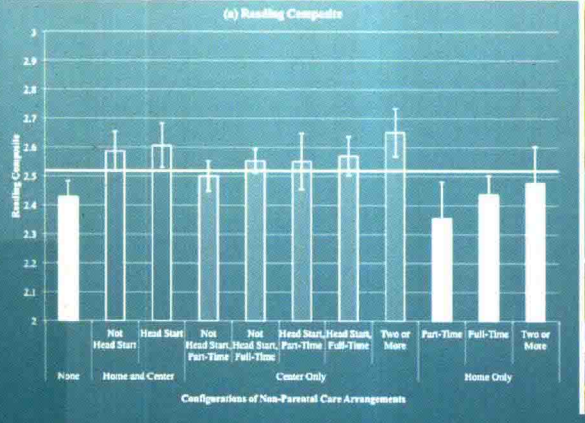
REGRESSION ANALYSIS FOR THE SOCIAL SCIENCES

SECOND EDITION



RACHEL A. GORDON

ROUTLEDGE

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

Operators for Stata Expressions		
	Meaning	**Symbols in a Stata Expression**
Comparison Operators		
	Equal to	==
	Not equal to	~=
		!=
	Greater than	>
	Greater than or equal to	>=
	Less than	<
	Less than or equal to	<=
Logical Operators		
	And	&
	Or	|
	Not	~
Arithmetic Operators		
	Multiplication	*
	Division	/
	Addition	+
	Subtraction	-
Note. The double equal sign (==) is used in expressions to test equality (a single equal sign is used in other places, such as when creating variables, as in `generate newvar = oldvar if othervar==1`)		

Additional Notes about Stata Commands	
System missing	• System . missing is treated as the highest positive value.
Case sensititivity	• Stata is case sensitive. • Commands must be typed in lower case, and can be abbreviated to the shortest unique letters (usually the first three letters of the command name).
End of line delimiter	• Stata uses the carriage return (end of line) as a delimiter (i.e., to indicate where one command stops and another starts). No symbol (like a period or semi-colon) is needed to end a line. To accommodate a long command that extends over multiple lines: o Change the delimiter to a semicolon with the command `#delimit ;` Change the delimiter back the carriage return with `#delimit cr` to use long lines only temporarily. o Use three forward slashes /// as a continuation comment. These are placed at the end of each continuing line (the final line ends with a carriage return).
Comments	• Stata has three ways of indicating comments. One way is to enclose the comment in between the symbols `/*` and `*/` This type of comment can extend over multiple lines. A single line comment can also begin with either with an asterisk or with two forward slashes //

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

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**.

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.