

Contents

Preface to Third Edition	xxiii
Preface of Second Edition	xxvii
Acknowledgments	xxxix
Author	xxxix
1. Introduction	1
1.1 The Personal Computer and Statistics	1
1.2 Statistics and Data Analysis	3
1.3 EDA	4
1.4 The EDA Paradigm	5
1.5 EDA Weaknesses	6
1.6 Small and Big Data	7
1.6.1 Data Size Characteristics	7
1.6.2 Data Size: Personal Observation of One	8
1.7 Data Mining Paradigm	8
1.8 Statistics and Machine Learning	9
1.9 Statistical Data Mining	10
References	11
2. Science Dealing with Data: Statistics and Data Science	13
2.1 Introduction	13
2.2 Background	13
2.3 The Statistics and Data Science Comparison	15
2.3.1 Statistics versus Data Science	15
2.4 Discussion: Are Statistics and Data Science Different?	21
2.4.1 Analysis: Are Statistics and Data Science Different?	22
2.5 Summary	23
2.6 Epilogue	23
References	23
3. Two Basic Data Mining Methods for Variable Assessment	25
3.1 Introduction	25
3.2 Correlation Coefficient	25
3.3 Scatterplots	27
3.4 Data Mining	28
3.4.1 Example 3.1	28
3.4.2 Example 3.2	29
3.5 Smoothed Scatterplot	30
3.6 General Association Test	33
3.7 Summary	34
References	35

4. CHAID-Based Data Mining for Paired-Variable Assessment	37
4.1 Introduction.....	37
4.2 The Scatterplot.....	37
4.2.1 An Exemplar Scatterplot.....	38
4.3 The Smooth Scatterplot.....	38
4.4 Primer on CHAID.....	39
4.5 CHAID-Based Data Mining for a Smoother Scatterplot.....	40
4.5.1 The Smoother Scatterplot.....	42
4.6 Summary.....	45
Reference.....	45
5. The Importance of Straight Data Simplicity and Desirability for Good Model-Building Practice	47
5.1 Introduction.....	47
5.2 Straightness and Symmetry in Data.....	47
5.3 Data Mining Is a High Concept.....	48
5.4 The Correlation Coefficient.....	48
5.5 Scatterplot of (xx3, yy3).....	50
5.6 Data Mining the Relationship of (xx3, yy3).....	50
5.6.1 Side-by-Side Scatterplot.....	53
5.7 What Is the GP-Based Data Mining Doing to the Data?.....	53
5.8 Straightening a Handful of Variables and a Baker's Dozen of Variables.....	53
5.9 Summary.....	54
References.....	54
6. Symmetrizing Ranked Data: A Statistical Data Mining Method for Improving the Predictive Power of Data	55
6.1 Introduction.....	55
6.2 Scales of Measurement.....	55
6.3 Stem-and-Leaf Display.....	57
6.4 Box-and-Whiskers Plot.....	58
6.5 Illustration of the Symmetrizing Ranked Data Method.....	58
6.5.1 Illustration 1.....	59
6.5.1.1 Discussion of Illustration 1.....	59
6.5.2 Illustration 2.....	61
6.5.2.1 <i>Titanic</i> Dataset.....	62
6.5.2.2 Looking at the Recoded <i>Titanic</i> Ordinal Variables CLASS_, AGE_, GENDER_, CLASS_AGE_, and CLASS_GENDER_.....	62
6.5.2.3 Looking at the Symmetrized-Ranked <i>Titanic</i> Ordinal Variables rCLASS_, rAGE_, rGENDER_, rCLASS_AGE_, and rCLASS_GENDER_.....	64
6.5.2.4 Building a Preliminary <i>Titanic</i> Model.....	65
6.6 Summary.....	68
References.....	68
7. Principal Component Analysis: A Statistical Data Mining Method for Many-Variable Assessment	69
7.1 Introduction.....	69
7.2 EDA Reexpression Paradigm.....	69

7.3	What Is the Big Deal?.....	70
7.4	PCA Basics	70
7.5	Exemplary Detailed Illustration	71
7.5.1	Discussion.....	71
7.6	Algebraic Properties of PCA	72
7.7	Uncommon Illustration.....	73
7.7.1	PCA of R_CD Elements ($X_1, X_2, X_3, X_4, X_5, X_6$).....	74
7.7.2	Discussion of the PCA of R_CD Elements	74
7.8	PCA in the Construction of Quasi-Interaction Variables	76
7.8.1	SAS Program for the PCA of the Quasi-Interaction Variable	78
7.9	Summary	80
8.	Market Share Estimation: Data Mining for an Exceptional Case	81
8.1	Introduction	81
8.2	Background.....	81
8.3	Data Mining for an Exceptional Case	82
8.3.1	Exceptional Case: Infant Formula YUM.....	82
8.4	Building the RAL-YUM Market Share Model	83
8.4.1	Decile Analysis of YUM_3mos MARKET-SHARE Model	92
8.4.2	Conclusion of YUM_3mos MARKET-SHARE Model	92
8.5	Summary	93
	Appendix 8.A Dummify PROMO_Code.....	93
	Appendix 8.B PCA of PROMO_Code Dummy Variables.....	94
	Appendix 8.C Logistic Regression YUM_3mos on PROMO_Code Dummy Variables	94
	Appendix 8.D Creating YUM_3mos_wo_PROMO_CodeEff	94
	Appendix 8.E Normalizing a Variable to Lie Within [0, 1]	95
	References	96
9.	The Correlation Coefficient: Its Values Range between Plus and Minus 1, or Do They?.....	97
9.1	Introduction	97
9.2	Basics of the Correlation Coefficient	97
9.3	Calculation of the Correlation Coefficient.....	99
9.4	Rematching	99
9.5	Calculation of the Adjusted Correlation Coefficient.....	101
9.6	Implication of Rematching	102
9.7	Summary	102
10.	Logistic Regression: The Workhorse of Response Modeling	105
10.1	Introduction	105
10.2	Logistic Regression Model.....	106
10.2.1	Illustration.....	106
10.2.2	Scoring an LRM	107
10.3	Case Study.....	109
10.3.1	Candidate Predictor and Dependent Variables.....	110
10.4	Logits and Logit Plots.....	110
10.4.1	Logits for Case Study	111
10.5	The Importance of Straight Data	112

10.6	Reexpressing for Straight Data	112
10.6.1	Ladder of Powers	113
10.6.2	Bulging Rule	114
10.6.3	Measuring Straight Data	114
10.7	Straight Data for Case Study	115
10.7.1	Reexpressing FD2_OPEN	116
10.7.2	Reexpressing INVESTMENT	116
10.8	Techniques when the Bulging Rule Does Not Apply	118
10.8.1	Fitted Logit Plot.....	118
10.8.2	Smooth Predicted-versus-Actual Plot.....	119
10.9	Reexpressing MOS_OPEN.....	119
10.9.1	Plot of Smooth Predicted versus Actual for MOS_OPEN	120
10.10	Assessing the Importance of Variables.....	123
10.10.1	Computing the G Statistic	123
10.10.2	Importance of a Single Variable.....	124
10.10.3	Importance of a Subset of Variables.....	124
10.10.4	Comparing the Importance of Different Subsets of Variables	124
10.11	Important Variables for Case Study	125
10.11.1	Importance of the Predictor Variables.....	126
10.12	Relative Importance of the Variables	127
10.12.1	Selecting the Best Subset	127
10.13	Best Subset of Variables for Case Study	128
10.14	Visual Indicators of Goodness of Model Predictions	129
10.14.1	Plot of Smooth Residual by Score Groups.....	130
10.14.1.1	Plot of the Smooth Residual by Score Groups for Case Study	130
10.14.2	Plot of Smooth Actual versus Predicted by Decile Groups.....	132
10.14.2.1	Plot of Smooth Actual versus Predicted by Decile Groups for Case Study	132
10.14.3	Plot of Smooth Actual versus Predicted by Score Groups	134
10.14.3.1	Plot of Smooth Actual versus Predicted by Score Groups for Case Study	134
10.15	Evaluating the Data Mining Work	136
10.15.1	Comparison of Plots of Smooth Residual by Score Groups: EDA versus Non-EDA Models.....	137
10.15.2	Comparison of the Plots of Smooth Actual versus Predicted by Decile Groups: EDA versus Non-EDA Models.....	139
10.15.3	Comparison of Plots of Smooth Actual versus Predicted by Score Groups: EDA versus Non-EDA Models	140
10.15.4	Summary of the Data Mining Work	141
10.16	Smoothing a Categorical Variable	141
10.16.1	Smoothing FD_TYPE with CHAID.....	142
10.16.2	Importance of CH_FTY_1 and CH_FTY_2	144
10.17	Additional Data Mining Work for Case Study	145
10.17.1	Comparison of Plots of Smooth Residual by Score Group: 4var-EDA versus 3var-EDA Models	146
10.17.2	Comparison of the Plots of Smooth Actual versus Predicted by Decile Groups: 4var-EDA versus 3var-EDA Models	147

10.17.3	Comparison of Plots of Smooth Actual versus Predicted by Score Groups: 4var-EDA versus 3var-EDA Models	147
10.17.4	Final Summary of the Additional Data Mining Work	149
10.18	Summary	150
11	Predicting Share of Wallet without Survey Data	151
11.1	Introduction	151
11.2	Background	151
11.2.1	SOW Definition	152
11.2.1.1	SOW _q Definition	152
11.2.1.2	SOW _q Likelihood Assumption	152
11.3	Illustration of Calculation of SOW _q	153
11.3.1	Query of Interest	153
11.3.2	DOLLARS and TOTAL DOLLARS	153
11.4	Building the AMPECS SOW _q Model	158
11.5	SOW _q Model Definition	159
11.5.1	SOW _q Model Results	160
11.6	Summary	161
	Appendix 11.A Six Steps	162
	Appendix 11.B Seven Steps	164
	References	167
12	Ordinary Regression: The Workhorse of Profit Modeling	169
12.1	Introduction	169
12.2	Ordinary Regression Model	169
12.2.1	Illustration	170
12.2.2	Scoring an OLS Profit Model	171
12.3	Mini Case Study	172
12.3.1	Straight Data for Mini Case Study	172
12.3.1.1	Reexpressing INCOME	174
12.3.1.2	Reexpressing AGE	175
12.3.2	Plot of Smooth Predicted versus Actual	177
12.3.3	Assessing the Importance of Variables	178
12.3.3.1	Defining the F Statistic and R-Squared	179
12.3.3.2	Importance of a Single Variable	179
12.3.3.3	Importance of a Subset of Variables	179
12.3.3.4	Comparing the Importance of Different Subsets of Variables	180
12.4	Important Variables for Mini Case Study	180
12.4.1	Relative Importance of the Variables	181
12.4.2	Selecting the Best Subset	181
12.5	Best Subset of Variables for Case Study	182
12.5.1	PROFIT Model with gINCOME and AGE	183
12.5.2	Best PROFIT Model	185
12.6	Suppressor Variable AGE	185
12.7	Summary	186
	References	187

13. Variable Selection Methods in Regression: Ignorable Problem, Notable Solution	189
13.1 Introduction	189
13.2 Background	189
13.3 Frequently Used Variable Selection Methods	192
13.4 Weakness in the Stepwise	193
13.5 Enhanced Variable Selection Method	194
13.6 Exploratory Data Analysis	196
13.7 Summary	200
References	200
14. CHAID for Interpreting a Logistic Regression Model	203
14.1 Introduction	203
14.2 Logistic Regression Model	203
14.3 Database Marketing Response Model Case Study	204
14.3.1 Odds Ratio	205
14.4 CHAID	205
14.4.1 Proposed CHAID-Based Method	206
14.5 Multivariable CHAID Trees	208
14.6 CHAID Market Segmentation	210
14.7 CHAID Tree Graphs	213
14.8 Summary	216
15. The Importance of the Regression Coefficient	219
15.1 Introduction	219
15.2 The Ordinary Regression Model	219
15.3 Four Questions	220
15.4 Important Predictor Variables	220
15.5 p-Values and Big Data	221
15.6 Returning to Question 1	222
15.7 Effect of Predictor Variable on Prediction	222
15.8 The Caveat	223
15.9 Returning to Question 2	225
15.10 Ranking Predictor Variables by Effect on Prediction	225
15.11 Returning to Question 3	226
15.12 Returning to Question 4	227
15.13 Summary	227
References	228
16. The Average Correlation: A Statistical Data Mining Measure for Assessment of Competing Predictive Models and the Importance of the Predictor Variables	229
16.1 Introduction	229
16.2 Background	229
16.3 Illustration of the <i>Difference</i> between Reliability and Validity	231
16.4 Illustration of the <i>Relationship</i> between Reliability and Validity	231
16.5 The Average Correlation	232
16.5.1 Illustration of the Average Correlation with an LTV5 Model	232

16.5.2	Continuing with the Illustration of the Average Correlation with an LTV5 Model.....	236
16.5.3	Continuing with the Illustration with a Competing LTV5 Model	236
16.5.3.1	The Importance of the Predictor Variables.....	237
16.6	Summary.....	237
	Reference.....	237
17.	CHAID for Specifying a Model with Interaction Variables	239
17.1	Introduction.....	239
17.2	Interaction Variables.....	239
17.3	Strategy for Modeling with Interaction Variables.....	240
17.4	Strategy Based on the Notion of a Special Point	240
17.5	Example of a Response Model with an Interaction Variable.....	241
17.6	CHAID for Uncovering Relationships.....	242
17.7	Illustration of CHAID for Specifying a Model.....	243
17.8	An Exploratory Look.....	246
17.9	Database Implication.....	247
17.10	Summary.....	248
	References	249
18.	Market Segmentation Classification Modeling with Logistic Regression.....	251
18.1	Introduction.....	251
18.2	Binary Logistic Regression.....	251
18.2.1	Necessary Notation	252
18.3	Polychotomous Logistic Regression Model.....	252
18.4	Model Building with PLR.....	253
18.5	Market Segmentation Classification Model	254
18.5.1	Survey of Cellular Phone Users.....	254
18.5.2	CHAID Analysis.....	255
18.5.3	CHAID Tree Graphs.....	258
18.5.4	Market Segmentation Classification Model.....	261
18.6	Summary.....	263
19.	Market Segmentation Based on Time-Series Data Using Latent Class Analysis	265
19.1	Introduction.....	265
19.2	Background.....	265
19.2.1	K-Means Clustering.....	265
19.2.2	PCA.....	266
19.2.3	FA.....	266
19.2.3.1	FA Model.....	267
19.2.3.2	FA Model Estimation.....	267
19.2.3.3	FA versus OLS Graphical Depiction.....	268
19.2.4	LCA versus FA Graphical Depiction.....	268
19.3	LCA.....	270
19.3.1	LCA of Universal and Particular Study.....	270
19.3.1.1	Discussion of LCA Output.....	270
19.3.1.2	Discussion of Posterior Probability	271
19.4	LCA versus k-Means Clustering.....	272

19.5	LCA Market Segmentation Model Based on Time-Series Data	274
19.5.1	Objective.....	274
19.5.2	Best LCA Models	276
19.5.2.1	Cluster Sizes and Conditional Probabilities/Mean.....	278
19.5.2.2	Indicator-Level Posterior Probabilities.....	281
19.6	Summary.....	282
Appendix 19.A	Creating Trend ³ for UNITS.....	282
Appendix 19.B	POS-ZER-NEG Creating Trend ⁴	284
References	285
20.	Market Segmentation: An Easy Way to Understand the Segments	287
20.1	Introduction	287
20.2	Background.....	287
20.3	Illustration.....	288
20.4	Understanding the Segments	289
20.5	Summary	290
Appendix 20.A	Dataset SAMPLE	290
Appendix 20.B	Segmentor-Means	291
Appendix 20.C	Indexed Profiles.....	291
References	292
21.	The Statistical Regression Model: An Easy Way to Understand the Model	293
21.1	Introduction.....	293
21.2	Background.....	293
21.3	EZ-Method Applied to the LR Model	294
21.4	Discussion of the LR EZ-Method Illustration.....	296
21.5	Summary.....	299
Appendix 21.A	M65-Spread Base Means X10–X14.....	299
Appendix 21.B	Create Ten Datasets for Each Decile.....	301
Appendix 21.C	Indexed Profiles of Deciles.....	302
22.	CHAID as a Method for Filling in Missing Values	307
22.1	Introduction	307
22.2	Introduction to the Problem of Missing Data	307
22.3	Missing Data Assumption.....	309
22.4	CHAID Imputation.....	310
22.5	Illustration.....	311
22.5.1	CHAID Mean-Value Imputation for a Continuous Variable	312
22.5.2	Many Mean-Value CHAID Imputations for a Continuous Variable ...	313
22.5.3	Regression Tree Imputation for LIFE_DOL	314
22.6	CHAID Most Likely Category Imputation for a Categorical Variable	316
22.6.1	CHAID Most Likely Category Imputation for GENDER.....	316
22.6.2	Classification Tree Imputation for GENDER	318
22.7	Summary.....	320
References	321
23.	Model Building with Big Complete and Incomplete Data	323
23.1	Introduction	323
23.2	Background.....	323

23.3	The CCA-PCA Method: Illustration Details	324
23.3.1	Determining the Complete and Incomplete Datasets	324
23.4	Building the RESPONSE Model with Complete (CCA) Dataset	326
23.4.1	CCA RESPONSE Model Results	327
23.5	Building the RESPONSE Model with Incomplete (ICA) Dataset	328
23.5.1	PCA on BICA Data	329
23.6	Building the RESPONSE Model on PCA-BICA Data	329
23.6.1	PCA-BICA RESPONSE Model Results	330
23.6.2	Combined CCA and PCA-BICA RESPONSE Model Results	331
23.7	Summary	332
	Appendix 23.A NMISS	333
	Appendix 23.B Testing CCA Samsizes	333
	Appendix 23.C CCA-CIA Datasets	333
	Appendix 23.D Ones and Zeros	333
	Reference	334
24.	Art, Science, Numbers, and Poetry	335
24.1	Introduction	335
24.2	Zeros and Ones	336
24.3	Power of Thought	336
24.4	The Statistical Golden Rule: Measuring the Art and Science of Statistical Practice	338
24.4.1	Background	338
24.4.1.1	The Statistical Golden Rule	339
24.5	Summary	340
	Reference	340
25.	Identifying Your Best Customers: Descriptive, Predictive, and Look-Alike Profiling	341
25.1	Introduction	341
25.2	Some Definitions	341
25.3	Illustration of a Flawed Targeting Effort	342
25.4	Well-Defined Targeting Effort	343
25.5	Predictive Profiles	345
25.6	Continuous Trees	348
25.7	Look-Alike Profiling	350
25.8	Look-Alike Tree Characteristics	353
25.9	Summary	353
26.	Assessment of Marketing Models	355
26.1	Introduction	355
26.2	Accuracy for Response Model	355
26.3	Accuracy for Profit Model	356
26.4	Decile Analysis and Cum Lift for Response Model	358
26.5	Decile Analysis and Cum Lift for Profit Model	359
26.6	Precision for Response Model	360
26.7	Precision for Profit Model	362
26.7.1	Construction of SWMAD	363
26.8	Separability for Response and Profit Models	363

26.9	Guidelines for Using Cum Lift, HL/SWMAD, and CV.....	364
26.10	Summary.....	364
27.	Decile Analysis: Perspective and Performance.....	367
27.1	Introduction.....	367
27.2	Background.....	367
27.2.1	Illustration.....	369
27.2.1.1	Discussion of Classification Table of RESPONSE Model.....	370
27.3	Assessing Performance: RESPONSE Model versus Chance Model.....	371
27.4	Assessing Performance: The Decile Analysis.....	372
27.4.1	The RESPONSE Decile Analysis.....	372
27.5	Summary.....	377
Appendix 27.A	Incremental Gain in Accuracy: Model versus Chance.....	378
Appendix 27.B	Incremental Gain in Precision: Model versus Chance.....	379
Appendix 27.C	RESPONSE Model Decile PROB_est Values.....	380
Appendix 27.D	2 × 2 Tables by Decile.....	382
	References.....	385
28.	Net T-C Lift Model: Assessing the Net Effects of Test and Control Campaigns.....	387
28.1	Introduction.....	387
28.2	Background.....	387
28.3	Building TEST and CONTROL Response Models.....	389
28.3.1	Building TEST Response Model.....	390
28.3.2	Building CONTROL Response Model.....	392
28.4	Net T-C Lift Model.....	394
28.4.1	Building the Net T-C Lift Model.....	395
28.4.1.1	Discussion of the Net T-C Lift Model.....	395
28.4.1.2	Discussion of Equal-Group Sizes Decile of the Net T-C Lift Model.....	397
28.5	Summary.....	398
Appendix 28.A	TEST Logistic with Xs.....	400
Appendix 28.B	CONTROL Logistic with Xs.....	402
Appendix 28.C	Merge Score.....	405
Appendix 28.D	NET T-C Decile Analysis.....	406
	References.....	410
29.	Bootstrapping in Marketing: A New Approach for Validating Models.....	413
29.1	Introduction.....	413
29.2	Traditional Model Validation.....	413
29.3	Illustration.....	414
29.4	Three Questions.....	415
29.5	The Bootstrap Method.....	416
29.5.1	Traditional Construction of Confidence Intervals.....	416
29.6	How to Bootstrap.....	417
29.6.1	Simple Illustration.....	418
29.7	Bootstrap Decile Analysis Validation.....	419
29.8	Another Question.....	420

29.9	Bootstrap Assessment of Model Implementation Performance.....	421
29.9.1	Illustration.....	424
29.10	Bootstrap Assessment of Model Efficiency	426
29.11	Summary	428
	References	428
30.	Validating the Logistic Regression Model: Try Bootstrapping	429
30.1	Introduction	429
30.2	Logistic Regression Model.....	429
30.3	The Bootstrap Validation Method	429
30.4	Summary	430
	Reference	430
31.	Visualization of Marketing Models: Data Mining to Uncover Innards of a Model.....	431
31.1	Introduction	431
31.2	Brief History of the Graph	431
31.3	Star Graph Basics.....	432
31.3.1	Illustration.....	433
31.4	Star Graphs for Single Variables	434
31.5	Star Graphs for Many Variables Considered Jointly	435
31.6	Profile Curves Method	437
31.6.1	Profile Curves Basics	437
31.6.2	Profile Analysis	438
31.7	Illustration.....	438
31.7.1	Profile Curves for RESPONSE Model	440
31.7.2	Decile Group Profile Curves	442
31.8	Summary	444
	Appendix 31.A Star Graphs for Each Demographic Variable about the Deciles.....	445
	Appendix 31.B Star Graphs for Each Decile about the Demographic Variables.....	447
	Appendix 31.C Profile Curves: All Deciles	450
	References	452
32.	The Predictive Contribution Coefficient: A Measure of Predictive Importance.....	453
32.1	Introduction	453
32.2	Background.....	453
32.3	Illustration of Decision Rule.....	455
32.4	Predictive Contribution Coefficient.....	457
32.5	Calculation of Predictive Contribution Coefficient.....	458
32.6	Extra-Illustration of Predictive Contribution Coefficient.....	459
32.7	Summary	462
	Reference	463
33.	Regression Modeling Involves Art, Science, and Poetry, Too.....	465
33.1	Introduction	465
33.2	Shakespearean Modelogue.....	465
33.3	Interpretation of the Shakespearean Modelogue.....	466
33.4	Summary	469
	References	469

34. Opening the Dataset: A Twelve-Step Program for Dataholics	471
34.1 Introduction	471
34.2 Background	471
34.3 Stepping	471
34.4 Brush Marking	473
34.5 Summary	474
Appendix 34.A Dataset IN	474
Appendix 34.B Samsize Plus	475
Appendix 34.C Copy-Pasteable	475
Appendix 34.D Missings	475
References	476
35. Genetic and Statistic Regression Models: A Comparison	477
35.1 Introduction	477
35.2 Background	477
35.3 Objective	478
35.4 The GenIQ Model, the Genetic Logistic Regression	478
35.4.1 Illustration of "Filling Up the Upper Deciles"	479
35.5 A Pithy Summary of the Development of Genetic Programming	480
35.6 The GenIQ Model: A Brief Review of Its Objective and Salient Features	482
35.6.1 The GenIQ Model Requires Selection of Variables and Function: An Extra Burden?	482
35.7 The GenIQ Model: How It Works	483
35.7.1 The GenIQ Model Maximizes the Decile Table	485
35.8 Summary	486
References	486
36. Data Reuse: A Powerful Data Mining Effect of the GenIQ Model	487
36.1 Introduction	487
36.2 Data Reuse	487
36.3 Illustration of Data Reuse	488
36.3.1 The GenIQ Profit Model	488
36.3.2 Data-Reused Variables	489
36.3.3 Data-Reused Variables GenIQvar_1 and GenIQvar_2	490
36.4 Modified Data Reuse: A GenIQ-Enhanced Regression Model	491
36.4.1 Illustration of a GenIQ-Enhanced LRM	491
36.5 Summary	493
37. A Data Mining Method for Moderating Outliers Instead of Discarding Them	495
37.1 Introduction	495
37.2 Background	495
37.3 Moderating Outliers Instead of Discarding Them	496
37.3.1 Illustration of Moderating Outliers Instead of Discarding Them	496
37.3.2 The GenIQ Model for Moderating the Outlier	498
37.4 Summary	499
Reference	499

38. Overfitting: Old Problem, New Solution	501
38.1 Introduction	501
38.2 Background.....	501
38.2.1 Idiomatic Definition of Overfitting to Help Remember the Concept	502
38.3 The GenIQ Model Solution to Overfitting.....	503
38.3.1 RANDOM_SPLIT GenIQ Model	505
38.3.2 RANDOM_SPLIT GenIQ Model Decile Analysis	505
38.3.3 Quasi N-tile Analysis	507
38.4 Summary.....	508
39. The Importance of Straight Data: Revisited	509
39.1 Introduction	509
39.2 Restatement of Why It Is Important to Straighten Data	509
39.3 Restatement of Section 12.3.1.1 “Reexpressing INCOME”	510
39.3.1 Complete Exposition of Reexpressing INCOME.....	510
39.3.1.1 The GenIQ Model Detail of the gINCOME Structure	511
39.4 Restatement of Section 5.6 “Data Mining the Relationship of (xx3, yy3)”	511
39.4.1 The GenIQ Model Detail of the GenIQvar(yy3) Structure	511
39.5 Summary	512
40. The GenIQ Model: Its Definition and an Application	513
40.1 Introduction	513
40.2 What Is Optimization?	513
40.3 What Is Genetic Modeling?	514
40.4 Genetic Modeling: An Illustration.....	515
40.4.1 Reproduction	517
40.4.2 Crossover.....	518
40.4.3 Mutation	518
40.5 Parameters for Controlling a Genetic Model Run.....	519
40.6 Genetic Modeling: Strengths and Limitations	519
40.7 Goals of Marketing Modeling.....	520
40.8 The GenIQ Response Model.....	520
40.9 The GenIQ Profit Model.....	521
40.10 Case Study: Response Model	522
40.11 Case Study: Profit Model.....	524
40.12 Summary	527
Reference	527
41. Finding the Best Variables for Marketing Models	529
41.1 Introduction	529
41.2 Background.....	529
41.3 Weakness in the Variable Selection Methods	531
41.4 Goals of Modeling in Marketing	532
41.5 Variable Selection with GenIQ.....	533
41.5.1 GenIQ Modeling	535
41.5.2 GenIQ Structure Identification	537
41.5.3 GenIQ Variable Selection.....	539

41.6	Nonlinear Alternative to Logistic Regression Model	542
41.7	Summary	545
	References	546
42.	Interpretation of Coefficient-Free Models	547
42.1	Introduction	547
42.2	The Linear Regression Coefficient.....	547
42.2.1	Illustration for the Simple Ordinary Regression Model	548
42.2.2	Illustration for the Simple Logistic Regression Model.....	548
42.3	The Quasi-Regression Coefficient for Simple Regression Models.....	549
42.3.1	Illustration of Quasi-RC for the Simple Ordinary Regression Model	549
42.3.2	Illustration of Quasi-RC for the Simple Logistic Regression Model	550
42.3.3	Illustration of Quasi-RC for Nonlinear Predictions.....	551
42.4	Partial Quasi-RC for the Everymodel	553
42.4.1	Calculating the Partial Quasi-RC for the Everymodel	554
42.4.2	Illustration for the Multiple Logistic Regression Model.....	555
42.5	Quasi-RC for a Coefficient-Free Model	560
42.5.1	Illustration of Quasi-RC for a Coefficient-Free Model.....	560
42.6	Summary	567
43.	Text Mining: Primer, Illustration, and TXTDM Software	569
43.1	Introduction	569
43.2	Background.....	569
43.2.1	Text Mining Software: Free versus Commercial versus TXTDM.....	570
43.3	Primer of Text Mining	571
43.4	Statistics of the Words	573
43.5	The Binary Dataset of Words in Documents	574
43.6	Illustration of TXTDM Text Mining	575
43.7	Analysis of the Text-Mined GenIQ_FAVORED Model.....	584
43.7.1	Text-Based Profiling of Respondents Who Prefer GenIQ	584
43.7.2	Text-Based Profiling of Respondents Who Prefer OLS-Logistic.....	585
43.8	Weighted TXTDM.....	585
43.9	Clustering Documents	586
43.9.1	Clustering GenIQ Survey Documents.....	586
43.9.1.1	Conclusion of Clustering GenIQ Survey Documents.....	592
43.10	Summary	593
	Appendix.....	593
	Appendix 43.A Loading Corpus TEXT Dataset	594
	Appendix 43.B Intermediate Step Creating Binary Words	594
	Appendix 43.C Creating the Final Binary Words	595
	Appendix 43.D Calculate Statistics TF, DF, NUM_DOCS, and N (=Num of Words).....	596
	Appendix 43.E Append GenIQ_FAVORED to WORDS Dataset	597
	Appendix 43.F Logistic GenIQ_FAVORED Model.....	598
	Appendix 43.G Average Correlation among Words	599
	Appendix 43.H Creating TF-IDF.....	600
	Appendix 43.I WORD_TF-IDF Weights by Concat of WORDS and TF-IDF.....	602
	Appendix 43.J WORD_RESP WORD_TF-IDF RESP	604

Appendix 43.K Stemming 604

Appendix 43.L WORD Times TF-IDF 604

Appendix 43.M Dataset Weighted with Words for Profile 605

Appendix 43.N VARCLUS for Two-Class Solution 606

Appendix 43.O Scoring VARCLUS for Two-Cluster Solution 606

Appendix 43.P Direction of Words with Its Cluster 1..... 607

Appendix 43.Q Performance of GenIQ Model versus Chance Model 609

Appendix 43.R Performance of Liberal-Cluster Model versus Chance Model 609

References 610

44. Some of My Favorite Statistical Subroutines..... 611

44.1 List of Subroutines 611

44.2 Smoothplots (Mean and Median) of Chapter 5—X1 versus X2..... 611

44.3 Smoothplots of Chapter 10—Logit and Probability 615

44.4 Average Correlation of Chapter 16—Among Var1 Var2 Var3..... 618

44.5 Bootstrapped Decile Analysis of Chapter 29—Using Data from Table 23.4 620

44.6 H-Spread Common Region of Chapter 42..... 627

44.7 Favorite—Proc Corr with Option Rank, Vertical Output 630

44.8 Favorite—Decile Analysis—Response 631

44.9 Favorite—Decile Analysis—Profit..... 635

44.10 Favorite—Smoothing Time-Series Data (Running Medians of Three)..... 638

44.11 Favorite—First Cut Is the Deepest—Among Variables with Large
Skew Values 643

Index..... 645

Chapter 45, *Some of My Favorite Statistical Subroutines*, contains 11 chapters that are inserted between the chapters on the statistical software. The chapters are arranged in the order of continuity of material. I will not discuss the details of the chapters here. Chapter 45.1, *List of Subroutines*, follows Chapter 1 (Introduction). Chapter 45.2, *Smoothplots (Mean and Median) of Chapter 5—X1 versus X2*, is entitled *Smoothplots (Mean and Median) of Chapter 5—X1 versus X2*. If one were not looking, then it would seem that the book is about the delete key on statistics and statistics and replaced them by mean and median. I investigate whether the recently coined term data science is a more developed and expanded domain or if data science is just a new name for the current state of statistics.

Chapter 45.3, *Smoothplots of Chapter 10—Logit and Probability*, follows the chapter on logistic regression. Chapter 45.4, *Average Correlation of Chapter 16—Among Var1 Var2 Var3*, follows the chapter on correlation analysis (PCA). In this chapter, a market share estimation model is used to fit the usual survey-based market share scenario, but the PCA is used to estimate market share for a real exceptional case study. Chapter 45.5, *Bootstrapped Decile Analysis of Chapter 29—Using Data from Table 23.4*, follows the chapter on bootstrapping used in building the market share model for the exceptional case study.

Chapter 45.6, *H-Spread Common Region of Chapter 42*, follows the chapter on logistic regression. Chapter 45.7, *Favorite—Proc Corr with Option Rank, Vertical Output*, is with survey data. The chapter is arranged with reluctance because survey work is time-consuming, and I do not have enough data. I provide a two-step method for predicting SOW (Share of Wallet) using PCA and using simulation for estimating total dollars spent. Chapter 45.8, *Favorite—Decile Analysis—Response*, uses fractional logistic regression to predict SOW. Fractional logistic regression is a generalized linear model (GLM) dependent variables that