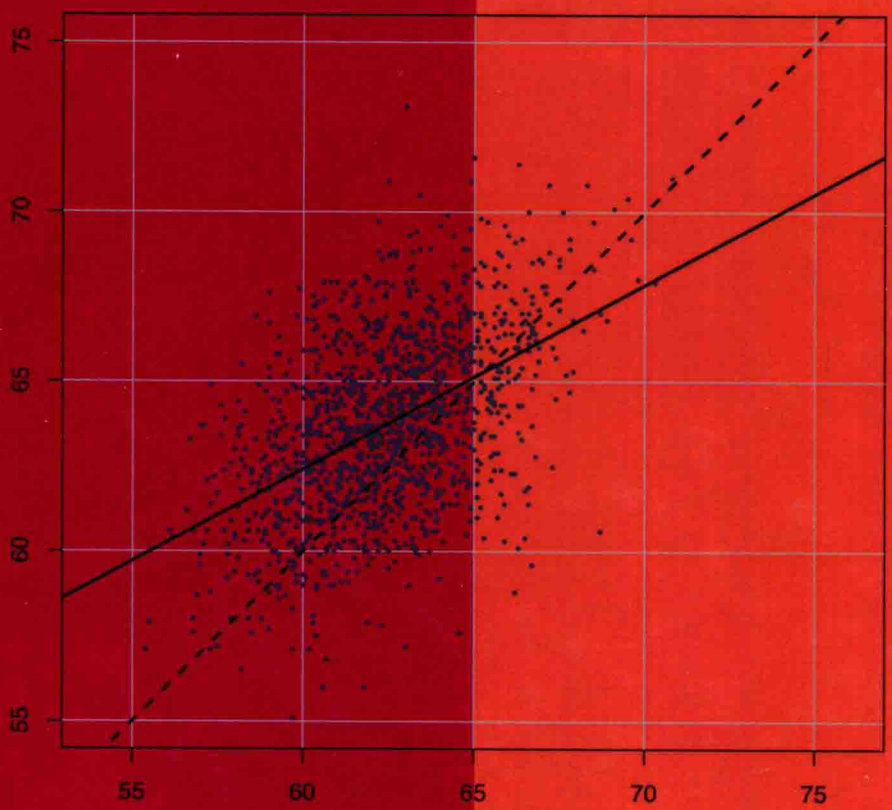
Wiley Series in Probability and Statistics

Applied Linear Regression

SANFORD WEISBERG

FOURTH EDITION



WWW. LINK AVAILABLE

WILEY

Applied Linear Regression

Fourth Edition

SANFORD WEISBERG

School of Statistics
University of Minnesota
Minneapolis, MN

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data:

Weisberg, Sanford, 1947–
Applied linear regression / Sanford Weisberg, School of Statistics, University of Minnesota, Minneapolis, MN.—Fourth edition.
pages cm
Includes bibliographical references and index.
ISBN 978-1-118-38608-8 (hardback)
1. Regression analysis. I. Title.
QA278.2.W44 2014
519.5′36–dc23

2014026538

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Applied Linear Regression

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice, Harvey Goldstein, Iain M. Johnstone, Geert Molenberghs, David W. Scott, Adrian F. M. Smith, Ruey S. Tsay, Sanford Weisberg*
Editors Emeriti: *Vic Barnett, J. Stuart Hunter, Joseph B. Kadane, Jozef L. Teugels*

A complete list of the titles in this series appears at the end of this volume.

To Carol, Stephanie,
and
the memory of my parents

Preface to the Fourth Edition

This is a *textbook* to help you learn about applied linear regression. The book has been in print for more than 30 years, in a period of rapid change in statistical methodology and particularly in statistical computing. This fourth edition is a thorough rewriting of the book to reflect the needs of current students. As in previous editions, the overriding theme of the book is to help you learn to do data analysis using linear regression. Linear regression is a excellent model for learning about data analysis, both because it is important on its own and it provides a framework for understanding other methods of analysis.

This edition of the book includes the majority of the topics in previous editions, although much of the material has been rearranged. New methodology and examples have been added throughout.

- Even more emphasis is placed on graphics. The first two editions stressed graphics for diagnostic methods (Chapter 9) and the third edition added graphics for understanding data before any analysis is done (Chapter 1). In this edition, *effects plots* are stressed to summarize the fit of a model.
- Many applied analyses are based on understanding and interpreting parameters. This edition puts much greater emphasis on parameters, with part of Chapters 2–3 and all of Chapters 4–5 devoted to this important topic.
- Chapter 6 contains a greatly expanded treatment of testing and model comparison using both likelihood ratio and Wald tests. The usefulness and limitations of testing are stressed.
- Chapter 7 is about the variance assumption in linear models. The discussion of weighted least squares has been been expanded to cover problems of ecological regressions, sample surveys, and other cases. Alternatives such as the bootstrap and heteroskedasticity corrections have been added or expanded.
- Diagnostic methods using transformations (Chapter 8) and residuals and related quantities (Chapter 9) that were the heart of the earlier editions have been maintained in this new edition.

- The discussion of variable selection in Chapter 10 has been updated from the third edition. It is designed to help you understand the key problems in variable selection. In recent years, this topic has morphed into the area of *machine learning* and the goal of this chapter is to show connections and provide references.
- As in the third edition, brief introductions to nonlinear regression (Chapter 11) and to logistic regression (Chapter 12) are included, with Poisson regression added in Chapter 12.

Using This Book

The website for this book is `http://z.umn.edu/alr4ed`.

As with previous editions, this book is not tied to any particular computer program. A primer for using the free `R` package (R Core Team, 2013) for the material covered in the book is available from the website. The primer can also be accessed directly from within `R` as you are working. An optional published companion book about `R` is Fox and Weisberg (2011).

All the data files used are available from the website and in an `R` package called `alr4` that you can download for free. Solutions for odd-numbered problems, all using `R`, are available on the website for the book[1]. You cannot learn to do data analysis without working problems.

Some advanced topics are introduced to help you recognize when a problem that looks like linear regression is actually a little different. Detailed methodology is not always presented, but references at the same level as this book are presented. The bibliography, also available with clickable links on the book's website, has been greatly expanded and updated.

Mathematical Level

The mathematical level of this book is roughly the same as the level of previous editions. Matrix representation of data is used, particularly in the derivation of the methodology in Chapters 3–4. Derivations are less frequent in later chapters, and so the necessary mathematics is less. Calculus is generally not required, except for an occasional use of a derivative. The discussions requiring calculus can be skipped without much loss.

ACKNOWLEDGMENTS

Thanks are due to Jeff Witmer, Yuhong Yang, Brad Price, and Brad's Stat 5302 students at the University of Minnesota. New examples were provided by April Bleske-Rechek, Tom Burk, and Steve Taff. Work with John Fox over the last few years has greatly influenced my writing.

For help with previous editions, thanks are due to Charles Anderson, Don Pereira, Christopher Bingham, Morton Brown, Cathy Campbell, Dennis Cook,

[1]All solutions are available to instructors using the book in a course; see the website for details.

Stephen Fienberg, James Frane, Seymour Geisser, John Hartigan, David Hinkley, Alan Izenman, Soren Johansen, Kenneth Koehler, David Lane, Michael Lavine, Kinley Larntz, Gary Oehlert, Katherine St. Clair, Keija Shan, John Rice, Donald Rubin, Joe Shih, Pete Stewart, Stephen Stigler, Douglas Tiffany, Carol Weisberg, and Howard Weisberg.

Finally, I am grateful to Stephen Quigley at Wiley for asking me to do a new edition. I have been working on versions of this book since 1976, and each new edition has pleased me more that the one before it. I hope it pleases you, too.

SANFORD WEISBERG

St. Paul, Minnesota
September 2013

Contents

Preface to the Fourth Edition xv

1 Scatterplots and Regression 1

1.1 Scatterplots, 2
1.2 Mean Functions, 10
1.3 Variance Functions, 12
1.4 Summary Graph, 12
1.5 Tools for Looking at Scatterplots, 13
1.5.1 Size, 14
1.5.2 Transformations, 14
1.5.3 Smoothers for the Mean Function, 14
1.6 Scatterplot Matrices, 15
1.7 Problems, 17

2 Simple Linear Regression 21

2.1 Ordinary Least Squares Estimation, 22
2.2 Least Squares Criterion, 24
2.3 Estimating the Variance σ^2, 26
2.4 Properties of Least Squares Estimates, 27
2.5 Estimated Variances, 29
2.6 Confidence Intervals and t-Tests, 30
2.6.1 The Intercept, 30
2.6.2 Slope, 31
2.6.3 Prediction, 32
2.6.4 Fitted Values, 33
2.7 The Coefficient of Determination, R^2, 35
2.8 The Residuals, 36
2.9 Problems, 38

3 Multiple Regression **51**

3.1 Adding a Regressor to a Simple Linear Regression Model, 51
3.1.1 Explaining Variability, 53
3.1.2 Added-Variable Plots, 53
3.2 The Multiple Linear Regression Model, 55
3.3 Predictors and Regressors, 55
3.4 Ordinary Least Squares, 58
3.4.1 Data and Matrix Notation, 60
3.4.2 The Errors e, 61
3.4.3 Ordinary Least Squares Estimators, 61
3.4.4 Properties of the Estimates, 63
3.4.5 Simple Regression in Matrix Notation, 63
3.4.6 The Coefficient of Determination, 66
3.4.7 Hypotheses Concerning One Coefficient, 67
3.4.8 *t*-Tests and Added-Variable Plots, 68
3.5 Predictions, Fitted Values, and Linear Combinations, 68
3.6 Problems, 69

4 Interpretation of Main Effects **73**

4.1 Understanding Parameter Estimates, 73
4.1.1 Rate of Change, 74
4.1.2 Signs of Estimates, 75
4.1.3 Interpretation Depends on Other Terms in the Mean Function, 75
4.1.4 Rank Deficient and Overparameterized Mean Functions, 78
4.1.5 Collinearity, 79
4.1.6 Regressors in Logarithmic Scale, 81
4.1.7 Response in Logarithmic Scale, 82
4.2 Dropping Regressors, 84
4.2.1 Parameters, 84
4.2.2 Variances, 86
4.3 Experimentation versus Observation, 86
4.3.1 Feedlots, 87
4.4 Sampling from a Normal Population, 89

4.5 More on R^2, 91
4.5.1 Simple Linear Regression and R^2, 91
4.5.2 Multiple Linear Regression and R^2, 92
4.5.3 Regression through the Origin, 93
4.6 Problems, 93

5 Complex Regressors **98**

5.1 Factors, 98
5.1.1 One-Factor Models, 99
5.1.2 Comparison of Level Means, 102
5.1.3 Adding a Continuous Predictor, 103
5.1.4 The Main Effects Model, 106
5.2 Many Factors, 108
5.3 Polynomial Regression, 109
5.3.1 Polynomials with Several Predictors, 111
5.3.2 Numerical Issues with Polynomials, 112
5.4 Splines, 113
5.4.1 Choosing a Spline Basis, 115
5.4.2 Coefficient Estimates, 116
5.5 Principal Components, 116
5.5.1 Using Principal Components, 118
5.5.2 Scaling, 119
5.6 Missing Data, 119
5.6.1 Missing at Random, 120
5.6.2 Imputation, 122
5.7 Problems, 123

6 Testing and Analysis of Variance **133**

6.1 *F*-Tests, 134
6.1.1 General Likelihood Ratio Tests, 138
6.2 The Analysis of Variance, 138
6.3 Comparisons of Means, 142
6.4 Power and Non-Null Distributions, 143
6.5 Wald Tests, 145
6.5.1 One Coefficient, 145
6.5.2 One Linear Combination, 146
6.5.3 General Linear Hypothesis, 146
6.5.4 Equivalence of Wald and Likelihood-Ratio Tests, 146

6.6 Interpreting Tests, 146
6.6.1 Interpreting p-Values, 146
6.6.2 Why Most Published Research Findings Are False, 147
6.6.3 Look at the Data, Not Just the Tests, 148
6.6.4 Population versus Sample, 149
6.6.5 Stacking the Deck, 149
6.6.6 Multiple Testing, 150
6.6.7 File Drawer Effects, 150
6.6.8 The Lab Is Not the Real World, 150
6.7 Problems, 150

7 Variances **156**

7.1 Weighted Least Squares, 156
7.1.1 Weighting of Group Means, 159
7.1.2 Sample Surveys, 161
7.2 Misspecified Variances, 162
7.2.1 Accommodating Misspecified Variance, 163
7.2.2 A Test for Constant Variance, 164
7.3 General Correlation Structures, 168
7.4 Mixed Models, 169
7.5 Variance Stabilizing Transformations, 171
7.6 The Delta Method, 172
7.7 The Bootstrap, 174
7.7.1 Regression Inference without Normality, 175
7.7.2 Nonlinear Functions of Parameters, 178
7.7.3 Residual Bootstrap, 179
7.7.4 Bootstrap Tests, 179
7.8 Problems, 179

8 Transformations **185**

8.1 Transformation Basics, 185
8.1.1 Power Transformations, 186
8.1.2 Transforming One Predictor Variable, 188
8.1.3 The Box–Cox Method, 190
8.2 A General Approach to Transformations, 191
8.2.1 The 1D Estimation Result and Linearly Related Regressors, 194
8.2.2 Automatic Choice of Transformation of Predictors, 195

8.3 Transforming the Response, 196
8.4 Transformations of Nonpositive Variables, 198
8.5 Additive Models, 199
8.6 Problems, 199

9 Regression Diagnostics **204**

9.1 The Residuals, 204
9.1.1 Difference between $\hat{e}$ and e, 205
9.1.2 The Hat Matrix, 206
9.1.3 Residuals and the Hat Matrix with Weights, 208
9.1.4 Residual Plots When the Model Is Correct, 209
9.1.5 The Residuals When the Model Is Not Correct, 209
9.1.6 Fuel Consumption Data, 211
9.2 Testing for Curvature, 212
9.3 Nonconstant Variance, 213
9.4 Outliers, 214
9.4.1 An Outlier Test, 215
9.4.2 Weighted Least Squares, 216
9.4.3 Significance Levels for the Outlier Test, 217
9.4.4 Additional Comments, 218
9.5 Influence of Cases, 218
9.5.1 Cook's Distance, 220
9.5.2 Magnitude of D_i, 221
9.5.3 Computing D_i, 221
9.5.4 Other Measures of Influence, 224
9.6 Normality Assumption, 225
9.7 Problems, 226

10 Variable Selection **234**

10.1 Variable Selection and Parameter Assessment, 235
10.2 Variable Selection for Discovery, 237
10.2.1 Information Criteria, 238
10.2.2 Stepwise Regression, 239
10.2.3 Regularized Methods, 244
10.2.4 Subset Selection Overstates Significance, 245
10.3 Model Selection for Prediction, 245
10.3.1 Cross-Validation, 247
10.3.2 Professor Ratings, 247
10.4 Problems, 248

11 Nonlinear Regression **252**

11.1 Estimation for Nonlinear Mean Functions, 253
11.2 Inference Assuming Large Samples, 256
11.3 Starting Values, 257
11.4 Bootstrap Inference, 262
11.5 Further Reading, 265
11.6 Problems, 265

12 Binomial and Poisson Regression **270**

12.1 Distributions for Counted Data, 270
12.1.1 Bernoulli Distribution, 270
12.1.2 Binomial Distribution, 271
12.1.3 Poisson Distribution, 271
12.2 Regression Models for Counts, 272
12.2.1 Binomial Regression, 272
12.2.2 Deviance, 277
12.3 Poisson Regression, 279
12.3.1 Goodness of Fit Tests, 282
12.4 Transferring What You Know about Linear Models, 283
12.4.1 Scatterplots and Regression, 283
12.4.2 Simple and Multiple Regression, 283
12.4.3 Model Building, 284
12.4.4 Testing and Analysis of Deviance, 284
12.4.5 Variances, 284
12.4.6 Transformations, 284
12.4.7 Regression Diagnostics, 284
12.4.8 Variable Selection, 285
12.5 Generalized Linear Models, 285
12.6 Problems, 285

Appendix **290**

A.1 Website, 290
A.2 Means, Variances, Covariances, and Correlations, 290
A.2.1 The Population Mean and E Notation, 290
A.2.2 Variance and Var Notation, 291
A.2.3 Covariance and Correlation, 291
A.2.4 Conditional Moments, 292
A.3 Least Squares for Simple Regression, 293

A.4 Means and Variances of Least Squares Estimates, 294
A.5 Estimating E($Y|X$) Using a Smoother, 296
A.6 A Brief Introduction to Matrices and Vectors, 298
A.6.1 Addition and Subtraction, 299
A.6.2 Multiplication by a Scalar, 299
A.6.3 Matrix Multiplication, 299
A.6.4 Transpose of a Matrix, 300
A.6.5 Inverse of a Matrix, 301
A.6.6 Orthogonality, 302
A.6.7 Linear Dependence and Rank of a Matrix, 303
A.7 Random Vectors, 303
A.8 Least Squares Using Matrices, 304
A.8.1 Properties of Estimates, 305
A.8.2 The Residual Sum of Squares, 305
A.8.3 Estimate of Variance, 306
A.8.4 Weighted Least Squares, 306
A.9 The QR Factorization, 307
A.10 Spectral Decomposition, 309
A.11 Maximum Likelihood Estimates, 309
A.11.1 Linear Models, 309
A.11.2 Logistic Regression, 311
A.12 The Box–Cox Method for Transformations, 312
A.12.1 Univariate Case, 312
A.12.2 Multivariate Case, 313
A.13 Case Deletion in Linear Regression, 314

References 317

Author Index 329

Subject Index 331

CHAPTER 1

Scatterplots and Regression

Regression is the study of dependence. It is used to answer interesting questions about how one or more predictors influence a response. Here are a few typical questions that may be answered using regression:

- Are daughters taller than their mothers?
- Does changing class size affect success of students?
- Can we predict the time of the next eruption of Old Faithful Geyser from the length of the most recent eruption?
- Do changes in diet result in changes in cholesterol level, and if so, do the results depend on other characteristics such as age, sex, and amount of exercise?
- Do countries with higher per person income have lower birth rates than countries with lower income?
- Are highway design characteristics associated with highway accident rates? Can accident rates be lowered by changing design characteristics?
- Is water usage increasing over time?
- Do conservation easements on agricultural property lower land value?

In most of this book, we study the important instance of regression methodology called *linear regression*. This method is the most commonly used in regression, and virtually all other regression methods build upon an understanding of how linear regression works.

As with most statistical analyses, the goal of regression is to summarize observed data as simply, usefully, and elegantly as possible. A theory may be available in some problems that specifies how the response varies as the values

Applied Linear Regression, Fourth Edition. Sanford Weisberg.