

## Regression for Health and Social Science

This textbook for students in nontechnical scientific fields covers the basics of linear model methods with a minimum of mathematics, assuming only a precalculus background. Numerous examples drawn from the news and current events, with an emphasis on health issues, illustrate the concepts in an immediately accessible way. Methods covered include linear regression models, Poisson regression, logistic regression, proportional hazards regression, survival analysis, and nonparametric regression.

The author emphasizes interpretation of computer output in terms of the motivating example. All of the **R** code is provided and carefully explained, allowing readers to quickly apply the methods to their own data. Plenty of exercises help students to think about the issues involved in the analysis and its interpretation.

Code and datasets are available for download from the book's website at [www.cambridge.org/zelterman](http://www.cambridge.org/zelterman)

**Daniel Zelterman**, PhD, is Professor Emeritus, Department of Biostatistics, at Yale University. His application areas include work in clinical trial designs for cancer studies. Before moving to Yale in 1995, he was on the faculty of the University of Minnesota and at the State University of New York at Albany. He is an elected Fellow of the American Statistical Association. In his spare time he plays oboe and bassoon and has backpacked hundreds of miles of the Appalachian Trail.



Cambridge University Press & Assessment  
978-1-108-47818-2 — Regression for Health and Social Science  
Daniel Zelterman  
Frontmatter  
[More Information](#)

---

# Regression for Health and Social Science

Applied Linear Models with **R**

DANIEL ZELTERMAN  
*Yale University, Connecticut*

Cambridge University Press & Assessment  
978-1-108-47818-2 — Regression for Health and Social Science  
Daniel Zelterman  
Frontmatter  
[More Information](#)

## CAMBRIDGE UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,  
New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

[www.cambridge.org](http://www.cambridge.org)

Information on this title: [www.cambridge.org/highereducation/isbn/9781108478182](http://www.cambridge.org/highereducation/isbn/9781108478182)

DOI: 10.1017/9781108784504

© Daniel Zelterman 2022

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published 2022

Printed in the United Kingdom by TJ Books Limited, Padstow, Cornwall, 2022

*A catalogue record for this publication is available from the British Library.*

ISBN 978-1-108-47818-2 Hardback

Additional resources for this publication at [www.cambridge.org/zelterman](http://www.cambridge.org/zelterman)

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

## Contents

	<i>Preface</i>	page xi
	<i>Acknowledgments</i>	xv
<b>1</b>	<b>Introduction</b>	1
	1.1 What Is Statistics?	1
	1.2 Statistics in the News: the Weather Map	4
	1.3 Mathematical Background	6
	1.4 Calculus	7
	1.5 Calculus in the News: New-Home Construction	9
	1.6 A Cautionary Tale	11
	1.7 Exercises	13
	1.7.1 Motorcycle Accidents	14
	1.7.2 Olympic Records	14
	1.7.3 Gasoline Consumption	15
	1.7.4 Foreign Owners of US Treasury Debt	17
	1.7.5 US Presidents and Stock Market Returns	17
<b>2</b>	<b>Principles of Statistics</b>	21
	2.1 The Binomial Distribution	21
	2.2 Confidence Intervals and the Hubble Constant	27
	2.3 The Normal Distribution	29
	2.4 Hypothesis Tests	32
	2.5 The Student t-Test	36
	2.5.1 An Example in Practice	37
	2.5.2 Read the Data and Perform Simple Checks	40
	2.5.3 Run and Interpret the t-Test	42
	2.6 The Chi-Squared Test and $2 \times 2$ Tables	45
	2.7 What Are Degrees of Freedom?	51
	2.8 <b>R</b> , in a Nutshell	52
	2.9 Survey of the Remainder of the Book	56

2.10	Exercises	57
2.10.1	Maintaining Balance	60
2.10.2	Reading Scores	61
2.10.3	A Helium-Filled Football	61
2.10.4	Reexamine the Fusion Times	62
<b>3</b>	<b>Introduction to Linear Regression</b>	<b>64</b>
3.1	Low-Birth-Weight Infants	64
3.2	The Least-Squares Regression Line	65
3.3	Regression in <b>R</b>	69
3.4	Statistics in the News: Future Healthcare Costs	71
3.5	Exercises	72
3.5.1	Statistics in the News: Savings for Medicare	75
3.5.2	Arsenic in Drinking Water	76
3.5.3	Dermatologists' Fees	78
3.5.4	Breast Cancer Survival and Climate	79
3.5.5	Cancer Mortality in Florida	80
3.5.6	Vital Rates	82
<b>4</b>	<b>Assessing the Regression</b>	<b>84</b>
4.1	Correlation	84
4.2	Statistics in the News: Correlations of the Global Economy	86
4.3	Analysis of Variance	87
4.4	Model Assumptions and Residual Plots	91
4.5	Exercises	95
4.5.1	Food Imports	96
4.5.2	US Homicide Rates	97
4.5.3	Statistics in the News: Women Managers	98
4.5.4	Statistics is More Than Just Numbers	99
<b>5</b>	<b>Multiple Regression and Diagnostics</b>	<b>101</b>
5.1	Example: Maximum January Temperatures	101
5.2	Graphical Displays of Multivariate Data	104
5.3	Leverage and the Hat Matrix Diagonal	106
5.4	Jackknife Diagnostics	110
5.5	Partial Correlation	113
5.6	Model-Building Strategies	116
5.7	Exercises	120
5.7.1	University Endowments	120
5.7.2	Maximum January Temperatures	122
5.7.3	Heart Surgery Mortality	123
5.7.4	Characteristics of Cars, 1974	124
5.7.5	Statistics in Advertising: Wine Prices	125
5.7.6	Statistics in Finance: Mutual Fund Returns	128

<b>6</b>	<b>Indicators, Interactions, and Transformations</b>	130
	6.1 Indicator Variables	130
	6.2 Drug Interactions	138
	6.3 Interactions of Explanatory Variables	140
	6.4 Transformations	144
	6.5 Additional Topics: Longitudinal Data	149
	6.6 Exercises	151
	6.6.1 More on Wine Prices	152
	6.6.2 Nicotine Levels in Cigarettes	153
	6.6.3 The Speed of a Reaction	153
	6.6.4 Tumor Growth in Mice	154
	6.6.5 Used Car Prices	155
	6.6.6 Percent Body Fat	156
	6.6.7 Fertility Rates in Switzerland	157
	6.6.8 ELISA	158
<b>7</b>	<b>Nonparametric Statistics</b>	160
	7.1 A Test for Medians	160
	7.2 Elementary School Math Achievement Scores	165
	7.3 Rank Sum Test	167
	7.4 Ranking and the Healthiest State	169
	7.5 Nonparametric Regression: LOESS	170
	7.6 Exercises	173
	7.6.1 Cloth Run-Up	175
	7.6.2 Prices of Beanie Babies	176
	7.6.3 The Cracker Diet	177
<b>8</b>	<b>Logistic Regression</b>	178
	8.1 Example: an Insecticide Experiment	178
	8.2 The Logit Transformation	179
	8.3 Logistic Regression in <b>R</b>	182
	8.4 The New York Mets	186
	8.5 Key Points	187
	8.6 Exercises	188
	8.6.1 A Phase I Clinical Trial in Cancer	189
	8.6.2 Toxoplasmosis in El Salvador	190
	8.6.3 Estimation of the $ED_{01}$	192
	8.6.4 Super Bowl XXXVIII	193
<b>9</b>	<b>Diagnostics for Logistic Regression</b>	195
	9.1 A Larger Example	195
	9.2 Residuals for Logistic Regression	197
	9.3 Influence in Logistic Regression	201

9.4	Exercises	205
9.4.1	Statistics in the News: Sex and Violins	206
9.4.2	Glove Use among Nurses	207
9.4.3	Statistics in Sports: Pittsburgh Steelers' Rushing Game	209
9.4.4	Climate Records in Washington, DC	210
<b>10</b>	<b>Poisson Regression</b>	212
10.1	Lottery Winners	212
10.2	Poisson Distribution Basics	212
10.3	Statistics in the News: Terror Attacks	215
10.4	Regression Models for Poisson Data	216
10.5	The Offset	221
10.6	Exercises	223
10.6.1	Coronary Bypass Mortality, Revisited	223
10.6.2	Cases of Mental Illness	223
10.6.3	Airlines Bump Passengers	224
10.6.4	Lottery Winners	226
10.6.5	Species on the Galápagos Islands	226
10.6.6	Sports Statistics: Pro Bowl Appearances	227
10.6.7	Cancer Rates in Japan	229
10.6.8	Tourette's Syndrome	231
<b>11</b>	<b>Survival Analysis</b>	233
11.1	Censoring	233
11.2	The Survival Curve and its Estimate	235
11.3	The Log-Rank Test	240
11.4	Exercises	243
11.4.1	Cancer of the Bile Duct	243
11.4.2	Survival of Centenarians	244
<b>12</b>	<b>Proportional Hazards Regression</b>	245
12.1	The Hazard Function	245
12.2	The Model of Proportional Hazards Regression	247
12.3	Proportional Hazards Regression in <b>R</b>	249
12.4	Exercises	251
12.4.1	Survival of Halibut	251
12.4.2	Stanford Heart Transplant Survival	253
12.4.3	Primary Biliary Cirrhosis	253
12.4.4	Multiple Myeloma	254



<b>13</b>	<b>Review of Methods</b>	256
	13.1 The Appropriate Method	256
	13.2 Other Review Questions	258
<b>Appendix</b>	<b>Statistical Distributions</b>	263
	A.1 Normal Distribution	263
	A.2 Chi-Squared Distribution	264
	<i>Selected Solutions and Hints</i>	266
	<i>References</i>	272
	<i>Index</i>	274

Cambridge University Press & Assessment  
978-1-108-47818-2 — Regression for Health and Social Science  
Daniel Zelterman  
Frontmatter  
[More Information](#)

---

## Preface

Linear models are a powerful and useful set of methods in a large number of settings. Briefly, there is some important outcome measurement and we want to explain variations in its values in terms of other measurements in the data. The heights of several trees can be explained in terms of the trees' ages, for example. It is not a straight-line relationship, of course, but knowledge of a tree's age offers us a large amount of explanatory value. We might also want to take into account the effects of measurements on the amount of light, water, nutrients, and weather conditions experienced by each tree. Some of these measurements will have greater explanatory value than others, and we may want to quantify the relative usefulness of these different measures. Even after we are given all of this information, some trees will appear to thrive and others will remain stunted, when all are subjected to identical conditions. Understanding this type of variability is the whole reason for the existence of statistics as a scientific discipline. We usually try to avoid use of the word "prediction" because this assumes there is a cause-and-effect relationship. A tree's age does not directly cause it to grow, for example, but rather, a cumulative process associated with many environmental factors results in increasing height and continued survival. The best estimate we can make is a statement about the behavior of the average tree under identical conditions.

Many of my students go on to work in the pharmaceutical or healthcare industries after graduating with a master's degree. Consequently, the choice of examples in this book has a decidedly health or medical bias. We expect our students to be useful to their employers the day they leave our program, so there is not a lot of time to spend on advanced theory that is not directly applicable. Not all of the examples are from the health sciences. Diverse examples such as the number of lottery winners and temperatures in various US cities are part of our common knowledge. Such examples do not need a lengthy explanation in order for the reader to appreciate many of the aspects of the data being presented.

How is this book different from the many available on the market? The mathematical content and notation are kept to an absolute minimum. To paraphrase the noted physicist Steven Hawking, who wrote extensively for the popular audience, every equation loses half of your audience. There is really no need for formulas and their derivations in a book of this type if we rely on the computer to calculate quantities of interest. Long gone are the days of doing statistics with calculators or on the back of an envelope. Students of mathematical statistics should be able to provide

the derivations of the formulas, but they represent a very different audience. All the important formulas are programmed in software so there is no need for the general user to know these.

The three important skills needed by a well-educated student of applied statistics are as follows.

1. Recognize the appropriate method needed in a given setting.
2. Have the necessary computer skills to perform the analysis.
3. Be able to interpret the output and draw conclusions in terms of the original data.

This book gives examples to introduce the reader to a variety of commonly encountered settings and provides guidance through these to complete the three goals. Not all possible situations can be described, of course, but the chosen settings include a broad survey of the types of problems the student of applied statistics is likely to run into.

What do I ask of my readers? We still need to use a lot of mathematical concepts such as the connection between a linear equation and drawing the line on  $X$ - $Y$  coordinates. There will be algebra and special functions such as square roots and logarithms. Logarithms, while we are on the subject, are always to the base  $e$  ( $= 2.718\dots$ ) in this book and not base 10.

We will also need a nodding acquaintance with the concepts of calculus. Many of us took calculus in college a long time ago and have not had much need to use it in the years since. Perhaps we intentionally chose a course of study avoiding abstract mathematics. Even so, calculus represents an important and useful tool. The definitions of the derivative of a function (What does this new function represent?) and integral (What does *this* new function represent?) are required, although we will never actually need to find a derivative or an integral. The necessary refresher to these important concepts is given in Section 1.4.

Also helpful is a previous course in statistics. The reader should be familiar with the mean and standard deviation, normal and binomial distributions, and hypothesis tests in general and the chi-squared and  $t$ -tests specifically. These important concepts are reviewed in Chapter 2, but an appreciation of these basic ideas is almost a full course in itself. There is a large reliance on  $p$ -values in scientific research, so it is important to know exactly what these represent.

There are a number of excellent general purpose statistical software packages available. We have chosen to illustrate our examples using **R** because of its wide acceptance and use in many industries but especially those of healthcare and pharmaceutical. Most of the examples given here are small, to emphasize interpretation and encourage practice. These data sets could be examined by most software packages. **R**, however, is capable of handling much larger data sets so the skills learned here can easily be used if and when much larger projects are encountered later.

The reader should already have some rudimentary familiarity with running **R** on a computer. This would include using the editor to change the program, submitting the program, retrieving and then printing the output. There are also popular point-and-click approaches to data analysis. While these are quick and acceptable, their ease of

use comes with the price of not always being able to repeat the analysis because of the lack of a printed record of the steps taken. Data analysis, then, should be reproducible.

We will review some of the basics of **R** but a little hand-holding will prevent some of the agonizing frustrations frequently occurring when first starting out. Running the computer, and more generally doing the exercises in this book, are a very necessary part of learning statistics. Just as you cannot learn to play the piano simply by reading a book, statistical expertise and the accompanying computer skills can only be obtained by hours of actively using them. Again, much like the piano, the instrument is not damaged by playing a wrong note. Nobody will laugh at you if you try something truly outlandish on the computer either. Perhaps something better will come from a new look at a familiar setting. Similarly, the reader is encouraged to look at the data and try a variety of different ways of looking, plotting, modeling, transforming, and manipulating. Unlike a mathematical problem with only one correct solution (contrary to many of our preconceived notions) there is often a lot of flexibility in the way statistics can be applied to summarize a set of data. As with yet another analogy to music, there are many ways to play the same song.

Cambridge University Press & Assessment  
978-1-108-47818-2 — Regression for Health and Social Science  
Daniel Zelterman  
Frontmatter  
[More Information](#)

---

## Acknowledgments

Thanks to the many students and teaching assistants who have provided useful comments and suggestions to the exposition as well as the computer assignments. Also to Beth Nichols, Chang Yu, and Steven Schwager for their careful readings of early drafts of the manuscript. Lauren Cowles and her staff at Cambridge University Press provided innumerable improvements and links to useful websites.

### Figure Credits

Figure 1.1: Courtesy of Pennsylvania State University, Department of Meteorology.

Figure 1.4: From the US Census.

Figure 1.5: From Stuckler D, King LP, and Basu S (2008). International Monetary Fund programs and tuberculosis outcomes in post-communist countries. *PLoS Medicine*. Available online at doi:10.1371/journal.pmed.0050143

Figure 1.6: From *The New York Times* (August 15, 2008). © 2008 The New York Times Company. All rights reserved. Used under license.

Figure 1.7: From *The New York Times* (August 4, 2008). © 2008 The New York Times Company. All rights reserved. Used under license.

Figure 1.8: From *The New York Times* (August 23, 2008). © 2008 The New York Times Company. All rights reserved. Used under license.

Figure 1.10: From *The New York Times* (October 14, 2008). © 2008 The New York Times Company. All rights reserved. Used under license.

Table 2.1: From Frisby JP and Clatworthy JL (1975). Learning to see complex random-dot stereograms. *Perception* 4: 173–8. © Sage Publishing with permission from Pion Ltd.

Figure 2.2: Courtesy of John Huchra.

Table 2.4: On the cognitive penetrability of posture control, N. Teasdale, C. Bard, J. Larue et al., *Experimental Aging Research*, © 1993 Taylor & Francis, reprinted by permission of the publisher (Taylor & Francis Group, [www.informaworld.com](http://www.informaworld.com)).

Table 3.1: Data from US Census and used with permission of The Baseball Cube.

Figure 3.4: Prepared by US Drought Monitor, University of Nebraska–Lincoln and updated daily.

Table 3.7: Data obtained from the US National Centers for Environmental Information [www.ncdc.noaa.gov](http://www.ncdc.noaa.gov) and the Florida annual cancer registry [www.floridahealth.gov/](http://www.floridahealth.gov/).

Table 4.5: Data from the World Economic Forum and also appeared in [www.reuters.com/article/us-japan-companies-women/women-in-management-at-japan-firms-still-a-rarity-reuters-poll-idUSKCN1LT3GF](http://www.reuters.com/article/us-japan-companies-women/women-in-management-at-japan-firms-still-a-rarity-reuters-poll-idUSKCN1LT3GF)

Figure 4.5: From the Bureau of Transportation Statistics.

Figures 6.1: reprinted from Cokol M, Chua HN, Tasan M, *et al.* (2011). Systematic exploration of synergistic drug pairs. *Molecular Systems Biology* **7**: Article number 544; [/http//doi:10.1038/msb.2011.71](http://doi:10.1038/msb.2011.71)

Figures 6.2: Reprinted from Cokol M, Chua HN, Tasan M, *et al.* (2011). Systematic exploration of synergistic drug pairs. *Molecular Systems Biology* **7**: Article number 544; [/http//doi:10.1038/msb.2011.71](http://doi:10.1038/msb.2011.71)

Table 6.3: Data from the Energy Information Administration.

Table 6.5: Data from the Massachusetts Department of Public Health.

Table 6.7: From Koziol JA, Maxwell DA, Fukushima M, Colmerauer MEM, and Pilch YH (1981). A distribution-free test for tumor-growth curve analysis with application to an animal tumor immunotherapy experiment. *Biometrics* **37**: 383–90. Reprinted with permission of Wiley.

Table 7.1: From *Introduction To Generalized Linear Models*, second edition by Dobson AJ. © 2001 by CRC Press. Reproduced with permission of Taylor & Francis Group LLC.

Figure 7.2: From the New York State Education Department.

Table 9.5: Data obtained from the US National Weather Service.

Figure 10.2: From Datagraver.com based on data from the START Global Terrorism Database.

Table 10.6: From van Watum PJ, Chappell PB, Zelterman D, Scahill LD, and Lecktmann JF (2000). Patterns of response to acute Naxolone infusion in Tourette’s Syndrome. *Movement Disorders* **15**: 1252–4. Reprinted with permission of Oxford University Press.

Figure 11.2: Wheler J, Tsimberidou AM, Hong D, *et al.* (2009). Survival of patients in a Phase I clinic. *Cancer* **115**(7): 1091–9. © 2009 American Cancer Society.

Table 11.3: From Fleming T, O’Fallon JR, O’Brien PD, and Harrington DP (1980). Modified Kolmogorov–Smirnov test procedures with application to arbitrarily right-censored data. *Biometrics* **36**: 607–25. Reprinted with permission of Wiley.