0.25, \text{and } 0.4)$ . For the small effect, the R code would look like:

```
pwr.anova.test(k = 3, n = 100, f = 0.1,
sig.level = 0.05, power = NULL)
```

Giving you the following output:

```
pwr.anova.test(k = 3, n = 100, f = 0.1,
sig.level = 0.05, power = NULL)
```

Balanced one-way analysis of variance power calculation

```
k = 3
n = 100
f = 0.1
sig.level = 0.05
power = 0.318643
```

NOTE: n is number in each group

Alternatively, if you want to obtain the power estimate for the medium and large effect sizes, you would simply change the **f** argument to reflect these values.

It is also possible to estimate the power for a given effect size over a range of group sizes. This entails storing a sequence of values in an object to be later used in the **n** argument of the function. For example, to determine the power to detect a small effect ( $f = 0.1$ ) in a study with three groups starting at 100 observations per group to 500 observations per group (in increments of ten observations per group), you would write as follows:

```
n <- seq(100,500,10)
pwr.anova.test(k = 3, n = n, f = 0.1,
sig.level = 0.05, power = NULL)
```

If we then wanted to view the results of this analysis in a table, we can save the results of our test in an R object, and print a data frame of values of interest to us. Below is the same code as above, yet now the results of the test are stored in the object *results*. Using the **data.frame()** function, we can then print a data frame of the significance level, power, n, effect size, and number of groups.

```
n <- seq(100,500,10)
results <- pwr.anova.test(k = 3, n = n, f = 0.1,
sig.level = 0.05, power = NULL)
data.frame(alpha = results$sig.level,
power = results$power, n = n,
delta=results$f, N_g = results$k)
```

Running the code above produces the following result:

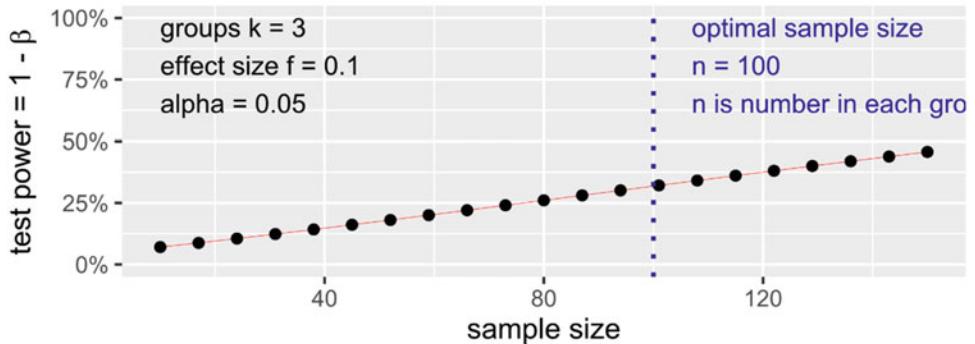
```
> n <- seq(100,500,10)
>
> results <- pwr.anova.test(k = 3, n = n, f = 0.1,
+                             sig.level = 0.05, power = NULL)
>
> data.frame(alpha = results$sig.level,
+             power = results$power, n = n,
+             delta=results$f, N_g = results$k)

  alpha   power     n delta N_g
1 0.05 0.3186430 100  0.1  3
2 0.05 0.3472508 110  0.1  3
3 0.05 0.3755599 120  0.1  3
4 0.05 0.4034710 130  0.1  3
5 0.05 0.4308970 140  0.1  3
6 0.05 0.4577623 150  0.1  3
7 0.05 0.4840022 160  0.1  3
8 0.05 0.5095622 170  0.1  3
9 0.05 0.5343969 180  0.1  3
10 0.05 0.5584699 190  0.1  3
11 0.05 0.5817527 200  0.1  3
12 0.05 0.6042240 210  0.1  3
13 0.05 0.6258690 220  0.1  3
14 0.05 0.6466789 230  0.1  3
15 0.05 0.6666505 240  0.1  3
16 0.05 0.6857850 250  0.1  3
17 0.05 0.7040879 260  0.1  3
18 0.05 0.7215685 270  0.1  3
19 0.05 0.7382392 280  0.1  3
20 0.05 0.7541152 290  0.1  3
```

It is also possible to plot results from the power analysis using the **plot()** function:

```
results2 <- pwr.anova.test(k = 3, n = 100, f = 0.1,
                           sig.level = 0.05, power = NULL)
plot(results2)
```

## Balanced one-way analysis of variance power calculation



If instead, you are interested in estimating the sample size required for a particular level of statistical power for a one-way ANOVA, you can do this by specifying a value for the **power** argument and leaving the **n** argument blank:

```
pwr.anova.test(k = 3, n = NULL, f = 0.1, power = 0.8,
                 sig.level = 0.05)
```

This results in the following output:

```
> pwr.anova.test(k = 3, n = NULL, f = 0.1,
                  power = 0.8, +sig.level = 0.05)
```

Balanced one-way analysis of variance power calculation

```
k = 3
n = 322.157
f = 0.1
sig.level = 0.05
power = 0.8
```

NOTE: n is number in each group

So, the estimated sample size needed to attain 80% power, with this effect size, is  $322 \times 3 =$  roughly 966 observations.

### Correlation

To compute power estimates for correlation coefficients in R, you could use the **pwr.r.test()** function in the *pwr* package.

```
pwr.r.test(n = NULL, r = NULL, sig.level = 0.05,
            power = NULL,
            alternative = c("two.sided", "less", "greater"))
```

where **n** is the number of observations, **r** is the linear correlation coefficient, **sig.level** is the significance level or alpha, **power** is the level of power, and **alternative** references the alternative hypothesis, the default being "two.sided". This function uses the *z*-transformation of correlation coefficients in its calculations. As with the other power functions discussed, the value that you want to estimate (e.g. power, sample size) should be left blank or null.

To conduct analysis of power estimates for correlation coefficients for a sample size of 100, we would enter the following code in R to estimate the power to detect a weak correlation:

```
pwr.r.test(n = 100, r = 0.1, sig.level = 0.05,
            power = NULL, alternative= "greater")
```

The estimated power received is approximately 0.26:

```
> pwr.r.test(n = 100, r = 0.1, sig.level = 0.05, +
              power = NULL, + alternative= "greater")
```

approximate correlation power calculation (arctanh transformation)

```
n = 100
r = 0.1
sig.level = 0.05
power = 0.2574446
alternative = greater
```

You can see how the estimated power changes for different magnitudes of association by adjusting the **r** value to reflect different correlation coefficients. If you want to see what sample size is required to attain a certain level of statistical power for a given level of association, you would include a value for the **power** argument, and set the **n** argument to NULL or empty. For example, if we wanted to estimate the sample size needed to detect a medium correlation (i.e., **r** = 0.3) with a power level of 0.80, we would enter the following command:

```
pwr.r.test(n = NULL, r = 0.3, sig.level = 0.05,
            power = 0.8, alternative= "greater")
```

In this case, the approximate sample size needed is at least 66.6 observations.

### *OLS Regression*

Power and sample sizes can also be computed for an  $R^2$  test in the general linear regression model. The **pwr.f2.test()** function from the *pwr* package allows for this:

```
pwr.f2.test(u = NULL, v = NULL, f2 = NULL,
             sig.level = 0.05, power = NULL)
```

where **u** refers to the degrees of freedom for the numerator (equal to the number of coefficients being estimated in the model minus the constant), **v** is the degrees of freedom for the denominator, **f2** is the effect size, **sig.level** is the significance level or alpha level, and **power** refers to the power of the test. Note that the standardized effect size (**f2**) is equivalent to  $R^2/(1 - R^2)$ , and this calculation can be included directly in the argument for **f2=**. The degrees of freedom for the denominator (**v**) is equivalent to the sample size minus the degrees of freedom for the numerator (**u**) minus 1.

To calculate the power needed for an OLS regression with eight variables being tested, 100 observations, and a weak effect ( $R^2 = 0.02$ ), we would run the following code:

```
pwr.f2.test(u = 8, v = 91, f2 = .02/(1-.02),
             sig.level = .05, power = NULL)
```

The power estimate we receive is 0.125. Adjusting the **f2** argument will allow you to estimate power for different values of  $R^2$ . To compute the sample size needed to achieve a designated level of statistical power, we would simply set the **v** argument to null or empty and include a power estimate:

```
pwr.f2.test(u = 8, v = NULL, f2 = .02/(1-.02),
             sig.level = .05, power = .8)
```

Recall that **v** is equivalent to **n - u - 1**, so this should be taken into account when trying to calculate needed sample size. We find that the estimated sample size needed to detect a weak effect ( $R^2 = 0.02$ ) is about 744 ( $734.8 + 8 + 1 = 743.8$ ).

### Problems

1. Compute the estimates of statistical power for each of the four scenarios provided below. Which scenario has the highest level of statistical power? Explain why.

Scenario 1: One-tailed test, 0.01 significance level

Sample size = 100 each

$\mu_1 = 15, \mu_2 = 10, \sigma = 2$

Scenario 2: One-tailed test, 0.05 significance level

Sample size = 100 each

$\mu_1 = 15, \mu_2 = 10, \sigma = 2$

Scenario 3: One-tailed test, 0.01 significance level

Sample size = 100 each

$\mu_1 = 15, \mu_2 = 10, \sigma = 2$

Scenario 4: One-tailed test, 0.05 significance level

Sample size = 100 each

$\mu_1 = 15, \mu_2 = 10, \sigma = 2$

2. Compute the estimates of statistical power for each of the four scenarios provided below. Which scenario has the highest level of statistical power? Explain why.

Scenario 1: Two-tailed test, 0.05 significance level

Sample size = 100 each

$\mu_1 = 15, \mu_2 = 14, \sigma = 5$

Scenario 2: One-tailed test, 0.05 significance level

Sample size = 100 each

$\mu_1 = 15, \mu_2 = 16, \sigma = 5$

Scenario 3: One-tailed test, 0.01 significance level

Sample size = 100 each

$\mu_1 = 15, \mu_2 = 10, \sigma = 5$

Scenario 4: One-tailed test, 0.05 significance level

Sample size = 100 each

$\mu_1 = 15, \mu_2 = 9, \sigma = 5$

3. Compute the estimates of statistical power for each of the following one-way ANOVA studies. (For all scenarios, assume that the researcher is trying to detect a small effect.)

- Scenario 1: Three groups with 75 cases per group.
- Scenario 2: Four groups with 60 cases per group.
- Scenario 3: Five groups with 55 cases per group.

Which scenario would have the highest level of statistical power? Explain why.

4. In attempting to design a correlation study looking at academic performance and delinquency, a researcher expects a small-to-moderate correlation among a population of adolescents he or she will sample from.
- (a) If he or she computes estimates of statistical power assuming a two-tailed test, what size sample would he or she need to detect a small correlation? Medium correlation?
  - (b) Do you think he or she could justify a one-tail test of the correlation? If a one-tail test was used, how does the estimated sample size change for both the small and medium correlations?
5. A research team is preparing to launch a statewide survey to gauge public sentiment about the incarceration of juvenile offenders, focusing primarily on support for more lenient punishments. Consistent with much public opinion research, expectations are that a combination of ten independent variables is likely to explain about 15% of the variation in views about juvenile punishment.

- (a) What size sample would the researchers need to have to achieve a power of 0.80? 0.90?
- (b) Of particular interest to the researchers is the effect of three different measures of experience with the justice system, but their expectation is that the overall effect of these three measures will be small. What size sample would the researchers need to achieve a power of 0.80? 0.90?
- (c) What size sample should the researchers try to obtain? Explain why.

## References

---

- Borenstein, M., Rothstein, H., & Cohen, J. (2001). *Power and precision*. Englewood, NJ: Biostat, Inc..
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Dattalo, P. (2008). *Determining sample size*. New York, NY: Oxford University Press.
- Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G\*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175–191.
- Hayes, J. P., & Steidl, R. J. (1997). Statistical power analysis and amphibian population trends. *Conservation Biology*, 11(1), 273–275.
- Kraemer, H. C., & Thomann, S. (1987). *How many subjects: Statistical power analysis in research*. Newbury Park, CA: Sage.
- Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.
- Maltz, M. D. (1994). Deviating from the mean: The declining significance of significance. *Journal of Research in Crime and Delinquency*, 31, 434–463.
- Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for accuracy in parameter estimation. *Annual Review of Psychology*, 59, 537–563.
- Murphy, K. R., & Myors, B. (2003). *Statistical power analysis* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Petersilia, J. (1989). Randomized experiments: Lessons from BJA's Intensive Supervision Project. *Evaluation Review*, 13, 435–458.
- Thomas, L. (1997). Retrospective power analysis. *Conservation Biology*, 11(1), 276–280.
- Weisburd, D. (1991). Design sensitivity in criminal justice experiments. *Crime and Justice*, 17, 337–379.
- Weisburd, D., Lum, C. M., & Yang, S. M. (2003). When can we conclude that treatments or programs “don't work”? *The Annals of the American Academy of Political and Social Science*, 587(1), 31–48.

## Randomized Experiments

### **W h a t   D o e s   a   R a n d o m i z e d   E x p e r i m e n t   L o o k   L i k e ?**

---

What are the Advantages of Randomized Experiments?

How Do Randomized Experiments Maximize Internal Validity?

What are Some of the Key Design Types and Associated Analyses in Randomized Experiments?

### **W h a t   i s   a   B l o c k   R a n d o m i z e d   T r i a l ?**

---

How Does Block Randomization Help Increase Equivalence and Statistical Power?

How Can Covariates be Used to Help Increase Statistical Power?

### **W h a t   i s   a   F a c t o r i a l   E x p e r i m e n t ?**

---

How Do You Distinguish Between Within- and Between-Group Factors?

**R**ANDOMIZED EXPERIMENTS have become a key method for identifying the effects of treatments or programs on outcomes in criminology and criminal justice. They have also become an important method for identifying how people's perceptions and attitudes change in different scenarios that are created in the laboratory. We begin the chapter by describing the structure of a randomized experiment and then illustrate why randomized experiments provide a very strong ability to make causal inferences without concern for confounding. Indeed, many scholars have taken the position that only randomized experiments can provide valid conclusions regarding the impacts of treatments and programs (see Boruch et al. 2000; Campbell and Boruch 1975; Cook et al. 1979; Feder et al. 2000; Flay and Best 1982; Weisburd 2000; Weisburd et al. 2001; Wilkinson 1999). Joan McCord (2003) argued, for example, that crime and justice evaluations should employ random assignment *whenever possible*. We then turn to selected design types and associated analyses. We pay particular attention to block randomized studies and illustrate how they help the researcher to maximize equivalence and statistical power in randomized experiments. Finally, we discuss the approach of using covariates in experimental studies as a method of increasing statistical power.

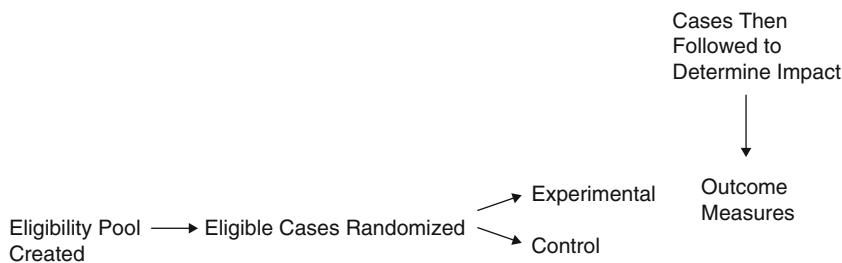
## The Structure of a Randomized Experiment<sup>1</sup>

---

The general structure of experiments in criminology is usually similar in design regardless of the area or the question of interest. Generally, experiments in criminology and criminal justice start with an eligibility pool, randomization, and posttest measures relevant to the dependent variable of interest (Fig. 9.1).

---

<sup>1</sup>For this section and the section on internal validity, we draw heavily from Weisburd et al. (2013).

**Figure 9.1***Diagram of the typical criminological experiment*

The **eligibility pool** is made up of those participants or units that are eligible for the experiment. Units of analysis can be individuals or aggregated groups or other entities that often are found in clusters. For example, in an experiment that evaluates the impact of increased foot patrol on crime rates, the eligibility pool may be individual officers who walk the beat in a specific area and who have a specific number of years of experience. Or the unit of analysis could be the geographic area or the beat that will be assigned to different conditions. The eligibility pool thus comprises those patrol officers (or patrol beats) that meet the criteria for inclusion in the study. The eligibility pool may also be a convenience sample of suitable individuals, such as university undergraduates recruited for a laboratory-based study on jury decision making.

Next, researchers randomly assign members from this pool of eligible participants or units to the study conditions—often a **treatment group** and a **control group**. However, there may be multiple groups. For example, an experiment might compare two different forms of a school-based drug prevention program to a no-treatment-control group. Even more complex designs are possible, such as factorial design where each factor (an independent variable of interest) represents two or more conditions. For example, a 2 by 2 factorial design on wrongful convictions might manipulate whether a participant is guilty or innocent (the first factor) by whether the interrogation method was accusatory or information gathering (the second factor) (For an example, see Meissner et al. 2014). The essential feature is that there are at least two conditions.

Study units are then randomly assigned to the experimental conditions. There are many ways to randomize study units, but most often researchers rely on computerized statistical software to carry out **randomization**. Some researchers may simply use the rule of odds and evens—that is, assigning every other case to one particular group. This is often referred to as alternation and is considered quasi-random assignment as the assignment of the numbers used is not actually random. When assignment is not

completely random, there is always a possibility that a systematic factor of bias has been added to the study.

The most critical factor in randomization is that each case has a known nonzero probability or likelihood of being assigned to each of the experimental groups or conditions. Typically, this is an equal probability or 50/50 random chance that any unit is assigned to the experimental or control group, assuming a two-group design. However, it is possible to have differential probabilities of assignment to each group.

In the usual criminological experiment, eligible cases are randomly assigned to one of the two groups—treatment or control. Experiments may have more than two groups. But typically, an experiment comprises a group that receives the treatment or the intervention and a control group that does not. It is also quite common in a criminological experiment for the control group to actually receive some type of conventional intervention. For example, in a foot patrol experiment, the control group may receive treatment as usual or the same number of foot patrol officers as typically employed. This provision of treatment as usual is often required because of the ethical challenges of not providing some type of treatment to people or places that are plagued by crime or other problems. However, the use of conventional treatment as a control may lead us to underestimate the actual impacts of an intervention (Ariel et al. 2020).

Experimental designs can include any number of outcomes. If the randomization was implemented with fidelity, it should produce two groups that are roughly equivalent on the pretest or baseline measures related to the outcome of interest.<sup>2</sup> Differences that do exist are random, reflecting sampling error. However, it is important to note that simple randomization schemes do not guarantee equivalence; rather, they allow the researcher to assume that there is no systematic reason for the groups to differ. Later in the chapter, we describe methods that increase the likelihood that the assumption of equivalence will be evident in the distribution of the sample cases. Finally, the researcher conducts analyses to determine if the intervention had any impact on the **posttest measures** or follow-up measures of the outcomes of interest, above and beyond what might be expected to result from random processes.

---

<sup>2</sup>Statistically, the groups are equivalent in the sense that the expected value of the mean for any baseline characteristics is the same across conditions. However, the observed groups will differ, but these differences will conform to known probability distributions, enabling us to differentiate between outcome differences that were plausibly due to these random differences or likely due to the experimental manipulation (e.g., treatment).

## The Main Advantage of Experiments: Isolating Causal Effects

---

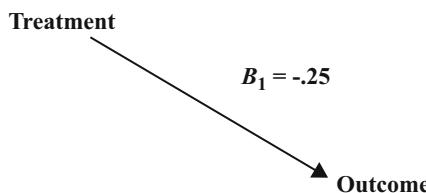
In identifying whether a variable has a causal influence, or evaluating the outcomes of treatments or programs, the key issue for researchers is getting an unbiased estimate of the treatment or the intervention effect. Without that any other consideration, such as the ability to generalize results, is superfluous. For example, suppose an evaluator was asked to assess whether an intervention for drug-involved offenders provided an effective deterrent to future offending. In the study employed, the treatment group was found to be half as likely to recidivate as the control condition not receiving the intervention. This would ordinarily lead the evaluator to report that the intervention was a success. But what if it was difficult to believe the result that was gained in the study? Suppose that the design of the study did not allow the evaluator to assume that the observed effect was actually the result of the intervention. In this scenario, the evaluator could not be sure that it was the intervention that caused the change or something else that was common to the treatment group but not to the control group. In such a situation, it does not do much good to argue about whether the results can be generalized to a specific population of interest. The results themselves are not believable.

The main problem researchers face in producing believable results is that treatments are often confounded with other factors. For example, suppose that the reason for the outcome observed above was that the evaluator had not taken into account the fact that the treated drug offenders were volunteers. Volunteers in turn are more likely to be highly motivated to succeed in such programs than individuals who are not volunteers (see De Leon et al. 2000; Taxman 1998; Rosenthal 1965). This is often termed *creaming* in the identification of the subjects in the treatment condition. The reason why the intervention group had lower recidivism rates in this case could easily be explained by the fact that they were on average more motivated to be rehabilitated than drug offenders in the control condition.

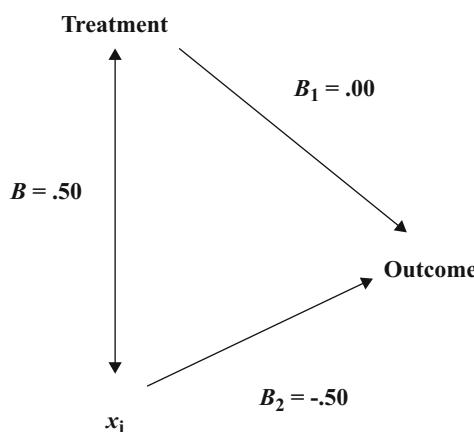
All research studies that seek to establish causation between a specific variable or treatment and an outcome must deal with this problem of confounding, and it stands as the major barrier to drawing believable conclusions in criminological studies. Nonexperimental methods, such as regression techniques using observational data that we reviewed in earlier chapters, rely on statistical controls to solve the problem of confounding. The logic is easily stated: If we know what the factors are that confound treatment (or the variable of interest), we can take them into account, such as including them in a regression model. In other words, nonexperimental methods, as we noted in Chap. 2, rely on a knowledge solution to the problem of confounding.

**Figure 9.2**

*Example of the bias in the estimate of a treatment effect caused by the exclusion of an unknown or an unmeasured factor ( $x_j$ )*



(a) Estimate of  $B_1$  in the case where the factor ( $x_j$ ) is unmeasured and excluded from the model.  
Estimate of  $B_1$  is  $-.25$ .



(b) Estimate of  $B_1$  in the case where the factor ( $x_j$ ) is included in the model. Estimate of  $B_1$  is  $.00$ .

But how does knowledge solve the problem of confounding? Let us take the example of regression analyses as described in Chap. 2 using observational data examining the question of the effect of a drug intervention program on recidivism. Figure 9.2a shows the effect of the intervention on recidivism using a standardized regression coefficient approach. Here, we have the simple relationship between treatment and recidivism with a coefficient value of  $-.25$ , a result suggesting that the intervention decreases recidivism. However, when we include in our analysis the **confounding factor**—level of motivation—the relationship between treatment and recidivism changes (see Fig. 9.2b). Taking into account the effect of level of motivation, the effect of the intervention becomes 0 in this illustration.

The observed effect was not due to the intervention but rather due to the confounding of the intervention with motivation of offenders.

Notice that two extra coefficients are included in Fig. 9.2b. The first represents the relationship between treatment (the variable of interest) and motivation (the confounding variable). This standardized coefficient is .50 and represents the extent to which treatment and motivation are related or confounded. The second additional coefficient (−.50) represents the relationship between motivation and recidivism. Together, these two relationships detail the extent of confounding that is clouding our ability to estimate the treatment or the program effect. Confounding in this context takes into account the extent to which the confounding factor is related to the outcome of interest, and the degree to which it is related to or confounded with the treatment variable. And indeed, we can estimate the overall confounding by simply multiplying them. This gives us a value of −.25, the value of the simple relationship observed in Fig. 9.2a. This can be defined as the degree of bias. In this case, the degree of bias is equal to the observed effect. Had we not taken motivation into account, we would have erroneously concluded that the treatment leads to lower rates of recidivism, when it is actually the motivation of offenders (represented by  $x_1$  below) that is responsible for this result.

The way in which multiple regression approaches allow us to control for confounding is illustrated in Eq. 9.1 and described in Chap. 2. Here, we show the computation of the regression coefficient  $b$  for a treatment variable ( $t$ , coded as 1 for the treatment and 0 for the control) controlling for a confounding factor ( $x_1$ , in this case motivation):

$$b_t = \left( \frac{r_{y,t} - (r_{y,x_1} r_{t,x_1})}{1 - r_{t,x_1}} \right) \left( \frac{s_y}{s_t} \right)$$
Equation 9.1

The key part of the equation for our interest is the numerator in the first part of the equation:  $r_{y,t} - (r_{y,x_1} r_{t,x_1})$ . Note that it includes the simple correlation between the treatment variable and the outcome measure ( $r_{y,t}$ ). Subtracted from that is the product of the correlation between the outcome measure and the confounding variable ( $r_{y,x_1}$ ) and the treatment and confounding variable ( $r_{t,x_1}$ )—the two components of confounding we have just described. In this context, we can statistically unconfound our estimate of the treatment if we have knowledge of the confounding factor. In our example, the computation would be  $-.25 - (.50 * -.50)$  or  $-.25 + .25$ , or 0, suggesting a null effect for treatment.

This solution is also key to the myriad of approaches that have been developed for other types of nonexperimental (i.e., quasi-experimental) approaches. All of them rely on knowledge about confounding. For

example, matching of subjects on known characteristics begins with the basic assumption that we have enough knowledge regarding confounding to create equivalence of units in the treatment and control conditions. Because the subjects in the matched groups are assumed to be alike, we make an assumption that confounders are not influencing our observations of a treatment effect. Note that in this case we are trying to statistically control for such confounding factors, by making the treatment and control groups alike on these factors. In such a case, the correlation between treatment and confounding variables is assumed to be 0. When it is, as illustrated in Eq. 9.2 for the bivariate regression coefficient, the effect of the treatment breaks down to the simple correlation between treatment and outcome:

$$b_t = (r_{y,t}) \left( \frac{s_y}{s_t} \right) \quad \text{Equation 9.2}$$

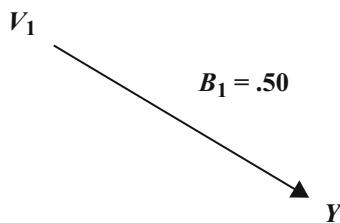
The problem with these methods is that if we want to get an unbiased estimate of treatment, we would in theory have to identify all confounding causes. Using the regression approach, which in some sense provides the most transparent form of nonexperimental methods, we would need to identify all confounding factors (factors related to participation in the treatment) that also have meaningful impacts on the outcome measure and include them in the regression. This would mean both that we would have to have knowledge about all such confounding factors and that we would be able to measure them in a research study.

Randomized experiments start with a different logic. If we cannot control out for confounding, we can make it irrelevant for the problem at hand. This is done through the process of randomizing treatment. If treatment is randomized, then there is no reason to suspect systematic biases. This can be illustrated by returning to the simple path diagrams we used earlier. In Fig. 9.3a, we show the simple relationship between a treatment and outcome. In Fig. 9.3b, we include a potential confounding variable. Note that the confounding factor has a strong standardized relationship with the outcome variable ( $B = .50$ ). However, using the theory of randomization, we can assume that there is no systematic relationship between the treatment and the confounder. This is the case because treatment has been randomly allocated. In theory, it is not going to be related systematically to other factors such as gender, race, age, and attitudes.

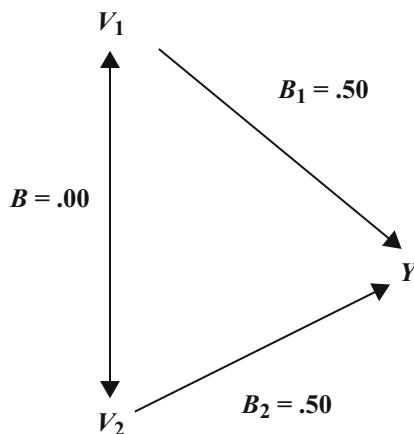
What this means is that the relationship between any potential confounder and the treatment can be assumed to be 0. By chance, fluctuations will occur, and there will be confounding relationships observed, but these can in this case be assumed to balance out in the long run. Or at least there is no reason to assume that they will not. Equally important, our statistical

**Figure 9.3**

*Example of the lack of confounding in the treatment effect when the treatment ( $V_1$ ) and potential confounder ( $V_2$ ) have no relationship*



(a) The model excluding a potential confounder,  $V_2$ .



(b) The model including a potential confounder, ( $V_2$ ) but no relationship between the treatment ( $V_1$ ) and the confounder because of randomization. If  $V_2$  is excluded, the bias =  $B \times B_2 = .00 \times .50 = .00$ .

methods provide a basis for determining how large any chance confounding relationship might be. Thus, if our observed treatment effect is larger than what might occur due to chance confounding, then we can have confidence that the treatment does have an effect on the outcome.

## Internal Validity

---

Our discussion so far is often subsumed under the heading of **internal validity** in methodological texts in criminology. A research design in which the impact of the intervention can be clearly distinguished from other

observed factors is known as having high internal validity. If there are confounding factors involved in the impact of the intervention, then the evaluation design is considered to have low internal validity. Shadish et al. (2002) have identified the most common threats to internal validity:

1. *Selection*: The pre-existing differences between treatment and control subject or units.
2. *History*: An external event occurring at the same time of the study that may influence impact.
3. *Maturation*: Changes in subjects or units between measurements of the dependent variable. These changes may be of natural evolution (e.g., aging) or due to time-specific incidences (e.g., fatigue and illness).
4. *Testing*: Measurement at pretest impacts measurement at posttest.
5. *Instrumentation*: Changes to the instrument or the method of measurement in posttest measures.
6. *Regression to the mean*: Natural trends may cause extreme subjects or units who score extremely high or low during the pretest to score closer to the mean at posttest.
7. *Differential attrition*: The differential loss of subjects or units from the treatment group compared to the control group.
8. *Causal order*: The certainty that the intervention did in fact precede the outcome of interest.

To further illustrate the importance of internal validity, let us suppose a researcher is interested in evaluating the impact of a youth court on juvenile recidivism. The internal validity is considered high if, at the end of the evaluation, the researcher can show that the change in juvenile recidivism among the intervention group is due only to the intervention (i.e., youth court) and no other confounding factors were at play. The researcher must show through either research design or analytical procedures that all confounding factors are accounted for in the measurement of outcomes. If the researcher is unable to account for other factors such as seriousness of first offense or the maturation of the study population, he or she must note that the observed effects may be due to other factors. If threats to validity (or potential confounding factors) are not accounted for, the internal validity of the study would be considered low.

Generally speaking, a randomized experiment has the highest possible internal validity, because as we illustrated above, this approach allows the researcher to assume that other confounding causes of the outcome of interest, known and unknown, are not systematically influencing the study results (i.e., they have been balanced between the groups). High internal validity in randomized experiments is gained through the process

of randomly allocating the treatment or the intervention to the experimental and control groups. Through random assignment, the researcher is not just randomizing the treatment. He or she is randomizing all other factors that may influence the outcome of the treatment. Thus, there is no systematic bias that increases the odds of one-unit's assignment to the treatment group and another unit's assignment to the control group. This is not to imply that the groups are the same on every characteristic—it is very possible that differences may occur; however, these differences can be assumed to be randomly distributed and are accounted for in the probability distributions that underlie statistical tests of significance. Regardless, neither the treatment group nor the control group should have an advantage over the other on the basis of known or unknown variables. Thus, randomized experiments are one of only a few designs that allow the researcher to assume statistically unbiased effects.<sup>3</sup>

The goal of most randomized experiments in criminology and criminal justice, as in other social science fields, is to disentangle the impact of the treatment or the intervention from the impact of other factors on the outcomes that are to be tested. A randomized experiment allows the researcher to attribute differences between the groups from pretest to posttest to the treatments or the interventions that are applied. At the conclusion of the study, the researcher is able to assert, with confidence, that the differences are likely a result of the treatment and not due to other confounding factors, within the bounds of statistical probability. It is more difficult for nonrandomized studies, even a high-quality quasi-experimental design, to make this assertion. This advantage is underscored by Farrington (1983):

The unique advantage of randomized experiments over other methods is high internal validity. There are many threats to internal validity that are eliminated in randomized experiments but are serious in nonexperimental research. In particular, selection effects, owing to differences between the kinds of persons in one condition and those in another, are eliminated.

## **Selected Design Types and Associated Statistical Methods**

---

Experimental designs include a large collection of possible design variants. We will focus on the most common of these and the statistical methods most relevant to each. Before we discuss these, it is important to distinguish between manipulated factors, observed or nonmanipulated factors,

---

<sup>3</sup>See Boruch (1997). Econometrics includes other methods, such as instrumental variable analysis, that allow for an unbiased estimation of a treatment effect. These are beyond the scope of this text. However, these methods often rely on naturally occurring random processes, thus mimicking what is discussed here.

blocking factors, and covariates. It is also important to differentiate **between-subjects** and **within-subjects** factors, as described below.

A factor is simply a nominal independent variable with two or more categories. For example, a simple two-group randomized experiment with a treatment group and a control group has a single factor representing the construct of treatment with two conditions, treatment and control. A factor is manipulated if it is under the control of the researcher. That is, if the researcher controls through randomization which level of a factor study units are exposed to, then it is controlled by the researcher and hence a manipulated factor. For a study to be experimental, it must have at least one such factor. A study may also have one or more observed or nonmanipulated factors, such as sex or risk level. Study units cannot be randomly assigned to the levels of an observed factor.

A block or blocking factor is a non-manipulated factor that is incorporated into the randomization design. A blocking factor is an observed characteristic of the study units, such as the risk level of an individual being released from prison or the baseline crime level of a hot spot. What makes it a blocking factor is that units are blocked or stratified according to this factor prior to randomization and randomization occurs within each level of the blocking factor, ensuring balance on this factor in each level of the manipulated factor, that is, the factor that is the focus of the study.

Factors may be either between-subjects or within-subjects. With the former, study units are randomly assigned to only one level of the factor. For example, a hot spot might be assigned to either the enhanced policing or the routine policing conditions but not both. Most factors typically will be of this type. With a within-subjects factor, each study unit gets assigned to all levels of that factor. A common example is repeated measurements reflecting a time factor with two or more levels. However, a within-subjects factor might also represent different treatments. For example, a study might examine the effects of three different doses of caffeine on cognitive performance. The study might randomly assign the order that each subject receives the different doses (or use another counterbalancing mechanism) such that each subject gets each dose.

Finally, a covariate is simply an observed characteristic at baseline on a scaled variable (i.e., ordinal or higher level of measurement). Thus, the only difference between a covariate and an observed factor is the level of measurement. As discussed below, covariates can be incorporated into the analysis of experimental data, primarily to increase statistical power.

These components can be combined in different ways to create a design that addresses the specific needs of a research study. The interested reader may wish to consult Maxwell et al. (2017) and Kirk (2013) for more information and discussion of numerous specific design types. Below, we will explore how to analyze the data from a few of the more common variants relevant to experimentation in criminology and criminal justice.

### The Two-Group Randomized Design

The two-group randomized experiment is often called a randomized controlled trial when used to compare two treatments or a treatment to a no-treatment or routine treatment control. The analysis of the treatment or experimental effect from these designs is straightforward. In the case of a scaled dependent variable, an independent sample  $t$ -test provides an unbiased test of the difference between the mean for each group, assuming the study was carried out with fidelity to the original design and randomization. The standard equation for the independent  $t$ -test is shown here:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2(n_1-1) + s_2^2(n_2-1)}{n_1+n_2-2}}}, \quad \text{Equation 9.3}$$

where  $\bar{x}$  is the mean,  $s$  is the standard deviation,  $n$  is the sample size, and the subscripts 1 and 2 indicate the treatment and control conditions, respectively. The degrees of freedom for this  $t$ -test is  $n_1 + n_2 - 2$ .

For a binary outcome, such as success/failure on some variable of interest, the difference in the success (or failure) rate between the two conditions can be tested using a  $\chi^2$  test of independence based on the 2 by 2 contingency table. If we label the frequencies of the 2 by 2 contingency table as  $a$ ,  $b$ ,  $c$ , and  $d$ , we can compute the  $\chi^2$  as:

$$\chi^2 = \frac{(ad - bc)^2(a + b + c + d)}{(a + b)(c + d)(a + c)(b + d)} \quad \text{Equation 9.4}$$

The  $\chi^2$  in Eq. 9.4 has 1 degree of freedom.

For other dependent variable types, such as count data or a nominal dependent variable with 3 or more categories, an appropriate generalized linear model, such as a negative binomial or multinomial logistic regression model, can be used with a dummy indicator for treatment as the independent variable.

### Three or More Group Randomized Design

The two-group randomized experiment can be extended to three or more groups. For example, a study might randomly assign study units to one of two treatments or a control condition. The analysis of such designs for a scaled outcome variable is the one-way analysis of variance (ANOVA), which extends the  $t$ -test to the comparison of three or more means. We can write the linear model for a one-way ANOVA as follows:

$$y_{ij} = \mu + \alpha_j + \epsilon_{ij},$$

where  $y_{ij}$  is the dependent variable for subject (or study unit)  $i$  within condition  $j$ ,  $\mu$  is the overall or grand mean,  $\alpha_j$  is the deviation of the mean for group  $j$  from the overall mean, and  $\epsilon_{ij}$  is the residual or deviation of each subject from their group mean. The ANOVA compares the variability across the means (variability associated with  $\alpha_j$ ) to the residual variability (variability associated with  $\epsilon_{ij}$ ). The former is called the  $MS_{\text{between}}$ , and the latter is called the  $MS_{\text{within}}$  where  $MS$  is the **mean squares**. Equations (9.5) and (9.6) show how to compute each of these:

$$MS_{\text{between}} = \frac{\sum n_j (\bar{y}_j - \bar{y})^2}{a - 1}, \quad \text{Equation 9.5}$$

$$MS_{\text{within}} = \frac{\sum \sum (y_{ij} - \bar{y}_j)^2}{N - a}, \quad \text{Equation 9.6}$$

where  $\bar{y}_j$  is the mean for each condition,  $\bar{y}$  is the overall mean,  $n_j$  is the sample size for each condition,  $N$  is the total sample size, and  $a$  is the number of conditions. The  $F$ -test is the ratio of these two values:

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}, \quad \text{Equation 9.7}$$

where the degrees of freedom is  $a - 1$  for the numerator and  $N - a$  for the denominator. Notice that the numerator for Eq. (9.5) is the sum-of-squares (SS) for each mean relative to the overall mean. The squared difference between the group mean and overall mean is multiplied by the sample size for the corresponding group so that we are summing this deviation once for every study unit within each group. The numerator for Eq. (9.6) is the sum-of-squares within each group or the residual sums of squares.

Assuming the  $F$  is significant, we can then run pairwise or other planned comparisons to better understand the pattern of evidence across the group means. This is covered in the chapter on one-way ANOVA in *Basics Statistics in Criminology and Criminal Justice*, the companion volume to this book.

For outcome measures that are not scaled, such as binary, count, and nominal dependent variables, suitable regression methods can be used combined with dummy variables encoding group membership. Note that

dummy coding will provide tests of paired comparisons, comparing each group to the reference or omitted group. To get an overall test of whether the results differ across conditions, such as what is produced in a one-way  $F$ , the likelihood ratio (LR) test can be used, contrasting the null model with the model that includes all dummy variables related to group membership.

### Factorial Design

A **factorial design** has two or more independent variables being compared simultaneously. For example, laboratory-based studies of the effect of alcohol on aggression typically have a 2 by 2 factorial design (meaning that there are four conditions being compared). The first factor is whether the study participants (college undergraduates) are given a drink that contains alcohol or does not contain alcohol. The second factor is whether the participant is told that the drink contains alcohol or not. The participants then participate in a teacher–learner or competitive reaction time exercise, and their level of aggression is measured. For more information on these studies, see Lipsey et al. (2002).

A common factorial design in program evaluation research is a two-group repeated measures design. This design has a between-subjects factor that reflects group assignment, such as treatment and control, and a within-subjects factor reflecting time, such as baseline and posttest or baseline, posttest, and follow-up.

Factorial designs can be analyzed using ANOVA methods as well as regression-based methods. The former simplifies the testing of interactions, although both share a common underlying statistical framework. Because of the dependent nature of the data, within-subjects factors must be handled differently than between-subjects factors. We will first explore the application of ANOVA to a fully between-subjects factorial design before addressing the complexities of designs with a within-subjects factor.

### *Two-Way ANOVA for Between-Subjects Designs*

In a two-way ANOVA, we can have three hypotheses. The first two relate to the main effects of each factor. The third relates to the interaction between the two factors and addresses whether the effect of one factor is conditional on the levels of the other factor. This is best illustrated with a simple example. Table 9.1 presents the means for a fictitious 2 by 3 factorial design, with Factor A having 2 levels and Factor B having 3 levels. The sample size for each cell is 30 for a total sample size of 180.

The main effect for Factor A is the difference between the means for the two levels or conditions of this factor, that is, 40 versus 25. These are the means for levels 1 and 2 of this factor, collapsing over the levels of Factor B. Main effects are sometimes called marginal effects as they are the effect *in the margin*. The main effect for Factor B is the differences across the three means for Factor B, collapsing over the levels of Factor A, or the differences

**Table 9.1**

Means for a 2 by 3 factorial design

FACTOR A	FACTOR B			TOTAL
	1	2	3	
1	30	40	50	40
2	30	25	20	25
Total	30	32.5	35	32.5

between the means of 30, 32.5, and 35 in this fictitious example. We can express this as a linear model, shown below:

$$\gamma_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk},$$

where  $\beta_k$  is the deviation from the overall mean for each level,  $k$ , of the second factor, and  $(\alpha\beta)_{jk}$  reflects the interaction or deviation of each cell after accounting for the main effects.

The ANOVA estimates a separate  $F$ -statistic for each of the above effects based on the ratio of the variance (mean squares) for each factor relative to the mean-square residual. Each of these mean squares and associated  $F$ -values is computed using the equations below:

$$MS_A = \frac{na \sum (\bar{y}_j - \bar{y})^2}{a-1},$$

$$MS_B = \frac{nb \sum (\bar{y}_k - \bar{y})^2}{b-1},$$

and

$$MS_{AB} = \frac{n \sum \sum (\bar{y}_{jk} - \bar{y}_j - \bar{y}_k + \bar{y})^2}{(a-1)(b-1)},$$

where  $\bar{y}_j$  is the mean for each condition of Factor A,  $\bar{y}_k$  is the mean for each condition of Factor B,  $\bar{y}_{jk}$  is the mean for each combination of the conditions of Factor A and B,  $n$  is the sample size for each cell of the factorial design,  $a$  and  $b$  are the number of conditions for Factors A and B, respectively. Note that we are assuming a balanced design, that is, a design with equal sample

sizes in each cell. See below for a discussion of the implication of an unbalanced design.

The mean square within or residual is the pooled variability within each cell of the factorial design, computed as follows:

$$\text{MS}_{\text{within}} = \frac{\sum \sum \sum (y_{ijk} - \bar{y}_{jk})^2}{ab(n-1)},$$

where  $y_{ijk}$  is each observation. The three  $F$ -values for the ANOVA are the ratios of the respective mean squares over the mean squares within:

$$F_A = \frac{\text{MS}_A}{\text{MS}_{\text{within}}},$$

$$F_B = \frac{\text{MS}_B}{\text{MS}_{\text{within}}},$$

and

$$F_{AB} = \frac{\text{MS}_{AB}}{\text{MS}_{\text{within}}}.$$

The denominator degrees of freedom is  $ab(n-1)$  for all three. The numerator degrees of freedom is  $a-1$ ,  $b-1$ , and  $(a-1)(b-1)$ , respectively.

A significant main effect indicates that the means for that factor vary. A significant interaction indicates that the effect of one factor is conditional on the levels of the other factor. In the above example, a significant interaction would indicate that the effect of Factor A differs across levels of Factor B and vice versa. We can see that this is case in Table 9.1. The effect of Factor A is 0, -15, and -30 across levels 1, 2, and 3 of Factor B. Thus, the effect of A clearly depends on the level of Factor B. A common mistake in interpreting an interaction is to state that it is occurring at a specific location. For example, it may be intuitive to think that the interaction is at level 3 (or levels 2 and 3) of Factor B because this is the location of the biggest difference. This is incorrect. The interaction reflects that there is significant variability in the effect of A across levels of B or that the effects of B differ between the two groups of A. For example, the effect increases across the levels of B for group 1 of Factor A, whereas it decreases across the levels of B for group 2 of Factor B.

*An Example: Perceptions of Children During a Police Interrogation*

Redlich et al. (2008) conducted a laboratory-based factorial design experiment of the perceptions of a child interviewed by police. In this study, participants read a transcript of a police officer questioning a child. The design had three manipulated factors, each manipulating a different characteristic of the transcript: whether the child was a victim or a suspect (Factor A), the age of the child (7, 11, and 14; Factor B), and whether the child admitted involvement in the incident (never admitted or admitted and recanted; Factor C). Thus, this is a 2 by 3 by 2 factorial design. The 229 study participants were randomly assigned to one of the 12 conditions produced by the combinations of these three factors. There were four dependent variables: child credibility, police fairness, child interview understanding, and child suggestibility. The published study used a multivariate ANOVA to provide an omnibus test of the effects across all four of these dependent variables as well as separate ANOVA models for each. The analyses also included an observed factor of participant sex (male or female) based on explicit a priori hypotheses related to this variable. Below, we focus on just one ANOVA model using the study participants' perceptions of the child's credibility and exclude the observed factor of sex to simplify the analysis for illustrative purposes. As such, our results differ slightly, but not meaningfully, from those of the published study.

The two-way ANOVA can easily be extended to a three- or four-way ANOVA. Note that the sample size requirements grow multiplicatively as you add factors. You want to ensure that you have an adequate sample size per condition. The number of possible interactions also increases as you add factors. In this example, there are two-way interactions between each pair of factors and a three-way interaction among all three factors. The latter was suppressed from the analysis as our interest here is on the three main effects and the associated two-way interactions.

The child's credibility score ranged from 1.375 to 5.375 with a mean of 3.50 and a standard deviation of 0.86. The three-way ANOVA results are shown in Table 9.2. Two of the three main effects are significant. The younger the child, the more credible he seemed (means were 3.65, 3.60, 3.27 for the age 7, 11, and 14 conditions). Scenarios where the child never admitted involvement in the incident were perceived as more credible than where the child admitted and then recanted involvement (means were 3.75 and 3.28, respectively). Whether the child in the scenario was a victim or a suspect was unrelated to perceptions of credibility (means of 3.54 and 3.47, respectively).

Two of the three interactions were significant. Interactions are easier to interpret graphically. All three of these interactions are shown in Fig. 9.4. The interaction between age and admission status shown in panel (a) is not significant. Notice that we can see the main effect of age (the lines decrease

**Table 9.2**

Three-way ANOVA for perception of child's credibility

EFFECT	SUMS OF SQUARES	df	MS	F	p
Status	0.189	1	0.189	0.293	0.589
Admit	12.076	1	12.076	18.672	<0.005
Age	7.711	2	3.856	5.962	0.003
Status by admit	3.483	1	3.483	5.385	0.021
Status by age	4.144	2	2.072	3.204	0.043
Admit by age	1.608	2	0.804	1.243	0.290
Within	141.643	219	0.647		

as age increases) and admission status (never admitted is higher than admitted and recanted). However, the lines are roughly parallel with only a slightly larger effect for admit at age 7 relative to ages 11 and 14. The interaction for child status (victim or suspect) and admissions status is shown in panel (b) and is significant. We see that the effect of admission status is larger for victims than for suspects. Flipping this around, we see that victims who never admit are seen as slightly more credible than suspects who never admit but that this is opposite for victims and suspects who admit and then recant.

Notice that degrees of freedom for age and interactions involving age have two degrees of freedom in the numerator. If this model were expressed as a linear regression model, each of these effects would have two regression coefficients. This illustrates an advantage of ANOVA over regression: It provides the overall effect for each factor even when the factor has 3 or more categories.

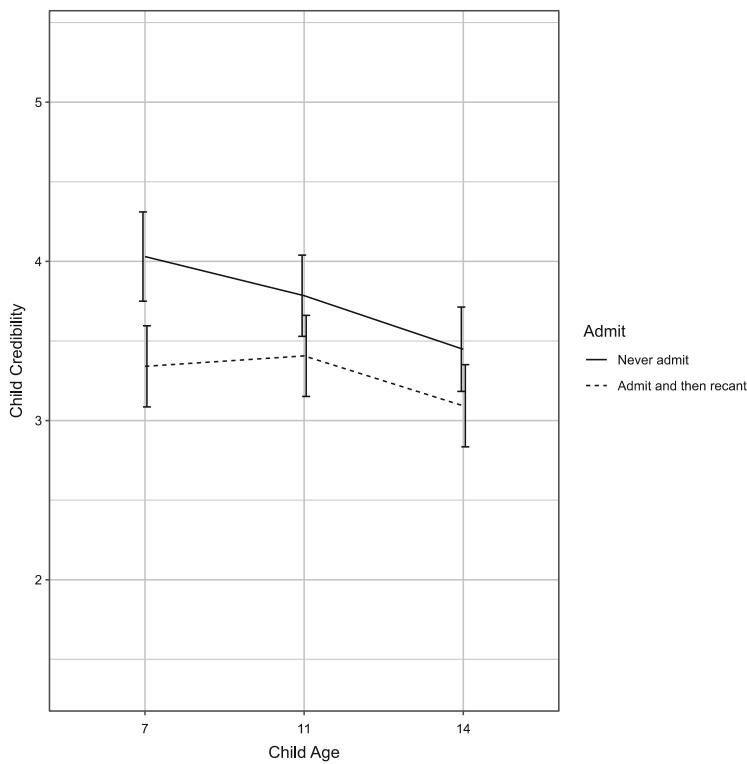
### Mixed Within- and Between-Subjects Factorial Designs

Experimental designs often mix within- and between-subjects factors. Recall that a within-subjects factor is one where each subject or unit of interest is exposed to each level of a factor. In field experiments, a common example is repeated measurements of the outcome, such as measures taken at pretest, posttest, and follow-up. In laboratory-based experiments, it is common to have experimental manipulations that are within-subjects. This only makes sense, of course, if any effects of an experimental manipulation are short-lived. For example, a common design for studies of the effect of sugar on children has the same child come into the laboratory on different days to perform a set of tasks or to be observed in free play following a lunch that is either high in sugar or low in sugar (For a meta-analysis of such studies, see Wolraich et al. 1995). In these designs, the sugar level is a within-subjects factor.

Any analysis of the data from an experimental study with a within-subjects factor must take into account the dependent nature of the data. ANOVA can handle completely within-subjects factorial designs as well as

Figure 9.4

*Mean perception of child's credibility by condition from the Redlich et al. (2008) study (a-c)*

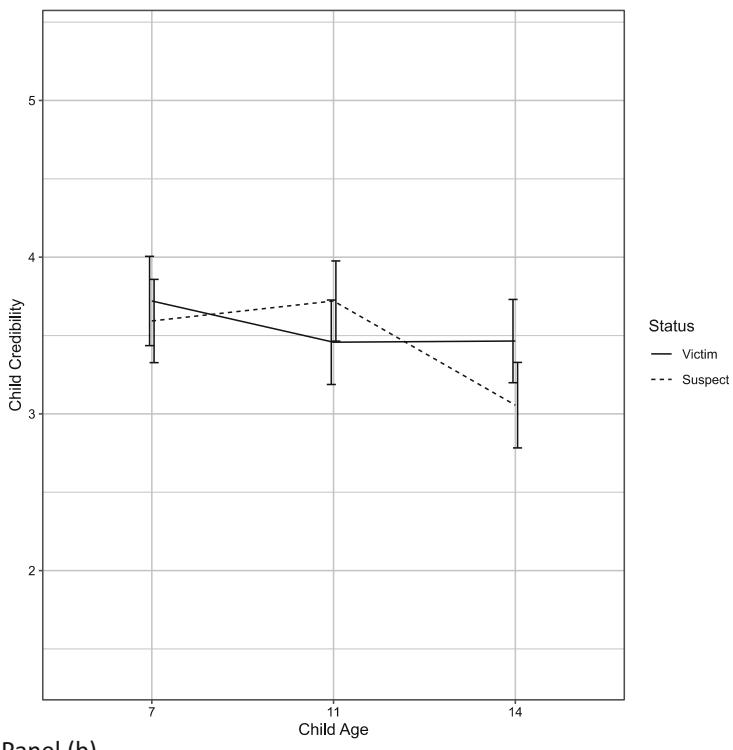


Panel (a)

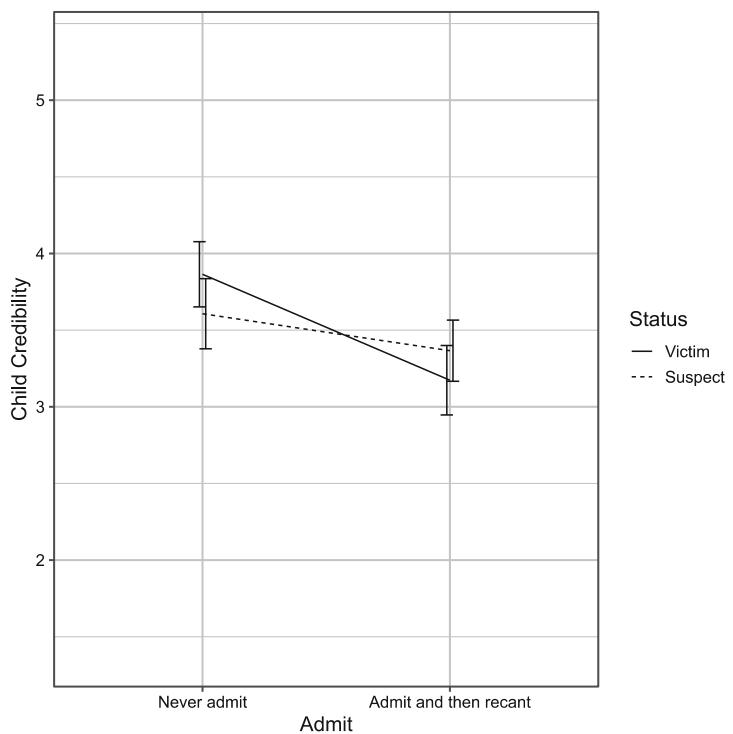
mixed within- and between-subjects designs. Alternatively, the multi-levels models discussed in Chap. 7 can be used. The ANOVA models for within-subjects' factors are a natural extension of the paired *t*-test. We illustrate below how to conduct such an analysis for the simple case of a 2 by 2 within- and between-subjects design, but the principles are the same for more complex designs.

The main difference in the ANOVA with a within-subjects factor compared to a factorial ANOVA with all between-subjects factors is changing the denominator for the *F*-test, called the error term, to reflect the nested or clustered nature of the data. For a 2 by 2 ANOVA with A as the between-subjects factor and B as the within-subjects factor, the *F*-tests are computed as:

$$F_A = \frac{MS_A}{MS_{\text{subjects}}},$$

**Figure 9.4***(continued)*

Panel (b)



Panel (c)

$$F_B = \frac{MS_B}{MS_{B \times \text{subjects}}},$$

and

$$F_{AB} = \frac{MS_{AB}}{MS_{B \times \text{subjects}}}.$$

Note that the numerators remain the same as defined previously for the between-subjects ANOVA but that the denominators have changed. Equations (9.8) and (9.9) provide the formulas for these new denominators:

$$MS_{\text{subjects}} = \frac{b \sum (\bar{y}_{ij} - \bar{y}_j)^2}{N - a}, \quad \text{Equation 9.8}$$

$$MS_{B \times \text{subjects}} = \frac{nb \sum (y_{ijk} - \bar{y}_{ij} - \bar{y}_k + \bar{y}_j)^2}{(N - a)(b - 1)}, \quad \text{Equation 9.9}$$

where  $y_{ijk}$  is the observation for each subject within each condition,  $\bar{y}_{ij}$  is the mean for each subject,  $\bar{y}_j$  is the mean for each level of A,  $\bar{y}_k$  is the mean for each level of B,  $N$  is the number of subjects,  $a$  is the number of conditions for factor A, and  $b$  is the number of conditions for factor B.

We will illustrate this with some fictitious data from a 2 by 2 design reflecting treatment (treatment and control) and time (pretest and posttest). The outcome is a scaled variable with a range from 1.43 to 8.29. Lower scores are more desirable. The treatment and control group pretest means were 5.16 and 5.32, respectively, whereas the posttest means were 4.66 and 5.61. Thus, we can see that the treatment group decreased from pretest to posttest, whereas the control group increased.

The ANOVA results are shown in Table 9.3 and show that the interaction between treatment and time is significant [ $F(1, 66) = 6.747, p = 0.012$ ]. This indicates a statistically significant difference in the change over time between the groups, favoring the treatment condition. Notice that in this design, the interaction effect is our test of the treatment effect and is a difference-in-differences analysis.

An alternative analytic approach with within-subjects designs is the use of multilevel or mixed-effects regression, discussed in Chap. 7. There are two advantages of the multilevel modeling approach. First, it can handle missing observations on the within-subjects factor. For example, if you have

**Table 9.3**

Two-way ANOVA with a within-subjects (time) and between-subjects (treatment) Factor

EFFECT	SUMS OF SQUARES	df	MS	F	p
Treatment	10.677	1	10.677	3.154	0.080
Subjects	223.44	66	3.385		
Time	0.399	1	0.399	0.513	0.476
Time by treatment	5.245	1	5.245	6.747	0.012
Subjects by time	51.31	66	0.777		

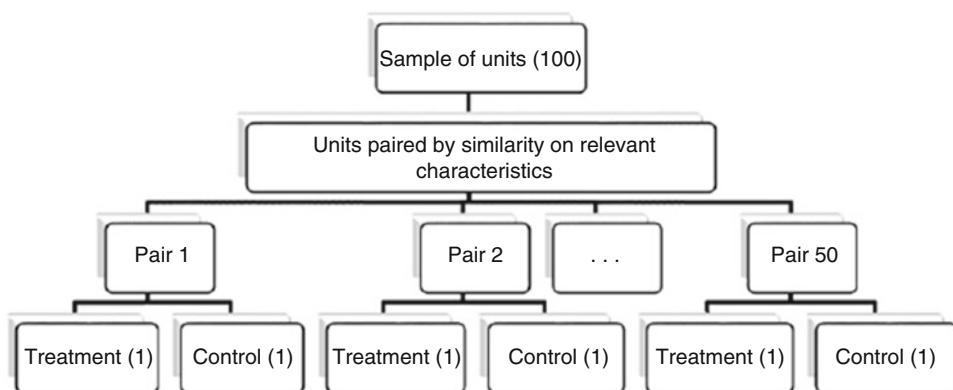
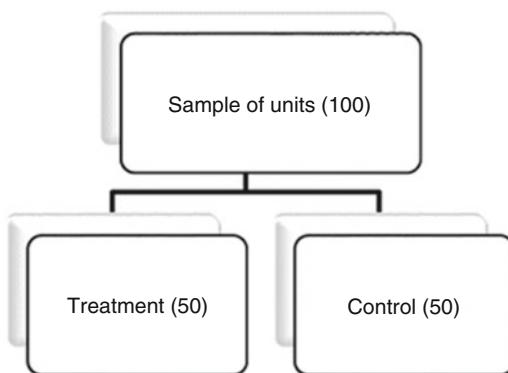
four repeated measurement time points, ANOVA would require each subject to have data for each time point. However, multilevel modeling can handle missing data of this type. In ANOVA, only cases with full data can be used. Second, multilevel modeling can handle dependent variables not suitable for ANOVA and OLS regression, such as counts and binary variables.

## Block Randomized Designs

---

A **block randomized** design is a special case of a factorial design where one factor is a baseline characteristic on which participants or other units of study are blocked. Randomization into the experimental factor then occurs within each level of the blocking factor. The first such study that we are aware of in criminology is the Cambridge-Somerville Youth Study (See Powers and Witmer 1951). In that study, problem youths were paired on age, social background, biological somatype, and temperament. Each of these pairs represents a block. A matched pairs design such as this is blocking taken to its natural limit. The researchers used matched pairs because the experimental treatment was lengthy and complex, so they sought to maximize the equivalence of the comparisons they could make. Their design is illustrated in Fig. 9.5. The researchers matched the youths into pairs on these characteristics and then randomly allocated them within the pairs into treatment and control conditions.

What advantage does this approach give over naïve or simple randomization? Naïve randomization, such as the two-group randomized design discussed earlier and illustrated in Fig. 9.6, is a common approach to field-based program evaluation experiments in criminal justice. This design assigns the total sample under study to treatment or control conditions without restrictions. Every subject in this case has an equal probability of being assigned to either the treatment or the control condition (note that there are statistical advantages to ensure that this results in equal sample sizes between groups). Naïve randomization relies on the assumption that

**Figure 9.5***Fully blocked (matched pairs) random assignment***Figure 9.6***Naïve (balanced) random assignment*

there are no systematic reasons for the treatment and control subjects to differ (since every subject had an equal probability of assignment to each condition), a key *raison d'être* for experimental studies in the first place. But it does not guarantee equivalence, simply that any differences will be the result of chance or the randomization process (i.e., that any nonequivalence is random and thus can be estimated using statistical probability distributions). When samples are large, this assumption is reasonable because large differences between the groups are unlikely by chance.

But why shouldn't we increase equivalence if we can, especially in small studies where chance differences between control and treatment groups might be large in the case of naïve randomization? Fully blocked randomized designs like the Cambridge-Somerville Youth Study assume that we have knowledge about the subjects or the units in an experiment that can help us create equivalence on factors that are related to the outcomes observed. Age and social background were considered key predictors of delinquency by the Cambridge-Somerville researchers, and their introduction as factors to match the youths in the study was seen as a direct way of making sure that the treatment and control conditions were similar on important influences of treatment success.

However, the benefit of equivalence gained through blocked randomization comes at a statistical price. For each limitation (i.e., block) on randomization in a study, there is a fine in terms of degrees of freedom. For example, in the Cambridge-Somerville Youth Study, 650 boys were matched into pairs. In a naïve randomization design with 325 cases per group, the study would have had 648 degrees of freedom for the statistical test of treatment effect ( $n_1 + n_2 - 2$ ). Using the matched pair design, the degrees of freedom of the tests declined to 324 ( $n_{pairs} - 1$ ).

The loss of degrees of freedom is meaningful because it changes the distribution of the test statistic. For example, as illustrated by Eq. (9.10) (dependent samples) and (9.11) (independent samples) below, in a *t*-test, the estimated standard deviations are divided by the degrees of freedom. This means that as the degrees of freedom of a test gets smaller the *t*-value observed also gets smaller:

$$t = \frac{\bar{x}_d}{\sqrt{\frac{s_d^2}{df}}} \quad \text{Equation 9.10}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{df}\right)\left(\frac{n_1+n_2}{n_1n_2}\right)}} \quad \text{Equation 9.11}$$

In turn, the value of the statistic needed to achieve statistical significance will be larger as the degrees of freedom for a test gets smaller (see the *t* distribution example in Table 9.4 below). This difference is not meaningful in the case of relatively large sample studies. For example, in the Cambridge-Somerville experiment, with 324 degrees of freedom, the critical value of the *t*-test (with standard criteria of  $p < .05$  and a non-directional test) is about 1.967, almost the same as the 1.960 in the *z* normal distribution without adjustment. But when the degrees of freedom is reduced to 100, the critical value for the *t*-test becomes 1.984 and at 50 degrees of freedom,

**Table 9.4**

Critical values for the  $t$ -distribution (two-tailed,  $\alpha = .05$ )

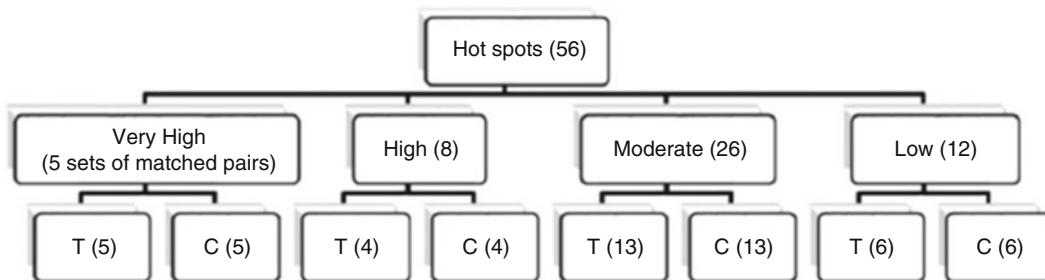
**DEGREES OF FREEDOM  $t$  CRITICAL VALUE (TWO-TAILED,  $\alpha = .05$ )**

10	2.228
20	2.086
50	2.009
100	1.984
200	1.972
324	1.967
500	1.965
648	1.964

2.009. (Recall that  $t$  is just a special case of  $F$  with 1 degree of freedom in the numerator and  $t^2 = F$ . Thus, this discussion applies equal to the ANOVA.)

The balance between loss of degrees of freedom and greater equivalence depends on the strength of the relationship between the blocking factor and the outcome. This is the case because the benefits of the blocked randomized design are greatest when each loss of degrees of freedom is accompanied by a gain in the equivalence of the treatment and control conditions on factors that are related (i.e., correlated) to treatment outcomes. If treatment outcomes are conditioned by such factors, then blocking will decrease the heterogeneity of outcomes in the study. Looking at Eqs. (9.10) and (9.11) above, this would mean that the numerators of the standard errors are made smaller and accordingly the  $t$ -values observed are larger. The same would be true for a two-way ANOVA where the blocking factor would pull variance from the within groups or residual mean squares, increasing the  $F$  for the experimental or treatment factor. This makes intuitive sense because if the groups are more similar in terms of what would have been expected absent treatment, then it should be easier to identify a treatment outcome. In statistical terms, there is likely to be less noise in identifying that outcome. In the case of a blocked design in which treatment outcomes were not related to the blocking factors, the standard deviations would remain the same as in a naïve design, while there would be a loss of degrees of freedom. This would mean that a price was paid for the blocked design without a corresponding benefit. Stated simply, to gain benefit from a blocked design, the blocking variable must be nontrivially correlated with the outcome.

In the Cambridge-Somerville study, blocking was taken to its maximum by creating pairs such that each block or pair had one treatment and one control youth. This effectively cut the degrees of freedom in half. Creating such complete blocking may be risky if you do not have sufficient knowledge regarding the correlation between the trait that you are blocking on and the outcome. Blocking, however, can be on any discrete characteristic,

**Figure 9.7***Blocked random assignment (Jersey City Drug Market Analysis Experiment)*

even a binary variable such as male and female. A block randomized design makes no assumptions regarding the number of blocks or groups identified at the outset. Rather, the number of blocks is determined by the researcher's assessment of the ability of known data to group units in ways that maximize their similarities on a key variable (or variables in the case of matched pairs). Cases are placed within the specified randomization blocks and then randomized within those blocks (see Fig. 9.7). The blocks do not have to be of equal size, but the number of cases in each block must be even to allow for equal randomization and balance within blocks (this can be relaxed so long as the assignment to conditions within blocks is random, although there are statistical advantages to balanced sample sizes between conditions within blocks, as discussed below).

The Jersey City Drug Market Experiment (JCE) provides a substantive example that illustrates the benefits of a block randomized design for place-based studies (Weisburd and Green 1995). The JCE evaluated an innovative drug enforcement strategy involving police crackdowns along with citizen and local business engagement in controlling crime at drug markets. A total of 56 high drug activity hot spots were randomly assigned in equal numbers to receive either the experimental program or regular, unsystematic enforcement on an ad hoc basis. Most of the drug market hot spots included fewer than four street segments and intersections, though two places included more than ten street segments. Police emergency calls for service for a variety of crime and disorder-related issues were measured for 7-month pre- and post-intervention periods. We focus below on three main outcome measures for disorder measured in the study: suspicious persons, public morals, and police assistance.

Knowing that there was considerable variation in criminal activity even across the sample of hot spots, the study authors were concerned at the outset that the prior level of crime would influence the effect of treatment. Given the small sample of drug hot spots that could be identified in Jersey

City, the authors were also concerned that naïve randomization might lead to nonequivalent groups. At the same time, there was concern that each loss of degrees of freedom in the experiment would substantially impact the results, since the total number of available cases was only 56. The solution in the JCE was to examine the distribution of both emergency calls for service and arrests and then to identify natural cutting points.

In this way, the researchers believed that they could gain greater equivalence between the groups without a large loss of degrees of freedom (28) that would have ensued if the fully blocked randomized design was adopted. The assumption here was that prior crime and disorder would have a general impact on the effects of treatment but would not be specific enough to distinguish sites in a way that would justify a fully paired randomized design. The researchers identified eight statistical blocks for randomization. The ten highest activity hot spots were randomized in pairs because of large gaps between them; these five pairs represented the five *very high activity* statistical blocks. Of the rest of the sample of hot spots, eight were grouped into a *high activity* block, 26 hot spots were classified as a *medium activity* block, and 12 were classified as a *low activity* block.

How much does this approach improve equivalence in a small  $N$  experiment? One approach to examining the contribution of statistical blocking to equivalence in the Jersey City Drug Market Analysis Experiment is to compare the equivalence gained between the treatment and control conditions on key baseline (pretest) measures. However, the Jersey City study is only one specific draw of randomization. By definition, any specific draw of a sample is going to be different from another draw. The statistical concern is whether on average, a draw using the block randomization procedure is likely to produce a more equivalent outcome than a draw using a simple randomization procedure. To examine this question, Weisbord and Gill (2014) developed 10,000 simulations of both naïve randomization and block randomization using the Jersey City data.<sup>4</sup> We focus on baseline calls for service for the three key disorder outcomes in the study (suspicious persons, public morals, and assistance). Table 9.5 reports the baseline information from the original study, the simulation results for the blocking approach, and the simulation outcomes of a naïve randomization approach. In the case of the simulations, we report the number of

---

<sup>4</sup>See Weisbord and Gill (2014). Stata programs were developed to run a randomization sequence (blocked or naïve) on the JCE dataset and then run a  $t$ -test comparing the treatment and control group means at baseline on the three outcomes of interest. Stata's simulation function was then used to run each program 10,000 times and create a dataset containing the group means,  $t$ -values,  $p$ -values, an indicator showing whether or not the two groups were significantly different at baseline for each iteration, and the absolute average mean group difference across all iterations.

**Table 9.5**

Calls for service at baseline in block and naïve randomizations of JCE data

	SUSPICIOUS PERSONS	PUBLIC MORALS	ASSISTANCE
<i>Original JCE block randomized data</i>			
Treatment mean (SD)	17.00 (16.15)	9.32 (10.58)	43.86 (43.40)
Control mean (SD)	17.93 (21.16)	9.11 (13.36)	42.11 (43.05)
Mean difference (SE)	0.93 (5.03)	0.21 (3.22)	1.75 (11.55)
<i>10,000 simulations block randomized data (N = 56)</i>			
Average mean difference across samples (SD)	2.67 (1.95)	1.29 (0.94)	7.11 (5.05)
Percent of samples with significant difference at baseline ( $p \leq 0.10$ )	0.93%	0.04%	2.18%
<i>10,000 simulations naïve randomized data (N = 56)</i>			
Average mean difference across samples (SD)	3.98 (2.89)	2.56 (1.86)	9.13 (6.71)
Percent of samples with significant difference at baseline ( $p \leq 0.10$ )	9.55%	9.89%	9.66%

samples that have significant differences at baseline and the overall absolute mean difference found in the 10,000 simulation samples.

Table 9.5 suggests the importance of blocking for creating equivalence. While the Jersey City randomization was a relatively lucky draw, it is clear from Table 9.5 that the procedure used was likely to produce much more equivalent groups than a simple randomization procedure. In the 10,000 simulations of the JCE block randomization procedure, less than 1% (0.93%) produced significantly different outcomes ( $p < .10$ ) for treatment and control conditions at baseline for suspicious persons calls, 0.04% for public moral calls, and rough 2% for assistance calls. In contrast, using the simple randomization approach on the same 56 cases, roughly 10% (9.55%) of the samples produced significant differences for suspicious person calls, 9.89% for public moral calls, and 9.66% for assistance calls.<sup>5</sup> These differences are of substantial magnitude and are also reflected in the average mean difference across all of the simulation samples. For suspicious persons the mean differences were almost 50% larger in the naïve randomization sample, for public morals about twice as large and for assistance almost a third larger.

### Block Randomization and Statistical Power

Despite the distinct advantages of randomized studies, it is often difficult to gain a large number of cases in a randomized experiment. Sometimes, this is true because it is difficult to identify a large number of subjects who can be made eligible for randomization into treatment and control conditions.

---

<sup>5</sup>Of course, this is about what we would have expected given a .10 significance threshold and a fair randomization procedure. But the important point is that the block randomization approach allows us to do better.

Sometimes, this is the case because treatment conditions or data collection are expensive, and each new case will increase the cost of the study. These problems are particularly acute in place-based randomized trials since the number of places with a specific crime problem is generally limited (See Braga et al. 1999; Sherman et al. 1989). Moreover, place-based trials ordinarily demand significant treatment resources per site, and accordingly, it is expensive for agencies to treat a large number of sites at one time (See Boruch et al. 2004; Weisburd 2005).

Block randomized experiments, when designed correctly, not only create greater equivalence, they also increase the statistical power of the study. A simple or a naïve randomized experiment presents a model for understanding outcomes where systematic variation is determined only by treatment. Accordingly, the linear model can be expressed as follows:

$$\gamma_{ij} = \mu + \alpha_j + \epsilon_{ij}, \quad \text{Equation 9.12}$$

where  $\alpha_j$  is the variability due to treatment and  $\epsilon_{ij}$  is the residuals or random error associated with individual observations ( $i$ ) within each group ( $j$ ). Thus, we have partitioned the total variability into the influence of the treatment ( $\alpha_j$ ) and the error variability ( $\epsilon_{ij}$ ). Remember that the  $F$ -test for the statistical significance test of the treatment or experimental effect is the ratio of these two variances, with the error term in the denominator. As such, anything that reduces the error term without affecting the treatment effect will increase the test statistic.

With the introduction of a blocking factor, an additional source of variability is taken into account in the model, as illustrated in Eq. (9.13):

$$\gamma_{ijk} = \mu + \alpha_j + \beta_k + \epsilon_{ijk}, \quad \text{Equation 9.13}$$

where  $\beta_k$  is the variability in  $y$  associated with the blocking factor. Any variability associated with the blocking factor is removed from the error term. The stronger the relationship between the blocking factor and the outcome, the more variance is pulled out of the error term. This is the case because a fully balanced (equal sample size in each treatment and control and block combination) or partially balanced (equal sample size in the treatment and control conditions within each block but different across blocks) block randomized model forces any relationship between the treatment and the block to equal zero (that is, for the factors to be orthogonal or independent). Because they are independent, the inclusion of the blocking factor in the model will not impact the size of the treatment effect. Accordingly, any variability associated with the block effect will be drawn out of the error term for the model and because the error term is a key element of the denominator of the test statistic (the degrees of freedom

**Table 9.6**

Univariate analysis of variance for treatment and treatment-block effects (JCE)

ANALYSIS OF VARIANCE MODELS	SUSPICIOUS PERSONS	PUBLIC MORALS	ASSISTANCE
<i>Treatment-only model</i>			
MS <sub>group</sub> (df = 1)	516.0710**	129.0180	355.0180
MS <sub>error</sub> (df = 54)	118.0304	53.4094	275.4782
MS <sub>total</sub> (df = 55)	125.2675	54.7841	276.9244
<i>F(p) for group effect</i>	4.372 (.041)**	2.416 (.126)	1.289 (.261)
<i>Treatment and block model</i>			
MS <sub>group</sub> (df = 1)	516.0710**	129.0180*	355.0180
MS <sub>block</sub> (df = 7)	193.1537	114.7637	869.5800
MS <sub>error</sub> (df = 47)	106.8418	44.2715	186.9949
MS <sub>total</sub> (df = 55)	125.2675	54.7841	276.9244
<i>F(p) for group effect</i>	4.830 (.033)**	2.914 (.094)*	1.899 (.175)

Notes: \* $p < .10$ , \*\* $p < .05$ 

being a second key element), its reduction without any change in the treatment effect will lead to a more significant outcome (i.e., a more powerful statistical outcome) than a naïve model.

As an illustration, Weisbord and Gill estimated ANOVA models using just the treatment factor and separately with the treatment and blocking factors (Table 9.6). Following our assumptions, the total variability and variability due to treatment are the same in both models. The total variability in the model is constant irrespective of model specification, and the effect of treatment is not influenced because of the balanced randomization of cases within blocks. However, the error variance declines in the analyses that include blocking as a factor. Table 9.6 shows the mean squares (variances) and *F*-tests for these models. For suspicious persons, the decline is from 118.0304 to 106.8418, for public morals from 53.4094 to 44.2715, and for assistance from 275.4782 to 186.9949. Note as well that there is a corresponding decrease in the degrees of freedom of the error term in the block randomized design (from 54 to 47 in all the models), reflecting the price of this approach. Importantly, the loss of statistical power generated by the reduction of degrees of freedom of the error term is less than the gain in statistical power from the inclusion of the blocking in the design and randomization. When we combine treatment and block effects in the model, all three comparisons show larger *F*-statistics. The observed *p*-value for suspicious persons declines from .041 to .033, for public morals from .126 to .094, and for assistance from .261 to .175.

### Examining Interaction in a Block Randomized Experiment

Just as we examined interactions between manipulated factors in discussing multiway factorial designs, we can also explore interactions between an experimental or manipulated factor and a blocking factor.

What if we wanted to know whether the effect of treatment in the Jersey City Study varied across blocks in the study? The blocks reflected to a great degree the intensity of crime at the hot spots. Was there an interaction between treatment and block? Does the inclusion of an interaction term impact the significance of the results we examine?

In the case of a partially balanced design, like that in Jersey City, additional statistical complexities arise. A partially balanced design is balanced within each level of one factor relative to the other factor but unbalanced across the levels of the first. Partially balanced designs are common in block randomization where the number of persons or other units within each block is unbalanced, often reflecting the natural variation on this factor in the population from which the sample was drawn. Balance is typically maintained, however, for the manipulated factor for each level of the block. For example, the number of persons is unbalanced across blocks, but there is an equal number of persons within the treatment and control conditions within each block.<sup>6</sup>

There are three model types that we can use in estimating the significance of treatment and interaction effects often termed Type I, II, and III models. Most criminologists use Type III models and have little contact with Type I and Type II models (not to be confused with Type I and II error). However, these latter methods for estimating *F* tests become important in the case where the researcher is estimating an interaction. When only including the main effects, the three model types will yield the same significance statistics for treatment impacts. In a partially balanced design, Type I and Type II will be equivalent, but the results for the Type III model may differ substantially from those if the blocks are of very different sizes. Type I and Type II use weighted means, while Type III uses unweighted means, where the weights are the cell sample sizes. The difference between these model types is that Type I is a sequential approach where each factor in the model is conditioned only on variables already in the model (i.e., adjusted for the prior factor), whereas a Type II model is partially sequentially, with factors or variables conditioned for all effects at the same level or lower (all main effects, main effects and two-way interaction, etc.). These model types are illustrated below for a two-factor model. Notice that the test of the interaction effect is the same for each model type, but the test associated with each main effect is different.

---

<sup>6</sup>In factorial experimental designs, it is ideal to have fully balanced designs. This both simplifies the analysis, as explained below, and maximizes statistical power given a fixed sample size. For block randomized designs, the sample sizes across the levels of the blocking factor are typically unequal. However, it is ideal to ensure balance on the experimental or treatment factor within each level of the blocking factor. That is, block randomized designs are ideally at least partially balanced.

**Type I Models (Sequential)**

SS(A) = SS for factor A

SS(B | A) = SS for factor B, conditioned on factor A

SS(AB | A, B) = SS for the interaction of A and B, conditioned on the main effects for A and B

**Type II Models (Partial)**

SS(A | B) = SS for factor A, conditioned on factor B

SS(B | A) = SS for factor B, conditioned on factor A

SS(AB | A, B) = SS for the interaction of A and B, conditioned on the main effects for A and B

**Type III Models (Simultaneous)**

SS(A | B, AB) = SS for factor A, conditioned on factor B and the interaction of A and B

SS(B | A, AB) = SS for factor B, conditioned on factor A and the interaction of A and B

SS(AB | A, B) = SS for the interaction of A and B, conditioned on the main effects for A and B

In a reanalysis of the Jersey City data conducted by Weisburd, Wilson, and Mazerolle, they compared the results across different model approaches (Weisburd et al. 2020). It is apparent from Table 9.7 that there are significant interaction effects for several of the outcomes. In the case suspicious persons, the interaction effect is highly significant ( $p < .001$ ). To interpret the interaction in this case, one would need to compare the relative treatment effects across blocks. In this case, it is clear that the largest impacts in the study are found in the blocks with the highest number of crimes. In some sense, this is precisely what the police in Jersey City wanted

**Table 9.7**

Significance ( $p$ -values) of main effects and interactions for selected outcomes for the Jersey City study under different sums of squares models (Type I vs. Type III)

OUTCOME	GROUP MAIN EFFECT <sup>a</sup>		
	TYPE III SS <sup>b</sup>	TYPE I SS <sup>c</sup>	INTERACTION <sup>d</sup>
Disorder (total)	.003	.088	.015
Nuisance	.060	.403	.039
Suspicious persons	.000	.002	.000
Public morals	.016	.037	.067
Assistance	.026	.074	.081

<sup>a</sup>One-tailed  $p$ -values for group main effects. Interaction  $p$ -values are two-tailed given that an  $F$  with 2 or more degrees of freedom in the numerator is nondirectional (i.e., it reflects the degree to which the individual block effects differ from the overall main effect in either direction)

<sup>b</sup>Model includes block, group, and block by group interaction

<sup>c</sup>Model includes block, group, and block by group interaction, entered into the model in that order

<sup>d</sup>Interaction between block and group. This is unaffected by sums of squares model type

to achieve, since the problems in such places are much more consequential. But as can also be seen in the table, the *p*-value for the main effect of group (hot spot versus traditional patrol) varies greatly between Model I and Model III approaches. This is the case because of the imbalance in the number of cases in each block.

In general, caution should be used in interpreting the main effects of treatment in the presence of an interaction. Even in a balanced design, the interaction indicates differential impacts across the different blocks, which makes the main effect either uninterpretable or at a minimum conditional on block. The choice of model type will depend on the assumptions that the researchers bring. In general, when the blocks are highly imbalanced, it will be preferable to use the Type I or Type II method. Alternatively, the researchers might wish to treat each block as equally important and that any difference in the sample sizes across blocks does not reflect the underlying population distribution of block stratification. In that case, the Type III method can be justifiable. The bottom line is that including an interaction in a partially balanced design will raise key statistical issues that must be examined carefully.<sup>7</sup>

## Using Covariates to Increase Statistical Power in Experimental Studies

---

Another technique for increasing statistical power in experimental studies follows the statistical logic of block randomization but does not balance the blocking characteristics at the outset. It relies heavily on the logic of randomization that we have already described. As noted earlier, if the cases are randomized to treatment and control conditions, then we can assume that there is no correlation between treatment and possible confounding factors. That means that the inclusion of additional covariates in an analysis will not, in theory, affect the estimate of the treatment effect. Since that is the case, we should in theory be able to include covariates without creating any bias in our assessment of the influence of the experimental variable. And in turn, we should gain a power benefit because such variables may help us reduce the error variance in our model.

---

<sup>7</sup>For a factorial experiment where two (or more) factors are manipulated, a Type III model will usually be preferred. In this situation, the unbalanced nature of the design is merely an experimental artifact and should be small in magnitude. Thus, any difference in the sample sizes across cells is random, and giving each cell equal weight in the analysis makes the most sense. That is, conceptually, we are interested in the effects that would be estimated if the design were balanced. However, if the main effects for a Type II versus Type III ANOVA differ, it is wise to explore why that is the case and carefully assess which makes most conceptual sense for your research question.

Let us again use the approach of examining the sums of squares of our equation. Suppose we convert the JCE to a simple naïve randomization sequence. In this case, our model includes treatment and error as the only variables. If we add covariates, the error term for the model should decline, while we have no reason to expect that the effect of treatment (i.e., group) will change. That is, adding covariates should reduce the error term but leave the treatment effect mostly unaffected. However, because we did not block on the covariate, ensuring at least a partially balanced design, there may be an observed correlational between the covariate and treatment assignment, even though in the population the correlation would be zero, given randomization. As an example, let us add as covariates variables that should be related to the dependent variables: The pre-experiment calls for service for robbery and aggravated assault (collectively the baseline violent crime calls for service) and the baseline calls for service for each respective outcome measure. Thus, for each outcome, we include three covariates: robbery at baseline, aggravated assault at baseline, and the outcome at baseline (e.g., suspicious person calls at baseline for the suspicious person outcome).

In Table 9.8, we show the results using the simple naïve design, as well as the results we would gain taking into account the three covariates for each outcome. As can be seen from the table, the statistical significance of the results including the covariates is considerably lower than when no covariates are included. For public morals, for example, the *p*-value for the group effect has dropped from a nonstatistically significant .126 in the naïve example to a significant .041 when including the covariates. For all three outcomes, we have substantially lowered the  $MS_{\text{error}}$  by adding the covariates. Even though we paid a price in degrees of freedom for using

**Table 9.8**

Univariate analysis of variance for treatment and treatment-covariate effects (JCE)

UNIVARIATE ANALYSIS OF VARIANCE MODELS	SUSPICIOUS PERSONS	PUBLIC MORALS	ASSISTANCE
<i>Treatment-only model</i>			
$MS_{\text{group}}$ (df = 1)	516.0710**	129.0180	355.0180
$MS_{\text{error}}$ (df = 54)	118.0304	53.4094	275.4782
$MS_{\text{total}}$ (df = 55)	125.2675	54.7841	276.9244
$F(p)$ for group effect	4.372 (.041)**	2.416 (.126)	1.289 (.261)
<i>Treatment and covariate model</i>			
$MS_{\text{group}}$ (df = 1)	559.666**	162.019**	638.146*
$MS_{\text{covariate}}$ (pre-outcome) (df = 1)	678.443**	751.986**	305.731
$MS_{\text{covariate}}$ (pre-robbery) (df = 1)	752.979**	160.373**	2403.850**
$MS_{\text{covariate}}$ (pre-assault) (df = 1)	55.281	15.301	405.114
$MS_{\text{error}}$ (df = 51)	104.604	36.7676	188.1661
$MS_{\text{total}}$ (df = 55)	125.2675	54.7841	276.9244
$F(p)$ for group effect	5.350 (.025**)	4.407 (.041**)	3.391 (.071*)

Notes: \* $p < .10$ , \*\* $p < .05$

three covariates, the benefit in terms of reducing the error and increasing the significance of our group findings outweighs the cost.

However, we should be cautious in including these covariates. While we would expect, in theory, that the effect of treatment to remain similar between the simple model and the model with covariates, this is not always the case. This is largely true for *suspicious persons* and *public morals* where the  $MS_{group}$  remains fairly similar in both sections of Table 9.8. For assistance, however, there is a large increase in the  $MS_{group}$  potentially suggesting that we may have introduced some level of bias into the model with our choice of covariates.

As is apparent, there is much to be gained by including covariates in an experimental analysis. However, as in other statistical procedures, covariates can be manipulated in ways that affect the validity of your results. Randomization allows us to assume that there is no relationship between the covariate and the treatment or the variable interest. But this does not mean that there is not in the sample of interest a spurious relationship that is observed—for example in the case of Suspicious Persons in our analysis. In any randomization, there are likely to be some measures that by chance are related to the treatment. The researcher can in the end manipulate the impacts of treatment simply by cherry-picking covariates that not only decrease the error term but also increase spuriously the impact of treatment. Here, as in multiple regression models more generally, the researcher can *go fishing* until the result they are looking for is gained.

A general rule that will protect you from the danger of manipulation of results is for the researcher to define at the outset which covariates will be used in analyzing the outcomes. In this way, the researcher cannot manage results post facto on the basis of knowledge of sample characteristics. Clearly, one should not run a large number of regressions with different covariates included until a so-called good result is gained. The process of selecting variables before the results of an experiment are known is in our view a good rule to follow. But more generally, if an experiment has sufficient statistical power, the researcher should use the simple analysis approach, in which covariates are not included. This is the only way to guarantee that the results are not being manipulated in a way that might lead to spurious findings.

## Chapter Summary

---

Randomized experiments provide higher levels of **internal validity** than observational studies in terms of determining the impacts of a treatment or an intervention. Through the process of determining an **eligibility pool**

and using **randomization** to allocate the eligible participants or units to the **treatment** or **control group(s)**, researchers can better deal with the problem of confounding in **posttest measures** relevant to the dependent variable of interest. Randomized experiments have the highest possible internal validity as they allow us to assume that confounding causes of the dependent variable are not a concern. Since treatment has been allocated randomly, we can assume that possible **confounding factors** are not systematically related to treatment.

Randomized experiments can be simple, involving one or two conditions such as an experimental and control condition, or complex with multiple factors, each having two or more conditions or groups. Factors may also be **between-subjects** or **within-subjects**. In the former, participants or other units of study are randomly assigned to only one condition of a factor, whereas for the latter participants experience each condition. Experiments can also have a **factorial design** in which two or more factors (independent variables) are being examined.

Experiments that rely on **block randomization** provide significant advantages over experiments that use simple or naïve randomization, especially when samples are small as in the case of many place-based experiments. Block randomization can ensure greater equivalence of experimental and control conditions and will, if carried out correctly, increase the statistical power of studies. Block randomization also allows the examination of interaction between treatment and block, which can add important information regarding whether treatment varies across the stratifications represented by the blocks. However, the inclusion of interaction in a statistical model makes the interpretation of the main effect of treatment much more complex.

Another way of increasing the statistical power in experimental studies is with the inclusion of covariates; however, covariates should be used cautiously as they can allow the researcher to manipulate the results of the study.

## Key Terms

---

**Between-subjects** In an ANOVA, a between-subjects factor or independent variable is one for which each subject or observation is in only one of the categories that makes up the factor or independent variable.

**Block randomization** A type of randomization whereby cases are first sorted into like groups and then afterwards randomly allocated into treatment and control conditions.

**Confounding factors** Variables associated with treatments and/or outcomes that can bias overall results if not controlled for statistically or by research design.

**Control group** The group that eligible cases are randomly assigned to which does not receive the treatment or the intervention being evaluated. In many criminological experiments, the control group may receive existing interventions in contrast to the innovative treatment.

**Eligibility pool** Participants or units that are eligible for an experiment.

**Factorial design** A factorial design is a type of experiment with two or more factors. Each factor is an independent variable with two or more categories or conditions.

**Internal validity** Whether the research design has allowed for the impact of the intervention or the treatment to be clearly distinguished from other factors.

**Mean squares** In ANOVA, a mean square is a variance associated with each factor or element of a research design.

**Posttest measures** Analyses conducted by the researcher to determine if the intervention had any impact on the outcome measures of interest.

**Randomization** The process of randomly assigning members from the pool of eligible participants or units to the study conditions—often a treatment group and a control group.

**Treatment group** One group that eligible cases are randomly assigned to which receives the treatment or the intervention being evaluated.

**Within-subjects** In an ANOVA, a within-subjects factor or independent variable is one for which each subject or observation is in each of the categories that makes up the factor or independent variable, such as with a repeated measure.

## Symbols and Formulas

---

$r$	Correlation coefficient
$s$	Standard deviation
$n_j$	Group sample size
$a, b, c, d$	Cell frequencies of a 2 by 2 contingency table
$y_{ij}$	Observation of an individual ( $i$ ) within a group ( $j$ )
$\bar{y}_j$	Group ( $j$ ) mean
$N$	Total sample size
$a$	Number of categories for Factor A
$b$	Number of categories for Factor B

To compute the regression coefficient  $b$  for a treatment variable ( $t$ ) controlling for a confounding factor ( $x_1$ ):

$$b_t = \left( \frac{r_{y,t} - (r_{y,x_1} r_{t,x_1})}{1 - r_{t,x_1}} \right) \left( \frac{s_y}{s_t} \right)$$

With the basic assumption that we have enough knowledge to create equivalence of units in the treatment and control conditions:

$$b_t = (r_{y,t}) \left( \frac{s_y}{s_t} \right)$$

Chi-square for 2 by 2 frequency table:

$$\chi^2 = \frac{(ad - bc)^2(a + b + c + d)}{(a + b)(c + d)(a + c)(b + d)}$$

Linear model for one-way ANOVA:

$$y_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

Mean-squared between for one-way ANOVA:

$$MS_{\text{between}} = \frac{\sum n_j (\bar{y}_j - \bar{y})^2}{a - 1}$$

Mean-squared within for one-way ANOVA:

$$MS_{\text{within}} = \frac{\sum \sum (y_{ij} - \bar{y}_j)^2}{N - a}$$

$F$ -test for one-way ANOVA:

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$

Linear model for two-way ANOVA (between-subjects):

$$y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk}$$

Mean squares for factor A for two-way ANOVA (between-subjects):

$$MS_A = \frac{na \sum (\bar{y}_j - \bar{y})^2}{a - 1}$$

Mean squares for factor B for two-way ANOVA (between-subjects):

$$MS_B = \frac{nb \sum (\bar{y}_k - \bar{y})^2}{b - 1}$$

Mean squares for factor A by factor B interaction for two-way ANOVA (between-subjects):

$$MS_{AB} = \frac{n \sum \sum (\bar{y}_{jk} - \bar{y}_j - \bar{y}_k + \bar{y})^2}{(a - 1)(b - 1)}$$

Mean squares within for two-way ANOVA (between-subjects):

$$MS_{\text{within}} = \frac{\sum \sum \sum (y_{ijk} - \bar{y}_{jk})^2}{ab(n - 1)}$$

*F*-test for factor A for two-way ANOVA (between-subjects):

$$F_A = \frac{MS_A}{MS_{\text{within}}}$$

*F*-test for factor B for two-way ANOVA (between-subjects):

$$F_B = \frac{MS_B}{MS_{\text{within}}}$$

*F*-test for factor A by factor B interaction for two-way ANOVA (between-subjects):

$$F_{AB} = \frac{MS_{AB}}{MS_{\text{within}}}$$

*F*-test for factor A for two-way repeated measures ANOVA (A between-subjects and B within-subjects factor):

$$F_A = \frac{MS_A}{MS_{\text{subjects}}}$$

*F*-test for factor B for two-way repeated measures ANOVA (A between-subjects and B within-subjects factor):

$$F_B = \frac{MS_B}{MS_{B \times \text{subjects}}}$$

*F*-test for factor A by factor B interaction for two-way repeated measures ANOVA (A between-subjects and B within-subjects factor):

$$F_{AB} = \frac{MS_{AB}}{MS_{B \times \text{subjects}}}$$

Mean-squared for subjects for two-way repeated measures ANOVA (A between-subjects and B within-subjects factor):

$$MS_{\text{subjects}} = \frac{b \sum (\bar{y}_{ij} - \bar{y}_j)^2}{N - a}$$

Mean-squared for subjects by factor B interaction for two-way repeated measures ANOVA (A between-subjects and B within-subjects factor):

$$MS_{B \times \text{subjects}} = \frac{nb \sum (y_{ijk} - \bar{y}_{ij} - \bar{y}_k + \bar{y}_j)^2}{(N - a)(b - 1)}$$

## Exercises

---

- 9.1. Darnell randomly allocates 30 students to either a treatment group that receives new instructional program or a control group that receives standard instruction. After randomization, he compares the characteristics of the treatment and control groups and finds that the treatment group is significantly different from the control group in two characteristics (age, reading level).
  - (a) Do these significant differences between the treatment and control group indicate that randomization failed?
  - (b) What about Darnell's sample might explain these significant differences?
- 9.2. Mark finds in a simple regression analysis that a drug treatment program has a significant impact on reducing the likelihood that patients will relapse. The standardized beta coefficient is  $-0.50$ . Mark concludes that the drug treatment program is effective. Brent, however, argues that Mark's results are confounded because he did not account for patients' level of motivation. Brent notes that motivation and likelihood of relapse are highly related ( $r = 0.50$ ). He reruns the regression results controlling for the level of motivation and finds that the impact of treatment has declined. The new standardized beta coefficient is  $-0.125$ .
  - (a) Diagram the impact of treatment on likelihood of relapse based on Mark's initial result.
  - (b) Diagram the impact of treatment on likelihood of relapse using Brent's analyses.
  - (c) What is the level of bias Mark has introduced by not including this confounder? What is the estimated  $r$  between the level of motivation and treatment?
- 9.3. Darcy wants to test the effectiveness of a new police training program on domestic violence. She identifies the officers with the least knowledge of domestic violence and administers the training to this group because she believes that it will be most worthwhile since they have the most to learn. She tests this group on domestic violence knowledge before and after the training. She also tests a control group of officers who did not receive the training. She finds a major jump in knowledge in the trained officers compared to the nontrained officers and concludes that her training program was effective.
  - (a) Are Darcy's conclusions warranted? Are there any threats to internal validity in her research design?
  - (b) Design an alternative study to test the effectiveness of the training program that has a higher level of internal validity than Darcy's study.

- 9.4. Adrian is designing a randomized trial to examine the effectiveness of a program designed to reduce recidivism in offenders. He has a sample of 200 prisoners that will all be released from prison on the same day and can be randomly allocated to a treatment group receiving the program or a control group that does not receive the program.
- If Adrian uses a naïve randomization procedure, how many prisoners will be in each group? What will be the total degrees of freedom for the research design?
  - If Adrian uses a fully blocked randomization procedure, how many pairs of prisoners will be randomized? What will be the total degrees of freedom for the research design?
  - If Adrian wants to use a partially blocked randomization procedure, what might be one prisoner characteristic he uses to create statistical blocks? What are the statistical consequences if this prisoner characteristic does not end up being related to the effectiveness of the program?
- 9.5. Sharon is analyzing data from a large randomized trial of the impact of afterschool programs on juvenile delinquency. After completing the experiment, she has been considering adding a number of different covariates to her overall analysis to minimize the error and improve her ability to identify a treatment effect.
- Do you have any concerns about the approach Sharon is taking to analyzing the experimental data? If so, what would be a better approach?
  - If Sharon has chosen good covariates, what should happen to the  $MS_{total}$  in the model? What should happen to the  $MS_{error}$ ? What should happen to the  $MS_{group}$ ?
  - With the large sample size, the statistical power in Sharon's experiment is estimated to be about 0.9. Does this affect whether she should consider using covariates?

## Computer Exercises

The section below reviews how to conduct independent sample *t*-tests, one-way ANOVA, and two-way factorial (Type I, II, and III sums of squares) in SPSS, Stata, and R

### SPSS

#### *Independent Sample t-Test*

To conduct an independent sample *t*-test in SPSS, you can use the T-TEST function. The independent variable is specified after the GROUPS= argument. Then, in parentheses, you indicate how the dichotomous variable is coded. The name of the dependent variable, which is dv in the example below, is added after the

/VARIABLES= argument. The alpha level can also be adjusted with the /CRITERIA= argument.

```
T-TEST GROUPS=factor1(0 1)
/MISSING=ANALYSIS
/VARIABLES=dv
/CRITERIA=CI(.95).
```

### *One-Way ANOVA*

A one-way ANOVA is conducted in SPSS using the ONEWAY function. The dependent variable is specified first, and the independent variable is specified second, with the two being separated using the BY argument. The subcommand /STATISTICS DESCRIPTIVES can be added to produce a descriptive table, which includes means and standard deviations broken down by group. The confidence level can also be specified by the user.

```
ONEWAY dv BY factor1
/STATISTICS DESCRIPTIVES
/CRITERIA=CILEVEL(0.95).
```

### *Two-Way Factorial (Type I SS)*

A two-way ANOVA is conducted in SPSS using the UNIANOVA function. Similar to one-way ANOVA, the dependent variable is specified first, and the BY argument separates the dependent variable and the independent variable(s). The METHOD subcommand is used to specify the sums of squares type of the model. SSTYPE(1) denotes Type III sums of squares. An intercept can be included, and your desired alpha can be specified. The example below has two factors, a and b, as well as the interaction between the two factors. The subcommand /PRINT DESCRIPTIVE can be added to print descriptive statistics.

```
UNIANOVA dv BY factor1 factor2
/METHOD=SSTYPE(1)
/INTERCEPT=INCLUDE
/PRINT DESCRIPTIVE
/CRITERIA=ALPHA(0.05)
/DESIGN= factor1 factor2 factor1*factor2.
```

Alternatively, we can include one factor and one covariate, as well as the interaction between the two. Notice that the WITH argument has been added and the argument proceeds the covariate of interest.

```
UNIANOVA dv BY factor1 WITH x
/METHOD=SSTYPE(3)
/INTERCEPT=INCLUDE
```

```
/PRINT DESCRIPTIVE
/CRITERIA=ALPHA(0.05)
/DESIGN= factor1 x factor1*x.
```

### *Two-Way Factorial (Type II SS)*

It is simple to specify Type II sums of squares in SPSS. You just need to change the SSTYPE(#) to 2.

```
UNIANOVA dv BY factor1 factor2
/METHOD=SSTYPE(2)
/INTERCEPT=INCLUDE
/PRINT DESCRIPTIVE
/CRITERIA=ALPHA(0.05)
/DESIGN= factor1 factor2.
```

### *Two-Way Factorial (Type III SS)*

Similarly to the models above, you change the SSTYPE(#) to 3 for Type III sums of squares.

```
UNIANOVA dv BY factor1 factor2
/METHOD=SSTYPE(3)
/INTERCEPT=INCLUDE
/PRINT DESCRIPTIVE
/CRITERIA=ALPHA(0.05)
/DESIGN= factor1 factor2.
```

## **Stata**

### *Independent Sample t-Test*

The two-sample *t*-test for differences in group means is performed by using the **ttest** command:

```
ttest dv, by(factor1)
```

### *One-Way ANOVA*

To compute a one-way ANOVA model in Stata, you will use the **oneway** command. The output will present the ANOVA table. If the **tabulate** option is included on the command line, Stata will generate group means, standard deviations, and counts for the number of cases on the variable of interest. If you omit the **tabulate** option, **oneway** will simply generate an ANOVA table.

```
oneway dv factor1, tabulate
```

### *Two-Way Factorial (Type I SS)*

A two-way factorial with Type I sums of squares can be conducted in Stata using the **anova** command and by adding the **sequential** argument to the right of the comma. The dependent variable is specified first, followed by the independent variable(s). Interaction terms can be easily added by using **##** between the two factors. Remember, the order of the variables matters.

```
anova dv factor1 factor2 factor1##factor2, sequential
```

### *Two-Way Factorial (Type II SS)*

In Stata, you can conduct a two-way factorial with Type I and Type III sums of squares but not Type II. The only way to accomplish that in Stata is to run your models with main effects first and then the second with the main effects and interactions.

### *Two-Way Factorial (Type III SS)*

Conducting an ANOVA with Type III sums of squares is similar to conducting ANOVA with Type I sums of squares. You still use the **anova** function, but you will want to remove the sequential argument.

```
anova dv factor1##factor2
```

## R

### *Independent Sample t-Test*

The code to conduct an independent sample *t*-test is easy in R. You can do this by using the **t.test()** function, specifying the dependent variable (ratio/interval variable of interest) and then specifying the grouping variable (binary independent variable). The default is that equal variance will not be assumed.

```
t.test(dv ~ factor1, data= dataset_name)
```

If your grouping variable is not already defined as a factor, you will want to use the **as.factor()** function to transform it.

### *One-Way ANOVA*

The command **aov()** in base R will allow you to easily conduct a one-way ANOVA. You can use the **summary()** command to get more detailed results.

```
MyAnova <- aov(dv ~ factor1, data=dataset_name)
summary(MyAnova)
```

Note that R does not provide a table of means and standard deviations like SPSS and Stata do. There are multiple ways this can be done in R, but we are going to rely on the *dplyr* package to do this. Make sure the package is installed, and then, you can

run the following code to get the means, standard deviations, and number of cases for each group:

```
dv %>%
  group_by(factor1) %>%
  summarise(
    count = n(),
    mean_dv = mean(dv, na.rm = TRUE),
    sd_dv = sd(dv, na.rm = TRUE)
  )
```

### *Two-Way Factorial (Type I SS)*

To obtain an ANOVA with Type I sums of squares, you may use both the **lm()** and **anova()** functions. Interaction terms can easily be added by using: (a colon) between the two factors. Remember to make sure that your factors are of class factor in R. You can change it using the **factor()** function around the name of the factor, if needed.

```
MyAnova<-lm(dv ~ factor1 + factor2 + factor1:factor2,
              data=dataset_name)
anova(MyAnova)
```

### *Two-Way Factorial (Type II SS)*

One way of conducting Type II sums of squares in R is to use both the **lm()** function and **Anova()** function. Note that the **anova()** function and **Anova()** functions are different. The **Anova()** function that we demonstrate is from the *car* package so make sure to install and load the package before using the **Anova()** command. Within the **Anova()** function, specify **type=2** for Type II SS.

```
library(car)
MyAnova <- lm(dv ~ factor(factor1) + factor(factor2),
               data=dataset_name)
Anova(MyAnova, type=2)
```

### *Two-Way Factorial (Type III SS)*

ANOVA with Type III sums of squares is done similar to the Type II sums of squares from the *car* package where you use both the **lm()** and **Anova()** functions. As such, remember to load and install the *car* package first. You will need to change the *type* argument to **type=3**. However, we will first use the **options()** function and **contrasts=** argument for Type III sums of squares comparisons.

```

library(car)
options(contrasts = c("contr.sum", "contr.poly"))
MyAnova <- lm(dv ~ factor(factor1) + factor(factor2),
               data=dataset_name)
Anova(model, type=3)

```

### Problems

The questions below use two data files, both of which are available in SPSS (*ProbationExperiment.sav* and *PolicingBlockedRct.sav*) and Stata (*ProbationExperiment.dta* and *PolicingBlockedRct.dta*) formats. You may also import these datasets into R using `read_sav()` or `read_dta()` from the *haven* package. The *ProbationExperiment* data file contains data on a simulated randomized experiment where drug-involved probationers were randomly assigned to three different drug treatment regimens and two different intensity levels. The *PolicingBlockedRct* data file contains data on a simulated block randomized experiment where cities were randomly assigned to receive a new policing intervention or no intervention (control group).

1. Open the *ProbationExperiment* data file and do the following:
  - (a) Assess the *Group* variable that (denotes which participants were randomly assigned to which type of drug treatment regimen) and the *Intensity* variable (specifies the intensity of the treatment regimen that the participant was randomly assigned to it). Discuss whether the experimental design is balanced.
  - (b) Select the appropriate ANOVA model to assess whether there is an interaction between *Group* and *Intensity* on the effect of the number of positive drug tests (*PosDrugTests* variable). Write out your hypothesis about their relationship. Then, conduct the ANOVA model and write a summary of the results of your test (make sure to include the test statistic and *p*-value in your discussion).
  - (c) Now, rerun the model as an additive ANOVA model to obtain Type II sums of squares. Write a summary of the results of your test (make sure to include the test statistic and *p*-value in your discussion).
2. Open the *PolicingBlockedRct* data file and do the following:
  - (a) Assess the *Condition* variable that denotes which locations implemented the new policing intervention (with the other cities being randomized to the control group) and the *Block* variable, which specifies the assigned block. Discuss whether the experimental design is balanced.
  - (b) Select the appropriate ANOVA model to assess whether there is an interaction between *Condition* and *Block* on the effect of change in the crime rate (*ChangeCrimeRate* variable). Write out your hypothesis about their

relationship. Then, conduct the ANOVA model and write a summary of the results of your test (make sure to include the test statistic and *p*-value in your discussion).

- (c) Now, rerun the model as an additive ANOVA model to obtain Type II sums of squares. Write a summary of the results of your test (make sure to include the test statistic and *p*-value in your discussion).

## References

---

- Ariel, B., Sherman, L. W., & Newton, M. (2020). Testing hot-spots police patrols against no-treatment controls: Temporal and spatial deterrence effects in the London Underground experiment. *Criminology*, 58, 101–128. <https://doi.org/10.1111/1745-9125.12231>.
- Boruch, R., May, H., Turner, H., Lavenberg, J., Petrosino, A., De Moya, D., et al. (2004). Estimating the effects of interventions that are deployed in many places: Place-randomized trials. *American Behavioral Scientist*, 47(5), 608–633.
- Boruch, R., Snyder, B., & DeMoya, D. (2000). The importance of randomized field trials. *Crime & Delinquency*, 46(2), 156–180.
- Boruch, R. F. (1997). *Randomized experiments for planning and evaluation: A practical guide* (Vol. 44). Thousand Oaks, CA: Sage.
- Braga, A. A., Weisburd, D. L., Waring, E. J., Mazerolle, L. G., Spelman, W., & Gajewski, F. (1999). Problem-oriented policing in violent crime places: A randomized controlled experiment. *Criminology*, 37(3), 541–580.
- Campbell, D. T., & Boruch, R. F. (1975). Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In C. A. Bennett & A. A. Lumsdaine (Eds.), *Evaluation and experiment: Some critical issues in assessing social programs* (pp. 195–296). New York, NY: Academic Press.
- Cook, T. D., Campbell, D. T., & Day, A. (1979). *Quasi-experimentation: Design & analysis issues for field settings* (p. 351). Boston, MA: Houghton Mifflin.
- De Leon, G., Melnick, G., Thomas, G., Kressel, D., & Wexler, H. K. (2000). Motivation for treatment in a prison-based therapeutic community. *The American Journal of Drug and Alcohol Abuse*, 26(1), 33–46.
- Farrington, D. P. (1983). Randomized experiments on crime and justice. *Crime and Justice*, 4, 257–308.
- Feder, L., Jolin, A., & Feyerherm, W. (2000). Lessons from two randomized experiments in criminal justice settings. *Crime & Delinquency*, 46(3), 380–400.
- Flay, B. R., & Best, J. A. (1982). Overcoming design problems in evaluating health behavior programs. *Evaluation & the Health Professions*, 5(1), 43–69.
- Kirk, R. E. (2013). Research strategies and the control of nuisance variables. In *Experimental design: Procedures for the behavioral sciences* (pp. 1–30). Thousand Oaks, CA: Sage.
- Lipsey, M. W., Wilson, D. B., Cohen, M. A., & Derzon, J. H. (2002). Is there a causal relationship between alcohol use and violence? In *Recent developments in alcoholism* (pp. 245–282). Boston, MA: Springer.
- Maxwell, S. E., Delaney, H. D., & Kelley, K. (2017). *Designing experiments and analyzing data: A model comparison perspective*. London: Routledge.
- McCord, J. (2003). Cures that harm: Unanticipated outcomes of crime prevention programs. *The Annals of the American Academy of Political and Social Science*, 587(1), 16–30.

- Meissner, C. A., Redlich, A. D., Michael, S. W., Evans, J. R., Camilletti, C. R., Bhatt, S., & Brandon, S. (2014). Accusatorial and information-gathering interrogation methods and their effects on true and false confessions: A meta-analytic review. *Journal of Experimental Criminology*, 10(4), 459–486.
- Powers, E., & Witmer, H. (1951). An experiment in the prevention of delinquency. In *The Cambridge-Somerville Youth Study*. New York, NY: Columbia University Press.
- Redlich, A. D., Quas, J. A., & Ghetti, S. (2008). Perceptions of children during a police interrogation: Guilt, confessions, and interview fairness. *Psychology, Crime & Law*, 14 (3), 201–223.
- Rosenthal, R. (1965). The volunteer subject. *Human Relations*, 18(4), 389–406.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). In W. R. Shadish, T. D. Cook, & D. T. Campbell (Eds.), *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.
- Sherman, L. W., Gartin, P. R., & Buerger, M. E. (1989). Hot spots of predatory crime: Routine activities and the criminology of place. *Criminology*, 27(1), 27–56.
- Taxman, F. S. (1998). *Reducing recidivism through a seamless system of care: Components of effective treatment, supervision, and transition services in the community*. Washington, DC: Bureau of Governmental Research.
- Weisburd, D. (2000). Randomized experiments in criminal justice policy: Prospects and problems. *Crime & Delinquency*, 46(2), 181–193.
- Weisburd, D. (2005). Hot spots policing experiments and criminal justice research: Lessons from the field. *The Annals of the American Academy of political and social science*, 599(1), 220–245.
- Weisburd, D., & Green, L. (1995). Policing drug hot spots: The Jersey City drug market analysis experiment. *Justice Quarterly*, 12(4), 711–735. <https://doi.org/10.1080/07418829500096261>.
- Weisburd, D., & Gill, C. (2014). Block randomized trials at places: Rethinking the limitations of small N experiments. *Journal of Quantitative Criminology*, 30(1), 97–112.
- Weisburd, D., Lum, C. M., & Petrosino, A. (2001). Does research design affect study outcomes in criminal justice? *The Annals of the American Academy of Political and Social Science*, 578(1), 50–70.
- Weisburd, D., Petrosino, A., & Fronius, T. (2013). Randomized experiments in criminology and criminal justice. In D. Weisburd & G. Bruinsma (Eds.), *Encyclopedia of criminology and criminal justice*. New York, NY: Springer Verlag.
- Weisburd, D., Wilson, D. B., & Mazerolle, L. (2020). Analyzing block randomized studies: The example of the Jersey City drug market analysis experiment. *Journal of Experimental Criminology*, 16(2), 265–287.
- Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54(8), 594–604.
- Wolraich, M. L., Wilson, D. B., & White, J. W. (1995). The effect of sugar on behavior or cognition in children: A meta-analysis. *JAMA*, 274(20), 1617–1621.

## Chapter ten

---

# Propensity Score Matching

### **What Is Propensity Score Modeling?**

---

What are Key Limitations for Estimating Treatment Effects in Traditional Regression Modeling?

How does Propensity Score Matching (PSM) Overcome these limitations?

What is the Logic Model for PSM?

### **How Do We Carry Out a Propensity Score Model?**

---

How Do We Identify Propensity Scores?

How Do We Choose Comparison Cases for Matching?

What is a Caliper?

How Do We Assess Whether the PSM has Produced an Adequate Model?

### **Limitations of Propensity Score Matching**

---

What is the Approach for Assessing the Sensitivity of PSM to Bias from Excluded Variables?

What are Key Limitations of the Propensity Score Approach?

**S**CHOLARS GENERALLY AGREE THAT randomized experiments provide the best method for isolating a causal effect. Unfortunately, it is not always possible to carry out experiments in applied field settings. This means that researchers in criminology and criminal justice must often find other ways of answering key theoretical and policy questions. We have focused a good deal in our text on the ways in which we can use generalized linear models to isolate specific causes at the same time that we control for possible confounding of other variables. But there is a large array of what is sometimes termed *matching* approaches to achieving the goal of identifying a valid causal effect. Most of these methods are outside the framework of our book and are generally seen as falling under the topic of research designs.<sup>1</sup> However, we focus in this chapter on a specific type of matching approach that is built on the generalized linear model approach—**propensity score matching (PSM)**. The approach was first developed by Rosenbaum and Rubin (1983) and has become widely used in criminology and criminal justice.

This chapter follows the chapter on experimental studies because PSM uses a conceptual framework that is similar to that used in experimental research. It seeks to create a comparison group that is equivalent to those in the treatment group. Remember that experiments rely on randomization to allow the researcher to assume that there are no systematic differences between the groups (any differences that exist are random). PSM, like the multiple regression approaches we examined in earlier chapters, seeks to create equivalence using statistical methods in the absence of randomization, based on observed or measured characteristics of a sample of individuals, some of whom participated in the treatment and some of whom did not. The reasons for participation or nonparticipation in treatment will vary depending on the research context. The goal, as with more common multiple regression methods, is to statistically adjust for selection bias, that is, any difference between the groups that is also related to the

---

<sup>1</sup>For discussions of these approaches, see Shadish et al. (2002).

outcome. A primary advantage of PSM over a simpler multiple regression-based approach is the ability to balance the groups on a large number of covariates without losing a large number of degrees of freedom, reducing the statistical power of the test for a treatment effect.

There are four main steps involved in propensity score matching analyses: (1) selecting covariates for predicting likelihood (propensity) of participating in treatment, (2) selecting a method for matching treatment participants to comparison observations based on the predicted propensities, (3) assessing the quality of the matches, and (4) using the treatment and matched comparison observations to estimate the treatment effect. This process, and the underlying logic of PSM, is described below.

## The Underlying Logic Behind Propensity Score Matching

---

A good way to understand PSM and how it provides distinct advantages for isolating treatment effects using observational data is to return to the traditional multiple regression solution. Using two independent variables, our equation for isolating the impact of treatment ( $x_i$ ) is expressed as follows:

$$y_i = b_0 + b_1x_{1i} + b_2x_{2i} + e_i$$

Using this approach, we can control for the possible confounding of a second variable,  $x_2$ . And this model can be extended to a large number of confounding variables, which makes it possible to control in a single equation for numerous potential threats to the validity of our estimate of treatment. As we detailed in Chap. 2, the estimate of the treatment effect ( $b_1$ ) we gain here is purged of the bias of the correlation between the two independent variables.

The problem is that this does not mean we now have an unbiased estimate of the treatment impact. There may be other variables that we have not measured that are also confounded with treatment. For the regression coefficient to provide an unbiased estimate of the treatment impact, all such confounding variables must be accounted for.<sup>2</sup> In theory, every possible cause of the outcome we examine that is also correlated with the treatment must be taken into account. Some statistics texts describe this assumption simply by noting that the covariance of the treatment variable and the error term in the population should be equal to 0. We discussed this

---

<sup>2</sup>Of course, there are other issues that might impact the validity of the treatment effect, such as measurement error.

issue in Chap. 2 in relationship to the assumption of the normality of the distribution of errors in a regression. Because any variable that is not accounted for in the equation is found in the error term, by saying we want the error term and the treatment variable not to be systematically related, we are also assuming that all variables correlated with the treatment variable are taken into account in the model. It is difficult to meet this assumption. This is perhaps the key objection to using multiple regression to identify causal effects of treatment. It is very difficult to argue that all meaningful confounding variables are taken into account.

This means that we must be very cautious when we estimate a treatment impact using traditional regression methods absent randomization. PSM is an attempt to improve on our ability to estimate that treatment effect using observational data. With multivariate models, we cannot keep adding independent variables to a model to account for bias because with each covariate that we incorporate, there is a statistical cost. With each variable you adjust for, your model becomes less precise—confidence intervals around your estimate get wider. In contrast, PSM summarizes all this information in only one estimate (the **propensity score**). While traditional multiple regression models are generally built to estimate the underlying models that create outcomes, including a wide array of potential causes, PSM, like randomized experiments, focuses on the specific effect of treatment. This focus allows us to improve on our ability to specify valid treatment impacts.

The distinct feature of PSM is to reframe the statistical approach as two models, rather than a single model that estimates the treatment effect controlling for a set of baseline covariates. In PSM, the first model predicts membership in the treatment group using the available baseline covariates, generally using logistic regression. From this model, we can generate probabilities, or propensities within the vernacular of PSM, for each observation associated with the likelihood of that observation receiving treatment. We then use these propensities to match treatment and comparison cases or stratify (group) the cases by propensity. The second model then tests for the treatment effect. In the case of matching, this can be done with a simple *t*-test or chi-square, depending on the nature of the dependent variable. In the case of stratification, described below, the treatment effect is tested using a blocked ANOVA type model or logistic regression with dummy codes for the strata and treatment condition. Cases with the same propensity score are presumed to be similar across all of the independent variables included in the first model. Notice that in the standard regression model, we do not *match* cases to create valid comparison groups, rather we statistically control for possible confounding in the estimate of the effect of treatment on an outcome.

## Selection of Model for Predicting Propensity for Treatment

---

We have already noted that in a traditional regression approach, we would have to assume that all variables that have a meaningful joint relationship with treatment and the outcome of interest are taken into account in the equation. If not, the estimate of treatment is likely biased. Using the PSM approach, our assumption is not that all such variables are taken into account, but that our estimate of the propensity for treatment is a valid one. That is, that we have a reasonable model of the selection mechanism.

This is a very difficult assumption to test and one that has led to statistical approaches that attempt to assess the impact of bias (as we describe later in the chapter). But it is not as stringent an assumption as we have in the traditional regression approach. This is the case, in part, because our model of propensity can benefit from the correlations of observed measures with those that we do not observe. For example, let us assume that having a college degree and socioeconomic status are both relevant causes of treatment and that they are both strongly correlated with each other, but that we were only able to gain a measure of socioeconomic status. Because having a college degree and socioeconomic status are highly correlated, even if we do not have data on whether someone has a college degree, the measure of socioeconomic status will reflect some of the causal influence of having a college degree.

Similarly, we can use measures that have strong correlations with the propensity for treatment, even if they combine different aspects or elements of the mechanisms that are generating propensity. This is because we are not interested in the impact of any specific variable per se but rather in the development of a strong and valid prediction of the propensity for treatment. It is for this reason that PSM generally includes all available baseline variables, assuming a sufficient sample size for such a model, and no serious multicollinearity issues exist. We are not interested in whether any of these variables are individually significant.

This does not mean that propensity score models are free from the wider criticism that they have not taken into account important variables in the development of PSM models. There are now a number of articles that advise caution in drawing conclusions from PSM models because of excluded variable bias.<sup>3</sup> However, it is important to note that because PSM changes the assumptions that are key to developing unbiased estimates of treatment (focusing on the prediction of propensity rather than controlling for confounding causes), it provides advantages over the traditional regression approach.

---

<sup>3</sup>For discussions of the limitations of the PSM approach, see Harrell and Slaughter (2020), and Loughran et al. (2015).

A second way that PSM improves our ability to draw causal conclusions about treatment has to do with the cases that it includes in a final analysis of treatment effects. The first step in PSM modeling is to develop a propensity scores for each individual in the sample—both treatment and comparison observations. This is typically accomplished using logistic regression with the participation in treatment as the dependent variable. Remember from Chap. 4 that this probability is generated by the following equation:

$$P(Y = 1) = \frac{1}{1 + e^{-Xb}}$$

where  $Xb$  is the result of the logistic regression model (covariates,  $X$ , and coefficients,  $b$ ). In practice, the propensities are generally produced by statistical software packages developed specifically for PSM (see Exercises).

## Matching Methods

---

Once propensities are assigned to every individual, both those who have received treatment and those who have not received treatment, the researcher decides on a matching algorithm. This algorithm matches one or more individuals from the comparison group to each individual in the treatment group, although it is possible to have some unmatched cases. The matching is based on the propensity score. The score has incorporated the multidimensional set of covariates into a single-dimensional score. Cases with the same propensity score should be balanced or have similar values across those covariates. There are several different methods of matching, all of which involve one or more comparison case matched with a treatment case or group of treatment cases.

There are many more matching algorithms than we can discuss in this chapter (e.g. Apel and Sweeten 2010; Austin 2014; Caliendo and Kopeinig 2008). The main approaches are nearest neighbor, caliper matching, and stratification matching. The first of these, **nearest neighbor matching**, matches each treatment case with a comparison case that has the closest propensity score. This can be done with and without replacement. With replacement ensures that each treatment case is matched with the closest comparison case but may have lower statistical power given that a single comparison case may be matched to more than one treatment case, reducing the sample size of the comparison condition. When possible, we recommend the without replacement approach, though when there are not enough potential comparisons, it makes sense to allow replacement or to use an alternative approach, such as stratification matching. A complication of the not replacing comparison cases that match a treatment cases is

that the order of the treatment cases matters because once a comparison case is matched with a treatment case, it is no longer available for matching with another treatment case, even if that would reflect a better match. Typical approaches involve sorting the treatment cases from low to high, high to low, or randomly, and then matching based on this ordering. A variant of nearest neighbor matching is to match multiple comparison cases to each treatment cases (i.e., one-to-many matches), assuming there is a large pool of such cases to draw from.

A weakness of the simple nearest neighbor approaches is that they may result in bad matches, that is, treatment and comparison cases with dissimilar propensities. Using **caliper matching** addresses this by putting a maximum propensity distance (**caliper**) as a criterion for matches. The matching follows any of the nearest neighbor variants (e.g., with and without replacement, one-to-one or one-to-many, and different sort orders). A clear advantage of the caliper matching approach is that it ensures that all matches have sufficiently close propensities. A challenge with this method is selecting an optimal caliper width. The caliper is sometimes expressed in units of the standard deviation of the logits of the propensities. Sometimes, it is expressed in the propensity score metric (e.g., see Stata), reflecting the difference in the probabilities between treated and untreated units. Computer simulations by Wang et al. (2013) suggest that a caliper width of 0.2 of the standard deviation of the logits of the propensities is often ideal, but that values between 0.1 and 0.8 may be suitable depending on the distribution of propensities across the treatment and comparison groups. Using the propensity score metric calipers between .01 and .05 are seen as rigorous criteria. It is important to note that with this method, some treatment cases might not be matched with any comparison cases because there are no comparison cases within that region of the propensity scores. This reduces the sample size for analysis but ensures that only treatment and comparison cases that are matched are used to test the treatment effect. As discussed below, this is called the **region of common support**.

In the traditional regression approach, we include the entire sample of cases in our assessment of treatment. PSM recognizes that treatment is generally not delivered randomly across the population of cases. In generating a sample for evaluating a treatment, researchers often include all cases within the comparison group, however constructed. For example, in a correctional program, the researcher might draw all inmates that are in the same prison that were eligible for treatment but did not participate in the treatment program. The problem with this approach is that it is likely that certain types of prisoners who were very appropriate candidates for the program almost always were chosen. In turn, there are also likely to be prisoners who were not chosen because they were not very good candidates for the program. At these extremes, there are likely to be few valid

matches for treatment cases. In regression, this is sometimes termed an empty cell problem. The regression is trying to estimate treatment outcomes under all conditions. But it does not have outcomes for both groups under these conditions.

In PSM, the caliper allows us to avoid this problem because it does not include any treatment case for which there is not an appropriate comparison case, and vice versa. When there is no comparison case within the caliper defined, that treatment case is excluded from the final analysis. If we examine the impact of treatment on recidivism, accordingly, we are only comparing treatment and comparison cases that were similar in their propensities to receive treatment. Of course, when we gain an advantage like this in statistics, we generally also pay a price. The price in this case is that the sample used to evaluate treatment impacts may not reflect the broader population of cases. To the extent that matches within the caliper cannot be gained, the researcher's ability to generalize to the population of treatment cases declines. This becomes a more and more serious problem as a larger and larger proportion of treatment cases are dropped in the PSM analysis. When this pruning is substantial, it can introduce bias into the treatment effect estimate (King and Nielsen 2016). This is one reason why researchers will sometimes increase the caliper used in the study.

The **stratification matching** method involves first trimming the cases to only include values of the propensity within the region of common support (the range of propensity scores that includes both treatment and comparison cases). The remaining set of cases is stratified by the propensity score. The general advice is to create at least 5 strata and no more than 10 (Neuhäuser et al. 2018). For example, if there are 250 cases total within the region of common support, these would be divided into stratum of 50 each based on the propensity score. The analysis would then treat the strata as blocks in a blocked ANOVA type of analysis. Notice that this method uses all treatment and comparison cases within the region of common support and does not require equal sample sizes per condition.

Once the PSM model and matching has been achieved, it can be used to examine each of the outcome measures of interest within a study. And in such analyses, the appropriate statistical tests will be based on the nature of the distribution of the outcome measures used, but may be as simple as a *t*-test or chi-square test of independence. Weighting may be required if matching with replacement or one-to-many matching was used. PSM provides a strong quantitative method for developing matched comparison groups to assess treatment outcomes. It is important to note that the PSM approach can also be used for matching outside of the program evaluation context, though the first step in any PSM is to define the propensity that is of interest, that is, the two groups that you as the researcher wish to compare.

### The Case of Work Release in Prison: A Substantive Example

To provide a concrete example of how propensity score matching can be used in criminology and criminal justice, we examine an evaluation of a work release program in the Israeli Prison Service (Weisburd et al. 2017). The program integrated work releases with social support elements in a specific living area for the treated inmates. Their main research question was whether this integration of multiple program elements led to benefits in terms of reduced re-incarceration and rearrests.

The study examined 712 prisoners who had participated in the work release program between 2004 and 2011 (allowing a maximum follow-up period of 5 years). To develop a valid comparison condition, the researchers sought to identify similar prisoners from the 65,893 sentenced prisoners under IPS supervision during this period. Information regarding these prisoners was obtained through the Israeli Prison Service (IPS) data system and incorporates various socio-demographic characteristics, as well as criminal history information and information on incarcerations. The system also collects data on prisoners' conduct throughout their incarceration, including visitations, furloughs, disciplinary hearings, educational programs, and vocational programs. A large array of potential matching variables prior to entering the program, as is the case here, is key to establishing strong PSM models.

The calculation of the propensity scores of the prisoners was conducted in two stages. The first stage included filtering out from the general pool of prisoners all prisoners that did not meet the basic requirements of the program as stated by the IPS during the study period and for which quantitative data were unavailable: being male; serving at least a quarter of their sentence; having two furloughs of at least 24 h; having at least 6 months left until release; under the age of 67; not being involved in criminal activity (that the prison authority or police are aware of), nor evidencing a disciplinary offense. All prisoners that did not meet these standards were deleted from the pool of potential comparison subjects. They then excluded all prisoners that were released after 2011, leaving a database of 2615 prisoners who did not take part in the work release program but met the basic requirements of the program.

In the second stage of the calculation, they identified variables that would be used to build the propensity score models. As noted earlier, the measures included in the PSM are critical to the validity of the matching process. The database available included a relatively large number of possible matching variables. The researchers included those that had been used in prior models in correctional treatment evaluations and that had been found to strongly predict recidivism outcomes and that seemed important in the particular Israeli context (see Table 10.1). The PSM model should include as many relevant baseline variables as possible.

**Table 10.1**

Treatment and comparison groups compared before and after propensity score matching

VARIABLE	BEFORE MATCHING (N = 3327)			AFTER MATCHING (N = 1094)		
	TREATMENT	COMPARISON	BIAS (%)	TREATMENT	COMPARISON	BIAS (%)
			(%)			(%)
<i>Socio-demographic variables</i>						
Age at program entry	35.76	35.5	2.6	35.88	35.10	8.06
New immigrant	16.9%	17.4%	0.19	16.3%	17.6%	3.47
Married	47.2%	36.8%	2.93**	45.5%	43.3%	4.43
Number of children	1.56	1.43	6.64	1.58	1.50	4.12
Nationality (Jewish)	58.3%	56.9%	0.34	58.1%	55.4%	5.45
Number of years schooling	10.56	9.83	27.18**	10.36	10.31	1.84
<i>Criminal background</i>						
No. of incarcerations	2.17	3.07	40.07***	2.33	2.31	0.99
Age at first incarceration	28.28	26.01	25.31***	28.15	27.54	6.55
<i>Current offense</i>						
Incarceration length	49.97	39.51	31.13***	46.36	44.47	5.92
Violent offense	48.9%	43.8%	10.24*	46.3%	47%	1.40
Drug offense	32.4%	37.4%	10.51	32%	34%	4.25
Property offense	30.5%	40.0%	20.00***	34.7%	33.3%	2.96
Sexual offense	4.2%	5.9%	7.75	5.1%	3.8%	6.3
Year of release	2007	2008	63.9***	2007	2007	0
<i>Prisoner profile</i>						
Criminal violence	7.4%	4.0%	14.69***	6.8%	7.3%	1.96
Domestic violence	1.1%	.2%	11.0**	.9%	.7%	2.22
Chronically ill	4.2%	5.3%	5.18	4%	3.8%	1.03
<i>Prison experience</i>						
Standardized no. of furloughs	2.52	2.33	12.6**	2.54	2.54	0
Standardized no. of hearings w/o isolation	.27	.42	10.00**	.21	.28	13.88*
Standardized no. of hearings with isolation	.07	.30	42.5***	.08	.12	11.16
Potential/time from last hearing to program entry	156.49	318.69	44***	179.33	191.8	3.67
Standardized no. of formal courses	.40	.44	5.45	.43	.45	2.59
Standardized no. of informal courses	1.78	1.09	37***	1.47	1.53	3.3
Standardized no. of contract Jobs	.04	.03	7.73*	.04	.03	8.08
Standardized no. of factory Jobs	.19	.16	12.2**	.19	.20	3.89
Standardized no. of privileges revoked	.33	.92	59.4***	.40	.42	3.02
Potential/time from last revocation to rehabilitation	156.49	232.92	22.4***	180.90	192.6	3.61
Standardized no. of professional training	.04	.06	10.4*	.05	.05	0
Standardized no. of nonprofessional training	.04	.04	0	.04	.04	0

Age for comparison group is the age of the prisoner in the last date he was still eligible to enter the program by the IPS criteria (minimum of 6 month left to release)

Note. \* $p < 0.05$ , \*\* $p < 0.01$ , and \*\*\* $p < 0.001$

They then used a nearest neighbor approach with caliper matching with a 0.01 caliper (based on the propensity score metric) and with no replacement. The treatment cases were randomly ordered. The first case within this random ordering was matched to the comparison case with the closest propensity score that was within the caliper. Once matched, the comparison cases were removed from consideration for matching and the next

treatment case was matched to a remaining comparison case and so on. This produced 547 matched treatment and comparison pairs. There were 165 treatment cases that remained unmatched because no comparison case was within the caliper region. Stated differently, these cases had no suitable match.

## Assessing the Quality of the Matches

---

The propensity scores were based on a logistic regression in which all of the variables in Table 10.1 were included as independent variables and participation in the work release program (1=yes, 0=no) was the dependent variable. The program they used (SPSS) then matched the cases using a .01 caliper as described above. An important first step in this process is to assess to what extent the PSM approach has created treatment and comparison groups that are well-matched. One way to do this is to compare the propensities of the treatment and comparison conditions before and after PSM has been carried out. Before matching, the treated group (work release program) ( $n = 712$ ) had a mean propensity score of 0.401 ( $SD = 0.232$ ,  $\min = 0.002$ ,  $\max = 0.988$ ). The nontreated or comparison group prisoners (those who did not take part in the work release program) ( $n = 2615$ ) had a mean propensity score of 0.163 ( $SD = 0.152$ ,  $\min = 0.000$ ,  $\max = 0.963$ ). After matching, the comparison group ( $n = 547$ ) had a mean propensity score of 0.318 ( $SD = 0.186$ ,  $\min = 0.002$ ,  $\max = 0.906$ ). The treatment group ( $n = 547$ ) had the same propensity score of 0.318 ( $SD = 0.001$ ,  $\min = 0.046$ ,  $\max = 0.905$ ) after matching. The fact that the means of propensity scores of the treatment and comparison groups are equal illustrates the degree to which the PSM approach is successful at defining groups that are comparable, at least in terms of the data that are known to the researchers.

A comparison of the pre- and post-matched groups on variables included in the PSM (and tests of statistical significance) used for matching is displayed in Table 10.1. As is apparent from this table, the PSM procedure produced groups that were roughly balanced on known included characteristics. Before matching, the groups differed substantially on most of the traits examined. After matching, the differences that remained were small. Only one characteristic has a statistically significant difference (standardized number of hearings without isolation); however, this difference is small. To observe only one significant result in such large number of tests is not surprising. Indeed, just by chance with a .05 significance criterion, this number of comparisons would be expected to produce at least one

significant difference. However, if after matching, there are a number of variables that yield statistically significant differences, you should critically examine the approach used.

Furthermore, it is critical to examine the size of the bias, even if not statistically significant to ensure that any bias is small relative to the expected treatment effects. Rosenbaum and Rubin have created a statistic, **standard absolute bias**, to identify when a scaled variable in the PSM model is considered to be unbalanced (Rosenbaum and Rubin 1985). It is calculated using the following equation:

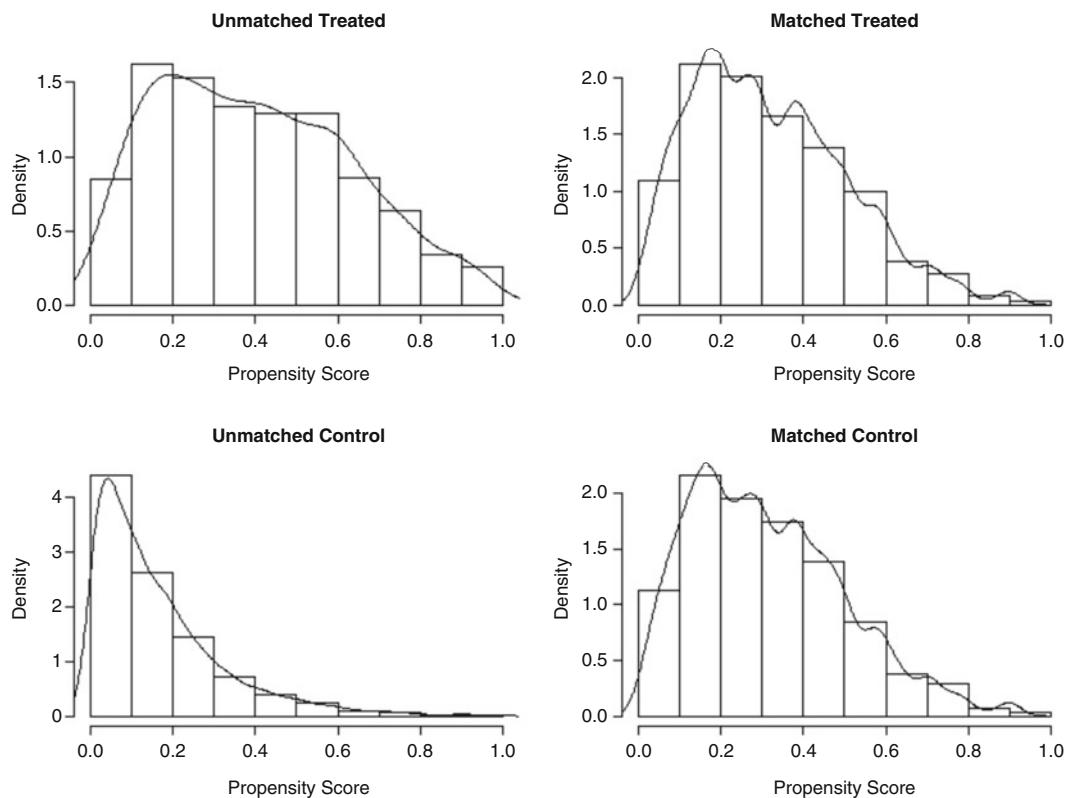
$$\text{Bias} = \frac{100(\bar{x}_t - \bar{x}_c)}{\sqrt{\left(\frac{s_t^2 + s_c^2}{2}\right)}} \quad \text{Equation 10.1}$$

where  $\bar{x}$  is the group mean,  $s$  is the group standard deviation, and the subscripts  $t$  and  $c$  reference the treatment and comparison groups, respectively. In the case of a binary variable, bias is simply the difference in the two percentages, as can be seen in Table 10.1. When standard absolute bias is greater than 20, Rosenbaum and Rubin argue that the variable is considered to be unbalanced. In this case, the researcher will have to construct their analyses to control for the unbalanced factor. This can be done simply by predicting the outcome of interest with this measure as a control variable. As is apparent from Table 10.1, none of the variables used in the PSM had values above 20.

If the samples are well-matched, the distributions of the treatment and comparison samples after matching should also be similar. This can be examined by graphing the distributions of the samples before and after matching. This is done for these data in Fig. 10.1. As can be seen from the figure, before matching the distribution of the propensity scores of treated and untreated sample, subjects are very different. But after matching, the two distributions are very similar. So accordingly, it is not just that the means are very similar after matching, but the distribution of cases in both groups across propensities is very similar. As noted earlier, the area in the distribution where there are both treatment and comparison cases that would be eligible for matching is called the region of common support. This gives you important information about your ability to develop adequate matches between treatment and comparison, and the extent to which comparison cases provide matches across the distribution. Looking at the cases before matching, it is clear that there is a wide spread of possible comparison cases, though as expected the potential for matching declines at the two extremes of the distribution. If there is a small region of common support, PSM may not be an appropriate tool.

**Figure 10.1**

*Comparison of distributions of propensities for treatment and comparison before and after matching and the region of common support*

**Table 10.2**

Cumulative re-incarceration rates for work release and comparison group prisoners

RECIDIVISM	PROGRAM		COMPARISON		CHI-SQUARE	
	WORK RELEASE					
	%	N	%	N		
One year	6.2	547	10.8	547	7.34**	
Two years	11.7	547	18.5	547	9.77**	
Three years	15.3	547	24.6	544	14.09***	
Four years	18.8	547	29.3	544	16.48***	
Five years	22.5	523	33.1	506	14.22***	

Note: \*\*  $p < .01$ , \*\*\*  $p < .001$

Table 10.2 shows the differences between treatment and comparison subjects in terms of re-incarceration rates over 5 years. The researchers used a chi-square test to assess whether the differences in recidivism were significant for each period. As is apparent from Table 10.2, in each year, there was a statistically significant difference between treatment and comparison cases. Overall, the proportion of re-incarcerations was higher in the comparison group.

The researchers noted that prisoners who took part in the program were 42.6% less likely to be re-incarcerated in the first year after their release than those who did not participate in the program:

$$\begin{aligned}\text{Proportional Decline} &= \frac{(\text{Comparison \%} - \text{Treatment \%})}{\text{Comparison \%}} \\ &= \frac{10.8 - 6.2}{10.8} \\ &= 0.4259\end{aligned}$$

Following this, the cumulative risk of re-incarceration for those who participated in the program was lower than the comparison group for each year of the remaining 4 years: After 2 years, their cumulative risk of re-incarceration was 36.7% lower (18.5 vs. 11.7%); after 3 years, 37.8% lower (24.6 vs. 15.3%); after 4 years, 35.8% lower (29.3 vs. 18.8%); and after 5 years, their cumulative risk of re-incarceration was 32% lower than those of the comparison group (33.1 vs. 22.5%).

The results were similar for rearrests (see Table 10.3). Chi-square tests found differences between the treatment and comparison conditions to be significant throughout the entire follow-up period. Prisoners who took part in the program were 40.9% (23.0 vs. 13.6%) less likely to be rearrested in the first year after their release than those who did not participate in the program. A meaningful gap remained for the three additional follow-up years: After 2 years, program participants were 41.4% (36.5 vs. 21.4%) less likely to be rearrested than nonparticipants; after 3 years, they were 35% (40.8 vs. 25.5%) less likely; and after 4 years, they were 30% (46.3 vs. 32.4%) less likely.

Overall, the researchers conclude:

Our study assessed whether the Israeli work release program reduced recidivism. We found that it did, and that effects observed are much stronger than those that have been observed in evaluations of work release in the U.S. We argue that the strength of the program's impacts was reinforced by its broader and more integrative approach in Israel. Work release in Israel is not simply work release but involves a positive social environment, a high dosage of counseling and therapy, and more general privileges for inmates including furloughs and cultural activities. This suggests more generally the importance of a broader more integrative

**Table 10.3**

Cumulative rearrest rates for work release and comparison group prisoners

RECIDIVISM	PROGRAM		COMPARISON		CHI-SQUARE	
	WORK RELEASE		COMPARISON			
	%	N	%	N		
One year	13.6	547	23.0	547	7.94**	
Two years	21.4	547	36.5	547	14.89***	
Three years	26.5	547	40.8	544	24.97***	
Four years	32.4	547	46.3	544	22.08***	

Note: \*\* $p < .01$ , \*\*\* $p < 0.001$ 

approach to work release. Stronger effects can be achieved by adding program elements that reinforce the logic model behind the program's impacts. (p. 258)

For our purposes, this example illustrates one of the distinct advantages of the PSM approach over traditional regression analyses. Though multiple regression is the key driver of PSM, the actual findings in terms of the outcomes can be presented in a simple table format. One can use regression analyses for PSM, for example, if single variables are found to be unbalanced. But in general, you are able to present relatively simple tables to describe your results, as was the case here. This is because the more complex statistical approaches are taken in the stage of identifying matched groups. After that, as we noted earlier, PSM analyses become similar to analyses presented in randomized experiments.

## Sensitivity Analysis for Average Treatment Effects

As we noted earlier, PSM, like other observational data analysis methods, is constrained by the degree to which variables that differentiate selection factors can be identified. Only in experiments can a strong assumption of *ignorability* of confounding be made. At the same time, given a large number of identified and relevant covariates, an assumption of equivalence between the groups has been shown to be reasonable (Shadish 2013). Nonetheless, there has been growing concern regarding overstatement of the validity of PSM models, especially in the case of weak covariate data sources (See, for example, Loughran et al. 2015).

Rosenbaum provides an approach that allows us to assess how sensitive study results are to selection bias due to the exclusion of key (unmeasured) variables in the development of a propensity model (Rosenbaum 2002). In a case where the outcome is binary, the **MHbounds (Mantel and Haenszel bounds) test** provides an overall estimate of the sensitivity of the model to bias (Aakvik 2001; Mantel and Haenszel 1959). This test only

refers to the sensitivity of the study to possible unobserved variables. It makes no statement about the actual degree of unobserved variable bias in the models estimates. This is an important caveat since we have no way of knowing to what extent there is bias in a PSM. We can argue that our models include a large number of covariates and that those covariates reflect theoretical understandings of the propensity we are examining. However, absent the true propensity model, we have no method for gaining an estimate of the actual bias in our model.

MHbounds provides a method of seeing how sensitive our model is to bias, if there is bias. When running the MHbounds, two different scenarios can be estimated,  $Q_{MH^+}$  and  $Q_{MH^-}$ . The former refers to a situation in which we have estimated a positive treatment effect and the latter a situation in which we have estimated a negative treatment effect.<sup>4</sup>

The question that MHbounds answers for us is how much bias would have to be present in our model for the results we observe to become insignificant. The extent of bias is estimated using **gamma ( $\gamma$ )** which represents the odds of differential assignment due to unobserved factors. A gamma of 1 indicates no bias. This is the scenario in which the PSM is correctly specified. With a gamma of 1.5, the bias caused by omitted variables in our PSM is assumed to increase the odds of entering treatment by 50%. One way to think about this is that with a gamma of 1.5 we are considering what our results would look like if our treated subjects had a much lower likelihood of recidivating, even if they did not participate in the program.

Table 10.4 shows the degree of sensitivity to the results under gamma levels ranging from 1.3 to 1.6. The table reports the gamma level, as well as the significance level of the estimates under the assumption of a gamma level of bias. The extent to which the results reported are robust to omitted variable biases varies across the comparisons that are made.

Using a .10 significance threshold, the results generally stay significant with gamma values of up to 1.5 for re-incarceration. This suggests that the problem of bias will substantively alter the interpretation of the treatment impact only if there are unobserved variables that bias selection (in terms of lower recidivism) into the treatment group by 50%. If a more stringent significance threshold is used, the tests are much more sensitive to bias. When referring to rearrests, the results are significant at slightly higher values of gamma.

Of course, these results only tell us what would happen if there was bias. They do not tell us how much bias there actually is. And this is a very critical distinction. A study might not yield significant results under assumptions of gamma of 1.1 or 1.2. And this might lead the researcher to conclude that the

---

<sup>4</sup>For details on the computation of MHbounds, see Becker and Caliendo (2007).

**Table 10.4**

Mantel and Haenszel bounds for the treatment effect for different follow-ups

	GAMMA ( $\Omega_{MH^-}$ )	<i>p</i> critical ( $p_{MH^-}$ )
Re-incarcerations		
1 year (incarcerations)	1.3	.08
	1.4	.14
2 years (incarcerations)	1.3	.05
	1.4	.12
3 years (incarcerations)	1.4	.06
	1.5	.14
4 years (incarcerations)	1.4	.05
	1.5	.12
5 years (incarcerations)	1.3	.03
	1.4	.1
Rearrests		
1 year (rearrests)	1.3	.07
	1.4	.15
2 years (rearrests)	1.5	.07
	1.6	.16
3 years (rearrests)	1.5	.04
	1.6	.1
4 years (rearrests)	1.5	.07
	1.6	.15

results are highly sensitive to excluded variable bias. But, in this case, the models may be particularly well specified which means that the results are ones that can be seen as valid. In contrast, a study with very strong outcomes may have a great deal of omitted variable bias. But the study is likely to yield strong results even if the actual bias is large. Our point is simply that this approach is useful, but must be interpreted cautiously.

In turn, there is no clear definition of what thresholds are required for a model to meet a strong ignorability assumption—an assumption that the model is robust even if there is a good deal of omitted variable bias. The results here are consistent with other published studies in criminology and criminal justice, though smaller than studies in fields such as medicine. Whatever the comparison, these results suggest that these models, like those of other observational studies, should be interpreted with some caution. In the absence of experimental data, we need to recognize that the exclusion of key factors from the PSM might increase or decrease odds of selection into the treatment and comparison groups.

## Limitations of Propensity Score Matching

Propensity score matching (PSM) has become a common approach to the construction of a quasi-experimental comparison groups for testing the effectiveness of a criminal justice programs, policies, or practices. When

suitable to the particular research problem and available data, it is a powerful method and has several advantages over multiple regression. However, it is not without its limitations.

The loss of treatment and/or comparison cases due to a failure to identify suitable matches is a key potential limitation of PSM and can result in both a reduction in statistical power and a more limited generalization of any observed treatment effect. You cannot generalize to the types of cases that were dropped from the analysis. However, the ability to assess the overlap between the treatment and comparison cases in terms of their distribution of covariates through an examination of the region of common support allows you to assess this limitation. This is also an advantage in that it brings to light any issue you have with nonoverlap in the linear combination of some covariates between the treatment and comparison group that would be obscured in a standard multiple regression approach.

The propensity score does not always ensure balance between the treatment and comparison group on critical baseline variables. Some of the literature on PSM recommends including selected baseline variables (and possibly squared and cubed versions of one or more of these variables) into the model that tests for the treatment effect, even though these are also part of the propensity score. Variables to include would be key predictors of the outcome or prognostic indicators. You can also include the propensity score (actually, the logit of the propensity score) in the model as well to help improve the statistical adjustment. If the sample size is small, these additional variables may reduce your statistical power.

Matching in PSM is sensitive to the order of the treatment cases when nonreplacement of comparison cases are used. For example, as noted with the nearest neighbor method, the first treatment case is matched with its nearest neighbor within the region of the caliper size. Once the comparison case is matched to this treatment case, it is no longer available to be matched with another case. However, it might be a better match for a different treatment case and had that treatment case been selected earlier in the matching processes, and the final matches may have been different. There are some matching algorithms available that try to find the optimal set of matches by iterating the matching process with different random orders of the treatment cases. The order that minimizes the average absolute within-pairs difference is considered optimal. Such an algorithm is computationally intensive. The broader point is to recognize that PSM does not produce a unique set of matches and the decisions regarding matching are arbitrary and may in some situations affect the results.

Finally, as noted earlier, PSM is only as good as the quality of the baseline measures. If important baseline variables are omitted (and they are not accounted for by their correlation with included covariates), the results will be biased. The goal is to have variables that are related to the selection process for who gets treatment and who does not, such as

baseline measures for the outcome or other selection features. Simply adjusting for readily available demographic variables will rarely produce an unbiased estimate of the treatment effect (Shadish and Steiner 2010). No statistical method makes up for insufficient data for addressing the research question of interest.

## Chapter Summary

---

This chapter examined a common method for creating matched comparison samples for assessing the impacts of treatments or interventions when trying to adjust for selection bias in absence of randomization of individuals to treatment. The approach, termed **propensity score matching (PSM)**, uses regression models to create estimates for the propensity of individuals treated or untreated to be selected into treatment using *one* score—a **propensity score**. In this sense, it focuses its main interest on the mechanisms that underlie selection, and its success in creating equivalent groups is based on the researcher’s ability to successfully model the selection mechanism. The first step in carrying out PSM is to identify a sample of cases that received and did not receive treatment. The researcher then develops a regression model that gives each subject in the sample, whether they received treatment or not, a propensity score for the likelihood of receiving treatment.

One advantage of PSM in evaluation studies is that it does not require that the researcher identify all variables that are meaningfully correlated with treatment and outcome, but rather that a valid estimate of propensity is gained. A second advantage of PSM is that it only creates matches for cases that have similar propensities, thus avoiding the empty cell problem common in regression analyses that seek to identify treatment outcomes. Once propensities have been assigned, the researcher identifies a matched case for each treatment case, which can be done a number of ways. A common method is **caliper matching**, which uses a specified bandwidth (referred to as a **caliper**) whereby cases are only matches if the propensity of the treatment and comparison case(s) falls within the specified caliper. Another matching technique is **nearest neighbor matching** which randomly orders the treated and comparison subjects; then, an individual from the comparison group is selected as a matched partner for an individual in the treatment group. Matching cases to their nearest neighbor could potentially result in bad matches, so this is avoided by putting a maximum propensity distance (caliper) as a criterion for matches. The **stratification matching** method involved first trimming the cases to only include values of the propensity within the **region of common support** (the range of propensity scores that includes both treatment and comparison cases), and then,

the remaining set of cases are stratified by the propensity score into a given stratum.

Having assigned the comparison cases, the analyses for PSM follow those of traditional experimental studies. This is another advantage of PSM as the reporting of outcomes is usually straight forward and easily communicated. As in other regression approaches, the availability of a wide array of relevant covariates is essential for producing a valid set of outcomes. The **standard absolute bias** is used to identify when a scaled variable in the PSM model is considered to be unbalanced (bias score greater than 20 is considered to be unbalanced). The **MHbounds test** has been developed to assess how sensitive the study findings are to bias from exclusion of relevant variables in the selection model. **Gamma ( $\gamma$ )** is the measure of bias that is used. When gamma equals 1, it means that you are assuming that there is no hidden bias, whereas as the gamma increases or decreases, it means that you are assuming that there are unobserved variables that can influence your results.

## Key Terms

---

**Caliper** The bandwidth used for the selection of comparison cases in propensity score modeling. This ensures that matched cases have a propensity scores that differ by no more than the caliper width.

**Caliper Matching** A matching method that uses a specified bandwidth (referred to as a caliper) whereby cases are only matched if the propensity of the treatment and comparison case(s) falls within the specified caliper.

**Gamma ( $\gamma$ )** The standardized measure of bias used to assess how sensitive PSM models are to excluded variables.

**MHbounds (Mantel and Haenszel bounds) test** A test for assessing how sensitive PSM results are to the bias of excluding key (unmeasured) measures in the selection model.

**Nearest neighbor matching** An approach to matching in PSM that matches the

treatment case to the not treated sample case with the smallest probability distance.

**Propensity score matching (PSM)** A commonly used method to identify matched cases when the researcher cannot gain experimental data and wants to assess the impacts of outcomes. It models the mechanism of selection into treatment as a method for matching cases.

**Propensity score** A single score that is the probability of a case receiving treatment given a set of measured covariates. This value is determined during the propensity score matching process.

**Region of common support** Area of the propensity score distribution for which there is overlap between treatment and comparison cases measured before matching. This is usually identified through a visual inspection.

**Standard absolute bias** A measure used to identify when a variable in the PSM model is considered to be unbalanced.

**Stratification matching** Treatment and comparison cases are stratified into 5 to

10 groups after trimming the cases to only include those within the region of common support.

## Symbols and Formulas

---

$\gamma$  Measure of bias to assess PSM

Standard Absolute Bias:

$$\text{Bias} = \frac{100(\bar{x}_t - \bar{x}_c)}{\sqrt{\left(\frac{s_t^2 + s_c^2}{2}\right)}}$$

## Exercises

---

- 10.1. Describe the advantages of propensity score matching approaches over conditioning on covariates to reduce bias when you are estimating the effect of a given treatment from observational data. What advantage does propensity score matching have over multivariate regression techniques?
- 10.2. Describe the methods used to assess the covariate balance of a matched sample.
- 10.3. If you determine that there is significant bias in your matched sample, what steps can be taken to reduce bias?
- 10.4. What is the ignorability assumption when it comes to propensity score matching methods?
- 10.5. What are the steps involved in conducting propensity score matching analysis?
- 10.6. How does one determine which matching technique is to be used when conducting propensity score matching? What criteria are used to gauge the believability of findings rendered?
- 10.7. List the propensity score matching techniques that were reviewed in this chapter. How do the techniques differ from one another?
- 10.8. What is a propensity score?

- 10.9. What are the shortcomings of the techniques used to assess covariate imbalance?
- 10.10. Describe the purpose of assessing sensitivity of propensity score matching methods.

## Computer Exercises

The data file used to illustrate propensity score matching in this chapter is in Stata's data format (*psm.dta*). This can be imported into R using the *read\_dta()* function from the *haven* package. SPSS does not have a built-in option for propensity score matching so the tutorial below will be reviewing propensity score matching in Stata and R.

### Stata

The propensity score matching examples will be relying upon the *psmatch2* module. If it is not installed already, run the following Stata code to install the module (which relies on the **ssc install** function).

```
ssc install psmatch2
```

After you open your dataset in Stata, make sure to address cases with missing data. If you would like to remove cases in your dataset that have missing data, you can rely on the **drop** function. You may specify one or multiple variables. The example below drops cases from the dataset that have missing data on the two specified variables named *var1* and *var2*.

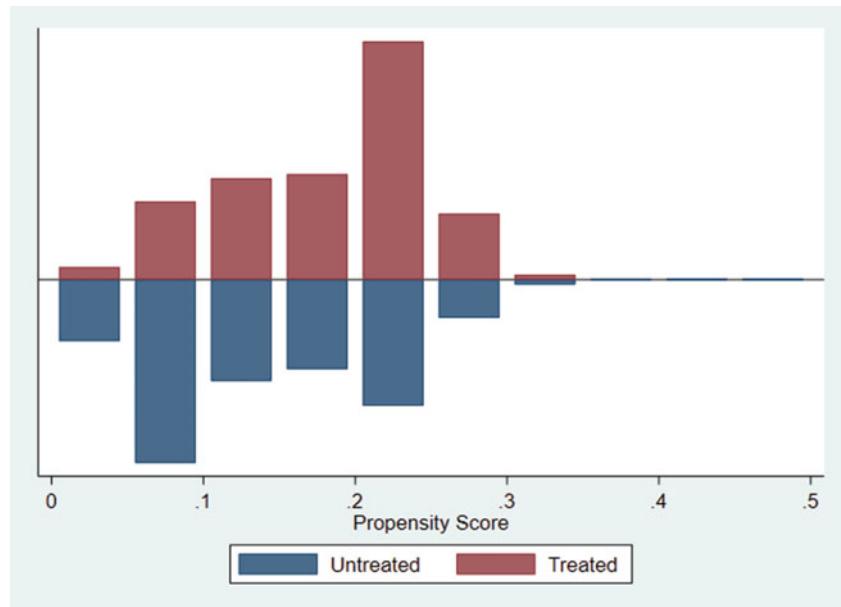
```
drop if missing (var1 var2)
```

### Estimating Propensity Score

As you will see in the next section, the **psmatch2** function from the *psmatch2* package calculates propensity scores during the matching process for us, but you can do it yourself if you would like by using the logit and predict functions. Here, we estimate a propensity score and store it in a variable that we name *ps\_score*. The *treat* variable in the model is the exposure to treatment variable (coded as 0 or 1), and *iv1*, *iv2*, and *iv3* are the independent variables used to predict propensity for treatment. Ideally, there would be a large number of independent variables listed, but working with a small number is sufficient to understand the basic steps and how the code works.

```
logit treat iv1 iv2 iv3
predict ps_score
```

You can examine the region of common support with the estimated propensity score very easily in Stata via the **psgraph** function. And as shown in the next section, you will obtain more information on the region of common support from the **psmatch2** function.

**psgraph****Matching Cases**

You can specify matching criteria with the **psmatch2** function by using a *caliper* via the **caliper** option, which restricts matching based on a specified number of standard deviations of the propensity score. In this example, we use the **psmatch2** function to conduct 1:1 matching with a **caliper** of 0.02. The first example relies on the **ps\_score** variable that we estimated in the prior section and uses the **pscore** function:

```
psmatch2 treat, pscore(ps_score) caliper(.02)
```

The second example estimates the propensity score when the propensity variable has not been calculated yet. If you are using this approach, you do not need to estimate the propensity score as you did in the prior section. Instead, you use the *logit* option to the right of the comma, which specifies that the propensity score be estimated with logistic regression.

```
psmatch2 treat iv1 iv2 iv3, logit caliper(.02)
```

Nearest neighbor matching is done by using the *neighbor* option, whereby you specify the number of nearest neighbors in the parentheses. You can match without replacement by specifying the *noreplace* argument to the right of the comma. A caliper can be used as well, if desired, as we did in the prior example.

*Nearest neighbor 1:1 matching without replacement:*

```
psmatch2 treat iv1 iv2 iv3, neighbor(1) noreplace
```

The argument **common** can be added to drop treated cases whose propensity score is higher than the maximum control case or lower than the minimum control case.

```
psmatch2 treat iv1 iv2 iv3, neighbor(1) noreplace common
```

Alternatively, you can *stratify* cases by propensity level. To do this, you need to install *egenmore* using the **ssc install** function (if you do not have it installed already). In the example below, 10 strata are created by using the option **n(10)**.

```
logit treat iv1 iv2 iv3  
predict ps_score
```

```
ssc install egenmore  
egen group = xtile(ps_score), n(10)
```

### *Assessing Matches*

The balance between the treated and control cases can be assessed with *psmatch2* by specifying the **pstest** immediately after the *psmatch2* function. This function provides the percent bias for the covariates and conducts *t*-tests for matched and unmatched cases between treatment and control on the covariates specified in the model.

### **pstest**

Assessing matching when propensity scores were created using stratification is also simple. Here, we do this by using the **pbalchk** function, as we are relying on our grouping variable that we created above named *group*. If you have not already installed it, you should first do so by using the **findit** function.

```
findit pbalchk  
pbalchk treat iv1 iv2 iv3, strata(group)
```

### *Estimating Treatment Effect*

The *psmatch2* function offers an outcome option that can be added when estimating propensity scores/matching to estimate the treatment effect. The name of the dependent variable in the example below is *outcome*. This option reports the ATT (average treatment effect of the treated). Add the *ate* option after the comma makes it so the ATE (average treatment effect of the population) and ATU (average treatment effect of the untreated) are also reported in the Stata output.

```
psmatch2 treat iv1 iv2 iv3, logit outcome(outcome) ate
```

With the psmatch2 function, only matched cases will receive a value on the `_weight` variable that is generated in the dataset. All cases in the treated group receive a `1` on this variable. For the control cases, the weight assigned is the number of observations from the treated group for which the observation is matched. You can assess the treatment effect by employing the weighted variable (as in the case of 1:many matching) or unweighted (as with 1:1 matching).

```
/* Weighted, 1:many matching */
reg outcome treat [fweight=_weight]

/* Unweighted, 1:1 matching, excluding unmatched cases */
reg outcome treat if !missing(_weight)
```

## R

There are a number of functions and packages to conduct propensity score matching in R. We will be relying on the `matchit()` function in the **MatchIt** package. So, you will need to install and load the *MatchIt* package, as follows:

```
install.packages("MatchIt")
library(MatchIt)
```

You may import your dataset into R using the `read_dta()` function from the *baren* package. Once loaded into R, make sure to address cases with missing data. If you would like to remove cases in your dataset that have missing data, you can rely on the following code from the *dplyr* package:

```
df <- dataset_w_missing %>%
  na.omit()
```

### *Estimating Propensity Score*

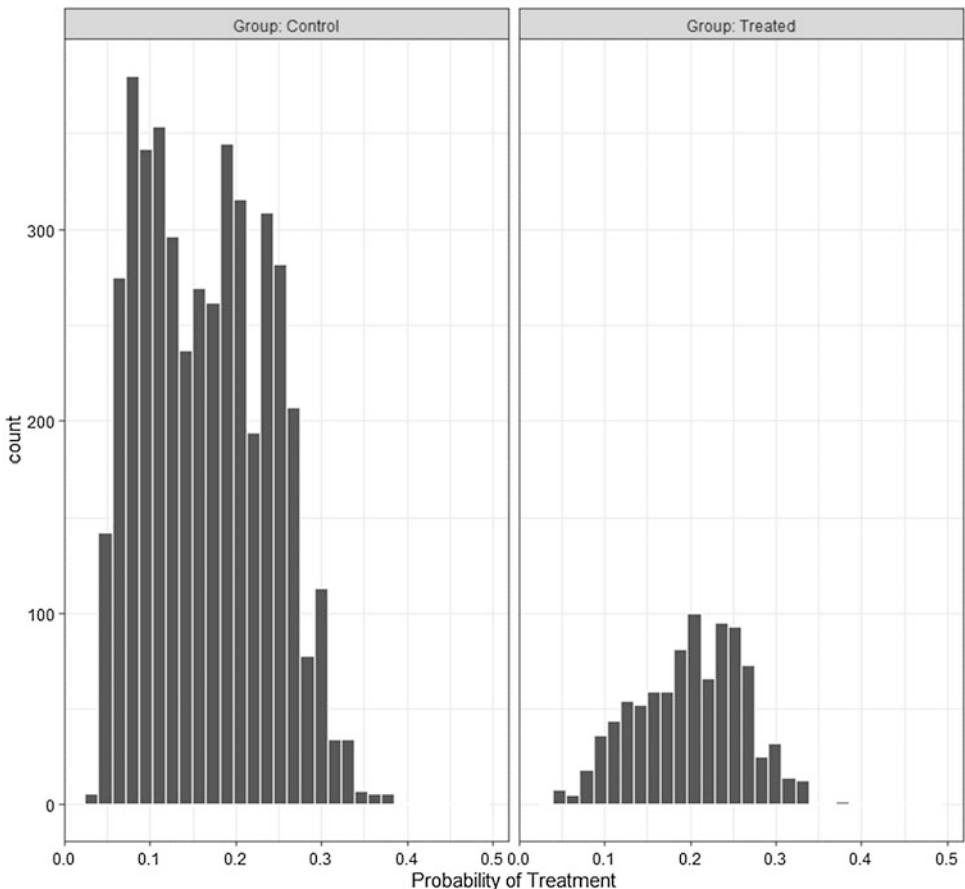
As you will see in the next section, the `matchit()` function that we will be using calculates propensity scores during the matching process for us, but you can do it yourself by using the `glm()` and `predict()` functions. Here, we estimate a propensity score and store it in a variable named `ps_score`. The `treat` variable in the model is the exposure to treatment variable, which is dichotomous so make sure to specify **family = binomial**. To use the `glm()` and `predict()` functions, make sure to first install and load the `stats` and `car` packages.

```
logit <- glm(treat ~ inv1 + inv2 + inv3,
  data = df_name, family = binomial)
df_name$ps_score<-predict(logit, type = "response")
```

You can examine the region of common support with the estimated propensity score via the `ggplot()` function. If you have not done so already, install and load the `ggplot2` package before running the syntax.

```
labs <- paste("Group:", c("Treated", "Control"))
```

```
df_name %>%
  mutate(treat = ifelse(treat == 1, labs[1], labs[2])) %>%
  ggplot(aes(x = ps_score)) +
  geom_histogram(color = "white") +
  facet_wrap(~treat) +
  xlab("Probability of Treatment") +
  theme_bw()
```



### *Matching Cases*

You can match using a *caliper* via the **caliper** = option, which restricts matching based on a specified number of standard deviations of the propensity score. In this

example, we use the `matchit()` function to conduct 1:1 matching with a caliper of .025 set:

```
match.it <- matchit(treat ~ inv1 + inv2 + inv3,
                     data = df_name, caliper = .025)
```

The ratio = option allows you to specify the *maximum* number of control cases to match to treated cases. A ratio of 1 (default) produces 1:1 matching, a ratio of 2 produces 2:1 matching, and so forth. You may also specify if you want a control cases to be matched with more than one treated case via the **replace** = option. The default is no replacement.

*Nearest neighbor* with replacement and a caliper:

```
match.it <- matchit(treat ~ inv1 + inv2 + inv3,
                     data = df_name, replace = TRUE, caliper = .025)
```

Matching treated and control cases when the propensity scores match *exactly*:

```
match.it <- matchit(treat ~ inv1 + inv2 + inv3,
                     data = df_name, method = "exact")
```

Alternatively, you can *stratify* cases by propensity level too. If the number of subclasses is not specified, then 6 subclasses will be created. In the example below, 10 strata are created.

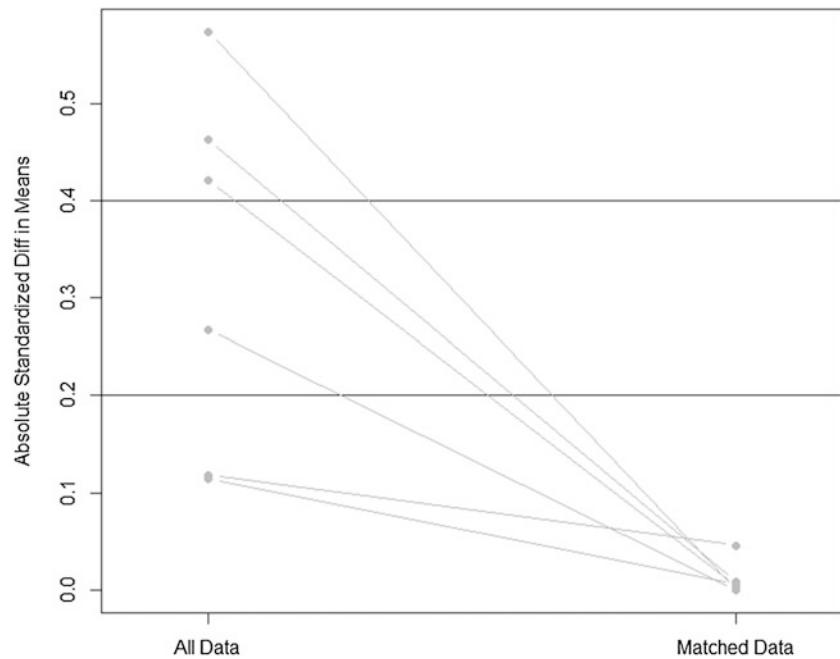
```
match.it <- matchit(treat ~ inv1 + inv2 + inv3,
                     data = df_name, method = "subclass", subclass=10)
```

### *Assessing Matches*

There are several options for examining matches using the `summary()` or `plot()` functions.

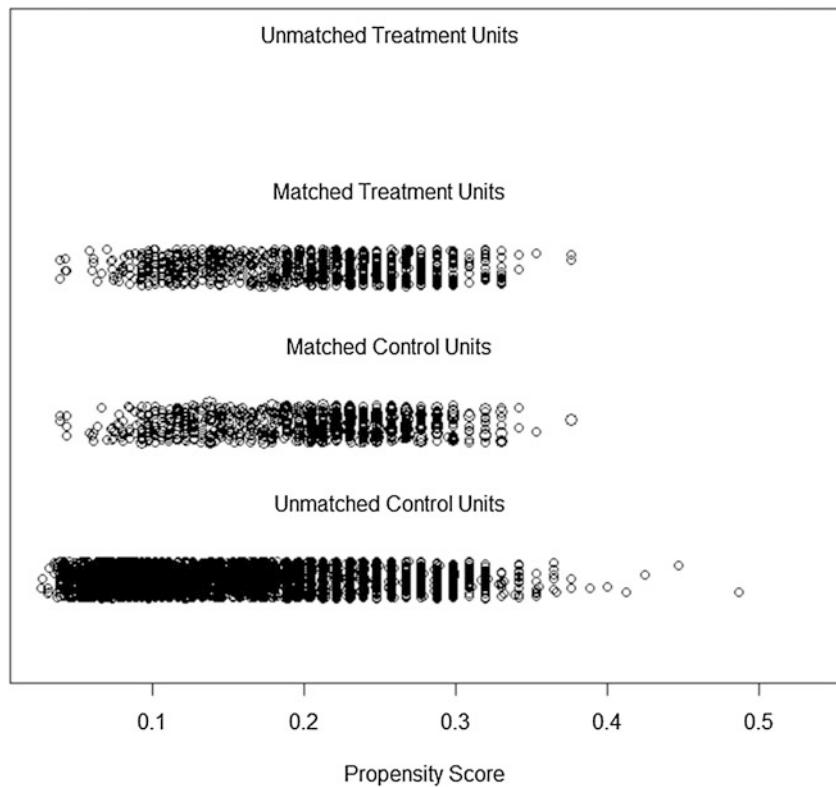
```
#Provides balance table
summary(match.it, standardize=T)
```

```
# Plots absolute std. mean diff before and after matching
plot(summary(match.it, standardize=T), interactive=F)
```

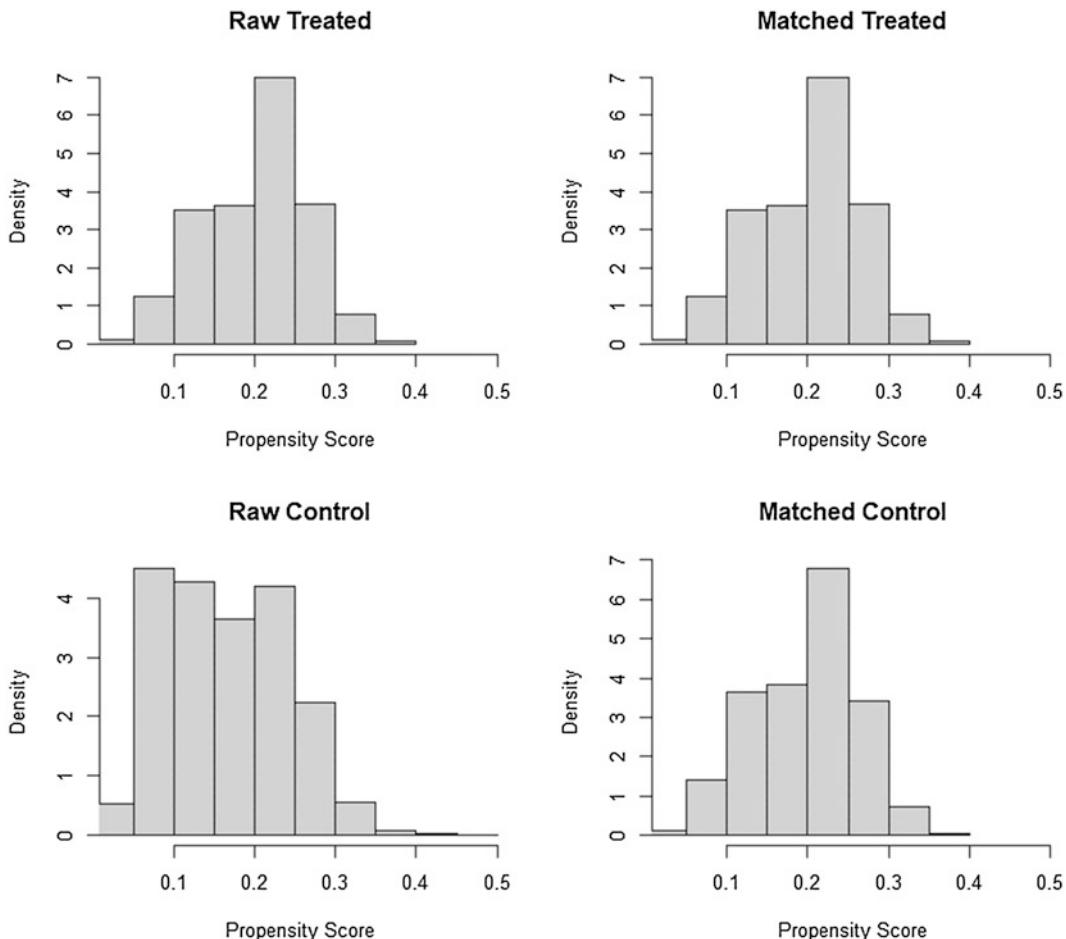


```
# Plot of propensity scores using jitter
plot(match.it, type = "jitter", interactive = F)
```

### Distribution of Propensity Scores



```
# Propensity score histogram plot
plot(match.it, type = "hist")
```



Once we have created matches, we can put matched treated and control cases in a single data frame. Here, we have the results from the `matchit()` function stored in an object named `match.it`, and we are putting the matched data into an object called `my.matches` by using the `match.data()` function from the *MatchIt* packages.

```
my.matches <- match.data(match.it)
```

Now, you can use the data frame containing the matched data to assess equivalence on the covariates by whether the cases were treated using the `t.test()` function. This example examines the difference in means between matched treated and control cases (`match` variable) on one of the independent variables (`iv1` variable).

```
with(my.matches, t.test(iv1 ~ match))
```

You can examine mean differences when using stratification by conducting the *t*-test for a selected subclass, or you can examine *t*-tests for each subclass by using a for-loop.

```
# t-test for subclass 1
t.test(iv1~treat, data=my.matches, subset=subclass==1)

# t-test for all 10 subclasses
for (x in 1:10) {
  print(t.test(iv1~treat, data= matched.data,
  subset=subclass==x))
}
```

### *Estimating Treatment Effect*

One way to assess the treatment effect on the outcome is using the `t.test()` function. Make sure that the subset of data you are working with only contains the matched cases (they should already be since we selected them out in the prior section).

```
t.test(outcome~treat, data=matched.data)

# Stratified sample (stratum #1)
t.test(outcome~treat, data=matched.data, subset=subclass==1)
```

You also have the option to assess the treatment effect using the `glm()` function. Remember to add *family = binomial* to the `glm()` function if your outcome is binary. Also, as noted, make sure that you are using the subset of data that contains only matched cases. Last, use the `summary()` function to view details of the regression model you estimated.

```
my.glm <- glm(outcome ~ treat, data = matched.data)

# Stratified sample
my.glm <- glm(outcome ~ treat + factor(subclass),
  data = matched.data)

# Binary outcome
my.glm <- glm(outcome ~ treat,
  data = matched.data, family = binomial)

summary(my.glm)
```

### Problems

Open the dataset provided for this chapter (*psm.dta*) to complete Exercises 10.1–10.3. *Psm.dta* is a simulated dataset of a sample of inmates who are incarcerated for a drug offense or drug-involved offense. Some of the inmates in the dataset participated in a prison program designed to reduce overall criminal thinking. Those inmates who participated in the program will have a 1 on the *treat* variable and will receive a 0 otherwise. The key variable of interest is the variable named *outcome*, a score that serves as a measure of criminal thinking that was obtained using a survey of inmates after the program ended. The dataset also contains variables to indicate the inmates' race/ethnicity (dummy variables: *race\_white*, *race\_black*, *eth\_hispanic*, *race\_asian*), level of income (ordinal variable: *income*), age (age variable), highest level of education completed (*education*), whether the inmate lived below the poverty level prior to their current incarceration (*poverty*), measure of their drug addiction severity (*drugadd*), whether they had a prior incarceration (*prior\_inc*), whether they participated in other prison programs during their current incarceration (*other\_prog*), and their level of association with antisocial peers (*antisocial\_assoc*).

1. Create a propensity score model with 3–4 covariates of your choice from the dataset to estimate the effect of the prison program on criminal thinking. Once you have completed the model, report the following:
  - (a) Details about the model (covariates selected, type of match type used)
  - (b) Details about covariance balance between the treated group and matched controls (e.g., percent bias, differences in means, *t*-test results)
  - (c) Visualization of the region of common support
  - (d) Your findings regarding the estimated treatment effect
2. Specify the same propensity score model as you did in Problem 1, but add other covariates of your choice to your model.
  - (a) Details about the model (covariates selected, type of match type used)
  - (b) Details about covariance balance between the treated group and matched controls (e.g., percent bias, differences in means, *t*-test results)
  - (c) Visualization of the region of common support
  - (d) Your findings regarding the estimated treatment effect
  - (e) Describe how the revised model compares to the one you ran in 1. How did the additional covariates affect the level of bias? Does the revised model change your overall conclusions on the effectiveness of the prison program?

3. Conduct another propensity score model with at least three covariates, but select a matching technique that is different from the one you employed in Problems 1 and 2.
  - (a) Details about the model (covariates selected, type of match type used)
  - (b) Details about covariance balance between the treated group and matched controls (e.g., percent bias, differences in means, *t*-test results)
  - (c) Visualization of the region of common support
  - (d) Your findings regarding the estimated treatment effect
  - (e) Describe how the revised model compares to the models you ran for 1 and 2. How did the additional covariates affect the level of bias? Does the revised model change your overall conclusions on the effectiveness of the prison program?

## References

---

- Aakvik, A. (2001). Bounding a matching estimator: the case of a Norwegian training program. *Oxford Bulletin of Economics and Statistics*, 63(1), 115–143.
- Apel, R. J., & Sweeten, G. (2010). Propensity score matching in criminology and criminal justice. In A. Piquero & D. Weisburd (Eds.), *Handbook of quantitative criminology*. New York, NY: Springer. [https://doi.org/10.1007/978-0-387-77650-7\\_26](https://doi.org/10.1007/978-0-387-77650-7_26).
- Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*, 33(6), 1057–1069.
- Becker, S. O., & Caliendo, M. (2007). *MHbounds – Sensitivity analysis for average treatment effects*. Retrieved from <http://ftp.iza.org/dp2542.pdf>.
- Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1), 31–72.
- Harrell, F.E., Jr., & Slaughter, J.C. (2020). *Biostatistics for biomedical research*. Retrieved from <http://hbiostat.org/doc/bbr.pdf>.
- King, G., & Nielsen, R. (2016). *Why propensity scores should not be used for matching*. Download Paper, 378. Retrieved from <http://j.mp/1sexgVw>.
- Loughran, T. A., Wilson, T., Nagin, D. S., & Piquero, A. R. (2015). Evolutionary regression? Assessing the problem of hidden biases in criminal justice applications using propensity scores. *Journal of Experimental Criminology*, 11(4), 631–652.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4), 719–748.
- Neuhäuser, M., Thielmann, M., & Ruxton, G. D. (2018). The number of strata in propensity score stratification for a binary outcome. *Archives of Medical Science: AMS*, 14(3), 695.
- Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York, NY: Springer.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.

- Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1), 33–38.
- Shadish, W. R. (2013). Propensity score analysis: promise, reality and irrational exuberance. *Journal of Experimental Criminology*, 9(2), 129–144.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, MI: Mifflin and Company.
- Shadish, W. R., & Steiner, P. M. (2010). A primer on propensity score analysis. *Newborn and Infant Nursing Reviews*, 10(1), 19–26.
- Wang, Y., Cai, H., Li, C., Jiang, Z., Wang, L., Song, J., & Xia, J. (2013). Optimal caliper width for propensity score matching of three treatment groups: A Monte Carlo study. *PLoS ONE*, 8(12), e81045.
- Weisbord, D., Hasisi, B., Shoham, E., Aviv, G., & Haviv, N. (2017). Reinforcing the impacts of work release on prisoner recidivism: The importance of integrative interventions. *Journal of Experimental Criminology*, 13(2), 241–264.

## Chapter eleven

---

# Meta-analysis

---

## Why Use Meta-analysis to Synthesize Results Across Studies?

---

What is the logic of meta-analysis?

What advantages does it have over traditional methods of reviewing studies?

---

## Effect Sizes

---

What are the different types of effect sizes commonly used in meta-analysis?

What type of data is appropriate for each commonly used effect size?

How do you convert among effect size types?

---

## The Average Effect Size

---

How do we compute a mean effect size?

Why do we weight by the inverse variance?

What is the difference between a fixed-effect and random-effects model?

How do we test for the heterogeneity of effect sizes?

How do we handle statistically dependent effect sizes?

## **Moderator Analysis**

---

What is the purpose of moderator analysis?

How do we assess the relationship between a categorical study feature and effect size?

How do we assess the relationship between a scaled variable or multiple variables and effect size?

## **Publication Selection Bias**

---

What is publication selection bias?

How do we minimize and assess for publication selection bias?

**M**ETA-ANALYSIS IS A STATISTICAL METHOD for synthesizing results across studies that examines a common research question. It is the analysis of analyses. The fundamental logic of meta-analysis shifts the focus from the individual outcomes and statistical significance of the studies being synthesized to the direction and magnitude of the overall effect of the outcome of interest. It is, after all, the latter and not the former that we are truly interested in. Each study in a meta-analysis is an observation, and the analysis examines the pattern of results across studies. In addition to producing an average effect, the methods allow for an examination of the consistency of results across studies and the relationship between study features and observed results.

This chapter presents the basic statistical methods that are part of a **meta-analysis**. However, a meta-analysis must be the result of a **systematic review**. Systematic review methods developed as part of the methods of meta-analysis but reflect the nonstatistical aspects of a meta-analysis. The essential features of a systematic review are as follows: (1) a systematic and documented search for published and unpublished studies, the latter often being called the *gray literature*; (2) explicit and detailed study inclusion and exclusion criteria that reflect the goals of the review; (3) a detailed coding protocol for extracting information from the eligible studies, including both study characteristics and effect size data; (4) double coding of all studies to ensure coding reliability; (5) an assessment of study validity, such as an assessment of risk-of-bias; and (6) the systematic and transparent presentation of results. Ideally, a systematic review would also include, if possible, a meta-analysis of study-level effect sizes. The goal is to apply the principles of good research, that is, replicability and transparency, to the task of synthesizing findings across studies. We also recommend registering an *a priori* protocol for the review methods, either with an organization such as the Campbell Collaboration<sup>1</sup> or with an online repository such as PROS-

---

<sup>1</sup><http://campbellcollaboration.org>

PERO.<sup>2</sup> The written manuscript should follow one of the various guidelines for reporting the results of a meta-analysis, such as the Meta-Analysis Reporting Standard<sup>3</sup> (MARS) or PRISMA.<sup>4</sup> Both the Campbell and Cochrane Collaborations also have standards for conducting and reporting of systematic reviews and meta-analyses.

## A Historical Note

---

In 1976, Gene Glass (1976) coined the term meta-analysis. Shortly thereafter, he and Mary Lee Smith (Smith and Glass 1977) published two seminal examples, one on the effectiveness of psychotherapy and another on the effects of class size on academic achievement (Glass and Smith 1979). Other scholars were independently developing methods of statistically synthesizing results across studies. Schmidt and Hunter (1977) published a review of the predictive validity of employment tests using a method now referred to as the Hunter–Schmidt method of psychometric meta-analysis. A third but similar method was developed by Rosenthal and Rubin (1978), resulting in a meta-analysis of 345 studies from psychology examining interpersonal expectancy effects.

In his history of meta-analysis, Hunt (1997, see also, O’Rourke 2007) argues that the roots of meta-analysis are deep and significantly pre-date the work of Glass, Hunter and Schmidt, and Rosenthal and Rubin. For example, in 1904, Karl Pearson averaged the correlations of inoculation for typhoid fever with both mortality and infection in British soldiers across numerous studies. Other statisticians, including R. A. Fisher, recognized the need to combine results across studies and contributed ideas and statistical methods that were incorporated into modern meta-analysis. The popularity of the method has seen explosive growth since the late 1970s and is widely used throughout the social sciences, medicine, and ecology. Examples can also be found in natural science disciplines other than ecology along with alternative but statistically similar methods for synthesizing results across studies (see, for example, Baker and Jackson 2013).

---

<sup>2</sup><https://www.crd.york.ac.uk/PROSPERO/>

<sup>3</sup><https://wmich.edu/sites/default/files/attachments/u58/2015/MARS.pdf>

<sup>4</sup><http://primsa-statement.org>

## The Logic of Meta-analysis

---

As a method of taking stock of the evidence on a topic, meta-analysis stands in contrast to classic narrative review methods. Narrative reviews draw upon the expertise of the reviewers to assess and evaluate the research literature and generally include more qualitative assessments of what the research tells us, often relying on the statistical significance of individual studies to draw conclusions. Most journal articles include some type of narrative review in introducing their study and contextualizing it in the research literature. However, narrative reviews that are systematic and comprehensive in their scope are produced by such groups as the National Academy of Sciences. Narrative reviews typically lack replicable and transparent methods.

The reliance on statistical significance of individual studies in a typical narrative review is a particularly problematic approach. While statistical significance is a useful tool for ensuring caution in interpreting the results from a single study, its usefulness breaks down when trying to make sense of a collection of results. The fundamental problem with statistical significance is the asymmetric nature of null hypothesis significance testing: A significant finding allows for a stronger conclusion than a nonsignificant finding. Recall that a statistically significant effect allows us to reject the null hypothesis. The likelihood of our doing so by mistake equals our *a priori* significance level, typically set at either 1% or 5%. Thus, assuming we have satisfied the assumptions of the statistical test, we can confidently reject the null hypothesis with a known level of uncertainty. A statistically nonsignificant effect, however, does not allow us to accept the null; rather, we *fail to reject* the null. By itself, a statistically nonsignificant *p*-value provides *no* information about the plausibility that the null is true. Across any collection of studies examining the same research question, there is bound to be a mix of statistically significant and statistically nonsignificant effects. The larger the number of studies, the more mixed the evidence is likely to appear from the perspective of statistical significance. How are we to balance *x* number of statistically significant findings against *y* number of statistically nonsignificant findings? There is no statistically defensible method for doing so.

A study with a statistically nonsignificant finding does, however, contain valuable information and that information may or may not be in support of the null. That information is the size of the observed effect and the associated precision of that effect, often reflected as a confidence interval. A null finding with an observed effect very close to the null value and a very tight confidence interval provides evidence that the true effect may be zero or close enough to zero to be of no substantive value. However, a null effect that is of a meaningful magnitude with a large confidence interval provides weak evidence against the null. Notice that the focus has shifted from

statistical significance to the direction and magnitude of the effect and the precision of that effect. It is this focus on the actual effects observed across a collection of studies that is the key to meta-analysis.

By focusing on the direction and magnitude of effects, we avoid the problem of the asymmetric nature of significance tests. In addition to examining the average effect across studies, we can statistically assess the consistency of effects as well as explore potential explanations for inconsistencies based on coded study features. This provides a statistically justifiable basis for drawing inferences about why some studies may find larger effects than others. Meta-analysis is the statistical method that makes all of this possible.

## The Effect Size

---

The **effect size** is the key to meta-analysis. We are using the term *effect size* in a generic sense and are not referring to any particular statistical index. We are also not using the word *effect* to connote causation, as in cause-and-effect, although in some cases the effect size may have a causal interpretation based on the type of research design being synthesized. There are many well-established effect sizes, and the effect size of choice should be selected to fit the nature of the data within the collection of studies. The essential feature is that the effect size encodes the research result of interest in such a way that it is numerically comparable across studies.

The most widely used effect sizes for meta-analysis are the standardized mean difference, the correlation coefficient, the risk ratio, and the odds ratio. Each is best suited to a specific type of research design and associated data, as summarized below:

- **Standardized mean difference:** The standardized mean difference, commonly called Cohen's  $d$  or Hedges'  $g$ , standardizes the difference between two means relative to the pooled standard deviation within the groups (see Chap. 8). Thus, it reflects the difference between the groups in units of the standard deviation on the dependent variable. Hedges'  $g$  is Cohen's  $d$  adjusted for small sample size bias. This effect size is suitable for studies comparing two groups on a scaled dependent variable, that is, a dependent variable on which computing a mean is appropriate and meaningful. The two groups can reflect either experimentally created conditions, such as a treatment versus a control group, or naturally occurring groups, such as boys versus girls.
- **Correlation coefficient:** The correlation coefficient is a standardized index of the strength of the relationship between two variables, with values that can range between  $-1$  and  $+1$ . Generally, both variables will be measured on a scale, although one or both may be binary. This effect size

is suitable for meta-analyzing correlational studies, that is, studies that are examining the relationship between two observed variables. Lots of examples of such meta-analyses can be found in social and industrial/organizational psychology.

- **Risk ratio:** The risk ratio is the ratio of the probability of success (or failure) in one group relative to another. As such, it is suitable for studies comparing two groups, such as treatment and control, on a binary outcome variable.
- **Odds ratio:** The odds ratio is similar to the risk ratio but uses the *odds* of success (or failure) instead of the *probability* of success (or failure) in constructing the ratio. An odds is the probability of success over the probability of failure. Given this similarity, the odds ratio is also suitable for studies comparing two groups on a binary outcome variable along with the risk ratio. However, unlike the risk ratio that is only suitable for prospective research designs, the odds ratio can be used for retrospective case-control studies. These studies select a sample based on the outcome and then observe the status of each observation on a prior risk or some other independent variable. For example, a study by Needleman et al. (2002) selected a sample of adjudicated delinquents (cases) and a sample of non-delinquents (controls) and measured their bone lead levels, finding that adjudicated delinquents had significantly higher bone lead levels relative to controls.

Less commonly used effect size indices include simple point estimates, such as a mean or a proportion, standardized gain or change scores, and standardized regression coefficients. Two examples of the latter are a meta-analysis of deterrence theory by Pratt et al. (2006) and a meta-analysis of effect of race on an officer's arrest decision Kochel et al. (2011). There are several statistical complications with meta-analyzing results from multiple regression that go beyond this introductory chapter (see Kim 2011).

The sections below present the computations of the four more widely used effect size indices along with any statistical adjustments or transformation that are needed prior to analysis.

### **The Standardized Mean Difference: Cohen's *d* and Hedges' *g***

The standardized mean difference effect size is based on the means, standard deviations, and sample sizes of the two groups being compared, as shown in Eq. (11.1):

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}}, \quad \text{Equation 11.1}$$

where  $\bar{x}_1$  is the mean of the first group, such as the treatment group,  $\bar{x}_2$  is the mean of the second group, such as the control group, and  $s_{pooled}$  is the pooled within-groups standard deviation. The latter is computed using Eq. (11.2):

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}, \quad \text{Equation 11.2}$$

where  $s^2$  and  $n$  are the standard deviation and sample sizes for the respective groups. Notice that  $s_{pooled}$  is simply the square root of the weighted average of the two variances. Thus, it will always produce a value that is between the two standard deviations for the groups. This differs from the overall standard deviation by removing any variability associated with the group contrast (e.g., treatment). It is the denominator of Eq. (11.1) that *standardizes* the mean difference such that the difference between the means is expressed in standard deviation units. In this way, we can compare effects across studies that used different measures of a common underlying construct of interest.

Hedges and Olkin (1985) established that Cohen's  $d$  is upwardly biased when based on small sample sizes. This becomes more pronounced with group sample sizes less than 20. Equation (11.3) provides a close approximation to the correction factor, and Eq. (11.4) adjusts  $d$  by this factor, producing Hedges'  $g$ :

$$J = 1 - \left[ \frac{3}{4(n_1 + n_2) - 9} \right], \quad \text{Equation 11.3}$$

$$g = J \times d. \quad \text{Equation 11.4}$$

It has become standard practice in meta-analysis to use Hedges'  $g$  over Cohen's  $d$ ; although in most situations in criminology, the difference is trivial.

Studies do not always report all of the data needed to compute  $d$  and  $g$  using the above equations. Fortunately, there are other ways to compute these effect sizes from related statistical information. For example,  $d$  can easily be computed from an independent  $t$ -test, as shown in Eq. (11.5):

$$d = t \sqrt{\frac{n_1 + n_2}{n_1 n_2}}. \quad \text{Equation 11.5}$$

This equality reflects the similarity of the equations for  $d$  and for  $t$ . Notice that Eq. (11.6) for the independent  $t$ -test is the same as the composite of Eqs. (11.1) and (11.2) with the addition of the right most radical in the denominator. Equation (11.5) backs out that component of the equation, producing  $d$ .

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \sqrt{\frac{n_1+n_2}{n_1n_2}}} \quad \text{Equation 11.6}$$

There are numerous other methods of computing  $d$  and by extension  $g$  from other statistical results reported in studies. These are detailed by Lipsey and Wilson (2001a) and Borenstein et al. (2011).

In order to conduct a meta-analysis, we also need the standard error of  $g$ , as this serves as the basis for the inverse-variance weight needed for analysis. This is computed using Eq. (11.7):

$$se_g = \sqrt{\frac{n_1 + n_2}{n_1n_2} + \frac{g^2}{2(n_1 + n_2)}}. \quad \text{Equation 11.7}$$

### Risk Ratio

The risk ratio is suitable for comparing two groups or conditions on a binary outcome variable. Typically, the risk ratio is used in the context of an experimental study comparing two treatment conditions or a treatment versus a control condition. We can represent data from such a study in a 2 by 2 contingency table, as shown in Table 11.1, where  $a$ ,  $b$ ,  $c$ , and  $d$  are the number of individuals or other units in each cell of the table.

The risk ratio is the ratio of the probabilities (i.e., proportions) of success ( $p$ ) for each group, as shown in Eq. (11.8):

$$\text{RR} = \frac{a/(a+b)}{c/(c+d)} = \frac{p_1}{p_2}, \quad \text{Equation 11.8}$$

where  $a$ ,  $b$ ,  $c$ , and  $d$  are the cell frequencies and  $p_1 = a/(a+b)$  and  $p_2 = c/(c+d)$ . Notice that the risk ratio can be computed from either the raw frequencies or the two probabilities. A risk ratio of 1 indicates that both groups have the same probability of success. Thus, 1 is the null value of no effect. Values less than 1 indicate a higher success probability for the

**Table 11.1**

Example 2 by 2 frequency table

	OUTCOME	
	SUCCESS	FAILURE
Treatment	$a$	$b$
Control	$c$	$d$

treatment condition relative to the control condition with values greater than 1 indicating the opposite (this can be flipped if that makes more sense in a particular research context). The distribution of effects is asymmetrical: negative effects (i.e.,  $RR < 1$ ) increase from 1 to 0, whereas positive effects (i.e.,  $RR > 1$ ) increase from 1 to  $\infty$ . A negative effect of  $RR = 0.50$  indicates that the probability of success in the control condition was half that of the treatment condition. Flipping this around would produce an  $RR = 2.00$ , indicating that the treatment condition had twice the probability of failure relative to the control condition.

The asymmetric nature of the risk ratio creates a problem for meta-analyzing these effect sizes. Principally, they are not on a linear scale and there is no calculable standard error. The solution is to take the natural logarithm. Effects on the logged risk ratio scale grow linearly and have an easily calculable standard error. The null value of the logged risk ratio is 0 (the natural log of 1 is 0), and the distribution of effects is symmetrical. For example, the natural log of 0.50 is  $-0.69$  and the natural log of 2.00 is  $+0.69$ . Equation (11.9) provides the standard error of the logged risk ratio:

$$se_{\ln(RR)} = \sqrt{\frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}} \quad \text{Equation 11.9}$$

Meta-analysis is performed on the logged risk ratios, and final results are converted back into risk ratios by taking the exponent (i.e., reversing the log transformation).

### Odds Ratio

The odds ratio is similar to the risk ratio and can be used in all situations where the risk ratio is suitable. The odds ratio is also suitable for case-control designs, whereas the risk ratio is not. A disadvantage of the odds ratio is that it is more difficult to interpret, at least for most people. That is because it reflects the effect in terms of the relative odds of success versus failure in each condition rather than the relative probability. The computation of the odds ratio is shown in Eq. (11.10):

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1(1-p_2)}{p_2(1-p_1)} = \frac{ad}{bc}, \quad \text{Equation 11.10}$$

where the terms are defined as above. As with the risk ratio, the odds ratio has a null value of 1 and is asymmetrically distributed. This is addressed in the same manner as with the risk ratio. Thus, analyses are performed on the logged odds ratios which have a standard error shown in Eq. (11.11):

$$se_{\ln(\text{OR})} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

**Equation 11.11**

Final results from a meta-analysis of logged odds ratios are converted back into odds ratios by reversing the natural log function through exponentiation.

Computing a risk ratio or odds ratio from a primary study is generally straightforward. The data from a study comparing two groups on a binary outcome will typically be reported either as a 2 by 2 frequency table, as shown in Table 11.1, or as the proportion or percent of successes or failures within each group, along with the group sample sizes. Quasi-experimental studies comparing two groups on a binary outcome will often perform a logistic regression analysis that includes observed baseline variables. The regression coefficient for the treatment dummy variable is a logged odds ratio and can be used directly in a meta-analysis. In these cases, the standard error from the regression model should be used, rather than Eq. (11.11).

### **Correlation Coefficient**

Lots of studies in the social sciences, including criminology and criminal justice, are correlational in nature. A natural effect size for such studies is simply the correlation coefficient, an index that is standardized with 0 indicating no relationship between the variables and  $-1$  and  $+1$  indicating a perfect negative and positive relationship, respectively. Thus, correlation coefficients can be compared across studies, assuming that they reflect a common relationship of interest. Studies of this type will typically report the correlation directly, greatly ease the coding and data extraction process.

A complication with the correlation coefficient is that it does not have a calculable standard error given that its possible values are constrained within the range of  $-1$  and  $+1$ , inclusively. The solution to this is to use Fisher's  $Z_r$  transformation, shown in Eq. (11.12):

$$Z_r = \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right).$$

**Equation 11.12**

This transformation stretches out the tails of the distribution with the values of  $-1$  and  $+1$  equalling  $-\infty$  and  $+\infty$ . (This value should not be confused with  $z$  from the standardized normal distribution, as they are distinctly different.) This can be seen in Table 11.2. For small to modest correlations, this transformation increases  $r$  slightly but for large correlations the increase is substantial. For example, for an  $r$  of 0.20,  $Z_r$  equals 0.203, whereas for an

**Table 11.2**Relationship between  $r$  and Fisher's  $Z_r$ 

$r$	$Z_r$
0.000	0.000
0.100	0.100
0.200	0.203
0.300	0.310
0.400	0.424
0.500	0.549
0.600	0.693
0.700	0.867
0.750	0.973
0.800	1.099
0.850	1.256
0.900	1.472
0.950	1.832
0.960	1.946
0.970	2.092
0.980	2.298
0.990	2.647
0.999	3.800
0.9999	4.952
1.000	$+\infty$

$r$  of 0.75,  $Z_r$  equals 0.973. Also, notice that as  $r$  approaches 1,  $Z_r$  grows dramatically.

An important benefit of this transformation is an easy to calculate standard error, shown in Eq. (11.13):

$$se_{Z_r} = \frac{1}{\sqrt{n-3}}, \quad \text{Equation 11.13}$$

where  $n$  is the sample size for the correlation. The results from a meta-analysis of  $Z_r$  effect sizes can be converted back into  $r$  values through exponentiation using Eq. (11.14):

$$r = \frac{e^{2(Z_r)-1}}{e^{2(Z_r)+1}}. \quad \text{Equation 11.14}$$

### Converting Between Effect Size Indices

A common problem when conducting a meta-analysis is that not all effects across studies will conform to a single effect size type. This is most likely to occur when using Cohen's  $d$ /Hedges'  $g$  as the effect size of choice where a subset of studies will report the outcome of interest dichotomized, rather than on a scaled (multivalued) measure. For example, in a meta-analysis

examining the effectiveness of a cognitive-behavioral program for depression, depression may be measured on a scale across most studies but dichotomously (depressed versus not depressed) for a small subset. Dropping such studies from the meta-analysis seems wasteful. Fortunately, there are methods for converting among the effect sizes discussed above, providing a method of computing a common effect size type based on scaled and dichotomous measures. Note, however, that these conversions are approximations. That is, the conversion methods for estimating Cohen's  $d$  from binary outcome data only approximates what the  $d$  would have been had the outcome been measured and reported on a scaled measure.

### *Converting Effect Sizes into Cohen's d*

There are three basic methods for computing Cohen's  $d$  from dichotomized or binary outcome data. Two of these rely on converting from the logged odds ratio. The logged odds ratio follows the logistic distribution. This distribution is symmetrical and roughly normal, although it is slightly leptokurtic. The standard deviation of the logistic distribution is as follows:

$$\sqrt{\frac{\pi^2}{3}}. \quad \text{Equation 11.15}$$

Because Cohen's  $d$  is scaled in terms of standard deviation units based on the standardized normal distribution, we can approximate Cohen's  $d$  by rescaling the logged odds ratio from the logistic distribution to the standardized normal distribution. This conversion is shown in Eq. (11.16):

$$d = \frac{\ln(\text{OR})}{\sqrt{\frac{\pi^2}{3}}} = \frac{\ln(\text{OR})}{1.814}. \quad \text{Equation 11.16}$$

The standard error is similarly converted, as shown in Eq. (11.17):

$$se_d = \sqrt{\frac{se_{\ln(\text{OR})}^2}{\left(\frac{\pi^2}{3}\right)}} = \sqrt{\frac{se_{\ln(\text{OR})}^2}{1.814^2}} \quad \text{Equation 11.17}$$

Because the logistic distribution is more peaked with fatter tails than the normal distribution, Cox proposed an alternative conversion that provides better overlap between the two distributions in the tails. This method divides the logged odds ratio by 1.65 rather than 1.814, as shown in Eq. (11.18):

$$d = \frac{\ln(\text{OR})}{1.65}.$$
Equation 11.18

The calculation of the standard error for the Cox method of computing Cohen's  $d$  is shown below as Eq. (11.19):

$$se_d = \sqrt{\frac{se_{\ln(\text{OR})}^2}{1.65^2}}.$$
Equation 11.19

A final method computes Cohen's  $d$  using probits. A probit is the  $z$  value from the standardized normal distribution with an area under the curve in the left tail that equals the probability of success or failure for a group. This method is shown in Eq. (11.20):

$$d = \text{probit}(p_1) - \text{probit}(p_2) = z_1 - z_2.$$
Equation 11.20

The standard error of  $d$  using the probit method is shown below as Eq. (11.21):

$$se_d = \sqrt{\frac{2\pi p_1(1-p_1)e^{z_1^2}}{n_1} + \frac{2\pi p_2(1-p_2)e^{z_2^2}}{n_2}}.$$
Equation 11.21

These three methods generally produce fairly similar results, with the logit and Cox logit methods being computationally simpler. For example, if the failure rates for the treatment and control groups are 0.25 and 0.33, the logged odds ratio is  $-0.39$ . The respective logit, Cox logit, and probit conversions are  $-0.215$ ,  $-0.237$ , and  $-0.235$ . In general, the Cox logit and probit methods will be most similar and the choice among these will rarely produce substantively different results. Sánchez-Meca and colleagues performed computer simulations on these transformation methods and showed that they work reasonably well under most situations (Sánchez-Meca et al. 2003).

It is also possible to convert correlation coefficients into Cohen's  $d$ . This should only be done if one of the variables that is part of the correlation is binary, reflecting two groups. Otherwise, the resulting  $d$  will not be meaningful. These are called *point-biserial* correlations, and Eq. (11.22) provides the conversion of such an  $r$  to  $d$ :

$$d = \frac{2r}{\sqrt{1 - r^2}}.$$
Equation 11.22

Equation (11.23) shows how to calculate the standard error of a  $d$  based on Eq. (11.22) converted from the standard error for  $r$ :

$$se_d = \sqrt{\frac{4se_r^2}{(1 - r^2)^3}}.$$
Equation 11.23

### *Converting Effect Sizes into Odds Ratios*

Just as it is possible to convert a logged odds ratio into a Cohen's  $d$ , it is also possible to reverse this conversion and convert Cohen's  $d$  into a logged odds ratio. The need to do this might occur in a meta-analysis where most of the studies measure the outcome construct of interest dichotomously, but a few studies measure it on a scaled variable. By first calculating Cohen's  $d$ , the logged odds ratio and associated standard error can be approximated using the inverse of the logit or Cox logit methods, as shown in Eqs. (11.24–11.27) below:

$$\ln(\text{OR}) = \frac{d}{0.551},$$
Equation 11.24

$$se_{\ln(\text{OR})} = \sqrt{\frac{se_d^2}{0.551^2}},$$
Equation 11.25

$$\ln(\text{OR}) = \frac{d}{0.606},$$
Equation 11.26

and

$$se_{\ln(\text{OR})} = \sqrt{\frac{se_d^2}{0.606^2}},$$
Equation 11.27

There is no comparable inverse for the probit method. It is also not possible to convert Cohen's  $d$  into a logged risk ratio. However, it is possible to convert a risk ratio into an odds ratio, as shown in Eq. (11.28):

$$\text{OR} = \frac{\text{RR} \times p_2 \times (1 - p_2)}{p_2 \times (1 - \text{RR} \times p_2)}, \quad \text{Equation 11.28}$$

where  $p_2$  is the probability of failure in the control group. This assumes that the risk ratio was calculated as the treatment probability over the control probability. If this is not the case, then the inverse of the risk ratio must be used.

### *Converting Effect Sizes into Risk Ratios*

The odds ratio can be converted into a risk ratio using Eq. (11.29) shown below:

$$\text{RR} = \frac{\text{OR}}{1 - p_2 + p_2(\text{OR})}, \quad \text{Equation 11.29}$$

where  $p_2$  is the probability of failure in the control group. It is not sensible to convert other effect sizes, such as Cohen's  $d$  into a risk ratio.

### *Converting Effect Sizes into Correlations*

Just as a correlation coefficient can be converted into a Cohen's  $d$ , the reverse is also true. In this case, the resulting correlation coefficient is a point-biserial correlation. The equation for this conversion is as follows:

$$r = \frac{d}{\sqrt{d^2 + \frac{n_1+n_2}{n_1 n_2}}}. \quad \text{Equation 11.30}$$

If the sample sizes for the two groups are equal, this simplifies to:

$$r = \frac{d}{\sqrt{d^2 + 4}}. \quad \text{Equation 11.31}$$

The standard error for  $r$  based on Cohen's  $d$  is computed as:

$$se_r = \sqrt{\frac{\left(\frac{n_1+n_2}{n_1 n_2}\right) se_d^2}{\left[se_d^2 + \left(\frac{n_1+n_2}{n_1 n_2}\right)\right]^3}}, \quad \text{Equation 11.32}$$

which simplifies to the following if the sample sizes for the two groups are equal:

$$se_r = \sqrt{\frac{4se_d^2}{(se_d^2 + 4)^3}}, \quad \text{Equation 11.33}$$

The odds ratio can be converted into a correlation by first converting it into a Cohen's  $d$  and then proceeding as above.

## Meta-analysis of Effect Sizes

---

The most basic meta-analysis estimates the mean effect size across a collection of studies along with associated statistics. The latter include the standard error of the mean effect size, a test of significance for the mean effect size, a confidence interval, and an assessment of heterogeneity, or the variability of effects across studies. More complex meta-analytic analyses, called moderator analyses, explore the relationship between study features and effect size.

A complication with effect size data is that each effect size is typically based on a different sample size and effect sizes from larger studies presumably provide a more precise estimate of the overall mean effect. Intuitively, we might solve this by weighting each study by its associated sample size, and this was a common practice in early applications of meta-analysis, such as in the early 1980s. A better index of the precision of an effect size is its standard error, not its sample size. However, standard errors get smaller as an effect size becomes more precise. As such, we want a weight that is the inverse of the standard error. As established by Hedges and Olkin (1985), the optimal weight is the inverse of the standard error squared. A squared standard error is a variance. As such, this is called inverse-variance weighting, and for Cohen's  $d$  and  $r$  effect sizes, these weights are highly correlated with sample size, as can be seen in Eqs. (11.7) and (11.13). For the odds ratio and the risk ratio, the weight is still strongly correlated with the sample size but is also affected by how close the success or failure probabilities are to 0 or 1. In addition to being optimal, inverse-variance weights provide a statistical foundation for determining the standard error of the mean effect size and other valuable statistics.

There are two main models in meta-analysis: **fixed-effect models** and **random-effects models**. Each makes different assumptions about the true effect that underlies each study. In the fixed-effect model, we assume that each study is estimating a common underlying effect size (common to all studies within the meta-analysis) with the difference between the observed effect size in a study and this true underlying effect size being a function

solely of sampling error. Stated simply, we assume there is one true population effect size value for our collection of studies. In the random-effects model, we assume that there is a distribution of true effects underlying the observed effects. That is, we assume that there is heterogeneity or variability in the true effects being estimated by the studies contributing to a meta-analysis. For example, in a meta-analysis of a treatment program, the true effects may differ across variations in the particular implementation of the treatment program, sample characteristics, and a host of other potential sources, some that may be knowable and others that may be unknowable. If we imagine an infinite number of studies suitable for our meta-analysis, the true effects of those studies would form a distribution. Under a random-effects model, we assume that our collection of studies is a random subset of this distribution of possible studies.<sup>5</sup>

The fixed-effect model is generally too restrictive except in the case of pure replications that are only possible in tightly controlled laboratory settings. The assumptions of the random-effects model are more likely to be consistent with the actual sampling distribution of effect sizes for meta-analyses in criminology and criminal justice. Furthermore, the random-effects model converges on, or becomes, the fixed-effect model in the absence of observed heterogeneity. Thus, even in the case of pure replications, starting with a random-effects model is generally recommended. We will, however, present the computation methods for the fixed-effect model first as they serve as the building blocks for the random-effects model.

### Fixed-Effect Meta-analysis

In the notation below, we will use  $y_i$  to represent each effect size where  $i = 1 \dots k$ , with  $k$  being the number of effect sizes. We are doing so because the methods are the same independent of the type of effect size. That is, the computation methods below are the same for Cohen's  $d$ , the logged odds ratio, logged risk ratio, Fisher's Z-transformed correlation, as well as any other effect size index of interest. We have also selected  $y$  to represent the effect size to reinforce the idea that it is the dependent variable in a meta-analysis.

The analyses below all assume that we are dealing with an *independent* set of effect sizes. Generally, this means one effect size per study. Within this context, a study is not a publication but a unique sample. Thus, we can treat each site of a multisite trial as a unique sample for meta-analysis even though these were part of a single study. We can also relax this requirement to independent subsets of the sample, such as effect sizes for males and

---

<sup>5</sup>For a detailed discussion of these models, see Borenstein et al. (2010).

females, although this should only be done for a moderator analysis comparing these subsets and not for an overall analysis of effects.

Most studies, however, report results on multiple dependent variables, enabling the computation of several effect sizes per study. The first approach to dealing with this is to sort these into distinct constructs and analyze each construct separately. When there remain multiple effect sizes per construct, there are three main options. The first is to select one effect size per construct based on a theoretically derived decision rule. The second is to average the multiple effect size before conducting the meta-analysis. A complication with this approach is that there is uncertainty as to the proper inverse-variance weight of the averaged effect size. Because the weights across these effect sizes are usually very similar, given a common sample size, the typical solution is to average the weights. A third option is to use robust standard errors that account for the dependent or clustered nature of the effect sizes in a fashion similar to multilevel models. This last method will be briefly discussed later in this chapter.

### *The Mean Effect Size and Associated Statistics*

Under a fixed-effect model, the mean effect size is computed as the weighted mean, weighting by the inverse-variance weight. This weight is based on the standard error ( $se$ ) for each effect size and is computed as:

$$w_i = \frac{1}{se_i^2}. \quad \text{Equation 11.34}$$

Using these weights, the mean is computed using Eq. (11.35):

$$\bar{y} = \frac{\sum_{i=1}^k w_i y_i}{\sum_{i=1}^k w_i}, \quad \text{Equation 11.35}$$

where  $y_i$  is the collection of independent effect sizes.

The standard error of the mean effect size is a function of the weights. Recall that these are based on the standard errors for the individual effect sizes. Thus, it should be intuitive that the precision of the mean effect size is a function of the precision of the individual effect sizes contributing to it. The more precise the individual effect sizes, the more precise the overall mean effect size. The standard error of the mean effect size,  $\bar{y}$ , is shown below:

$$se_{\bar{y}} = \sqrt{\frac{1}{\sum_{i=1}^k w_i}}. \quad \text{Equation 11.36}$$

The statistical significance of the mean effect size can be assessed using the standard error to compute a  $z$ -test. The null hypothesis for this test is that the mean equals zero. This is shown in Eq. (11.37):

$$z = \frac{\bar{y}}{se_{\bar{y}}}. \quad \text{Equation 11.37}$$

The  $p$ -value or statistical significance of  $z$  can be determined from a standardized normal distribution. The two-tailed 0.05 critical value is  $\pm 1.96$ . Thus, if  $z$  is greater than  $+1.96$  or less than  $-1.96$ , then the mean effect size is statistically significant at the 0.05 level.

We can also use the standard error to construct a confidence interval. For a 95% confidence interval, we would use a  $z$  critical value ( $z_{CV}$ ) of 1.96. The confidence interval is constructed in the usual way, as shown here:

$$\bar{y}_{\text{lower}} = \bar{y} - z_{CV} se_{\bar{y}}, \quad \text{Equation 11.38}$$

$$\bar{y}_{\text{upper}} = \bar{y} + z_{CV} se_{\bar{y}}. \quad \text{Equation 11.39}$$

### *Homogeneity Testing*

An important aspect of a meta-analysis is the assessment of the heterogeneity of effects across studies. This is done using the homogeneity test. The test examines whether the variability across effect sizes is larger than what would be expected based on sampling error alone. The null hypothesis is that the distribution is homogeneous. The test statistic is  $Q$  and is distributed as a chi-square with degrees of freedom equal to the number of effect sizes minus one ( $k - 1$ ). Equation (11.40) provides the formula for  $Q$ :

$$Q = \sum_{i=1}^k w_i (y_i - \bar{y})^2. \quad \text{Equation 11.40}$$

Notice that this is simply a weighted sum-of-squares. Equation (11.40) defines  $Q$ , but Eq. (11.41) provides a computationally easier method:

$$Q = \sum_{i=1}^k w_i y_i^2 - \frac{\left(\sum_{i=1}^k w_i y_i\right)^2}{\sum_{i=1}^k w_i}.$$
Equation 11.41

A weakness of the  $Q$  statistics is that it is statistically underpowered when the number of studies is small, particularly if the sample sizes for those studies is also small. As such, a nonsignificant  $Q$  should not be interpreted as evidence of homogeneity. Rather, it is merely the absence of sufficient evidence to reject homogeneity as a possibility. If  $Q$  is greater than its degrees of freedom, then there is more variability across effect sizes than would be expected due solely to sampling error, although the amount may not be sufficient to reject the null of homogeneity. To help address this, Higgins and Thompson (2002) developed  $I^2$ . This index provides the percentage of the total variance in effects due to heterogeneity rather than sampling error. The calculation of  $I^2$  is shown in Eq. (11.42),

$$I^2 = \left( \frac{Q - df}{Q} \right) \times 100,$$
Equation 11.42

where  $Q$  is the  $Q$  statistic and  $df$  is the degrees of freedom associated with  $Q$ . Higgins and Thompson provide the following thresholds as a rough guide to interpreting the inconsistency across effect sizes: 0–40% might not be important, 30–60% may represent moderate heterogeneity, 50–90% may represent substantial heterogeneity, and 75–100% may represent considerable heterogeneity. Notice that these ranges overlap, reflecting the rough nature of these guidelines. More generally, the higher  $I^2$ , the greater the proportion of variability across the effect sizes that is likely due to true differences in the underlying population distribution of effects.

### *The Random-Effect Model*

The random-effects model assumes that there are two sources of variability affecting each effect size: study-level sampling error and between-study differences in the true underlying effect, that is, heterogeneity. A significant  $Q$  is clear indication of heterogeneity and some early texts on conducting meta-analysis recommended using this test to determine whether the fit a fixed-effect or random-effects model. This is unwise given that  $Q$  is statistically underpowered when there are a small number of studies and doing so amounts to accepting the null hypothesis of homogeneity. If  $Q$  is larger than its associated degrees of freedom, then there is at least some level of observed heterogeneity, even if it is insufficient to reject homogeneity. The assumptions of the random-effects model are almost always more

plausible. Furthermore, the random-effects model becomes the fixed-effect model if the distribution becomes homogeneous.

In the fixed-effect model, the weights are solely a function of study-level sampling error (i.e., the standard error is a measure of sampling error). The key to a random-effects model is to modify the weights so that they reflect both sampling error and study-level heterogeneity. This is done by estimating  $\tau^2$ , the random-effects variance component, and incorporating it into the weights. Thus, the difference between a fixed-effect and random-effects model is entirely in the definition of the inverse-variance weight. The fixed-effects weight is the inverse of the squared standard error, as shown in Eq. (11.34), whereas the random-effects weight is the inverse of the squared standard error plus study-level variability, as shown here in Eq. (11.43):

$$w_i = \frac{1}{se_i^2 + \tau^2}. \quad \text{Equation 11.43}$$

How do we estimate  $\tau^2$ ? There are several methods. A common method that is easily calculated is based on the method-of-moments and was developed by DerSimonian and Laird (1986). This method has a closed mathematical solution and is based on the relationship between  $Q$  and its associated degrees of freedom. This estimator for  $\tau^2$  is shown in Eq. (11.44):

$$\tau^2 = \frac{Q - df_Q}{\sum_{i=1}^k w_i - \frac{\sum_{i=1}^k w_i^2}{\sum_{i=1}^k w_i}}. \quad \text{Equation 11.44}$$

Note that  $\tau^2$  increases in value as  $Q$  becomes larger relative to its degrees of freedom. If  $Q$  equals its degrees of freedom, then  $\tau^2$  will equal zero, indicating no observed heterogeneity. Because a variance cannot be negative, when Eq. (11.44) returns a value less than zero, it is set to zero, producing weights equivalent to those under a fixed-effect model. The mean effect size, standard error of the mean,  $z$ -test, and confidence intervals are computed as before, using Equations (11.35–11.39), just with these new weights, as defined in Eq. (11.43).

There are several other estimates of  $\tau^2$  (Veroniki et al. 2016). Two that are more widely available in software designed for meta-analysis are full information maximum likelihood and restricted maximum likelihood. These methods are iterative and as such are computationally more demanding. Both are asymptotically unbiased and more efficient than the method-of-moments estimator. However, the full information maximum likelihood estimator tends to be downwardly biased (too small) when there are a

small number of effect sizes. This is partially corrected for in the restricted maximum likelihood estimator. In general, the method-of-moments estimator tends to perform adequately (Veroniki et al. 2016) and is considered a safe choice. In a large meta-analysis, all three will tend to produce fairly similar results. As of this writing, the method-of-moments estimator, while not ideal, is generally considered to be a sensible choice as is the restricted maximum likelihood estimator, particularly for continuous outcomes, such as outcomes that would be appropriate for a Hedges'  $g$  type effect size.

It is not meaningful to perform a separate homogeneity test under a random-effects model. The  $Q$  test performed using the fixed-effects weights is the general test of heterogeneity. This test is also the significance test for  $\tau^2$  estimated using the Dersimonian and Laird method-of-moments estimator. Other estimates of  $\tau^2$  have standard error formulas that can be used for significance testing or constructing a confidence interval around the random-effects variance component.

### **Example: Police-Led Diversion of Youth**

Wilson et al. (2018a) conducted a meta-analysis of police-led diversion programs for youth. As described by the authors, “police diversion schemes are a collection of strategies police can apply as an alternative to court processing of youth. Diversion as an option is popular among law enforcement officers, as it provides an option between ignoring youth engaged in minor wrongdoing and formally charging such youth with a crime. Police-led diversion has the potential to reduce reoffending by limiting the exposure of low-risk youth to potentially harmful effects of engagement with the criminal justice system.” The focus of this review was on the subset of programs for which the diversion occurs before formal charges, either before arrests or before the imposition of formal charges.

The systematic review identified 26 eligible documents representing 19 unique studies. Several of these studies include multiple eligible treatment comparisons, such as studies occurring in different locations or multiple types of diversion programs relative to a control condition. For the meta-analysis, these were treated as independent, producing 31 treatment-comparison contrasts for analysis.

The primary outcome of interest was delinquent behavior. Studies often included both self-reported and official measures of delinquency, although some only reported data from official measures. For the analyses below, only effect sizes based on a dichotomous official measure were used. Because multiple such outcomes were often reported, the authors selected one per study using the following decision rule: selected the effect closest to 12 months, selected arrest over court appearances, and for quasi-experimental studies, select regression adjusted effects over raw effects. For two studies, multiple effect sizes remained. The average of these effects was used in the analysis below.

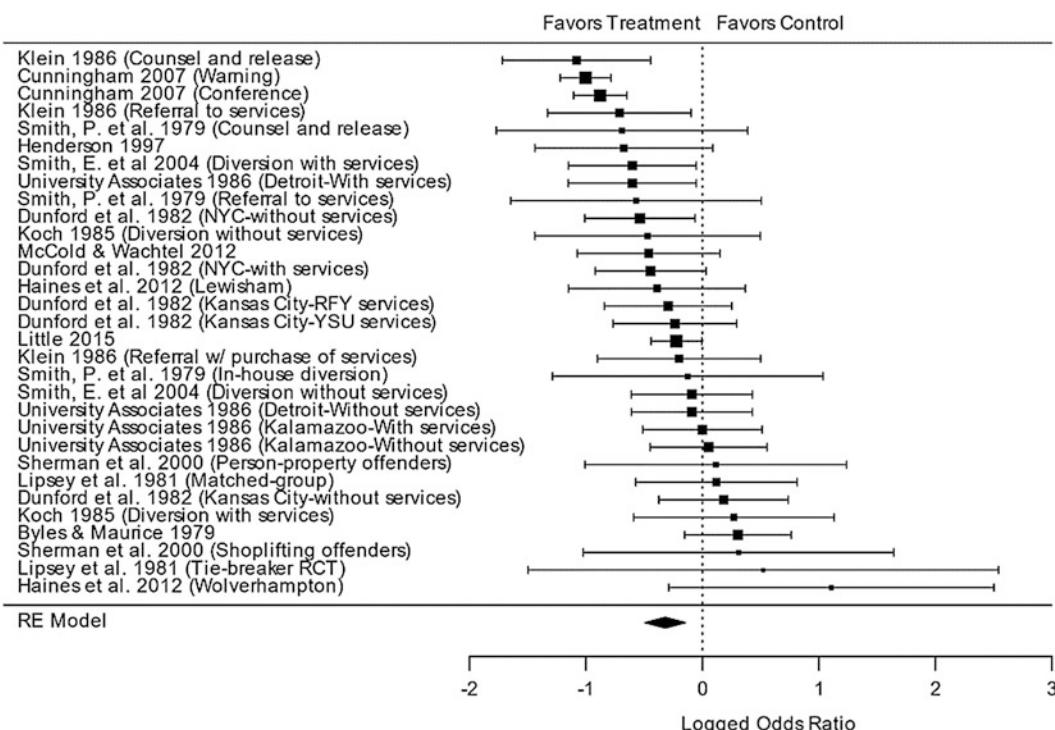
Across the 31 treatment–comparisons, the odds ratios ranged from 0.34 to 3.03. The effects were coded such that values below 1 indicated that the diversion condition had a lower odds of a new delinquent incident relative to the control condition. The authors assumed a random-effects variance model a priori and estimated  $\tau^2$  using the method-of-moments estimator. The analyses were performed on logged odds ratios. The mean effect size was  $-0.32$  with a 95% confidence interval of  $-0.49$  to  $-0.15$ . Converting these values back into odds ratios through exponentiation produces: OR = 0.72 with a 95% CI = 0.61–0.85. Thus, youth diverted from formal processing appear to have close to a 28% lower odds of recidivism relative to controls ( $1 - 0.72$ ). The  $z$ -test for the mean logged odds ratio is  $-4.00$ , which is statistically significant at the 0.05 alpha level. Importantly, the homogeneity test indicates statistically significant heterogeneity:  $Q = 88.92$ ,  $df = 30$ ,  $p < .0001$ . The  $I^2$  is 60% suggesting that 60% of the variability across the logged odds ratios reflects true differences in effects and not simply sampling error. Thus, the effects, while showing benefits of diversion, are inconsistent across studies. Some true effects are larger than others, and some may be null or even slightly harmful. Below, we will present methods of moderator analyses design to explore potential explanations for this variability and then apply those methods to these data.

## Forest Plots

---

A useful graphical method for visualizing the results of a meta-analysis is a **forest plot**. An example using the police-led diversion data is shown in Fig. 11.1. Each row in this figure is a study or in this case a treatment–control contrast. Each square represents the effect size for each row, and the horizontal line is the 95% confidence interval for that effect size. The bottom row shows the mean effect size under the random-effects (RE) model using the method-of-moments estimator. This is represented as a diamond. The width of the diamond indicates the 95% confidence interval. The size of each box indicates the weight associated with each effect size. The bigger the box, the greater the weight. Forest plots may also contain additional information, such as the underlying data on which the effect sizes are based, study sample size, and other relevant information.

There are several things worth noting in this figure. The first is the overall pattern of results that favors the treatment (diversion) condition over the control. We noted at the outset that this was one of the most important contributions of meta-analysis to the science of reviewing studies. It is interesting to note that most of the effects across these studies are not individually statistically significant. Only eight of these 31 effect sizes are statistically significant at the .05 significance level, as can be seen by the

**Figure 11.1***Example forest plot from police-led diversion of youth meta-analysis*

95% confidence intervals that do not overlap with the vertical line at the null value. A vote-counting narrative review of these studies would likely conclude that the results are equivocal. However, the overall pattern and mean effect size show that the evidence supports the conclusion that these programs are effective, on average. Also, several of these studies have inadequate statistical power as evidenced by the very large confidence intervals, particularly the two largest effects favoring the control condition.

## Moderator Analysis

Moderator analysis in the context of meta-analysis is the examination of study features as predictors of effect size variability. That is, do coded characteristics of studies explain heterogeneity in results? There are two main approaches. The first focuses on a single categorical variable and compares the mean effect size across the categories. For example, in a

meta-analysis of intervention studies, we might compare the mean effect sizes for studies that used a random assignment design to those that used a quasi-experimental comparison group design. Alternatively, we might compare the means across different variants of the intervention or different operationalizations of the outcome construct. These analyses are similar to a one-way ANOVA or in the case of the comparison of two means, an independent *t*-test. Thus, this statistical method is called the analog-to-the ANOVA.

The second moderator analysis method relies on regression adapted for meta-analysis and as such is often called *meta-regression*. With this type of moderator analysis, scaled variables can be examined as well as multiple moderator variables in a single analysis, assuming you have a sufficient number of studies. For example, an analysis of this type might explore the relationship between the year the data were collected and effect size or the intensity of the intervention, such as the number of sessions for a group-based program, and effect size. A meta-regression model might also combine two or more moderator variables, such as design type and program type.

### Analog-to-the-ANOVA Moderator Analysis

Conceptually, the analog-to-the-ANOVA method of moderator analysis is straightforward. The effect size data are divided into mutually exclusive and exhaustive subgroups based on a categorical variable of interest. Under a fixed-effect model, the mean effect size and associated statistics are computed as previously. Under a random-effects model,  $\tau^2$  must be re-estimated based on the excess variability that remains after accounting for the variability explained by the moderator variable.

As with a one-way ANOVA, a moderator analysis of a categorical variable partitions variability across effect sizes into the portion explained by the moderator variable and the portion remaining unexplained. This is done by partitioning the  $Q$  statistic into a  $Q_{\text{between}}$  and a  $Q_{\text{within}}$ , corresponding to the explained and unexplained variances. The equation for the latter is as follows:

$$Q_{\text{within}} = \sum w_{ij}(\bar{y}_{ij} - \bar{y}_j)^2, \quad \text{Equation 11.45}$$

where  $\bar{y}_j$  is the mean effect size within each group of the categorical variable. Notice that this is the sum of the  $Q$ s within each group. The easiest way to compute  $Q_{\text{between}}$  is to subtract  $Q_{\text{within}}$  from the overall or total  $Q$ , as shown below in Eq. (11.46):

$$Q_{\text{between}} = Q - Q_{\text{within}}.$$

**Equation 11.46**

Each of these is chi-square distributed with the degrees of freedom for the  $Q_{\text{within}}$  equals the number of effect sizes ( $k$ ) minus the number of categories ( $j$ ), and for the  $Q_{\text{between}}$  equals the number of categories minus 1, or  $j - 1$ . Note that the sum of these two degrees of freedom equals the degrees of freedom for the overall  $Q$ .  $Q_{\text{between}}$  is interpreted in much the same way that the  $F$ -test is in a one-way ANOVA. If it is significant, then we can reject the null that the mean effect sizes across the groups are equal. If there are three or more groups, then the confidence intervals can be used as a guide to identifying meaningful differences. In the case of only two means, this test is comparable to a  $t$ -test between means and indicates that there is a statistically significant difference between them.

The  $Q_{\text{within}}$  provides a test of the residual homogeneity under a fixed-effect model. This indicates whether significant heterogeneity remains after accounting for the moderator variable. Little weight should be placed on this statistic. A nonsignificant  $Q_{\text{within}}$  should not be interpreted as evidence of homogeneity given that, as noted earlier, in most situations it is statistically underpowered. Also, it is strongly recommended to assume a random-effects model a priori, even for a moderator analysis. Under a random-effects model,  $Q_{\text{within}}$  is not meaningful.

The random-effects analog-to-the-ANOVA uses weights with a  $\tau^2$  based on the  $Q_{\text{within}}$  from a fixed-effect analog-to-the-ANOVA. The method-of-moments estimator shown in Eq. (11.44) changes slightly. In the numerator,  $Q$  is replaced with  $Q_{\text{within}}$  and its degrees of freedom. The denominator becomes a bit more complex and involves matrix algebra (Raudenbush 2009). It is best to rely on software implementation for this computation. A moderator analysis of this type fit under a random-effect model is often called a *mixed-effects* model. This reflects that the moderator is treated as a fixed effect, whereas the residual variance is assumed to be random (i.e., there remains unexplained heterogeneity across the effect sizes within each group).

### **Example Analog-to-the-ANOVA Moderator Analysis: Police-Led Diversion of Youth**

The majority of studies included in the meta-analysis of police-led diversion programs for youth used a random assignment to condition design (24 of the 31 treatment–comparison contrasts). To assess whether the results presented earlier showing positive benefits from these programs were biased as a result of including a subset of nonrandomized studies, we conducted a moderator analysis with design type as the categorical variable with two groups: random and nonrandom. This analysis was performed under a random-effects model using the method-of-moments estimator of  $\tau^2$ .

**Table 11.3**

Example analog-to-the ANOVA moderator analysis

GROUP	MEAN	95% CI		z	p	k
	ODDS RATIO	LOWER	UPPER			
Random	0.77	0.64	0.93	-2.73	0.01	24
Nonrandom	0.60	0.44	0.82	-3.24	0.00	7

$$\tau^2 = 0.0988 \text{ estimated via method-of-moments}$$

$$Q_{\text{between}} = 1.789, \text{ df} = 1, p = 0.18105$$

The results of this analysis are shown in Table 11.3. This shows that the mean odds ratio is slightly larger in terms of the effect (further from 1) for the nonrandom studies (odds ratio = 0.60) than for the randomized studies (odds ratio = 0.77). Both means are statistically significant with 95% confidence intervals that exclude the null value of one. The  $Q_{\text{between}}$  equals 1.789 with one degree of freedom. Thus, the difference between these two mean odds ratios is not statistically significant. The mean effect for the nonrandom studies appears to be somewhat larger, but this difference may simply be due to chance.

### Meta-regression Moderator Analysis

A second approach to moderator analysis relies on regression methods and is more flexible than the analog-to-the-ANOVA, allowing for scaled moderator variables or multiple variables. Conceptually, *meta-regression* is simply a weighted regression model, weighting by the inverse variance. Statistically, however, it is a bit more complicated given the meta-analytic nature of the data. The standard errors take into account the precision of the individual effect sizes and under a random-effects model any unexplained heterogeneity. Thus, it is important to use software implementations designed for meta-analytic data and not simply use a weight least squares procedure.

The output from a meta-regression model includes a  $Q$  for the model that is similar to the overall  $F$ -test in a multiple regression model. That is, it indicates whether the model overall accounts for a statistically significant portion of variability across effect sizes. In practice, this typically means that one or more of the regression coefficients is statistically significant. For a regression model that dummy codes a single categorical moderator variable,  $Q_{\text{model}}$  will equal  $Q_{\text{between}}$  from an ANOVA type model.

There are two important considerations for fitting regression models to meta-analytic data. The first is the often highly confounded nature of study characteristics.<sup>6</sup> Study features tend to cluster together; studies with one

---

<sup>6</sup>For of discussion of this issue, see Lipsey and Wilson (2001b).

particular feature in common may have other features in common. For example, in the police-led diversion example, only random assignment studies reported self-reported delinquency measures (these studies also reported official measures used in the above analyses). Thus, design and self-report are confounded. As such, it is critical to understand how moderator variables interrelate and whether there are empty cells for combinations of categorical moderators that may be included in a single model.

The second important consideration is the number of effect sizes relative to the number of independent variables. The sample size requirements are lower than for OLS regression. Because each effect size has a measure of precision associated with it, a model can be successfully fit with only a few cases within the combinations of the values of the moderator variables. However, meta-regression models with a small number of effect sizes suffer from lower statistical power, making it difficult to detect meaningful moderator relationships. Meta-regression is best suited to a large meta-analysis (e.g., 20+ studies) and generally becomes the main analytic focus in very large meta-analyses (e.g., 100+ studies).

### **Example Meta-regression Moderator Analysis: Restorative Justice Programs for Youth**

Wilson et al. (2018b) conducted a meta-analysis of restorative justice programs for adolescent youth, including programs with any component consistent with restorative justice principles. This meta-analysis coded 90 unique treatment–comparison contrasts. One aspect of this meta-

**Table 11.4**

Meta-regression model predicting delinquency effect sizes (Hedges'  $g$ ) based on restorative justice program components

VARIABLE	REGRESSION COEFFICIENT	p
Restorative justice component		
Apology (written/verbal)	-0.11	0.35
Community service/restitution	-0.10	0.43
Follow-up compliance	0.11	0.28
Face-to-face/facilitator/mediator	-0.11	0.45
Family involved/present	-0.01	0.87
Restorative agreement	-0.06	0.64
Victim present	-0.04	0.69
Pre-conference/pre-mediation meeting	0.31	0.01
Community involvement	-0.12	0.22
Other		
Random assignment design	-0.19	0.08
Constant	0.45	0.00

Notes: Models based on 90 treatment–comparison contrasts. Random-effects model estimated via method-of-moments. Model statistics:  $Q_{model} = 15.94$ , df = 9,  $p = .07$ ,  $\tau^2 = 0.087$

analysis was the use of meta-regression methods to identify which program elements, rather than program types, might contribute to higher reductions in future delinquent behavior.

The results from this meta-regression are shown in Table 11.4. Nine different components, each coded yes (1) or no (0), were included in this model. The model also included whether the research design involved random assignment to conditions. The meta-analysis used Hedge's  $g$  effect sizes coded such that positive values indicated better outcomes (less delinquency). The only element with a statistically significant additive effect was the presence of a pre-conference or a pre-mediation meeting. This was present in 25% of the programs included in the meta-analysis. It appears that programs with this element have a larger positive effect than programs without this element, adjusting for the other elements included in this model. However, the  $Q_{\text{model}}$  is not significant, suggesting that this model does a poor job, overall, at explaining heterogeneity in effect sizes. The distribution remains highly heterogeneous.

## **Handling Statistically Dependent Effect Sizes: Robust Standard Errors**

---

A common complication in a meta-analysis, particularly those in the social sciences, is multiple effect sizes related to a common dependent variable or construct of interest. For example, in the police-led diversion meta-analysis discussed earlier, 67 effect sizes of delinquent behavior were coded across the 31 treatment-comparison contrasts. The multiple effect sizes per unique sample are statistically dependent or correlated. Thus, only one effect size per treatment-comparison contrast was selected for the analyses presented above. However, these analyses ignore a potential source of dependence: Some of these treatment-comparison contrasts shared a control group. Hedges et al. (2010) developed a method for analyzing statistically dependent effect sizes, extending the analytic possibilities when confronted with a situation such as the one described here.

An important feature of the Hedges et al. method is that it does not require knowledge of the correlation or more accurately the covariances between statistically dependent effect sizes. It produces robust standard errors with clustered effect sizes and can handle multiple types of clustering, such as in our example with multiple effect sizes for unique samples and for unique treatment conditions compared to a shared control condition. This method can be used to compute an overall mean effect size or to perform regression-type moderator analysis. There are implementations of this method available for R, Stata, SPSS, and SAS.

Returning to the police-led diversion of youth meta-analysis, we can estimate the mean effect size clustering effects within the 19 unique studies. Doing so produces results that are similar to the original results presented earlier but with a slightly larger 95% confidence interval. The mean odds ratio was 0.77 compared to 0.73 for the original analysis. The 95% confidence ranges from 0.63 to 0.96 compared to 0.61–0.86 for the original. The overall conclusion is the same, although the estimated effect is slightly smaller (closer to 1). The results remain statistically significant. The larger confidence interval, however, shows the importance of accounting for any statistical dependency in effect size data. Failing to do so may lead to overly optimistic conclusions.

What about analyzing all 67 effect sizes? Doing so reduces the effect slightly to a mean odds ratio of 0.82, suggesting only an 18% reduction in delinquent behavior for the diversion condition relative to the control condition of routine processing. The result is no longer significant at a conventional level ( $p = 0.056$ ) with a 95% confidence interval that ranges from 0.66 to 1.01. This suggests that the omitted effect sizes were slightly smaller on average than those selected for analysis. The main selection criterion was to use official measures and not self-reported measures of delinquency. Maybe self-reported measures show a smaller effect? We can explore this with a moderator analysis.

In a cluster or hierarchical analysis, moderator variables may vary at the cluster or unique study-level or the effect size level within studies. For the latter, we can examine the effects both between and within studies. This is done by creating a group mean and a group centered version of the moderator variable. The former is the mean of the covariate within each group (i.e., study). The latter is the deviation in the covariate for each observation from its group mean. The between-group covariate assesses whether the studies that reported self-report effect sizes reported smaller or larger effect sizes on average. The within-groups covariate assesses whether self-reported and official measures differ within studies on average. It is generally the within-groups covariate that is of greatest interest as it is unconfounded with other study features.

The results of this analysis are reported in Table 11.5. These analyses were run on the logged odds ratio, and negative values reflect better outcomes for the diverted youth. Neither covariate is statistically significant. However, the between-study (group mean) version of the covariate for self-report is slightly positive (0.0425), indicating that the mean moves toward zero for those studies that included self-reported measures. All of the random assignment studies reported both self-report and official measures of delinquency, whereas only a subset of nonrandom assignment provided self-report measures. This covariate will equal zero for those studies that only reported official measures and a number between zero and one for those that reported both. The latter reflects the proportion of effect sizes

**Table 11.5**

Robust standard errors for clustered effect size meta-regression: police-led diversion for youth

VARIABLE	REGRESSION COEFFICIENT	<i>p</i>
Intercept	-0.2074	0.105
Self-report (between study)	0.0425	0.883
Self-report (within study)	0.1445	0.525

Notes: Models based on 67 effect sizes across unique 19 studies. Analyses performed on logged odds ratios  
 $\tau^2 = 0.1632$

within that study based on self-reported data. Within those studies reporting both, the effect size for self-reported measures was closer to zero, on average, compared to the official measures. These effects are not statistically significant and as such should be viewed cautiously. However, they do explain why the model with all 67 effect sizes that included the self-reported measures produced a slightly smaller mean odds ratio relative to the mean odds ratio based solely on official measures.

## Publication Selection Bias

An important threat to the validity of any meta-analysis (or any other form of reviewing) is **publication selection bias**. This is the tendency of studies that are statistically significant to have a higher likelihood of being published in a peer-reviewed journal. This is a well-established phenomenon and starts with an author's enthusiasm, or lack thereof, for writing up results (i.e., some studies never get written up), to peer reviewers being more critical of studies with nonsignificant results, to editors tending to favor studies with significant results. The particular outcomes and analyses reported within a study are also often those that showed statistical significance. The cumulative effect of this is that the literature identified for a meta-analysis might be biased toward larger, more statistically significant, effects.

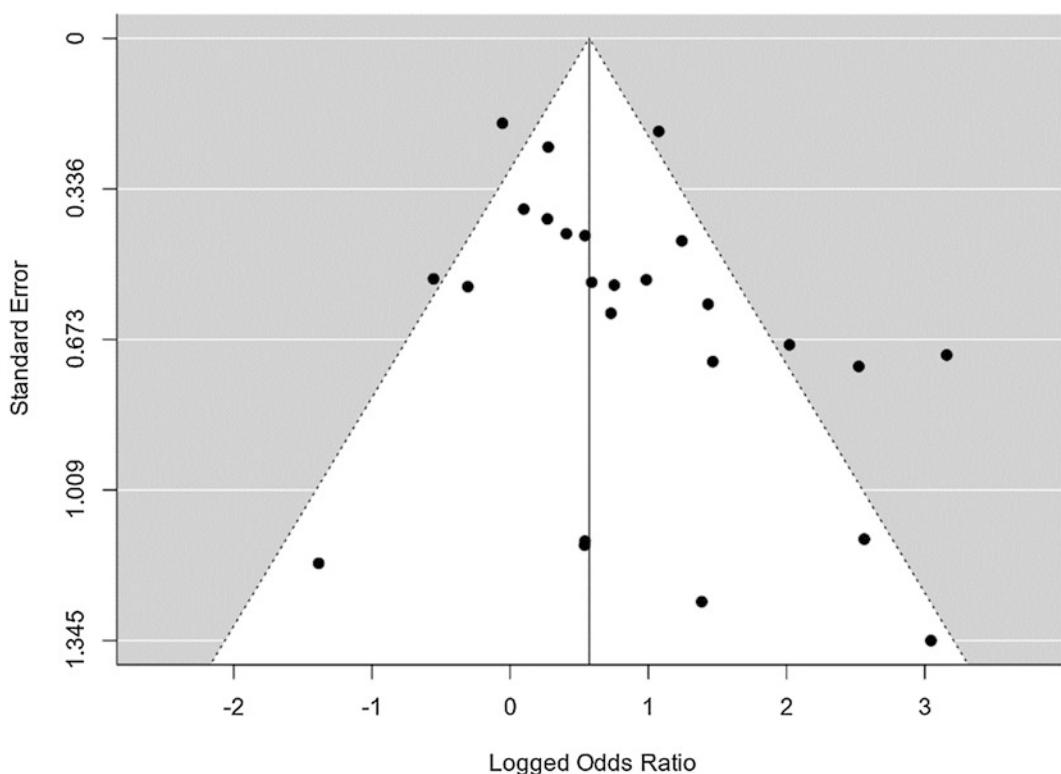
The first line of defense against publication selection bias is searching for and included eligible gray literature studies, such as theses and dissertations, technical reports, conference presentations, and other less formally published works. Additionally, several methods have been developed to look for evidence of publication selection bias in effect size data. The simplest approach is to compare the mean effect size for published studies versus unpublished (or less formally published) studies. This of course presumes that at least some gray literature studies were identified and eligible.

Other methods include a visual inspection of a funnel plot, Duvall and Tweedie's trim-and-fill, and Egger's test. All three of these are based on the increased likelihood of results being censored as the sampling size decreases. The smaller the study, the less likely it is to find a statistically significant result, all other things being equal, and as such less likely to be published. Furthermore, it has been found that larger studies are more likely to be published even if their results are not statistically significant.

The **funnel plot** illustrates this relationship between sample size, or more precisely, the inverse of the standard error, and the likelihood of an effect being censored. In the absence of publication selection bias, we would expect the effect to be symmetrical around the mean effect size but more spread-out for smaller studies and more compact around the mean for larger studies. Thus, if we plot the standard error on the  $y$ -axis flipped (smaller standard errors at the top) and the effect size in the  $x$ -axis, then we would expect the plot to look like an upside-down funnel, narrow at the top and wide at the bottom. In the presence of publication selection

**Figure 11.2**

*Example funnel plot*



bias, we would expect a pattern of missingness in the lower part of the plot for effects that would not have been statistically significant. If the mean effect size is positive, this would create asymmetry with missing effects to the left of the mean in the lower part of the plot.

An example of this is shown for a sample of 31 effect sizes in Fig. 11.2. We can see clear asymmetry in this plot with more effects in the middle and lower sections of the plot to the right of the horizontal line than to the left. The dashed vertical lines are the approximate cut-points for where effects would be statistically significant. This plot shows evidence of “missing” nonsignificant effects. Two problems with this method are first that it depends on a visual assessment, rather than a statistical test, and second, that asymmetry may arise for other reasons, such as meaningful differences between small and large studies. It is possible that smaller studies genuinely have large true effects on average.

A statistical test related to the funnel plot is the Duval and Tweedie (2000) trim-and-fill method. This is a nonparametric method that initially trims effect sizes from the plot until it becomes symmetrical and then returns these effects along with filled effects that are equal distance to the left of the horizontal line. A new mean effect size is then computed that includes these filled effect sizes. This provides a rough gauge of the potential seriousness of the publication selection bias.

For the data shown in Fig. 11.2, the trim-and-fill method suggests that there are six missing studies to the left of the mean producing an adjusted mean effect size of 0.44 compared to 0.57 on the original data. This suggests that these data are meaningfully upwardly biased due to publication selection but that the overall effect is still positive.

Another statistical approach is the Egger’s test (Egger et al. 1997). This test is also based on funnel plot asymmetry. However, it does this using regression to examine if effect sizes are related to their standard errors. The dependent variable is the effect size divided by its standard error, and the independent variable is the inverse of the standard error. A significant Egger’s test indicates evidence of publication selection bias. The test is underpowered when the number of effect sizes is small. For the data shown in Fig. 11.2, Egger’s test produces a  $t = 2.047$  with 23 degrees of freedom. This has a  $p$  value of 0.05. This provides additional evidence of publication selection bias.

With highly heterogeneous data, the above methods may fail to detect publication select bias. This should *not* be interpreted as evidence that publication bias does not exist in a particular meta-analysis. Publication selection bias is almost always present to some degree. The focus should be on trying to establish the plausible seriousness of this threat to the credibility of the results from a meta-analysis.

## Chapter Summary

---

Meta-analysis is a statistical method of synthesizing results across studies examining a common research question, which were obtained using **systematic review** methods. It accomplishes this by encoding study findings as an **effect size** index reflecting the statistical parameter of interest. Common effect sizes include the standardized mean difference, odds ratio, risk ratio, and correlation coefficient. Several less common effect size indices are available for studies with designs and measures not suitable for these commonly used ones.

To be credible, a meta-analysis needs to approach the tasks of searching for eligible studies, assessing for study eligibility, coding and extraction of study features and effect sizes, and the analysis of effect size data in a systematic, transparent, and replicable manner. This process should also explicitly search for and include eligible unpublished studies, such as technical reports and unpublished manuscripts, as protection against **publication selection bias**.

Meta-analysis methods focus on estimating the mean effect size, along with an associated confidence interval, and an assessment of the consistency or inconsistency in the effects across studies. Effect sizes and their confidence intervals can be visualized using a **forest plot**, while a **funnel plot** can be used to depict the relationship between effect size and its standard error for assessing publication selection bias. Variability in effect sizes can also be explored through **moderator analyses** that either partition the studies into subgroups or examines the relationship between study features and effect size.

There are two main types of meta-analysis models: **fixed-effect models** and **random-effects models**. The former assumes that the collection of studies being meta-analyzed shares a single common underlying true effect size. The random-effects model assumes that there is a distribution of true effects, and the collection of studies being meta-analyzed represents a sample from that distribution. That is, the random-effects model assumes that the true effects differ across studies. The assumptions of the random-effects model are almost always more plausible than those of the fixed-effect model except in very limited situations. Thus, the random-effects model should generally be used. An implication of this is larger standard errors and confidence intervals. Random-effects models are more conservative because they are accounting for both within-study sampling error and between-study variability in estimating the model.

Traditional reviewing methods focus on statistical significance. Meta-analysis focuses on the direction and magnitude of the results in the studies and provides a more credible examination of the pattern of evidence. It is often also useful at identifying areas where there is a need for additional research and can highlight common methodological weaknesses across studies.

## Key Terms

---

**Effect size** A statistical index that encodes the findings of a study in a way that allows for comparison across studies via meta-analysis. Common effect sizes include the standardized mean difference, the correlation coefficient, the odds ratio, and the risk ratio.

**Fixed-effect model** A fixed-effect meta-analysis model assumes that each study contributing to a meta-analysis shares a common underlying true population effect.

**Forest plot** A forest plot is a graphic display of effect sizes with their confidence intervals, along with the overall results of a meta-analysis.

**Funnel plot** A type of scatterplot used to assess for publication selection bias. This plot shows the relationship between an effect size and its standard error. Asymmetry in this plot is evidence of publication selection bias.

**Moderator analysis** Moderator analysis is the examination of the relationship between study features and effect sizes. Two main types are the analog-to-the-ANOVA, which is

similar to a one-way ANOVA model, and meta-regression.

**Publication selection bias** Publication selection bias is the tendency for studies with statistically significant results to have a greater likelihood of being published and hence included in a meta-analysis, potentially upwardly biasing the results.

**Random-effects model** A random-effects model in meta-analysis assumes that there is variability in the true population effects being estimated by a collection of studies and incorporates this heterogeneity into the model.

**Systematic review** A set of systematic, documented, and replicable methods for reviewing the literature (published and unpublished), on a topic. The methods include: (1) a systematic search for all eligible studies; (2) detailed eligibility criteria; (3) systematic coding of study features; (4) an assessment of risk-of-bias; and (5) a credible method synthesizing findings across studies, such as through meta-analysis.

## Symbols and Formulas

---

*d* Cohen's standardize mean difference effect size

*g* Hedges' small-sample size bias adjusted standardized mean difference effect size

*k* Number of effect sizes in an analysis

*I*<sup>2</sup> Higgins' heterogeneity index

OR Odds ratio effect size

$Q$  Homogeneity statistic

RR Risk ratio effect size

$\tau^2$  Random effects variance estimate

$Z_r$  Fisher's transformed  $r$

Cohen's  $d$  standardized mean difference effect size:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}}$$

The pooled within-groups standard deviation:

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Hedges' small sample size bias correction factor:

$$J = 1 - \left[ \frac{3}{4(n_1 + n_2) - 9} \right]$$

Hedges'  $g$  standardized mean difference effect size:

$$g = J \times d$$

Standard error for Hedge's  $g$ :

$$se_g = \sqrt{\frac{n_1 + n_2}{n_1 n_2} + \frac{g^2}{2(n_1 + n_2)}}$$

Risk ratio effect size:

$$\text{RR} = \frac{a/(a+b)}{c/(c+d)} = \frac{p_1}{p_2}$$

Standard error of the logged risk ratio:

$$se_{\ln(\text{RR})} = \sqrt{\frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}}$$

Odds ratio effect size:

$$\text{OR} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1(1-p_2)}{p_2(1-p_1)} = \frac{ad}{bc}$$

Standard error of the logged odds ratio:

$$se_{\ln(\text{OR})} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Fisher's Z-transformation for the correlation coefficient:

$$Z_r = \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right)$$

Standard error for Fisher's Z-transformed correlation coefficient:

$$se_{Z_r} = \frac{1}{\sqrt{n-3}}$$

Fixed-effect model inverse-variance weight:

$$w_i = \frac{1}{se_i^2}$$

Mean effect size:

$$\bar{y} = \frac{\sum_{i=1}^k w_i y_i}{\sum_{i=1}^k w_i}$$

Standard error of the mean effect size:

$$se_{\bar{y}} = \sqrt{\frac{1}{\sum_{i=1}^k w_i}}$$

*z*-Test for the mean effect size:

$$z = \frac{\bar{y}}{se_{\bar{y}}}$$

*Q*-test of homogeneity:

$$Q = \sum_{i=1}^k w_i y_i^2 - \frac{\left( \sum_{i=1}^k w_i y_i \right)^2}{\sum_{i=1}^k w_i}$$

*I*<sup>2</sup> index of heterogeneity:

$$I^2 = \left( \frac{Q - df_Q}{Q} \right) \times 100$$

Random-effects inverse-variance weight:

$$w_i = \frac{1}{se_i^2 + \tau^2}$$

Dersimonian and Laird method-of-moments estimator of the random-effect variance component ( $\tau^2$ ):

$$\tau^2 = \frac{Q - df_Q}{\sum_{i=1}^k w_i - \frac{\sum_{i=1}^k w_i^2}{\sum_{i=1}^k w_i}}$$

$Q_{\text{within}}$  groups for the analog-to-the-ANOVA:

$$Q_{\text{within}} = \sum w_{ij} (y_{ij} - \bar{y}_j)^2$$

$Q_{\text{between}}$  groups for the analog-to-the-ANOVA:

$$Q_{\text{between}} = Q - Q_{\text{within}}$$

## Exercises

---

- 11.1. A study reports the mean self-esteem score on a standardized scale (higher values indicating more self-esteem) for the treatment group as 127.8 ( $n = 25$ ) and for the comparison group as 132.3 ( $n = 30$ ). The standard deviations are 10.4 and 9.8, respectively.
  - (a) Calculate Hedges'  $g$ .
  - (b) Calculate the standard error for this Hedges'  $g$ .
- 11.2. A study reports a  $t$ -value of 1.68 favoring the treatment group. The treatment group has 10 respondents and the comparison group has 12.
  - (a) Calculate Hedges'  $g$ .
  - (b) Calculate the standard error for this Hedges'  $g$ .
- 11.3. A study of adult psychotherapy reports that 32% and 37% of clients in the treatment and control conditions, respectively, were rated as *improved* post-treatment. There were 42 patients in the treatment group and 29 in the control group.
  - (a) Use the Cox logged odds ratio method to calculate Hedges'  $g$ .
  - (b) Calculate the standard error for this Hedges'  $g$ .
  - (c) Calculate the odds ratio for these data.
  - (d) Calculate the standard error of the logged odds ratio for these data.
  - (e) Calculate the risk ratio for these data.
  - (f) Calculate the standard error for the logged risk ratio for these data.

- 11.4. You have the following six Hedges'  $g$  effect sizes and associated standard errors:

<b><i>g</i></b>	<b><i>se</i></b>
-0.23	0.32
0.25	0.31
0.08	0.11
0.10	0.26
0.20	0.17
0.22	0.24

- (a) Calculate the fixed-effect mean effect size.
- (b) Calculate the standard error of the mean effect size.
- (c) Construct a 95% confidence interval around the mean effect size.
- (d) Compute  $Q$  and  $I^2$  and interpret.

## Computer Exercises

### SPSS

An author of this chapter, David Wilson, has created macros that enable SPSS to perform meta-analysis, but only through syntax and not through the point-and-click menu system. The macros are available at <http://mason.gmu.edu/~dwilsonb> in the file called *spss\_macros.zip*. There is an *spss\_readme.txt* file in that zip file that explains how to install and use these macros. These instructions will be summarized here.

To use a macro, it must be read into SPSS. There is no need to edit or modify the macros themselves. Reading in a macro into SPSS is done using the “INCLUDE” command, and this only needs to be run once per SPSS session (i.e., if you close SPSS and restart it, you need to rerun the INCLUDE statement to enable the macro). Note that only two of the three macros (MEANES.SPS, METAF.SPS, and METAREG.SPS) can be used in any one session. Trying to use all three in a single session will fail.

Here is an example of the syntax for initializing the MEANES.SPS macro:

```
INCLUDE "[::DRIVE::]::[::PATH::]MEANES.SPS" .
```

The [:DRIVE:] and [:PATH:] specify the location on your hard drive where the macros have placed (where you extracted them to when you unzipped the zip file; note that you can place these macros in any location that suits you). For example, if you placed the macros in: "C:\Users\David\Documents\Macros", then the INCLUDE command would be:

```
INCLUDE "C:\Users\David\Documents\Macros\MEANES.SPS" .
```

METAF.SPS and METAREG.SPS are loaded into SPSS in the same fashion. Only load what you need for a particular session.

Once initialized, a macro can be used like any other SPSS procedure. Unfortunately, these macros are less stable than built-in features and return a long list of error messages if you make an syntax mistake, such as mistype a variable name.

The MEANES.SPS macro computes the mean effect size and associated statistics under both a fixed-effect model and random-effects model. You must specify the name of the effect size variable with “ES = variable name” and the name of the inverse-variance weight variable with “W = variable name”. For example, if you have named your effect size “HedgesG” and the weight variables as “GWeight”, the syntax would read:

```
MEANES ES = HedgesG / W = GWeight.
```

The METAF.SPS macro performs the analog-to-the-ANOVA analysis. The syntax is similar to the above, but you must now also specify the categorical variable by adding “GROUP = variable name”. For this command, must also specify the model type, with fixed effect being the default if unspecified. The options are “MODEL = MM” for method-of-moments, “MODEL = ML” for full information maximum likelihood, and “MODEL = REML” for restricted maximum likelihood. The examples below use the variable “TXTYPE” as the categorical moderator.

```
METAF ES = HedgesG / W = GWeight / GROUP = TXTYPE .
METAF ES = HedgesG / W = GWeight / GROUP = TXTYPE /
    MODEL = MM .
METAF ES = HedgesG / W = GWeight / GROUP = TXTYPE /
    MODEL = ML .
METAF ES = HedgesG / W = GWeight / GROUP = TXTYPE
    / MODEL = REML .
```

The METAREG.SPS macro works in a similar fashion, changing “GROUP” to “IVS” and allowing for one or more independent variables. Examples are shown below.

```
METAREG ES = HedgesG / W = GWeight / GROUP = RANDOM
    TX1 TX2 .
METAREG ES = HedgesG / W = GWeight / GROUP = RANDOM
    TX1 TX2
    / MODEL = MM .
METAREG ES = HedgesG / W = GWeight / GROUP = RANDOM
    TX1 TX2
    / MODEL = ML .
METAREG ES = HedgesG / W = GWeight / GROUP = RANDOM
    TX1 TX2
    / MODEL = REML .
```

These macros also have some print options useful for when the effect size is a Z-transformed correlation or a logged odds ratio or logged risk ratio. The "/ PRINT IVZR" option will convert *Zr* results into correlations. The "/ PRINT EXP" will exponentiate the results, converting logged odds ratio and logged risk ratios into odds ratio and risk ratios.

### Stata

For Stata, you have two options. As of version 16, Stata ships with built-in meta-analysis commands. You can also use the commands created by David Wilson and available on his website at the link above. The latter work in a fashion similar to the SPSS macros. For example, assuming that we have named our effect size **g** and our weight **wg**, the overall mean effect size and associated statistics are generated with the following command.

```
means g [w=wg]
```

Assuming our categorical moderator variable named **txtype**, then an analog-to-the-ANOVA analysis is performed with the following commands, depending on model type (fe = fixed effect, mm = random effects using method-of-moments, ml = random effects using full information maximum likelihood):

```
metaf g txtype [w=wgl], model(fe)
metaf g txtype [w=wgl], model(mm)
metaf g txtype [w=wgl], model(ml)
metaf g txtype [w=wgl], model(ml) print(ivzr)
metaf g txtype [w=wgl], model(mm) print(exp)
```

The macro for meta-regression, **metareg**, is syntactically similar to the **metaf** command, accommodating multiple independent variables, as shown below.

```
metareg g random tx1 tx2 [w=wgl], model(fe)
metareg g random tx1 tx2 [w=wgl], model(mm)
metareg g random tx1 tx2 [w=wgl], model(ml)
```

The new as of version 16 Stata commands can compute the effect sizes as part of the analysis, assuming all studies provides uniform data (e.g., means, standard deviations, and sample sizes for a Hedges' *g* meta-analysis) or allows the user to provide already computed effect sizes and standard errors. In the case of the latter, you must first declare the meta-analysis data, specifying the variable name for the effect size and standard error (or confidence intervals if you prefer). For example, if the effect size is **g** and the standard error is **gse**, then you would declare this as shown below.

```
meta set g gse
```

You can then generate overall summary statistics with the **meta summarize** command. The “meta summarize” command can also be used to perform an analog-to-the-ANOVA model by specifying subgroups. Examples are shown below (reml = restricted maximum likelihood, mle = full information maximum likelihood, dlaird = method-of-moments).

```
meta summarize
meta summarize, subgroup(txtype)
meta summarize, subgroup(txtype) random(reml)
meta summarize, subgroup(txtype) random(mle)
meta summarize, subgroup(txtype) random(dlaird)
```

To perform a meta-regression, you can use the **meta regress** command, as shown below.

```
meta regress random tx1 tx2
meta regress random tx1 tx2, random(reml)
meta regress random tx1 tx2, random(mle)
meta regress random tx1 tx2, random(dlaird)
```

A forest plot can be produced with the command:

```
meta forestplot
```

To run a trim-and-fill analysis or generated a funnel plot, install the following two add-on programs:

```
ssc install metatrim
ssc install metafunnel
```

Examples of using each are below:

```
metatrim g gse
metafunnel g gse
```

## R

There are several packages for performing meta-analysis using R. We will focus on using **metafor**. This is installed and loaded in the usual fashion.

```
install.packages("metafor")
library("metafor")
```

This package has extensive features. We will only discuss the basics for performing the analyses discussed in this chapter. The mean effect size and associated statistics is generated with the **rma()** function as shown below.

```
rma(effect_size, variance_of_effect_size,
method="FE", data=data_frame)
```

If our data frame is called `ex1` and our effect size and associated variance (standard error squared) are `g` and `v`, this command would read:

```
rma(g, v, method="FE", data=ex1)
rma(g, v, method="REML", data=ex1)
```

There are numerous method options, a few of which are as follows: FE = fixed effect, DL = method-of-moments, ML = full information maximum likelihood, REML = restricted maximum likelihood).

Extending this to a meta-regression simply involves adding the regression formula to the above command. For example, if `random`, `tx1`, and `tx2` are our independent variables, then the command would read as follows:

```
rma(g ~ random + tx1 + tx2, v, method="REML", data=ex1)
```

The `metafor` package does not explicitly perform an analog-to-the-ANOVA type analysis. However, this is just meta-regression model with different output, that is, the means for each category of the moderator variable. You can get these means using the `predict()` function, as shown below.

```
model1 <- rma(g ~ random, v, method="REML", data=ex1)
print(model1)
predict(model1, newmods=c(0,1))
```

This assumes that `random` is either a binary 0/1 variable or a binary factor. If “`random`” was a three level factor, then the `predict` command would read as follows:

```
predict(model1, newmods=matrix(c(0,0,1,0,0,1),
nrow=3, byrow=TRUE))
```

This can be adapted to additional numbers of levels. Note that it is important that the independent variable is specified as an unordered factor.

A forest plot can be generated with the `forestplot()` function. To get labels for each row, an label must be added to the `rma()` function, as shown below, using the `slab=` option.

```
model1 <- rma(g, v, method="REML", data=ex1,
               slab=AuthorLabel)
forestplot(model1)
```

See the documentation for lots options and how to tweak the aesthetics of the plot.

You can generate a trim-and-fill analysis and a funnel plot based on a meta-analysis object (i.e., a model associated with the “`rma()`” command) as follows:

```
trimfill(model1)
funnel(model1)
```

### Problems

For these problems, use the dataset *cbt*. This data file contains studies that evaluated the effectiveness in reducing recidivism of a group-based cognitive-behavioral program for adult offenders. Perform the following analyses. The effect size variable is *g*, and the variance is *v*.

1. Generate the mean effect size and associated statistics. Interpret the results.
2. Perform a trim-and-fill analysis, and generate a funnel plot, if using Stata or R. Interpret the results.
3. Perform an analog-to-the-ANOVA type analysis with *random* as the moderator. This variable indicates whether the study used random assignment to conditions or not.
4. Perform an analog-to-the-ANOVA type analysis with *txtype* as the moderator. This variable has three categories: moral reconation therapy, reasoning and rehabilitation program, and other cognitive-behavioral program.
5. Repeat Problem 3 using meta-regression. In Stata and SPSS, you will need to generate a new variable that is numeric and coded 0 and 1. Interpret the results. How do these results compare with those from Problem 3?

### References

---

- Baker, R. D., & Jackson, D. (2013). Meta-analysis inside and outside particle physics: Two traditions that should converge? *Research Synthesis Methods*, 4(2), 109–124.
- Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, 1(2), 97–111.
- Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. Hoboken, NJ: John Wiley & Sons.
- DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3), 177–188.
- Duval, S., & Tweedie, R. (2000). A nonparametric “trim and fill” method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, 95 (449), 89–98.
- Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109), 629–634.
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5, 3–8.
- Glass, G. V., & Smith, M. L. (1979). Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy Analysis*, 1(1), 2–16.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic.
- Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1(1), 39–65.

- Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11), 1539–1558.
- Hunt, M. (1997). *How science takes stock: The story of meta-analysis*. New York, NY: Russell Sage Foundation.
- Kim, R. S. (2011). *Standardized regression coefficients as indices of effect sizes in meta-analysis*. Tallahassee, FL: Doctoral dissertation, Florida State University.
- Kochel, T. R., Wilson, D. B., & Mastrofski, S. D. (2011). Effect of suspect race on officers' arrest decision. *Criminology*, 49(2), 473–512.
- Lipsey, M. W., & Wilson, D. B. (2001a). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
- Lipsey, M. W., & Wilson, D. B. (2001b). The way in which intervention studies have "personality" and why it is important to meta-analysis. *Evaluation & The Health Professions*, 24(3), 236–254.
- Needleman, H. L., McFarland, C., Ness, R. B., Fienberg, S. E., & Tobin, M. J. (2002). Bone lead levels in adjudicated delinquents: A case control study. *Neurotoxicology and Teratology*, 24(6), 711–717.
- O'Rourke, K. (2007). A historical perspective on meta-analysis: dealing quantitatively with varying study results. *Journal of the Royal Society of Medicine*, 100(12), 579–582.
- Pratt, T. C., Cullen, F. T., Blevins, K. R., Daigle, L. E., & Madensen, T. D. (2006). The empirical status of deterrence theory: A meta-analysis. In F. T. Cullen, J. P. Wright, & K. R. Blevins (Eds.), *Taking stock: The status of criminological theory—Advances in criminological theory* (pp. 367–395). New Brunswick, NJ: Transaction.
- Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. The handbook of research synthesis and meta-analysis. In H. M. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *Handbook of research synthesis and meta-analysis* (2nd ed., pp. 295–316). New York, NY: Russell Sage.
- Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences*, 1(3), 377–386.
- Sánchez-Meca, J., Marín-Martínez, F., & Chacón-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods*, 8(4), 448.
- Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, 62(5), 529.
- Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 32(9), 752.
- Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., et al. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, 7(1), 55–79.
- Wilson, D. B., Brennan, I., & Olaghore, A. (2018a). Police-initiated diversion for youth to prevent future delinquent behavior: A systematic review. *Campbell Systematic Reviews*, 14(1), 1–88.
- Wilson, D. B., Olaghore, A., & Kimbrell, C. S. (2018b). *Effectiveness of restorative justice principles in juvenile justice: A meta-analysis (Technical report)*. NCJ 250872.

## Spatial Regression

### **M o d e l i n g   S p a t i a l   D a t a**

---

What is Spatial Autocorrelation?

Why Can't We Use Ordinary Least Squares Regression for Spatial Data?

What is Spatial Regression?

What Types of Spatial Regression Are There?

How Do We Determine the Appropriate Type of Spatial Regression to Use?

### **S p a t i a l   L a g   M o d e l**

---

What is a Spatial Lag Regression Model?

When Do We Use Spatial Lag Regression Models?

How Do You Run a Spatial Lag Regression Model?

How Are Spatial Lag Regression Coefficients Interpreted?

### **S p a t i a l   E r r o r   M o d e l**

---

What is a Spatial Error Regression Model?

When Do We Use Spatial Error Regression Models?

How Do You Run a Spatial Error Regression Model?

How Are Spatial Error Regression Coefficients Interpreted?

**C**RIMINOLOGISTS OFTEN RELY ON SPATIALLY REFERENCED data (e.g., the street address or geographic coordinates of where a stop-question-frisk was conducted) or focus on a unit of analysis that is organized on a spatial landscape (e.g., Census blocks, counties, states). Spatially referred data are likely to be spatially autocorrelated, meaning that a given observation is similar to nearby observations, with **spatial autocorrelation** referring to the strength of association between cases across space (*how* similar observations are to their neighbors). Crime data, when considered across space, consistently exhibit very strong spatial autocorrelation. For instance, the highest crime police beats in a city are most likely going to be surrounded by beats with a similarly high level of crime.

While the term spatial autocorrelation describes the strength of association between cases across space, it does not describe the process by which these dependencies arise. There are two types of spatial processes that cause spatially autocorrelated data: **spatial dependence** (spatial interaction) and **spatial heterogeneity** (spatial structure). In the literature, you may find the terms spatial dependence and spatial heterogeneity used interchangeably. Albeit both are types of spatial processes and both can lead to violating assumptions in regression analysis but—do not be mistaken—they are not the same. The ability to differentiate these two can be difficult but it is important because, as you will learn below, the type of spatial process at play in your data often informs the type of spatial regression model to be used.

Spatial dependence causes clusters of events across a landscape that arises due to an active process/interaction among observations that are geographically near (e.g., the spatial spread of a virus across a city due to the transmission of the virus through interacting with an infected person). In other words, the intensity of cases at a given location affects the intensity of cases at neighboring locations. Spatial heterogeneity, in contrast, causes a patchy distribution of events across a landscape that arises due to phenomena that vary across space. For example, we may naturally see increases in violent crime rates for Southern states in the USA during

summer months because of their shared warm climate. This topic will be discussed in detail below, but it is important that you understand that if your data have significant spatial autocorrelation, whether arising from spatial dependence or spatial heterogeneity, the issue of space when conducting analysis should not be ignored.

## Why Can't We Use OLS Regression with Spatial Data?

---

When analyzing spatially autocorrelated data, we must bear in mind one of the most fundamental assumptions of spatial theory—Tobler's (1970) first law of geography. This law provides that, "[e]verything is related to everything else, but near things are more related than distant things." That being the case, traditional regression analysis for tackling research questions is no longer appropriate with spatially autocorrelated data because they violate key assumptions underlying most statistical tools. Ignoring such dependencies may lead us to draw conclusions that are biased or even flat-out wrong. This relationship between cases violates the assumption in OLS regression that observations are independent. Relatedly, if the regression model does not capture the spatial relationship between observations, there is a tendency for residuals to be spatially autocorrelated. For example, imagine a criminologist is exploring socioeconomic predictors of crime rates at the census tract level using OLS regression. Then, imagine that she maps the unexplained residual (difference between the observed and predicted crime rate). She will not want to see any systematic pattern in the residuals across space. However, if she finds census tracts tend to have residuals similar to their neighbors (residuals being spatially autocorrelated), then the residuals of her regression model are no longer random and, therefore, violate the assumption in OLS regression. Additionally, spatial data are more prone to heteroscedasticity due to varying sizes of the spatial units so she would need to assess whether the error term is homoscedastic.

And be aware that regression models that take into account nesting of groups defined by geography, such as multilevel models (MLMs) or regression that uses a clustered sandwich estimator, do not address this issue. Just like OLS, nested or clustered models are *aspatial* models; the location of the cases and the distance between them are not considered. The problem lies in the issue of space because if you were to rearrange the nested units in space, your result would not change. As you will learn below, in models that incorporate spatial effects, the dependence is dealt with by weighting observations by their neighbors. Therefore, a given case may serve as a weight to multiple different cases in the sample. This is problematic with MLM because groups of observations must be distinct and not be placed into more than one group.

The consequence of violating the assumption of independence by ignoring the spatial aspect of data in OLS regression is that it increases our chances of making a Type I error. Put differently, we are putting ourselves at greater risk for *mistakenly* concluding that a relationship exists, meaning that our model is less precise. Spatially correlated data can lead to confidence intervals that are too small (the confidence interval gets smaller as the level of spatial autocorrelation in the data increases). The reason for this is that if the regression model ignores the spatial autocorrelation, the estimated variance will not be accurate regardless of the sample size, where the true variance will be larger in the presence of positive spatial autocorrelation (Cressie 1993). This is illustrated in an example on estimating gross domestic product per capita (GDP) and level of democracy for several countries provided by Ward and Gleditsch (2008), whereby  $\rho$  represents the level of spatial autocorrelation observed in GPD across countries:

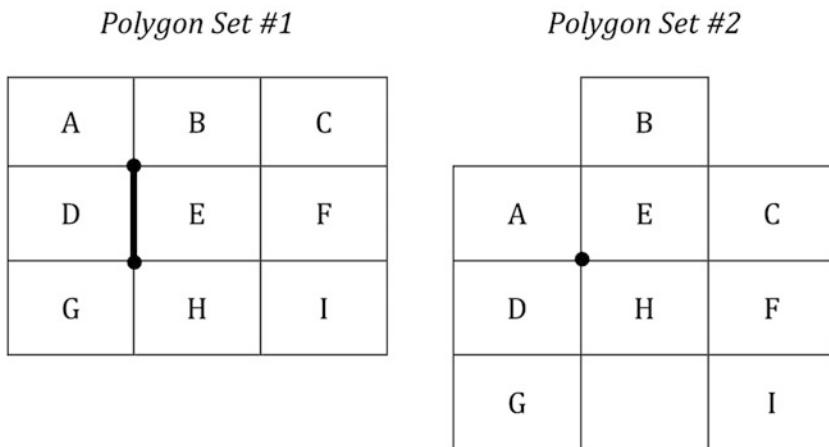
[F]or a sample of 158 observations on GDP, the 95% confidence band under an assumption of normality would be  $(1.96 \times \sigma)/\sqrt{n}$ , but if there were a spatial correlation of 0.65—the actual value of  $\hat{\rho}$  for GDP from [their presented example]—the correct confidence interval would be approximately 4.22 instead of 1.96, over twice as large. In the case of the level of democracy,  $\hat{\rho}$  is 0.47, which leads to a 95% confidence band that is  $(3.26 \times \sigma)/\sqrt{n}$ , which is almost 70% wider. (pp. 9–10)

So, it is imperative to recognize that, even though your research question may not directly ask questions about anything spatial, you may have to employ a statistical technique that you would not have otherwise to appropriately accommodate the spatial autocorrelation in your data. Spatial regression adjusts for spatial dependence by incorporating spatial effects into a regression model. This is most commonly done with regression in one of two ways, spatially lagging the dependent variable by using a spatial lag model or spatially lagging the error term by using a spatial error model. These regression techniques will be described in detail in the next section of the chapter.

## How Do We Define Spatial Relationships?

---

As you will learn in the discussion below, a **spatial weights matrix** is essential to detect and correct for spatial autocorrelation. A spatial weights matrix is a way to numerically represent neighboring relationships among *points* (set of  $x$ - $y$  coordinates where a point represents one geographic location) or *polygons* (set of closed shapes that are defined by connecting lines, commonly representing a geographic boundary). Neighbors can be defined according to *contiguity*, *distance*, or *the number of neighbors*.

**Figure 12.1***Illustration of an edge and vertex in two sets of polygons*

Contiguity-based spatial weight matrixes concern the issue of adjacency. Generally, contiguity-based relationships are used only for polygons. And not only are they concerned with whether the observations touch (referred to as being contiguous), but they also are concerned with how they are touching. This is determined by whether the polygons share edges or vertexes. For instance, let us describe the spatial relationship between polygons D and E in Polygon Set #1 of Fig. 12.1. The bolded line is one (of many) edges in the figure. Polygons D and E share that edge and that is the only edge they share. The bolded points have been added to highlight two (of many) vertexes of Polygon Set #1 in the figure. D and E share those two vertexes (and they are the only two vertexes they share). If we were to consider the spatial relationships of only the top vertex, we find that polygons B, C, E, and F all share that same vertex. Notice, however, that while they all share at least one vertex, they do not all share edges. Now, let us contrast this with Polygon Set #2 in Fig. 12.1, where the center three polygons have shifted up a row. Here, we find that polygons D and E no longer share an edge, but they still share a vertex (though, it is only one vertex now).

With contiguity-based matrixes, *edges* and *vertexes* are used to delineate the type of contiguity by which we want to define neighbors. So, in what way do we want them to be touching for them to be considered neighbors? The two contiguity types are depicted in Fig. 12.2, where you will see that their names derive from the movement of chess pieces. With the **queen contiguity**, neighbors are those sharing at least one edge or at least one vertex. With the **rook contiguity**, neighbors are those sharing at least one

Figure 12.2

Illustration of contiguity types

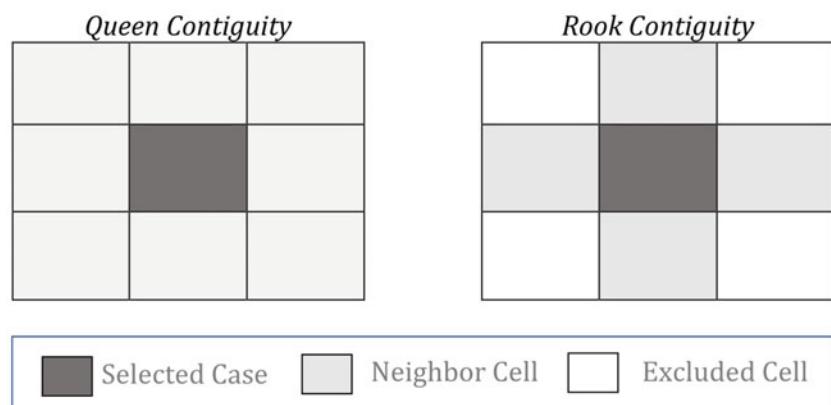
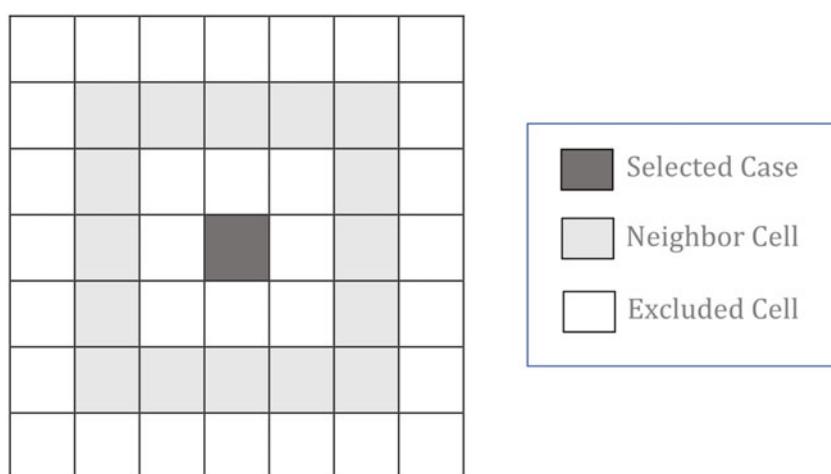


Figure 12.3

Illustration of second-order neighbors for queen contiguity



edge (shared vertexes are no longer considered). The user can also define the *order of contiguity*, which is how many cells away the given cell (number of lags). In Fig. 12.2, the cells shaded in light gray are considered first-order neighbors for both queen and rook contiguity, respectively. In Fig. 12.3, the cells shaded in light gray are the second-order neighbors for the queen contiguity (notice that first-order neighbors are no longer included automatically unless it is specified that lower orders be included).

As noted, an alternative to contiguity-based relationships is to define neighbors via distance or the number of neighbors. With a distance-based weight matrix, you define neighbors by those falling within a specified distance from each given point (e.g., a 500-foot radius). Distance can be calculated a number of ways but the two most common are **Euclidean distance** (straight-line distance; as a crow flies) or **Manhattan distance** (using a road network).

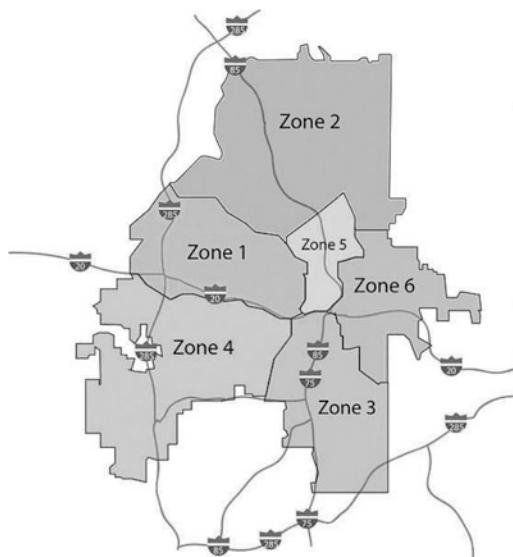
With a matrix based on number of neighbors, the user specifies that each observation should be given  **$k$ -nearest neighbors**, where  $k$  indicates the number of neighbors we want to be included in the analysis. This has the closest neighbor being assigned first, then the second closest, and so forth until  $k$  neighbors are defined for each observation. For example, specifying a spatial weight matrix with two nearest neighbors would mean that the two closest neighbors of each observation would be assigned. Note that while the neighbors may be contiguous,  $k$ -nearest neighbors do not require contiguity (e.g., consider  $k$ -nearest neighbor when it comes to defining neighbors with points).  $K$ -nearest neighbor weights do not rely on a distance threshold. A potential issue occurs when more than one neighbor falls within the same distance away from a given observation, but solutions exist for addressing such ties (e.g., randomly selecting).

Let us assume for the moment that we decided to define neighbors based on first-order rook contiguity. This information would be in the form of a list. For instance, with Polygon Set #1 in Fig. 12.1, there are nine polygons so the list would have nine rows. Polygon A would be in the first row (left column) and another column (on the right) would list polygon A's neighbors, which would be B and D. However, for analysis purposes, this list must be transformed so it is numerically stored in an  $n \times n$  matrix, where the spatial weights,  $w_{ij}$ , are defined for each pair of spatial units  $i$  and  $j$ . The relationships of Polygon Set #1 in Fig. 12.1 would be a  $9 \times 9$  matrix (A through I horizontally and then A through I vertically). The structure of this matrix is similar to a correlation matrix in the sense that the diagonal reflects each spatial unit with itself will always be zero; it is not a neighbor with itself.

The simplest way to quantify these relationships is by using a *binary* weighting strategy, where when a pair  $(i, j)$  of spatial units are considered neighbors,  $w_{ij} = 1$ , and otherwise  $w_{ij} = 0$ . An example of this weighting scheme using Atlanta Police Department's patrol zones is presented in Fig. 12.4 (top matrix). The first row identifies neighbors of Zone 1, with Zones 2–5 assigned a 1 since they border Zone 1, while Zone 6 is assigned a 0 because it does not border Zone 1. If you were to add the values on each row (or column), they are equal to the total number of assigned neighbors for that zone so Zone 1 having a total of four neighbors. The use of a *row-standardized* weighting scheme is standard practice, which takes the binary weights and divides them by the total number of weights for that

Figure 12.4

$w_{ij}$  connectivity matrixes of the Atlanta Police Department patrol zones by weighting scheme



		Binary Weighting Scheme						
		Zone						
Zone		1	2	3	4	5	6	Sum
		1	0	1	1	1	0	4
2	1	0	0	0	1	1	3	
3	1	0	0	1	1	1	4	
4	1	0	1	0	0	0	2	
5	1	1	1	0	0	1	4	
6	0	1	1	0	1	0	3	

		Row-Standardized Weighting Scheme						
		Zone						
Zone		1	2	3	4	5	6	Sum
		1	0	.25	.25	.25	.25	1
2	.33	0	0	0	.33	.33		1
3	.25	0	0	.25	.25	.25		1
4	.50	0	.50	0	0	0		1
5	.25	.25	.25	0	0	.25		1
6	0	.33	.33	0	.33	0		1

given row. Let us take a look at row #1 of the contiguity matrix in Fig. 12.4 (bottom table). Here, we divided the 1s by 4 since Zone 4 has a total of 4 neighbors, resulting in four separate weights of .25. Zone 4, in contrast, only has two neighbors so both neighbors are assigned a weight of .50. Accordingly, if you add the row-standardized weights of each row, they will each equal 1 (so all are proportionate). Furthermore, *inverse weighted distance* can be used to apply a distance decay to the weights so neighbors farther away are weighted less than nearby neighbors.

A spatial weights matrix is used to compute one of the oldest indicators of spatial autocorrelation, **Moran's I** (or more formally, **Moran's Index**). Moran's I is a correlation coefficient that quantifies the strength of spatial autocorrelation. It ranges from  $-1$  (where a negative spatial autocorrelation value generally indicates spatial dispersion but note an exception, discussed with Eq. (12.2)) to  $+1$  (where positive spatial autocorrelation indicates spatial clustering).

Moran's I can be defined as:

$$I = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{W \Sigma (x - \bar{x})^2} \quad \text{Equation 12.1}$$

where  $x$  is the number of spatial units,  $x_i$  is the value of an interval or ratio variable at a particular location  $i$ ,  $x_j$  is the value of the variable at another location  $j$ ,  $\bar{x}$  is the mean of that variable, and  $W$  is the sum of all spatial weights  $w_{ij}$ . The Moran's I when no spatial autocorrelation exists is the following:

$$E(I) = \frac{-1}{n - 1} \quad \text{Equation 12.2}$$

While the expected Moran's I will always be close to 0, notice that the expected Moran's I value,  $E(I)$ , is always negative so 0 cannot be used to distinguish positive and negative autocorrelation. Albeit, observed Moran's I values close to 0 indicate no spatial autocorrelation. Also, note that the expected Moran's I is inversely related to sample size, where the larger the sample size, the smaller the expected Moran's I value. A random permutation test is then used where the attribute values are randomly assigned a number of times (99 permutations for 0.01 precision, 999 permutations for 0.001 precision), and for each permutation, Moran's I is calculated to obtain a distribution of Moran's I values under *complete spatial randomness*. Then, the statistical significance of the observed Moran's I is determined by comparing it to the distribution of permuted Moran's I values.

Moran's I is normally computed using only one variable (though, a bivariate test is available that is interpreted similarly to a correlation coefficient) and a global Moran's I value is computed (one value summarizing the strength across all observations). For instance, imagine that we computed the Moran's I of the crime rate measured at the census tract level in Philadelphia. The one variable we are examining is the crime rate, and the unit of observation is census tracts. A Moran's I value in this instance would provide one value to indicate how similar (or dissimilar) a given census tract's crime rate is to its nearby neighbors on average.

Another example of a global Moran's I is if we were examining police calls for service in Washington, DC, a global Moran's I would be a single value that summarizes the observed spatial autocorrelation of property crime rates in Washington, DC, among zip codes on average. It would likely be positive, indicating that a given zip code has a property crime rate similar to that of its neighboring zip codes as compared to zip codes farther away. Alternatively, Moran's I can be computed locally. A local Moran's I is considered a local indicator of spatial association (LISA). A local Moran's I would illustrate how spatial autocorrelation varies across Washington, DC, at a microlevel (akin to a thematic map that is shaded according to the expected level of snowfall across a region). A local Moran's I would only depict areas with significant spatial autocorrelation, and it would also depict the type of the relationship: high-high (neighboring areas have

*similarly high* values), low-low (neighboring areas have *similarly low* values), high-low (neighboring areas have *dissimilarly low* values), and low-high (neighboring areas have *dissimilarly high* values) clusters.

Since we are going to employ a global Moran's I test to determine whether there is a spatial pattern to the residuals of an OLS regression model, let us walk through an example of computing a global Moran's I where we use the Atlanta Police Department zones and its binary connectivity matrix (where if  $i$  and  $j$  are neighbors, the  $w_{ij}$  will be 1 and 0 otherwise) presented in Fig. 12.4, as well as the total number of crime counts in each zone (Zone 1 = 1582; Zone 2 = 5827; Zone 3 = 1438; Zone 4 = 2382; Zone 5 = 3498; Zone 6 = 4510). The numerator in Eq. (12.1) is based on the cross-product of differences between neighboring values and the overall mean. While this is similar to Pearson's correlation coefficient, the covariance is only counted if spatial units  $i$  and  $j$  are neighbors, and normally only one variable is being considered with respect to Moran's I. The deviation in Eq. (12.1) is the sum of the squared deviations, which is scaled by the total weight of the matrix.

For the binary spatial weights matrix in Fig. 12.4, the total number of zones in this example,  $n$ , is 6, and the sum of the weights is 20. Table 12.1 presents the crime counts for each zone,  $x$ , the deviations from the mean,  $x - \bar{x}$ , and the squared mean deviations,  $(x - \bar{x})^2$ . The mean number of crimes is 3206, and the sum of the mean-squared deviations is 15,097,497 (part of the Moran's I denominator). The cross-products of these deviations for pairs are presented in Table 12.2. The sum of the cross-products is 4,244,420.

Now, we can use the information presented in Tables 12.1 and 12.2 to compute the observed global Moran's I for crime in the patrol zones.

**Table 12.1**

Mean crime deviations and square of the mean crime deviations

ZONES	CRIME COUNTS ( $x$ )	$x - \bar{x}$	$(x - \bar{x})^2$
1	1582	-1624	2,637,928
2	5827	2621	6,868,750
3	1438	-1768	31,26,425
4	2382	-824	679,256
5	3498	292	85,165
6	4510	1304	1,699,973
Total	19,237		15,097,497
Mean ( $\bar{x}$ )	3206		

**Table 12.2**

Spatial weights multiplied by the cross-product of the mean deviations

		ZONES						
		1	2	3	4	5	6	TOTAL
ZONES	DEVIATIONS	-1624	2621	-1768	-824	292	1304	
1	-1624		0	-1624 (2621)	-1624 (-1768)	-1624 (-824)	-1624 (292)	0 -520,254
2	2621	2621 (-1624)		0	0	0	2621 (292)	2621 (1304) -74,720
3	-1768	-1768 (-1624)		0	0	-1768 (-824)	-1768 (292)	-1768 (1304) 1,507,683
4	-824	-824 (-1624)		0	-824 (-1768)	0	0	0 2,795,865
5	292	292 (-1624)	292 (2621)	292 (-1768)		0	0	292 (1304) 155,347
6	1304		0	1304 (2621)	1304 (-1768)	0	1304 (292)	0 380,499
Total:								4,244,420

**Working It Out**

$$I = \frac{6(4,244,420)}{20(15,097,497)}$$

$$I = \frac{25,466,520}{301,949,940}$$

$$I = 0.084$$

We find that the observed Moran's I for crime in the Atlanta Police Department zones is 0.084. That observed value is being compared to the expected Moran's I for a random pattern, which would be the following for the patrol zones:

**Working It Out**

$$E(I) = \frac{-1}{6 - 1}$$

$$E(I) = -0.200$$

We find that our observed Moran's I of 0.084 is larger than our expected Moran's I of  $-0.200$ . This indicates that the crime counts in the patrol zones are more positively spatially autocorrelated than a random pattern that has no spatial autocorrelation.

## What Is Spatial Regression?

---

It is assumed that you have an understanding of ordinary least squares (OLS) regression but let us review some of the basic aspects of the technique. That way, the difference between OLS regression and spatial regression will become clearer to you later on, and you will be better able to understand why OLS regression is generally inappropriate for spatially referenced data.

OLS regression estimates a linear relationship between one or many independent variables and a dependent variable. A simplified version of the population model for this equation is as follows:

$$y = x\beta + e \quad \text{Equation 12.3}$$

where  $y$  is the dependent variable,  $x$  is the independent variable(s),  $\beta$  is the slope regression coefficient(s), and  $e$  is the random error term. The regression equation represents the line that is able to best summarize the relationship between the dependent variable,  $y$ , and a linear composite of the independent variable(s),  $x$ . Of course, the regression line will not be perfect. This difference between our observed and predicted estimates is considered a residual,  $e$ . OLS regression draws this best-fit line by minimizing the sum of the squared prediction errors. In other words, given that we want to quantify how far our observed points fall from our best-fit line, we cannot just add up how far each observation is from the line. Some of our observed values will fall higher than our best-fit line (overestimated the predicted value so it will have a positive value), while others fall lower than our best-fit line (underestimated the predicted value so it will have a negative value). As such, we need to square the predicted errors, so the positive and negative values do not cancel each other out. For this reason, OLS regression estimates regression coefficients based on the *least squares of error*.

One of the assumptions of OLS regression, which we touched upon at the beginning of this chapter and explored in-depth in Chap. 2, is that the residuals have equal variance (referred to as the homoscedasticity assumption). For instance, imagine you are searching through different radio stations to find something you want to listen to. Some of those stations

will just be static white noise. That is akin to what we want our error to be. A reason why spatially autocorrelated data may cause our error no longer to be stochastic is that if you mapped the error from your regression model, you may see a spatial patterning to it. For example, imagine I have quantified each country's effort to combat human trafficking into a scale (e.g., those countries with adequate human trafficking legislation and actively prosecute traffickers will have higher scores on this scale). If we were to conduct OLS regression to predict this score and then map the error of each country's observed value from the observed one, we may find that our regression model was more accurate in certain geographic areas. For instance, our model may have been better able to predict values of countries that fall into the European Union (that will normally border one another) since they follow one guiding legislation. You must also consider the fact that some countries may compel the countries they border (so as to reduce trafficking across borders) to heighten their anti-trafficking efforts, which may result in the error of our model being similar for geographically near countries.

The **heteroscedasticity** in the error of our model (residuals have unequal variance) violates an underlying assumption of OLS regression. Heteroscedasticity can oftentimes be a problem in regression with spatial data, especially if the size of the spatial unit of interest varies considerably. Thus, the presence of heteroscedastic errors may be an indication that spatial effects need to be accounted for, but spatial autocorrelation can be a problem without the presence of heteroscedasticity (and vice versa). There are many tools to identify heteroscedasticity, such as the *Breusch–Pagan* test (detects a linear form of heteroscedasticity), the *Koenker–Bassett* test (similar to Breusch–Pagan but is robust to outliers), and the *White* test (detects a more general form of heteroscedasticity).

While a scatterplot may indicate that the residuals are homoscedastic, you may find that the residuals vary across space (meaning that the errors exhibit spatial pattern). With autocorrelated spatial errors, a positive or negative error in one area is correlated with a positive or negative error in nearby areas. If spatial autocorrelation is present (which it likely is), this violates the independence assumption of OLS regression, and spatial regression should be used since significance tests based on OLS will be misleading (biased coefficients or errors, significance tests, and goodness-of-fit measures). While we use OLS regression as an example, the problems we note regarding the error terms apply more generally to regression approaches.

The two most common types of spatial regression are spatial lag models (SAR) and spatial error models (SEM). These models do not rely upon the least squares of errors method in estimating regression coefficients; instead, these spatial regression models select parameters that are the most probable with the highest likelihood (referred to as *maximum likelihood*)

**Table 12.3**

Comparison of OLS, spatial lag, and spatial error regression models

	<b>OLS</b>	<b>SPATIAL LAG</b>	<b>SPATIAL ERROR</b>
Estimation process	Least squares of errors	Maximum likelihood	Maximum likelihood
Location of spatial lag	N/A	Dependent variable as separate coefficient	Within error term
Equation	$y = x\beta + e$	$y = \rho Wy + x\beta + e$	$y = x\beta + e$ , where $e = \lambda We + \xi$
Adjusts for	None	Spatial dependence	Spatial heterogeneity
Neighbor influence	 No neighbor influence	 Dependent variable influenced by neighbors	 Residuals influenced by neighbors

*estimation*). Maximum likelihood is used for spatial regression because you are estimating the spatial relationships (lag or error) in addition to the other variables, and it does this by optimizing the beta and rho/lambda (described further below), depending whether a lag or error model is used. This requires an iterative process and not the closed-form mathematical solution of OLS. Furthermore, OLS regression cannot be used with a spatially lagged covariate as that leads to *simultaneity bias* (an explanatory variable, the spatial lag, is determined by and determines the dependent variable, causing the spatially lagged covariate to be correlated with the error term).<sup>1</sup> A summary of the differences between OLS regression, spatial lag regression, and spatial error regression is presented in Table 12.3. The differences between these models will be discussed further in the two sections below.

### What Is a Spatial Lag Model?

**Spatial lag models** are used for regression when you believe that your dependent variable is actively influenced by neighbors. It is also referred to as a spatial autoregressive model so you may see it go by the acronym, SAR. SARs are the most common way to model spatial dependence. This type of model deals with spatial dependence by incorporating a spatially lagged

---

<sup>1</sup>This issue can be addressed by maximum-likelihood estimation or through instrumental variable methods (see Anselin 2001).

dependent variable. You will see that the equation for a SAR model is similar to OLS regression (Eq. 12.3), but the spatially lagged dependent variable has been added to the right-hand side of the equation:

$$y = \rho W y + x\beta + e \quad \text{Equation 12.4}$$

where  $y$  is an  $n$  by 1 vector of observations on the dependent variable across space. Added to the OLS regression equation is neighboring values of  $y$  that are weighted by spatial weights matrix  $W$ , and then multiplied by  $\rho$  (termed *rho*), which is a spatial autoregressive coefficient for the dependent variable.

Rho indicates the strength of spatial autocorrelation present in the dependent variable, whereby the null hypothesis is a rho of 0 (no spatial autocorrelation).  $x$  is an  $n$  by  $k$  matrix of independent variables,  $\beta$  is a  $k$  by 1 vector of respective regression coefficients, and  $e$  is an  $n$  by 1 vector of independently and normally distributed residuals.

Unlike OLS regression, note that the estimation of  $y$  is dependent upon both  $x\beta$  and  $\rho W$ . This means that  $x\beta$  does not have the same straightforward interpretation as it does in OLS regression since it is difficult to tease apart the associations due to the covariates with that of  $Wy$ . That is to say that in OLS regression, the effect of the independent variable on the dependent variable is constant across observations. This is not the case when spatial effects are incorporated into the model because they vary as a result of different neighbors for each observation in your dataset. Albeit, you can read the coefficients the same, but keep in mind that the effect of the independent variable on the dependent variable includes the direct effect (as with OLS regression) and the indirect effect (average impact of the observations' neighbors). Since the effect of neighbors is provided as a coefficient, spatial autocorrelation in SARs is considered substantive. This is different from SEMs because SEM model spatial autocorrelation as a nuisance (something that only needs to be corrected for). Therefore, if you are interested in discovering and understanding the underlying processes behind the spatial interaction and the effect of neighbors on the dependent variable, you will want to use a SAR model or a spatial Durbin model (incorporate spatial effects in both the dependent variable and all the independent variables using spatial lags), depending on which variables you are interested in. This lag coefficient can be either positive (as with attraction) or negative (referred to as backwash or spread effect), although caution should be taken when interpreting the spatial lag coefficient since the spatial scale at which the data are measured is not always the same as the phenomena causing the spatial dependence.

### What Is a Spatial Error Model?

**Spatial error models** (SEMs) are used for regression when you believe that the observed spatial autocorrelation is due to a spatial process whose intensity varies across space (spatial heterogeneity). This type of model deals with spatial heterogeneity by estimating a spatial coefficient within the regression error term of the model. The SEM equation is similar to that of OLS regression (Eq. 12.3), but SEMs include an error term for neighbors (defined by the spatial weight matrix,  $W$ ), as well as the usual error term that is left after accounting for spatial autocorrelation:

$$y = x\beta + e,$$

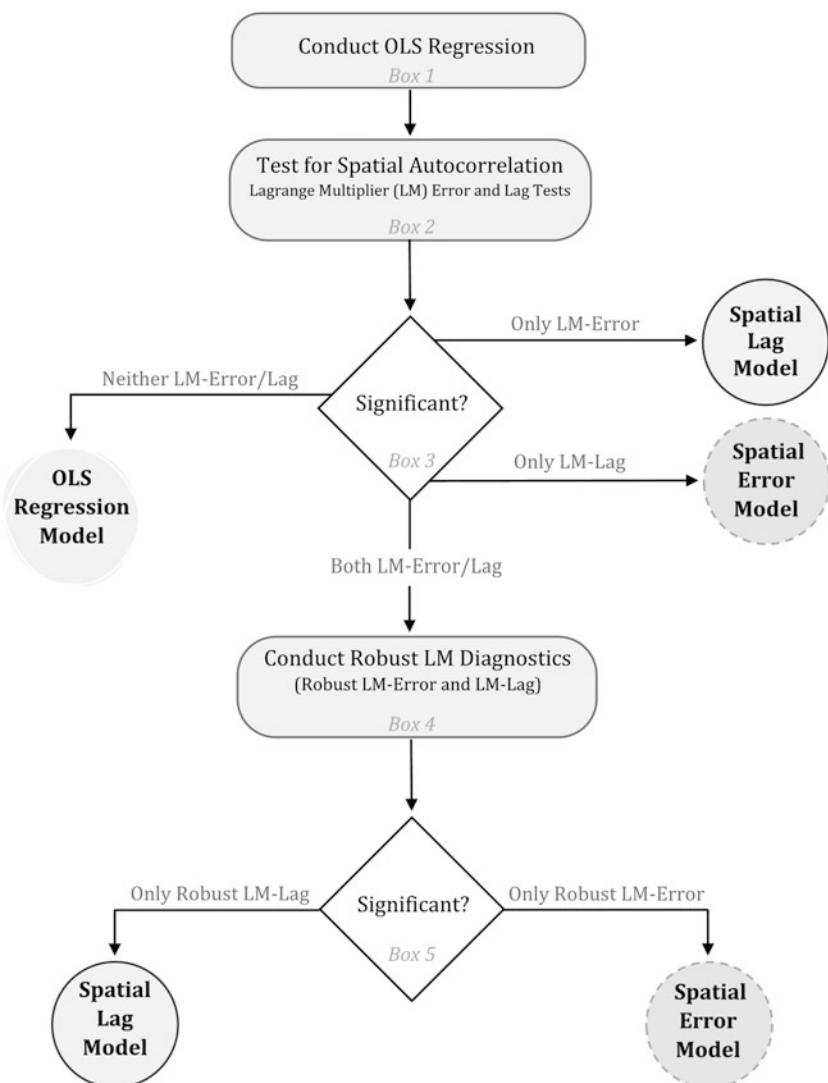
where  $e = \lambda We + \xi$ .

The equation is similar to that of OLS regression (Eq. 12.3); however,  $e$  is now comprised of  $\lambda$  (a spatial autoregressive coefficient for the residuals, referred to as *lambda*),  $We$  (residuals in neighboring locations that are weighted by the spatial weights matrix), and  $\xi$  (vector of independent and identically distributed errors; akin to white noise). Notice that in a SEM, the spatial coefficient to indicate the effect of neighbors on the dependent variable is not estimated separately like it is in a SAR model since it is incorporated into the error term. In that regard,  $\lambda$  is considered a nuisance that needs to be corrected for and it is of little interest to the research question at hand.

---

## Which Type of Spatial Regression Should I Use?

This section of the chapter outlines a process for conducting spatial regression. Since it can be challenging to select the appropriate spatial regression model if you are not quite sure whether the observed spatial autocorrelation is due to spatial dependence or spatial heterogeneity, we will be demonstrating the use of a decision process outlined by Luc Anselin for selecting the appropriate type of spatial regression model (Anselin 2005). The decision tree for this process is presented in Fig. 12.5. After these steps are reviewed below, we will walk through each step of the process. R code for conducting the analyses, as well as some application exercises, is provided at the end of the chapter. Step-by-step instructions for how to conduct these analyses in *Geoda* is also provided in the supplemental materials.

**Figure 12.5***Anselin's steps of decision making for spatial regression*

### **Assess Spatial Autocorrelation**

The need for a spatial regression model is based upon whether the residuals of an OLS regression model are spatially dependent. As such, you will first want to conduct your OLS regression model (box 1 in Fig. 12.5). Ignore the spatial autocorrelation issue for the moment, but you will want to otherwise follow the best practices for specifying the model that are

presented in Chap. 2 of this textbook (e.g., make sure a linear relationship is being modeled, no extreme outliers/multicollinearity present). While Anselin's decision-making process relies on two Lagrange Multiplier tests (described below) for spatial dependence, it is still common to save the residuals out as a new column in your dataset when you estimate your OLS regression model and then assess the residuals for spatial autocorrelation using Moran's I. Regardless of which spatial autocorrelation you conduct, you will need to create a spatial weights matrix to assess spatial dependence.

*What type of spatial weights matrix should I use?* You will need to first create a spatial weights matrix. Your results depend on the specification of your matrix so it may be beneficial to explore the effect of using different weighting schemes on your findings. A useful starting point for how to define your spatial weights matrix may be based on theory or empirical findings of previous studies. For example, in studying the criminogenic nature of facilities on nearby locations (examining 400–1200 ft distance thresholds from the facility), Groff and Lockwood (2014) relied upon Manhattan distances when calculating distance from the facility since crime pattern theory stresses the importance of the road network on crime. Their findings also suggest that crime type should also be considered when selecting a distance threshold of a spatial weight since the criminogenic effect of facilities decayed differently across space based upon the type of crime being examined (e.g., within a 400-ft distance from the facility, bars increase violent crime by 9%, property crime by 14%, and disorder-related crime by 49%). Further, when you select how you are defining the spatial weight matrix for your analysis, be mindful that if you are using a distance weight, you may want to determine the minimum distance needed for all observations to have at least one neighbor.

*Is a spatial regression model needed?* Once you have created your spatial weights matrix, you are able to move on to checking for spatial autocorrelation (box 2 in Fig. 12.5). You may want to start by creating a map the model residuals to visualize any spatial patterning. While Moran's I is a powerful tool for detecting spatial autocorrelation in residuals, it is not very helpful for providing guidance as to what type of spatial regression model is needed. That being the case, Anselin has proposed the use of two Lagrange Multiplier (LM) tests to assess spatial autocorrelation: LM spatial error dependence (LM-error) and LM spatial lag dependence (LM-lag). These tests are similar to Moran's I in that they are able to identify spatial autocorrelation, but they are also more specific as they offer insight as to whether a spatial lag model or spatial error model is best (if spatial autocorrelation is a problem, that is). LM-error tests for spatial error autocorrelation, while LM-lag tests for a missing spatially lagged dependent variable. There are also variants of these two tests that are robust to the presence of each other, robust LM-error, and robust LM-lag. As with Moran's I, the LM

tests require the use of a spatial weights matrix for estimation. If you find that any of the aforementioned tests reveal statistically significant spatial autocorrelation (box 3 in Fig. 12.5), a spatial regression model is appropriate, and you should proceed with assessing whether a spatial error or spatial lag model is appropriate. However, if you find no statistically significant spatial autocorrelation, you should keep the results of the OLS regression model and spatial regression is not needed.

### Which Type of Spatial Regression Model Should I Conduct?

If you have determined that a spatial regression model is appropriate, you will need to identify which type of spatial regression to conduct. To do this, you will once again rely upon the LM-error and LM-lag diagnostics to identify the best spatial regression model to conduct (box 3 in Fig. 12.5). If the LM-error diagnostic is statistically significant but the LM-lag diagnostic is not, then a spatial error regression model is appropriate. In the reverse, if the LM-lag diagnostic is statistically significant but the LM-error diagnostic is not, it is suggesting that a spatial lag model is a better model fit. In the event that both LM-error and LM-lag tests are statistically significant, then defer to the robust LM diagnostics to aid model selection (box 4 in Fig. 12.5). And only the robust versions of the LM diagnostics should be considered when both nonrobust LM diagnostics are statistically significant. As with the nonrobust LM when only one of the two diagnostics is statistically significant, use a spatial error model if the robust LM-error is statistically significant and use a spatial lag model if the robust LM-lag is significant (box 5 in Fig. 12.5). In the event that both robust diagnostics tests are statistically significant, it is advised to explore whether there is a strong theoretical rationale to inform model selection. If there is not, you can conduct both models and select the model with the best goodness-of-fit or select the model that has the highest LM value.

Once you have conducted the diagnostic tests, move on to re-estimating your model with the spatial regression technique(s) that you have selected and assess its fit. If you would like to compare fit between model types, keep in mind that OLS regression uses the least squares of errors for model fitting and spatial regression models use maximum likelihood. The difference in estimation has implications for model comparison because the pseudo  $R^2$  in the spatial regression models cannot be directly compared to the  $R^2$  in OLS regression. The three goodness-of-fit measures to use when comparing OLS to SAR/SEM are the following: log-likelihood, Akaike information criteria (AIC), and the Schwarz Bayesian information criterion (represented by the acronym SC or BIC). With the log-likelihood, higher values indicate a better model fit. Lower AIC and BIC/SC estimates indicate a better fit. The  $R^2$  can be used when comparing SAR and SEM since they are both based on maximum likelihood estimation. Additionally, the Likelihood Ratio (LR) test can be used to determine whether the spatial

regression model is a significantly better fit over OLS regression model. We will walk through an example of this process in the next section, as well as review model interpretation.

## Spatial Regression Example

---

We will be using the software R for our spatial regression example, and the R code is reviewed at the end of the chapter. The data we will be analyzing derive from a large National Institutes of Health study of a sample of hot spots and non-hot spots in Baltimore, Maryland (Weisburd et al. 2011). The aim of the original study was to understand how living in drug or violent crime hot spots (defined as a street segment from intersection to intersection in a city) influences personal health, drug use, and crime. Further, it sought to develop knowledge on why places become drug or crime hot spots and how characteristics of street segments and their residents' impact developmental trends of health, drug use, and crime. The sample includes 3738 face-to-face surveys of residents living on 449 street segments in Baltimore City, MD. The 449 street segments were identified by randomly selecting street segments from the top 3.0% of residential streets with drug crime calls and violent crime calls for the entire city, and residential streets with low crime were selected for comparison purposes. Face-to-face surveys were then conducted at a random sample of residences on each street with a goal of at least seven surveys per street segment. Respondents were asked about their physical and mental health and individual perceptions about their street block and local community (e.g., views on the police, physical/social disorder, collective action, crime). Next, we will demonstrate Anselin's spatial regression decision process that we reviewed above.

### Step 1: Conduct OLS Regression

We will be examining whether type of hot spot significantly predicts perceived level of social cohesion, which was measured by using a six-item scale. It ranges from 0 to 6, where higher values on the scale indicate higher social cohesion. These items asked respondents such things as whether their neighbors share the same values, can be trusted, get along, help each other, talk to one another, and watch out for each other. The street segment is the unit of analysis in the dataset, so the household survey responses were aggregated (mean) to the street segment level. As such, the dependent variable for the example reflects the mean perceived level of social cohesion of survey respondents, where the mean social cohesion level across street segments was 3.55 (min = 2.58, max = 4.40).

Our main independent variable of interest is hot spot type, which was measured by first categorizing segments by type of crime occurring on the

**Table 12.4**

OLS regression model predicting mean social cohesion

INDEPENDENT VARIABLES	REGRESSION COEFFICIENT	STANDARD ERROR	t	p
(Intercept)	3.6056***	0.0876	41.1416	<0.0001
Hot spot type (ref: violent crime only)				
Cold	0.4302***	0.0475	9.0624	<0.0001
Cool	0.1945***	0.0371	5.2387	<0.0001
Drug crime	0.0765*	0.0351	2.1778	0.0300
Drug/violent crime	-0.0132	0.0443	-0.2979	0.7659
Number of respondents	-0.0183	0.0098	-1.8612	0.0634
Number of dwelling units on segment	-0.0002	0.0002	-1.3260	0.1855
Log-Likelihood:	-49.58			
Akaike Info Criterion (AIC):	113.17			
Schwarz Criterion (SC/BIC):	141.92			

Adjusted  $R^2$ : 0.208,  $F(7, 442) = 20.60$ ,  $p < 0.001$ 

street: 47 cold spots (very low crime), 100 cool spots (low crime), 121 drug hot spots, 126 violent hot spots, and 55 drug/violent crime hot spots. This variable will be incorporated into the model using four dummy variables (cold spot, cool spot, drug spot, drug/violent crime spot dummies, with the violent crime hot spot being the reference category). We will also control for the number of survey respondents and the number of dwelling units on each street segment. The results of our OLS regression model are presented in Table 12.4. In running this model, we also saved the residuals as a new variable in the dataset. In order to determine whether spatial autocorrelation is a concern, we need to create a spatial weights matrix.

### Step 2: Construct a Spatial Weights Matrix

We must specify how spatial relationships should be defined. Once neighbors are assigned, the list of neighbors needs to be transformed with a weighting technique, so all spatial relationships are represented numerically in a spatial weights matrix (with a nonzero weight indicating that two given observations are neighbors). The specifications of our spatial weights matrix are presented in Table 12.5. Given that we are using the centroid (geographic center) of street segments in the sample as our unit of analysis (rather than polygons, which normally share borders with one another/contiguous), we would not want to define neighbors using a contiguity weight. For our analysis, we are going to be using a Euclidean distance whereby we are going to assign neighbors that fall within 6500 ft radius of each observation (based on the centroid/center of the street segment). This distance bandwidth ensures that all street segments have at least one assigned neighbor. We are going to employ a row-standardized weighting scheme. A summary of the spatial weights matrix that we created is also presented in Table 12.5. Based on the matrix we defined, the average

**Table 12.5**

Spatial weights matrix connectivity summary

SPECIFICATIONS	DISTANCE
<i>Weight</i>	
Distance metric	Euclidean
Distance bandwidth	6500 ft
Weighting strategy	Row-standardized
<i>Overall connectedness</i>	
Number of observations	449
Min/max neighbors	2/95
Mean neighbors	42.35
Percent nonzero links	9.43

**Table 12.6**

Moran's I and Lagrange multiple diagnostics

TEST	ESTIMATE	p
<i>Moran's I (global)</i>		
Dependent variable	0.101***	0.001
OLS residuals	0.034***	0.004
<i>Lagrange multiplier (LM)</i>		
LM-lag	13.81***	<0.001
LM-error	6.19*	0.013
Robust LM-lag	9.61**	0.002
Robust LM-error	1.99	0.158

number of links (neighbors) in our sample is 42.35, with the fewest number of neighbors of a given street segment being 2 and the maximum number being 95 neighbors.

### Step 3: Test for Spatial Autocorrelation

We can proceed to using our spatial weight matrix to assess spatial autocorrelation. We are going to do this in two ways—by calculating the global Moran's I of the OLS regression residuals and by calculating the Lagrange Multiplier tests. The results of these tests are presented in Table 12.6. (The robust LM estimates are included in the table in case we will need them later on.) The Moran's I is statistically significant, which means that the residuals of our OLS regression model are spatially autocorrelated. Furthermore, both the LM-lag and LM-error tests are statistically significant. These diagnostics suggest that spatial regression may be a better fit for our model. Therefore, we will proceed to the next step where we will identify the appropriate spatial regression model to conduct.

### Step 4: Select and Conduct Spatial Regression Model

We have determined that the residuals of our OLS regression model are spatially dependent, so we will need to identify which type of spatial regression to conduct. To do this, we will once again rely upon the LM-error and

**Table 12.7**

Spatial lag regression model predicting mean social cohesion

INDEPENDENT VARIABLES	REGRESSION COEFFICIENT	STANDARD ERROR	t	p
(Rho-spatially lagged DV)	0.3152**	0.1170	2.6938	0.0071
(Intercept)	2.5027***	0.4219	5.9326	<0.0001
Hot spot type (ref: violent crime only)				
Cold	0.3943***	0.0473	8.3315	<0.0001
Cool	0.1712***	0.0367	4.6611	<0.0001
Drug crime	0.0684*	0.0345	1.9848	0.0472
Drug/violent crime	-0.0096	0.0434	-0.2211	0.8251
Number of respondents	-0.0183	0.0097	-1.8953	0.0581
Number of dwelling units on segment	-0.0003	0.0002	-1.6247	0.1042
Log-likelihood:	-45.25			
Akaike info criterion (AIC):	106.51			
Schwarz criterion (SC/BIC):	139.36			
Likelihood ratio test:	8.66**			

DV dependent variable (social cohesion);  $R^2$ : 0.237 (Caution: can be compared to  $R^2$  of a SEM but DO NOT compare this to OLS regression  $R^2$ )

LM-lag diagnostics that are presented in Table 12.6. Referring back to Anselin's spatial regression decision process (box 3 in Fig. 12.5), we must rely upon the robust LM diagnostics to guide our model selection since both nonrobust LM measures are statistically significant. The decision process flowchart dictates that when only the robust LM-lag is statistically significant, we are to conduct a spatial lag regression model (box 5 in Fig. 12.5).

The spatial lag regression results are presented in Table 12.7. There are a few small changes in the beta coefficients when comparing the OLS and spatial lag models. In the OLS model, we observe small-to-moderate statistically significant beta coefficients for cold ( $b = 0.430$ ), cool ( $b = 0.195$ ), and drug ( $b = 0.076$ ) hot spots as compared to violent crime hot spots in predicting street segment level social cohesion. In contrast, these estimated coefficients in the spatial lag model are slightly lower (cold,  $b = 0.394$ ; cool,  $b = 0.171$ ; drug,  $b = 0.068$ ). As mentioned above, while the beta coefficients do not have the exact same interpretation as OLS regression, you can interpret them just as an OLS regression. But, keep in mind that the spatial lag model includes both the direct effect of the given segments' hot spot type and the effect of its neighboring street segments' hot spot type on level of social cohesion at the street segment. Accordingly, the average level of social cohesion across street segments in the sample after controlling for covariates and the effects due to neighboring street segments is 2.502. Cold and cool spots, respectively, have 0.394 and 0.171 significantly higher social cohesion scores across street segments in the sample in comparison with violent crime hot spots (both  $p < 0.001$ ), whereas drug hot spots have 0.068 higher social cohesion ( $p < 0.05$ ).

Finally, the coefficient for the spatially lagged dependent variable (rho) is positive and statistically significant ( $b = 0.315$ ;  $p < 0.01$ ), indicating that

when the social cohesion of surrounding street segments increases by 0.315, so does the cohesion level in each street segment. This reinforces the belief that adding the spatial lag of our dependent variable is important for specifying the distribution of perceived social cohesion across space. After taking it into account, the model diagnostics indicate that our spatial lag model is a notably better fit than the OLS regression model. The AIC (106.51 vs. 113.17) and BIC (139.36 vs. 141.92) are lower in the spatial lag model in comparison with the OLS regression model, which is desirable. The likelihood ratio test is also statistically significant and the global Moran's I of the spatial lag regression residuals is no longer statistically significant (Moran's I = -0.0026;  $p = 0.235$ ), both of which also indicate a better goodness-of-fit over the OLS regression model.

In following Anselin's process for selecting an appropriate spatial regression model, the robust LM tests indicated that a SAR would be best fit. However, as mentioned above, there are situations where both the robust LM tests are statistically significant. In which case, if there is not a strong theoretical rationale to inform model selection, you can conduct the lag and error models to determine the best model fit. As such, let us review the findings from a SEM model. The output for this model is presented in Table 12.8. In comparison with the OLS and SAR models, we find that drug crime hot spot type is no longer statistically significant. We also find that number of respondents on the street segment is now statistically significant; every additional respondent on a street segment decreases average social cohesion by -0.0192 ( $p = 0.048$ ).

The model diagnostics for all three types of regression models are presented in Table 12.9. The pseudo- $R^2$  for the SAR model is only slightly

**Table 12.8**

Spatial error regression model predicting mean social cohesion

INDEPENDENT VARIABLES	REGRESSION COEFFICIENT	STANDARD ERROR	t	p
(Intercept)	3.6285***	0.0876	41.4057	<0.0001
Hot spot type (ref: violent crime only)				
Cold	0.3996***	0.0480	8.3277	<0.0001
Cool	0.1702***	0.0371	4.5883	<0.0001
Drug crime	0.0616	0.0352	1.7486	0.0804
Drug/violent crime	-0.0114	0.0440	-0.2582	0.7963
Number of respondents	-0.0192*	0.0097	-1.9766	0.0481
Number of dwelling units on segment	-0.0003	0.0002	-1.6720	0.0945
Lambda-spatially correlated error	0.31*			
Log-likelihood:	-47.00			
Akaike info criterion (AIC):	108.01			
Schwarz criterion (SC/BIC):	136.76			
Likelihood ratio test:	5.15*			

DV dependent variable (social cohesion);  $R^2$ : 0.230 (Caution: can be compared to  $R^2$  of a SAR but DO NOT compare this to OLS regression  $R^2$ )

**Table 12.9**

Comparison of goodness of fit

	<b>OLS</b>	<b>SAR</b>	<b>SEM</b>	<b>BEST FIT?</b>
<i>Diagnostic</i>				
Pseudo- $R^2$	—	0.237	0.230	SAR
AIC	113.17	106.51	108.01	SAR
SC/BIC	141.92	139.36	136.76	SEM
Log-likelihood	-49.58	-45.25	-47.00	SAR
Likelihood ratio test	—	8.66**	5.15*	SAR

higher than the SEM (0.237 vs. 0.230), suggesting that the SAR has a higher percent of the variation in social cohesion explained. As noted, the  $R^2$  from the OLS regression model is not included in the table since it is not appropriate to compare it with the  $R^2$  from the spatial regression models. The SAR model has the most desirable AIC, while the SEM model has the most desirable SC/BIC out of the three models.

While both the SAR and SEM made a significantly improvement in the log-likelihood in comparison with the OLS, the SAR model had the most considerable change. These diagnostics reaffirm that the SAR model is the best-fitting model of the three. While the selection of a SAR model did not change the beta coefficients and their statistical significance considerably, it is important to be aware that is a possibility; therefore, spatial regression should be employed when it is known that the assumptions of OLS regression have been violated due to spatial autocorrelation.

## Chapter Summary

Spatially referenced data frequently display **spatial autocorrelation**, where observations will be similar to those observations nearby. **Spatial dependence** and **spatial heterogeneity** are the two processes that produce spatial autocorrelation. Spatial dependence is due to interaction that occurs in space (e.g., interaction among people leads to the spread of a virus), whereas spatial heterogeneity arises due to a shared spatial phenomenon (e.g., same climate). The use of spatially autocorrelated data commonly violates the assumptions of OLS regression because the data are not independent and more likely to display **heteroscedasticity**. Spatial autocorrelation can be quantified using **Moran's I**, whereby values closer to 1 indicate dependence and values closer to -1 indicate dispersion. **Spatial lag models** and **spatial error models** are two of the most common spatial regression modeling techniques, and both use maximum likelihood estimation. Spatial lag regression incorporates a spatially lagged dependent variable on the right regression equation as a covariate and is

used when adjusting for spatial dependence. Spatial error regression incorporates a spatial weight into the errors of the model and is used when adjusting for spatial heterogeneity. These regression techniques overcome the problem of spatial autocorrelation by using a **spatial weights matrix** during model estimation, which contains information about neighboring observations in the data. Matrixes can define neighbors a number of ways, such as via contiguity (**queen** or **rook contiguity**), distance (**Euclidean distance** or **Manhattan distance**), or number of nearby neighbors to include (**k-nearest neighbors**). Anselin has outlined a spatial regression decision process, which relies on Lagrange's Multiplier tests (LM-lag and LM-error) for helping to determine whether OLS regression, spatial lag regression, or spatial error regression is the most appropriate modeling technique.

## Key Terms

---

**Euclidean distance** Straight-line distance between two points (as the crow flies).

**Heteroscedasticity** A condition in which the variance of an error term is unequally distributed.

**k-nearest neighbors** A way to define neighbors, where  $k$  refers to the number of nearby neighbors to assign.

**Manhattan distance** Distance between two points that is measured on a road network (similar to a car moving through a city).

**Moran's Index** A coefficient ranging from  $-1$  (dispersion) to  $1$  (clustered) that quantifies spatial autocorrelation (also known as Moran's I).

**Queen contiguity** A type of spatial weight that defines neighbors as polygons sharing at least one edge or vertex.

**Rook contiguity** A type of spatial weight that defines neighbors as polygons sharing at least one edge.

**Spatial autocorrelation** The correlation of a variable's values over space, with strong

autocorrelation indicating that nearby values of the variable are similar to one another.

**Spatial dependence** Similarity in nearby observations due to the influence of neighboring locations via interaction.

**Spatial error model (SEM)** A regression technique that corrects for spatial autocorrelation in the residuals (spatial heterogeneity) by incorporating spatial effects via adding a spatial coefficient within the error term of the model.

**Spatial heterogeneity** Large-scale, regional variation due to a process that varies across space.

**Spatial lag model (SAR)** A regression technique that corrects for a dependent variable with spatial dependence by incorporating spatial effects via adding a spatially lagged dependent variable as a covariate.

**Spatial weights matrix** An  $n \times n$  matrix that quantifies spatial relationships, where  $n$  is the number of geographic features.

## Symbols and Formulas

---

I	Moran's
E(I)	Expected Moran's I
$\rho$	Spatial autoregressive coefficient
$W$	Spatial weights matrix

Computation of Moran's I:

$$I = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{W\Sigma(x - \bar{x})^2}$$

Computation of Expected Moran's I:

$$E(I) = \frac{-1}{n-1}$$

Simplified equation for OLS regression model:

$$y = x\beta + e$$

Equation for spatial lag regression model:

$$y = \rho W y + x\beta + e$$

Equation for spatial error regression model:

$$y = x\beta + e,$$

where  $e = \lambda We + \xi$ .

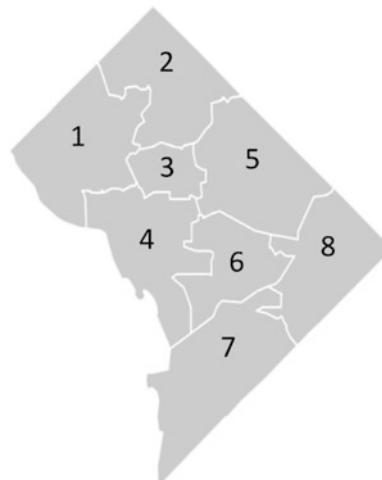
## Exercises

---

- 12.1. For 12.1a and 12.1b, list the neighbors for each polygon in Fig. 12.1 based on the specified contiguity type. You may want to create two lists like the ones below (one for Polygon Set #1 in 12.1a and one for Polygon Set #2 in 12.1b).
- List the neighboring polygons (letter of the alphabet) for each polygon in Polygon Set #1 according to first-order rook contiguity.
  - List the neighboring polygons (letter of the alphabet) for each polygon in Polygon Set #2 according to first-order queen contiguity.

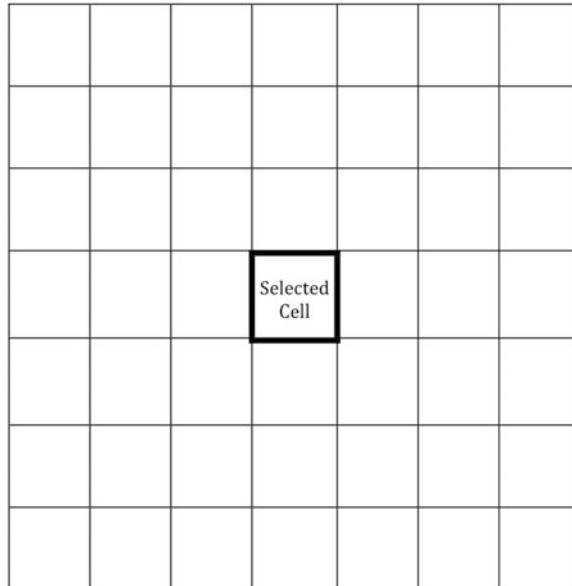
12.1A POLYGON SET #1		12.1B POLYGON SET #2	
POLY	NEIGHBORS	POLY	NEIGHBORS
A		A	
B		B	
C		C	
D		D	
E		E	
F		F	
G		G	
H		H	
I		I	

- 12.2. Use the geographic boundary map of the eight Wards in the District of Columbia to answer questions 12.2 (a) and 12.2 (b) below.



- Create a binary spatial weights matrix of the Wards. Which Ward(s) have the highest sum of weights?
- Create a row-standardized spatial weights matrix of the Wards. Which Ward(s) have the highest sum of weights?

- 12.3. Create two  $7 \times 7$  grids like the one below to answer questions 12.3(a) through 12.3(c).
- Shade in the third-order rook neighbors of the selected cell.
  - Shade in the first -, second-, and third-order rook neighbors of the selected cell.
  - Shade in second-order queen neighbors of the selected cell.



- 12.4. For questions 12.4(a) through 12.14(c), identify the correct regression technique to employ using the diagnostics for spatial autocorrelation provided: OLS, spatial lag, or spatial error regression. You may want to refer to boxes 3 through 5 in Fig. 12.5, as well as the in-text discussion, for guidance.

(a):

TEST	VALUE	<i>p</i>
Moran's I (error)	1.300	0.502
Lagrange multiplier (lag)	4.202	0.350
Robust LM (lag)	2.535	0.904
Lagrange multiplier (error)	3.640	0.242
Robust LM (error)	2.399	0.565

(b):

TEST	VALUE	<i>p</i>
Moran's I (error)	6.236	0.022
Lagrange multiplier (lag)	18.534	0.001
Robust LM (lag)	5.932	0.060
Lagrange multiplier (error)	29.538	0.001
Robust LM (error)	9.209	0.009

(c):

TEST	VALUE	<i>p</i>
Moran's I (error)	9.745	0.001
Lagrange multiplier (lag)	15.234	0.004
Robust LM (lag)	9.382	0.028
Lagrange multiplier (error)	24.242	0.003
Robust LM (error)	6.421	0.034

- 12.5. Provide a criminology/criminal justice-related example of spatial dependence not provided in-text.
- 12.6. Provide a criminology/criminal justice-related example of spatial heterogeneity not provided in-text.

## Computer Exercises

Since you are already familiar with OLS multivariate regression (Chap. 2 in this textbook), the R code to conduct spatial regression is generally straightforward and only requires a few more steps to calculate the weight matrix and spatial effects in the regression model. The R code to compute a spatial lag regression and a spatial error regression with the dataset used in the spatial regression example section of this chapter is described below. Adapt this R code using the publicly available shapefiles provided to complete the problems in the next section. But, just a quick note before delving into the R code... While R can accommodate spatially referenced data and offers many packages to conduct spatial analysis, it is not a spatial data analysis software. As such, you may want to use software that was designed as such since they are oftentimes more user-friendly with spatial analysis such as *Geoda*, which is free. A tutorial on how to conduct these analyses in *Geoda* is provided in the supplemental materials.

## R

Before you begin, you may find it helpful to turn off scientific notation. This can be done in R as follows:

```
options(scipen=999)
```

Next, open the shapefile you are working with in R. You will first want to specify your working directory using the `setwd()` function or by using the `here()` function within an R project. Then, add the shapefile to your R working environment using the `st_read()` function. To do this, you will want to install and load the package `sf` using `install.packages()` and `library()`, respectively. Here, we install and load the `sf` package, and then, we want to store the shapefile in an object named `df`:

```
install.packages("sf")
library(sf)
df <- st_read("ShapeFileName.shp")
```

Make sure to include quotation marks around the package name in the `install.packages()` function, but do not use quotation marks within the `library()` function. Also, make sure that you include the three-letter filetype `.shp`.

### OLS Regression

Once you have successfully loaded your shapefile, it is good practice to look over the data file using the `View(df)` command. You can also create a simple map to visualize how your dependent variable varies across space; we present this code later on when reviewing how to visualize residuals in R. For now, let us move on to running our OLS regression model. Remember to first assess for common violations of assumptions before specifying your model (e.g., multicollinearity, extreme outliers). To conduct our regression model, we are going to use the `lm()` function. Here, we are specifying our dependent variable first (variable named `Mn_Cohesion`), followed by our independent variables. Make sure to specify the dataset name with the `data=` command. The model results will be stored in an object named `ols`. Then, use the `summary()` function to view detailed results of your model.

```
ols <- lm(Mn_Cohesion ~ Cold + Cool + Drug +
  CombineVioDrug + N_Respondents + Sum_Dwelling,
  data=df)
summary(ols)
```

Since we want to review whether the model residuals are spatially autocorrelated, we are going to extract them using the `residuals()` function and save them in a new variable named `res_ols`.

```
df$res_ols <- residuals(ols)
```

The AIC and BIC of the OLS regression model can be extracted with the functions `AIC()` and `BIC()`, respectively. These functions can be found within the `stats` package.

```
ols.AIC<-AIC(ols)
ols.BIC<-BIC(ols)
```

### *Visualize OLS Regression Residuals Spatially*

Now, let us visualize our model residuals using a thematic map (and you may want to separately plot your dependent variable to view the distribution across space). We are mapping point data in our example, but you can use this code for the exercises when you are plotting polygons of Chicago census tracts. You will want to install and load the *RColorBrewer* package first. Then, we are going to convert the shapefile to a different format using the **as\_spatial()** function so we can plot the residuals using **plot()**.

Now, specify the color ramp you would like to use using **brewer.pal()**. We are using the one named *RdYlGn*, which includes the colors (in ascending order) red, yellow, and green. You can view other color ramps in the package by using the command **display.brewer.all()**. We are binning the residuals into 5 categories so we want to specify that we want 5 colors from the color palette, and then, we are going to specify that we want 5 cut-points made when using the **cut()** function. We need to use the *@data* argument with this package since we want to access variable information from a shapefile rather than a typical data frame.

```
df2<-as_Spatial(df)
map_colors <- brewer.pal(5, "RdYlGn")
map_colors <- colorRampPalette(map_colors)(5)

class_of_residuals <- cut(df2@data$res_ols, 5)
map_colors <- map_colors[as.numeric(class_of_residuals)]
plot(df2, col=map_colors)
```

### *Distance-Based Spatial Weights Matrix*

To assess whether spatial autocorrelation is present, a spatial weights matrix needs to be created. The example in this chapter defined neighbors using a distance-based matrix. This first part of creating a spatial weight can be implemented in R using the **dnearneigh()** function in the *spdep* package. This creates a list of each observation and their corresponding neighbors. After you have the package installed, you are ready to define your matrix. You need to specify the spatial object (which is *df* in our case), the lower distance bound, and the upper distance bound. With the code below, we are specifying that neighbors be defined as those falling within 6500 ft from the given observation, and we are placing this information in an object named *list\_neigh*. You can then use the **summary()** function to view this information. Be aware that the units of the shapefile in this example is in feet and is a points file (not polygons) so we are able to specify feet when delineating Euclidean distance. We can verify that our data are specified in feet by using the command *st\_geometry(df)*. If your data are in latitude/longitude, you will want to define your distance band within the **dnearneigh()** function in kilometers.

```
list_neigh <-dnearneigh(df, 0, 6500)
```

To get a summary of connectivity (e.g., average number of neighbors, min/max number of neighbors), use the **summary()** function. The **View()** function can also be used to open the actual list of neighbors that each case has.

```
summary(list_neigh)
View(list_neigh)
```

Then, the **nb2mat()** and **mat2listw()** functions transform the list into a format that is needed to employ the spatial weights matrix for the Moran's I test. The two most common style choices are *B* (binary) and *W* (row-standardized).

```
list_neigh_mtx <- nb2mat(list_neigh, style='W')
Dist_mtx <- mat2listw(list_neigh_mtx, style='W')
```

### *Contiguity-Based Spatial Weights Matrix*

If you want to create a spatial weights matrix based on adjacency, you can rely on functions within the *spdep* package. First, use the **poly2nb()** function to define first-order neighbors. If you want neighbors based on queen contiguity, specify the *queen=* argument as TRUE (set to FALSE for rook contiguity).

```
queen_wgt<-poly2nb(df, queen=TRUE)
```

You can query the list to have it report the assigned neighbors. For instance, the following will provide you the neighbors for those assigned to the first case in the *df* dataset. Note that it will return the row number of the neighbor (not the unique identifier from the data frame you are assigning neighbors from).

```
queen_wgt[2] # R command
[1] 3 6 7 # Output
```

The R output above indicates that rows 3, 6, and 7 from our data frame *df* are neighbors of row 2 in the same data frame. You can then use that information to query the ID field in *df* by specifying the variable name and then row(s).

```
df$street_ID[2] #ID of street we are identifying neighbors for
df$street_ID[c(3, 6, 7)] #ID of its neighbors
```

The **nblag()** function can be employed for higher order neighbors. In this example, first-order queen neighbors are defined in *queen\_wgt*, and then, second-order neighbors are defined.

```
queen_wgt2order <- nblag(queen_wgt, 2)
```

Note that if you use the command *summary(queen\_wgt2order)* and then *class(queen\_wgt2order)*, you will see that the **nblag()** function created two lists within a list (one for first-order and one for second-order neighbors). If you want both first- and second-order neighbors in one list, combine these objects using the **nblag\_cumul()** function. Then, as we did with the distance-based weights matrix, format the neighbor lists (or one, if that is all you created) into a spatial weights matrix using the

**nb2mat()**. As we did above, we are specifying the style as  $W$  for row-standardized (as opposed to  $B$  for binary).

```
cuml <- nblag_cumul(queen_wgt2order)
cuml_mtx <- nb2mat(cuml, style=' W')
```

You can then query this matrix to report the assigned weight. You can specify the row, column, or cell of the matrix. For example, see the following commands:

```
cuml_mtx[2,] # All weights for df$street_ID[2]
cuml_mtx[2,3] # Weight df$street_ID[2] has been given for
# neighbor df$street_ID[3]
```

You then need to convert the matrix into a different format for analysis purposes with the **mat2listw()** function.

```
culm_mtx <- mat2listw(cuml_mtx, style='W')
```

### *Moran's I Test of Residuals*

You can use the **lm.morantest()** function, which is in the *spdep* package. To conduct the test, you need to specify the object that contains your OLS result, the spatial weight matrix, your alternative hypothesis (*greater*, *less*, or *two-sided*), and specify with the *resfun* argument that we have weighted residuals.

```
lm.morantest(ols, Dist_mtx, alternative="two.sided",
             resfun=weighted.residuals)
```

Once run, this test will provide you with the observed global Moran's I of the residuals, the expected Moran's I, and the associated *p*-value of the Moran's I estimate.

### *Lagrange Multiplier Diagnostics*

You are able to conduct both the nonrobust and robust Lagrange Multiplier (LM) diagnostics simultaneously with **lm.LMtests()**. This function is from the *spdep* package. The type of tests is specified with the *test*= argument. And as with the Moran's I test, you need to specify the object that contains your OLS regression results, followed by the spatial weights matrix. With this code, we, respectively, conduct the LM-lag, LM-error, robust LM-lag, and robust LM-error tests.

```
lm.LMtests(ols, Dist_mtx,
            test = c("LMLag", "LMMerr", "RLMLag", "RLMMerr"))
```

### *Spatial Lag/Error Regression*

The code to conduct spatial lag regression and spatial error regression in R is very similar to OLS regression. The dependent variable is specified, followed by the independent variable. The difference being that a different function is used, and you also must specify the spatial weights matrix. Spatial lag regression is specified using

**lagsarlm()**, while spatial error regression is specified using **errorsarlm()**. These functions are from the *spdep* package.

```
lag_model<-lagsarlm(Mn_Cohesion ~ Cold + Cool + Drug +
  CombineVioDrug + N_Respondents + Sum_Dwelling,
  data=df, Dist_mtx)
error_model<-errorsarlm(Mn_Cohesion ~ Cold + Cool +
  Drug + CombineVioDrug + N_Respondents + Sum_Dwelling,
  data=df, Dist_mtx)
```

As with OLS regression, you can obtain a summary of the spatial regression results by using the **summary()** function:

```
summary(lag_model)
summary(error_model)
```

Note that the summary of the results contains the *LR test value* and *p-value*, which is the likelihood ratio test comparing the given model to OLS regression.

### Problems

Open the St. Louis county-level homicide shapefile (*stlouis.shp*) to complete questions 1 and 2 below. Use the FIPS variable as the state/county identifier.

1. Create a first-order queen contiguity matrix of St. Louis counties that has a binary weighting scheme.
  - (a) Provide the FIPS number for each assigned neighbor of Madison, Illinois (row 33 in the shapefile, FIPS = 17,119).
  - (b) Provide the numeric weight that Fayette, Illinois (row 27 in the shapefile, FIPS = 17,051), has been given for its neighbor, Shelby, Illinois (row 16 in the shapefile, FIPS = 17,173).
  - (c) Provide the numeric weight that Shelby, Illinois (row 16 in the shapefile, FIPS = 17,173), has been given for its neighbor, Fayette, Illinois (row 27 in the shapefile, FIPS = 17,051).
2. Create a first-order queen contiguity matrix of St. Louis counties that has a row-standardized weighting scheme.
  - (a) Provide the FIPS number for each assigned neighbor of Madison, Illinois (row 33 in the shapefile, FIPS = 17,119).
  - (b) Provide the numeric weight that Fayette, Illinois (row 27 in the shapefile, FIPS = 17,051) has been given for its neighbor, Shelby, Illinois (row 16 in the shapefile, FIPS = 17,173).

- (c) Provide the numeric weight that Shelby, Illinois (row 16 in the shapefile, FIPS = 17,173), has been given for its neighbor Fayette, Illinois (row 27 in the shapefile, FIPS = 17,051).

Open the crime dataset from Chicago (census tracts), which are stored in a shapefile named *foreclosures.shp*, for questions 3 through 7. Then, select a dependent variable that you will use for regression analysis, either *violent* (count of violent crimes) or *property* (count of property crimes). Then, select one or more covariates in the dataset that are from the U.S. Census 2007 to 2008: count of foreclosures (*est\_fcs*), number of mortgages (*est\_mtgs*), foreclosures divided by mortgages (*est\_fcs\_rt*), unemployment rate (*bls\_unemp*), and total population (*totpop*). When selecting these independent variables, keep in mind the guidelines for selecting covariates that you learned in prior chapters (e.g., be mindful of multicollinearity).

3. Conduct OLS regression.
  - (a) What are the standardized beta coefficients and corresponding *p*-values of the covariates in your model? Make sure to provide the variables names in your answer.
  - (b) Provide the log-likelihood for your model.
  - (c) Provide the AIC and BIC for your model.
4. Create a spatial weights matrix that is appropriate for polygons.
  - (a) Provide the type of spatial weights matrix you conducted and weighting scheme.
  - (b) Provide a summary of your spatial weights matrix connectivity.
5. Conduct the Lagrange Multiplier (LM) error and lag tests (nonrobust).
  - (a) What are the results of the LM tests?
  - (b) Based on the findings of the LM tests, which of the following is the appropriate next step and why? *Keep OLS regression results, conduct a spatial lag model, conduct a spatial error model, or conduct more diagnostic tests.*
  - (c) If more diagnostic tests are needed, conduct the robust LM-error and lag tests. What regression model do the results indicate that you should conduct and why?
6. Conduct the appropriate spatial regression model as indicated from question 5.

- (a) What are the standardized beta coefficients and corresponding *p*-values of the covariates in spatial regression model?
- (b) How do these standardized beta coefficients/*p*-values differ from that of your OLS regression model?
- (c) If you conducted a spatial lag regression model, provide an interpretation for rho.
- (d) Provide a direct interpretation of the statistically significant beta coefficients in the model.
- (e) Is the spatial regression model a better fit than the OLS regression model? Provide evidence from multiple goodness-of-fit indicators to support your answer.
7. Change the type and/or weighting scheme of your spatial weights matrix.
- (a) Rerun the LM-error and lag tests. Do the results of these tests differ from the weights matrix you specified in question 5? If so, how?
- (b) Conduct the same spatial regression model you conducted in 6 but use the new spatial weights matrix. How do the results differ from your findings in question 6?

## References

---

- Anselin, L. (2001). Spatial econometrics. In B. Baltagi (Ed.), *A companion to theoretical econometrics*. Malden, MA: Blackwell Publishers.
- Anselin, L. (2005). *Exploring spatial data with GeoDaTM: A workbook*. Santa Barbara, CA: Center for Spatially Integrated Social Science. Retrieved from <http://www.csiss.org>.
- Cressie, N. (1993). *Statistics for spatial data*. Hoboken, NJ: John Wiley & Sons.
- Groff, E. R., & Lockwood, B. (2014). Criminogenic facilities and crime across street segments in Philadelphia: Uncovering evidence about the spatial extent of facility influence. *Journal of Research in Crime and Delinquency*, 51(3), 277–314.
- Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(Supp 1), 234–240.
- Ward, M. D., & Gleditsch, K. S. (2008). *Spatial regression models*. Los Angeles, CA: Sage Publications.
- Weisburd, D., Lawton, B., Ready, J., & Haviland, A. (2011). *Longitudinal study of community health and anti-social behavior at drug hot spots*. Bethesda, MD: National Institute on Drug Abuse, National Institutes of Health. [Grant no. 5R01DA032639-03, 2012].

# Glossary

---

**Alternation** A type of quasi-random assignment in which researchers assign every other case to one particular group.

**Between effect** Effect of an independent variable on the dependent variable using the cluster as the unit of analysis—a regression of cluster-level averages across all the clusters included in the analysis.

**Between-subjects** In an ANOVA, a between-subjects factor or independent variable is one for which each subject or observation is in only one of the categories that makes up the factor or independent variable.

**Biased** Describing a statistic when its estimate of a population parameter does not center on the true value. In regression analysis, the omission of relevant independent variables will lead to bias in the estimate of  $Y$ . When relevant independent variables are omitted and those measures are related to an independent variable included in regression analysis, then the estimate of the effect of that variable will also be biased.

**Block randomization** A type of randomization whereby cases are first sorted into like groups and then afterwards randomly allocated into treatment and control conditions.

**Caliper** The bandwidth used for the selection of comparison cases in propensity score modeling. This ensures that matched cases have propensity scores that differ by no more than the caliper width.

**Cluster mean centering** Computed difference between the observed raw score on some variable for each observation in the sample and the cluster mean for that variable.

**Confounding factors** Variables associated with treatments and/or outcomes that can bias overall results if not controlled for statistically.

**Contrast coding** A method for recoding a multi-category nominal variable into multiple indicator variables (one less than the total number of categories), where the indicator category is coded as 1, the reference category is coded as -1, and all other categories are coded as 0. Contrast coding ensures that the sum of all the estimated effects for the indicator variable is equal to 0.

**Control group** The group that eligible cases are randomly assigned to which does not receive the treatment or the intervention being evaluated. In many criminological experiments, the control group may receive existing interventions in contrast to the innovative treatment.

**Correctly specified regression model** A regression model in which the researcher has taken into account all of the potential confounding variables that might account for the relationship between the dependent variable and independent variables of theoretical interest.

**Counts** A measure that reflects the number of events or some other countable entity. The values are discrete, whole numbers and may include 0 as a valid count, i.e., no observed events for a given observational unit.

**Cox and Snell's  $R^2$**  A commonly used pseudo  $R^2$  measure whose main component, as in other pseudo  $R^2$  statistics, is the log likelihood function ( $-2LL$ ).

**Cross-level interaction** An interaction effect included in a multilevel model between a level 1 independent variable and a level 2 cluster characteristic.

**Cumulative logistic probability function** A transformation of the logistic probability function that allows computation of the probability that  $y$  will occur, given a certain combination of characteristics of the independent variables.

**Derivative at mean (DM)** A measure that converts the nonlinear logistic regression coefficient to a simple linear regression coefficient, which may be interpreted as the change in  $y$  associated with a unit change in  $X$ .

**Design sensitivity** The statistical power of a research study. In a sensitive study design, statistical power will be maximized, and the statistical test employed will be more capable of identifying an effect.

**Dummy coding** A method for including nominal variables in a regression model that involves the creation of a series of binary 0/1 variables that indicate membership of an observation in each category of the nominal variable.

**Dummy variable** A binary nominal-level variable that is included in a multiple regression model.

**Effect size** A statistical index that encodes the findings of a study in a way that allows for comparison across studies via meta-analysis. Common effect sizes include the standardized mean difference, the correlation coefficient, the odds ratio, and the risk ratio.

**Eligibility pool** Participants or units that are eligible for an experiment.

**Euclidean distance** Straight-line distance between two points (as the crow flies).

**Exposure** A variable that reflects the exposure for a count. This might be the length of time, such as a

minute or year during which the counts were made, the population base for the counts, the size of a geographic area, or some combination of these. An exposure variable is the denominator in the ratio that converts a count into a meaningful rate, such as crimes per 100,000 persons per county per year.

**Factorial design** A factorial design is a type of experiment with two or more factors. Each factor is an independent variable with two or more categories or conditions.

**Fixed-effect model** A fixed-effect meta-analysis model assumes that each study contributing to a meta-analysis shares a common underlying true population effect.

**Fixed effects** A descriptive label for the regression coefficients ( $b_k$ ) estimated in a model with random effects. Fixed effects represent the average effects of the independent variables on the dependent variable across all individuals and clusters in a multilevel model.

**Forest plot** A forest plot is a graphic display of effect sizes with their confidence intervals, along with the overall results of a meta-analysis.

**Fully balanced** A fully balanced factorial design has an equal number of observations in each cell of the design, that is, the sample size is the same for each condition or combination of factor categories.

**Funnel plot** A type of scatterplot used to assess for publication selection bias. This plot shows the relationship between an effect size and its standard error. Asymmetry in this plot is evidence of publication selection bias.

**Gamma ( $\gamma$ )** The standardized measure of bias used to assess how sensitive PSM models are to excluded variables.

**Grand mean centering** Computed difference between the observed raw score on some variable for each observation in the sample and the overall sample mean for that variable.

**Group allocation** In criminological experiments, eligible cases are randomly assigned to two or more groups—typically treatment or control.

**Heteroskedasticity** A condition in which the variance of an error term is unequally distributed.

**Homoscedasticity** An assumption of multiple regression. When this assumption is met, the error variance is equal across all combinations of the independent variables.

**Incident rate ratio (IRR)** The incident rate ratio is an effect size for interpreting Poisson and negative binomial regression coefficients. It is the exponent of the coefficient that reflects the rate at which the count is increasing or decreasing for every one-unit change in the independent variable.

**Interaction effect** An interaction effect is present when the effect of one independent variable on the dependent variable is conditional on the level of a second independent variable.

**Internal validity** Whether the research design has allowed for the impact of the intervention or the treatment to be clearly distinguished from other factors.

**Intraclass correlation** A measure of association that measures the level of absolute agreement of values within each cluster.

**Iteration** Each time we identify another tentative solution and reestimate our logistic regression coefficients.

**k-nearest neighbors** A way to define neighbors, where  $k$  refers to the number of nearby neighbors to assign.

**Lack of convergence** Failure of a logistic regression analysis to reach a result that meets the criterion of reduction in the log likelihood function.

**Likelihood ratio chi-square test** A test for statistical significance that allows the researcher to examine whether a subset of independent variables in a logistic regression is statistically significant. It compares  $-2LL$  for a full model to  $-2LL$  for a reduced model.

**Log likelihood function** A measure of the probability of observing the results in the sample, given the coefficient estimates in the model. In logistic regression, the log likelihood function ( $-2LL$ ) is defined as  $-2$  times the natural logarithm of the likelihood function.

**Logarithm** The power to which a fixed number (the base) must be raised to produce another number.

**Logistic model curve** The form of the predicted outcomes of a logistic regression analysis. Shaped like an S, the logistic curve begins to flatten as it approaches 0 or 1, so it keeps coming closer to—but never actually reaches—either of these two values.

**Logistic regression analysis** A type of regression analysis that allows the researcher to make predictions about dichotomous dependent variables in terms of the log of the odds of  $Y$ .

**Logistic regression coefficient** The coefficient  $b$  produced in a logistic regression analysis. It may be interpreted as the change in the log of the odds of  $y$  associated with a one-unit increase in  $X$ .

**Manhattan distance** Distance between two points that is measured on a road network (similar to a car moving through a city).

**Maximum likelihood estimation** A technique for estimating the parameters or coefficients of a model that maximizes the probability that the estimates obtained will produce a distribution similar to that of the observed data.

**Mean squares** In ANOVA, a mean square is a variance associated with each factor or element of a research design.

**MHbounds (Mantel and Haenszel bounds) test** A test for assessing how sensitive PSM results are to the bias of excluding key (unmeasured) measures in the selection model.

**Model chi-square** The statistical test used to assess the statistical significance of the overall logistic regression model. It compares the  $-2\text{LL}$  for the full model with the  $-2\text{LL}$  calculated without any independent variables included.

**Moderator analysis** Moderator analysis is the examination of the relationship between study features and effect sizes. Two main types are the analog-to-the-ANOVA, which is similar to a one-way ANOVA model, and meta-regression.

**Moran's Index** A coefficient ranging from  $-1$  (dispersion) to  $1$  (clustered) that quantifies spatial autocorrelation (also known as Moran's I).

**Multicollinearity** Condition in a multiple regression model in which independent variables examined are very strongly intercorrelated. Multicollinearity leads to unstable regression coefficients.

**Multilevel data** Sample data where individual observations (level 1 data) are clustered within a higher-level sampling unit (level 2 data).

**Multinomial logistic regression** A statistical technique to predict the value of a dependent variable with three or more categories measured at the nominal level of measurement.

**Multiple regression** A technique for predicting change in a dependent variable, using more than one independent variable.

**Nagelkerke  $R^2$**  A pseudo  $R^2$  statistic that corrects for the fact that Cox and Snell's estimates, as well as many other pseudo  $R^2$  statistics, often have a maximum value of less than  $1$ .

**Natural logarithm of the odds of  $y$  (logit of  $y$ )** The outcome predicted in a logistic regression analysis.

**Nearest neighbor matching** An approach to matching in PSM that matches the treatment case to the not-treated sample case with the smallest probability distance.

**Negative binomial regression** A variant on Poisson regression that accounts for observed over-dispersion. This method estimates an over-dispersion parameter as part of the model and will produce regression coefficients that may differ from a Poisson model.

**Nonlinear relationship** Relationship between the dependent and the independent variable that is not captured by a straight line (linear) relationship.

**Odds ratio [Exp(B)]** A statistic used to interpret the logistic regression coefficient. It represents the

impact of a one-unit change in  $x$  on the ratio of the probability of  $y$ .

**Offset** A term in a count-based regression model used to convert the counts into meaningful rates. The offset is the natural log of exposure.

**Ordinal logistic regression (proportional odds model)** A statistical technique to predict the value of a dependent variable with three or more categories measured at the ordinal level of measurement.

**Over-dispersion** Variability in a distribution of counts that is in excess of what would be expected for a Poisson distribution where the variance equals the mean.

**Parallel slopes assumption** In an ordinal logistic regression model, the effect of each independent variable is assumed to be constant across all categories of the dependent variable.

**Partial proportional odds model** An ordinal logistic regression model that allows the effects of one or more of the independent variables to vary across the levels of the ordinal dependent variable. Useful when the parallel slopes assumption is violated.

**Partially balanced** A partially balanced factorial design is balanced within each level of one factor relative to the other factor but unbalanced across the levels of the first factor.

**Percent of correct predictions** A statistic used to assess how well a logistic regression model explains the observed data. An arbitrary decision point (usually  $0.50$ ) is established for deciding when a predicted value should be set at  $1$ , and then the predictions are compared to the observed data.

**Poisson regression** A regression method for count-based dependent variables. This method makes the assumption that the counts, after accounting for variability due to the independent variables, are Poisson distributed. That is, this method does not adjust for over-dispersion.

**Post-test measures** Analyses conducted by the researcher to determine if the intervention had any impact on the outcome measures of interest.

**Propensity score** A single score that is the probability of a case receiving treatment given a set of measured covariates. This value is determined during the propensity score matching process.

**Propensity score matching (PSM)** A commonly used method to identify matched cases when the researcher cannot gain experimental data and wants to assess the impacts of outcomes. It models the mechanism of selection into treatment as a method for matching cases.

**Pseudo  $R^2$**  The term generally used for a group of measures used in logistic regression to create an approximation of the OLS regression  $R^2$ . They are

generally based on comparisons of  $-2\text{LL}$  for a full model and a null model (without any independent variables).

**Publication selection bias** Publication selection bias is the tendency for studies with statistically significant results to have a greater likelihood of being published and hence included in a meta-analysis, potentially upwardly biasing the results.

**Quasi-Poisson regression** A variant on Poisson regression that accounts for observed over- or under-dispersion. The regression coefficients will be the same as with a Poisson model but the standard errors are adjusted for observed dispersion in the data.

**Queen contiguity** A type of spatial weight that defines neighbors as polygons sharing at least one edge or vertex.

**Random coefficient model** A linear regression model that allows the intercept and the effect of at least one independent variable to vary randomly across cluster—random effects are included for the model intercept and at least one independent variable.

**Random-effects model** A random-effects model in meta-analysis assumes that there is variability in the true population effects being estimated by a collection of studies and incorporates this heterogeneity into the model.

**Random effects** A descriptive label for the random error terms included in a multilevel model that allow for variation across cluster from the sample average estimated in the fixed effects. Random effects are assumed to be normally distributed in most multilevel models.

**Random intercept model** A linear regression model that allows the intercept to vary randomly across cluster—random effects are included for the model intercept.

**Randomization** The process of randomly assigning members from the pool of eligible participants or units to the study conditions—often a treatment group and a control group.

**Randomized experiment** A type of study in which the effect of one variable can be examined in isolation through random allocation of subjects to treatment and control, or comparison, groups.

**Region of common support** Area of the propensity score distribution for which there is overlap between treatment and comparison cases measured before matching. This is usually identified through a visual inspection.

**Regression coding** A method for recoding a multi-category nominal variable into multiple indicator dummy variables (one less than the total number of categories), where the indicator category is coded as 1 and all other categories are coded as 0. The reference category does not have an

indicator variable and is coded as a 0 on all the indicator dummy variables.

**Rook contiguity** A type of spatial weight that defines neighbors as polygons sharing at least one edge.

**Sensitivity analysis** The running of multiple regression models that allow the researcher to see how different specifications (such as inclusion and exclusion of outliers) impact the model results.

**Spatial autocorrelation** The correlation of a variable's values over space, with strong autocorrelation indicating that nearby values of the variable are similar to one another.

**Spatial dependence** Similarity in nearby observations due to the influence of neighboring locations via interaction.

**Spatial error model (SEM)** A regression technique that corrects for spatial autocorrelation in the residuals (spatial heterogeneity) by incorporating spatial effects via adding a spatial coefficient within the error term of the model.

**Spatial heterogeneity** Large-scale, regional variation due to a process that varies across space.

**Spatial lag model (SAR)** A regression technique that corrects for a dependent variable with spatial dependence by incorporating spatial effects via adding a spatially lagged dependent variable as a covariate.

**Spatial weights matrix** An  $n \times n$  matrix that quantifies spatial relationships, where  $n$  is the number of geographic features.

**Standard absolute bias** A measure used to identify when a scaled variable in the PSM model is considered to be unbalanced.

**Standardized logistic regression coefficient** A statistic used to compare logistic regression coefficients that use different scales of measurement. It is meant to approximate Beta, the standardized regression coefficient in OLS regression.

**Standardized regression coefficient (Beta)** Weighted or standardized estimate of  $b$  that takes into account the standard deviation of the independent and the dependent variables. The standardized regression coefficient is used to compare the effects of independent variables measured on different scales in a multiple regression analysis.

**Statistical power** One minus the probability of a Type II error. The greater the statistical power of a test, the less chance there is that a researcher will mistakenly fail to reject the null hypothesis.

**Stratification matching** Treatment and comparison cases are stratified into 5–10 groups after trimming the cases to only include those within the region of common support.

**Systematic review** A set of systematic, documented, and replicable methods for

reviewing the literature (published and unpublished, on a topic). The methods include (1) a systematic search for all eligible studies; (2) detailed eligibility criteria; (3) systematic coding of study features; (4) an assessment of risk-of-bias; and (5) a credible method synthesizing findings across studies, such as through meta-analysis.

**Thresholds** Points that mark the limits of the underlying continuum measured by an ordinal variable.

**Tolerance** A measure of the extent of the intercorrelations of each independent variable with all other independent variables. Tolerance may be used to test for multicollinearity in a multiple regression model.

**Transformation** Dependent and independent variables can be mathematically transformed, such as by taking the natural logarithm, squaring, taking the square-root, etc. This can improve normality and/or be used to fit a nonlinear relationship.

**Treatment group** One group that eligible cases are randomly assigned to which receives the treatment or the intervention being evaluated.

**Variance components model** A one-way analysis of variance model that includes random effects for each cluster that assesses whether there is random

variation in the mean of the dependent variable across the clusters included in the analysis.

**Variance inflation factor (VIF)** A measure of the extent to which a variable of interest is highly intercorrelated with other variables in the regression equation. The VIF is used to test for multicollinearity in a regression equation and is the inverse of tolerance.

**Wald statistic** A statistic used to assess the statistical significance of coefficients in a logistic regression model.

**Within effect** Effect of an independent variable on the dependent variable within each cluster and then averaged across all clusters or groups included in the analysis.

**Within-subjects** In an ANOVA, a within-subjects factor or independent variable is one for which each subject or observation is in each of the categories that make up the factor or independent variable, such as with a repeated measure.

**ZIP, ZINB regression** Zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models deal with count data that have an excess of zeros relative to expectation under a Poisson or negative binomial model.

# Index

## A

- Akaike information criteria (AIC), 517, 519, 521–523, 529, 534  
Analog-to-the ANOVA, 476–478, 486, 490, 492–495  
Analysis of variance (ANOVA), 16, 44, 67, 281–283, 340–342, 344, 349, 352–355, 358–361, 379–389, 405–407, 410–415, 424, 476–478  
Assess spatial autocorrelation, 515, 516, 520

## B

- Between effect, 292, 294, 310  
Between-subjects, 378, 381–383, 385–389, 403, 406, 407  
Biased, 48, 51, 60, 111, 275, 421, 434, 458  
Binary logistic regression model, 189–191, 209, 213  
Binomial data, 238  
Bivariate regression coefficient, 374  
Block randomization, 403  
    benefit of equivalence, 391  
    degrees of freedom, 391, 392  
    effects of treatment, 394  
    equal probability, 390  
    equal randomization, 393  
    equivalence, 394  
    factorial design, 389  
    interactions, 397, 398, 400  
    Naïve random assignment, 390  
    statistical power, 395–397  
     $t$  distribution, 391  
    treatment outcomes, 392  
    two-way ANOVA, 392  
Bonferroni correction, 39, 57

- Brant test, 213–216, 218, 224, 226, 227, 229  
Breusch–Pagan test, 511

## C

- Caliper, 423, 424, 426, 427, 434–436, 439, 442, 443  
Caliper matching, 422, 423, 426, 435  
Cambridge–Somerville study, 389, 391, 392  
Campbell Collaboration, 453  
Central limit theorem, 30, 36, 57, 136  
Chi-square statistic, 170, 177, 204, 205  
Clustered data, 28, 275, 276, 284  
Cluster-level characteristics, 300–303  
Cohen’s  $d$ , 332, 337, 338, 342, 348, 349, 357, 358, 456, 458, 462–468, 487  
    converting to Hedge’s  $g$ , 456  
Confidence interval, 5, 6, 41, 52, 171, 182, 202, 213, 420, 470, 474, 475, 478, 481, 485, 502  
Confounding factor, 372–374, 376, 377, 403, 404  
Contrast coding, 310  
Control group, 369–371, 377, 378, 403, 404, 457, 464, 537  
Control variables, 48, 306, 308, 343, 356, 428  
Correctly specified regression model, 17, 24, 48, 51, 60  
Correlation coefficient, 456, 464, 466  
    converting between effect size indices, 462  
    effect sizes into correlations, 466  
    effect sizes into odds ratios, 465, 466  
    effect sizes into risk ratios, 466  
    Fisher’s  $Z$ -transformation, 488  
    small to modest correlations, 461  
Count-based dependent variable, 21, 235, 237, 239, 250

- Count-based regression approaches, 260  
 Cox and Snell's  $R^2$ , 168, 169, 177, 182, 185  
 Cross-level interaction, 301, 304, 310  
 Cumulative logistic probability function, 139–146, 174
- D**  
 Data entry errors, 40  
 Data-generating process, 29  
 Derivative at mean (DM), 156, 157, 174, 175  
 Design sensitivity, 337, 347, 348  
 Discriminant analysis, 5  
 Discriminant functions, 5  
 Dummy coding, 75, 76, 112, 113, 118, 120, 122, 278  
 Dummy variable, 45, 53, 55, 61, 93–96, 99, 301
- E**  
 Effect size (ES), 241, 327, 331–335, 338, 340–344, 348, 359, 469–470, 473, 476, 481–486  
 cause-and-effect, 456  
 $Cohen's d$ , 462  
 converting between effect sizes, 462–467  
 correlation coefficient, 456, 461  
 correlations, 466–467  
 effect sizes, 456  
 meta-analysis, 456, 467  
     fixed-effect models, 467, 468  
     random-effect models, 467  
 odds ratio, 457  
 risk ratio, 457, 459–461  
 simple point estimates, 457  
 standardized mean difference, 456–458  
 statistical dependency, 480  
 Eligibility pool, 368, 369, 402, 404  
 Euclidean distance, 505, 519, 524, 530
- F**  
 Factorial design, 403, 404  
 laboratory-based studies, 381  
 mixed within-and between-subjects, 385, 386, 388, 389  
 participants, 381  
 program evaluation research, 381  
 two-way ANOVA, 381, 383  
 Fisher's Z-transformed correlation, 468, 488  
 Fixed effects, 310, 477, 492, 495
- Fixed-effect meta-analysis  
 heterogeneity, 470, 471  
 mean effect size, 469  
 multiple dependent variables, 469  
 random-effects model, 471–473  
 statistical significance, 470  
 Fixed-effect models, 467–469, 472, 476, 485, 486, 488, 492  
 Focal independent variables, 41  
 Forest plot, 474, 475, 485, 486, 494, 495  
 $F$ -test, 88, 243, 380, 386, 398, 405–407, 477  
     and ANOVA, 44  
     and multiple regression model, 62  
     in a regression model, 63  
     for a subset of variables, 67, 68, 70  
 Funnel plot, 483–486, 494
- G**  
 Gamma, 432, 433, 436  
 Generalized least squares, 136  
 Generalized linear model (GLM), 13, 136, 264–266, 379, 418  
     common variants, 12  
     independent variable, 8  
     maximum likelihood methods, 12  
     OLS regression model, 12  
     regression analyses, 2  
     statistical methods, 2  
 Grand mean centering, 289–291  
 Group mean centering, 289
- H**  
 Heterogeneity testing, 470, 471  
 Heteroscedasticity, 34, 35, 52, 56, 237, 501, 511, 523  
 Homoscedasticity, 33, 34, 56, 61, 135, 136, 146  
     of errors, 31
- I**  
 Imaginary populations, 29  
 Incidence rate ratio (IRR), 241  
     definition, 242, 261  
     fictitious Poisson model, 242  
     Poisson coefficient, 242  
     regression coefficients, 254  
 Independence, 28, 59, 61, 379, 424, 502, 511  
 Individual regression coefficient, 40, 41, 174  
 Interaction effects, 76, 290, 302, 304, 398, 399  
     additive effect, 92

- dependent variable, 112  
 dummy variable, 93–96, 99, 112  
 independent variable, 112  
 interpretation, 92, 112  
 multiplicative effect, 92  
 scaled variable, 93–96, 99, 112  
 two scaled variables, 99, 101–103, 105  
 Internal validity, 375–377, 402, 404  
 Intervention effect, 371  
 Intraclass correlation, 284, 310, 312, 313  
 Iteration, 145–148, 174, 292
- J**  
 Jersey City Drug Market Experiment (JCE), 393–395, 397, 401
- K**  
 $k$ -nearest neighbors, 505, 524  
*Koenker–Bassett* test, 511
- L**  
 Lack of convergence, 5, 146, 174  
 Lagrange multiplier (LM) tests, 516, 517, 520–522, 527, 528, 532, 534  
 Least squares regression, 343, 344, 346  
 Likelihood ratio chi-square test, 172, 174  
 Likelihood ratio (LR) test, 170, 202, 203, 250, 258, 269, 288, 517, 522, 533  
 Linearity assumption, 31, 34, 56  
 Local indicator of spatial association (LISA), 507  
 Logarithm, 88, 139, 145, 147, 151, 192, 198, 281  
 Logistic model curve, 136, 138, 173, 175  
 Logistic regression, 4, 5  
     analysis, 136  
     cumulative logistic probability function, 139–146  
     curve, 137, 138  
     dependent variable, 139  
     equation, 138  
     OLS regression methods, 136  
     R, 183, 184  
     SPSS, 181  
     Stata, 182  
     transformation, 138  
 Logistic regression analysis, 136, 137, 149, 173, 178, 201, 205, 461  
 Logistic regression coefficient, 148, 174  
     binary dummy variables, 158  
     Compstat model, 162, 163  
     derivative at mean (DM), 156, 157
- independent variables, 151  
 numeric interpretation, 151  
 odds ratio, 151–156  
 percent of correct predictions, 166, 167  
 probability estimates, 158–161  
 pseudo- $R^2$ , 168, 169  
 scales, 158  
 standardized, 164–166  
 statistical significance, 169–173  
 statistics, 158  
 variables, 151
- Logistic regression equation, 141  
 Log-likelihood, 145, 148, 168, 174, 202, 223, 253, 286, 288, 299, 517
- M**  
 Manhattan distance, 505, 516, 524  
 Matching approaches, 418  
 Maximum likelihood estimation, 12, 145, 174, 288, 511, 512, 517  
 Mean squares, 380, 404, 406, 407  
 Measurement error, 32, 43, 335, 419  
 Meta-analysis, 14, 453–485  
     to classic narrative review methods, 455  
     effect size (*see* Effect size)  
     forest plot, 474–475  
     history, 454  
     homogeneity testing, 470–471  
     moderator analysis, 475–480  
     meta-regression, 478–479  
     statistical significance, 455  
     systematic review methods, 453  
 Meta-regression, 476, 478–480, 482, 493, 494  
 MHbounds test, 431, 432, 436  
 Model assumptions, *see* Multiple regression model  
 Model building, 51, 52  
 Model chi-square, 170, 174, 177, 180, 204, 205  
 Model specification, 50, 61  
 Moderator analysis, 467, 469, 475, 476  
     analog-to-the ANOVA, 476–478  
     meta-regression, 478–480  
     study features and effect sizes, 486  
     variability in effect sizes, 485  
 Moran's I (Moran's Index), 506–510, 516, 520, 522, 523, 525  
 Multicategory nominal variable, 147, 172, 174, 207  
 Multicollinearity, 76, 109–111, 113, 118, 421, 534  
 Multilevel data, 275, 276, 292, 310

- Multilevel models (MLMs), 309, 388, 389, 501  
 cluster-level characteristics, 300, 301, 303  
 multilevel structure of the data, 277  
 simple (*see* Simple multilevel regression models)
- Multilevel negative binomial regression, 306, 308
- Multinomial logistic regression model  
 binary independent variable, 190  
 binary logistic regression model, 190  
 categories, 220  
 coefficients, 192, 220  
 dependent variable, 189, 192, 205  
 identity equation, 192  
 independent variable, 190, 220, 222  
 likelihood ratio (LR) test, 220  
 multiple independent variables, 189  
 outcome categories, 190, 205  
 outcome variable, 191  
 R, 228–230  
 significance levels, 193  
 simpler models, 189  
 SPSS, 225, 226  
 Stata, 226–228  
 transformation, 190
- Multiple coefficients, 171, 201–204, 220
- Multiple independent variables, 16, 27, 31, 75, 143, 189, 493
- Multiple OLS regression modeling, 45
- Multiple regression coefficient, 25, 27  
 and simple regression coefficient, 27
- Multiple regression methods, 4, 9, 12
- Multiple regression model, 16  
 assumptions  
   homoscedasticity of errors, 31  
   independence, 28–29  
   linearity assumption, 31  
   normally distributed errors, 29, 30  
 binary independent variable, 76  
 collinearity diagnostics, 120, 121  
 dependent variable, 75  
 descriptive statistics, 78  
 distributional differences, 79  
 dummy coding, 75, 76, 120  
 dummy variable, 53, 77, 79, 80  
 independent variables, 75  
 individual regression coefficient, 40, 41  
 interaction effects, 76  
 interaction terms, 119, 121  
 model fit and nested models, 41–44  
 multicollinearity, 76, 109–111  
 nonlinear relationships, 75  
 nonlinear terms, 118–121  
 OLS regression, 17
- outliers and influential cases, 38–40  
 R, 122, 123  
 reference category, 76, 77, 79  
 regression coefficient, 78  
 regression diagnostics, 33–35, 37  
 regression model, 77, 78, 80
- N**
- Nagelkerke  $R^2$ , 169  
 Natural logarithm of the odds of  $y$  (logit of  $y$ ), 138, 143, 174, 191  
 Nearest neighbor matching, 422, 423, 426, 434–436, 439  
 Negative binomial regression, 236, 239, 250, 261  
 count-based regression approaches, 260  
 GENLIN procedure, 264  
*MASS* package, 268  
 over-dispersion parameter, 251, 252  
 regression coefficients, 254  
 in Stata, 265  
 variant on Poisson regression, 261  
 ZINB models, 261
- Nominal independent variables, 12, 45, 112, 278, 378
- Nonexperimental methods, 371, 374
- Nonindependence, 28, 37, 38, 275, 287
- Nonlinear relationships, 75  
 coefficients, 86, 88  
 dependent variable, 88–91, 112  
 graphical assessment  
   dependent variable, 84  
   linear regression line, 84, 85  
   straight-line relationship, 84  
 hypothetical, 81  
 independent variable, 80, 112  
 interpretation of OLS regression, 80, 85, 86  
 quadratic equation, 81  
 statistical significance, 88  
 transformation, 84, 112
- Normally distributed errors, 29, 30, 36, 57
- Null hypothesis, 22, 40, 170, 173, 213, 214, 220, 322–324, 327–329, 331, 337, 339, 341, 343, 345, 347, 355
- O**
- Odds ratio, 151–156, 174  
 case-control designs, 460  
 computation, 460  
 effect size, 488  
 into a Cohen's  $d$ , 465  
 logged, 463, 464

- mean odds ratio, 481  
 and negative values, 481  
 retrospective case-control studies, 457  
 risk ratio, 460, 461  
 treatment comparisons, 474
- Offsets, 244–247, 261, 265–270
- Order of contiguity, 504
- Ordinal logistic regression  
 binary logistic model, 207  
 Brant test, 214, 215  
 coefficients, 209–212  
 cumulative probability, 208  
 equal intervals, 206  
 estimation approaches, 206  
 independent variables, 208  
 intercept-only model, 208  
 interpretation, 209, 210  
 logistic regression-type models, 208  
 multicategory nominal variable, 207  
 multinomial regression, 205  
 ordinal-level dependent variable, 205,  
 206  
 parallel slopes, 213, 220  
 partial proportional odds, 215, 216, 220  
 score test, 213, 214  
 statistical significance, 212, 213  
 statistical software programs, 208  
 thresholds, 206
- Ordinal logistic regression model,  
 206–210, 212, 214, 215, 220, 221,  
 223–226, 229–230
- Ordinary least squares (OLS) regression,  
 17, 41, 60, 189, 290, 510, 529, 530  
 additive model, 130  
 analysis tools, 129  
 binary dependent variable, 135  
 count data, 235  
 distribution, 135  
 homoscedasticity, 136  
 interval-and ratio-level measures, 130,  
 135  
 interval-and ratio-scale measures, 131  
 logical problem, 134  
 ordinal-and nominal-level variables, 130  
 regression approach, 135  
 regression error, 135
- Over-dispersion, 249–253, 261
- Poisson distribution, 236, 238, 249–251,  
 253, 255
- Poisson probability distribution, 236–238,  
 246, 251–253
- Poisson regression, 261  
 exposure and offsets, 244–246  
 GENLIN procedure, 264  
 IRRs, 241, 242  
 linear regression model, 239  
 logistic regression, 239  
 OLS and Poisson model, 240  
 OLS regression model, 239  
 over-dispersion in count data, 249, 250  
 risk levels, 240  
 significance testing, 243
- Poisson/negative binomial model, 250, 255
- Posttest measures, 370, 376, 403, 404
- Probability equation, 194, 195
- Propensity score, 436  
 logistic regression, 427  
 matching, 422  
 metric calipers, 423  
 PSM models, 421  
 in R, 441  
 regression models, 435  
 in Stata, 438  
 treatment and comparison groups, 426
- Propensity score matching (PSM), 13  
 advantages, 419  
 assessment, quality, 426, 427  
 final analysis, treatment effects, 422  
 gamma levels, 432  
 generalized linear model approach, 418  
 ignorability assumption, 431, 433  
 limitations, 433, 434  
 match treatment and comparison cases,  
 420  
 matching methods, 422–424  
 MHbounds test, 431, 432  
 multiple regression approaches, 418  
 selection mechanism, 421  
 in Stata, 438  
 steps, analysis, 419  
 treatment impact, 419, 420
- Pseudo- $R^2$ , 168, 169, 174, 176, 182, 517,  
 522, 523
- Publication selection bias, 482–486

**P**

- Parallel slopes assumption, 213–216, 218,  
 220, 221, 230, 231  
 Percent of correct predictions, 166–169,  
 174, 177  
 Point-biserial correlation, 464, 466

**Q**

- Quantile plot, 34, 35  
 Quantile–quantile plot (Q–Q plot), 34, 35,  
 38  
 Quasi-Poisson regression, 251, 252, 261,  
 264, 265, 267

- Quasi-Poisson regression (*cont.*)  
 coefficients, 252  
 errors, 251  
 negative binomial models, 253  
 negative binomial regression, 252  
 over-dispersion parameter, 251, 252
- Queen contiguity, 503, 504, 524, 526, 531, 533
- R**
- Random coefficient model, 295, 310  
 definition, 295  
 development, 295  
 variance of the random effects, 296
- Random effects, 278, 279, 284–288, 291, 295, 296, 302, 310, 315–317, 319, 471–473, 485, 493
- Random-effect models, 467, 471–473
- Random intercept model, 287, 295, 310  
 between and within effects, 292, 293  
 centering independent variables, 288–290  
 independent variables, 287  
 random intercept, 288  
 regression coefficients, 287  
 statistical significance, 287, 288  
 testing for between and within effects, 294
- Randomization, 369, 370, 374, 378, 394, 403, 404
- Randomized experiments, 13, 24, 60  
 associated statistical methods, 377, 378  
 covariates, 400–402  
 internal validity, 375–377  
 isolating causal effects, 371–375  
 R, 412, 413  
 selected design types, 377, 378  
 SPSS, 409, 411  
 Stata, 411, 412  
 statistical power, 368, 400–402  
 symbols and formulas, 404–407  
 treatments, 368  
 two-group, 379
- Random sampling, 28, 32, 36
- Reference category, 193, 195
- Region of common support, 423, 424, 428, 434, 435, 438, 441, 447
- Regression coding, 311
- Regression diagnostics, 33–35, 37
- Regression modeling, 4, 16, 29, 32, 34, 50–52  
 measurement error, 32
- Research hypothesis, 214, 220, 324, 325, 328, 334, 336, 337, 347
- Risk ratio, 195, 242, 265, 457, 459–461, 465–468, 487  
 asymmetric nature, 460  
 binary outcome variable, 459  
 effect size, 487  
 logged risk ratio, 460  
 and odds ratio, 457, 460, 465, 467  
 ratio of probabilities, 459
- Rook contiguity, 503–505, 524, 526
- S**
- Scaled variable, 93–96, 99–105, 112, 301, 388, 428, 465
- Schwarz Bayesian information criterion, 517
- Selection mechanism, 421, 435  
*See also* Propensity score matching (PSM)
- Sensitivity analysis, 40, 61
- Significance testing, 39, 51, 243, 334, 455
- Simple multilevel regression model, 277–279  
 fixed and random effects, 279  
 fixed-/random-effects model, 284, 285  
 intraclass correlation and explained variance, 283, 284  
 statistically significant, 285
- Simple random coefficient model, 301
- Simple random sampling, 28, 36
- Simple regression model  
 comparing regression coefficients, 46, 47  
 correlation coefficient, 21  
 definition, 18  
 independence, 28  
 independent variable, 18, 59  
 prediction error/residual, 21  
 regression coefficients, 18, 19, 21  
 slope and intercept, 18  
 visual inspection, 20
- Simultaneity bias, 512
- Single coefficients, 201, 203, 213, 214, 243
- Slope coefficient, 295–300, 309
- Spatial autocorrelation, 500, 502, 506, 520, 523
- Spatial data, 501, 511  
 OLS regression, 501, 502  
 order of, 504  
 queen, 503, 504  
 regression models, 501  
 rook, 503, 504  
 spatial autocorrelation, 500, 502  
 spatial dependence, 500  
 spatial error model, 502

- spatial heterogeneity, 500  
spatial landscape, 500  
spatial process, 500  
spatial theory, 501  
statistical tools, 501  
Spatial dependence, 500–502, 512–514, 516, 523, 528  
Spatial error models (SEMs), 502, 511, 514, 516, 523  
Spatial heterogeneity, 500, 514, 523, 528  
Spatial lag models (SAR), 511–514, 516, 517, 521–523  
Spatial regression  
    Anselin’s decision process, 521  
    autocorrelated spatial errors, 511  
    contiguity-based spatial weights matrix, 531, 532  
    diagnostic tests, 517  
    distance-based spatial weights matrix, 530  
    equal variance, 510  
    heteroscedasticity, 511  
    lagrange multiplier diagnostics, 532  
    least squares of error, 510  
    LM-error, 517  
    maximum likelihood estimation, 512  
    model diagnostics, 522  
    Moran’s I test, 532  
    OLS, 510, 518, 529, 530  
    R, 528, 529  
    simultaneity bias, 512  
    spatial autocorrelation, 520  
    spatial lag regression model, 521, 532  
    weights matrix, 519, 520  
Spatial relationships  
    binary weighting strategy, 505  
    complete spatial randomness, 507  
    contiguity types, 504  
    cross-product, 509  
    Euclidean distance, 505  
    inverse weighted distance, 506  
     $k$ -nearest neighbors, 505  
    Manhattan distance, 505  
    mean crime deviations, 508  
    Pearson’s correlation coefficient, 508  
    polygons, 503  
    queen contiguity, 503, 504  
    rook contiguity, 503  
    row-standardized, 505  
    spatial autocorrelation, 506, 507  
    spatial pattern, 508  
    spatial units, 507  
    spatial weights matrix, 502  
    weighting scheme, 506  
Spatial weights matrix, 502, 506, 508, 513, 514, 516, 519, 520, 524, 530–532  
Standard absolute bias, 428, 436, 437  
Standardized coefficient, 46, 63, 66, 68, 69, 164, 373  
Standardized logistic regression  
    coefficient, 164–166, 168, 176, 177, 180  
Standardized mean difference, 338, 456–458  
Standardized regression coefficients (betas), 46–48, 54, 61, 63  
    for SPSS, 66  
    in R, 69  
    in Stata, 68  
Standardized residuals, 68  
Statistical inference  
    multiple coefficients, 202–204  
    overall model, 204, 205  
    single coefficients, 201  
Statistical power, 13  
    analysis of variance, 336  
    ANOVA, 340, 341, 352–355  
    Cohen’s  $d$  effect size, 348, 349  
    components, 336  
        directional hypotheses, 328  
        effect size (ES), 331–334  
        sample size, 329, 330  
        statistical significance, 327  
    computation, 337  
    correlation coefficient, 342, 355, 356  
    design sensitivity, 348  
    effect size, 337, 347  
    elements, 336  
    implications, 325  
    least squares regression, 343, 344, 346  
    level, 326  
    means test, 338, 339  
    noncentral distribution, 324, 337  
    nondirectional test, 336  
    null hypothesis, 322, 324  
    OLS regression, 356, 357  
R  
    ANOVA models, 358–361  
    correlation coefficients, 361–363  
    two-sample difference of means tests, 357, 358  
    research design, 324  
    retrospective/post-hoc, 335  
    sample size, 335, 347  
    Stata, 350, 351  
    statistical significance, 322  
    statistical software tools, 335  
    Type I and Type II errors, 323–325  
    Type II error, 322  
Statistical significance, 22, 40, 44, 88, 169, 170, 172, 174, 201, 203, 204, 212–213, 243, 285, 287–288, 299, 318, 322, 323, 327, 391, 401, 453, 455, 456

- Stratification matching, 422, 424, 435, 437  
 Studentized residual, 39, 57  
 Superpopulation, 28  
 Systematic review, 473, 485  
     essential features, 453
- T**  
 Theory-testing context, 41  
 Tolerance, 109–111, 113, 114, 120, 121, 123–125  
 Traditional multiple regression models, 419, 420  
 Traditional regression methods, 420  
 Transformation, 75, 81, 87, 88, 91, 99, 112  
 Treatment group, 323, 338, 340, 369, 371, 377, 378, 388, 404, 418, 420, 435, 457  
*t*-Tests, 16, 22, 31, 39
- U**  
 Unconditional multilevel model, 308
- V**  
 Variance components model, 279, 285–287, 309, 311–313
- Variance inflation factor (VIF), 110, 113, 114, 123  
 Visual inspection, 20, 39, 56, 483  
 Vuong test, 258
- W**  
 Wald statistic, 170, 171, 174, 176, 177, 182, 184, 201, 203, 210, 212, 213, 226, 229  
 White test, 511  
 Within effect, 292–294, 311  
 Within-subjects, 378, 381, 385, 386, 388, 389, 403, 404, 407
- Z**  
 Zero-inflated negative binomial (ZINB) model, 250, 255, 257, 258, 260, 261, 266  
 Zero-inflated over-dispersion, 250  
 Zero-inflated Poisson (ZIP) model, 250  
     negative binomial model, 257  
     nonzero-inflated model, 258  
     over-dispersion, 255  
 Poisson/negative binomial models, 260  
 regression models, 255  
 ZINB models, 261