

Wiley Series in Probability and Statistics

Introduction to
**Linear
Regression
Analysis**

FIFTH EDITION

DOUGLAS C. MONTGOMERY
ELIZABETH A. PECK
G. GEOFFREY VINING

 WILEY

ftp://
SITE AVAILABLE

INTRODUCTION TO LINEAR REGRESSION ANALYSIS

Fifth Edition

DOUGLAS C. MONTGOMERY

Arizona State University
School of Computing, Informatics, and Decision Systems Engineering
Tempe, AZ

ELIZABETH A. PECK

The Coca-Cola Company (retired)
Atlanta, GA

G. GEOFFREY VINING

Virginia Tech
Department of Statistics
Blacksburg, VA



A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2012 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
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 Section 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, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. 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, or online at <http://www.wiley.com/go/permissions>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Montgomery, Douglas C.

Introduction to linear regression analysis / Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining. – 5th ed.

p. cm. – (Wiley series in probability and statistics ; 821)

Includes bibliographical references and index.

ISBN 978-0-470-54281-1 (hardback)

1. Regression analysis. I. Peck, Elizabeth A., 1953– II. Vining, G. Geoffrey, 1954– III. Title. QA278.2.M65 2012

519.5'36–dc23

2012003859

Printed in the United States of America.

10 9 8 7 6 5 4 3

INTRODUCTION TO LINEAR REGRESSION ANALYSIS

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.

PREFACE

Regression analysis is one of the most widely used techniques for analyzing multi-factor data. Its broad appeal and usefulness result from the conceptually logical process of using an equation to express the relationship between a variable of interest (the response) and a set of related predictor variables. Regression analysis is also interesting theoretically because of elegant underlying mathematics and a well-developed statistical theory. Successful use of regression requires an appreciation of both the theory and the practical problems that typically arise when the technique is employed with real-world data.

This book is intended as a text for a basic course in regression analysis. It contains the standard topics for such courses and many of the newer ones as well. It blends both theory and application so that the reader will gain an understanding of the basic principles necessary to apply regression model-building techniques in a wide variety of application environments. The book began as an outgrowth of notes for a course in regression analysis taken by seniors and first-year graduate students in various fields of engineering, the chemical and physical sciences, statistics, mathematics, and management. We have also used the material in many seminars and industrial short courses for professional audiences. We assume that the reader has taken a first course in statistics and has familiarity with hypothesis tests and confidence intervals and the normal, t , χ^2 , and F distributions. Some knowledge of matrix algebra is also necessary.

The computer plays a significant role in the modern application of regression. Today even spreadsheet software has the capability to fit regression equations by least squares. Consequently, we have integrated many aspects of computer usage into the text, including displays of both tabular and graphical output, and general discussions of capabilities of some software packages. We use Minitab®, JMP®, SAS®, and R for various problems and examples in the text. We selected these packages because they are widely used both in practice and in teaching regression and they have good regression. Many of the homework problems require software

for their solution. All data sets in the book are available in electronic form from the publisher. The ftp site ftp://ftp.wiley.com/public/sci_tech_med/introduction_linear_regression hosts the data, problem solutions, PowerPoint files, and other material related to the book.

CHANGES IN THE FIFTH EDITION

We have made extensive changes in this edition of the book. This includes the reorganization of text material, new examples, new exercises, a new chapter on time series regression, and new material on designed experiments for regression models. Our objective was to make the book more useful as both a text and a reference and to update our treatment of certain topics.

Chapter 1 is a general introduction to regression modeling and describes some typical applications of regression. Chapters 2 and 3 provide the standard results for least-squares model fitting in simple and multiple regression, along with basic inference procedures (tests of hypotheses, confidence and prediction intervals). Chapter 4 discusses some introductory aspects of model adequacy checking, including residual analysis and a strong emphasis on residual plots, detection and treatment of outliers, the PRESS statistic, and testing for lack of fit. Chapter 5 discusses how transformations and weighted least squares can be used to resolve problems of model inadequacy or to deal with violations of the basic regression assumptions. Both the Box–Cox and Box–Tidwell techniques for analytically specifying the form of a transformation are introduced. Influence diagnostics are presented in Chapter 6, along with an introductory discussion of how to deal with influential observations. Polynomial regression models and their variations are discussed in Chapter 7. Topics include the basic procedures for fitting and inference for polynomials and discussion of centering in polynomials, hierarchy, piecewise polynomials, models with both polynomial and trigonometric terms, orthogonal polynomials, an overview of response surfaces, and an introduction to nonparametric and smoothing regression techniques. Chapter 8 introduces indicator variables and also makes the connection between regression and analysis-of-variance models. Chapter 9 focuses on the multicollinearity problem. Included are discussions of the sources of multicollinearity, its harmful effects, diagnostics, and various remedial measures. We introduce biased estimation, including ridge regression and some of its variations and principal-component regression. Variable selection and model-building techniques are developed in Chapter 10, including stepwise procedures and all-possible-regressions. We also discuss and illustrate several criteria for the evaluation of subset regression models. Chapter 11 presents a collection of techniques useful for regression model validation.

The first 11 chapters are the nucleus of the book. Many of the concepts and examples flow across these chapters. The remaining four chapters cover a variety of topics that are important to the practitioner of regression, and they can be read independently. Chapter 12 introduces nonlinear regression, and Chapter 13 is a basic treatment of generalized linear models. While these are perhaps not standard topics for a linear regression textbook, they are so important to students and professionals in engineering and the sciences that we would have been seriously remiss without giving an introduction to them. Chapter 14 covers regression

models for time series data. Chapter 15 includes a survey of several important topics, including robust regression, the effect of measurement errors in the regressors, the inverse estimation or calibration problem, bootstrapping regression estimates, classification and regression trees, neural networks, and designed experiments for regression.

In addition to the text material, Appendix C contains brief presentations of some additional topics of a more technical or theoretical nature. Some of these topics will be of interest to specialists in regression or to instructors teaching a more advanced course from the book. Computing plays an important role in many regression courses. Minitab, JMP, SAS, and R are widely used in regression courses. Outputs from all of these packages are provided in the text. Appendix D is an introduction to using SAS for regression problems. Appendix E is an introduction to R.

USING THE BOOK AS A TEXT

Because of the broad scope of topics, this book has great flexibility as a text. For a first course in regression, we would recommend covering Chapters 1 through 10 in detail and then selecting topics that are of specific interest to the audience. For example, one of the authors (D.C.M.) regularly teaches a course in regression to an engineering audience. Topics for that audience include nonlinear regression (because mechanistic models that are almost always nonlinear occur often in engineering), a discussion of neural networks, and regression model validation. Other topics that we would recommend for consideration are multicollinearity (because the problem occurs so often) and an introduction to generalized linear models focusing mostly on logistic regression. G.G.V. has taught a regression course for graduate students in statistics that makes extensive use of the Appendix C material.

We believe the computer should be directly integrated into the course. In recent years, we have taken a notebook computer and computer projector to most classes and illustrated the techniques as they are introduced in the lecture. We have found that this greatly facilitates student understanding and appreciation of the techniques. We also require that the students use regression software for solving the homework problems. In most cases, the problems use real data or are based on real-world settings that represent typical applications of regression.

There is an instructor's manual that contains solutions to all exercises, electronic versions of all data sets, and questions/problems that might be suitable for use on examinations.

ACKNOWLEDGMENTS

We would like to thank all the individuals who provided helpful feedback and assistance in the preparation of this book. Dr. Scott M. Kowalski, Dr. Ronald G. Askin, Dr. Mary Sue Younger, Dr. Russell G. Heikes, Dr. John A. Cornell, Dr. André I. Khuri, Dr. George C. Runger, Dr. Marie Gaudard, Dr. James W. Wisnowski, Dr. Ray Hill, and Dr. James R. Simpson made many suggestions that greatly improved both earlier editions and this fifth edition of the book. We particularly appreciate the many graduate students and professional practitioners who provided feedback,

often in the form of penetrating questions, that led to rewriting or expansion of material in the book. We are also indebted to John Wiley & Sons, the American Statistical Association, and the Biometrika Trustees for permission to use copyrighted material.

DOUGLAS C. MONTGOMERY
ELIZABETH A. PECK
G. GEOFFREY VINING

CONTENTS

PREFACE	xiii
1. INTRODUCTION	1
1.1 Regression and Model Building / 1	
1.2 Data Collection / 5	
1.3 Uses of Regression / 9	
1.4 Role of the Computer / 10	
2. SIMPLE LINEAR REGRESSION	12
2.1 Simple Linear Regression Model / 12	
2.2 Least-Squares Estimation of the Parameters / 13	
2.2.1 Estimation of β_0 and β_1 / 13	
2.2.2 Properties of the Least-Squares Estimators and the Fitted Regression Model / 18	
2.2.3 Estimation of σ^2 / 20	
2.2.4 Alternate Form of the Model / 22	
2.3 Hypothesis Testing on the Slope and Intercept / 22	
2.3.1 Use of t Tests / 22	
2.3.2 Testing Significance of Regression / 24	
2.3.3 Analysis of Variance / 25	
2.4 Interval Estimation in Simple Linear Regression / 29	
2.4.1 Confidence Intervals on β_0 , β_1 and σ^2 / 29	
2.4.2 Interval Estimation of the Mean Response / 30	
2.5 Prediction of New Observations / 33	
2.6 Coefficient of Determination / 35	

- 2.7 A Service Industry Application of Regression / 37
- 2.8 Using SAS® and R for Simple Linear Regression / 39
- 2.9 Some Considerations in the Use of Regression / 42
- 2.10 Regression Through the Origin / 45
- 2.11 Estimation by Maximum Likelihood / 51
- 2.12 Case Where the Regressor x is Random / 52
 - 2.12.1 x and y Jointly Distributed / 53
 - 2.12.2 x and y Jointly Normally Distributed:
Correlation Model / 53
- Problems / 58

3. MULTIPLE LINEAR REGRESSION

67

- 3.1 Multiple Regression Models / 67
- 3.2 Estimation of the Model Parameters / 70
 - 3.2.1 Least-Squares Estimation of the Regression
Coefficients / 71
 - 3.2.2 Geometrical Interpretation of Least Squares / 77
 - 3.2.3 Properties of the Least-Squares Estimators / 79
 - 3.2.4 Estimation of σ^2 / 80
 - 3.2.5 Inadequacy of Scatter Diagrams
in Multiple Regression / 82
 - 3.2.6 Maximum-Likelihood Estimation / 83
- 3.3 Hypothesis Testing in Multiple Linear Regression / 84
 - 3.3.1 Test for Significance of Regression / 84
 - 3.3.2 Tests on Individual Regression Coefficients
and Subsets of Coefficients / 88
 - 3.3.3 Special Case of Orthogonal Columns in \mathbf{X} / 93
 - 3.3.4 Testing the General Linear Hypothesis / 95
- 3.4 Confidence Intervals in Multiple Regression / 97
 - 3.4.1 Confidence Intervals on the Regression Coefficients / 98
 - 3.4.2 CI Estimation of the Mean Response / 99
 - 3.4.3 Simultaneous Confidence Intervals on Regression
Coefficients / 100
- 3.5 Prediction of New Observations / 104
- 3.6 A Multiple Regression Model for the Patient
Satisfaction Data / 104
- 3.7 Using SAS and R for Basic Multiple Linear Regression / 106
- 3.8 Hidden Extrapolation in Multiple Regression / 107
- 3.9 Standardized Regression Coefficients / 111
- 3.10 Multicollinearity / 117
- 3.11 Why Do Regression Coefficients Have the Wrong Sign? / 119
- Problems / 121

4. MODEL ADEQUACY CHECKING	129
4.1 Introduction / 129	
4.2 Residual Analysis / 130	
4.2.1 Definition of Residuals / 130	
4.2.2 Methods for Scaling Residuals / 130	
4.2.3 Residual Plots / 136	
4.2.4 Partial Regression and Partial Residual Plots / 143	
4.2.5 Using Minitab®, SAS, and R for Residual Analysis / 146	
4.2.6 Other Residual Plotting and Analysis Methods / 149	
4.3 PRESS Statistic / 151	
4.4 Detection and Treatment of Outliers / 152	
4.5 Lack of Fit of the Regression Model / 156	
4.5.1 Formal Test for Lack of Fit / 156	
4.5.2 Estimation of Pure Error from Near Neighbors / 160	
Problems / 165	
5. TRANSFORMATIONS AND WEIGHTING TO CORRECT MODEL INADEQUACIES	171
5.1 Introduction / 171	
5.2 Variance-Stabilizing Transformations / 172	
5.3 Transformations to Linearize the Model / 176	
5.4 Analytical Methods for Selecting a Transformation / 182	
5.4.1 Transformations on y : The Box–Cox Method / 182	
5.4.2 Transformations on the Regressor Variables / 184	
5.5 Generalized and Weighted Least Squares / 188	
5.5.1 Generalized Least Squares / 188	
5.5.2 Weighted Least Squares / 190	
5.5.3 Some Practical Issues / 191	
5.6 Regression Models with Random Effect / 194	
5.6.1 Subsampling / 194	
5.6.2 The General Situation for a Regression Model with a Single Random Effect / 198	
5.6.3 The Importance of the Mixed Model in Regression / 202	
Problems / 202	
6. DIAGNOSTICS FOR LEVERAGE AND INFLUENCE	211
6.1 Importance of Detecting Influential Observations / 211	
6.2 Leverage / 212	
6.3 Measures of Influence: Cook's D / 215	
6.4 Measures of Influence: $DFFITs$ and $DFBETAs$ / 217	
6.5 A Measure of Model Performance / 219	

- 6.6 Detecting Groups of Influential Observations / 220
- 6.7 Treatment of Influential Observations / 220
- Problems / 221

7. POLYNOMIAL REGRESSION MODELS 223

- 7.1 Introduction / 223
- 7.2 Polynomial Models in One Variable / 223
 - 7.2.1 Basic Principles / 223
 - 7.2.2 Piecewise Polynomial Fitting (Splines) / 229
 - 7.2.3 Polynomial and Trigonometric Terms / 235
- 7.3 Nonparametric Regression / 236
 - 7.3.1 Kernel Regression / 237
 - 7.3.2 Locally Weighted Regression (Loess) / 237
 - 7.3.3 Final Cautions / 241
- 7.4 Polynomial Models in Two or More Variables / 242
- 7.5 Orthogonal Polynomials / 248
- Problems / 254

8. INDICATOR VARIABLES 260

- 8.1 General Concept of Indicator Variables / 260
- 8.2 Comments on the Use of Indicator Variables / 273
 - 8.2.1 Indicator Variables versus Regression on Allocated Codes / 273
 - 8.2.2 Indicator Variables as a Substitute for a Quantitative Regressor / 274
- 8.3 Regression Approach to Analysis of Variance / 275
- Problems / 280

9. MULTICOLLINEARITY 285

- 9.1 Introduction / 285
- 9.2 Sources of Multicollinearity / 286
- 9.3 Effects of Multicollinearity / 288
- 9.4 Multicollinearity Diagnostics / 292
 - 9.4.1 Examination of the Correlation Matrix / 292
 - 9.4.2 Variance Inflation Factors / 296
 - 9.4.3 Eigensystem Analysis of $X'X$ / 297
 - 9.4.4 Other Diagnostics / 302
 - 9.4.5 SAS and R Code for Generating Multicollinearity Diagnostics / 303
- 9.5 Methods for Dealing with Multicollinearity / 303
 - 9.5.1 Collecting Additional Data / 303
 - 9.5.2 Model Respecification / 304
 - 9.5.3 Ridge Regression / 304

9.5.4	Principal-Component Regression / 313	
9.5.5	Comparison and Evaluation of Biased Estimators / 319	
9.6	Using SAS to Perform Ridge and Principal-Component Regression / 321	
	Problems / 323	
10.	VARIABLE SELECTION AND MODEL BUILDING	327
10.1	Introduction / 327	
10.1.1	Model-Building Problem / 327	
10.1.2	Consequences of Model Misspecification / 329	
10.1.3	Criteria for Evaluating Subset Regression Models / 332	
10.2	Computational Techniques for Variable Selection / 338	
10.2.1	All Possible Regressions / 338	
10.2.2	Stepwise Regression Methods / 344	
10.3	Strategy for Variable Selection and Model Building / 351	
10.4	Case Study: Gorman and Toman Asphalt Data Using SAS / 354	
	Problems / 367	
11.	VALIDATION OF REGRESSION MODELS	372
11.1	Introduction / 372	
11.2	Validation Techniques / 373	
11.2.1	Analysis of Model Coefficients and Predicted Values / 373	
11.2.2	Collecting Fresh Data—Confirmation Runs / 375	
11.2.3	Data Splitting / 377	
11.3	Data from Planned Experiments / 385	
	Problems / 386	
12.	INTRODUCTION TO NONLINEAR REGRESSION	389
12.1	Linear and Nonlinear Regression Models / 389	
12.1.1	Linear Regression Models / 389	
12.2.2	Nonlinear Regression Models / 390	
12.2	Origins of Nonlinear Models / 391	
12.3	Nonlinear Least Squares / 395	
12.4	Transformation to a Linear Model / 397	
12.5	Parameter Estimation in a Nonlinear System / 400	
12.5.1	Linearization / 400	
12.5.2	Other Parameter Estimation Methods / 407	
12.5.3	Starting Values / 408	
12.6	Statistical Inference in Nonlinear Regression / 409	
12.7	Examples of Nonlinear Regression Models / 411	
12.8	Using SAS and R / 412	
	Problems / 416	

13. GENERALIZED LINEAR MODELS	421
13.1 Introduction / 421	
13.2 Logistic Regression Models / 422	
13.2.1 Models with a Binary Response Variable / 422	
13.2.2 Estimating the Parameters in a Logistic Regression Model / 423	
13.2.3 Interpretation of the Parameters in a Logistic Regression Model / 428	
13.2.4 Statistical Inference on Model Parameters / 430	
13.2.5 Diagnostic Checking in Logistic Regression / 440	
13.2.6 Other Models for Binary Response Data / 442	
13.2.7 More Than Two Categorical Outcomes / 442	
13.3 Poisson Regression / 444	
13.4 The Generalized Linear Model / 450	
13.4.1 Link Functions and Linear Predictors / 451	
13.4.2 Parameter Estimation and Inference in the GLM / 452	
13.4.3 Prediction and Estimation with the GLM / 454	
13.4.4 Residual Analysis in the GLM / 456	
13.4.5 Using R to Perform GLM Analysis / 458	
13.4.6 Overdispersion / 461	
Problems / 462	
14. REGRESSION ANALYSIS OF TIME SERIES DATA	474
14.1 Introduction to Regression Models for Time Series Data / 474	
14.2 Detecting Autocorrelation: The Durbin-Watson Test / 475	
14.3 Estimating the Parameters in Time Series Regression Models / 480	
Problems / 496	
15. OTHER TOPICS IN THE USE OF REGRESSION ANALYSIS	500
15.1 Robust Regression / 500	
15.1.1 Need for Robust Regression / 500	
15.1.2 M -Estimators / 503	
15.1.3 Properties of Robust Estimators / 510	

15.2	Effect of Measurement Errors in the Regressors / 511	
15.2.1	Simple Linear Regression / 511	
15.2.2	The Berkson Model / 513	
15.3	Inverse Estimation—The Calibration Problem / 513	
15.4	Bootstrapping in Regression / 517	
15.4.1	Bootstrap Sampling in Regression / 518	
15.4.2	Bootstrap Confidence Intervals / 519	
15.5	Classification and Regression Trees (CART) / 524	
15.6	Neural Networks / 526	
15.7	Designed Experiments for Regression / 529	
	Problems / 537	
APPENDIX A.	STATISTICAL TABLES	541
APPENDIX B.	DATA SETS FOR EXERCISES	553
APPENDIX C.	SUPPLEMENTAL TECHNICAL MATERIAL	574
C.1	Background on Basic Test Statistics / 574	
C.2	Background from the Theory of Linear Models / 577	
C.3	Important Results on SS_R and SS_{Res} / 581	
C.4	Gauss-Markov Theorem, $\text{Var}(\epsilon) = \sigma^2 I$ / 587	
C.5	Computational Aspects of Multiple Regression / 589	
C.6	Result on the Inverse of a Matrix / 590	
C.7	Development of the PRESS Statistic / 591	
C.8	Development of $S_{(i)}^2$ / 593	
C.9	Outlier Test Based on R -Student / 594	
C.10	Independence of Residuals and Fitted Values / 596	
C.11	Gauss-Markov Theorem, $\text{Var}(\epsilon) = V$ / 597	
C.12	Bias in MS_{Res} When the Model Is Underspecified / 599	
C.13	Computation of Influence Diagnostics / 600	
C.14	Generalized Linear Models / 601	
APPENDIX D.	INTRODUCTION TO SAS	613
D.1	Basic Data Entry / 614	
D.2	Creating Permanent SAS Data Sets / 618	
D.3	Importing Data from an EXCEL File / 619	
D.4	Output Command / 620	
D.5	Log File / 620	
D.6	Adding Variables to an Existing SAS Data Set / 622	

APPENDIX E. INTRODUCTION TO R TO PERFORM LINEAR REGRESSION ANALYSIS	623
E.1 Basic Background on R / 623	
E.2 Basic Data Entry / 624	
E.3 Brief Comments on Other Functionality in R / 626	
E.4 R Commander / 627	
REFERENCES	628
INDEX	642