

Applied Regression Analysis

WILEY SERIES IN PROBABILITY AND STATISTICS
TEXTS AND REFERENCES SECTION

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *Vic Barnett, Ralph A. Bradley, Noel A. C. Cressie, Nicholas I. Fisher,
Iain M. Johnstone, J. B. Kadane, David G. Kendall, David W. Scott,
Bernard W. Silverman, Adrian F. M. Smith, Jozef L. Teugels, Geoffrey S. Watson;
J. Stuart Hunter, Emeritus*

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

Applied Regression Analysis

THIRD EDITION

Norman R. Draper

Harry Smith



A Wiley-Interscience Publication

JOHN WILEY & SONS, INC.

New York · Chichester · Weinheim · Brisbane · Singapore · Toronto

This book is printed on acid-free paper. ∞

Copyright © 1998 by John Wiley & Sons, Inc. All rights reserved.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008.

Library of Congress Cataloging-in-Publication Data:

Draper, Norman Richard.

Applied regression analysis / N.R. Draper, H. Smith. — 3rd ed.

p. cm. — (Wiley series in probability and statistics. Texts and references section)

“A Wiley-Interscience publication.”

Includes bibliographical references (p. —) and index.

ISBN 0-471-17082-8 (acid-free paper)

1. Regression analysis. I. Smith, Harry, 1923– . II. Title.

III. Series.

QA278.2.D7 1998

519.5'36—dc21

97-17969

CIP

Printed in the United States of America.

Contents

Preface	xiii
About the Software	xvii
0 Basic Prerequisite Knowledge	1
0.1 Distributions: Normal, t , and F , 1	
0.2 Confidence Intervals (or Bands) and t -Tests, 4	
0.3 Elements of Matrix Algebra, 6	
1 Fitting a Straight Line by Least Squares	15
1.0 Introduction: The Need for Statistical Analysis, 15	
1.1 Straight Line Relationship Between Two Variables, 18	
1.2 Linear Regression: Fitting a Straight Line by Least Squares, 20	
1.3 The Analysis of Variance, 28	
1.4 Confidence Intervals and Tests for β_0 and β_1 , 34	
1.5 F -Test for Significance of Regression, 38	
1.6 The Correlation Between X and Y , 40	
1.7 Summary of the Straight Line Fit Computations, 44	
1.8 Historical Remarks, 45	
Appendix 1A Steam Plant Data, 46	
Exercises are in “Exercises for Chapters 1–3”, 96	
2 Checking the Straight Line Fit	47
2.1 Lack of Fit and Pure Error, 47	
2.2 Testing Homogeneity of Pure Error, 56	
2.3 Examining Residuals: The Basic Plots, 59	
2.4 Non-normality Checks on Residuals, 61	
2.5 Checks for Time Effects, Nonconstant Variance, Need for Transformation, and Curvature, 62	
2.6 Other Residuals Plots, 67	

- 2.7 Durbin–Watson Test, 69
- 2.8 Reference Books for Analysis of Residuals, 70
- Appendix 2A Normal Plots, 70
- Appendix 2B MINITAB Instructions, 76
- Exercises are in “Exercises for Chapters 1–3”, 96

3 Fitting Straight Lines: Special Topics **79**

- 3.0 Summary and Preliminaries, 79
- 3.1 Standard Error of \hat{Y} , 80
- 3.2 Inverse Regression (Straight Line Case), 83
- 3.3 Some Practical Design of Experiment Implications of Regression, 86
- 3.4 Straight Line Regression When Both Variables Are Subject to Error, 89
- Exercises for Chapters 1–3, 96

4 Regression in Matrix Terms: Straight Line Case **115**

- 4.1 Fitting a Straight Line in Matrix Terms, 115
- 4.2 Singularity: What Happens in Regression to Make $\mathbf{X}'\mathbf{X}$ Singular? An Example, 125
- 4.3 The Analysis of Variance in Matrix Terms, 127
- 4.4 The Variances and Covariance of b_0 and b_1 from the Matrix Calculation, 128
- 4.5 Variance of \hat{Y} Using the Matrix Development, 130
- 4.6 Summary of Matrix Approach to Fitting a Straight Line (Nonsingular Case), 130
- 4.7 The General Regression Situation, 131
- Exercises for Chapter 4, 132

5 The General Regression Situation **135**

- 5.1 General Linear Regression, 135
- 5.2 Least Squares Properties, 137
- 5.3 Least Squares Properties When $\epsilon \sim N(0, \mathbf{I}\sigma^2)$, 140
- 5.4 Confidence Intervals Versus Regions, 142
- 5.5 More on Confidence Intervals Versus Regions, 143
- Appendix 5A Selected Useful Matrix Results, 147
- Exercises are in “Exercises for Chapters 5 and 6”, 169

6 Extra Sums of Squares and Tests for Several Parameters Being Zero **149**

- 6.1 The “Extra Sum of Squares” Principle, 149
- 6.2 Two Predictor Variables: Example, 154
- 6.3 Sum of Squares of a Set of Linear Functions of Y ’s, 162

Appendix 6A	Orthogonal Columns in the \mathbf{X} Matrix, 165	
Appendix 6B	Two Predictors: Sequential Sums of Squares, 167	
	Exercises for Chapters 5 and 6, 169	
7	Serial Correlation in the Residuals and the Durbin–Watson Test	179
7.1	Serial Correlation in Residuals, 179	
7.2	The Durbin–Watson Test for a Certain Type of Serial Correlation, 181	
7.3	Examining Runs in the Time Sequence Plot of Residuals: Runs Test, 192	
	Exercises for Chapter 7, 198	
8	More on Checking Fitted Models	205
8.1	The Hat Matrix \mathbf{H} and the Various Types of Residuals, 205	
8.2	Added Variable Plot and Partial Residuals, 209	
8.3	Detection of Influential Observations: Cook’s Statistics, 210	
8.4	Other Statistics Measuring Influence, 214	
8.5	Reference Books for Analysis of Residuals, 214	
	Exercises for Chapter 8, 215	
9	Multiple Regression: Special Topics	217
9.1	Testing a General Linear Hypothesis, 217	
9.2	Generalized Least Squares and Weighted Least Squares, 221	
9.3	An Example of Weighted Least Squares, 224	
9.4	A Numerical Example of Weighted Least Squares, 226	
9.5	Restricted Least Squares, 229	
9.6	Inverse Regression (Multiple Predictor Case), 229	
9.7	Planar Regression When All the Variables Are Subject to Error, 231	
	Appendix 9A Lagrange’s Undetermined Multipliers, 231	
	Exercises for Chapter 9, 233	
10	Bias in Regression Estimates, and Expected Values of Mean Squares and Sums of Squares	235
10.1	Bias in Regression Estimates, 235	
10.2	The Effect of Bias on the Least Squares Analysis of Variance, 238	
10.3	Finding the Expected Values of Mean Squares, 239	
10.4	Expected Value of Extra Sum of Squares, 240	
	Exercises for Chapter 10, 241	
11	On Worthwhile Regressions, Big F’s, and R^2	243
11.1	Is My Regression a Useful One?, 243	
11.2	A Conversation About R^2 , 245	

Appendix 11A How Significant Should My Regression Be?, 247	
Exercises for Chapter 11, 250	
12 Models Containing Functions of the Predictors, Including Polynomial Models	251
12.1 More Complicated Model Functions, 251	
12.2 Worked Examples of Second-Order Surface Fitting for $k = 3$ and $k = 2$ Predictor Variables, 254	
12.3 Retaining Terms in Polynomial Models, 266	
Exercises for Chapter 12, 272	
13 Transformation of the Response Variable	277
13.1 Introduction and Preliminary Remarks, 277	
13.2 Power Family of Transformations on the Response: Box–Cox Method, 280	
13.3 A Second Method for Estimation λ , 286	
13.4 Response Transformations: Other Interesting and Sometimes Useful Plots, 289	
13.5 Other Types of Response Transformations, 290	
13.6 Response Transformations Chosen to Stabilize Variance, 291	
Exercises for Chapter 13, 294	
14 “Dummy” Variables	299
14.1 Dummy Variables to Separate Blocks of Data with Different Intercepts, Same Model, 299	
14.2 Interaction Terms Involving Dummy Variables, 307	
14.3 Dummy Variables for Segmented Models, 311	
Exercises for Chapter 14, 317	
15 Selecting the “Best” Regression Equation	327
15.0 Introduction, 327	
15.1 All Possible Regressions and “Best Subset” Regression, 329	
15.2 Stepwise Regression, 335	
15.3 Backward Elimination, 339	
15.4 Significance Levels for Selection Procedures, 342	
15.5 Variations and Summary, 343	
15.6 Selection Procedures Applied to the Steam Data, 345	
Appendix 15A Hald Data, Correlation Matrix, and All 15 Possible Regressions, 348	
Exercises for Chapter 15, 355	
16 Ill-Conditioning in Regression Data	369
16.1 Introduction, 369	
16.2 Centering Regression Data, 371	

- 16.3 Centering and Scaling Regression Data, 373
- 16.4 Measuring Multicollinearity, 375
- 16.5 Belsley's Suggestion for Detecting Multicollinearity, 376
- Appendix 16A Transforming \mathbf{X} Matrices to Obtain Orthogonal Columns, 382
- Exercises for Chapter 16, 385

17 Ridge Regression 387

- 17.1 Introduction, 387
- 17.2 Basic Form of Ridge Regression, 387
- 17.3 Ridge Regression of the Hald Data, 389
- 17.4 In What Circumstances Is Ridge Regression Absolutely the Correct Way to Proceed?, 391
- 17.5 The Phoney Data Viewpoint, 394
- 17.6 Concluding Remarks, 395
- Appendix 17A Ridge Estimates in Terms of Least Squares Estimates, 396
- Appendix 17B Mean Square Error Argument, 396
- Appendix 17C Canonical Form of Ridge Regression, 397
- Exercises for Chapter 17, 400

18 Generalized Linear Models (GLIM) 401

- 18.1 Introduction, 401
- 18.2 The Exponential Family of Distributions, 402
- 18.3 Fitting Generalized Linear Models (GLIM), 404
- 18.4 Performing the Calculations: An Example, 406
- 18.5 Further Reading, 408
- Exercises for Chapter 18, 408

19 Mixture Ingredients as Predictor Variables 409

- 19.1 Mixture Experiments: Experimental Spaces, 409
- 19.2 Models for Mixture Experiments, 412
- 19.3 Mixture Experiments in Restricted Regions, 416
- 19.4 Example 1, 418
- 19.5 Example 2, 419
- Appendix 19A Transforming k Mixture Variables to $k - 1$ Working Variables, 422
- Exercises for Chapter 19, 425

20 The Geometry of Least Squares 427

- 20.1 The Basic Geometry, 427
- 20.2 Pythagoras and Analysis of Variance, 429
- 20.3 Analysis of Variance and F -Test for Overall Regression, 432
- 20.4 The Singular $\mathbf{X}'\mathbf{X}$ Case: An Example, 433

- 20.5 Orthogonalizing in the General Regression Case, 435
- 20.6 Range Space and Null Space of a Matrix \mathbf{M} , 437
- 20.7 The Algebra and Geometry of Pure Error, 439
- Appendix 20A Generalized Inverses \mathbf{M}^+ , 441
- Exercises for Chapter 20, 444

21 More Geometry of Least Squares 447

- 21.1 The Geometry of a Null Hypothesis: A Simple Example, 447
- 21.2 General Case $H_0: \mathbf{A}\boldsymbol{\beta} = \mathbf{c}$: The Projection Algebra, 448
- 21.3 Geometric Illustrations, 449
- 21.4 The F -Test for H_0 , Geometrically, 450
- 21.5 The Geometry of R^2 , 452
- 21.6 Change in R^2 for Models Nested Via $\mathbf{A}\boldsymbol{\beta} = \mathbf{0}$, Not Involving β_0 , 452
- 21.7 Multiple Regression with Two Predictor Variables as a Sequence of Straight Line Regressions, 454
- Exercises for Chapter 21, 459

22 Orthogonal Polynomials and Summary Data 461

- 22.1 Introduction, 461
- 22.2 Orthogonal Polynomials, 461
- 22.3 Regression Analysis of Summary Data, 467
- Exercises for Chapter 22, 469

23 Multiple Regression Applied to Analysis of Variance Problems 473

- 23.1 Introduction, 473
- 23.2 The One-Way Classification: Standard Analysis and an Example, 474
- 23.3 Regression Treatment of the One-Way Classification Example, 477
- 23.4 Regression Treatment of the One-Way Classification Using the Original Model, 481
- 23.5 Regression Treatment of the One-Way Classification: Independent Normal Equations, 486
- 23.6 The Two-Way Classification with Equal Numbers of Observations in the Cells: An Example, 488
- 23.7 Regression Treatment of the Two-Way Classification Example, 489
- 23.8 The Two-Way Classification with Equal Numbers of Observations in the Cells, 493
- 23.9 Regression Treatment of the Two-Way Classification with Equal Numbers of Observations in the Cells, 494
- 23.10 Example: The Two-Way Classification, 498

23.11 Recapitulation and Comments, 499

Exercises for Chapter 23, 500

24 An Introduction to Nonlinear Estimation 505

24.1 Least Squares for Nonlinear Models, 505

24.2 Estimating the Parameters of a Nonlinear System, 508

24.3 An Example, 518

24.4 A Note on Reparameterization of the Model, 529

24.5 The Geometry of Linear Least Squares, 530

24.6 The Geometry of Nonlinear Least Squares, 539

24.7 Nonlinear Growth Models, 543

24.8 Nonlinear Models: Other Work, 550

24.9 References, 553

Exercises for Chapter 24, 553

25 Robust Regression 567

25.1 Least Absolute Deviations Regression (L_1 Regression), 567

25.2 M -Estimators, 567

25.3 Steel Employment Example, 573

25.4 Trees Example, 575

25.5 Least Median of Squares (LMS) Regression, 577

25.6 Robust Regression with Ranked Residuals (rreg), 577

25.7 Other Methods, 580

25.8 Comments and Opinions, 580

25.9 References, 581

Exercises for Chapter 25, 584

26 Resampling Procedures (Bootstrapping) 585

26.1 Resampling Procedures for Regression Models, 585

26.2 Example: Straight Line Fit, 586

26.3 Example: Planar Fit, Three Predictors, 588

26.4 Reference Books, 588

Appendix 26A Sample MINITAB Programs to Bootstrap Residuals
for a Specific Example, 589

Appendix 26B Sample MINITAB Programs to Bootstrap Pairs for a
Specific Example, 590

Additional Comments, 591

Exercises for Chapter 26, 591

Bibliography 593

True/False Questions 605

Answers to Exercises 609

Tables	684
Normal Distribution, 684	
Percentage Points of the t -Distribution, 686	
Percentage Points of the χ^2 -Distribution, 687	
Percentage Points of the F -Distribution, 688	
Index of Authors Associated with Exercises	695
Index	697

Preface to the Third Edition

The second edition had 10 chapters; this edition has 26. On the whole (but not entirely) we have chosen to use smaller chapters, and so distinguish more between different types of material. The tabulation below shows the major relationships between second edition and third edition sections and chapters.

Material dropped consists mainly of second edition Sections 6.8 to 6.13 and 6.15, Sections 7.1 to 7.6, and Chapter 8. New to this edition are Chapters 16 on multicollinearity, 18 on generalized linear models, 19 on mixture ingredients, 20 and 21 on the geometry of least squares, 25 on robust regression, and 26 on resampling procedures. Small revisions have been made even in sections where the text is basically unchanged. Less prominence has been given to printouts, which nowadays can easily be generated due to the excellent software available, and to references and bibliography, which are now freely available (either in book or computer form) via the annual updates in *Current Index to Statistics*. References are mostly given in brief either in situ or close by, at the end of a section or chapter. Full references are in a bibliography but some references are also given in full in sections or within the text or in exercises, whenever this was felt to be the appropriate thing to do. There is no precise rule for doing this, merely the authors' predilection. Exercises have been grouped as seemed appropriate. They are intended as an expansion to the text and so most exercises have full or partial solutions; there are a very few exceptions. One hundred and one true/false questions have also been provided; all of these are in "true" form to prevent readers remembering erroneous material. Instructors can reword them to create "false" questions easily enough. Sections 24.5 and 24.6 have some duplication with work in Chapter 20, but we decided not to eliminate this because the sections contain some differences and have different emphases. Other smaller duplications occur; in general, we feel that duplication is a good feature, and so we do not avoid it.

Our viewpoint in putting this book together is that it is desirable for students of regression to work through the straight line fit case using a pocket calculator and then to proceed quickly to analyzing larger models on the computer. We are aware that many instructors like to get on to the computer right away. Our personal experience is that this can be unwise and, over the years, we have met many students who enrolled for our courses saying "I know how to put a regression on the computer but I don't understand what I am doing." We have tried to keep such participants constantly in mind.

We have made no effort to explain any of the dozens of available computing systems. Most of our specific references to these were removed after we received reviews of an earlier draft. Reviewers suggested we delete certain specifics and replace them by others. Unfortunately, the reviewers disagreed on the specifics! In addition, many specific program versions quickly become obsolete as new versions are issued. Quite often students point out to us in class that “the new version of BLANK does (or doesn’t!) do that now.” For these reasons we have tried to stay away from advocating any particular way to handle computations. A few mild references to MINITAB (used in our University of Wisconsin classes) have been retained but readers will find it easy to ignore these, if they wish.

We are grateful for help from a number of people, many of these connected with N. R. Draper at the University of Wisconsin. Teaching assistants contributed in many ways, by working new assignments, providing class notes of lectures spoken but not recorded, and discussing specific problems. Former University of Wisconsin student Dennis K. J. Lin, now a faculty member at Pennsylvania State University, contributed most in this regard. More generally, we profited from teaching for many years from the excellent Wiley book *Linear Regression Analysis*, by George A. F. Seber, whose detailed algebraic treatment has clearly influenced the geometrical presentations of Chapters 20 and 21.

N. R. Draper is grateful to the University of Wisconsin and to his colleagues there for a timely sabbatical leave, and to Professor Friedrich Pukelsheim of the University of Augsburg, Germany, for inviting him to spend the leave there, providing full technical facilities and many unexpected kindnesses as well. Support from the German Alexander von Humboldt Stiftung is also gratefully acknowledged. N. R. Draper is also thankful to present and former faculty and staff at the University of Southampton, particularly Fred (T. M. F.) Smith, Nye (J. A.) John (now at Waikato University, New Zealand), Sue Lewis, Phil Prescott, and Daphne Turner, all of whom have made him most welcome on annual visits for many years. The enduring influence of R. C. Bose (1901–1987) is also gratefully acknowledged.

The staff at the Statistics Department, Mary Esser (staff supervisor, retired), Candy Smith, Mary Ann Clark (retired), Wanda Gray (retired), and Gloria Scalissi, have all contributed over the years. Our special thanks go to Gloria Scalissi who typed much of a difficult and intricate manuscript.

For John Wiley & Sons, the effects of Bea Shube’s help and wisdom linger on, supplemented more recently by those of Kate Roach, Jessica Downey, and Steve Quigley. We also thank Alison Bory on the editorial side and Production Editor Lisa Van Horn for their patience and skills in the final stages.

We are grateful to all of our reviewers, including David Belsley and Richard (Rick) Chappell and several anonymous ones. The reviews were all very helpful and we followed up most of the suggestions made, but not all. We ourselves have often profited by reading varying presentations in different places and so we sometimes resisted changing our presentation to conform to presentations elsewhere.

Many others contributed with correspondence or conversation over the years. We do not have a complete list, but some of them were Cuthbert Daniel, Jim Durbin, Xiaoyin (Frank) Fan, Conrad Fung, Stratis Gavaris, Michael Haber, Brian Joiner, Jane Kawasaki, Russell Langley, A. G. C. Morris, Ella Munro, Vedula N. Murty, Alvin P. Rainosek, J. Harold Ranck, Guangheng (Sharon) Shen, Jake Sredni, Daniel Weiner, William J. Welch, Yonghong (Fred) Yang, Yuyun (Jessie) Yang, and Lisa Ying. Others are mentioned within the text, where appropriate. We are grateful to them all.

To notify us of errors or misprints, please e-mail to draper@stat.wisc.edu. An updated list of such discrepancies will be returned e-mail, if requested. For a hardcopy of the list, please send a stamped addressed envelope to N. R. Draper, University of Wisconsin Statistics Department, 1210 West Dayton Street, Madison, WI 53706, U.S.A.

NORMAN R. DRAPER
HARRY SMITH

Relationships of Second Edition and Third Edition Text Material

Topic	Sections		Topic	Sections	
	Second Edition	Third Edition		Second Edition	Third Edition
Straight line fit	1.0–1.4	1.0–1.5	Polynomial models	5.1, 5.2	12.1
Pure error	1.5	2.1–2.2	Transformations	5.3	13
Correlation	1.6	1.6	Dummy variables	5.4	14
Inverse regression	1.7	3.2	Centering and scaling	5.5	16.2, 16.3
Practical implications	1.8	3.3	Orthogonal polynomials	5.6	22.2
			Orthogonalizing \mathbf{X}	5.7	16A
			Summary data	5.8	22.3
Straight line, matrices	2.0–2.5	4			
General regression	2.6	5	Selection procedures	6.0–6.6, 6.12	15
Extra SS	2.7–2.9	6.1, 6.2, 6A	Ridge regression	6.7	17
General linear hypothesis	2.10	9.1	Ridge, canonical form	6A	17A
Weighted least squares	2.11	9.2, 9.3, 9.4	Press	6.8	—
Restricted least squares	2.13	9.5	Principal components	6.9	—
Inverse regression	2.15	9.6	Latent root regression	6.10	—
Errors in multiple X 's	—	9.7	Stagewise regression	6.13	—
Bias in estimates	2.12	10	Robust regression	6.14	25
Errors in X and Y	2.14	3.4			
Inverse regression	2.15	9.6	Data example	7.0–7.6	—
Matrix results	2A	5A	Polynomial example	7.7	12.2
E (Extra SS)	2B	10.4			
How significant?	2C	11	Model building talk	8	—
Lagrange's multipliers	2D	9A			
			ANOVA models	9	23
Residuals plots	3.1–3.8	2	Nonlinear estimation	10	24
Serial correlation	3.9–3.11	7			
Influential observations	3.12	8.3, 8.4	Multicollinearity	—	16
Normal plots	3A	2A	GLIM	—	18
Two X 's example	4.0, 4.2	6.3	Mixtures models	—	19
Geometry	4.1	21.7	Geometry of LS	—	20
			More geometry	—	21.1–21.6
			Robust regression	—	25
			Resampling methods	—	26

About the Software

The diskette that accompanies the book includes data files for the examples used in the chapters and for the exercises. These files can be used as input for standard statistical analysis programs. When writing program scripts, please note that descriptive text lines are included above data sections in the files.

The data files are included in the REGRESS directory on the diskette, which can be placed on your hard drive by your computer operating system's usual copying methods. You can also use the installation program on the diskette to copy the files by doing the following.

1. Type `a:install` at the Run selection of the File menu in a Windows 3.1 system or access the floppy drive directory through a Windows file manager and double click on the `INSTALL.EXE` file.
2. After skipping through the introductory screens, select a path for installing the files. The default directory for the file installation is `C:\REGRESS`. You may edit this selection to choose a different drive or directory. Press Enter when done.
3. The files will be installed to the selected directory.

Applied Regression Analysis