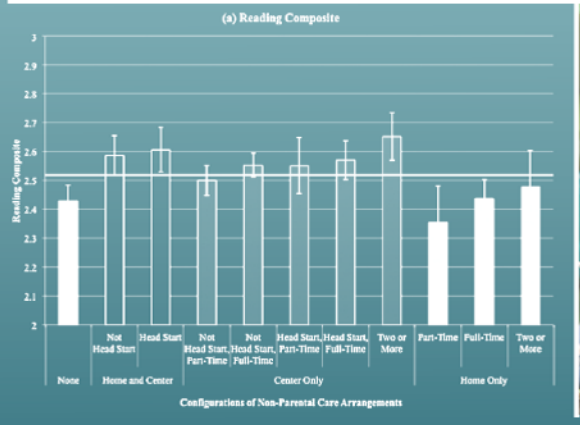
REGRESSION ANALYSIS FOR THE SOCIAL SCIENCES

SECOND EDITION



RACHEL A. GORDON

ROUTLEDGE

REGRESSION ANALYSIS FOR THE SOCIAL SCIENCES

This book provides graduate students in the social sciences with the basic skills that they need in order to estimate, interpret, present, and publish basic regression models using contemporary standards.

Key features of the book include:

- interweaving the teaching of statistical concepts with examples developed for the course from publicly available social science data or drawn from the literature;
- thorough integration of teaching statistical theory with teaching data processing and analysis using Stata;
- use of chapter exercises in which students practice programming and interpretation on the same data set; and course exercises in which students can choose their own research questions and data set.

Rachel A. Gordon is Professor in the Department of Sociology and Associate Director of the Institute of Government and Public Affairs at the University of Illinois at Chicago. Professor Gordon has multidisciplinary substantive and statistical training and a passion for understanding and teaching applied statistics.

TITLES OF RELATED INTEREST

Applied Statistics for the Social and Health Sciences by Rachel A. Gordon
Contemporary Social Theory by Anthony Elliot
GIS and Spatial Analysis for the Social Sciences by Robert Nash Parker and Emily K. Asencio
Statistical Modelling for Social Researchers by Roger Tarling
Social Statistics: Managing Data, Conducting Analyses, Presenting Results, Second Edition by Thomas J. Linneman
Principles and Methods of Social Research, Third Edition by William D. Crano, Marilynn B. Brewer, Andrew Lac
IBM SPSS for Intermediate Statistics, Fifth Edition by Nancy L. Leech, Karen C. Barrett, George A. Morgan
The Essence of Multivariate Thinking, Second Edition by Lisa L. Harlow
Understanding the New Statistics by Geoff Cumming

REGRESSION ANALYSIS FOR THE SOCIAL SCIENCES

Second Edition

Rachel A. Gordon
University of Illinois at Chicago

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

First published 2015
by Routledge
711 Third Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

Library of Congress Cataloging in Publication Data
Gordon, Rachel A.
Regression analysis for the social sciences / Rachel A. Gordon.—Second edition.
pages cm
Summary: "This book provides graduate students in the social sciences with the basic skills that they need in order to estimate, interpret, present, and publish basic regression models using contemporary standards. Key features of the book include: – interweaving the teaching of statistical concepts with examples developed for the course from publicly available social science data or drawn from the literature; – thorough integration of teaching statistical theory with teaching data processing and analysis using Stata; – use of chapter exercises in which students practice programming and interpretation on the same data set and course exercises in which students can choose their own research questions and data set" – Provided by publisher.
Includes bibliographical references and index.
1. Social sciences. 2. Regression analysis. I. Title.
H61.G578 2015
300—dc23
2014030561

ISBN: 978-1-138-81053-2 (hbk)
ISBN: 978-1-138-81251-2 (pbk)
ISBN: 978-1-315-74878-8 (ebk)

Typeset in Times New Roman
by RefineCatch Limited, Bungay, Suffolk

List of Trademarks that feature in the text

Stata	Microsoft Word
SAS	Adobe Acrobat
Microsoft Excel	Notepad
TextPad	DBMS/Copy
UltraEdit	SPSS
StatTransfer	R
LISREL	Minitab
AMOS	S-Plus
Mplus	Systat
EQS	TextEdit

Go to **www.routledge.com/cw/Gordon** for an invaluable set of resources associated with *Regression Analysis for the Social Sciences, Second Edition* by Rachel A. Gordon.

For instructors interested in expanding the coverage, *Regression Analysis for the Social Sciences* is also available with the following additional material: Basic Descriptive and Inferential Statistics and The Generalized Linear Model. For more information on ordering contact: saleshss@taylorandfrancis.com

PREFACE TO REVISED EDITION

This is a revision to *Regression Analysis for the Social Sciences* (2010). Major changes from the first edition include:

- Exclusive focus on Stata 13, and incorporation on new Stata commands, including accessible introductions to the `recode` command, factor variables, `margins`, and `marginsplot`.
- An "at your fingertips" summary of Stata syntax located inside the front and back covers of the book.
- Movement of all Stata examples from Appendices into the chapters, so they don't require flipping to the back of the book as you read.
- Inclusion of all analysis data sets for Stata examples, making it even easier for instructors and students to replicate results.
- New literature excerpts in Chapter 1, featuring recent studies published by graduate students and new scholars, including international studies.
- All new Chapter Exercises for homework problems, drawing on the National Household Interview Survey.

Like the first edition, this text is intended for graduate students in the social sciences. Both aimed to fill a gap in regression textbooks aimed at graduate students in the social sciences. We target the social science branches such as sociology, human development, psychology, education, and social work to which students bring a wide range of mathematical skills and have a wide range of methodological affinities. For many of these students, a successful course in regression will not only offer statistical content but will also help them to overcome any apprehensions about math and statistics and to develop an appreciation for how regression models might answer some of the research questions of interest to them.

To meet these objectives, the second edition, like the first, uses numerous examples of interest to social scientists including literature excerpts, drawn from a range of journals and a range of subfields, and examples from real data sets, including two data sets carried throughout the book (the National Survey of Families and Households; the National Health Interview Survey). The book also thoroughly integrates teaching of statistical theory with teaching data processing and analysis using Stata. We strategically choose which equations and math to highlight, aiming to discuss those that we do present slowly and deeply enough in order to reveal how they are relevant to the applied scholar.

THE IMPETUS FOR THIS TEXTBOOK

Since the publication of the first edition, modern computing power and data sharing capacities continue to change the landscape of social science research. Many large, secondary data sets are readily accessible to answer a host of research questions. There is increasing access to big data outside of academia as well, including by advocates, government officials, and citizens themselves. Well-designed studies, and succinct, clear presentations are required for results to stand out from a flood of information.

Instructors in graduate programs in the social sciences, however, have not always had access to a book aimed at their students' level, interests, and niche. Texts aimed at the undergraduate level often do not meet the goals and coverage of graduate sequences intended to prepare students to understand primary sources and conduct their own publishable research. These texts are sometimes used because they are at the right level for graduate students who are less mathematically inclined, but they do not fully satisfy the needs of graduate students and the faculty. Texts aimed at other disciplines are also problematic because they do not connect with students' substantive training and interests. For example, econometrics texts typically use economic examples and often assume more advanced mathematical understanding than is typical of other social science disciplines.

Like the first edition, this text aims to address this current landscape. The goal of the book is to provide graduate students with the basic skills that they need to estimate, interpret, present, and publish regression models using contemporary standards. Key features of the book include:

- interweaving the teaching of statistical concepts with examples developed for the course from publicly available social science data or drawn from the literature;
- thorough integration of teaching statistical theory with teaching data processing and analysis using Stata;

- use of chapter exercises in which students practice programming and interpretation on the same data set, and of course exercises in which students can choose their own research questions and data set.

THE AUDIENCE FOR THE BOOK

This book is designed for a semester-long course in graduate-level regression analyses for the social sciences. We assume that students will already have basic training in descriptive and inferential statistics, and some may go on to take other advanced courses (see Gordon 2012, for a year-long book that also covers these basics and some advanced topics, including maximum likelihood, logit, ordered logit, and multinomial logit).

Graduate regression courses typically occur in the first or second year of graduate study, following a course in basic descriptive and inferential statistics. The skills, motivations, and interests of students vary considerably.

For some students, anxiety is high, and this core sequence comprises the only statistics course that they plan to take. These students will become better engaged in the course if the concepts and skills are taught in a way that recognizes their possible math anxiety, is embedded in substantive examples, connects with the students' research interests, and helps them to feel that they can "do quantitative research." Part of the challenge of connecting with students' research interests, though, is that they are typically just starting their graduate programs when they take their statistics sequence, so the course needs to explicitly make connections to students' budding interests.

Other students in the course are eager to gain a deep understanding of statistical concepts and sophisticated skills in data management and analysis so that they can quickly move on to and excel in advanced techniques. Many of these students will come into their programs believing that quantitative research would be a major part of their career. Some want to use the skills they learn in the course to secure coveted research assistant positions. Many of these students enter the program with solid math skills, prior success in statistics courses, and at least some experience with data management and analysis. For these students, the course will be frustrating and unfulfilling if it doesn't challenge them, build on their existing knowledge and skills, and set them on a path to take advanced courses and learn sophisticated techniques.

Students also vary in their access to resources for learning statistics and statistical packages beyond the core statistics sequence. In some departments, strategies for locating data, organizing a research project, and presenting results in a manuscript are easily learned

from mentors and research teams (including through research assistantships) and through informal conversations with fellow students. Some programs also have separate "capstone" courses that put statistics into practice, typically following the core sequence. For other students, there are few such formal and informal opportunities. These students will struggle with implementing the concepts learned in statistics courses without answers to practical questions such as "Where can I find data?" "How do I get the data into Stata format?" "How do I interpret a codebook?" "How should I organize my files?" "How do I present my results in my manuscript?" Integrating this practical training within the core statistics sequence meets the needs of students (and faculty) in programs with few formal and informal learning opportunities for such practical skills. We also use this integrated approach in the book to help students practice the statistical concepts they are learning with real data, in order to help reinforce their learning, engage them in the course, and give them confidence in conducting quantitative research.

THE GOALS OF THE BOOK

The goals of the book are to prepare students to:

1. conduct a research project from start to finish using basic regression analyses;
2. have the basic tools necessary to be a valuable beginning research assistant;
3. have the basic knowledge and skills needed to take advanced courses that build on basic regression models; and
4. intelligently and critically read publications in top journals that utilize basic regression models.

We focus especially on concepts and techniques that are needed either to publish basic regression analyses in major journals in the relevant fields (for goals 1–3) or read publications using these models in those journals (for goal 4).

At every stage of the book, we attempt to look through the lens of the social scientist in training: Why do I need to know this? How is it useful to me? The book is applied in orientation and frequently makes concepts concrete through examples based on social science data and excerpts from recent journal publications.

Although the book is applied, we introduce key mathematical concepts aiming to provide sufficient explanation in order to accommodate students with weaker math backgrounds. For example, students are taught to find meaning in equations. Throughout the text, students are shown how to manipulate equations in order to facilitate understanding, with detailed in-text explanations of each step. The goal is to help all students feel comfortable reading equations, rather than leaving some to skip over them.

For more advanced students, or students returning to the book later in their careers, we provide references for additional details. We also attempt to present concepts and techniques deeply and slowly, using concrete examples for reinforcement. Our goal is for students to learn an idea or skill well enough that they remember it and how to apply it. This pace and approach allows sufficient time for students who struggle with learning statistics to "really get it" and allows sufficient time for students who learn statistics easily to achieve a more fundamental understanding (including references to more advanced topics/readings).

As part of this approach, we unpack ideas and look at them from multiple angles (again with a goal toward what is needed when preparing a manuscript for publication or reading a published article). For example, we spend considerable time on understanding how to test and interpret interactions (e.g., plotting predicted values, testing differences between points on the lines, calculating conditional slopes).

We assume that students have had an introductory course in research methods and in descriptive and inferential statistics, although we review concepts typically covered in these courses when we first use them (Gordon 2012 offers a more in-depth treatment of these topics).

THE CHAPTERS OF THE BOOK

The first part of the book introduces regression analysis through a number of literature excerpts and teaches students how to locate data, use statistical software, and organize a quantitative research project. The second part covers basic ordinary least squares (OLS) regression models in detail. The final chapter pulls together the earlier material, including providing a roadmap of advanced topics and revisiting the examples used in earlier chapters.

Part 1: Getting Started

Part 1 of the book aims to get students excited about using regression analysis in their own research and to put students on common ground by exposing them to literature excerpts, data sets, statistical packages, and strategies for organizing a quantitative research project. As noted above, this leveling of the playing field is important because students will vary in the prior statistics courses that they have taken and their prior experience of analyzing data as well as in opportunities in their program to learn how to put statistical concepts into practice.

- Chapter 1 introduces the basic ideas of regression analysis using four literature excerpts. By using a range of substantive applications and a range of data sources,

a major goal of the excerpts is to get students excited about applying regression analysis to their own work. In this chapter, the examples were also selected because they were completed when the authors were graduate students or new scholars, thus giving students attainable role models. The examples are also meant to begin to help students read and interpret published regression results (beyond their experiences reading articles that report regression analyses in substantive courses) and to preview some of the central topics to be covered in later chapters (e.g., controlling for confounds, examining mediation, testing for interactions) and others of which will be pointed to in the roadmap in the last chapter of the book (e.g., negative binomial models, propensity score models).

- Chapter 2 discusses strategies for organizing a research project. Especially with large secondary data sets with numerous variables, it is easy to get lost "playing with the data." To avoid this trap we encourage students to keep theoretical ideas and a long-range perspective in mind throughout a project. This chapter directly addresses the variability in formal and informal opportunities for research experiences mentioned above, and attempts to pull together various "words of wisdom" about planning and documenting a project and locating data that some students might otherwise miss. The chapter also exposes students to a breadth of secondary data sets, which can provide the knowledge and comfort needed to access secondary data as their interests develop over the years of graduate study. The chapter teaches students basic skills in understanding documentation for secondary data sources and selecting data sets. The data set carried throughout the in-text examples, the National Survey of Families and Households (NSFH), is introduced in the chapter.
- Chapter 3 introduces the basic features of data documentation and statistical software. The chapter begins with basic concepts of how data sets are stored in the computer and read by statistical packages. The rationale for using Stata is provided, along with its basic file types. The chapter also covers how to organize files in a project and how to identify relevant variables from large existing data sets. The display uses the data set carried throughout the in-text examples (NSFH).
- Chapter 4 teaches students how to write basic statistical programs. The chapter begins with the basics of the Stata interface and syntax. We then cover how to create new variables and to keep a subset of cases. The chapter ends with recommendations for organizing files (including comments and spacing) and for debugging programs (identifying and fixing errors).

Part 2: Ordinary Least Squares Regression with Continuous Outcome Variables

- Chapter 5 covers basic concepts of bivariate regression. Interpretation of the intercept and slope is emphasized through examining the regression line in detail,

first generally with algebra and geometry, and then concretely with examples drawn from the literature and developed for the course. We look at the formulas for the slope coefficient and its standard error in detail, emphasizing what factors affect the size of the standard error. We discuss hypothesis testing and confidence intervals for testing statistical significance and rescaling and effect sizes for evaluating substantive significance.

- Chapter 6 covers basic concepts of multiple regression. We look in detail at a model with two predictors, using algebra, geometry, and concrete examples to offer insights into interpretation. We look at how the formulas for the slope coefficients and their standard errors differ from the single predictor variable context, emphasizing how correlations among the predictors affect the size of the standard error. We cover joint hypothesis testing and introduce the general linear F-test. We again use algebra, illustrations, and examples to reinforce a full understanding of the F-test, including its specific uses for an overall model F-test and a partial F-test. We re-emphasize statistical and substantive significance and introduce the concepts of R-squared and Information Criteria.
- Chapter 7 covers dummy variable predictors in detail, starting with a model with a single dummy predictor and extending to (1) models with multiple dummies that represent one multicategory variable, and (2) models with multiple dummies that represent two multicategory variables. We look in detail at why dummy variables are needed, how they are constructed, and how they are interpreted. We present three approaches for testing differences among included categories.
- Chapter 8 covers interactions in detail, including an interaction between two dummy variables, between a dummy and interval variable, and between two interval variables. We present the Chow test and fully interacted regression model. We look in detail at how to interpret and present results, building on the three approaches for testing among included categories presented in Chapter 7.
- Chapter 9 covers nonlinear relationships between the predictor and outcome. We discuss how to specify several common forms of nonlinear relationships between an interval predictor and outcome variable using the quadratic function and logarithmic transformation. We discuss how these various forms might be expected by conceptual models and how to compare them empirically. We also show how to calculate and plot predictions to illustrate the estimated forms of the relationships. And, we also discuss how to use dummy variables to estimate a flexible relationship between a predictor and the outcome.
- Chapter 10 examines how adding variables to a multiple regression model affects the coefficients and their standard errors. We cover basic concepts of path analysis, including total, direct, and indirect effects. We relate these ideas to the concept of omitted variable bias, and discuss how to anticipate the direction of bias from omitted variables. We discuss the challenge of distinguishing between mediators and confounds in cross-sectional data.

- Chapter 11 encompasses outliers, heteroskedasticity, and multicollinearity. We cover numerical and graphical techniques for identifying outliers and influential observations. We also cover the detection of heteroskedasticity, implications of violations of the homoskedasticity assumption, and calculation of robust standard errors. Finally, we discuss three strategies for detecting multicollinearity: (1) variance inflation factors, (2) significant model F but no significant individual coefficients, and (3) rising standard errors in models with controls. We also discuss strategies for addressing multicollinearity based on answers to two questions: Are the variables indicators of the same or different constructs? How strongly do we believe the two variables are correlated in the population versus our sample (and why)?

Part 3: Wrapping Up

The final chapter provides a roadmap of topics that students may want to pursue in the future to build on the foundation of regression analysis taught in this book. The chapter organizes a range of advanced topics and briefly mentions their key features and when they might be used (but does not teach how to implement those techniques). Students are presented with ideas about how to learn these topics as well as gaining more skill with Stata (e.g., searching at their own or other local universities; using summer or other short course opportunities; using online tutorials). The chapter also revisits the Literature Excerpts featured in the first chapter of the book.

SOME WAYS TO USE THE BOOK

The author has used the complete textbook in a 15-week semester with two 75-minute lectures and a weekly lab session. Typically, chapters can be covered in a week, although a bit less time is needed for Part 1 (usually accomplished in the first two to three weeks) and extra time is often taken with the earliest chapters in Part 2 (two weeks each on the basics of bivariate regression, the basics of multiple regression, dummy variables, and interactions).

With a few exceptions, each chapter has a common set of materials at the end: key terms, review questions, review exercises, chapter exercises, and a course exercise.

- **Key Terms** are in bold within the chapter and defined in the glossary index.
- **Review Questions** allow students to demonstrate their broad understanding of the major concepts introduced in the chapter.
- **Review Exercises** allow students to practice the concepts introduced in the chapter by working through short, standalone problems.
- **Chapter Exercises** allow students to practice the applied skills of analyzing data and interpreting the results. The chapter exercises carry one example throughout

the book allowing students to ask questions easily of one another, the teaching assistant, and instructor as they all work with the same data set. The goal of the chapter exercises is to give students confidence in working with real data, which may encourage them to continue to do so to complement whatever other research approaches they use in the future.

- The **Course Exercise** allows students to select a data set to apply the concepts learned in each chapter to a research question of interest to them. The author has used this option with students who have more prior experience than the average student, who ask for extra practice because they know that they want to go on to advanced courses, or who are retaking the course as they begin to work on a masters or dissertation. The course exercises help students to gain confidence in working independently with their own data.

Answers to the review questions, review exercises, and chapter exercises, including the batch programs and results for the chapter exercises, are available to instructors on the textbook web site. The data sets, programs, and results from the in-text examples are also available on the textbook web site **www.routledge.com/cw/Gordon**.

This page intentionally left blank

ACKNOWLEDGMENTS

This book reflects many individuals' early nurturing and continued support of my own study of statistics, beginning at Penn State University, in the psychology, statistics, computer science and human development departments, and continuing at the University of Chicago, in the schools of public policy and business, in the departments of statistics, sociology, economics, and education, and at the social sciences and public policy computing center. I benefited from exposure to numerous faculty and peers who shared my passion for statistics, and particularly its application to examining research questions in the social sciences.

UIC's sociology department, in the College of Liberal Arts and Sciences, was similarly flush with colleagues engaged in quantitative social science research when I joined the department in 1999 and has provided me with the opportunity to teach graduate statistics over the last decades. This book grew out of my lecture notes for our graduate statistics sequence and benefits from numerous interactions with students and colleagues over the years. Kathy Crittenden deserves special thanks, as she planted the idea of this book and connected me with my publisher. I also have benefitted from interacting with my colleagues at the University of Illinois' Institute of Government and Public Affairs, especially Robert Kaestner, as I continued to study statistics from multiple disciplinary vantage points.

My publisher, Steve Rutter, was instrumental in taking me over the final hurdle in deciding to write this book and has been immensely supportive throughout the process of writing the first and second editions. He has ably provided advice and identified excellent reviewers for input as the book took shape. The reviewers' comments also importantly improved the book, including early reviews of the proposal, detailed reviews of all chapters of the first edition, and feedback from instructors who used the first edition. I also want to thank all of the staff at Routledge who helped

produce the book, especially Leah Babb-Rosenfeld, Mhairi Bennett, and Margaret Moore. Any remaining errors or confusions in the book are my own.

I also thank my husband, Kevin, and daughter, Ashley, for their support and for enduring the intense periods of work on the book.

Every effort has been made to trace and contact copyright holders. The publishers would be pleased to hear from any copyright holders not acknowledged here, so that this acknowledgement page may be amended at the earliest opportunity.

TABLE OF CONTENTS IN BRIEF

REGRESSION ANALYSIS FOR THE SOCIAL SCIENCES

PART 1: GETTING STARTED

PART 2: THE REGRESSION MODEL

PART 3: WRAPPING UP

TABLE OF CONTENTS IN DETAIL

REGRESSION ANALYSIS FOR THE SOCIAL SCIENCES

PART 1: GETTING STARTED

PART 2: THE REGRESSION MODEL

PART 3: WRAPPING UP

APPENDICES

This page intentionally left blank

Part 1

GETTING STARTED

This page intentionally left blank

Chapter 1

EXAMPLES OF SOCIAL SCIENCE RESEARCH USING REGRESSION ANALYSIS

CHAPTER 1: EXAMPLES OF SOCIAL SCIENCE RESEARCH USING REGRESSION ANALYSIS

> "Statistics present us with a series of techniques that transform raw data into a form that is easier to understand and to communicate or, to put it differently, that make it easy for the data to tell their story."
>
> Jan de Leeuw and Richard Berk (2004)
> Introduction to the Series
> Advanced Quantitative Techniques in the Social Sciences

Regression analysis,* a subfield of statistics, is a means to an end for most social scientists. Social scientists use regression analysis to explore research questions and to test hypotheses. This statement may seem obvious, but it is easy to get sidetracked in the details of the theory and practice of the method, and lose sight of this bigger picture (especially in introductory statistics courses). To help keep the big picture in sight, this chapter provides excerpts from the social science literature.

These excerpts also help to focus attention on how regression analyses are used in journal articles, consistent with two of the major reasons graduate students learn about regression analysis: (a) to be able to read the literature, and (b) to be able to contribute to the literature. You have surely already read at least some articles and books that report the results of regression analyses (if not, you likely will be doing so in the coming weeks in your substantive courses). Examining such literature excerpts in the context of a course on regression analysis provides a new perspective, with an eye toward how the technique facilitates exploring the question or testing the hypotheses at hand, what choices the researchers make in order to implement the model, and how the results are interpreted. At this point, you do not have the skills to understand fully the regression results presented in the excerpts (otherwise this book would not be needed!), so the purpose of this chapter is to present these features in such a way that they overview what later chapters will cover, and why. We will revisit these excerpts in the final chapter of the book, to help to reinforce what we have covered (and what advanced topics you might still want to pursue).

We have purposefully chosen excerpts in this chapter that were written by young scholars from a broad array of subfields and using different types of data sources.

* **Terms in color** in the text are defined in the glossary/index.

1.1: WHAT IS REGRESSION ANALYSIS?

Later chapters will develop the statistical details of regression analysis. But, in order to provide some guideposts to the features we will examine in the literature excerpts, it is helpful first to briefly consider what regression analysis is conceptually, as well as some of its key elements.

Why is it called regression, and why is it so widely used in the social sciences? The term regression is attributed to Francis Galton (Stigler 1986). Galton was interested in heredity and gathered data sets to understand how traits are passed down across generations. The data he gathered ranged from measures of parents' and children's heights to assessments of sweet peas grown from seeds of varying size. His calculations showed that the height or size of the second generation was closer to the sample average than the height or size of the first generation (i.e., it reverted, or regressed, to the mean). His later insights identified how a certain constant factor (such as exposure to sunlight) might affect average size, with dispersion around that group average, even as the entire sample followed a normal distribution around the overall mean. Later scientists formalized these concepts mathematically and showed their wide applicability. Ultimately, regression analysis provided the breakthrough that social scientists needed in order to study social phenomena when randomized experiments were not possible. As Stephen Stigler puts it in his *History of Statistics* "beginning in the 1880s . . . a series of remarkable men constructed an empirical and conceptual methodology that provided a surrogate for experimental control and in effect dissipated the fog that had impeded progress for a century" (Stigler 1986, 265).

Regression analysis allows scientists to quantify how the average of one variable systematically varies according to the levels of another variable. The former variable is often called a **dependent variable** or **outcome variable** and the latter an **independent variable**, **predictor variable**, or **explanatory variable**. For example, a social scientist might use regression analysis to estimate the size of the gender wage gap (how different are the mean wages between women and men?), where wage is the dependent variable and gender the independent variable. Or, a scientist might test for an expected amount of returns to education in adults' incomes, looking for a regular increment in average income (outcome) with each additional year of schooling (predictor). When little prior research has addressed a topic, the regression analyses may be exploratory, but these variables are ideally identified through theories and concepts applied to particular phenomena. Indeed, throughout the text, we encourage forward thinking and conceptual grounding of your regression models. Not only is this most consistent with the statistical basis of hypothesis testing, but thinking ahead (especially based on theory) can facilitate timely completion of a project, easier interpretation of the output, and stronger contributions to the literature.

An important advantage of regression analysis over other techniques (such as bivariate *t*-tests or correlations) is that additional variables can be introduced into the model to help to determine if a relationship is genuine or spurious. If the relationship is spurious, then a third variable (a confounder, common cause, or extraneous variable) causes both the predictor and outcome; and, adjusting for the third variable in a regression model should reduce the association between the original predictor of focal interest and outcome to near zero. In some cases, the association may not be erased completely, and the focal predictor may still have an association with the outcome, but part of the initial association may be due to the third variable. For example, in initial models, teenage mothers may appear to attain fewer years of schooling than women who do not have a child until adulthood. If family socioeconomic status leads to both teenage motherhood and academic achievement, then adjusting for the family of origin's education, occupation, and income should substantially reduce the average difference in school attainment between teenage and adult mothers. In Chapters 6 and 10, we will discuss how these adjustments are accomplished, and their limitations.

Such statistical adjustments for confounding variables are needed in the social sciences when randomized experiments cannot be conducted due to ethical and cost concerns. For example, if we randomly assigned some teenagers to have a child and others not to, then we could be assured that the two groups were statistically equivalent except for their status as teenage mothers. But, of course, doing so is not ethical. Although used less often than in the physical sciences to test basic research questions, experiments are more frequently used in certain social science subfields (e.g., social psychology) and applications (e.g., evaluations of social programs). When experiments are not possible, social scientists rely on statistical adjustments to **observational data** (data in which people were not assigned experimentally to treatment and control groups, such as population surveys or program records). Each literature excerpt we show in this chapter provides examples of using control variables in observational studies in an attempt to adjust for such confounding variables.

Regression models also allow scientists to examine the **mechanisms** that their theories and ideas suggest explain the association between a particular predictor variable and an outcome (often referred to as mediation). For example, how much of the wage gap between men and women is due to discriminatory practices on the part of employers and how much is due to differences in family responsibilities of men and women (such as child-rearing responsibilities)? If the mediators—discriminatory practices and family responsibilities—are measured, regression models can be used to examine the extent to which they help to explain the association between the focal predictor of interest—in this case, gender—and the outcome—wages. In Chapter 10, we will discuss how these mechanisms can be identified in regression models and some of the challenges

that arise in interpreting them. Some of the literature excerpts we discuss below illustrate the use of mediators.

Social scientists also use regression models to examine how two predictors jointly associate with an outcome variable (often referred to as moderation or interaction). For example, is the gender wage gap larger for African Americans than for whites? Are the returns to education larger for workers who grew up in lower versus higher income families? In Chapter 8, we will discuss how such research questions can be examined with regression models, and below we illustrate their use in the literature.

Regression models rest on a number of assumptions, which we will discuss in detail in later chapters. It is important for social scientists to understand these assumptions so that they can test whether they are met in their own work and know how to recognize the signs that they are violated when reading others' work. Although the basic regression model is linear, it is possible to model nonlinear relationships (as we discuss in Chapter 9) and important to check for cases that have a great effect on the slope of the regression line, referred to as outliers and influential observations (as we discuss in Chapter 11). Other technical terms that you may have seen in reading about regression models in the past, and that we will unpackage in Chapter 11, include the problems of heteroskedasticity and multicollinearity. Although these terms undoubtedly merely sound like foreign jargon now, we will spend considerable time discussing what these terms mean and how to test for and correct for them.

It is also helpful from the outset to recognize that this book will help you to understand a substantial fraction, but by no means all, of what you will later read (or potentially need to do) with regression models. Like all social scientists, you may need to take additional courses, engage in independent study, or seek out consultation when the research questions you pose require you to go beyond the content of this textbook. An important skill for a developing social scientist is to be able to distinguish what you know from what you need to know, and find out how to fill the gaps in your knowledge. We use the literature excerpts in this chapter, and the excerpts in future chapters, to begin to help you to distinguish between the two. One of our goals is to provide a solid foundation that adequately prepares you for such advanced study and help seeking. Chapter 12 provides a roadmap of advanced topics and ideas about how to locate relevant courses and resources.

In the interest of space and to meet our objectives, we extract only certain details from each article in the extracts below. Reading the entire article is required for a full picture of the conceptual framework, method and measures, results and interpretations.

1.2: LITERATURE EXCERPT 1.1

Reilly, David. 2012. "Gender, Culture, and Sex-Typed Cognitive Abilities." *PLOS ONE*, 7: e39904.

We begin with an example of a descriptive study in which graduate student David Reilly relied on the large, representative Programme for International Student Assessment (PISA) to examine gender gaps in students' reading, math, and science skills. The study, published in the open access journal *PLOS ONE*, does not present regression models per se, but offers a helpful example of the basic concepts of simple bivariate regression models which we will consider in Chapter 5.

Reilly lays out the rationale for his study in terms of the potential for variation across countries in gender achievement gaps and how these may shed light on the extent of ingrained or malleable gender differences. A wide range in gender gaps across cultures—especially results showing men outperforming women in some countries but women outperforming men in others—would suggest that environmental factors either produce gender differences or affect their expression.

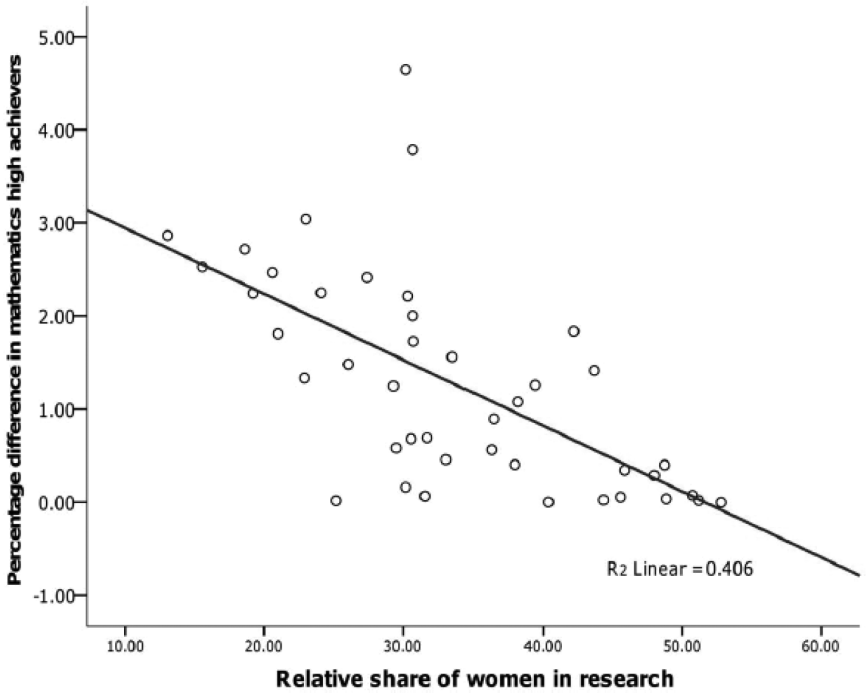
The PISA is sponsored by the OECD (Organisation for Economic Co-operation and Development) and has been repeated every three years since 2000. In 2009, the year analyzed by Reilly, nearly half a million 15-year-old students were surveyed in 65 OECD member and non-member nations.

Reilly calculated gaps between girls' and boys' reading, math, and science scores within each country, and also examined correlations between these gaps and other characteristics of the countries, including the share of women holding research positions. Although descriptive (not causal) his findings are intriguing. For instance, girls outperformed boys in reading in every country. In contrast, although boys outperformed girls in math and science in many countries, girls matched or outperformed boys in other countries.

The paper illustrates the type of careful descriptive analysis that can informatively begin any study and that provides an importance basis to any field of inquiry. For instance, Reilly's Figure 3, reproduced in Literature Excerpt 1.1a, shows a scatterplot between gender gaps in mathematics—in this case the ratio of boys to girls among high math achievers—and the relative share of women in research. Each circle indicates a country, and the downward sloping line represents the best-fitting regression line, which we will discuss in Chapter 5. The direction of the line indicates that among countries with a higher relative share of women in research, boys and girls are about equally represented among high math achievers. In contrast, in countries where relatively few

Literature Excerpt 1.1a

Figure 3. Relationship between women in research and gender ratios of high-achievers in mathematics literacy. doi:10.1371/journal.pone.0039904.g003



Source: Reilly, D. (2012). "Gender, Culture, and Sex-Typed Cognitive Abilities." *PLOS ONE*, 7: e39904.

women hold research positions, a higher share of boys than girls are found among high math achievers. We will discuss the value *R2 Linear* also listed on the graph in Chapter 6, which indicates that the R-squared value for the regression is 0.406, meaning that about 40% of the variation in the male:female gender ratio among high math achievers is explained by the relative share of women in research.

Examining such scatterplots between an outcome of interest and various predictors is an important starting point for any regression analysis. We demonstrate the utility of such graphs in Chapter 11, where we discuss testing the assumptions underlying regression analyses and issues related to their fit and specification. Reilly's Figure 3 illustrates one such issue—notice the point in the top middle of the graph that is positioned quite high in terms of the gender ratio; for this country, there is nearly a

five-fold ratio of boys to girls among high achievers in math. Such an outlying observation has the potential to greatly influence a regression line, with results possibly differing greatly when the outlying case is included versus excluded from the analysis. We will learn in Chapter 11 how to identify such cases and think through whether and how to adjust our analyses when they are found. In Reilly's case, the outlier's influence is diminished because it falls closer to the middle than either end of the distribution of the other variable, *Relative share of women in research*.

1.3: LITERATURE EXCERPT 1.2

Eng, Sothy. 2012. "Cambodian Early Adolescents' Academic Achievement: The Role of Social Capital." *Journal of Early Adolescence* 33(3): 378–403.

We now turn to an example of the application of regression analyses in a relatively small sample collected specifically to answer a new scholar's research questions. The paper by Sothy Eng relies on a survey of approximately two hundred Cambodian adolescents and their parents. The sample was carefully chosen to include students in Grades 5 and 6 of seven elementary schools, two located in urban and five located in rural areas. Because the author was interested in understanding factors that affected these early adolescents' academic achievement, he strategically chose students so that half came from the bottom and half came from the top half of the distribution of first semester grades within each school.

As Eng notes "empirical research related to Cambodian educational attainment and achievement is sparse" (p. 379) and he aims to help fill this gap by understanding how parents' attitudes and behaviors affected their children's school-related outcomes. He makes a case for the importance of the study, noting that Cambodia continued to recover from war and genocide in the 1970s, and that, despite recognition that investing in education might help the country rebuild and advance, educational attainment remains low and poverty widespread. For instance, he cites low literacy rates—just four-fifths of males and two-thirds of females were literate—and school enrollment rates—just over half of boys and two-fifths of girls were in school in late adolescence. The author argues that there is a need to pay more attention to family factors explaining educational attainment, including parents' beliefs about the importance of fate and traditional gender roles and their involvement and support of their children's education. He focuses on early adolescence because this time period has been identified as the point at which many students are vulnerable to dropping out of school.

Literature Excerpt 1.2a shows Eng's Table 3, which presents results from a regression model predicting students' grades as reported by school administrators (scored from

Literature Excerpt 1.2a

Table 3. OLS (Ordinary Least Square) Regression Predicting Academic Achievement From Social Capital Variables.

	Model 1			Model 2			Model 3		
	B	SE	β	B	SE	β	B	SE	β
R^2 Change (total)		.15			.06(21)			.00 (.21)	
F for change in R^2		6.31***			2.07*			0.30 (ns)	
Step 1 (control variables)									
Adolescent's gender (Female = 0, Male = 1)	–0.89	.22	–.27***	–0.82	.24	–.24***	–1.23	.84	–.37
Hours of extra classes	0.06	.03	.16*	0.05	.03	.16*	0.05	.03	.15*
Family wealth	–0.12	.09	–.10	–0.10	.09	–.08	–0.10	.09	–.08
Home-school distance	–0.38	.24	–.11	–0.38	.24	–.11	–0.39	.24	–.11
Education of father	0.04	.11	.02	0.09	.18	.04	0.10	.18	.04
Education of mother	0.52	.20	.19*	0.54	.21	.20*	0.53	.21	.20*
Step 2 (social capital variables)									
Length of residence				0.02	.01	.18*	0.02	.01	.18*
Number of son(s)				–0.07	.08	–.06	–0.07	.08	–.06
Number of daughter(s)				–0.04	.08	–.03	–0.04	.08	–.04
Parents' fatalistic beliefs				–2.21	.10	–.16*	–2.23	.99	–.16*
Parents' academic involvement				–0.33	.27	–.09	–0.35	.27	–.09
Parents' academic inspirations				0.37	.28	.08	0.38	.29	.09
Parents' gender role attitudes				–0.02	.22	.01	–0.10	.31	–.03
Step 3									
Parents' Gender							0.22	.43	.13
Role × Child's									
Gender									

Note: N = 202.
*p < .05. **p < .01. ***p < .001.

Source: Eng, Sothy. 2012. "Cambodian Early Adolescents' Academic Achievement: The Role of Social Capital." *Journal of Early Adolescence* 33(3): 378–403.

zero to 10). Since this is the first regression table we examine, let's consider its structure before we delve into its meaning. The title tells us that Eng used ordinary least squares regression, the type of analysis we cover in this book. The table has four main columns, with the first column primarily comprised of the names of predictor variables and the remaining columns containing results for three different models. Within each major heading (model), there are three smaller columns, labeled B, SE, and β. By convention, scholars understand that these symbols represent the regression coefficients (B for unstandardized and β for standardized) and their standard errors (SE). We will learn

how to calculate and interpret each of these in Chapter 5. The notes to the table tell us that the sample size (N) is 202. The asterisks indicate different p-values, and we will also review in Chapter 5 how to interpret these values and how to use them to make conclusions about hypothesis tests.

Returning to the interior of the table, let's consider the top two rows labeled R^2 *Change (total)* and *F for change in* R^2. These results offer an assessment of the overall quality of each model, as we will discuss in Chapter 6. The results tell us that the set of predictors in Model 1 and in Model 2 each help explain variation in students' grades, but the variable added to Model 3 in *Step 3* (see again first column) does not. This lack of significance in Model 3 is also evident by the absence of any asterisks on the numbers in the bottom row of the table. Within the first two columns, we can look at the asterisks to see which individual predictor variables significantly contribute to explaining grades. The variables with asterisks include the *Adolescent's gender (Female = 0, Male =1), Hours of extra classes, Education of mother, Length of residence*, and *Parents' fatalistic beliefs*.

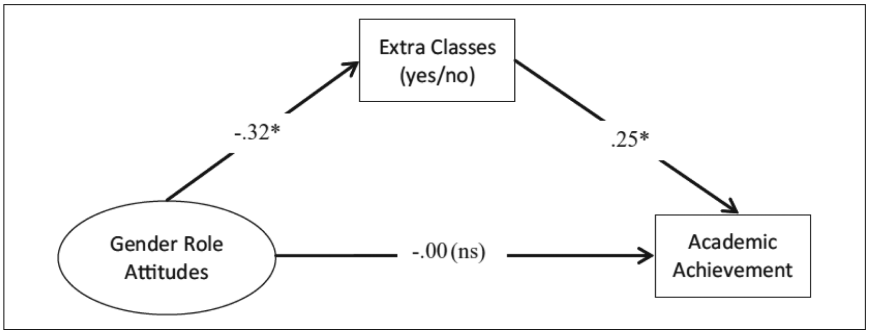
We will have much more to say about interpreting the substantive size and meaning of these significant results in later chapters. For now, we will draw out a few points. One is that the sign on the coefficients informs us about the direction of the associations. For instance, Eng indicates in the table that gender was coded such that boys were represented by the value one and girls by the value zero; as we will learn in Chapter 7, this lets us interpret the negative sign of the coefficient as meaning that boys' average grades are lower than girls', controlling for the other variables in the model. The magnitude of the association is close to one point in unstandardized form (–0.89 in Model 1, –0.82 in Model 2) and just over one-quarter point in standardized form (–0.27 in Model 1, –0.24 in Model 2). We will discuss these values in depth in Chapter 5. For now, suffice it to say that the unstandardized coefficient is interpreted relative to the outcome's natural units. To consider the substantive importance of the association, we might ask ourselves: Is nearly one point on the zero-to-ten scale of grades in these Cambodian schools large? The standardized coefficient can help in such substantive interpretations as well, relating this amount to the variable's standard deviation. Elsewhere in the paper, Eng tells us that grades average 6.23 with a standard deviation of 1.68. Based on this information, it appears that Eng is reporting the completely standardized coefficient in his table. We will see in Chapter 5 that a semi-standardized coefficient would relate the –0.89 value from Model 1 directly to the standard deviation of grades, revealing that boys average nearly half a standard deviation lower than girls in grades –0.89/1.68 = –0.53, a moderately sized association. We will also have more to say in Chapter 10 about how and why the coefficients change between Models 1 and 2 (i.e., why does the understandized coefficient for gender fall from –0.89 to –0.82?).

Of particular interest to Eng's substantive questions are the results for parents' fatalistic beliefs and gender role attitudes, which are added to the model in Step 2. In these results, we see a negative sign for fatalism, telling us that parents who hold more fatalistic beliefs have children with lower grades, on average after adjusting for the other variables in the model. Notice that whereas the unstandardized coefficient for fatalism, –2.21 in Model 2, is larger than the unstandardized coefficient we saw above for gender, the standardized coefficient for fatalism, –0.16, is smaller than the standardized coefficient for gender. We will discuss in Chapter 5 the ways in which the standardized coefficients help place different variables on "equal footing." In short, the unstandardized coefficient represents the expected change in the outcome variable for a one-unit increase in the predictor variable. Since one more can mean very different things across different variables (e.g., one more dollar of annual earnings is much smaller than one more bitcoin, at the time of this writing was valued at several hundred dollars). Scales like Eng's fatalism measure also lack intrinsic meaning. The standardized coefficient allows us a more unitless interpretation, telling us that one more standard deviation in parental fatalism is associated with about one-sixth of a standard deviation in lower grades.

Eng's Figure 1, shown in Literature Excerpt 1.2b, also provides an illustration of the concepts of path analysis that we will introduce in Chapter 10. Eng conducted additional

Literature Excerpt 1.2b

Figure 1. Structural model predicting early adolescents' academic achievement from parents' traditional gender role attitudes and adolescents' extra classes.



Notes: A model for girls ($n = 104$); oval-shaped *Gender Role Attitudes* indicates a latent construct. Fit indices: $\chi^2 = 22.04$, $df = 19$, $p = .28$, comparative fit index (CFI) = 0.95, Tucker-Lewis Index (TLI) = 0.93, root mean square error of approximation (RMSEA) = 0.039.
*$p < .05$.

Source: Eng, Sothy. 2012. "Cambodian Early Adolescents' Academic Achievement: The Role of Social Capital." *Journal of Early Adolescence* 33(3): 378–403.

analyses to demonstrate that, although parents' gender role attitudes were not directly associated with grades (the small value of –0.02 in Model 2, with no asterisk), his Figure 1 shows that such attitudes appear to be indirectly associated with grades through children's enrollment in extra classes (some Cambodian parents pay teachers to provide extra hours of tutoring to individual students). Of conceptual interest, this association was evident only for girls. The statistical evidence for differences between boys and girls was limited, however. We will discuss in Chapter 8 how to test for such interactions. Eng's Step 3 in Model 3 shows the kind of interaction term we will use, showing no interaction between parents' gender role attitudes and children's gender. Eng also reports in the text that although the pathways shown in Figure 1 were significant only for girls, *differences* between these pathways and the pathways estimated for boys were not significant. Many find such nuance difficult to understand, and we will have much more to say in Chapter 8 about what these results mean.

As Eng acknowledges in his discussion, it is also the case that his study has some weaknesses, including its relatively small size and cross-sectional, non-experimental design. Of particular importance to the results shown in Figure 1, the design is strengthened by the fact that grades were gathered from an independent source (school administrators) different from the predictors (parent reports). However, because both sources were interviewed at the same time, and since parents reported both their gender role attitudes and their children's extra tutoring, we cannot be sure that the direction of the arrows shown in Figure 1 is correct (it is possible, for instance, that children's high grades lead parents to invest in extra classes, or that some third variable, not measured by the study, predicts both parents' gender attitudes and willingness to pay for tutoring). We will have more to say, especially in Chapter 10, about the importance of such confounds and considerations of time ordering.

1.4: LITERATURE EXCERPT 1.3

Hawkinson, L.E., A.S. Griffen, N. Dong, and R.A. Maynard (2013). "The Relationship between Child Care Subsidies and Children's Cognitive Development." *Early Childhood Research Quarterly*, 28: 388–404.

We next consider a study based on analyses of a large, nationally representative survey of U.S. children. As we discuss in Chapter 2, numerous large-scale data sets are now available, following individuals over time. These data sets are particularly well suited to regression analyses because they are drawn to represent a population (often with a stratified, clustered design; we will consider some of these design implications in Chapters 11 and 12; see also Gordon 2012 for a fuller treatment of regression analyses with complex sampling designs). These public-use data sets often have large enough

sample sizes to allow the estimation of complicated models with numerous variables to measure both conceptually relevant constructs and important confounds.

In Literature Excerpt 1.3 we show how Laura Hawkinson and her colleagues took advantage of one such data set to examine the extent to which young children's cognitive development is associated with their families' receipt of subsidies to cover the cost of their child care. Importantly, the authors leverage the large data set and its numerous variables in order to gain greater purchase on the extent to which any observed associations might reflect causal mechanisms (although, as they recognize, their estimates are still not definitive evidence of causality, since the design is not experimental).

The data set that they use is the Early Childhood Longitudinal Study-Birth Cohort (ECLS-B), one of several large national data sets collected by the U. S. Department of Education. The ECLS-B began in 2001 by sampling approximately 14,000 newborns from vital statistics in a nationally representative set of communities across the United States. Importantly for Hawkinson and colleagues' interests, children were followed across time, with interviews and assessments at 9 months, 2 years, 4 years, and kindergarten. The focal predictor of interest to Hawkinson and her co-authors was parents' reports of whether they received help from a social service or welfare agency in paying for their 4-year-old (preschool-aged) child's care arrangements. The focal outcomes in the authors' study were children's scores on reading and math assessments conducted when they were in kindergarten. The time ordering of these assessments (subsidies during preschool, outcomes during kindergarten) helps the authors assure the time ordering needed for interpreting a potential causal association. They also estimate a series of alternative models to try to gain insight into this question of causality, adding and removing variables from the model and altering the way the model is estimated. For instance, the authors sometimes draw on preschool (age 4) assessments of children's reading and math as control variables, and they also use a technique called propensity score matching as an alternative way of identifying children who do and do not receive subsidies but are equivalent in other ways (we preview this technique in the roadmap to advanced techniques found in Chapter 12).

In Literature Excerpt 1.3a we reproduce Table 3 from Hawkinson and colleagues' paper. The structure of the table is similar to that seen in Eng's paper, although Hawkinson and colleagues build their models by including their focal predictor of interest *Preschool subsidy receipt* in Model 1 and then adding control variables, first preschool math and reading scores in Model 2 and then a broader set of controls in Model 3. The final Model 4 has the full set of controls, but the sample now comprises lower-income children rather than the full sample. They show two columns of results for each model, one each for math and reading scores. Their table note tells us that the values are

Literature Excerpt 1.3a

Table 3 Results of OLS regressions of subsidy receipt on kindergarten math and reading scale scores, main results with covariates.

	Full sample						Under 185% poverty	
	(1) No control variables		(2) Only preschool cognitive controls		(3) Full set of control variables		(4) Full set of control variables	
	Math score	Reading score	Math score	Reading score	Math score	Reading score	Math score	Reading score
Preschool subsidy receipt	–4.39**	–5.46**	–2.44**	–2.61**	–1.62**	–1.87*	–1.96*	–2.09*
	(0.86)	(1.21)	(0.72)	(1.01)	(0.58)	(0.87)	(0.66)	(0.93)
Pre-test measures of cognitive skills								
Preschool math score	–	–	0.45**	0.36**	0.46**	0.39**	0.46**	0.41**
			(0.026)	(0.035)	(0.023)	(0.033)	(0.035)	(0.052)
Preschool reading score	–	–	0.18**	0.52**	0.13**	0.47**	0.15**	0.43**
			(0.024)	(0.035)	(0.021)	(0.033)	(0.033)	(0.057)
2 years mental score	–	–	–	–	0.090**	0.048*	0.098**	0.050
					(0.015)	(0.022)	(0.025)	(0.032)
9 months mental score	–	–	–	–	0.0020	–0.010	0.0066	0.0021
					(0.014)	(0.021)	(0.020)	(0.030)
Child characteristics								
Birth weight[a]	–	–	–	–	0.54**	0.069	0.62*	0.083
					(0.18)	(0.28)	(0.27)	(0.42)
Percent male	–	–	–	–	0.57**	–0.63	–0.029	–0.91
					(0.26)	(0.39)	(0.40)	(0.59)
Percent black[b]	–	–	–	–	–0.73*	1.40*	–0.42	1.35
					(0.35)	(0.55)	(0.49)	(0.74)
Percent Hispanic[b]	–	–	–	–	–0.23	0.39	–0.16	0.25
					(0.39)	(0.64)	(0.57)	(0.84)
Percent other race[b]	–	–	–	–	0.20	1.19	0.10	0.71
					(0.42)	(0.69)	(0.57)	(0.98)
Percent started K in 2007	–	–	–	–	5.23**	5.83**	4.55**	4.85**
					(0.49)	(0.73)	(0.75)	(1.11)
Age (months) of K assessment	–	–	–	–	0.46**	0.67**	0.53**	0.72**
					(0.049)	(0.072)	(0.077)	(0.11)
Family characteristics at age 2								
Family income[c]	–	–	–	–	0.039	0.041	–0.0024	0.057
					(0.036)	(0.059)	(0.094)	(0.16)
Percent welfare recipients	–	–	–	–	–0.65	–2.24**	–0.67	–1.83*
					(0.55)	(0.78)	(0.59)	(0.82)

Percent WIC recipients	–	–	–	–	–0.45	–0.14	0.045	0.41
					(0.32)	(0.51)	(0.44)	(0.69)
Percent single parents	–	–	–	–	–0.64	–1.14*	–0.52	–0.88
					(0.35)	(0.57)	(0.45)	(0.69)
Mother's years of education	–	–	–	–	0.37**	0.52**	0.49**	0.98**
					(0.067)	(0.11)	(0.12)	(0.18)
Mother's age at child's birth	–	–	–	–	0.0052	–0.024	0.021	0.012
					(0.024)	(0.039)	(0.035)	(0.055)
Percent speak English at home	–	–	–	–	–0.19	–1.44*	–0.53	–2.14*
					(0.39)	(0.63)	(0.56)	(0.88)
Constant	44.9**	45.0**	27.0**	21.1**	–23.4**	–36.4**	–31.6**	–46.6**
	(0.19)	(0.27)	(0.54)	(0.72)	(3.76)	(5.69)	(5.98)	(8.45)
N	5650	5650	5650	5650	5650	5650	2450	2450
r^2	0.008	0.006	0.348	0.339	0.556	0.493	0.466	0.405
Cohen's d	0.31	0.27	0.17	0.13	0.12	0.09	0.15	0.11

Note: Coefficients are unstandardized and include standard errors in parentheses.
[a] Birthweight is measured in 1000 g units:
[b] referent group is white;
[c] income is measured in $10,000 units
* $p < 0.05$.
** $p < 0.01$.

Source: Hawkinson, L.E., A.S. Griffen, N. Dong, and R.A. Maynard (2013). "The Relationship between Child Care Subsidies and Children's Cognitive Development." *Early Childhood Research Quarterly*, 28: 388–404.

unstandardized coefficients and that standard errors are shown in parentheses. Asterisks indicate different *p*-value levels, as defined in the table notes. The values in the bottom rows of the table provide the sample sizes (smaller than the total study sample size primarily because only a subsample of children were followed into kindergarten, due to funding constraints) as well as the R-squared value and a measure called Cohen's *d* which relates to the effect sizes that we will learn about in Chapter 5.

The way the authors set up their series of models allows us to see the extent to which the coefficients for *Preschool subsidy receipt* change as more control variables are introduced into the model. For instance, the unstandardized coefficients for math scores fall from –4.39 in Model 1, to –2.44 in Model 2, to –1.62 in Model 3. As we discuss in later chapters (especially Chapters 6 and 10), these declines reflect the ways in which regression analyses help us narrow to unique effects of each variable, effects not shared by other variables in the model. These control variables help us adjust for confounds and get closer to causal estimates (although estimates are still never truly causal with non-experimental data). Doing so is important in Hawkinson and colleagues'

application, since confounding characteristics may be associated with whether parents decide to use child care subsidies and with the types of care arrangements that they use. Elsewhere in the paper, the authors reported additional sensitivity analyses that examined how robust the results were to various alternative ways of specifying the models. Such sensitivity analyses are an important part of careful regression modeling. For instance, the authors demonstrated that the results generally remained significant and of similar magnitude when they excluded children attending higher-quality programs (like Head Start) and when they excluded children who had received subsidies prior to the preschool assessment point.

Stepping back to think about the substantive meaning of the authors' findings, it may seem surprising that receipt of subsidies is negatively associated with children's development. However, as the authors discuss, this result likely reflects the goals of the subsidy program and the mixed market for child care. In fact, families use many different kinds of care arrangements for young children, including care from family, friends, and neighbors as well as care in child care facilities, churches, and schools. Some of this care is designed with a primary goal of enhancing children's development, and sometimes used part-time even when a parent is available to provide care. Other care is designed with a primary goal of supporting parental employment, with only minimal regulations often focused on health and safety. The current rules for child care subsidies also prioritize parental choice in this mixed market, but do not always offer high enough reimbursement rates to cover the highest quality care. Parents' choices may be constrained further because of their schedules and available options in their local area. In short, Hawkinson and colleagues' findings contribute to a small set of studies documenting such a negative association between subsidy receipt and child development, and suggest a need for more research into the reasons for this negative association.

1.5: LITERATURE EXCERPT 1.4

Lyons, Christopher J. 2007. "Community (Dis)Organization and Racially Motivated Crime." *American Journal of Sociology,* 113(3): 815–63.

We end with a paper by Christopher Lyons that was published in the *American Journal of Sociology* and illustrates the innovative use of several existing data sources to examine the correlates of hate crimes. Lyons (2007, 816) summarizes his research questions this way: "Are racial hate crimes the product of socially disorganized communities low in economic and social capital? Or are racially motivated crimes more likely in communities with substantial resources to exclude outsiders?" Lyons creatively

used multiple existing data sources to examine these questions. He took advantage of the Project on Human Development in Chicago Neighborhoods (PHDCN) which had gathered data that precisely measured concepts about social organization for the same time period in which he was able to obtain Chicago Police Department statistics on hate crimes for the same communities. The availability of these data provided a unique opportunity to go beyond other community measures that would be less close to the concept of social organization (e.g., measures of the poverty rate or mobility rate in the area from the US Census Bureau).

Bringing together these various data sources is an important innovation of the study. On the other hand, using existing data sources means the fit with the research questions is not as precise as it could be with newly gathered data (e.g., as Lyons discusses, the community areas might ideally be smaller in size than the 77 Chicago community areas, which average some 40,000 people, and the police districts and PHDCN data had to be aggregated to that level with some imprecision; pp. 829, 833–4). But, the creative combination of existing data sources allowed a novel set of research questions to be tested with moderate time and cost by a graduate student. The results raise some questions that might be best answered with in-depth methods targeted at the problem at hand. Yet, the existing data provide an important first look at the questions, with the results suggesting a number of next steps that could productively be pursued by any scholar, including a post-doc or young assistant professor building a body of work.

Lyons identifies three specific hypotheses that are motivated by theory and are testable with regression analyses. Importantly, the third hypothesis produces expectations different from the first two, increasing intellectual interest in the results. If hate crimes are similar to general crime, then Lyons (p. 818) expects that two traditional models would apply:

1. *Social disorganization theory* predicts more crimes in disadvantaged areas with low levels of social capital.
2. *Resource competition theories* specify that crimes are most likely when competition between racial groups increases, especially during economic downturns when resources are scarce.

The third model is unique to hate crimes, and differs from the others in its predictions:

3. The *defended community perspective* implies that interracial antagonism is most likely in economically and socially *organized* communities that are able to use these resources to exclude racial outsiders.

Lyons also posed a subquestion that is an example of a statistical interaction (a situation in which the relationship between the predictor and outcome variables differs for certain subgroups). In particular, Lyons expected that the defended community result will be particularly likely in a subset of communities: "Social cohesion and social control may be leveraged for hate crime particularly (or perhaps exclusively) in racially homogenous communities that are threatened by racial invasion, perhaps in the form of recent in-migration of racial outgroups." (p. 825).

Lyons separately coded antiblack and antiwhite hate crimes because of theoretically different expectations for each and because antiwhite hate crimes have been understudied. Literature Excerpt 1.4a provides the results for the regression analyses of antiblack hate crime (Table 6 from the article).

The table follows a pattern similar to those we examined in the earlier excerpts. The key predictor variables from the research questions and other variables that adjust for possible confounders are listed in the leftmost column. The outcome variable is indicated in the title ("antiblack hate crimes"). Several models are run, and presented in different columns, to allow us to see how the results change when different variables are included or excluded. In the final chapter of the book, we will return to this table to link its various pieces to what we have learned (such as the term "unstandardized coefficients") or what is left to be learned, in our roadmap of advanced topics (such as the term "overdispersion"). For now, we will consider the major evidence related to Lyons' research questions based on the significance (indicated by asterisks) and sign (positive or negative) of the relationship of key predictor variables.

In Lyons' Table 6, a blue circle in the last column (Model II) encloses the coefficient estimate of a three-way interaction between social control, percentage white in 1990, and percentage change in black between 1990 and 2000 which is relevant to the defended communities hypothesis. The asterisks indicate that the interaction is significant. It is difficult to interpret this interaction based only on the results in Lyons' Table 6. His Table 7 (also reproduced in Literature Excerpt 1.4a) provides additional results from the model that aid in interpretation of the interaction results (we will detail how to calculate such results in Part 2).

Each cell value in Lyons' Table 7 is the number of antiblack hate crimes that the model predicts for a cell with the listed row and column characteristics. The results show that when white communities are under threat (have experienced black in-migration of 15 percent or higher) and are high in social control, they have substantially more antiblack hate crimes (predicted number of 34.7, circled in blue) than any other communities. This is consistent with the author's expectation about which communities would be most likely to conform to the defended communities perspective: racially

Literature Excerpt 1.4a

Table 6. Negative Binomial Regressions: Community Characteristics and "Bonafide" Antiblack Hate Crimes, 1997–2002

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11
Constant	–4.01*	–6.36**	–6.29**	–6.38**	–5.64**	–14.78***	–11.15**	–10.11**	–11.10**	–10.18**	–12.67**
	(2.32)	(2.39)	(2.39)	(2.66)	(2.60)	(3.71)	(4.05)	(4.12)	(4.76)	(4.79)	(5.75)
Ln population 1990	.46**	.67**	.67**	.60**	.67**	.88**	.67**	.70**	.68**	.75**	.78**
	(.22)	(.23)	(.23)	(.25)	(.26)	(.25)	(.26)	(.26)	(.26)	(.26)	(.24)
Spatial proximity	.67**	.31	.30	.25	.26	.24	.20	.22	.19	.21	–.15
	(.34)	(.31)	(.31)	(.30)	(.30)	(.32)	(.29)	(.29)	(.29)	(.29)	(.28)
Disadvantage		–.65***	–.60**	–.22	–.05				.03	.19	.17
		(.18)	(.20)	(.34)	(.32)				(.37)	(.37)	(.35)
Stability		.11	.13	.17	.24				.03	.11	.16
		(.15)	(.15)	(.15)	(.16)				(.18)	(.18)	(.17)
White unemployment			–.02								
			(.02)								
%white 1990				.013			.01*		.013		.03
				(.01)			(.006)		(.01)		(.06)
%black 1990					–.018**			–.01**		–.018**	
					(.01)			(.006)		(.01)	
%Hispanic 1990				.01	–.007		.01	–.002	–.012	–.005	.02**
				(.01)	(.01)		(.01)	(.01)	(.01)	(.01)	(.01)
Informal social control						1.50**	1.34**	1.35**	1.31**	1.29**	1.48**
						(.60)	(.59)	(.58)	(.67)	(.66)	(.90)
Social cohesion						–.22	–.23	–.26	–.24	–.26	–.43
						(.47)	(.56)	(.53)	(.59)	(.57)	(.50)

(continued)

Literature Excerpt 1.4a—continued

Table 6. Negative Binomial Regressions: Community Characteristics and "Bonafide" Antiblack Hate Crimes, 1997–2002

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11
%change in black population, 1990–2000											–.52 (.44)
Informal social control × %white											–.002 (.02)
Informal social control × %change in black population											.14 (.13)
%white 1990 × %change in black population											–.02* (.01)
Informal social control × %white × %change in black											.005** (.002)
Overdispersion	1.23 (31.00)	.87 (.25)	.87 (.25)	.80 (.24)	.76 (.23)	.84 (.25)	.71 (.23)	.70 (.22)	.71 (.23)	.67 (.22)	.31 (–.16)
Log likelihood	146.17	139.10	–138.87	137.73	–136.46	–139.27	–136.19	–135.21	–36.17	–134.90	–126.23

Note.—$N = 77$ Chicago community areas; unstandardized coefficients; SEs are in parentheses.
* $P < .10$.
** $P < .05$.
*** $P < .001$.

(continued)

Literature Excerpt 1.4a—continued

Table 7. Predicted Antiblack Hate Crime, Chicago Communities, 1997–2002

	White Communities (85% White)		Nonwhite Communities (10% White)	
	Threat[a]	No Threat[b]	Threat[a]	No Threat[b]
High informal social control	34.7	3.9	3.1	2.1
Low informal social control	.1	.9	.1	.3

Note.—See table 6, model 11. Except for %white, informal social control, and change in %black, all variables held at mean values. High informal social control: 1 SD above mean; low informal social control: 1 SD below mean.
[a] Black in-migration 15%.
[b] Black in-migration 0%.

Source: Lyons, Christopher J. (2007). "Community (Dis)Organization and Racially Motivated Crime." *American Journal of Sociology*, 113(3): 815–63.

homogeneous areas with recent in-migration of racial outgroups. Although less extreme, it is also the case that other communities conform to the defended communities perspective: within each column, communities with high social control have more predicted antiblack hate crimes than those low in social control (i.e., two to four versus fewer than one antiblack hate crime predicted by the model).

Literature Excerpt 1.4b shows the regression results for antiwhite crime, Table 8 in Lyons' article.

A key finding here is that the basic results differ substantially from those seen for antiblack crime (compare results enclosed by black circles in Literature Excerpts 1.4a and 1.4b). For antiwhite crime, the results are consistent with social disorganization theories: in Model 2, economic disadvantage associates with more antiwhite hate crimes (asterisks and positive sign) and residential stability associates with less antiwhite hate crimes (asterisks and negative sign; see again the black circle in Literature Excerpt 1.4b). In contrast, for antiblack crime, economic disadvantage is associated with less antiblack hate crime (asterisks and negative sign) and residential stability is not associated with antiblack hate crime (no asterisks; see again the black circle in Literature Excerpt 1.4a).

Literature Excerpt 1.4b

Table 8. Negative Binomial Regressions: Community Characteristics and "Bonafide" Antiwhite Hate Crime, 1997–2002

	All Chicago[a]		Excluding Outliers[b]								
	Model 1	Model 2	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
Constant	–6.12**	–5.34**	–5.82**	–5.85	–6.24**	–5.97**	–2.31	–3.95	–3.63	–7.83**	–5.59*
	(1.84)	(1.87)	(1.83)	(1.84)	(1.89)	(1.97)	(2.68)	(2.96)	(3.17)	(3.77)	(3.13)
Ln population 1990	.64***	.56**	.59**	.59	.64**	.63**	.58**	.68***	.66***	.69**	.69**
	(.18)	(.18)	(.17)	(.17)	(.19)	(.18)	(.18)	(.19)	(.19)	(.20)	(.20)
Spatial proximity	69**	.73**	.57**	.58	.54**	.53**	.59**	.46**	.49**	.46*	.45*
	(.22)	(.24)	(.24)	(.24)	(.24)	(.24)	(.21)	(.23)	(.24)	(.24)	(.24)
Economic disadvantage		.13	.30**	.29	.26	.20				.33	.28
		(.12)	(.12)	(.14)	(.21)	(.21)				(.25)	(.25)
Residential stability		–.15	–.26**	–.25	–.26**	–.27**				–.28**	–.30**
		(.12)	(.12)	(.12)	(.13)	(.13)				(.13)	(.14)
Black unemployment 1990				.01							
				(.01)							
% black 1990					.001				.002	.004	
					(.01)				(.00)	(.01)	
% white 1990						–.003		–.005			–.01
						(.005)		(.01)			(.01)
% Hispanic 1990					–.004	–.005		–.007	–.005	.000	–.004
					(.01)	(.01)		(.01)	(.01)	(.01)	(.01)
Informal social control							–1.10*	–1.11*	–1.10*	–.54	–.56
							(.58)	(.58)	(.58)	(.61)	(.61)
Social cohesion							.18	.44	.30	.77	.85
							(.40)	(.50)	(.30)	(.51)	(.52)
Overdispersion	.36	.34	.24	.24	.23	.24	.24	.21	.21	.18	.18
	(.16)	(.16)	(.14)	(.14)	(.14)	(.14)	(.15)	(.14)	(.15)	(.13)	(.13)
Log likelihood	–130.39	–129.05	–116.68	–116.67	–116.31	–116.21	–118.32	–117.63	–117.91	–115.13	–114.80

Note.—Unstandardized coefficients, SEs are in parentheses.
[a] $N = 77$ Chicago community areas.
[b] $N = 74$ Chicago community areas.
* $P < .10$.
** $P < .05$.
*** $P < .001$.

Source: Lyons, Christopher J. (2007). "Community (Dis)Organization and Racially Motivated Crime." *American Journal of Sociology*, 113(3): 815–63.

1.6: SUMMARY

SUMMARY 1

The four literature excerpts in this chapter cover a diverse range of applications of regression analysis to examine interesting research questions. All represent research completed when the authors were graduate students or new scholars.

Reilly's study illustrated the potential for descriptive results to be illuminating, and to spark new insights and research ideas. He found that gender gaps in adolescents' reading, math, and science achievement varied across countries. Whereas girls outperformed boys in reading in all countries, boys did not always outperform girls in math and science. Reilly argues that these results suggest important ways in which context shapes stratification by gender and the ways in which policies might reduce gender disparities.

Eng analyzed data that he collected to understand the ways in which parental attitudes and investments affect academic achievement in Cambodia. Framed by the country's continuing struggle to emerge from the effects of war and genocide, Eng's results suggest important ways in which parents' fatalistic beliefs might limit their children's achievement; these results reflect a common Cambodian saying, reported by the author, that "Human strengths cannot change destiny." Eng's findings also support the possibility that parental beliefs about traditional gender roles may limit their daughters' academic achievement, especially by reducing investments in tutoring for girls.

Hawkinson and colleagues used a large U.S. sample and found that kindergartners' math and reading scores were lower when their parents had used public subsidies to pay for their child care prior to school entry. Their findings held up to numerous controls and alternative estimation strategies; along with their consistency with a handful of other studies, these sensitivity checks lend support to the veracity of the finding. The results suggest ways in which policymakers are challenged to support the dual roles of child care as both an employment support and early childhood investment.

Lyons' research on hate crimes shows that a seemingly positive aspect of community—social capital—can work to achieve ingroup goals at the cost of outgroup members. Socially organized areas are observed to have more antiblack hate crime. This is especially evident in primarily white communities that have experienced high levels of black in-migration. A different process applies for antiwhite hate crimes, which correlate with characteristics of neighborhoods as emphasized by general theories of crime, especially residential instability.

In future chapters we will dig deeper to understand questions these excerpts may have raised for you. For example: how do we know to interpret the significance and

size of coefficients as we do? Exactly how and why do coefficients change when other variables are controlled? How is it that the regression model tests for an interaction and what do the results mean? We hope that these excerpts spark enthusiasm for answering these questions and learning how to better understand the regression results that you read in the literature and to implement regression models in your own work.

KEY TERMS 1

KEY TERMS

Dependent Variable (or Outcome Variable)

Independent Variable (Predictor Variable or Explanatory Variable)

Mechanisms

Observational Data

Regression Analysis

CHAPTER EXERCISES 1

CHAPTER EXERCISES

1.1. Locate a journal article using regression analyses, preferably on a topic of interest to you. You can use an article you already have, or draw something from the syllabus of one of your substantive classes, or search a bibliographic database. Read (or reread) the article, paying particular attention to the research questions or hypothesis statements and how these are examined in the regression models. Examine at least one table of regression results in detail, and pull out some of the features that were discussed in this chapter (e.g., significance and sign of coefficients relevant to the study's research questions and hypotheses). Write a short paragraph discussing what is easiest and most difficult to understand about the article's regression analysis. Keep a copy of this article to revisit in the chapter exercise in the final chapter of the book.

1.2. Answer the following questions in order to help your instructor get to know you better and to help you think about your goals for the course. Some questions may be difficult to answer, especially if your research interests and empirical approach are still developing, so answer as best you can. In addition to helping your instructor tailor lectures and interactions, this exercise can also help you to think about how to get the most out of the course.

(a) Have you taken previous statistics courses? (If yes, what/when? How well do you feel you mastered the material?)

(b) Have you worked, or are you currently working, as a research assistant on any research projects (qualitative or quantitative)?

(c) Have you collected your own data or analyzed existing data for a prior paper, including for an undergraduate senior thesis or a graduate master's thesis (qualitative or quantitative)?

(d) What do you hope to get out of the current class?

(e) Do you see yourself more as a qualitative researcher, a quantitative researcher, both, neither, or are you unsure?

(f) What is the substantive area of your research (or what major topics do you think you would like to study as a graduate student)?

(g) Provide an example of a regression-type relationship from your area of research; that is, list an "outcome" (dependent variable) with one or more "predictors" (independent variables).

COURSE EXERCISE

COURSE EXERCISES 1

Write three research questions in your area of interest. Be sure to identify clearly the outcome (dependent variable) and one or more predictors of interest (independent variables) for each question. Ideally, your outcome could be measured continuously for at least one research question so that it will be appropriate for the techniques learned in Part 2 of the book.

If you are able to write your research question ("Do earnings differ by gender?") as a hypothesis statement (e.g., "Women earn less than men.") you should do so, but if prior research and theory do not offer enough insights for a directional hypothesis statement, a nondirectional research question is fine at this stage.

If your interests are still developing, you may want to get ideas by scanning a number of articles, chapters or books from one of your substantive classes, perusing the table of contents of recent top journals in your field, or by conducting a few bibliographic literature searches with key terms of broad interest to you.

The research questions should be of interest to you and something that you would like to examine as you move forward with the course. They could be already well

studied in your field (i.e., you do not need to identify a dissertation-type question that makes a novel contribution to the literature). You may modify and refine your questions/hypotheses as you progress, especially so that you can apply the techniques learned in future chapters (e.g., modeling different types of variable, testing for mechanisms, etc.).

Chapter 2

PLANNING A QUANTITATIVE RESEARCH PROJECT WITH EXISTING DATA

CHAPTER 2: PLANNING A QUANTITATIVE RESEARCH PROJECT WITH EXISTING DATA

A primary goal of this chapter is to help you to plan for a quantitative research project. Whether you collect your own data, or use existing data, careful planning will produce a more efficient project (see also Long 2009 for an excellent treatment of the *Workflow* of social science research). You will be able to complete the project more quickly and to interpret the results more easily than you could otherwise. At the planning stage you should think through questions such as: What variables are needed to operationalize my concepts? Can I predict the direction of association between each predictor variable and the outcome? Do I anticipate that any predictor variables' associations with the outcome may change when confounding variables are included in the models? Do I anticipate that any predictor variables' associations with the outcome will differ for various subgroups? Doing so can help you to choose from among existing data sets (selecting the one that has the right variables and sample coverage). Forward thinking should save you time in the long run, because you will be less likely to have to backtrack to create additional variables or to rerun earlier analyses. Forward thinking will also help you to avoid being unable to examine particular research questions because subsamples are too small or constructs were not measured.

We focus on existing data in this book. We expose you to numerous data sets that are readily available online. Knowing about these data sets should make it easier for you to use regression analyses to examine research questions of interest to you in the short term. This chapter also contributes to our goal of leveling the playing field by offering information about where and how to look for such existing data, since not all students will have access to this information through mentors and peers.

Generally, existing data sources also have three distinct advantages over collecting new data:

1. They often provide designs and sample sizes that support complex regression models and hypothesis tests (e.g., confounds, mediators, moderators).
2. They allow research questions to be addressed when resources for collecting new data are limited.
3. Their use provides a greater return on public investment in large-scale data collection.

There are instances where regression analysis might be used on non-random samples (Berk 2004). But, in order to support statistical tests based on regression analyses—and thus test hypotheses or answer research questions—a sample of adequate size (e.g., often 100 or more cases) should be drawn randomly from a known population. More complicated models, with more variables and interactions among variables, require

even larger sample sizes, overall and within subgroups. Yet, gathering large-scale data sets drawn systematically from populations is expensive. If an existing data set has the relevant measures to answer a social scientist's research question, then the research can be conducted more quickly and with less cost than if the researcher attempted to gather new data. Turning first to existing data also best utilizes scarce resources for funding new studies. And, first testing a research question on existing data can provide the preliminary answers needed to modify the research questions and to demonstrate competence for seeking funding for new data collection.

That said, even though we focus on existing data, the general strategies about planning for the analysis at the data-gathering stage also apply generally for your own data collection (e.g., being sure you ask all the right questions and oversample subgroups when needed). You should, of course, consult with mentors and, where needed, take advanced courses to prepare to implement your own data collection. We assume that you have taken, or will take, a basic research methods class, which covers a range of data collection methodologies.

2.1: SOURCES OF EXISTING DATA

Existing data come from a number of sources, including:

- multi-data set **archives**;
- single-data set archives;
- individual researchers.

Table 2.1 provides links to example resources we discuss below in each of these categories. Before turning to that discussion, we first want to emphasize the importance of checking with your local Institutional Review Board (IRB) regarding necessary IRB approval if you plan to use one of these sources for research purposes (versus for educational purposes only, such as course exercises). It may seem that IRB approval is not needed since some existing data sets are prepared for **public release**, for example with careful removal of all identifying information, and can be downloaded from web sites. Indeed, some local IRBs have determined that publicly available data sets accessed from preapproved data archives do not involve human subjects and thus do not require IRB review. But, current human subjects' policy requires researchers to verify this with their local IRB, rather than making such determinations independently (see Box 2.1 for an example). Some existing

Box 2.1

The University of Chicago's Social and Behavioral Sciences Institutional Review Board provides a nice summary of its policies and procedures for existing data sets: *http://sbsirb.uchicago.edu/page/secondary-data-analysis.*

Table 2.1: Examples of Data Archives

General Multi-Data Set Archives	
Inter-University Consortium for Political and Social Research (ICPSR)	http://www.icpsr.org/
Henry A. Murray Research Archive at Harvard University	http://www.murray.harvard.edu/
Sociometrics	http://www.socio.com/
Government Multi-Data Set Archives	
US Census Bureau	
American FactFinder	http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml
Census Data Products	http://www.census.gov/mp/www/cat/
Integrated Public Use Microdata Series (IPUMS)	http://www.ipums.umn.edu/
NCHS Public Use Data	http://www.cdc.gov/nchs/data_access/ftp_data.htm
NCES Surveys and Programs	http://nces.ed.gov/pubsearch/surveylist.asp
NCHS Research Data Center	http://www.cdc.gov/rdc/
Census Research Data Centers	http://www.census.gov/ces/rdcresearch/
Single-Data Set Archives	
National Survey of Families and Households	http://www.ssc.wisc.edu/nsfh/
Urban Communes Data Set	http://home.uchicago.edu/~jlmartin/UCDS
Three City Study	http://www.jhu.edu/welfare

Notes: NCHS = National Center for Health Statistics; NCES = National Center for Education Statistics.

data sets also require special data use or data security agreements, for example that spell out special precautions for limiting access to the data. If you want to use these data for research, you should build in time for completing these agreements, which may require signatures from your adviser and administrators at your university, and for review by your local IRB. Some **restricted-use data sets** cannot be sent to you, but can be analyzed in special secure locations. These data sets require the greatest time lags for approvals, including local IRB review.

2.1.1: Multi-Data Set Archives

One of the oldest and largest multi-data set archives is the Inter-University Consortium for Political and Social Research (**ICPSR**; see Table 2.1). ICPSR began in 1962 to archive data from computer-based studies in political science. In the mid-1970s, the name "social" was added to the title to reflect the broader set of disciplines that were

archiving data (Vavra 2002). The founders recognized the need to store centrally the growing amounts of quantitative data being collected by political scientists across the country. This allowed other scientists not only to replicate the findings of the original scholar, but also to tap the data for additional purposes, beyond those within the interests and time limits of the original scholar (Vavra 2002; see Box 2.2). The archive is housed at the Institute for Social Research at the University of Michigan and hundreds of universities from across the world are institutional members, giving their faculty and students access to tens of thousands of data sets.

Box 2.2

Today's new user of ICPSR is used to having a wealth of data at her fingertips over the Internet. The remarkable achievement underlying this ease of access to data sets spanning centuries is easy to overlook. Especially in recent decades, the speed with which hardware and software advanced made data sets vulnerable to being inaccessible, even if their files were stored. In other words, data sets stored in formats written for now obsolete operating systems or statistical packages are not directly readable on today's computers. Fortunately, the archives had the foresight to convert these data sets to more general formats, making them still usable to today's researchers (Vavra 2002).

Data sharing is common today in part because funders often require it. For example, the National Institutes of Health (NIH) now requires that all proposals requesting $500,000 or more in direct costs in any single year must include a plan for sharing the data (with appropriate removal of identifying information to protect confidentiality). In its policy, NIH notes the intention to give broader access to the data, once the original researchers have had a chance to pursue their main research objectives: "initial investigators may benefit from first and continuing use but not from prolonged exclusive use" (NIH 2003). Thus, archives have become important to scholars, as investigators can turn to such archives to deposit their data and meet such requirements.

ICPSR's coverage is the broadest among today's major data archives. At the time of this writing, ICPSR used numerous thematic categories covering broad topics including wars, economic behavior, leadership, geography, mass political behavior, health care, social institutions, and organizational behavior. The subtopic of family and gender, within social institutions, returns over 75 data sets, ranging from one-time polls to multiyear surveys. A bibliography of publications using the archived data sets includes over tens of thousands of citations.

Two other data archives often used by social scientists are the Henry A. Murray Research Archive at Harvard University and the Sociometrics Archive (see again Table 2.1). The Murray Archive focuses on data sets that study human development across the lifespan from a range of disciplines. The Sociometrics archive focuses on nine topics: teen pregnancy, family, aging, disability, maternal drug abuse, HIV/AIDS/STI, contextual, child poverty, and alternative medicine.

Federal government agencies also provide direct access to data that they gather. Census data, aggregated to geographic areas such as states, counties, and census tracts, are available freely online or for purchase on DVD. Most students are familiar with the decennial census of population, but data are also available from the economic censuses, which survey business establishments every five years. Individual level data from the Decennial Census of Population are also available through the Integrated Public Use Microdata Series (IPUMS).[1] Similarly, the National Center for Health Statistics and the National Center for Education Statistics also make numerous data sets available, in public use and restricted formats.

The costs of accessing these data range from zero to several hundreds of dollars. Data in the ICPSR archive are free to faculty and students at ICPSR member institutions. In most cases, individual faculty and students can create an account using their institutional affiliation, and download data and documentation directly from the ICPSR web site.[2] Some of the other data resources are also freely available on the Web (e.g., Census American FactFinder, IPUMS); others have data set-specific use agreements and fees. Sometimes, multiple archives contain the same data set; at other times, a data set is available in only one archive, or is available in more detail in one archive. So, it is worth searching several archives when initially exploring data sets for a topic.

In addition to these resources, which allow a copy of the data to be analyzed at the researcher's place of work (sometimes with specific requirements to limit others' access to the data), the US Census Bureau and National Center for Health Statistics also make data available to researchers at Research Data Centers across the country (the Web addresses listed in Table 2.1 include descriptions of the data sets available at the centers). At these sites researchers can access data that are not included in public releases. For example, researchers can identify individuals at small levels of geography and match the individuals to information about these contexts. In addition, firm-level data from the economic censuses can be linked across time to study the births and deaths of organizations. Although the time to get approval is longer for these restricted-use data than with publicly available data sets, and the fees for using the center can raise the cost of a project, these centers allow scholars to pursue unique and innovative projects.

2.1.2: Single-Data Set Archives

Some investigators archive their data individually. Sometimes these data sets are also available through multi-data archives. For example, the National Survey of Families and Households (**NSFH**), which we will use for examples throughout the book, maintains its own web site from which data and documentation can be freely downloaded. Some of the data are also available through the ICPSR and Sociometrics archives.