

THIRD EDITION

Applied Regression Analysis and Other Multivariable Methods

David G. Kleinbaum

Emory University

Lawrence L. Kupper

University of North Carolina, Chapel Hill

Keith E. Muller

University of North Carolina, Chapel Hill

Azhar Nizam

Emory University

An Alexander Kugushev Book



Duxbury Press

An Imprint of Brooks/Cole Publishing Company

ITP® An International Thomson Publishing Company

Pacific Grove • Albany • Belmont • Bonn • Boston • Cincinnati • Detroit • Johannesburg • London
Madrid • Melbourne • Mexico City • New York • Paris • Singapore • Tokyo • Toronto • Washington

Contents

1 CONCEPTS AND EXAMPLES OF RESEARCH 1

- 1-1 Concepts 1
- 1-2 Examples 2
- 1-3 Concluding Remarks 5
- References 6

2 CLASSIFICATION OF VARIABLES AND THE CHOICE OF ANALYSIS 7

- 2-1 Classification of Variables 7
- 2-2 Overlapping of Classification Schemes 11
- 2-3 Choice of Analysis 11
- References 13

3 BASIC STATISTICS: A REVIEW 14

- 3-1 Preview 14
- 3-2 Descriptive Statistics 15
- 3-3 Random Variables and Distributions 16
- 3-4 Sampling Distributions of t , χ^2 , and F 19
- 3-5 Statistical Inference: Estimation 21
- 3-6 Statistical Inference: Hypothesis Testing 24

3-7 Error Rates, Power, and Sample Size	28
Problems	30
References	33

4 INTRODUCTION TO REGRESSION ANALYSIS 34

4-1 Preview	34
4-2 Association versus Causality	35
4-3 Statistical versus Deterministic Models	37
4-4 Concluding Remarks	38
References	38

5 STRAIGHT-LINE REGRESSION ANALYSIS 39

5-1 Preview	39
5-2 Regression with a Single Independent Variable	39
5-3 Mathematical Properties of a Straight Line	42
5-4 Statistical Assumptions for a Straight-line Model	43
5-5 Determining the Best-fitting Straight Line	47
5-6 Measure of the Quality of the Straight-line Fit and Estimate of σ^2	51
5-7 Inferences About the Slope and Intercept	52
5-8 Interpretations of Tests for Slope and Intercept	54
5-9 Inferences About the Regression Line $\mu_{Y X} = \beta_0 + \beta_1 X$	57
5-10 Prediction of a New Value of Y at X_0	59
5-11 Assessing the Appropriateness of the Straight-line Model	60
Problems	60
References	87

6 THE CORRELATION COEFFICIENT AND STRAIGHT-LINE REGRESSION ANALYSIS 88

6-1 Definition of r	88
6-2 r as a Measure of Association	89
6-3 The Bivariate Normal Distribution	90
6-4 r and the Strength of the Straight-line Relationship	93
6-5 What r Does Not Measure	95
6-6 Tests of Hypotheses and Confidence Intervals for the Correlation Coefficient	96
6-7 Testing for the Equality of Two Correlations	99
Problems	101
References	103

7 THE ANALYSIS-OF-VARIANCE TABLE 104

- 7-1 Preview 104
- 7-2 The ANOVA Table for Straight-line Regression Problems 104
108

8 MULTIPLE REGRESSION ANALYSIS:
GENERAL CONSIDERATIONS 111

- 8-1 Preview 111
- 8-2 Multiple Regression Models 112
- 8-3 Graphical Look at the Problem 113
- 8-4 Assumptions of Multiple Regression 115
- 8-5 Determining the Best Estimate of the Multiple Regression Equation 118
- 8-6 The ANOVA Table for Multiple Regression 119
- 8-7 Numerical Examples 121
Problems 123
References 135

9 TESTING HYPOTHESES IN MULTIPLE REGRESSION 136

- 9-1 Preview 136
- 9-2 Test for Significant Overall Regression 137
- 9-3 Partial F Test 138
- 9-4 Multiple Partial F Test 143
- 9-5 Strategies for Using Partial F Tests 145
- 9-6 Tests Involving the Intercept 150
Problems 151
References 159

10 CORRELATIONS: MULTIPLE,
PARTIAL, AND MULTIPLE PARTIAL 160

- 10-1 Preview 160
- 10-2 Correlation Matrix 161
- 10-3 Multiple Correlation Coefficient 162
- 10-4 Relationship of $R_{Y|X_1, X_2, \dots, X_k}$ to the Multivariate Normal Distribution 164
- 10-5 Partial Correlation Coefficient 165
- 10-6 Alternative Representation of the Regression Model 172

10-7	Multiple Partial Correlation	172
10-8	Concluding Remarks	174
	Problems	174
	Reference	185

11 CONFOUNDING AND INTERACTION IN REGRESSION 186

11-1	Preview	186
11-2	Overview	186
11-3	Interaction in Regression	188
11-4	Confounding in Regression	194
11-5	Summary and Conclusions	199
	Problems	199
	Reference	211

12 REGRESSION DIAGNOSTICS 212

12-1	Preview	212
12-2	Simple Approaches to Diagnosing Problems in Data	212
12-3	Residual Analysis	216
12-4	Treating Outliers	228
12-5	Collinearity	237
12-6	Scaling Problems	248
12-7	Treating Collinearity and Scaling Problems	248
12-8	Alternate Strategies of Analysis	249
12-9	An Important Caution	252
	Problems	253
	References	279

13 POLYNOMIAL REGRESSION 281

13-1	Preview	281
13-2	Polynomial Models	282
13-3	Least-squares Procedure for Fitting a Parabola	282
13-4	ANOVA Table for Second-order Polynomial Regression	284
13-5	Inferences Associated with Second-order Polynomial Regression	284
13-6	Example Requiring a Second-order Model	286
13-7	Fitting and Testing Higher-order Models	290
13-8	Lack-of-fit Tests	290
13-9	Orthogonal Polynomials	292

13-10	Strategies for Choosing a Polynomial Model	301
	Problems	302

14 DUMMY VARIABLES IN REGRESSION 317

14-1	Preview	317
14-2	Definitions	317
14-3	Rule for Defining Dummy Variables	318
14-4	Comparing Two Straight-line Regression Equations: An Example	319
14-5	Questions for Comparing Two Straight Lines	320
14-6	Methods of Comparing Two Straight Lines	321
14-7	Method I: Using Separate Regression Fits to Compare Two Straight Lines	322
14-8	Method II: Using a Single Regression Equation to Compare Two Straight Lines	327
14-9	Comparison of Methods I and II	330
14-10	Testing Strategies and Interpretation: Comparing Two Straight Lines	330
14-11	Other Dummy Variable Models	332
14-12	Comparing Four Regression Equations	334
14-13	Comparing Several Regression Equations Involving Two Nominal Variables	336
	Problems	338
	References	360

15 ANALYSIS OF COVARIANCE AND OTHER METHODS FOR ADJUSTING CONTINUOUS DATA 361

15-1	Preview	361
15-2	Adjustment Problem	361
15-3	Analysis of Covariance	363
15-4	Assumption of Parallelism: A Potential Drawback	365
15-5	Analysis of Covariance: Several Groups and Several Covariates	366
15-6	Comments and Cautions	368
15-7	Summary	371
	Problems	371
	Reference	385

16 SELECTING THE BEST REGRESSION EQUATION 386

16-1	Preview	386
16-2	Steps in Selecting the Best Regression Equation	387
16-3	Step 1: Specifying the Maximum Model	387
16-4	Step 2: Specifying a Criterion for Selecting a Model	390
16-5	Step 3: Specifying a Strategy for Selecting Variables	392

16-6	Step 4: Conducting the Analysis	401
16-7	Step 5: Evaluating Reliability with Split Samples	401
16-8	Example Analysis of Actual Data	403
16-9	Issues in Selecting the Most Valid Model	409
	Problems	409
	References	422

17 ONE-WAY ANALYSIS OF VARIANCE 423

17-1	Preview	423
17-2	One-way ANOVA: The Problem, Assumptions, and Data Configuration	426
17-3	Methodology for One-way Fixed-effects ANOVA	429
17-4	Regression Model for Fixed-effects One-way ANOVA	435
17-5	Fixed-effects Model for One-way ANOVA	438
17-6	Random-effects Model for One-way ANOVA	440
17-7	Multiple-comparison Procedures for Fixed-effects One-way ANOVA	443
17-8	Choosing a Multiple-comparison Technique	456
17-9	Orthogonal Contrasts and Partitioning an ANOVA Sum of Squares	457
	Problems	463
	References	483

18 RANDOMIZED BLOCKS: SPECIAL CASE OF TWO-WAY ANOVA 484

18-1	Preview	484
18-2	Equivalent Analysis of a Matched Pairs Experiment	488
18-3	Principle of Blocking	491
18-4	Analysis of a Randomized-blocks Experiment	493
18-5	ANOVA Table for a Randomized-blocks Experiment	495
18-6	Regression Models for a Randomized-blocks Experiment	499
18-7	Fixed-effects ANOVA Model for a Randomized-blocks Experiment	502
	Problems	503
	References	515

19 TWO-WAY ANOVA WITH EQUAL CELL NUMBERS 516

19-1	Preview	516
19-2	Using a Table of Cell Means	518
19-3	General Methodology	522
19-4	<i>F</i> Tests for Two-way ANOVA	527
19-5	Regression Model for Fixed-effects Two-way ANOVA	530
19-6	Interactions in Two-way ANOVA	534
19-7	Random- and Mixed-effects Two-way ANOVA Models	541
	Problems	544
	References	560

20 TWO-WAY ANOVA WITH UNEQUAL CELL NUMBERS 561

- 20-1 Preview 561
- 20-2 Problems with Unequal Cell Numbers: Nonorthogonality 563
- 20-3 Regression Approach for Unequal Cell Sample Sizes 567
- 20-4 Higher-way ANOVA 571
 - Problems 572
 - References 588

21 ANALYSIS OF REPEATED MEASURES DATA 589

- 21-1 Preview 589
- 21-2 Examples 590
- 21-3 General Approach for Repeated Measures ANOVA 592
- 21-4 Overview of Selected Repeated Measures Designs and ANOVA-based Analyses 594
- 21-5 Repeated Measures ANOVA for Unbalanced Data 611
- 21-6 Other Approaches to Analyzing Repeated Measures Data 612
 - Appendix 21-A Examples of SAS's GLM and MIXED Procedures 613
 - Problems 616
 - References 638

22 THE METHOD OF MAXIMUM LIKELIHOOD 639

- 22-1 Preview 639
- 22-2 The Principle of Maximum Likelihood 639
- 22-3 Statistical Inference via Maximum Likelihood 642
- 22-4 Summary 652
 - Problems 653
 - References 655

23 LOGISTIC REGRESSION ANALYSIS 656

- 23-1 Preview 656
- 23-2 The Logistic Model 656
- 23-3 Estimating the Odds Ratio Using Logistic Regression 658
- 23-4 A Numerical Example of Logistic Regression 664
- 23-5 Theoretical Considerations 671
- 23-6 An Example of Conditional ML Estimation
 - Involving Pair-matched Data with Unmatched Covariates 677
- 23-7 Summary 681
 - Problems 682
 - References 686

24 POISSON REGRESSION ANALYSIS 687

- 24-1 Preview 687
- 24-2 The Poisson Distribution 687
- 24-3 An Example of Poisson Regression 688
- 24-4 Poisson Regression: General Considerations 690
- 24-5 Measures of Goodness of Fit 694
- 24-6 Continuation of Skin Cancer Data Example 696
- 24-7 A Second Illustration of Poisson Regression Analysis 701
- 24-8 Summary 704
- Problems 705
- References 709

A APPENDIX A—TABLES 711

- A-1 Standard Normal Cumulative Probabilities 712
- A-2 Percentiles of the t Distribution 715
- A-3 Percentiles of the Chi-square Distribution 716
- A-4 Percentiles of the F Distribution 717
- A-5 Values of $\frac{1}{2} \ln \frac{1+r}{1-r}$ 724
- A-6 Upper α Point of Studentized Range 726
- A-7 Orthogonal Polynomial Coefficients 728
- A-8 Bonferroni Corrected Jackknife and Studentized Residual Critical Values 729
- A-9 Critical Values for Leverages 730
- A-10 Critical Values for the Maximum of N Values of Cook's $d(i)$ times $(n - k - 1)$ 731

B APPENDIX B—MATRICES AND THEIR
RELATIONSHIP TO REGRESSION ANALYSIS 732**C** APPENDIX C—ANOVA INFORMATION FOR FOUR
COMMON BALANCED REPEATED MEASURES DESIGNS 744

- C-1 Balanced Repeated Measures Design with One Crossover Factor (Treatments) 744
- C-2 Balanced Repeated Measures Design with Two Crossover Factors 746
- C-3 Balanced Repeated Measures Design with One Nest Factor (Treatments) 750
- C-4 Balanced Repeated Measures Design with One Crossover Factor and One Nest Factor 752
- C-5 Balanced Two-group Pre/Posttest Design 755
- References 757

D SOLUTIONS TO EXERCISES 758

INDEX 787