

SEVENTH EDITION

Using Multivariate Statistics

Barbara G. Tabachnick

Linda S. Fidell



SEVENTH EDITION

Using Multivariate Statistics

Barbara G. Tabachnick
California State University, Northridge

Linda S. Fidell
California State University, Northridge



330 Hudson Street, NY NY 10013

Portfolio Manager: *Tanimaa Mehra*
Content Producer: *Kani Kapoor*
Portfolio Manager Assistant: *Anna Austin*
Product Marketer: *Jessica Quazza*
Art/Designer: *Integra Software Services Pvt. Ltd.*
Full-Service Project Manager: *Integra Software Services Pvt. Ltd.*

Composer: *Integra Software Services Pvt. Ltd.*
Printer/Binder: *LSC Communications, Inc.*
Cover Printer: *Phoenix Color/Hagerstown*
Cover Design: *Lumina Datamatics, Inc.*
Cover Art: *Shutterstock*

Acknowledgments of third party content appear on pages within the text, which constitutes an extension of this copyright page.

Copyright © 2019, 2013, 2007 by Pearson Education, Inc. or its affiliates. All Rights Reserved. Printed in the United States of America. This publication is protected by copyright, and permission should be obtained from the publisher prior to any prohibited reproduction, storage in a retrieval system, or transmission in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise. For information regarding permissions, request forms and the appropriate contacts within the Pearson Education Global Rights & Permissions department, please visit www.pearsoned.com/permissions/.

PEARSON and ALWAYS LEARNING are exclusive trademarks owned by Pearson Education, Inc. or its affiliates, in the U.S., and/or other countries.

Unless otherwise indicated herein, any third-party trademarks that may appear in this work are the property of their respective owners and any references to third-party trademarks, logos or other trade dress are for demonstrative or descriptive purposes only. Such references are not intended to imply any sponsorship, endorsement, authorization, or promotion of Pearson's products by the owners of such marks, or any relationship between the owner and Pearson Education, Inc. or its affiliates, authors, licensees or distributors.

Many of the designations by manufacturers and seller to distinguish their products are claimed as trademarks. Where those designations appear in this book, and the publisher was aware of a trademark claim, the designations have been printed in initial caps or all caps.

Library of Congress Cataloging-in-Publication Data

Names: Tabachnick, Barbara G., author. | Fidell, Linda S., author.
Title: Using multivariate statistics/Barbara G. Tabachnick, California State University, Northridge, Linda S. Fidell, California State University, Northridge.
Description: Seventh edition. | Boston: Pearson, [2019] | Chapter 14, by Jodie B. Ullman.
Identifiers: LCCN 2017040173 | ISBN 9780134790541 | ISBN 0134790545
Subjects: LCSH: Multivariate analysis. | Statistics.
Classification: LCC QA278 .T3 2019 | DDC 519.5/35—dc23
LC record available at <https://lcn.loc.gov/2017040173>

Contents

Preface	xiv		
1 Introduction	1		
1.1 Multivariate Statistics: Why?	1		
1.1.1 The Domain of Multivariate Statistics: Numbers of IVs and DVs	2		
1.1.2 Experimental and Nonexperimental Research	2		
1.1.3 Computers and Multivariate Statistics	3		
1.1.4 Garbage In, Roses Out?	4		
1.2 Some Useful Definitions	5		
1.2.1 Continuous, Discrete, and Dichotomous Data	5		
1.2.2 Samples and Populations	6		
1.2.3 Descriptive and Inferential Statistics	7		
1.2.4 Orthogonality: Standard and Sequential Analyses	7		
1.3 Linear Combinations of Variables	9		
1.4 Number and Nature of Variables to Include	10		
1.5 Statistical Power	10		
1.6 Data Appropriate for Multivariate Statistics	11		
1.6.1 The Data Matrix	11		
1.6.2 The Correlation Matrix	12		
1.6.3 The Variance–Covariance Matrix	12		
1.6.4 The Sum-of-Squares and Cross-Products Matrix	13		
1.6.5 Residuals	14		
1.7 Organization of the Book	14		
2 A Guide to Statistical Techniques: Using the Book	15		
2.1 Research Questions and Associated Techniques	15		
2.1.1 Degree of Relationship Among Variables	15		
2.1.1.1 Bivariate r	16		
2.1.1.2 Multiple R	16		
2.1.1.3 Sequential R	16		
2.1.1.4 Canonical R	16		
2.1.1.5 Multiway Frequency Analysis	17		
2.1.1.6 Multilevel Modeling	17		
2.1.2 Significance of Group Differences	17		
2.1.2.1 One-Way ANOVA and t Test	17		
2.1.2.2 One-Way ANCOVA	17		
2.1.2.3 Factorial ANOVA	18		
2.1.2.4 Factorial ANCOVA	18		
2.1.2.5 Hotelling's T^2	18		
2.1.2.6 One-Way MANOVA	18		
2.1.2.7 One-Way MANCOVA	19		
2.1.2.8 Factorial MANOVA	19		
2.1.2.9 Factorial MANCOVA	19		
2.1.2.10 Profile Analysis of Repeated Measures	19		
2.1.3 Prediction of Group Membership	20		
2.1.3.1 One-Way Discriminant Analysis	20		
2.1.3.2 Sequential One-Way Discriminant Analysis	20		
2.1.3.3 Multiway Frequency Analysis (Logit)	21		
2.1.3.4 Logistic Regression	21		
2.1.3.5 Sequential Logistic Regression	21		
2.1.3.6 Factorial Discriminant Analysis	21		
2.1.3.7 Sequential Factorial Discriminant Analysis	22		
2.1.4 Structure	22		
2.1.4.1 Principal Components	22		
2.1.4.2 Factor Analysis	22		
2.1.4.3 Structural Equation Modeling	22		
2.1.5 Time Course of Events	22		
2.1.5.1 Survival/Failure Analysis	23		
2.1.5.2 Time-Series Analysis	23		
2.2 Some Further Comparisons	23		
2.3 A Decision Tree	24		
2.4 Technique Chapters	27		
2.5 Preliminary Check of the Data	28		
3 Review of Univariate and Bivariate Statistics	29		
3.1 Hypothesis Testing	29		
3.1.1 One-Sample z Test as Prototype	30		
3.1.2 Power	32		
3.1.3 Extensions of the Model	32		
3.1.4 Controversy Surrounding Significance Testing	33		
3.2 Analysis of Variance	33		
3.2.1 One-Way Between-Subjects ANOVA	34		
3.2.2 Factorial Between-Subjects ANOVA	36		
3.2.3 Within-Subjects ANOVA	38		
3.2.4 Mixed Between-Within-Subjects ANOVA	40		
3.2.5 Design Complexity	41		
3.2.5.1 Nesting	41		
3.2.5.2 Latin-Square Designs	42		
3.2.5.3 Unequal n and Nonorthogonality	42		
3.2.5.4 Fixed and Random Effects	43		
3.2.6 Specific Comparisons	43		
3.2.6.1 Weighting Coefficients for Comparisons	43		
3.2.6.2 Orthogonality of Weighting Coefficients	44		
3.2.6.3 Obtained F for Comparisons	44		
3.2.6.4 Critical F for Planned Comparisons	45		
3.2.6.5 Critical F for Post Hoc Comparisons	45		
3.3 Parameter Estimation	46		
3.4 Effect Size	47		

3.5 Bivariate Statistics: Correlation and Regression	48	5 Multiple Regression	99
3.5.1 Correlation	48	5.1 General Purpose and Description	99
3.5.2 Regression	49	5.2 Kinds of Research Questions	101
3.6 Chi-Square Analysis	50	5.2.1 Degree of Relationship	101
4 Cleaning Up Your Act: Screening Data Prior to Analysis	52	5.2.2 Importance of IVs	102
4.1 Important Issues in Data Screening	53	5.2.3 Adding IVs	102
4.1.1 Accuracy of Data File	53	5.2.4 Changing IVs	102
4.1.2 Honest Correlations	53	5.2.5 Contingencies Among IVs	102
4.1.2.1 Inflated Correlation	53	5.2.6 Comparing Sets of IVs	102
4.1.2.2 Deflated Correlation	53	5.2.7 Predicting DV Scores for Members of a New Sample	103
4.1.3 Missing Data	54	5.2.8 Parameter Estimates	103
4.1.3.1 Deleting Cases or Variables	57	5.3 Limitations to Regression Analyses	103
4.1.3.2 Estimating Missing Data	57	5.3.1 Theoretical Issues	103
4.1.3.3 Using a Missing Data Correlation Matrix	61	5.3.2 Practical Issues	104
4.1.3.4 Treating Missing Data as Data	61	5.3.2.1 Ratio of Cases to IVs	105
4.1.3.5 Repeating Analyses with and without Missing Data	61	5.3.2.2 Absence of Outliers Among the IVs and on the DV	105
4.1.3.6 Choosing Among Methods for Dealing with Missing Data	62	5.3.2.3 Absence of Multicollinearity and Singularity	106
4.1.4 Outliers	62	5.3.2.4 Normality, Linearity, and Homoscedasticity of Residuals	106
4.1.4.1 Detecting Univariate and Multivariate Outliers	63	5.3.2.5 Independence of Errors	108
4.1.4.2 Describing Outliers	66	5.3.2.6 Absence of Outliers in the Solution	109
4.1.4.3 Reducing the Influence of Outliers	66	5.4 Fundamental Equations for Multiple Regression	109
4.1.4.4 Outliers in a Solution	67	5.4.1 General Linear Equations	110
4.1.5 Normality, Linearity, and Homoscedasticity	67	5.4.2 Matrix Equations	111
4.1.5.1 Normality	68	5.4.3 Computer Analyses of Small-Sample Example	113
4.1.5.2 Linearity	72	5.5 Major Types of Multiple Regression	115
4.1.5.3 Homoscedasticity, Homogeneity of Variance, and Homogeneity of Variance–Covariance Matrices	73	5.5.1 Standard Multiple Regression	115
4.1.6 Common Data Transformations	75	5.5.2 Sequential Multiple Regression	116
4.1.7 Multicollinearity and Singularity	76	5.5.3 Statistical (Stepwise) Regression	117
4.1.8 A Checklist and Some Practical Recommendations	79	5.5.4 Choosing Among Regression Strategies	121
4.2 Complete Examples of Data Screening	79	5.6 Some Important Issues	121
4.2.1 Screening Ungrouped Data	80	5.6.1 Importance of IVs	121
4.2.1.1 Accuracy of Input, Missing Data, Distributions, and Univariate Outliers	81	5.6.1.1 Standard Multiple Regression	122
4.2.1.2 Linearity and Homoscedasticity	84	5.6.1.2 Sequential or Statistical Regression	123
4.2.1.3 Transformation	84	5.6.1.3 Commonality Analysis	123
4.2.1.4 Detecting Multivariate Outliers	84	5.6.1.4 Relative Importance Analysis	125
4.2.1.5 Variables Causing Cases to Be Outliers	86	5.6.2 Statistical Inference	128
4.2.1.6 Multicollinearity	88	5.6.2.1 Test for Multiple R	128
4.2.2 Screening Grouped Data	88	5.6.2.2 Test of Regression Components	129
4.2.2.1 Accuracy of Input, Missing Data, Distributions, Homogeneity of Variance, and Univariate Outliers	89	5.6.2.3 Test of Added Subset of IVs	130
4.2.2.2 Linearity	93	5.6.2.4 Confidence Limits	130
4.2.2.3 Multivariate Outliers	93	5.6.2.5 Comparing Two Sets of Predictors	131
4.2.2.4 Variables Causing Cases to Be Outliers	94	5.6.3 Adjustment of R^2	132
4.2.2.5 Multicollinearity	97	5.6.4 Suppressor Variables	133
		5.6.5 Regression Approach to ANOVA	134
		5.6.6 Centering When Interactions and Powers of IVs Are Included	135
		5.6.7 Mediation in Causal Sequence	137

5.7 Complete Examples of Regression Analysis	138		
5.7.1 Evaluation of Assumptions	139		
5.7.1.1 Ratio of Cases to IVs	139		
5.7.1.2 Normality, Linearity, Homoscedasticity, and Independence of Residuals	139		
5.7.1.3 Outliers	142		
5.7.1.4 Multicollinearity and Singularity	144		
5.7.2 Standard Multiple Regression	144		
5.7.3 Sequential Regression	150		
5.7.4 Example of Standard Multiple Regression with Missing Values Multiply Imputed	154		
5.8 Comparison of Programs	162		
5.8.1 IBM SPSS Package	163		
5.8.2 SAS System	165		
5.8.3 SYSTAT System	166		
6 Analysis of Covariance	167		
6.1 General Purpose and Description	167		
6.2 Kinds of Research Questions	170		
6.2.1 Main Effects of IVs	170		
6.2.2 Interactions Among IVs	170		
6.2.3 Specific Comparisons and Trend Analysis	170		
6.2.4 Effects of Covariates	170		
6.2.5 Effect Size	171		
6.2.6 Parameter Estimates	171		
6.3 Limitations to Analysis of Covariance	171		
6.3.1 Theoretical Issues	171		
6.3.2 Practical Issues	172		
6.3.2.1 Unequal Sample Sizes, Missing Data, and Ratio of Cases to IVs	172		
6.3.2.2 Absence of Outliers	172		
6.3.2.3 Absence of Multicollinearity and Singularity	172		
6.3.2.4 Normality of Sampling Distributions	173		
6.3.2.5 Homogeneity of Variance	173		
6.3.2.6 Linearity	173		
6.3.2.7 Homogeneity of Regression	173		
6.3.2.8 Reliability of Covariates	174		
6.4 Fundamental Equations for Analysis of Covariance	174		
6.4.1 Sums of Squares and Cross-Products	175		
6.4.2 Significance Test and Effect Size	177		
6.4.3 Computer Analyses of Small-Sample Example	178		
6.5 Some Important Issues	179		
6.5.1 Choosing Covariates	179		
6.5.2 Evaluation of Covariates	180		
6.5.3 Test for Homogeneity of Regression	180		
6.5.4 Design Complexity	181		
6.5.4.1 Within-Subjects and Mixed Within-Between Designs	181		
6.5.4.2 Unequal Sample Sizes	182		
6.5.4.3 Specific Comparisons and Trend Analysis	185		
6.5.4.4 Effect Size	187		
6.5.5 Alternatives to ANCOVA	187		
6.6 Complete Example of Analysis of Covariance	189		
6.6.1 Evaluation of Assumptions	189		
6.6.1.1 Unequal <i>n</i> and Missing Data	189		
6.6.1.2 Normality	189		
6.6.1.3 Linearity	191		
6.6.1.4 Outliers	191		
6.6.1.5 Multicollinearity and Singularity	192		
6.6.1.6 Homogeneity of Variance	192		
6.6.1.7 Homogeneity of Regression	193		
6.6.1.8 Reliability of Covariates	193		
6.6.2 Analysis of Covariance	193		
6.6.2.1 Main Analysis	193		
6.6.2.2 Evaluation of Covariates	196		
6.6.2.3 Homogeneity of Regression Run	196		
6.7 Comparison of Programs	200		
6.7.1 IBM SPSS Package	200		
6.7.2 SAS System	200		
6.7.3 SYSTAT System	200		
7 Multivariate Analysis of Variance and Covariance	203		
7.1 General Purpose and Description	203		
7.2 Kinds of Research Questions	206		
7.2.1 Main Effects of IVs	206		
7.2.2 Interactions Among IVs	207		
7.2.3 Importance of DVs	207		
7.2.4 Parameter Estimates	207		
7.2.5 Specific Comparisons and Trend Analysis	207		
7.2.6 Effect Size	208		
7.2.7 Effects of Covariates	208		
7.2.8 Repeated-Measures Analysis of Variance	208		
7.3 Limitations to Multivariate Analysis of Variance and Covariance	208		
7.3.1 Theoretical Issues	208		
7.3.2 Practical Issues	209		
7.3.2.1 Unequal Sample Sizes, Missing Data, and Power	209		
7.3.2.2 Multivariate Normality	210		
7.3.2.3 Absence of Outliers	210		
7.3.2.4 Homogeneity of Variance–Covariance Matrices	210		
7.3.2.5 Linearity	211		
7.3.2.6 Homogeneity of Regression	211		
7.3.2.7 Reliability of Covariates	211		
7.3.2.8 Absence of Multicollinearity and Singularity	211		
7.4 Fundamental Equations for Multivariate Analysis of Variance and Covariance	212		
7.4.1 Multivariate Analysis of Variance	212		

7.4.2	Computer Analyses of Small-Sample Example	218
7.4.3	Multivariate Analysis of Covariance	221
7.5	Some Important Issues	223
7.5.1	MANOVA Versus ANOVAs	223
7.5.2	Criteria for Statistical Inference	223
7.5.3	Assessing DVs	224
7.5.3.1	Univariate F	224
7.5.3.2	Roy–Bargmann Stepdown Analysis	226
7.5.3.3	Using Discriminant Analysis	226
7.5.3.4	Choosing Among Strategies for Assessing DVs	227
7.5.4	Specific Comparisons and Trend Analysis	227
7.5.5	Design Complexity	228
7.5.5.1	Within-Subjects and Between- Within Designs	228
7.5.5.2	Unequal Sample Sizes	228
7.6	Complete Examples of Multivariate Analysis of Variance and Covariance	230
7.6.1	Evaluation of Assumptions	230
7.6.1.1	Unequal Sample Sizes and Missing Data	230
7.6.1.2	Multivariate Normality	231
7.6.1.3	Linearity	231
7.6.1.4	Outliers	232
7.6.1.5	Homogeneity of Variance– Covariance Matrices	233
7.6.1.6	Homogeneity of Regression	233
7.6.1.7	Reliability of Covariates	235
7.6.1.8	Multicollinearity and Singularity	235
7.6.2	Multivariate Analysis of Variance	235
7.6.3	Multivariate Analysis of Covariance	244
7.6.3.1	Assessing Covariates	244
7.6.3.2	Assessing DVs	245
7.7	Comparison of Programs	252
7.7.1	IBM SPSS Package	252
7.7.2	SAS System	255
7.7.3	SYSTAT System	255

8 Profile Analysis: The Multivariate Approach to Repeated Measures 256

8.1	General Purpose and Description	256
8.2	Kinds of Research Questions	257
8.2.1	Parallelism of Profiles	258
8.2.2	Overall Difference Among Groups	258
8.2.3	Flatness of Profiles	258
8.2.4	Contrasts Following Profile Analysis	258
8.2.5	Parameter Estimates	258
8.2.6	Effect Size	259
8.3	Limitations to Profile Analysis	259
8.3.1	Theoretical Issues	259
8.3.2	Practical Issues	259
8.3.2.1	Sample Size, Missing Data, and Power	259

8.3.2.2	Multivariate Normality	260
8.3.2.3	Absence of Outliers	260
8.3.2.4	Homogeneity of Variance–Covariance Matrices	260
8.3.2.5	Linearity	260
8.3.2.6	Absence of Multicollinearity and Singularity	260
8.4	Fundamental Equations for Profile Analysis	260
8.4.1	Differences in Levels	262
8.4.2	Parallelism	262
8.4.3	Flatness	265
8.4.4	Computer Analyses of Small-Sample Example	266
8.5	Some Important Issues	269
8.5.1	Univariate Versus Multivariate Approach to Repeated Measures	269
8.5.2	Contrasts in Profile Analysis	270
8.5.2.1	Parallelism and Flatness Significant, Levels Not Significant (Simple-Effects Analysis)	272
8.5.2.2	Parallelism and Levels Significant, Flatness Not Significant (Simple-Effects Analysis)	274
8.5.2.3	Parallelism, Levels, and Flatness Significant (Interaction Contrasts)	275
8.5.2.4	Only Parallelism Significant	276
8.5.3	Doubly Multivariate Designs	277
8.5.4	Classifying Profiles	279
8.5.5	Imputation of Missing Values	279
8.6	Complete Examples of Profile Analysis	280
8.6.1	Profile Analysis of Subscales of the WISC	280
8.6.1.1	Evaluation of Assumptions	280
8.6.1.2	Profile Analysis	282
8.6.2	Doubly Multivariate Analysis of Reaction Time	288
8.6.2.1	Evaluation of Assumptions	288
8.6.2.2	Doubly Multivariate Analysis of Slope and Intercept	290
8.7	Comparison of Programs	297
8.7.1	IBM SPSS Package	297
8.7.2	SAS System	298
8.7.3	SYSTAT System	298

9 Discriminant Analysis 299

9.1	General Purpose and Description	299
9.2	Kinds of Research Questions	302
9.2.1	Significance of Prediction	302
9.2.2	Number of Significant Discriminant Functions	302
9.2.3	Dimensions of Discrimination	302
9.2.4	Classification Functions	303
9.2.5	Adequacy of Classification	303
9.2.6	Effect Size	303
9.2.7	Importance of Predictor Variables	303
9.2.8	Significance of Prediction with Covariates	304
9.2.9	Estimation of Group Means	304

9.3	Limitations to Discriminant Analysis	304	10	Logistic Regression	346
9.3.1	Theoretical Issues	304	10.1	General Purpose and Description	346
9.3.2	Practical Issues	304	10.2	Kinds of Research Questions	348
9.3.2.1	Unequal Sample Sizes, Missing Data, and Power	304	10.2.1	Prediction of Group Membership or Outcome	348
9.3.2.2	Multivariate Normality	305	10.2.2	Importance of Predictors	348
9.3.2.3	Absence of Outliers	305	10.2.3	Interactions Among Predictors	349
9.3.2.4	Homogeneity of Variance–Covariance Matrices	305	10.2.4	Parameter Estimates	349
9.3.2.5	Linearity	306	10.2.5	Classification of Cases	349
9.3.2.6	Absence of Multicollinearity and Singularity	306	10.2.6	Significance of Prediction with Covariates	349
10.2.7	Effect Size	349	10.3	Limitations to Logistic Regression Analysis	350
9.4	Fundamental Equations for Discriminant Analysis	306	10.3.1	Theoretical Issues	350
9.4.1	Derivation and Test of Discriminant Functions	307	10.3.2	Practical Issues	350
9.4.2	Classification	309	10.3.2.1	Ratio of Cases to Variables	350
9.4.3	Computer Analyses of Small-Sample Example	311	10.3.2.2	Adequacy of Expected Frequencies and Power	351
9.5	Types of Discriminant Analyses	315	10.3.2.3	Linearity in the Logit	351
9.5.1	Direct Discriminant Analysis	315	10.3.2.4	Absence of Multicollinearity	351
9.5.2	Sequential Discriminant Analysis	315	10.3.2.5	Absence of Outliers in the Solution	351
9.5.3	Stepwise (Statistical) Discriminant Analysis	316	10.3.2.6	Independence of Errors	352
9.6	Some Important Issues	316	10.4	Fundamental Equations for Logistic Regression	352
9.6.1	Statistical Inference	316	10.4.1	Testing and Interpreting Coefficients	353
9.6.1.1	Criteria for Overall Statistical Significance	317	10.4.2	Goodness of Fit	354
9.6.1.2	Stepping Methods	317	10.4.3	Comparing Models	355
9.6.2	Number of Discriminant Functions	317	10.4.4	Interpretation and Analysis of Residuals	355
9.6.3	Interpreting Discriminant Functions	318	10.4.5	Computer Analyses of Small-Sample Example	356
9.6.3.1	Discriminant Function Plots	318	10.5	Types of Logistic Regression	360
9.6.3.2	Structure Matrix of Loadings	318	10.5.1	Direct Logistic Regression	360
9.6.4	Evaluating Predictor Variables	320	10.5.2	Sequential Logistic Regression	360
9.6.5	Effect Size	321	10.5.3	Statistical (Stepwise) Logistic Regression	362
9.6.6	Design Complexity: Factorial Designs	321	10.5.4	Probit and Other Analyses	362
9.6.7	Use of Classification Procedures	322	10.6	Some Important Issues	363
9.6.7.1	Cross-Validation and New Cases	322	10.6.1	Statistical Inference	363
9.6.7.2	Jackknifed Classification	323	10.6.1.1	Assessing Goodness of Fit of Models	363
9.6.7.3	Evaluating Improvement in Classification	323	10.6.1.2	Tests of Individual Predictors	365
9.7	Complete Example of Discriminant Analysis	324	10.6.2	Effect Sizes	365
9.7.1	Evaluation of Assumptions	325	10.6.2.1	Effect Size for a Model	365
9.7.1.1	Unequal Sample Sizes and Missing Data	325	10.6.2.2	Effect Sizes for Predictors	366
9.7.1.2	Multivariate Normality	325	10.6.3	Interpretation of Coefficients Using Odds	367
9.7.1.3	Linearity	325	10.6.4	Coding Outcome and Predictor Categories	368
9.7.1.4	Outliers	325	10.6.5	Number and Type of Outcome Categories	369
9.7.1.5	Homogeneity of Variance–Covariance Matrices	326	10.6.6	Classification of Cases	372
9.7.1.6	Multicollinearity and Singularity	327	10.6.7	Hierarchical and Nonhierarchical Analysis	372
9.7.2	Direct Discriminant Analysis	327			
9.8	Comparison of Programs	340			
9.8.1	IBM SPSS Package	344			
9.8.2	SAS System	344			
9.8.3	SYSTAT System	345			

10.6.8	Importance of Predictors	373
10.6.9	Logistic Regression for Matched Groups	374
10.7	Complete Examples of Logistic Regression	374
10.7.1	Evaluation of Limitations	374
10.7.1.1	Ratio of Cases to Variables and Missing Data	374
10.7.1.2	Multicollinearity	376
10.7.1.3	Outliers in the Solution	376
10.7.2	Direct Logistic Regression with Two-Category Outcome and Continuous Predictors	377
10.7.2.1	Limitation: Linearity in the Logit	377
10.7.2.2	Direct Logistic Regression with Two-Category Outcome	377
10.7.3	Sequential Logistic Regression with Three Categories of Outcome	384
10.7.3.1	Limitations of Multinomial Logistic Regression	384
10.7.3.2	Sequential Multinomial Logistic Regression	387
10.8	Comparison of Programs	396
10.8.1	IBM SPSS Package	396
10.8.2	SAS System	399
10.8.3	SYSTAT System	400

11 Survival/Failure Analysis 401

11.1	General Purpose and Description	401
11.2	Kinds of Research Questions	403
11.2.1	Proportions Surviving at Various Times	403
11.2.2	Group Differences in Survival	403
11.2.3	Survival Time with Covariates	403
11.2.3.1	Treatment Effects	403
11.2.3.2	Importance of Covariates	403
11.2.3.3	Parameter Estimates	404
11.2.3.4	Contingencies Among Covariates	404
11.2.3.5	Effect Size and Power	404
11.3	Limitations to Survival Analysis	404
11.3.1	Theoretical Issues	404
11.3.2	Practical Issues	404
11.3.2.1	Sample Size and Missing Data	404
11.3.2.2	Normality of Sampling Distributions, Linearity, and Homoscedasticity	405
11.3.2.3	Absence of Outliers	405
11.3.2.4	Differences Between Withdrawn and Remaining Cases	405
11.3.2.5	Change in Survival Conditions over Time	405
11.3.2.6	Proportionality of Hazards	405
11.3.2.7	Absence of Multicollinearity	405
11.4	Fundamental Equations for Survival Analysis	405
11.4.1	Life Tables	406

11.4.2	Standard Error of Cumulative Proportion Surviving	408
11.4.3	Hazard and Density Functions	408
11.4.4	Plot of Life Tables	409
11.4.5	Test for Group Differences	410
11.4.6	Computer Analyses of Small-Sample Example	411
11.5	Types of Survival Analyses	415
11.5.1	Actuarial and Product-Limit Life Tables and Survivor Functions	415
11.5.2	Prediction of Group Survival Times from Covariates	417
11.5.2.1	Direct, Sequential, and Statistical Analysis	417
11.5.2.2	Cox Proportional-Hazards Model	417
11.5.2.3	Accelerated Failure-Time Models	419
11.5.2.4	Choosing a Method	423
11.6	Some Important Issues	423
11.6.1	Proportionality of Hazards	423
11.6.2	Censored Data	424
11.6.2.1	Right-Censored Data	425
11.6.2.2	Other Forms of Censoring	425
11.6.3	Effect Size and Power	425
11.6.4	Statistical Criteria	426
11.6.4.1	Test Statistics for Group Differences in Survival Functions	426
11.6.4.2	Test Statistics for Prediction from Covariates	427
11.6.5	Predicting Survival Rate	427
11.6.5.1	Regression Coefficients (Parameter Estimates)	427
11.6.5.2	Hazard Ratios	427
11.6.5.3	Expected Survival Rates	428
11.7	Complete Example of Survival Analysis	429
11.7.1	Evaluation of Assumptions	430
11.7.1.1	Accuracy of Input, Adequacy of Sample Size, Missing Data, and Distributions	430
11.7.1.2	Outliers	430
11.7.1.3	Differences Between Withdrawn and Remaining Cases	433
11.7.1.4	Change in Survival Experience over Time	433
11.7.1.5	Proportionality of Hazards	433
11.7.1.6	Multicollinearity	434
11.7.2	Cox Regression Survival Analysis	436
11.7.2.1	Effect of Drug Treatment	436
11.7.2.2	Evaluation of Other Covariates	436
11.8	Comparison of Programs	440
11.8.1	SAS System	444
11.8.2	IBM SPSS Package	445
11.8.3	SYSTAT System	445

12 Canonical Correlation 446

12.1	General Purpose and Description	446
12.2	Kinds of Research Questions	448

12.2.1	Number of Canonical Variate Pairs	448	13.3.2.5	Absence of Multicollinearity and Singularity	482
12.2.2	Interpretation of Canonical Variates	448	13.3.2.6	Factorability of R	482
12.2.3	Importance of Canonical Variates and Predictors	448	13.3.2.7	Absence of Outliers Among Variables	483
12.2.4	Canonical Variate Scores	449	13.4	Fundamental Equations for Factor Analysis	483
12.3	Limitations	449	13.4.1	Extraction	485
12.3.1	Theoretical Limitations	449	13.4.2	Orthogonal Rotation	487
12.3.2	Practical Issues	450	13.4.3	Communalities, Variance, and Covariance	488
12.3.2.1	Ratio of Cases to IVs	450	13.4.4	Factor Scores	489
12.3.2.2	Normality, Linearity, and Homoscedasticity	450	13.4.5	Oblique Rotation	491
12.3.2.3	Missing Data	451	13.4.6	Computer Analyses of Small-Sample Example	493
12.3.2.4	Absence of Outliers	451	13.5	Major Types of Factor Analyses	496
12.3.2.5	Absence of Multicollinearity and Singularity	451	13.5.1	Factor Extraction Techniques	496
12.4	Fundamental Equations for Canonical Correlation	451	13.5.1.1	PCA Versus FA	496
12.4.1	Eigenvalues and Eigenvectors	452	13.5.1.2	Principal Components	498
12.4.2	Matrix Equations	454	13.5.1.3	Principal Factors	498
12.4.3	Proportions of Variance Extracted	457	13.5.1.4	Image Factor Extraction	498
12.4.4	Computer Analyses of Small-Sample Example	458	13.5.1.5	Maximum Likelihood Factor Extraction	499
12.5	Some Important Issues	462	13.5.1.6	Unweighted Least Squares Factoring	499
12.5.1	Importance of Canonical Variates	462	13.5.1.7	Generalized (Weighted) Least Squares Factoring	499
12.5.2	Interpretation of Canonical Variates	463	13.5.1.8	Alpha Factoring	499
12.6	Complete Example of Canonical Correlation	463	13.5.2	Rotation	500
12.6.1	Evaluation of Assumptions	463	13.5.2.1	Orthogonal Rotation	500
12.6.1.1	Missing Data	463	13.5.2.2	Oblique Rotation	501
12.6.1.2	Normality, Linearity, and Homoscedasticity	463	13.5.2.3	Geometric Interpretation	502
12.6.1.3	Outliers	466	13.5.3	Some Practical Recommendations	503
12.6.1.4	Multicollinearity and Singularity	467	13.6	Some Important Issues	504
12.6.2	Canonical Correlation	467	13.6.1	Estimates of Communalities	504
12.7	Comparison of Programs	473	13.6.2	Adequacy of Extraction and Number of Factors	504
12.7.1	SAS System	473	13.6.3	Adequacy of Rotation and Simple Structure	507
12.7.2	IBM SPSS Package	474	13.6.4	Importance and Internal Consistency of Factors	508
12.7.3	SYSTAT System	475	13.6.5	Interpretation of Factors	509
13	Principal Components and Factor Analysis	476	13.6.6	Factor Scores	510
13.1	General Purpose and Description	476	13.6.7	Comparisons Among Solutions and Groups	511
13.2	Kinds of Research Questions	479	13.7	Complete Example of FA	511
13.2.1	Number of Factors	479	13.7.1	Evaluation of Limitations	511
13.2.2	Nature of Factors	479	13.7.1.1	Sample Size and Missing Data	512
13.2.3	Importance of Solutions and Factors	480	13.7.1.2	Normality	512
13.2.4	Testing Theory in FA	480	13.7.1.3	Linearity	512
13.2.5	Estimating Scores on Factors	480	13.7.1.4	Outliers	513
13.3	Limitations	480	13.7.1.5	Multicollinearity and Singularity	514
13.3.1	Theoretical Issues	480	13.7.1.6	Factorability of R	514
13.3.2	Practical Issues	481	13.7.1.7	Outliers Among Variables	515
13.3.2.1	Sample Size and Missing Data	481	13.7.2	Principal Factors Extraction with Varimax Rotation	515
13.3.2.2	Normality	482			
13.3.2.3	Linearity	482			
13.3.2.4	Absence of Outliers Among Cases	482			

13.8 Comparison of Programs	525
13.8.1 IBM SPSS Package	527
13.8.2 SAS System	527
13.8.3 SYSTAT System	527

14 Structural Equation Modeling by Jodie B. Ullman 528

14.1 General Purpose and Description	528
14.2 Kinds of Research Questions	531
14.2.1 Adequacy of the Model	531
14.2.2 Testing Theory	532
14.2.3 Amount of Variance in the Variables Accounted for by the Factors	532
14.2.4 Reliability of the Indicators	532
14.2.5 Parameter Estimates	532
14.2.6 Intervening Variables	532
14.2.7 Group Differences	532
14.2.8 Longitudinal Differences	533
14.2.9 Multilevel Modeling	533
14.2.10 Latent Class Analysis	533
14.3 Limitations to Structural Equation Modeling	533
14.3.1 Theoretical Issues	533
14.3.2 Practical Issues	534
14.3.2.1 Sample Size and Missing Data	534
14.3.2.2 Multivariate Normality and Outliers	534
14.3.2.3 Linearity	535
14.3.2.4 Absence of Multicollinearity and Singularity	535
14.3.2.5 Residuals	535
14.4 Fundamental Equations for Structural Equations Modeling	535
14.4.1 Covariance Algebra	535
14.4.2 Model Hypotheses	537
14.4.3 Model Specification	538
14.4.4 Model Estimation	540
14.4.5 Model Evaluation	543
14.4.6 Computer Analysis of Small-Sample Example	545
14.5 Some Important Issues	555
14.5.1 Model Identification	555
14.5.2 Estimation Techniques	557
14.5.2.1 Estimation Methods and Sample Size	559
14.5.2.2 Estimation Methods and Nonnormality	559
14.5.2.3 Estimation Methods and Dependence	559
14.5.2.4 Some Recommendations for Choice of Estimation Method	560
14.5.3 Assessing the Fit of the Model	560
14.5.3.1 Comparative Fit Indices	560
14.5.3.2 Absolute Fit Index	562

14.5.3.3 Indices of Proportion of Variance Accounted	562
14.5.3.4 Degree of Parsimony Fit Indices	563
14.5.3.5 Residual-Based Fit Indices	563
14.5.3.6 Choosing Among Fit Indices	564
14.5.4 Model Modification	564
14.5.4.1 Chi-Square Difference Test	564
14.5.4.2 Lagrange Multiplier (LM) Test	565
14.5.4.3 Wald Test	569
14.5.4.4 Some Caveats and Hints on Model Modification	570
14.5.5 Reliability and Proportion of Variance	570
14.5.6 Discrete and Ordinal Data	571
14.5.7 Multiple Group Models	572
14.5.8 Mean and Covariance Structure Models	573
14.6 Complete Examples of Structural Equation Modeling Analysis	574
14.6.1 Confirmatory Factor Analysis of the WISC	574
14.6.1.1 Model Specification for CFA	574
14.6.1.2 Evaluation of Assumptions for CFA	574
14.6.1.3 CFA Model Estimation and Preliminary Evaluation	576
14.6.1.4 Model Modification	583
14.6.2 SEM of Health Data	589
14.6.2.1 SEM Model Specification	589
14.6.2.2 Evaluation of Assumptions for SEM	591
14.6.2.3 SEM Model Estimation and Preliminary Evaluation	593
14.6.2.4 Model Modification	596
14.7 Comparison of Programs	607
14.7.1 EQS	607
14.7.2 LISREL	607
14.7.3 AMOS	612
14.7.4 SAS System	612

15 Multilevel Linear Modeling 613

15.1 General Purpose and Description	613
15.2 Kinds of Research Questions	616
15.2.1 Group Differences in Means	616
15.2.2 Group Differences in Slopes	616
15.2.3 Cross-Level Interactions	616
15.2.4 Meta-Analysis	616
15.2.5 Relative Strength of Predictors at Various Levels	617
15.2.6 Individual and Group Structure	617
15.2.7 Effect Size	617
15.2.8 Path Analysis at Individual and Group Levels	617
15.2.9 Analysis of Longitudinal Data	617
15.2.10 Multilevel Logistic Regression	618
15.2.11 Multiple Response Analysis	618

15.3	Limitations to Multilevel Linear Modeling	618	15.7.1.1	Sample Sizes, Missing Data, and Distributions	656
15.3.1	Theoretical Issues	618	15.7.1.2	Outliers	659
15.3.2	Practical Issues	618	15.7.1.3	Multicollinearity and Singularity	659
15.3.2.1	Sample Size, Unequal- <i>n</i> , and Missing Data	619	15.7.1.4	Independence of Errors: Intraclass Correlations	659
15.3.2.2	Independence of Errors	619	15.7.2	Multilevel Modeling	661
15.3.2.3	Absence of Multicollinearity and Singularity	620	15.8	Comparison of Programs	668
15.4	Fundamental Equations	620	15.8.1	SAS System	668
15.4.1	Intercepts-Only Model	623	15.8.2	IBM SPSS Package	670
15.4.1.1	The Intercepts-Only Model: Level-1 Equation	623	15.8.3	HLM Program	671
15.4.1.2	The Intercepts-Only Model: Level-2 Equation	623	15.8.4	MLwiN Program	671
15.4.1.3	Computer Analyses of Intercepts-Only Model	624	15.8.5	SYSTAT System	671
15.4.2	Model with a First-Level Predictor	627	16	Multway Frequency Analysis	672
15.4.2.1	Level-1 Equation for a Model with a Level-1 Predictor	627	16.1	General Purpose and Description	672
15.4.2.2	Level-2 Equations for a Model with a Level-1 Predictor	628	16.2	Kinds of Research Questions	673
15.4.2.3	Computer Analysis of a Model with a Level-1 Predictor	630	16.2.1	Associations Among Variables	673
15.4.3	Model with Predictors at First and Second Levels	633	16.2.2	Effect on a Dependent Variable	674
15.4.3.1	Level-1 Equation for Model with Predictors at Both Levels	633	16.2.3	Parameter Estimates	674
15.4.3.2	Level-2 Equations for Model with Predictors at Both Levels	633	16.2.4	Importance of Effects	674
15.4.3.3	Computer Analyses of Model with Predictors at First and Second Levels	634	16.2.5	Effect Size	674
15.5	Types of MLM	638	16.2.6	Specific Comparisons and Trend Analysis	674
15.5.1	Repeated Measures	638	16.3	Limitations to Multway Frequency Analysis	675
15.5.2	Higher-Order MLM	642	16.3.1	Theoretical Issues	675
15.5.3	Latent Variables	642	16.3.2	Practical Issues	675
15.5.4	Nonnormal Outcome Variables	643	16.3.2.1	Independence	675
15.5.5	Multiple Response Models	644	16.3.2.2	Ratio of Cases to Variables	675
15.6	Some Important Issues	644	16.3.2.3	Adequacy of Expected Frequencies	675
15.6.1	Intraclass Correlation	644	16.3.2.4	Absence of Outliers in the Solution	676
15.6.2	Centering Predictors and Changes in Their Interpretations	646	16.4	Fundamental Equations for Multway Frequency Analysis	676
15.6.3	Interactions	648	16.4.1	Screening for Effects	678
15.6.4	Random and Fixed Intercepts and Slopes	648	16.4.1.1	Total Effect	678
15.6.5	Statistical Inference	651	16.4.1.2	First-Order Effects	679
15.6.5.1	Assessing Models	651	16.4.1.3	Second-Order Effects	679
15.6.5.2	Tests of Individual Effects	652	16.4.1.4	Third-Order Effect	683
15.6.6	Effect Size	653	16.4.2	Modeling	683
15.6.7	Estimation Techniques and Convergence Problems	653	16.4.3	Evaluation and Interpretation	685
15.6.8	Exploratory Model Building	654	16.4.3.1	Residuals	685
15.7	Complete Example of MLM	655	16.4.3.2	Parameter Estimates	686
15.7.1	Evaluation of Assumptions	656	16.4.4	Computer Analyses of Small-Sample Example	690
			16.5	Some Important Issues	695
			16.5.1	Hierarchical and Nonhierarchical Models	695
			16.5.2	Statistical Criteria	696
			16.5.2.1	Tests of Models	696
			16.5.2.2	Tests of Individual Effects	696
			16.5.3	Strategies for Choosing a Model	696
			16.5.3.1	IBM SPSS HILOGLINEAR (Hierarchical)	697

16.5.3.2 IBM SPSS GENLOG (General Log-Linear)	697
16.5.3.3 SAS CATMOD and IBM SPSS LOGLINEAR (General Log-Linear)	697
16.6 Complete Example of Multiway Frequency Analysis	698
16.6.1 Evaluation of Assumptions: Adequacy of Expected Frequencies	698
16.6.2 Hierarchical Log-Linear Analysis	700
16.6.2.1 Preliminary Model Screening	700
16.6.2.2 Stepwise Model Selection	702
16.6.2.3 Adequacy of Fit	702
16.6.2.4 Interpretation of the Selected Model	705
16.7 Comparison of Programs	710
16.7.1 IBM SPSS Package	710
16.7.2 SAS System	712
16.7.3 SYSTAT System	713
17 Time-Series Analysis	714
17.1 General Purpose and Description	714
17.2 Kinds of Research Questions	716
17.2.1 Pattern of Autocorrelation	717
17.2.2 Seasonal Cycles and Trends	717
17.2.3 Forecasting	717
17.2.4 Effect of an Intervention	718
17.2.5 Comparing Time Series	718
17.2.6 Time Series with Covariates	718
17.2.7 Effect Size and Power	718
17.3 Assumptions of Time-Series Analysis	718
17.3.1 Theoretical Issues	718
17.3.2 Practical Issues	718
17.3.2.1 Normality of Distributions of Residuals	719
17.3.2.2 Homogeneity of Variance and Zero Mean of Residuals	719
17.3.2.3 Independence of Residuals	719
17.3.2.4 Absence of Outliers	719
17.3.2.5 Sample Size and Missing Data	719
17.4 Fundamental Equations for Time-Series ARIMA Models	720
17.4.1 Identification of ARIMA (p, d, q) Models	720
17.4.1.1 Trend Components, d : Making the Process Stationary	721
17.4.1.2 Auto-Regressive Components	722
17.4.1.3 Moving Average Components	724
17.4.1.4 Mixed Models	724
17.4.1.5 ACFs and PACFs	724
17.4.2 Estimating Model Parameters	729
17.4.3 Diagnosing a Model	729
17.4.4 Computer Analysis of Small-Sample Time-Series Example	734
17.5 Types of Time-Series Analyses	737
17.5.1 Models with Seasonal Components	737
17.5.2 Models with Interventions	738

17.5.2.1 Abrupt, Permanent Effects	741
17.5.2.2 Abrupt, Temporary Effects	742
17.5.2.3 Gradual, Permanent Effects	745
17.5.2.4 Models with Multiple Interventions	746
17.5.3 Adding Continuous Variables	747
17.6 Some Important Issues	748
17.6.1 Patterns of ACFs and PACFs	748
17.6.2 Effect Size	751
17.6.3 Forecasting	752
17.6.4 Statistical Methods for Comparing Two Models	752
17.7 Complete Examples of Time-Series Analysis	753
17.7.1 Time-Series Analysis of Introduction of Seat Belt Law	753
17.7.1.1 Evaluation of Assumptions	754
17.7.1.2 Baseline Model Identification and Estimation	755
17.7.1.3 Baseline Model Diagnosis	758
17.7.1.4 Intervention Analysis	758
17.7.2. Time-Series Analysis of Introduction of a Dashboard to an Educational Computer Game	762
17.7.2.1 Evaluation of Assumptions	763
17.7.2.2 Baseline Model Identification and Diagnosis	765
17.7.2.3 Intervention Analysis	766
17.8 Comparison of Programs	771
17.8.1 IBM SPSS Package	771
17.8.2 SAS System	774
17.8.3 SYSTAT System	774

18 An Overview of the General Linear Model 775

18.1 Linearity and the General Linear Model	775
18.2 Bivariate to Multivariate Statistics and Overview of Techniques	775
18.2.1 Bivariate Form	775
18.2.2 Simple Multivariate Form	777
18.2.3 Full Multivariate Form	778
18.3 Alternative Research Strategies	782

Appendix A

A Skimpy Introduction to Matrix Algebra	783
A.1 The Trace of a Matrix	784
A.2 Addition or Subtraction of a Constant to a Matrix	784
A.3 Multiplication or Division of a Matrix by a Constant	784
A.4 Addition and Subtraction of Two Matrices	785
A.5 Multiplication, Transposes, and Square Roots of Matrices	785

A.6 Matrix “Division” (Inverses and Determinants)	786
A.7 Eigenvalues and Eigenvectors: Procedures for Consolidating Variance from a Matrix	788

Appendix B

Research Designs for Complete Examples	791
B.1 Women’s Health and Drug Study	791
B.2 Sexual Attraction Study	793
B.3 Learning Disabilities Data Bank	794
B.4 Reaction Time to Identify Figures	794
B.5 Field Studies of Noise-Induced Sleep Disturbance	795
B.6 Clinical Trial for Primary Biliary Cirrhosis	795

B.7 Impact of Seat Belt Law	795
B.8 The Selene Online Educational Game	796

Appendix C

Statistical Tables	797
C.1 Normal Curve Areas	798
C.2 Critical Values of the t Distribution for $\alpha = .05$ and $.01$, Two-Tailed Test	799
C.3 Critical Values of the F Distribution	800
C.4 Critical Values of Chi Square (χ^2)	804
C.5 Critical Values for Squares Multiple Correlation (R^2) in Forward Stepwise Selection: $\alpha = .05$	805
C.6 Critical Values for $F_{MAX} (S^2_{MAX}/S^2_{MIN})$ Distribution for $\alpha = .05$ and $.01$	807
References	808
Index	815

Preface

Some good things seem to go on forever: friendship and updating this book. It is difficult to believe that the first edition manuscript was typewritten, with real cutting and pasting. The publisher required a paper manuscript with numbered pages—that was almost our downfall. We could write a book on multivariate statistics, but we couldn't get the same number of pages (about 1200, double-spaced) twice in a row. SPSS was in release 9.0, and the other program we demonstrated was BMDP. There were a mere 11 chapters, of which 6 of them were describing techniques. Multilevel and structural equation modeling were not yet ready for prime time. Logistic regression and survival analysis were not yet popular.

Material new to this edition includes a redo of all SAS examples, with a pretty new output format and replacement of interactive analyses that are no longer available. We've also re-run the IBM SPSS examples to show the new output format. We've tried to update the references in all chapters, including only classic citations if they date prior to 2000. New work on relative importance has been incorporated in multiple regression, canonical correlation, and logistic regression analysis—complete with demonstrations. Multiple imputation procedures for dealing with missing data have been updated, and we've added a new time-series example, taking advantage of an IBM SPSS expert modeler that replaces previous tea-leaf reading aspects of the analysis.

Our goals in writing the book remain the same as in all previous editions—to present complex statistical procedures in a way that is maximally useful and accessible to researchers who are not necessarily statisticians. We strive to be short on theory but long on conceptual understanding. The statistical packages have become increasingly easy to use, making it all the more critical to make sure that they are applied with a good understanding of what they can and cannot do. But above all else—what does it all mean?

We have not changed the basic format underlying all of the technique chapters, now 14 of them. We start with an overview of the technique, followed by the types of research questions the techniques are designed to answer. We then provide the cautionary tale—what you need to worry about and how to deal with those worries. Then come the fundamental equations underlying the technique, which some readers truly enjoy working through (we know because they helpfully point out any errors and/or inconsistencies they find); but other readers discover they can skim (or skip) the section without any loss to their ability to conduct meaningful analysis of their research. The fundamental equations are in the context of a small, made-up, usually silly data set for which computer analyses are provided—usually IBM SPSS and SAS. Next, we delve into issues surrounding the technique (such as different types of the analysis, follow-up procedures to the main analysis, and effect size, if it is not amply covered elsewhere). Finally, we provide one or two full-bore analyses of an actual real-life data set together with a Results section appropriate for a journal. Data sets for these examples are available at www.pearsonhighered.com in IBM SPSS, SAS, and ASCII formats. We end each technique chapter with a comparison of features available in IBM SPSS, SAS, SYSTAT and sometimes other specialized programs. SYSTAT is a statistical package that we reluctantly had to drop a few editions ago for lack of space.

We apologize in advance for the heft of the book; it is not our intention to line the coffers of chiropractors, physical therapists, acupuncturists, and the like, but there's really just so much to say. As to our friendship, it's still going strong despite living in different cities. Art has taken the place of creating belly dance costumes for both of us, but we remain silly in outlook, although serious in our analysis of research.

The lineup of people to thank grows with each edition, far too extensive to list: students, reviewers, editors, and readers who send us corrections and point out areas of confusion. As always, we take full responsibility for remaining errors and lack of clarity.

Barbara G. Tabachnick

Linda S. Fidell

Chapter 1

Introduction



Learning Objectives

- 1.1 Explain the importance of multivariate techniques in analyzing research data
- 1.2 Describe the basic statistical concepts used in multivariate analysis
- 1.3 Explain how multivariate analysis is used to determine relationships between variables
- 1.4 Summarize the factors to be considered for the selection of variables in multivariate analysis
- 1.5 Summarize the importance of statistical power in research study design
- 1.6 Describe the types of data sets used in multivariate statistics
- 1.7 Outline the organization of the text

1.1 Multivariate Statistics: Why?

Multivariate statistics are increasingly popular techniques used for analyzing complicated data sets. They provide analysis when there are many independent variables (IVs) and/or many dependent variables (DVs), all correlated with one another to varying degrees. Because of the difficulty in addressing complicated research questions with univariate analyses and because of the availability of highly developed software for performing multivariate analyses, multivariate statistics have become widely used. Indeed, a standard univariate statistics course only begins to prepare a student to read research literature or a researcher to produce it.

But how much harder are the multivariate techniques? Compared with the multivariate methods, univariate statistical methods are so straightforward and neatly structured that it is hard to believe they once took so much effort to master. Yet many researchers apply and correctly interpret results of intricate analysis of variance before the grand structure is apparent to them. The same can be true of multivariate statistical methods. Although we are delighted if you gain insights into the full multivariate general linear model,¹ we have accomplished our goal if you feel comfortable selecting and setting up multivariate analyses and interpreting the computer output.

Multivariate methods are more complex than univariate by at least an order of magnitude. However, for the most part, the greater complexity requires few conceptual leaps. Familiar concepts such as sampling distributions and homogeneity of variance simply become more elaborate.

Multivariate models have not gained popularity by accident—or even by sinister design. Their growing popularity parallels the greater complexity of contemporary research.

¹ Chapter 17 attempts to foster such insights.

In psychology, for example, we are less and less enamored of the simple, clean, laboratory study, in which pliant, first-year college students each provide us with a single behavioral measure on cue.

1.1.1 The Domain of Multivariate Statistics: Numbers of IVs and DVs

Multivariate statistical methods are an extension of univariate and bivariate statistics. Multivariate statistics are the *complete* or general case, whereas univariate and bivariate statistics are special cases of the multivariate model. If your design has many variables, multivariate techniques often let you perform a single analysis instead of a series of univariate or bivariate analyses.

Variables are roughly dichotomized into two major types—independent and dependent. Independent variables (IVs) are the differing conditions (treatment vs. placebo) to which you expose your research participants or the characteristics (tall or short) that the participants themselves bring into the research situation. IVs are usually considered predictor variables because they predict the DVs—the response or outcome variables. Note that IV and DV are defined within a research context; a DV in one research setting may be an IV in another.

Additional terms for IVs and DVs are predictor–criterion, stimulus–response, task–performance, or simply input–output. We use IV and DV throughout this book to identify variables that belong on one side of an equation or the other, without causal implication. That is, the terms are used for convenience rather than to indicate that one of the variables caused or determined the size of the other.

The term *univariate statistics* refers to analyses in which there is a single DV. There may be, however, more than one IV. For example, the amount of social behavior of graduate students (the DV) is studied as a function of course load (one IV) and type of training in social skills to which students are exposed (another IV). Analysis of variance is a commonly used univariate statistic.

Bivariate statistics frequently refers to analysis of two variables, where neither is an experimental IV and the desire is simply to study the relationship between the variables (e.g., the relationship between income and amount of education). Bivariate statistics, of course, can be applied in an experimental setting, but usually they are not. Prototypical examples of bivariate statistics are the Pearson product–moment correlation coefficient and chi-square analysis. (Chapter 3 reviews univariate and bivariate statistics.)

With multivariate statistics, you simultaneously analyze multiple dependent and multiple independent variables. This capability is important in both nonexperimental (correlational or survey) and experimental research.

1.1.2 Experimental and Nonexperimental Research

A critical distinction between experimental and nonexperimental research is whether the researcher manipulates the levels of the IVs. In an experiment, the researcher has control over the levels (or conditions) of at least one IV to which a participant is exposed by determining what the levels are, how they are implemented, and how and when cases are assigned and exposed to them. Further, the experimenter randomly assigns cases to levels of the IV and controls all other influential factors by holding them constant, counterbalancing, or randomizing their influence. Scores on the DV are expected to be the same, within random variation, except for the influence of the IV (Shadish, Cook, and Campbell, 2002). If there are systematic differences in the DV associated with levels of the IV, these differences are attributed to the IV.

For example, if groups of undergraduates are randomly assigned to the same material but different types of teaching techniques, and afterward some groups of undergraduates perform better than others, the difference in performance is said, with some degree of confidence, to be caused by the difference in teaching technique. In this type of research, the terms *independent* and *dependent* have obvious meaning: the value of the DV depends on the manipulated level of the IV. The IV is manipulated by the experimenter and the score on the DV depends on the level of the IV.

In nonexperimental (correlational or survey) research, the levels of the IV(s) are not manipulated by the researcher. The researcher can define the IV, but has no control over the assignment of cases to levels of it. For example, groups of people may be categorized into geographic area of residence (Northeast, Midwest, etc.), but only the definition of the variable is under researcher control. Except for the military or prison, place of residence is rarely subject to manipulation by a researcher. Nevertheless, a naturally occurring difference like this is often considered an IV and is used to predict some other nonexperimental (dependent) variable such as income. In this type of research, the distinction between IVs and DVs is usually arbitrary and many researchers prefer to call IVs *predictors* and DVs *criterion variables*.

In nonexperimental research, it is very difficult to attribute causality to an IV. If there is a systematic difference in a DV associated with levels of an IV, the two variables are said (with some degree of confidence) to be related, but the cause of the relationship is unclear. For example, income as a DV might be related to geographic area, but no causal association is implied.

Nonexperimental research takes many forms, but a common example is the survey. Typically, many people are surveyed, and each respondent provides answers to many questions, producing a large number of variables. These variables are usually interrelated in highly complex ways, but univariate and bivariate statistics are not sensitive to this complexity. Bivariate correlations between all pairs of variables, for example, could not reveal that the 20 to 25 variables measured really represent only two or three “supervariables.”

If a research goal is to distinguish among subgroups in a sample (e.g., between Catholics and Protestants) on the basis of a variety of attitudinal variables, we could use several univariate *t* tests (or analyses of variance) to examine group differences on each variable separately. But if the variables are related, which is highly likely, the results of many *t* tests are misleading and statistically suspect.

With the use of multivariate statistical techniques, complex interrelationships among variables are revealed and assessed in statistical inference. Further, it is possible to keep the overall Type I error rate at, say, 5%, no matter how many variables are tested.

Although most multivariate techniques were developed for use in nonexperimental research, they are also useful in experimental research, in which there may be multiple IVs and multiple DVs. With multiple IVs, the research is usually designed so that the IVs are independent of each other and a straightforward correction for numerous statistical tests is available (see Chapter 3). With multiple DVs, a problem of inflated error rate arises if each DV is tested separately. Further, at least some of the DVs are likely to be correlated with each other, so separate tests of each DV reanalyze some of the same variance. Therefore, multivariate tests are used.

Experimental research designs with multiple DVs were unusual at one time. Now, however, with attempts to make experimental designs more realistic, and with the availability of computer programs, experiments often have several DVs. It is dangerous to run an experiment with only one DV and risk missing the impact of the IV because the most sensitive DV is not measured. Multivariate statistics help the experimenter design more efficient and more realistic experiments by allowing measurement of multiple DVs without violation of acceptable levels of Type I error.

One of the few considerations not relevant to choice of *statistical* technique is whether the data are experimental or correlational. The statistical methods “work” whether the researcher manipulated the levels of the IV or not. But attribution of causality to results is crucially affected by the experimental–nonexperimental distinction.

1.1.3 Computers and Multivariate Statistics

One answer to the question “Why multivariate statistics?” is that the techniques are now accessible by computer. Only the most dedicated number cruncher would consider doing real-life-sized problems in multivariate statistics without a computer. Fortunately, excellent multivariate programs are available in a number of computer packages.

Two packages are demonstrated in this book. Examples are based on programs in IBM SPSS and SAS.

If you have access to both packages, you are indeed fortunate. Programs within the packages do not completely overlap, and some problems are better handled through one package than the other. For example, doing several versions of the same basic analysis on the same set of data is particularly easy with IBM SPSS, whereas SAS has the most extensive capabilities for saving derived scores from data screening or from intermediate analyses.

Chapters 5 through 17 (the chapters that cover the specialized multivariate techniques) offer explanations and illustrations of a variety of programs² within each package and a comparison of the features of the programs. We hope that once you understand the techniques, you will be able to generalize to virtually any multivariate program.

Recent versions of the programs are available in Windows, with menus that implement most of the techniques illustrated in this book. All of the techniques may be implemented through syntax, and syntax itself is generated through menus. Then you may add or change syntax as desired for your analysis. For example, you may “paste” menu choices into a syntax window in IBM SPSS, edit the resulting text, and then run the program. Also, syntax generated by IBM SPSS menus is saved in the “journal” file (statistics.jnl), which may also be accessed and copied into a syntax window. Syntax generated by SAS menus is recorded in a “log” file. The contents may then be copied to an interactive window, edited, and run. Do not overlook the help files in these programs. Indeed, SAS and IBM SPSS now provide the entire set of user manuals online, often with more current information than is available in printed manuals.

Our IBM SPSS demonstrations in this book are based on syntax generated through menus whenever feasible. We would love to show you the sequence of menu choices, but space does not permit. And, for the sake of parsimony, we have edited program output to illustrate the material that we feel is the most important for interpretation.

With commercial computer packages, you need to know which version of the package you are using. Programs are continually being changed, and not all changes are immediately implemented at each facility. Therefore, many versions of the various programs are simultaneously in use at different institutions; even at one institution, more than one version of a package is sometimes available.

Program updates are often corrections of errors discovered in earlier versions. Sometimes, a new version will change the output format but not its information. Occasionally, though, there are major revisions in one or more programs or a new program is added to the package. Sometimes defaults change with updates, so that the output looks different although syntax is the same. Check to find out which version of each package you are using. Then, if you are using a printed manual, be sure that the manual you are using is consistent with the version in use at your facility. Also check updates for error correction in previous releases that may be relevant to some of your previous runs.

Except where noted, this book reviews Windows versions of IBM SPSS Version 24 and SAS Version 9.4. Information on availability and versions of software, macros, books, and the like changes almost daily. We recommend the Internet as a source of “keeping up.”

1.1.4 Garbage In, Roses Out?

The trick in multivariate statistics is not in computation. This is easily done as discussed above. The trick is to select reliable and valid measurements, choose the appropriate program, use it correctly, and know how to interpret the output. Output from commercial computer programs, with their beautifully formatted tables, graphs, and matrices, can make garbage look like roses. Throughout this book, we try to suggest clues that reveal when the true message in the output more closely resembles the fertilizer than the flowers.

Second, when you use multivariate statistics, you rarely get as close to the raw data as you do when you apply univariate statistics to a relatively few cases. Errors and anomalies

² We have retained descriptions of features of SYSTAT (Version 13) in these sections, despite the removal of detailed demonstrations of that program in this edition.

in the data that would be obvious if the data were processed by hand are less easy to spot when processing is entirely by computer. But the computer packages have programs to graph and describe your data in the simplest univariate terms and to display bivariate relationships among your variables. As discussed in Chapter 4, these programs provide preliminary analyses that are absolutely necessary if the results of multivariate programs are to be believed.

There are also certain costs associated with the benefits of using multivariate procedures. Benefits of increased flexibility in research design, for instance, are sometimes paralleled by increased ambiguity in interpretation of results. In addition, multivariate results can be quite sensitive to which analytic strategy is chosen (cf. Section 1.2.4) and do not always provide better protection against statistical errors than their univariate counterparts. Add to this the fact that occasionally you still cannot get a firm statistical answer to your research questions, and you may wonder if the increase in complexity and difficulty is warranted.

Frankly, we think it is. Slippery as some of the concepts and procedures are, these statistics provide insights into relationships among variables that may more closely resemble the complexity of the “real” world. And sometimes you get at least partial answers to questions that could not be asked at all in the univariate framework. For a complete analysis, making sense of your data usually requires a judicious mix of multivariate and univariate statistics.

The addition of multivariate statistical methods to your repertoire makes data analysis a lot more fun. If you liked univariate statistics, you will love multivariate statistics!³

1.2 Some Useful Definitions

In order to describe multivariate statistics easily, it is useful to review some common terms in research design and basic statistics. Distinctions were made between IVs and DVs and between experimental and nonexperimental research in preceding sections. Additional terms that are encountered repeatedly in the book but not necessarily related to each other are described in this section.

1.2.1 Continuous, Discrete, and Dichotomous Data

In applying statistical techniques of any sort, it is important to consider the type of measurement and the nature of the correspondence between the numbers and the events that they represent. The distinction made here is among continuous, discrete, and dichotomous variables; you may prefer to substitute the terms *interval* or *quantitative* for *continuous* and *nominal*, *categorical* or *qualitative* for *dichotomous* and *discrete*.

Continuous variables are measured on a scale that changes values smoothly rather than in steps. Continuous variables take on any values within the range of the scale, and the size of the number reflects the amount of the variable. Precision is limited by the measuring instrument, not by the nature of the scale itself. Some examples of continuous variables are time as measured on an old-fashioned analog clock face, annual income, age, temperature, distance, and grade point average (GPA).

Discrete variables take on a finite and usually small number of values, and there is no smooth transition from one value or category to the next. Examples include time as displayed by a digital clock, continents, categories of religious affiliation, and type of community (rural or urban).

Sometimes discrete variables are used in multivariate analyses as if continuous if there are numerous categories and the categories represent a quantitative attribute. For instance, a variable that represents age categories (where, say, 1 stands for 0 to 4 years, 2 stands for 5 to 9 years, 3 stands for 10 to 14 years, and so on up through the normal age span) can be used because there are a lot of categories and the numbers designate a quantitative attribute (increasing age). But the same numbers used to designate categories of religious affiliation are not in appropriate

³ Don't even think about it.

form for analysis with many of the techniques⁴ because religions do not fall along a quantitative continuum.

Discrete variables composed of qualitatively different categories are sometimes analyzed after being changed into a number of dichotomous or two-level variables (e.g., Catholic vs. non-Catholic, Protestant vs. non-Protestant, Jewish vs. non-Jewish, and so on until the degrees of freedom are used). Recategorization of a discrete variable into a series of dichotomous ones is called *dummy variable coding*. The conversion of a discrete variable into a series of dichotomous ones is done to limit the relationship between the dichotomous variables and others to linear relationships. A discrete variable with more than two categories can have a relationship of any shape with another variable, and the relationship is changed arbitrarily if the assignment of numbers to categories is changed. Dichotomous variables, however, with only two points, can have only linear relationships with other variables; they are, therefore, appropriately analyzed by methods using correlation in which only linear relationships are analyzed.

The distinction between continuous and discrete variables is not always clear. If you add enough digits to the digital clock, for instance, it becomes for all practical purposes a continuous measuring device, whereas time as measured by the analog device can also be read in discrete categories such as hours or half hours. In fact, any continuous measurement may be rendered discrete (or dichotomous) with some loss of information, by specifying cutoffs on the continuous scale.

The property of variables that is crucial to the application of multivariate procedures is not the type of measurement so much as the shape of distribution, as discussed in Chapter 4 and in discussions of tests of assumptions in Chapters 5 through 17. Non-normally distributed continuous variables and dichotomous variables with very uneven splits between the categories present problems to several of the multivariate analyses. This issue and its resolution are discussed at some length in Chapter 4.

Another type of measurement that is used sometimes produces a rank order scale. This scale assigns a number to each case to indicate the case's position vis-à-vis other cases along some dimension. For instance, ranks are assigned to contestants (first place, second place, third place, etc.) to provide an indication of who is the best—but not by how much. A problem with rank order measures is that their distributions are rectangular (one frequency per number) instead of normal, unless tied ranks are permitted and they pile up in the middle of the distribution.

In practice, we often treat variables as if they are continuous when the underlying scale is thought to be continuous, but the measured scale actually is rank order, the number of categories is large—say, seven or more, and the data meet other assumptions of the analysis. For instance, the number of correct items on an objective test is technically not continuous because fractional values are not possible, but it is thought to measure some underlying continuous variable such as course mastery. Another example of a variable with ambiguous measurement is one measured on a Likert-type scale, in which consumers rate their attitudes toward a product as “strongly like,” “moderately like,” “mildly like,” “neither like nor dislike,” “mildly dislike,” “moderately dislike,” or “strongly dislike.” As mentioned previously, even dichotomous variables may be treated as if continuous under some conditions. Thus, we often use the term *continuous*, throughout the remainder of this book, whether the measured scale itself is continuous or the variable is to be treated as if continuous. We use the term *discrete* for variables with a few categories, whether the categories differ in type or quantity.

1.2.2 Samples and Populations

Samples are measured to make generalizations about populations. Ideally, samples are selected, usually by some random process, so that they represent the population of interest. In real life, however, populations are frequently best defined in terms of samples, rather than vice versa; the population is the group from which you were able to randomly sample.

⁴ Some multivariate techniques (e.g., logistic regression, SEM) are appropriate for all types of variables.

Sampling has somewhat different connotations in nonexperimental and experimental research. In nonexperimental research, you investigate relationships among variables in some predefined population. Typically, you take elaborate precautions to ensure that you have achieved a representative sample of that population; you define your population, and then do your best to randomly sample from it.⁵

In experimental research, you attempt to create different populations by treating subgroups from an originally homogeneous group differently. The sampling objective here is to ensure that all cases come from the same population before you treat them differently. Random sampling consists of randomly assigning cases to treatment groups (levels of the IV) to ensure that, before differential treatment, all subsamples come from the same population. Statistical tests provide evidence as to whether, after treatment, all samples still come from the same population. Generalizations about treatment effectiveness are made to the type of individuals who participated in the experiment.

1.2.3 Descriptive and Inferential Statistics

Descriptive statistics describe samples of cases in terms of variables or combinations of variables. Inferential statistical techniques test hypotheses about differences in populations on the basis of measurements made on samples of cases. If statistically significant differences are found, descriptive statistics are then used to provide estimations of central tendency, and the like, in the population. Descriptive statistics used in this way are called *parameter estimates*.

Use of inferential and descriptive statistics is rarely an either–or proposition. We are usually interested in both describing and making inferences about a data set. We describe the data, find statistically significant differences or relationships, and estimate population values for those findings. However, there are more restrictions on inference than there are on description. Many assumptions of multivariate statistical methods are necessary only for inference. If simple description of the sample is the major goal, many assumptions are relaxed, as discussed in Chapters 5 through 17.

1.2.4 Orthogonality: Standard and Sequential Analyses

Orthogonality is a perfect nonassociation between variables. If two variables are orthogonal, knowing the value of one variable gives no clue as to the value of the other; the correlation between them is zero.

Orthogonality is often desirable in statistical applications. For instance, factorial designs for experiments are orthogonal when two or more IVs are completely crossed with equal sample sizes in each combination of levels. Except for use of a common error term, tests of hypotheses about main effects and interactions are independent of each other; the outcome of each test gives no hint as to the outcome of the others. In orthogonal experimental designs with random assignment of cases, manipulation of the levels of the IV, and good controls, changes in value of the DV can be unambiguously attributed to various main effects and interactions.

Similarly, in multivariate analyses, there are advantages if sets of IVs or DVs are orthogonal. If all pairs of IVs in a set are orthogonal, each IV adds, in a simple fashion, to prediction of the DV. Consider income as a DV with education and occupational prestige as IVs. If education and occupational prestige are orthogonal, and if 35% of the variability in income may be predicted from education and a different 45% is predicted from occupational prestige, then 80% of the variance in income (the DV, Y) is predicted from education and occupational prestige together.

Orthogonality can easily be illustrated in Venn diagrams, as shown in Figure 1.1. Venn diagrams represent shared variance (or correlation) as overlapping areas between two (or more) circles. The total variance for income is one circle. The section with horizontal stripes represents the part of income predictable from education (the first IV, X_1), and the section with

⁵ Strategies for random sampling are discussed in many sources, including Levy and Lemenshow (2009), Rea and Parker (1997), and de Vaus (2002).

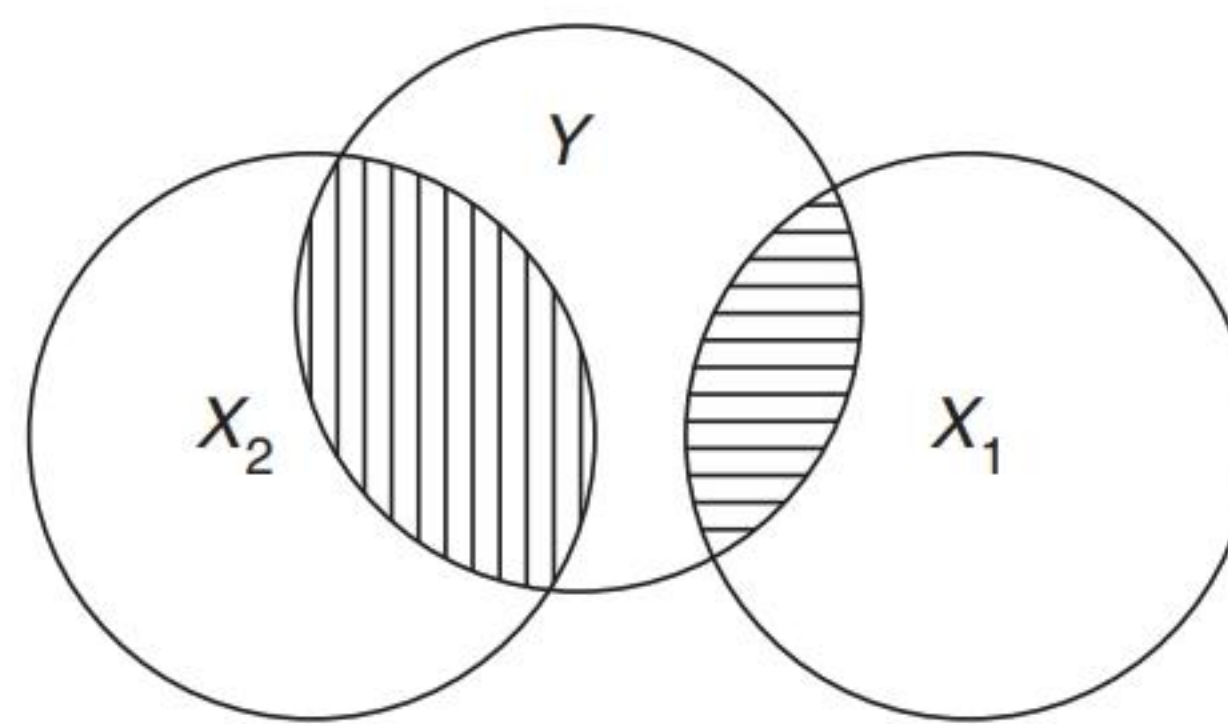


Figure 1.1 Venn diagram for Y (income), X_1 (education), and X_2 (occupational prestige).

vertical stripes represents the part predictable from occupational prestige (the second IV, X_2); the circle for education overlaps the circle for income 35% and the circle for occupational prestige overlaps 45%. Together, they account for 80% of the variability in income because education and occupational prestige are orthogonal and do not themselves overlap. There are similar advantages if a set of DVs is orthogonal. The overall effect of an IV can be partitioned into effects on each DV in an additive fashion.

Usually, however, the variables are correlated with each other (nonorthogonal). IVs in nonexperimental designs are often correlated naturally; in experimental designs, IVs become correlated when unequal numbers of cases are measured in different cells of the design. DVs are usually correlated because individual differences among participants tend to be consistent over many attributes.

When variables are correlated, they have shared or overlapping variance. In the example of Figure 1.2, education and occupational prestige correlate with each other. Although the independent contribution made by education is still 35% and that by occupational prestige is 45%, their joint contribution to prediction of income is not 80%, but rather something smaller due to the overlapping area shown by the arrow in Figure 1.2(a). A major decision for the multivariate analyst is how to handle the variance that is predictable from more than one variable. Many multivariate techniques have at least two strategies for handling it, but some have more.

In standard analysis, the overlapping variance contributes to the size of summary statistics of the overall relationship but is not assigned to either variable. Overlapping variance is disregarded in assessing the contribution of each variable to the solution. Figure 1.2(a) is a Venn diagram of a standard analysis in which overlapping variance is shown as overlapping areas in circles; the unique contributions of X_1 and X_2 to prediction of Y are shown as horizontal and vertical areas, respectively, and the total relationship between Y and the combination of X_1 and X_2 is those two areas plus the area with the arrow. If X_1 is education and X_2 is occupational prestige, then in standard analysis, X_1 is "credited with" the area marked by the horizontal lines and X_2 by the area marked by vertical lines. Neither of the IVs is assigned the area designated with the arrow. When X_1 and X_2 substantially overlap each other, very little horizontal or vertical area may be left for either of them, despite the fact that they are both related to Y . They have essentially knocked each other out of the solution.

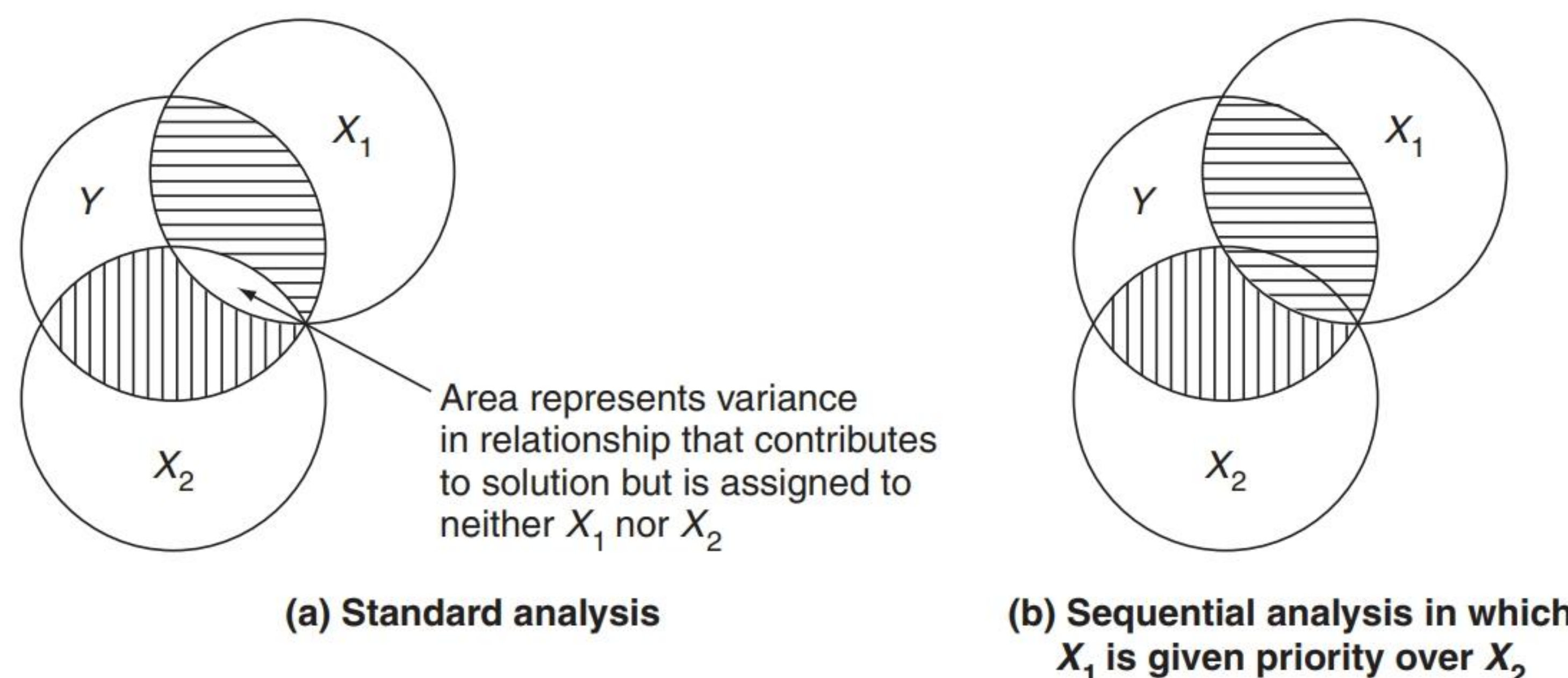


Figure 1.2 Standard (a) and sequential (b) analyses of the relationship between Y , X_1 , and X_2 . Horizontal shading depicts variance assigned to X_1 . Vertical shading depicts variance assigned to X_2 .

Sequential analyses differ, in that the researcher assigns priority for entry of variables into equations, and the first one to enter is assigned both unique variance and any overlapping variance it has with other variables. Lower-priority variables are then assigned on entry their unique and any remaining overlapping variance. Figure 1.2(b) shows a sequential analysis for the same case as Figure 1.2(a), where X_1 (education) is given priority over X_2 (occupational prestige). The total variance explained is the same as in Figure 1.2(a), but the relative contributions of X_1 and X_2 have changed; education now shows a stronger relationship with income than in the standard analysis, whereas the relation between occupational prestige and income remains the same.

The choice of strategy for dealing with overlapping variance is not trivial. If variables are correlated, the overall relationship remains the same, but the apparent importance of variables to the solution changes depending on whether a standard or a sequential strategy is used. If the multivariate procedures have a reputation for unreliability, it is because solutions change, sometimes dramatically, when different strategies for entry of variables are chosen. However, the strategies also ask different questions of the data, and it is incumbent on the researcher to determine exactly which question to ask. We try to make the choices clear in the chapters that follow.

1.3 Linear Combinations of Variables

Multivariate analyses combine variables to do useful work, such as predict scores or predict group membership. The combination that is formed depends on the relationships among the variables and the goals of analysis, but in most cases, the combination is linear. A linear combination is one in which each variable is assigned a weight (e.g., W_1), and then the products of weights and the variable scores are summed to predict a score on a combined variable. In Equation 1.1, Y' (the predicted DV) is predicted by a linear combination of X_1 and X_2 (the IVs).

$$Y' = W_1X_1 + W_2X_2 \quad (1.1)$$

If, for example, Y' is predicted income, X_1 is education, and X_2 is occupational prestige, the best prediction of income is obtained by weighting education (X_1) by W_1 and occupational prestige (X_2) by W_2 before summing. No other values of W_1 and W_2 produce as good a prediction of income.

Notice that Equation 1.1 includes neither X_1 nor X_2 raised to powers (exponents) nor a product of X_1 and X_2 . This seems to severely restrict multivariate solutions until one realizes that X_1 could itself be a product of two different variables or a single variable raised to a power. For example, X_1 might be education squared. A multivariate solution does not produce exponents or cross-products of IVs to improve a solution, but the researcher can include X s that are cross-products of IVs or are IVs raised to powers. Inclusion of variables raised to powers or cross-products of variables has both theoretical and practical implications for the solution. Berry (1993) provides a useful discussion of many of the issues.

The size of the W values (or some function of them) often reveals a great deal about the relationship between DVs and IVs. If, for instance, the W value for some IV is zero, the IV is not needed in the best DV–IV relationship. Or if some IV has a large W value, then the IV tends to be important to the relationship. Although complications (to be explained later) prevent interpretation of the multivariate solution from the sizes of the W values alone, they are nonetheless important in most multivariate procedures.

The combination of variables can be considered a supervariable, not directly measured but worthy of interpretation. The supervariable may represent an underlying dimension that predicts something or optimizes some relationship. Therefore, the attempt to understand the meaning of the combination of IVs is worthwhile in many multivariate analyses.

In the search for the best weights to apply in combining variables, computers do not try out all possible sets of weights. Various algorithms have been developed to compute the weights. Most algorithms involve manipulation of a correlation matrix, a variance–covariance matrix, or a sum-of-squares and cross-products matrix. Section 1.6 describes these matrices in very

simple terms and shows their development from a very small data set. Appendix A describes some terms and manipulations appropriate to matrices. In the fourth sections of Chapters 5 through 17, a small hypothetical sample of data is analyzed by hand to show how the weights are derived for each analysis. Though this information is useful for a basic understanding of multivariate statistics, it is not necessary for applying multivariate techniques fruitfully to your research questions and may, sadly, be skipped by those who are math averse.

1.4 Number and Nature of Variables to Include

Attention to the number of variables included in an analysis is important. A general rule is to get the best solution with the fewest variables. As more and more variables are included, the solution usually improves, but only slightly. Sometimes the improvement does not compensate for the cost in degrees of freedom of including more variables, so the power of the analyses diminishes.

A second problem is *overfitting*. With overfitting, the solution is very good; so good, in fact, that it is unlikely to generalize to a population. Overfitting occurs when too many variables are included in an analysis relative to the sample size. With smaller samples, very few variables can be analyzed. Generally, a researcher should include only a limited number of uncorrelated variables in each analysis,⁶ fewer with smaller samples. We give guidelines for the number of variables that can be included relative to sample size in the third sections of Chapters 5 through 17.

Additional considerations for inclusion of variables in a multivariate analysis include cost, availability, meaning, and theoretical relationships among the variables. Except in analysis of structure, one usually wants a small number of valid, cheaply obtained, easily available, uncorrelated variables that assess all the theoretically important dimensions of a research area. Another important consideration is reliability. How stable is the position of a given score in a distribution of scores when measured at different times or in different ways? Unreliable variables degrade an analysis, whereas reliable ones enhance it. A few reliable variables give a more meaningful solution than a large number of less reliable variables. Indeed, if variables are sufficiently unreliable, the entire solution may reflect only measurement error. Further considerations for variable selection are mentioned as they apply to each analysis.

1.5 Statistical Power

A critical issue in designing any study is whether there is adequate power. Power, as you may recall, represents the probability that effects that actually exist have a chance of producing statistical significance in your eventual data analysis. For example, do you have a large enough sample size to show a significant relationship between GRE and GPA if the actual relationship is fairly large? What if the relationship is fairly small? Is your sample large enough to reveal significant effects of treatment on your DV(s)? Relationships among power and errors of inference are discussed in Chapter 3.

Issues of power are best considered in the planning state of a study when the researcher determines the required sample size. The researcher estimates the size of the anticipated effect (e.g., an expected mean difference), the variability expected in assessment of the effect, the desired alpha level (ordinarily .05), and the desired power (often .80). These four estimates are required to determine the necessary sample size. Failure to consider power in the planning stage often results in failure to find a significant effect (and an unpublishable study). The interested reader may wish to consult Cohen (1988), Rossi (1990), Sedlmeier and Gigerenzer (1989), or Murphy, Myers, and Wolach (2014) for more detail.

There is a great deal of software available to help you estimate the power available with various sample sizes for various statistical techniques, and to help you determine necessary

⁶ The exceptions are analysis of structure, such as factor analysis, in which numerous correlated variables are measured.

sample size given a desired level of power (e.g., an 80% probability of achieving a significant result if an effect exists) and expected sizes of relationships. One of these programs that estimates power for several techniques is NCSS PASS (Hintze, 2017). Many other programs are reviewed (and sometimes available as shareware) on the Internet. Issues of power relevant to each of the statistical techniques are discussed in Chapters 5 through 17.

1.6 Data Appropriate for Multivariate Statistics

An appropriate data set for multivariate statistical methods consists of values on a number of variables for each of several participants or cases. For continuous variables, the values are scores on variables. For example, if the continuous variable is the GRE, the values for the various participants are scores such as 500, 420, and 650. For discrete variables, values are number codes for group membership or treatment. For example, if there are three teaching techniques, students who receive one technique are arbitrarily assigned a “1,” those receiving another technique are assigned a “2”.

1.6.1 The Data Matrix

The data matrix is an organization of scores in which rows (lines) represent participants and columns represent variables. An example of a data matrix with six participants⁷ and four variables is given in Table 1.1. For example, X_1 might be type of teaching technique, X_2 score on the GRE, X_3 GPA, and X_4 gender, with women coded 1 and men coded 2.

Data are entered into a data file with long-term storage accessible by computer in order to apply computer techniques to them. Each participant starts with a new row (line). Information identifying the participant is typically entered first, followed by the value of each variable for that participant.

Scores for each variable are entered in the same order for each student. If there are more data for each case that can be accommodated on a single line, the data are continued on additional lines, but all of the data for each case are kept together. All of the computer package manuals provide information on setting up a data matrix.

In this example, there are values for every variable for each student. This is not always the case with research in the real world. With large numbers of cases and variables, scores are frequently missing on some variables for some cases. For instance, respondents may refuse to answer some kinds of questions, or some students may be absent the day when a particular test is given, and so forth. This creates missing values in the data matrix. To deal with missing values, first build a data file in which some symbol is used to indicate that a value on a variable is missing in data for a case. The various programs have standard symbols, such as a dot (.), for this purpose. You can also use other symbols, but it is often just as convenient to use one of the default symbols. Once the data set is available, consult Chapter 4 for various options to deal with this messy (but often unavoidable) problem.

Table 1.1 A Data Matrix of Hypothetical Scores

Student	X_1	X_2	X_3	X_4
1	1	500	3.20	1
2	1	420	2.50	2
3	2	650	3.90	1
4	2	550	3.50	2
5	3	480	3.30	1
6	3	600	3.25	2

⁷Normally, of course, there are many more than six cases.

Table 1.2 Correlation Matrix for Part of Hypothetical Data for Table 1.1

	X_2	X_3	X_4	
$\mathbf{R} =$	X_2	1.00	.85	-.13
	X_3	.85	1.00	-.46
	X_4	-.13	-.46	1.00

1.6.2 The Correlation Matrix

Most readers are familiar with \mathbf{R} , a correlation matrix. \mathbf{R} is a square, symmetrical matrix. Each row (and each column) represents a different variable, and the value at the intersection of each row and column is the correlation between the two variables. For instance, the value at the intersection of the second row, third column, is the correlation between the second and the third variables. The same correlation also appears at the intersection of the third row, second column. Thus, correlation matrices are said to be symmetrical about the main diagonal, which means they are mirror images of themselves above and below the diagonal from top left to bottom right. Hence, it is common practice to show only the bottom half or the top half of an \mathbf{R} matrix. The entries in the main diagonal are often omitted as well, since they are all ones—correlations of variables with themselves.⁸

Table 1.2 shows the correlation matrix for X_2 , X_3 , and X_4 of Table 1.1. The value .85 is the correlation between X_2 and X_3 and it appears twice in the matrix (as do other values). Other correlations are as indicated in the table.

Many programs allow the researcher a choice between analysis of a correlation matrix and analysis of a variance–covariance matrix. If the correlation matrix is analyzed, a unit-free result is produced. That is, the solution reflects the relationships among the variables but not in the metric in which they are measured. If the metric of the scores is somewhat arbitrary, analysis of \mathbf{R} is appropriate.

1.6.3 The Variance–Covariance Matrix

If scores are measured along a meaningful scale, it is sometimes appropriate to analyze a variance–covariance matrix. A variance–covariance matrix, Σ , is also square and symmetrical, but the elements in the main diagonal are the variances of each variable, and the off-diagonal elements are covariances between pairs of different variables.

Variances, as you recall, are averaged squared deviations of each score from the mean of the scores. Since the deviations are averaged, the number of scores included in computation of a variance is not relevant, but the metric in which the scores are measured is relevant. Scores measured in large numbers tend to have large numbers as variances, and scores measured in small numbers tend to have small variances.

Covariances are averaged cross-products (product of the deviation between one variable and its mean and the deviation between a second variable and its mean). Covariances are similar to correlations except that they, like variances, retain information concerning the scales in which the variables are measured. The variance–covariance matrix for the continuous data in Table 1.1 appears in Table 1.3.

Table 1.3 Variance–Covariance Matrix for Part of Hypothetical Data of Table 1.1

	X_2	X_3	X_4	
$\Sigma =$	X_2	7026.66	32.80	–6.00
	X_3	32.80	.21	–.12
	X_4	–6.00	–.12	.30

⁸ Alternatively, other information such as standard deviations is inserted.

1.6.4 The Sum-of-Squares and Cross-Products Matrix

The matrix, \mathbf{S} , is a precursor to the variance–covariance matrix in which deviations are not yet averaged. Thus, the size of the entries depends on the number of cases as well as on the metric in which the elements were measured. The sum-of-squares and cross-products matrix for X_2 , X_3 , and X_4 in Table 1.1 appears in Table 1.4.

The entry in the major diagonal of the matrix \mathbf{S} is the sum of squared deviations of scores from the mean for that variable, hence, “sum of squares,” or SS. That is, for each variable, the value in the major diagonal is

$$SS(X_j) = \sum_{i=1}^N (X_{ij} - \bar{X}_j)^2 \quad (1.2)$$

where $i = 1, 2, \dots, N$
 $N =$ the number of cases
 $j =$ the variable identifier
 $X_{ij} =$ the score on variable j by case i
 $\bar{X}_j =$ the mean of all scores on the j th variable

For example, for X_4 , the mean is 1.5. The sum of squared deviations around the mean and the diagonal value for the variable is

$$\begin{aligned} \sum_{i=1}^6 (X_{i4} - \bar{X}_4)^2 &= (1 - 1.5)^2 + (2 - 1.5)^2 + (1 - 1.5)^2 + (2 - 1.5)^2 + (1 - 1.5)^2 + (2 - 1.5)^2 \\ &= 1.50 \end{aligned}$$

The off-diagonal elements of the sum-of-squares and cross-products matrix are the cross-products—the sum of products (SP)—of the variables. For each pair of variables, represented by row and column labels in Table 1.4, the entry is the sum of the product of the deviation of one variable around its mean times the deviation of the other variable around its mean.

$$SP(X_j X_k) = \sum_{i=1}^N (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k) \quad (1.3)$$

where j identifies the first variable, k identifies the second variable, and all other terms are as defined in Equation 1.1. (Note that if $j = k$, Equation 1.3 becomes identical to Equation 1.2.)

For example, the cross-product term for variables X_2 and X_3 is

$$\begin{aligned} \sum_{i=1}^N (X_{i2} - \bar{X}_2)(X_{i3} - \bar{X}_3) &= (500 - 533.33)(3.20 - 3.275) + (420 - 533.33)(2.50 - 3.275) \\ &+ \dots + (600 - 533.33)(3.25 - 3.275) = 164.00 \end{aligned}$$

Most computations start with \mathbf{S} and proceed to Σ or \mathbf{R} . The progression from a sum-of-squares and cross-products matrix to a variance–covariance matrix is simple.

$$\Sigma = \frac{1}{N - 1} \mathbf{S} \quad (1.4)$$

Table 1.4 Sum-of-Squares and Cross-Products Matrix for Part of Hypothetical Data of Table 1.1

	X_2	X_3	X_4	
$\mathbf{S} =$	X_2	35133.33	164.00	−30.00
	X_3	164.00	1.05	−0.58
	X_4	−30.00	−0.58	1.50

The variance–covariance matrix is produced by dividing every element in the sum-of-squares and cross-products matrix by $N - 1$, where N is the number of cases.

The correlation matrix is derived from an **S** matrix by dividing each sum-of-squares by itself (to produce the 1s in the main diagonal of **R**) and each cross-product of the **S** matrix by the square root of the product of the sum-of-squared deviations around the mean for each of the variables in the pair. That is, each cross-product is divided by

$$\text{Denominator}(X_j X_k) = \sqrt{\sum (X_{ij} - X_j)^2 \sum (X_{ik} - X_k)^2} \quad (1.5)$$

where terms are defined as in Equation 1.3.

For some multivariate operations, it is not necessary to feed the data matrix to a computer program. Instead, an **S** or an **R** matrix is entered, with each row (representing a variable) starting a new line. Often, considerable computing time and expense are saved by entering one or the other of these matrices rather than raw data.

1.6.5 Residuals

Often a goal of analysis or test of its efficiency is its ability to reproduce the values of a DV or the correlation matrix of a set of variables. For example, we might want to predict scores on the GRE (X_2) of Table 1.1 from knowledge of GPA (X_3) and gender (X_4). After applying the proper statistical operations—a multiple regression in this case—a predicted GRE score for each student is computed by applying the proper weights for GPA and gender to the GPA, and gender scores for each student. But because we already obtained GRE scores for the sample of students, we are able to compare the predicted score with the obtained GRE score. The difference between the predicted and obtained values is known as the *residual* and is a measure of error of prediction.

In most analyses, the residuals for the entire sample sum to zero. That is, sometimes the prediction is too large and sometimes it is too small, but the average of all the errors is zero. The squared value of the residuals, however, provides a measure of how good the prediction is. When the predictions are close to the obtained values, the squared errors are small. The way that the residuals are distributed is of further interest in evaluating the degree to which the data meet the assumptions of multivariate analyses, as discussed in Chapter 4 and elsewhere.

1.7 Organization of the Book

Chapter 2 gives a guide to the multivariate techniques that are covered in this book and places them in context with the more familiar univariate and bivariate statistics where possible. Chapter 2 includes a flow chart that organizes statistical techniques on the basis of the major research questions asked. Chapter 3 provides a brief review of univariate and bivariate statistical techniques for those who are interested.

Chapter 4 deals with the assumptions and limitations of multivariate statistical methods. Assessment and violation of assumptions are discussed, along with alternatives for dealing with violations when they occur. The reader is guided back to Chapter 4 frequently in Chapters 5 through 17.

Chapters 5 through 17 cover specific multivariate techniques. They include descriptive, conceptual sections as well as a guided tour through a real-world data set for which the analysis is appropriate. The tour includes an example of a Results section describing the outcome of the statistical analysis appropriate for submission to a professional journal. Each technique chapter includes a comparison of computer programs. You may want to vary the order in which you cover these chapters.

Chapter 18 is an attempt to integrate univariate, bivariate, and multivariate statistics through the multivariate general linear model. The common elements underlying all the techniques are emphasized, rather than the differences among them. Chapter 18 is meant to pull together the material in the remainder of the book with a conceptual rather than pragmatic emphasis. Some may wish to consider this material earlier, for instance, immediately after Chapter 2.

Chapter 2

A Guide to Statistical Techniques

Using the Book



Learning Objectives

- 2.1 Determine statistical techniques based on the type of research questions
- 2.2 Determine when to use specific analytical strategies
- 2.3 Use a decision tree to determine selection of statistical techniques
- 2.4 Outline the organization of the technique chapters
- 2.5 Summarize the primary requirements before selecting a statistical technique

2.1 Research Questions and Associated Techniques

This chapter organizes the statistical techniques in this book by major research questions. A decision tree at the end of this chapter leads you to an appropriate analysis for your data. On the basis of your major research question and a few characteristics of your data set, you determine which statistical technique(s) is appropriate. The first and the most important criterion for choosing a technique is the major research question to be answered by the statistical analysis. Here, the research questions are categorized into degree of relationship among variables, significance of group differences, prediction of group membership, structure, and questions that focus on the time course of events. This chapter emphasizes differences in research questions answered by the different techniques described in nontechnical terms, whereas Chapter 18 provides an integrated overview of the techniques with some basic equations used in the multivariate general linear model.¹

2.1.1 Degree of Relationship Among Variables

If the major purpose of analysis is to assess the associations among two or more variables, some form of correlation/regression or chi-square is appropriate. The choice among five different statistical techniques is made by determining the number of independent and dependent variables, the nature of the variables (continuous or discrete), and whether any of the independent variables (IVs) are best conceptualized as covariates.²

¹ You may find it helpful to read Chapter 18 now instead of waiting for the end.

² If the effects of some IVs are assessed after the effects of other IVs are statistically removed, the latter are called *covariates*.