Statistical and Machine-Learning Data Mining

Techniques for Better Predictive Modeling and Analysis of Big Data

Third Edition

Statistical and Machine-Learning Data Mining

Techniques for Better Predictive Modeling and Analysis of Big Data

Third Edition

Bruce Ratner

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper

International Standard Book Number-13: 978-1-4987-9760-3 (Hardback)

Library of Congress Cataloging-in-Publication Data

Names: Ratner, Bruce, author.
Title: Statistical and machine-learning data mining / Bruce Ratner.
Description: Third Edition. | Boca Raton, FL : CRC Press, 2017. | Revised edition of the author's Statistical and machine-learning data mining, c2003.
Identifiers: LCCN 2016048787 | ISBN 9781498797603 (978-1-4987-9760-3)
Subjects: LCSH: Database marketing--Statistical methods. | Data mining--Statistical methods. | Big data--Statistical methods.
Classification: LCC HF5415.126 .R38 2017 | DDC 658.8/72--dc23 LC record available at https://lccn.loc.gov/2016048787

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

This book is dedicated to:

My father, Isaac—always encouraging, my role model who taught
me by doing, not saying, other than to think positive.

My mother, Leah—always nurturing, my friend who taught me to love love and hate hate.

My daughter, Amanda—always for me, my crowning and most significant result.

Contents

Preface to Third Edition ... xxiii
Preface of Second Edition ... xxvii
Acknowledgments ... xxxi
Author ... xxxiii

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 *Titanic* Dataset 62
6.5.2.2 Looking at the Recoded *Titanic* Ordinal Variables CLASS_, AGE_, GENDER_, CLASS_AGE_, and CLASS_GENDER_ 62
6.5.2.3 Looking at the Symmetrized-Ranked *Titanic* Ordinal Variables rCLASS_, rAGE_, rGENDER_, rCLASS_AGE_, and rCLASS_GENDER_ 64
6.5.2.4 Building a Preliminary *Titanic* 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 Likeliness 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 *Difference* between Reliability and Validity 231
16.4 Illustration of the *Relationship* 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/Means 278
19.5.2.2 Indicator-Level Posterior Probabilities 281
19.6 Summary 282
Appendix 19.A Creating Trend3 for UNITS 282
Appendix 19.B POS-ZER-NEG Creating Trend4 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

Preface to Third Edition

Predictive analytics of big data has maintained a steady presence in the four years since the publication of the second edition. My decision to write this third edition is not a result of the success (units) of the second edition but is due to the countless positive feedback (personal correspondence from the readership) I have received. And, importantly, I have the need to share my work on problems that do not have widely accepted, reliable, or known solutions. As in the previous editions, John Tukey's tenets, necessary to advance statistics, flexibility, practicality, innovation, and universality, are the touchstones of each chapter's new analytic and modeling methodology.

My main objectives in preparing the third edition are to:

1. Extend the content of the core material by including strategies and methods for problems, which I have observed on the top of *statistics on the table* [1] by reviewing predictive analytics conference proceedings and statistical modeling workshop outlines.
2. Reedit current chapters for improved writing and tighter endings.
3. Provide the statistical subroutines used in the proposed methods of analysis and modeling. I use Base SAS© and STAT/SAS. The subroutines are also available for downloading from my website: http://www.geniq.net/articles.html#section9. The code is easy to convert for users who prefer other languages.

I have added 13 new chapters that are inserted between the chapters of the second edition to yield the greatest flow of continuity of material. I outline the new chapters briefly here.

The first new chapter, Chapter 2, follows Chapter 1 (Introduction). The chapter is entitled *Science Dealing with Data: Statistics and Data Science*. If one were not looking, then it would appear that someone hit the delete key on statistics and statisticians and replaced them by science and data scientists. I investigate whether the recently minted term data science implies statistics is a subset of a more developed and expanded domain or if data science a buzzed-up cloaking of the current state of statistics.

Chapter 8, *Market Share Estimation: Data Mining for an Exceptional Case*, follows the Chapter 7 about principal component analysis (PCA). In this chapter, a market share estimation model, unique in that it does not fit the usual survey-based market share scenario, uses PCA as the foundation for estimating market share for a real exceptional case study. I provide the SAS subroutines used in building the market share model for the exceptional case study.

Chapter 11, *Predicting Share of Wallet without Survey Data*, follows the chapter on logistic regression. The everyday approach for predicting share of wallet (SOW) is with survey data. This approach is always met with reluctance because survey work is time-consuming, expensive, and yields unreliable data. I provide a two-step method for predicting SOW without data, by defining a quasi-SOW and using simulation for estimating total dollars spent. The second step uses fractional logistic regression to predict SOW_q. Fractional response regression cleverly uses ordinary logistic regression for dependent variables that assume proportions or rates. I present a case study in detail as well as provide SAS subroutines that readers should find valuable for their toolkits.

Chapter 19, *Market Segmentation Based on Time-Series Data Using Latent Class Analysis,* follows the chapter on market segmentation via logistic regression. In this chapter, I propose the model-based clustering method of latent class analysis (LCA). The innovative strategy of this segmentation is in my use of time-series data. The times-series LCA model is radically distinctive and not equal, and it will prove to be a template for treating times-series data in cross-sectional datasets and the application of LCA instead of the popular data-based heuristic k-means. I provide SAS subroutines so that data miners can perform similar segmentations as presented, along with a unique way of incorporating time-series data in an otherwise cross-sectional dataset.

Chapter 20, *Market Segmentation: An Easy Way to Understand the Segments,* is well-placed after the LCA-based market segmentation. The literature is replete with clustering methodologies, of which any one can serve for conducting a market segmentation. In contrast, the literature is virtually sparse in the area of how to interpret the segmentation results. This chapter provides an easy way to understand the discovered customer segments. I illustrate the new method with an admittedly simple example that will not belie the power of the approach. I provide SAS subroutines for conducting the proposed technique so data miners can add this worthy statistical technique to their toolkits.

Chapter 21, *The Statistical Regression Model: An Easy Way to Understand the Model*, is an extension of the method of understanding a market segmentation presented in Chapter 20. Its purpose is to provide an easy way to understand the statistical regression model, that is, ordinary least squares and logistic regression (LR) models. I illustrate the proposed method with an LR model. The illustration brings out the power of the method, in that it imparts supplementary information, making up for a deficiency in the ever-relied-upon regression coefficient for understanding a statistical regression model. I provide the SAS subroutines, which serve as a valued addition to any bag of statistical methods.

Chapter 23, *Model Building with Big Complete and Incomplete Data*, follows the chapter that uses CHAD as a method for imputation. This chapter overhears missing data warning the statistician, "You can't win unless you learn how to accept me." Traditional data-based methods (complete case analysis), predating big data, are known to be problematic with virtually all datasets. These methods now open a greater concern as to their unknown ineffectiveness on big data. I propose a two-stage approach, in which modeling of response on complete-case data precedes modeling of response on incomplete-case data via PCA. The two models can be used separately or combined, depending on the goals of the task. I provide SAS subroutines for the proposed method, which should become a utile technique for the statistical model builder.

Chapter 24, *Art, Science, Numbers, and Poetry,* is a high-order blend of artwork, science, numbers, and poetry, all inspired by the Egyptian pyramids, da Vinci, and Einstein. Love it or hate it, this chapter makes you think.

Chapter 27, *Decile Analysis: Perspective and Performance*, complements the preceding chapter on assessment of marketing models. Marketers use decile analysis to assess predictive incremental gains of their response models over responses obtained by chance. I define two new metrics, response model decile-analysis precision and chance model decile precision, which allow marketers to make a more insightful assessment as to the incremental gain of a response model over the chance model. I provide the SAS subroutines for constructing the two new metrics and the proposed procedure, which will be a trusty tool for marketing statisticians.

Chapter 28, *Net T-C Lift Model: Assessing the Net Effects of Test and Control Campaigns*, extends the practice of assessing response models to the proper use of a control group (found in the literature under names such as the uplift or net lift model) instead of the chance model as discussed in Chapter 27. There is large literature, albeit confusing and conflicting, on the methodologies of net lift modeling. I propose another approach, the *net T-C lift* model, to moderate the incompatible literature on this topic by offering a simple, straightforward, reliable model that is easy to implement and understand. I provide the SAS subroutines for the Net T-C Lift Model to enable statisticians to conduct net lift modeling without purchasing proprietary software.

Chapter 34, *Opening the Dataset: A Twelve-Step Program for Dataholics*, has valuable content for statisticians as they embark on the first step of any journey with data. Set in prose, I provide a light reading on the expectant steps of what to do when cracking open the dataset. Enjoy. I provide SAS subroutines of the twelve-step program in case the reader wants to take a nip.

Chapter 43, *Text Mining: Primer, Illustration, and TXTDM Software*, has three objectives: First, to serve as a primer, readable, brief though detailed, about what text mining encompasses, and how to conduct basic text mining; second, to illustrate text mining with a small body of text, yet interesting in its content; and third, to make text mining available to interested readers, by providing my SAS subroutines, named *TXTDM*.

Chapter 44, *Some of My Favorite Statistical Subroutines*, includes specific subroutines referenced throughout the book and generic subroutines for some second-edition chapters for which I no longer have the data. Lastly, I provide some of my favorite statistical subroutines, helpful in almost all analyses.

If there are any corrections post-production of the text, I will post them to the errata link http://www.geniq.net/articles.html#section9.

Reference

1. Stigler, S. M., *Statistics on the Table*, Harvard University Press, Cambridge, MA, 2002.

Preface of Second Edition

This book is unique. It is the only book, to date, that distinguishes between statistical data mining and machine-learning data mining. I was an orthodox statistician until I resolved my struggles with the weaknesses of statistics within the big data setting of today. Now, as a reform statistician who is free of the statistical rigors of yesterday, with many degrees of freedom to exercise, I have composed by intellectual might the original and practical statistical data mining techniques in the first part of the book. The GenIQ Model, a machine-learning alternative to statistical regression, led to the creative and useful machine-learning data mining techniques in the remaining part of the book.

This book is a compilation of essays that offer detailed background, discussion, and illustration of specific methods for solving the most commonly experienced problems in predictive modeling and analysis of big data. The common theme among these essays is to address each methodology and assign its application to a specific type of problem. To better ground the reader, I spend considerable time discussing the basic methodologies of predictive modeling and analysis. While this type of overview has been attempted before, my approach offers a truly nitty-gritty, step-by-step approach that both tyros and experts in the field can enjoy playing with. The job of the data analyst is overwhelmingly to predict and explain the result of the target variable, such as RESPONSE or PROFIT. Within that task, the target variable is either a binary variable (RESPONSE is one such example) or a continuous variable (of which PROFIT is a good example). The scope of this book is purposely limited, with one exception, to dependency models, for which the target variable is often referred to as the "left-hand" side of an equation, and the variables that predict and/or explain the target variable is the "right-hand" side. This is in contrast to interdependency models that have no left- or right-hand side. I devote a chapter to one type of interdependency model, which is tied into a dependency model. Because interdependency models comprise a minimal proportion of the data analyst's workload, I humbly suggest that the focus of this book will prove utilitarian.

Therefore, these essays have been organized in the following fashion. Chapter 1 reveals the two most influential factors in my professional life: John W. Tukey and the personal computer (PC). The PC has changed everything in the world of statistics. The PC can effortlessly produce precise calculations and eliminate the computational burden associated with statistics. One need only provide the right questions. Unfortunately, the confluence of the PC and the world of statistics has turned generalists with minimal statistical backgrounds into quasi-statisticians and affords them a false sense of confidence.

In 1962, in his influential article, "The Future of Data Analysis" [1], John Tukey predicted a movement to unlock the rigidities that characterize statistics. It was not until the publication of *Exploratory Data Analysis* [2] in 1977 that Tukey led statistics away from the rigors that defined it into a new area, known as EDA (from the first initials of the title of his seminal work). At its core, EDA, known presently as data mining or formally as statistical data mining, is an unending effort of numerical, counting, and graphical detective work.

To provide a springboard to more esoteric methodologies, Chapter 2 covers the correlation coefficient. While reviewing the correlation coefficient, I bring to light several issues unfamiliar to many, as well as introduce two useful methods for variable assessment. Building on the concept of smooth scatterplot presented in Chapter 2, I introduce in Chapter 3 the smoother scatterplot based on CHAID (chi-squared automatic interaction detection). The new method has the potential of exposing a more reliable depiction of the unmasked relationship for paired-variable assessment than that of the smoothed scatterplot.

In Chapter 4, I show the importance of straight data for the simplicity and desirability it brings for good model building. In Chapter 5, I introduce the method of symmetrizing ranked data and add it to the paradigm of simplicity and desirability presented in Chapter 4.

Principal component analysis, the popular data reduction technique invented in 1901, is repositioned in Chapter 6 as a data mining method for many-variable assessment. In Chapter 7, I readdress the correlation coefficient. I discuss the effects the distributions of the two variables under consideration have on the correlation coefficient interval. Consequently, I provide a procedure for calculating an adjusted correlation coefficient.

In Chapter 8, I deal with logistic regression, a classification technique familiar to everyone, yet in this book, one that serves as the underlying rationale for a case study in building a response model for an investment product. In doing so, I introduce a variety of new data mining techniques. The continuous side of this target variable is covered in Chapter 9. On the heels of discussing the workhorses of statistical regression in Chapters 8 and 9, I resurface the scope of literature on the weaknesses of variable selection methods, and I enliven a notable solution for specifying a well-defined regression model in Chapter 10 anew. Chapter 11 focuses on the interpretation of the logistic regression model with the use of CHAID as a data mining tool. Chapter 12 refocuses on the regression coefficient and offers common misinterpretations of the coefficient that point to its weaknesses. Extending the concept of the coefficient, I introduce the average correlation coefficient in Chapter 13 to provide a quantitative criterion for assessing competing predictive models and the importance of the predictor variables.

In Chapter 14, I demonstrate how to increase the predictive power of a model beyond that provided by its variable components. This is accomplished by creating an interaction variable, which is the product of two or more component variables. To test the significance of the interaction variable, I make what I feel to be a compelling case for a rather unconventional use of CHAID. Creative use of well-known techniques is further carried out in Chapter 15, where I solve the problem of market segment classification modeling using not only logistic regression but also CHAID. In Chapter 16, CHAID is yet again utilized in a somewhat unconventional manner—as a method for filling in missing values in one's data. To bring an interesting real-life problem into the picture, I wrote Chapter 17 to describe profiling techniques for the marketer who wants a method for identifying his or her best customers. The benefits of the predictive profiling approach is demonstrated and expanded to a discussion of look-alike profiling.

I take a detour in Chapter 18 to discuss how marketers assess the accuracy of a model. Three concepts of model assessment are discussed: the traditional decile analysis, as well as two additional concepts, precision and separability. In Chapter 19, continuing in this mode, I point to the weaknesses in the way the decile analysis is used and offer a new approach known as the bootstrap for measuring the efficiency of marketing models.

The purpose of Chapter 20 is to introduce the principal features of a bootstrap validation method for the ever-popular logistic regression model. Chapter 21 offers a pair of graphics or visual displays that have value beyond the commonly used exploratory phase of analysis. In this chapter, I demonstrate the hitherto untapped potential for visual displays to describe the functionality of the final model once it has been implemented for prediction.

I close the statistical data mining part of the book with Chapter 22, in which I offer a data-mining alternative measure, the predictive contribution coefficient, to the standardized coefficient.

With the discussions just described behind us, we are ready to venture to new ground. In Chapter 1, I elaborated on the concept of machine-learning data mining and defined it as PC learning without the EDA/statistics component. In Chapter 23, I use a metrical modelogue, "To Fit or Not to Fit Data to a Model," to introduce the machine-learning method of GenIQ and its favorable data mining offshoots.

In Chapter 24, I maintain that the machine-learning paradigm, which lets the data define the model, is especially effective with big data. Consequently, I present an exemplar illustration of genetic logistic regression outperforming statistical logistic regression, whose paradigm, in contrast, is to fit the data to a predefined model. In Chapter 25, I introduce and illustrate brightly, perhaps, the quintessential data mining concept: data reuse. Data reuse is appending new variables, which are found when building a GenIQ Model, to the original dataset. The benefit of data reuse is apparent: The original dataset is enhanced with the addition of new, predictive-full GenIQ data-mined variables.

In Chapters 26–28, I address everyday statistics problems with solutions stemming from the data mining features of the GenIQ Model. In statistics, an outlier is an observation whose position falls outside the overall pattern of the data. Outliers are problematic: Statistical regression models are quite sensitive to outliers, which render an estimated regression model with questionable predictions. The common remedy for handling outliers is "determine and discard" them. In Chapter 26, I present an alternative method of moderating outliers instead of discarding them. In Chapter 27, I introduce a new solution to the old problem of overfitting. I illustrate how the GenIQ Model identifies a structural source (complexity) of overfitting, and subsequently instructs for deletion of the individuals who contribute to the complexity, from the dataset under consideration. Chapter 28 revisits the examples (the importance of straight data) discussed in Chapters 4 and 9, in which I posited the solutions without explanation as the material needed to understand the solution was not introduced at that point. At this point, the background required has been covered. Thus, for completeness, I detail the posited solutions in this chapter.

GenIQ is now presented in Chapter 29 as such a nonstatistical machine-learning model. Moreover, in Chapter 30, GenIQ serves as an effective method for finding the best possible subset of variables for a model. Because GenIQ has no coefficients—and coefficients furnish the key to prediction—Chapter 31 presents a method for calculating a quasi-regression coefficient, thereby providing a reliable, assumption-free alternative to the regression coefficient. Such an alternative provides a frame of reference for evaluating and using coefficient-free models, thus allowing the data analyst a comfort level for exploring new ideas, such as GenIQ.

References

1. Tukey, J.W., The future of data analysis, *Annals of Mathematical Statistics*, 33, 1–67, 1962.
2. Tukey, J.W., *Exploratory Data Analysis*, Addison-Wesley, Reading, MA, 1977.

Acknowledgments

This book, like all books—except the Bible—was written with the assistance of others. First and foremost, I acknowledge Hashem who has kept me alive, sustained me, and brought me to this season.

I am grateful to David Grubbs, my editor, who contacted me about outdoing myself by writing this book. I am indebted to the staff of the CRC Press/Taylor & Francis Group for their excellent work: Sherry Thomas, Editorial Assistant and Project Coordinator; Todd Perry, Project Editor; Victoria Jones, Copy Editor; Viswanath Prasanna, Senior Project Manager; Shanmuga Vadivu, Proofreader; Celia McCoy, Indexer, and Elise Weinger and Kevin Craig, Cover Designers.

Author

Bruce Ratner, PhD, The Significant Statistician™, is president and founder of DM STAT-1 Consulting, the boutique firm for statistical modeling and analysis, data mining, and machine-learning. Bruce specializes in all standard statistical techniques, as well as recognized and innovative machine-learning methods, such as the patented GenIQ Model. Bruce achieves the clients' goals across varied industries: direct and database marketing, banking, insurance, finance, retail, telecommunications, healthcare, pharmaceutical, publication and circulation, mass and direct advertising, catalog marketing, e-commerce, Web-mining, B2B, risk management, and nonprofit fundraising.

Bruce's par excellence the expertise is apparent as he is the author of the best-selling book, *Statistical Modeling and Analysis for Database Marketing: Effective Techniques for Mining Big Data*. Bruce ensures the optimal solution methodology for his clients' marketing problems with rapid startup and timely delivery of project results. Bruce executes the highest level of statistical practice for his clients' projects. He is an often-invited speaker at public industry events, such as the SAS Data Mining Conference, and private seminars at the request of *Fortune* magazine's top 100 companies.

Bruce has his footprint in the predictive analytics community as a frequent speaker at industry conferences and as the instructor of the advanced statistics course sponsored by the Direct Marketing Association for more than a decade. He is the author of more than 100 peer-reviewed articles on statistical and machine-learning procedures and software tools. He is a coauthor of the popular textbook, *The New Direct Marketing*, and is on the editorial board of the *Journal of Database Marketing and Customer Strategy*.

Bruce is also active in the online data mining industry. He is a frequent contributor to *KDnuggets Publications,* the top resource for the data mining community. His articles on statistical and machine-learning methodologies draw a huge monthly following. Other online venues in which Bruce participates are the networking sites LinkedIn and ResearchGate, in which his postings on statistical and machine-learning procedures for big data have sparked countless rich discussions. Also, he is the author of his own *DM STAT-1 Newsletter* on the web.

Bruce holds a doctorate in mathematics and statistics, with a concentration in multivariate statistics and response model simulation. His research interests include developing hybrid modeling techniques, which combine traditional statistics and machine-learning methods. He holds a patent for a unique application in solving the two-group classification problem with genetic programming.

19

Market Segmentation Based on Time-Series Data Using Latent Class Analysis

19.1 Introduction

Market segmentation is an often-used marketing model for efficient allocation of a company's resources. Market segmentation divides a population of customers into subpopulations—segments of customers. Customers within a segment are similar in their products and services, and customers across segments are dissimilar in their products and services. Market segmentation model implementation allows for effectively applying resources by targeting customers within their assigned segments. There are many statistical methods for market segmentation. A traditional and popular method is k-means clustering. A not-as-well-known method is latent class analysis (LCA). The purpose of this chapter is to present a novel approach to building a market segmentation model based on time-series data using LCA. I provide SAS© subroutines so data miners can perform segmentations similar to those presented, and I offer a unique way of incorporating time-series data in an otherwise cross-sectional dataset. The subroutines are also available for downloading from my website: http://www.geniq.net/articles.html#section9.

19.2 Background

I pedagogically outline this chapter. First, I concisely describe k-means clustering. Second, I cursorily review principal component analysis (PCA). I need to revisit PCA (from Chapter 7) because PCA and factor analysis (FA) are often confused, and FA is quite helpful in explaining LCA. Third, I present the LCA. Fourth, I compare LCA and k-means clustering. Lastly, I illustrate building a market segmentation model based on times-series data using LCA.

19.2.1 K-Means Clustering*

K-means clustering generates k mutually exclusive groups by distances computed from one or more quantitative variables (Xs) [1]. Every observation belongs to one and only

* This section draws on https://support.sas.com/rnd/app/stat/procedures/fastclus.html.

one group of the k clusters. Because the number of clusters k is unknown, the model builder performs as many cluster solutions as desired. The typical k-means procedure uses Euclidean distance (i.e., least-squares calculations) that yields cluster means (of the observations in a cluster for the Xs). If groups exist, then all distances among observations in a group are less than all distances among observations in a different group.

The k-means algorithm is heuristic, which involves the following steps:

1. Diversely select k initial seeds (i.e., random points in the X-space of the observations).
2. These points represent initial cluster means.
3. Assign each observation to the cluster whose mean is the closest to the observation.
4. When all observations are assigned, recalculate the k means.
5. Repeat Steps #2 and #3, until the k means are stable.

19.2.2 PCA

PCA transforms a set of p variables[*] X1, X2, …, Xj, ..., Xp into p linear combination variables PC1, PC2, ..., PCj, ..., PCp (PCj denotes the j-th principal component). The essential objective of PCA is to establish a smaller set of the new PCj variables that represents most of the information (variation) in the original set of variables. An attractive analytic feature of the PCs is that they are uncorrelated with each other. The PCs are defined as:

$$PC1 = a11*X1 + a12*X2 + \ldots + a1j*Xj + \ldots + a1p*Xp$$

$$PC2 = a21*X1 + a22*X2 + \ldots + a2j*Xj + \ldots + a2p*Xp$$

$$\vdots$$

$$PCi = ai1*X1 + ai2*X2 + \ldots + aij*Xj + \ldots + aip*Xp$$

$$\vdots$$

$$PCp = ap1*X1 + ap2*X2 + \ldots + apj*Xj + \ldots + app*Xp$$

where the aij's are the PC coefficients.

19.2.3 FA

There is confusion among too many statistics practitioners who do not understand the difference between PCA and FA. The reasons for the confusion is perhaps because:

1. PCA is often mentioned in textbooks as a particular case of FA.
2. Statistical computer packages treat PCA as an option in FA modules.

[*] For ease of presentation, the Xs are standardized.

3. PCA and FA both aim to reduce the dimensionality of the given dataset: PCA and FA are data reduction techniques. For example, a dataset with, say, 1,000 variables can be reduced to a statistically equivalent dataset with, say, only 150 variables.
4. PCA has been used extensively as a part of the FA solution.

19.2.3.1 FA Model

The FA model is defined as: p variables X1, X2, …, Xj, ..., Xp can be expressed (up to an error term) as a linear combination of m latent (unobservable) *continuous* variables or factor F1, F2, …, Fj, ..., Fm. Fs are defined as:

$$Xl = c11*Fl + c12*F2 + ... + c1j*Fj + ... + c1m*Fm$$

$$X2 = c21*Fl + c22*F2 + ... + c2j*Fj + ... + c2m*Fm$$

$$\vdots$$

$$Xi = ci1*F1 + ci2*F2 + ... + cij*Fj + ... + cim*Fm$$

$$\vdots$$

$$Xp = cp1*F1 + cp2*F2 + ... + cpj*Fj + ... + cpm*Fm$$

where cij's are like regression coefficients, called factor loadings.

The difference between PCA and FA is immediately apparent. Focus on the i-th PC and Xi:

$$PCi = ai1*X1 + ai2*X2 + \ldots + aij2*Xij + \ldots + aip*Xp$$

$$Xi = cil*Fl + ci2*F2 + ... + cij*Fj + \ldots + cim*Fm$$

The PCs are a linear combination of the original *observed* Xs. Each Xi is a linear combination of *unobserved* latent factors F1, F2, …, Fj, ..., Fm.

FA attempts to achieve a reduction from p to m dimensions by postulating a model relating p Xs to m latent factors. In contrast, PCA directly transforms the Xs into PCs, which possess the desirable properties cited in Chapter 7.

19.2.3.2 FA Model Estimation

At first sight, the FA model (with simplified subscripts) looks like a standard ordinary least squares (OLS) regression model.

$$FA: X = c1*Fl + c2*F2 + ... + cj*Fj + ... + cm*Fm \tag{19.1}$$

$$OLS: Y = b1*X1 + b2*X2 + \ldots bj*Kj + ... + bm*Xm \tag{19.2}$$

However, a closer inspection reveals a substantial difference. With FA, c's and Fs are both *unknown*. With OLS, b's are *unknown*, and the Xs are known.

The PCA solution to FA or the so-called principal FA involves the following steps:

1. Obtain initial estimates of the c's by taking the first m PCs from a PCA.
2. Obtain initial estimates of the Fs. X and c's are known. Thus, F can be initially determined.
3. Initial estimates are then fine-tuned, looping within Step #2, until optimal.

19.2.3.3 FA versus OLS Graphical Depiction

The typical graphical depictions of the FA and OLS models clearly indicate the theoretical framework of these two models in Figures 19.1 and 19.2.

The two models illustrated with three Xs are now visibly different in conceptual ways. The FA model has line arrows from the factor F to the Xs. The direction of the arrows indicates that the factor, the unobserved latent variable, affects the independent variables X1, X2, and X3. The lower case e's indicate the unknown errors associated with the measurement of the observed corresponding Xs. The OLS regression model has line arrows from the independent variables X1, X2, and X3 to the dependent variable Y. The direction of the arrows indicates the independent variables X1, X2, and X3 affect the dependent variable Y. There are no measurement errors assumed in OLS that influence Xs.

19.2.4 LCA versus FA Graphical Depiction

As a proper introduction to LCA is in the next section, I bring out the traditional setting of LCA, which is often considered similar to FA—with one difference. When the LCA model (Figure 19.3) is compared to the FA model (Figure 19.1), the similarity is visible. The significant and useful difference between the methods is the factor F of FA is a *continuous* latent variable, while the latent variable LC of LCA is *categorical*.

The usual graphic display for LCA is Figure 19.3. However, this display does not clearly portray the categorical nature of LC. I offer a refocused visual for an LCA 2-Class model in Figure 19.4, which clearly indicates LC is categorical. LC(1) and LC(2) are the two categories of LC. The crisscross-like arrows indicate each LC affects the independent variables X1, X2, and X3. Important to note, LCA nomenclature for an independent variable is *indicator*.

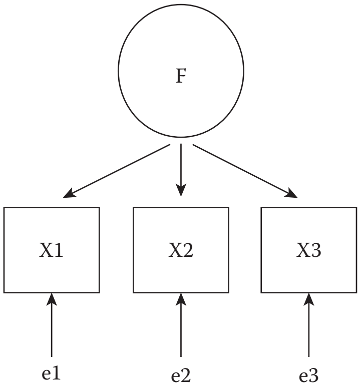

FIGURE 19.1
FA model graphic.

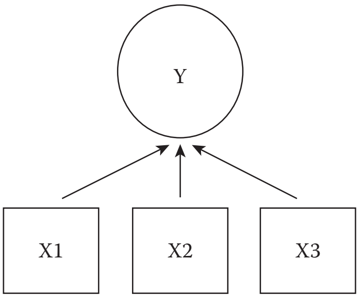

FIGURE 19.2
OLS model graphic.

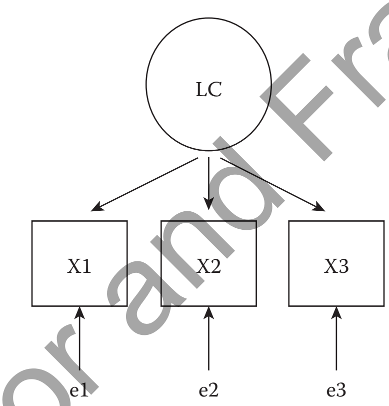

FIGURE 19.3
LCA model graphic.

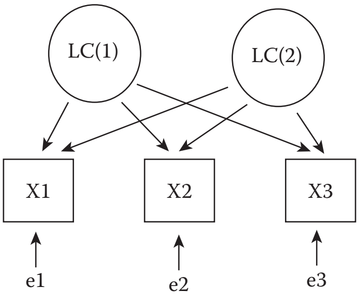

FIGURE 19.4
LCA 2-Class model graphic.

19.3 LCA

The concept of LCA, a statistical technique for identifying unobservable subpopulations within a population, was conceived in 1950 by Lazarsfeld, the father of this concept [2]. In 1968, Lazarsfeld (by then the *grandfather* of LCA) published the first comprehensive treatment of LCA with only categorical indicators, but he did not provide a reliable method for parameter estimation [3]. In 1974, Goodman solved the problem of obtaining maximum likelihood estimates* of LCA model parameters [4]. The traditional LCA was generalized by Vermunt in 1998 to include all scales of data—categorical, continuous, count, and ordinal [5].

19.3.1 LCA of Universal and Particular Study

As an illustration of how LCA works along with its output, I present a well-known and often revisited "Role Conflict and Personality" article, often referred to as the Universal and Particular Study, conducted in 1950 [6], which Goodman examined in 1974 [7]. In the 1950 study, 216 Harvard/Radcliffe undergraduates were asked how they would respond in four role-conflict situations. The premise of the role conflict study is "What right has your friend to expect you to protect him?" The four conflict scenarios† are:

1. You are riding in a car driven by a close friend, and he hits a pedestrian. You know that he is going at least 35 miles an hour (MPH) in a 20 MPH speed zone. There are no other witnesses. His lawyer says that if you testify under oath that the speed was only 20 MPH, it may save your friend from serious consequences.
   a. Universalistic response: "He has *no right* as a friend to expect me to testify to the lower figure."
   b. Particularistic response: "He has *a right* as a friend to expect me to give evidence to the lower figure."
2. As a physician, your friend asks you to "shade doubts" about a physical examination for an insurance policy.
3. As a drama critic, your friend asks you to "go easy on a review" of a bad play in which all of his savings are invested.
4. As a member of the board of directors, your friend intimates that you will "tip him off" about financially ruinous, though secret, company information.

The response patterns of the 216 respondents generate the 16 (= 2 × 2 × 2 × 2) row patterns in Table 19.1.‡ For each of the conflict scenarios (A, B, C, D), responses tending toward one that is universalistic are indicated by "+ ", and responses tending toward one that is particularistic are indicated by "–" [7].

19.3.1.1 Discussion of LCA Output

Hagenaars and McCutcheon perform a simple two-class LCA on the Universal-Particular study [8]. The vital statistics of LCA output in Table 19.2 consist of (1) latent class probabilities and (2) conditional probabilities.

* Maximum likelihood estimation, specifically, the expectation–maximization (EM) process.

† The descriptions of the scenarios are taken from the original study.

‡ Table is taken from Goodman 1974, p. 216.

TABLE 19.1

Response Patterns of the Four Conflict Scenarios

A	B	C	D	Observed Frequency	A	B	C	D	Observed Frequency
+	+	+	+	42	−	+	+	+	1
+	+	+	−	23	−	+	+	−	4
+	+	−	+	6	−	+	−	+	1
+	+	−	−	25	−	+	−	−	6
+	−	+	+	6	−	−	+	+	2
+	−	+	−	24	−	−	+	−	9
+	−	−	+	7	−	−	−	+	2
+	−	−	−	38	−	−	−	−	20

TABLE 19.2

LCA Output of Study

Observed Indicators	Universal	Particular
Auto passenger friend	0.993	0.714
Insurance doctor friend	0.939	0.329
Drama critic friend	0.929	0.354
Board of directors friend	0.769	0.132
Latent Class Size	0.280	0.720

The latent class probabilities are the sizes of the two latent classes, which are 0.280 and 0.720 for Universal and Particular, respectively. These probabilities show 28% and 72% of the study population are in the Universal and Particular classes, respectively.

Within the similarity of LCA and FA, a conditional probability is comparable to the factor loading in FA. Large conditional probabilities for a latent class imply the corresponding indicators are highly associated with the latent class and therefore define the latent class.

The large probabilities of the first class column, ranging from 0.993 to 0.769, confirm the correct labeling of the column as Universal. For the second class column, the large (0.714), moderate (0.329 and 0.354), and low (0.132) conditional probabilities reasonably define the Particular class. In other words, latent class Universal is reliably identified respondents, whereas Particular is moderately determined.

Within LCA, the conditional probability is the probability of an individual in a given class responding at a given level of the indicator. The conditional probability is 99.3% for a respondent in class Universal responding that there is no right to lie for the driver friend. The conditional probability is 71.4% for a respondent in Particular class responding that there is a right to lie for the driver friend.

Similarly, for a respondent who is in the Universal class and answers that there is no reason for doctor friend to lie, the conditional probability is 93.9%. And, the conditional probability is 32.9% for a respondent in Particular class responding that there is a reason for doctor friend to lie.

19.3.1.2 Discussion of Posterior Probability

The relevant statistic noticeably missing in Table 19.2 is the posterior probability—the probability of class membership *given* a respondent's answers to a given scenario. This statistic allows for probabilistically classifying a respondent into one of the latent classes. The posterior probability is notationally represented as Prob (LC = c | A = i, B = j, C = k, D = l), where c = 1, 2; i = yes, no; j = yes, no; k = yes, no; and l = yes, no. The symbol | represents the word *given*.

The calculation of posterior probability needs the LCA joint probability, Prob (A = i, B = j, C = k, D = l, CL = c), defined as the product of:

- Prob (individual at CL = c)
- Prob (individual at A = i | CL = c)
- Prob (individual at B = j | CL = c)
- Prob (individual at C = k | CL = c)
- Prob (individual at D = l | CL = c)

Thus, the posterior probability of an individual belonging to latent class c (CL = c) given A = i, B = j, C = k, D = l is in Equation 19.3:

$$\begin{aligned}&\mathrm{Prob}\left(\mathrm{LC}=\mathrm{c}\,|\,\mathrm{A}=\mathrm{i},\mathrm{B}=\mathrm{j},\mathrm{C}=\mathrm{k},\mathrm{D}=\mathrm{l}\right)\\&\quad=\mathrm{Prob}\left(\mathrm{A}=\mathrm{i},\mathrm{B}=\mathrm{j},\mathrm{C}=\mathrm{k},\mathrm{D}=\mathrm{l},\mathrm{CL}=\mathrm{c}\right)/\,\mathrm{sum}\left[\mathrm{Prob}\left(\mathrm{A}=\mathrm{i},\mathrm{B}=\mathrm{j},\mathrm{C}=\mathrm{k},\mathrm{D}=\mathrm{l},\mathrm{CL}=\mathrm{c}\right)\right],\text{ across c}\end{aligned} \tag{19.3}$$

Thus far, the discussion of LCA regards only categorical indicators. LCA with continuous and categorical indicators generates output virtually identical to that of LCA with only categorical indicators. Expectantly, the statistical computations of LCA with continuous variables add a level of detail and knowledge of the mathematical statistics underpinnings of LCA beyond the scope of this chapter. Accordingly, there is no discussion of the extended LCA, for which Vermunt (1998) is the definitive source.

19.4 LCA versus k-Means Clustering

LCA and k-means clustering have the same objective of dividing a population of individuals into k disjoint and exhaustive subpopulations (clusters) such that individuals within a group are as similar as possible, and the individuals among the groups are as dissimilar as possible. In statistics-speak, cluster construction seeks to maximize the between-cluster differences and minimize the within-cluster differences.

The fundamental underlying difference between LCA and k-means methodologies is that LCA is model-based, whereas k-means is a heuristic (technique). A heuristic is any approach to solving a problem that uses intuition based on the problem domain. Heuristics do not provide optimal solutions but rather good solutions. Markedly, LCA as a model-based statistical technique means that it posits a theoretical equation that represents the data-generating process of the population and drawn sample data. Moreover, the probability distribution of the statistical model is the distinguishing feature between model-based versus data-based heuristics.

K-means clustering is a popular, practical, and useful tool, especially for market segmentation. Similar to all techniques, k-means has strengths and weaknesses. The strengths (Items 1–4) and weaknesses (Items 5–12) are listed as follows:

1. K-means is easy to use, understand, and implement.
2. K-means performs best with many variables, which it can accommodate because it is computationally efficient.
3. K-means always produces a cluster solution.
4. K-means tends to produce tighter clusters than alternative techniques.
5. K-means can create a cluster with one or a handful of individuals.
6. K-means cannot suggest the optimal number of clusters.
7. K-means is sensitive to outliers. (When few individuals define a cluster, they are outliers.)
8. K-means is not a robust method in that pseudorandom seeds yield different solutions. Or, a holdout dataset can produce a different cluster solution.
9. The k-means algorithm does not reveal which variables are relevant.
10. The k-means algorithm is affected by variables with large variances. Accordingly, k-means often use standardized data.
11. K-means always produces a cluster solution. Always-a-solution gives a false sense of statistical security, in that k-means finds a good solution.
12. An objective criterion by which to assess the quality of the cluster solution does not exist.

LCA as a statistical model has the apparatus that indicates its strengths (first four items in the list that follows). Unfortunately, the apparatus does not eliminate weaknesses (last seven items in the list that follows).

1. There are objective criteria by which to assess the goodness-of-fit of the cluster solution.
2. Statistical testing exists for comparison between two or more candidate cluster solutions.
3. LCA solutions are not affected by the different scales and unequal or large variances of the indicators.
4. LCA, like all statistical models, produces residuals. Residual analysis is crucial in assessing remedies of lack-of-fit.
5. LCA cannot suggest the optimal number of clusters.
6. A serious problem with LCA is the assumption of conditional dependence, also known as local independence. Local independence means the indicators within a cluster class are independent of each other. The local independence is sometimes not a tenable assumption. The standard LCA model must be modified to account for this (http://john-uebersax.com/stat/faq.htm).
7. Methods for relaxing the conditional independence assumption are in-progress in recent years.
8. LCA uses maximum likelihood estimation like many statistical models. Thus, LCA solutions are subject to local maxima not necessarily the global maxima.

9. There are many goodness-of-fitness criteria, such as the basics: likelihood (L), log-likelihood (LL), and L-square (L^2).
10. Some criteria weight the fit and the parsimony of a model based on sample size and degrees of freedom. They are:
    a. Akaike Information Criteria (AIC):* AIC, AIC(L^2), AIC3(L^2), AIC(LL), and AIC3(LL).
    b. Bayesian Information Criteria (BIC):† BIC, BIC2(L^2), and BIC(LL).
    c. None of these criteria are universally deemed superior to another.
    d. The behavior of these statistics creates confusion and uncertainty as to which is the best cluster solution. For example, when comparing two models, it is not known how much difference in BIC is significant in order to choose one model over another [9].
11. Restrictions when assessing goodness-of-fit complicate building the model. Hypotheses are tested by imposing restrictions and determining how these limitations affect the fit of the model to the data. Two such restrictions include (1) equality constraint (e.g., parallel indicators or equal error rate) and (2) deterministic (e.g., setting conditional probability to a particular value—usually 1 or 0).
12. Sparseness, due to many indicators with many response options, leads to difficulties in model evaluation (e.g., determining the degrees of freedom).

19.5 LCA Market Segmentation Model Based on Time-Series Data

I present the building of a market segmentation model based on times-series data using LCA.

The proposed model is not to be confused with partitioning of a times series, producing a sequence of discrete segments to reveal the underlying structure of input time-series data. A typical method of a times-series segmentation is a piecewise linear regression, which forecasts, say, stock market trading, inputting time series into k straight lines, each of an equal length. The proposed LCA market segmentation model is a novel and efficient segmentation technique, based on time-series data that are used to create indicators, as required by LCA.

I outline the intended LCA time-series market segmentation model pedagogically in a step-by-step manner to show (1) the components of the time-series data preparation with the SAS subroutines used and (2) building the LCA segmentation. The LCA program used (not provided here) is among several commercial software packages available.

19.5.1 Objective

Hi-tech company PirSQ wants to build a market segmentation as an efficient way to target their best customers. The segmentation, providing segments or clusters, allows PirSQ to develop marketing strategies to optimize future unit orders.

* AIC= 2*Npar – 2*ln(L). AIC(L^2) = L^2 – 2*df. AIC3(L^2) = L^2 – 3*df. AIC(LL) = 2 log L + 2*Npar. AIC3(LL) = 2 log L + 3 Npar. Npar = number of parameters.

† BIC= -2*ln(L) + Npar*ln(N). BIC(L^2) =L^2 – log(N)*df. BIC(LL) = 2*log L + log(N)*Npar. Npar = number of parameters, and N = sample size.