

# Applied Statistics

## Analysis of Variance and Regression

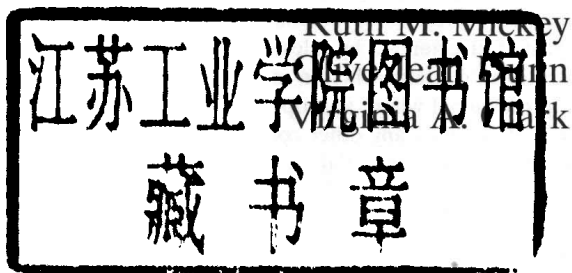
Third Edition

RUTH M. MICKEY  
OLIVE JEAN DUNN  
VIRGINIA A. CLARK

ftp://  
SITE AVAILABLE

*Applied Statistics*  
*Analysis of Variance and Regression*

Third Edition



 WILEY-  
INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

# *Applied Statistics*

Third Edition

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWART and SAMUEL S. WILKS

Editors: *David J. Balding, Noel A. C. Cressie, Nicholas I. Fisher,  
Iain M. Johnstone, J. B. Kadane, Geert Molenberghs, Louise M. Ryan,  
David W. Scott, Adrian F. M. Smith, Jozef L. Teugels*  
Editors Emeriti: *Vic Barnett, J. Stuart Hunter, David G. Kendall*

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

Copyright © 2004 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 978-646-8600, or on the web at [www.copyright.com](http://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.

**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 please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. 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, however, may not be available in electronic format.

***Library of Congress Cataloging-in-Publication Data***

Mickey, Ruth M., 1954–

Applied statistics : analysis of variance and regression.—3rd ed. / Ruth M. Mickey, Olive Jean Dunn, Virginia A. Clark.

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

Includes bibliographical references and index.

ISBN 0-471-37038-X (acid-free paper)

1. Analysis of variance. 2. Regression analysis. I. Dunn, Olive Jean. II. Clark, Virginia, 1928– III. Title. IV. Series

QA279.M45            2004

519.5'38—dc21

20033053461

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

# Preface

## What This Book Is About

The objectives of this third edition of *Applied Statistics: Analysis of Variance and Regression* are to be a textbook for a one year or, with omissions, a one-semester course in analysis of variance and regression and also to be a useful reference. In each chapter, we give an example, discuss how to summarize the data, state the model and the assumptions, give confidence intervals and tests, describe how to tell if the assumptions are satisfied, and offer advice on what to do if they are not satisfied. To illustrate the analysis of data that do not meet all the assumptions, in some of the examples the assumptions are not fully met. The statistical methods are given in a general context and are illustrated in terms of the chapter example. With this organization, the student is exposed to the whole data analysis process, starting from a data set to the presentation and interpretation of its statistical analyses. A low level of mathematics is assumed and clear statements of the model assumptions are made.

## What Changes Have Been Made in this 3rd Edition

There is now greater emphasis on data screening. This is introduced in the first chapter and continues to be addressed in subsequent chapters. Interpretation of computer program results are included so the reader will know how

to perform and interpret real life analyses. We have added sections on how to explain the statistical methods used and the results of the analyses.

We have included a wider variety of homework problems at the end of each chapter. Some problems make use of small, artificial data sets, which lend themselves to focusing on the statistical methods addressed in each chapter. These emphasize straightforward applications of the statistical methods or their verification by spreadsheet or by hand. Other problems make use of larger, real data sets; these are particularly useful for illustrating violations of the assumptions and other practical problems that can arise in real world applications. Finally, we have developed a limited number of problems that involve simulations. Large data sets are available on the Wiley ftp given below.

Some of the more technical material (such as general rules for deriving the formulae given in the chapters and some illustrative examples) is now in an appendix. Students can read through the material in a chapter and stay focused on the application and interpretation of the statistical methods; for interested students who need or want to verify formulae, the appendix is a good source of information. We reference relevant books and journal articles at the end of each chapter. We have tried to include both standard references as well as some very practical articles that have appeared in *The American Statistician*.

The first four chapters of the second edition were deleted as this introductory material would likely not now be covered in a course in analysis of variance and regression. However, some key topics in those deleted chapters are incorporated into other chapters or included in the appendix. We also dropped material that seems more appropriate for a course in experimental design: factorial designs with each factor at two levels and the Latin square design. Finally, we eliminated detailed computing formulas.

### Use of the Computer and the Wiley ftp Site

We recognize that there are numerous software packages that are available and have not written this book with any particular one in mind. We performed our data analyses using SAS and Minitab and occasionally S-PLUS; the graphs were generated using S-PLUS. We have tried to be aware of alternative names that the different packages use to describe the same thing and mention them in the text. The ftp site associated with this book is [ftp://ftp.wiley.com/public/sci\\_tech\\_med/applied\\_statistics](ftp://ftp.wiley.com/public/sci_tech_med/applied_statistics). We will use this site to store the larger data sets used in the text, selected computer programs that we created, additional homework problems, and errata.

## Acknowledgements

Many people contributed to this revision. Dr Mickey, who has used this material in her teaching, has taken a lead role in this new edition and she has been backstopped by Dr Clark. We both have benefited from the clarity of thinking and writing of Dr Dunn in earlier editions. We thank Dr Philip Ades, Dr Lorraine Berkett, Dr Richard Branda, Dr Elena Garcia, Neil Kamman, and Scott Pfister for generously giving us their data sets. We thank Welden Clark for his assistance with figures and tables and his help in putting the entire work together. We thank Steve Quigley, Executive Editor, and the editorial and production staff at Wiley—Heather Bergman, Rosalyn Farkas, and Susanne Steitz, for their patience and advice.

RUTH M. MICKEY (PROFESSOR, UNIV. OF VERMONT)

OLIVE JEAN DUNN (PROFESSOR EMERITA, UCLA)

VIRGINIA A. CLARK (PROFESSOR EMERITA, UCLA)



# Contents

<i>Preface</i>	xv
<b>1</b> <i>Data Screening</i>	<b>1</b>
1.1 <i>Variables and Their Classification</i>	2
1.2 <i>Describing the Data</i>	3
1.2.1 <i>Errors in the Data</i>	3
1.2.2 <i>Descriptive Statistics</i>	6
1.2.3 <i>Graphical Summarization</i>	9
1.3 <i>Departures from Assumptions</i>	14
1.3.1 <i>The Normal Distribution</i>	15
1.3.2 <i>The Normality Assumption</i>	15
1.3.3 <i>Transformations</i>	20
1.3.4 <i>Independence</i>	25
1.4 <i>Summary</i>	28
<i>Problems</i>	29
<i>References</i>	31
<b>2</b> <i>One-Way Analysis of Variance Design</i>	<b>33</b>
2.1 <i>One-Way Analysis of Variance with Fixed Effects</i>	34
2.1.1 <i>Example</i>	34

2.1.2	<i>The One-Way Analysis of Variance Model with Fixed Effects</i>	36
2.1.3	<i>Null Hypothesis: Test for Equality of Population Means</i>	39
2.1.4	<i>Estimation of Model Terms</i>	39
2.1.5	<i>Breakdown of the Basic Sum of Squares</i>	41
2.1.6	<i>Analysis of Variance Table</i>	43
2.1.7	<i>The F Test</i>	45
2.1.8	<i>Analysis of Variance with Unequal Sample Sizes</i>	49
2.2	<i>One-Way Analysis of Variance with Random Effects</i>	50
2.2.1	<i>Data Example</i>	50
2.2.2	<i>The One-Way Analysis of Variance Model with Random Effects</i>	51
2.2.3	<i>Null Hypothesis: Test for Zero Variance of Population Means</i>	52
2.2.4	<i>Estimation of Model Terms</i>	52
2.2.5	<i>The F Test</i>	53
2.3	<i>Designing an Observational Study or Experiment</i>	53
2.3.1	<i>Randomization for Experimental Studies</i>	54
2.3.2	<i>Sample Size and Power</i>	56
2.4	<i>Checking if the Data Fit the One-Way ANOVA Model</i>	57
2.4.1	<i>Normality</i>	58
2.4.2	<i>Equality of Population Variances</i>	59
2.4.3	<i>Independence</i>	60
2.4.4	<i>Robustness</i>	61
2.4.5	<i>Missing Data</i>	61
2.5	<i>What to Do if the Data Do Not Fit the Model</i>	62
2.5.1	<i>Making Transformations</i>	62
2.5.2	<i>Using Nonparametric Methods</i>	63
2.5.3	<i>Using Alternative ANOVAs</i>	64
2.6	<i>Presentation and Interpretation of Results</i>	64
2.7	<i>Summary</i>	65
	<i>Problems</i>	66
	<i>References</i>	69
3	<i>Estimation and Simultaneous Inference</i>	73
3.1	<i>Estimation for Single Population Means</i>	74
3.1.1	<i>Parameter Estimation</i>	74
3.1.2	<i>Confidence Intervals</i>	75

3.2	<i>Estimation for Linear Combinations of Population Means</i>	77
3.2.1	<i>Differences of Two Population Means</i>	78
3.2.2	<i>General Contrasts for Two or More Means</i>	80
3.2.3	<i>General Contrasts for Trends</i>	81
3.3	<i>Simultaneous Statistical Inference</i>	82
3.3.1	<i>Straightforward Approach to Inference</i>	83
3.3.2	<i>Motivation for Multiple Comparison Procedures and Terminology</i>	84
3.3.3	<i>The Bonferroni Multiple Comparison Method</i>	86
3.3.4	<i>The Tukey Multiple Comparison Method</i>	88
3.3.5	<i>The Scheffé Multiple Comparison Method</i>	89
3.4	<i>Inference for Variance Components</i>	91
3.5	<i>Presentation and Interpretation of Results</i>	91
3.6	<i>Summary</i>	92
	<i>Problems</i>	93
	<i>References</i>	93
4	<i>Hierarchical or Nested Design</i>	95
4.1	<i>Example</i>	96
4.2	<i>The Model</i>	98
4.3	<i>Analysis of Variance Table and F Tests</i>	100
4.3.1	<i>Analysis of Variance Table</i>	100
4.3.2	<i>F Tests</i>	101
4.3.3	<i>Pooling</i>	102
4.4	<i>Estimation of Parameters</i>	103
4.4.1	<i>Comparison with the One-Way ANOVA Model of Chapter 2</i>	106
4.5	<i>Inferences with Unequal Sample Sizes</i>	107
4.5.1	<i>Hypothesis Testing</i>	107
4.5.2	<i>Estimation</i>	110
4.6	<i>Checking If the Data Fit the Model</i>	110
4.7	<i>What to Do If the Data Don't Fit the Model</i>	111
4.8	<i>Designing a Study</i>	112
4.8.1	<i>Relative Efficiency</i>	112
4.9	<i>Summary</i>	113
	<i>Problems</i>	113
	<i>References</i>	115

5	<i>Two Crossed Factors: Fixed Effects and Equal Sample Sizes</i>	117
5.1	<i>Example</i>	118
5.2	<i>The Model</i>	119
5.3	<i>Interpretation of Models and Interaction</i>	121
5.4	<i>Analysis of Variance and F Tests</i>	126
5.5	<i>Estimates of Parameters and Confidence Intervals</i>	129
5.6	<i>Designing a Study</i>	134
5.7	<i>Presentation and Interpretation of Results</i>	137
5.8	<i>Summary</i>	138
	<i>Problems</i>	139
	<i>References</i>	142
6	<i>Randomized Complete Block Design</i>	143
6.1	<i>Example</i>	144
6.2	<i>The Randomized Complete Block Design</i>	145
6.3	<i>The Model</i>	147
6.4	<i>Analysis of Variance Table and F Tests</i>	149
6.5	<i>Estimation of Parameters and Confidence Intervals</i>	151
6.6	<i>Checking If the Data Fit the Model</i>	154
6.7	<i>What to Do if the Data Don't Fit the Model</i>	155
6.7.1	<i>Friedman's Rank Sum Test</i>	155
6.7.2	<i>Missing Data</i>	156
6.8	<i>Designing a Randomized Complete Block Study</i>	157
6.8.1	<i>Experimental Studies</i>	157
6.8.2	<i>Observational Studies</i>	158
6.9	<i>Model Extensions</i>	159
6.10	<i>Summary</i>	159
	<i>Problems</i>	160
	<i>References</i>	162
7	<i>Two Crossed Factors: Fixed Effects and Unequal Sample Sizes</i>	163
7.1	<i>Example</i>	164
7.2	<i>The Model</i>	165
7.3	<i>Analysis of Variance and F Tests</i>	166

7.4	<i>Estimation of Parameters and Confidence Intervals</i>	168
7.4.1	<i>Means and Adjusted Means</i>	168
7.4.2	<i>Standard Errors and Confidence Intervals</i>	172
7.5	<i>Checking If the Data Fit the Two-Way Model</i>	174
7.6	<i>What To Do If the Data Don't Fit the Model</i>	178
7.7	<i>Summary</i>	180
	<i>Problems</i>	180
	<i>References</i>	181
8	<i>Crossed Factors: Mixed Models</i>	183
8.1	<i>Example</i>	184
8.2	<i>The Mixed Model</i>	184
8.3	<i>Estimation of Fixed Effects</i>	188
8.4	<i>Analysis of Variance</i>	188
8.5	<i>Estimation of Variance Components</i>	189
8.6	<i>Hypothesis Testing</i>	192
8.7	<i>Confidence Intervals for Means and Variance Components</i>	193
8.7.1	<i>Confidence Intervals for Population Means</i>	193
8.7.2	<i>Confidence Intervals for Variance Components</i>	196
8.8	<i>Comments on Available Software</i>	197
8.9	<i>Extensions of the Mixed Model</i>	197
8.9.1	<i>Unequal Sample Sizes</i>	198
8.9.2	<i>Fixed, Random, or Mixed Effects</i>	198
8.9.3	<i>Crossed versus Nested Factors</i>	198
8.9.4	<i>Dependence of Random Effects</i>	199
8.10	<i>Summary</i>	199
	<i>Problems</i>	200
	<i>References</i>	201
9	<i>Repeated Measures Designs</i>	203
9.1	<i>Repeated Measures for a Single Population</i>	204
9.1.1	<i>Example</i>	204
9.1.2	<i>The Model</i>	206
9.1.3	<i>Hypothesis Testing: No Time Effect</i>	208

9.1.4	<i>Simultaneous Inference</i>	209
9.1.5	<i>Orthogonal Contrasts</i>	211
9.1.6	<i>F Tests for Trends over Time</i>	213
9.2	<i>Repeated Measures with Several Populations</i>	214
9.2.1	<i>Example</i>	215
9.2.2	<i>Model</i>	216
9.2.3	<i>Analysis of Variance Table and F Tests</i>	218
9.3	<i>Checking if the Data Fit the Repeated Measures Model</i>	220
9.4	<i>What to Do if the Data Don't Fit the Model</i>	222
9.5	<i>General Comments on Repeated Measures Analyses</i>	222
9.6	<i>Summary</i>	223
	<i>Problems</i>	224
	<i>References</i>	227
10	<i>Linear Regression: Fixed X Model</i>	229
10.1	<i>Example</i>	230
10.2	<i>Fitting a Straight Line</i>	233
10.3	<i>The Fixed X Model</i>	235
10.4	<i>Estimation of Model Parameters and Standard Errors</i>	237
10.4.1	<i>Point Estimates</i>	237
10.4.2	<i>Estimates of Standard Errors</i>	238
10.5	<i>Inferences for Model Parameters: Confidence Intervals</i>	240
10.6	<i>Inference for Model Parameters: Hypothesis Testing</i>	242
10.6.1	<i>t Tests for Intercept and Slope</i>	242
10.6.2	<i>Division of the Basic Sum of Squares</i>	244
10.6.3	<i>Analysis of Variance Table and F Test</i>	245
10.7	<i>Checking if the Data Fit the Regression Model</i>	245
10.7.1	<i>Outliers</i>	247
10.7.2	<i>Checking for Linearity</i>	247
10.7.3	<i>Checking for Equality of Variances</i>	248
10.7.4	<i>Checking for Normality</i>	249
10.7.5	<i>Summary of Screening Procedures</i>	252
10.8	<i>What to Do if the Data Don't Fit the Model</i>	253
10.9	<i>Practical Issues in Designing a Regression Study</i>	255
10.9.1	<i>Is Fixed X Regression an Appropriate Technique?</i>	255
10.9.2	<i>What Values of X Should Be Selected?</i>	257
10.9.3	<i>Sample Size Calculations</i>	257

10.10	<i>Comparison with One-Way ANOVA</i>	257
10.11	<i>Summary</i>	258
	<i>Problems</i>	259
	<i>References</i>	263
11	<i>Linear Regression: Random X Model and Correlation</i>	265
11.1	<i>Example</i>	266
	11.1.1 <i>Sampling and Summary Statistics</i>	266
11.2	<i>Summarizing the Relationship Between X and Y</i>	268
11.3	<i>Inferences for the Regression of Y on X</i>	272
	11.3.1 <i>Comparison of Fixed X and Random X Sampling</i>	274
11.4	<i>The Bivariate Normal Model</i>	276
	11.4.1 <i>The Bivariate Normal Distribution</i>	276
	11.4.2 <i>The Correlation Coefficient</i>	278
	11.4.3 <i>The Correlation Coefficient: Confidence Intervals and Tests</i>	280
11.5	<i>Checking if the Data Fit the Random X Regression Model</i>	285
	11.5.1 <i>Checking for High-Leverage, Outlying, and Influential Observations</i>	286
11.6	<i>What to Do if the Data Don't Fit the Random X Model</i>	291
	11.6.1 <i>Nonparametric Alternatives to Simple Linear Regression</i>	291
	11.6.2 <i>Nonparametric Alternatives to the Pearson Correlation</i>	291
11.7	<i>Summary</i>	293
	<i>Problems</i>	294
	<i>References</i>	298
12	<i>Multiple Regression</i>	301
12.1	<i>Example</i>	302
12.2	<i>The Sample Regression Plane</i>	305
12.3	<i>The Multiple Regression Model</i>	308
12.4	<i>Parameters, Standard Errors, and Confidence Intervals</i>	310
	12.4.1 <i>Prediction of <math>E(Y X_1, \dots, X_k)</math></i>	311
	12.4.2 <i>Standardized Regression Coefficients</i>	311
12.5	<i>Hypothesis Testing</i>	313
	12.5.1 <i>Test That All Partial Regression Coefficients Are 0</i>	313
	12.5.2 <i>Tests that One Partial Regression Coefficient is 0</i>	314

12.6	<i>Checking If the Data Fit the Multiple Regression Model</i>	319
12.6.1	<i>Checking for Outlying, High Leverage and Influential Points</i>	320
12.6.2	<i>Checking for Linearity</i>	321
12.6.3	<i>Checking for Equality of Variances</i>	321
12.6.4	<i>Checking for Normality of Errors</i>	322
12.6.5	<i>Other Potential Problems</i>	322
12.7	<i>What to Do If the Data Don't Fit the Model</i>	324
12.8	<i>Summary</i>	325
	<i>Problems</i>	325
	<i>References</i>	329
13	<i>Multiple and Partial Correlation</i>	331
13.1	<i>Example</i>	332
13.2	<i>The Sample Multiple Correlation Coefficient</i>	333
13.3	<i>The Sample Partial Correlation Coefficient</i>	335
13.4	<i>The Joint Distribution Model</i>	338
13.4.1	<i>The Population Multiple Correlation Coefficient</i>	340
13.4.2	<i>The Population Partial Correlation Coefficient</i>	340
13.5	<i>Inferences for the Multiple Correlation Coefficient</i>	341
13.6	<i>Inferences for Partial Correlation Coefficients</i>	342
13.6.1	<i>Confidence Intervals for Partial Correlation Coefficients</i>	343
13.6.2	<i>Hypothesis Tests for Partial Correlation Coefficients</i>	344
13.7	<i>Checking If the Data Fit the Joint Normal Model</i>	346
13.8	<i>What to Do If the Data Don't Fit the Model</i>	346
13.9	<i>Summary</i>	347
	<i>Problems</i>	347
	<i>References</i>	349
14	<i>Miscellaneous Topics in Regression</i>	351
14.1	<i>Models with Dummy Variables</i>	352
14.2	<i>Models with Interaction Terms</i>	355
14.3	<i>Models with Polynomial Terms</i>	356
14.3.1	<i>Polynomial Model</i>	358
14.4	<i>Variable Selection</i>	358
14.4.1	<i>Criteria for Evaluating and Comparing Models</i>	359



14.4.2	<i>Methods for Variable Selection</i>	362
14.4.3	<i>General Comments on Variable Selection</i>	365
14.5	<i>Summary</i>	369
	<i>Problems</i>	370
	<i>References</i>	372
15	<i>Analysis of Covariance</i>	375
15.1	<i>Example</i>	376
15.2	<i>The ANCOVA Model</i>	378
15.3	<i>Estimation of Model Parameters</i>	380
15.4	<i>Hypothesis Tests</i>	381
15.5	<i>Adjusted Means</i>	385
15.5.1	<i>Estimation of Adjusted Means and Standard Errors</i>	385
15.5.2	<i>Confidence Intervals for Adjusted Means</i>	387
15.6	<i>Checking If the Data Fit the ANCOVA Model</i>	388
15.7	<i>What to Do if the Data Don't Fit the Model</i>	392
15.8	<i>ANCOVA in Observational Studies</i>	392
15.9	<i>What Makes a Good Covariate</i>	394
15.10	<i>Measurement Error</i>	395
15.11	<i>ANCOVA versus Other Methods of Adjustment</i>	396
15.12	<i>Comments on Statistical Software</i>	397
15.13	<i>Summary</i>	398
	<i>Problems</i>	398
	<i>References</i>	400
16	<i>Summaries, Extensions, and Communication</i>	403
16.1	<i>Summaries and Extensions of Models</i>	403
16.2	<i>Communication of Statistics in the Context of a Research Project</i>	404
	<i>References</i>	408
Appendix A		409
A.1	<i>Expected Values and Parameters</i>	409
A.2	<i>Linear Combinations of Variables and Their Parameters</i>	410
A.3	<i>Balanced One-Way ANOVA, Expected Mean Squares</i>	414
A.3.1	<i>To Show <math>EMS(MS_a) = \sigma^2 + n \sum_{i=1}^a \alpha_i^2 / (a - 1)</math></i>	414
A.3.2	<i>To Show <math>EMS(MS_r) = \sigma^2</math></i>	415