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. ♾

Published simultaneously in Canada.

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