CONTENTS

1. INTRODUCTION	1
1.1 DATA COLLECTION	2
1.1.1 Data Base Limitations	3
1.1.2 Data-Conditioned Inferences	5
1.2 REGRESSION ANALYSIS	6
1.2.1 Linear Regression Models	6
1.2.2 Regression vs. Correlation	8
1.3 USES OF REGRESSION ANALYSIS	9
1.3.1 Prediction	9
1.3.2 Model Specification	10
1.3.3 Parameter Estimation	11
1.4 ABUSES OF REGRESSION ANALYSIS	12
1.4.1 Extrapolation	12
1.4.2 Generalization	15
1.4.3 Causation	17
2. INITIAL DATA EXPLORATION	19
2.1 PRELIMINARY DATA EDITING	23
2.1.1 Obvious Anomalies	23
2.1.2 Descriptive Statistics	26
2.1.3 Graphical Aids	28
2.2 INITIAL MODEL SPECIFICATION	33
2.2.1 Categorical Predictor Variables	33
2.2.2 Interactions	37
2.2.3 Smoothing	39

2.3 REEXPRESSING VARIABLES 42
2.3.1 Special Functions 42
2.3.2 Predictor Variable Transformations 45
2.3.3 Linearization of the Response 47
2.3.4 Prediction Versus Model Building 48
EXERCISES 50

3. SINGLE-VARIABLE LEAST SQUARES 52
3.1 APPROPRIATENESS 56
3.1.1 Theoretical Validity 57
3.1.2 Approximations in Restricted Regions 61
3.1.3 No-Intercept Models 63
3.2 LEAST SQUARES PARAMETER ESTIMATION 66
3.2.1 The Principle of Least Squares 66
3.2.2 Model Assumptions 71
3.2.3 Standardization 72
3.3 ADDITIONAL MEASURES OF FIT 77
3.3.1 Partitioning Variability 78
3.3.2 Coefficient of Determinatio 82
3.3.3 Residual Variability 83
APPENDIX 85
3.A Derivation of Least Squares Estimators 85
3.B Coefficient Estimates for Standardized Variables 87
3.C Partitioning the Total Sum of Squares 88
3.D Calculation of the Coefficient of Determination 89
EXERCISES 90

4. MULTIPLE-VARIABLE PRELIMINARIES 92
4.1 REVIEW OF MATRIX ALGEBRA 94
4.1.1 Notation 94
4.1.2 Vector and Matrix Operations 96
4.1.3 Model Definition 103
4.1.4 Latent Roots and Vectors 104
4.1.5 Rank of a Matrix 106
4.2 CARE IN MODEL BUILDING 108
4.2.1 Misspecification Bias 108
4.2.2 Overspecification Redundancy 110
4.3 STANDARDIZATION 111
4.3.1 Benefits 112
4.3.2 Correlation Form of $X'X$ 114

4.4 MULTICOLLINEARITY 115
4.4.1 Definition 115
4.4.2 Detection 118
4.4.3 Sample vs. Population Multicollinearities 120
APPENDIX 122
4.A Determinants and Inverse Matrices 122
4.B Determinants and Inverses Using Latent Roots and Latent Vectors 124
EXERCISES 126

5. MULTIPLE-VARIABLE LEAST SQUARES 128
5.1 PARAMETER ESTIMATION 131
5.1.1 Matrix Algebra Formulation 132
5.1.2 Fitting by Stages 135
5.1.3 Orthogonal Predictor Variables 140
5.1.4 Standardization 142
5.1.5 No-Intercept Models 148
5.2 INTERPRETATION OF FITTED MODELS 149
5.2.1 General Interpretation 150
5.2.2 Adjustment Computations 152
5.2.3 Special Case: Indicator Variables 153
5.3 INITIAL ASSESSMENT OF FIT 154
5.3.1 Analysis of Variance Table 154
5.3.2 Error Variance 158
APPENDIX 160
5.A Derivation of Least Squares Estimators 160
5.B Algebraic Derivation of Fitting by Stages 161
5.C Calculating SSR 164
5.D Equivalence of Residual Expressions 165
EXERCISES 166

6. INFERENCE 167
6.1 MODEL DEFINITION 169
6.1.1 Four Key Assumptions 169
6.1.2 Alternative Assumptions 172
6.2 ESTIMATOR PROPERTIES 174
6.2.1 Geometrical Representation 174
6.2.2 Expectation 178
6.2.3 Variances and Covariances 183
6.2.4 Probability Distributions 187

6.3 TESTS OF HYPOTHESIS 189
6.3.1 Tests on Individual Parameters 190
6.3.2 Analysis of Variance Tests 195
6.3.3 Repeated Predictor Variable Values (Lack of Fit Test) 198
6.4 INTERVAL ESTIMATION 202
6.4.1 Confidence Intervals and Regions 202
6.4.2 Response Intervals 206
APPENDIX 210
6.A Expectation of Least Squares Estimators 210
6.B Variances and Covariances of Least Squares Estimators 213
6.C Distribution of Least Squares Estimators 216
EXERCISES 218

7. RESIDUAL ANALYSIS 220
7.1 TYPES OF RESIDUALS 223
7.1.1 Distinction Between Residuals and Errors 224
7.1.2 Raw and Scaled Residuals 225
7.1.3 Deleted Residuals 229
7.2 VERIFICATION OF ERROR ASSUMPTIONS 231
7.2.1 Checks for Random Errors 232
7.2.2 Test for Serial Correlation 234
7.2.3 Detecting Heteroscedasticity 236
7.2.4 Normal Probability Plots 239
7.3 MODEL SPECIFICATION 241
7.3.1 Plots of Residuals Versus Predictor Variables 242
7.3.2 Partial Residual Plots 247
7.4 OUTLIER DETECTION 252
7.4.1 Plotting Techniques 252
7.4.2 Statistical Measures 255
APPENDIX 258
7.A Derivation of Deleted Residual 258
EXERCISES 259

8. VARIABLE SELECTION TECHNIQUES 262
8.1 BASIC CONSIDERATIONS 264
8.1.1 Problem Formulation 265
8.1.2 Additional Selection Criteria 267
8.2 SUBSET SELECTION METHODS 268
8.2.1 All Possible Regressions 269
8.2.2 Best Subset Regression 270
8.2.3 C_k Plots 272
8.2.4 Additional Search Routines 275

8.3 STEPWISE SELECTION METHODS 278
8.3.1 Forward Selection Method 278
8.3.2 Backward Elimination Method 282
8.3.3 Stepwise Procedure 284
APPENDIX 286
8.A Derivation of Equation (8.3.4) 286
EXERCISES 288

9. MULTICOLLINEARITY EFFECTS 290
9.1 COEFFICIENT ESTIMATORS 293
9.1.1 Variances and Covariances 293
9.1.2 Estimator Effects 299
9.2 INFERENCE PROCEDURES 302
9.2.1 *t* Statistics 302
9.2.2 Other Selection Criteria 304
9.2.3 Effects on Subset Selection Methods 306
9.3 POPULATION-INHERENT MULTICOLLINEARITIES 308
9.3.1 Variable Selection 308
9.3.2 Prediction 310
EXERCISES 313

10. BIASES REGRESSION ESTIMATORS 315
10.1 PRINCIPAL COMPONENT REGRESSION 317
10.1.1 Motivation 318
10.1.2 Analysis of Variance 322
10.1.3 Inference Techniques 326
10.2 LATENT ROOT REGRESSION ANALYSIS 329
10.2.1 Motivation 330
10.2.2 Predictive Multicollinearities 332
10.2.3 Analysis of Variance 335
10.2.4 Inference Techniques 338
10.3 RIDGE REGRESSION 340
10.3.1 Motivation 341
10.3.2 Selecting Ridge Parameters 343
10.3.3 Analysis of Variance 346
10.4 FINAL REMARKS 348
EXERCISES 350

APPENDIX A. DATA SETS ANALYZED IN THIS TEXT 352
A.1 Selected Demographic Characteristics of Countries of the World 358
A.2 Final High School Average Grades and First Year of College Average Grades 359

A.3 Homicides in Detroit, 1961-1973 360
A.4 Educational Status Attainment 361
A.5 Nitrous Oxide Emissions Modelling 362
A.6 Anthropometric and Physical Fitness Measurements on Police Department Applicants 363
A.7 Housing Rent Study 365
A.8 Body Measurements on Police Department Applicants 367
APPENDIX B. DATA SETS FOR FURTHER STUDY 368
B.1 Mortality and Pollution Study 370
B.2 Solid Waste Data 372
APPENDIX C. STATISTICAL TABLES 373
C.1 Cumulative Standard Normal Distribution 376
C.2 Cumulative Student *t* Distribution 377
C.3 Cumulative Chi-Square Distribution 378
C.4 Cumulative *F* Distribution 380
C.5 Critical Values for Runs Test 386
C.6 Critical Values for Durbin-Watson Test Statistic 388

BIBLIOGRAPHY 389

INDEX 398