



Fundamentals of Predictive Analytics with JMP®, Second Edition. Full book available for purchase [here](#).

Contents

About This Book	xiii
About These Authors	xvii
Acknowledgments	xiv
Chapter 1: Introduction	1
Historical Perspective	1
Two Questions Organizations Need to Ask	2
Return on Investment.....	2
Cultural Change	2
Business Intelligence and Business Analytics.....	2
Introductory Statistics Courses.....	4
The Problem of Dirty Data	5
Added Complexities in Multivariate Analysis	5
Practical Statistical Study	6
Obtaining and Cleaning the Data.....	6
Understanding the Statistical Study as a Story.....	7
The Plan-Perform-Analyze-Reflect Cycle	7
Using Powerful Software	8
Framework and Chapter Sequence	9
Chapter 2: Statistics Review	11
Introduction	11
Fundamental Concepts 1 and 2.....	12
FC1: Always Take a Random and Representative Sample	12
FC2: Remember That Statistics Is Not an Exact Science	13
Fundamental Concept 3: Understand a Z-Score	14
Fundamental Concept 4	15
FC4: Understand the Central Limit Theorem.....	15
Learn from an Example.....	17
Fundamental Concept 5	20
Understand One-Sample Hypothesis Testing	20

Consider p-Values	21
Fundamental Concept 6:	22
Understand That Few Approaches/Techniques Are Correct—Many Are Wrong	22
Three Possible Outcomes When You Choose a Technique	32
Chapter 3: Dirty Data	33
Introduction	33
Data Set.....	34
Error Detection	35
Outlier Detection	39
Approach 1	39
Approach 2	41
Missing Values	43
Statistical Assumptions of Patterns of Missing	44
Conventional Correction Methods.....	45
The JMP Approach	48
Example Using JMP	49
General First Steps on Receipt of a Data Set	54
Exercises.....	54
Chapter 4: Data Discovery with Multivariate Data	57
Introduction	57
Use Tables to Explore Multivariate Data	59
PivotTables.....	59
Tabulate in JMP	62
Use Graphs to Explore Multivariate Data	64
Graph Builder	64
Scatterplot.....	68
Explore a Larger Data Set	71
Trellis Chart.....	71
Bubble Plot.....	73
Explore a Real-World Data Set.....	78
Use Graph Builder to Examine Results of Analyses	79
Generate a Trellis Chart and Examine Results.....	80
Use Dynamic Linking to Explore Comparisons in a Small Data Subset.....	83
Return to Graph Builder to Sort and Visualize a Larger Data Set	84
Chapter 5: Regression and ANOVA	87
Introduction	87
Regression.....	88

Perform a Simple Regression and Examine Results	88
Understand and Perform Multiple Regression	89
Understand and Perform Regression with Categorical Data	101
Analysis of Variance.....	107
Perform a One-Way ANOVA	108
Evaluate the Model.....	109
Perform a Two-Way ANOVA.....	122
Exercises.....	129
Chapter 6: Logistic Regression	131
Introduction	131
Dependence Technique	132
The Linear Probability Model	132
The Logistic Function.....	133
A Straightforward Example Using JMP	135
Create a Dummy Variable	135
Use a Contingency Table to Determine the Odds Ratio	136
Calculate the Odds Ratio.....	138
A Realistic Logistic Regression Statistical Study.....	147
Understand the Model-Building Approach	147
Run Bivariate Analyses	150
Run the Initial Regression and Examine the Results.....	152
Convert a Continuous Variable to Discrete Variables	154
Produce Interaction Variables.....	155
Validate and Use the Model.....	157
Exercises.....	158
Chapter 7: Principal Components Analysis	159
Introduction	159
Basic Steps in JMP	160
Produce the Correlations and Scatterplot Matrix.....	160
Create the Principal Components	161
Run a Regression of y on Prin1 and Excluding Prin2	163
Understand Eigenvalue Analysis	163
Conduct the Eigenvalue Analysis and the Bartlett Test.....	164
Verify Lack of Correlation.....	164
Dimension Reduction	165
Produce the Correlations and Scatterplot Matrix.....	165
Conduct the Principal Component Analysis.....	165

Determine the Number of Principal Components to Select.....	166
Compare Methods for Determining the Number of Components.....	167
Discovery of Structure in the Data	167
A Straightforward Example	168
An Example with Less Well Defined Data	169
Exercises.....	171
Chapter 8: Least Absolute Shrinkage and Selection Operator and Elastic Net	173
Introduction	173
The Importance of the Bias-Variance Tradeoff	174
Ridge Regression	175
Least Absolute Shrinkage and Selection Operator	179
Perform the Technique	180
Examine the Results.....	180
Refine the Results	182
Elastic Net.....	182
Perform the Technique	182
Examine the Results.....	183
Compare with LASSO.....	184
Exercises.....	184
Chapter 9: Cluster Analysis.....	185
Introduction	185
Example Applications.....	186
An Example from the Credit Card Industry.....	186
The Need to Understand Statistics and the Business Problem	187
Hierarchical Clustering.....	187
Understand the Dendrogram.....	187
Understand the Methods for Calculating Distance between Clusters	187
Perform a Hierarchical Clustering with Complete Linkage	189
Examine the Results.....	191
Consider a Scree Plot to Discern the Best Number of Clusters.....	193
Apply the Principles to a Small but Rich Data Set	193
Consider Adding Clusters in a Regression Analysis.....	196
K-Means Clustering	197
Understand the Benefits and Drawbacks of the Method.....	197
Choose k and Determine the Clusters	198
Perform k-Means Clustering	200

Change the Number of Clusters	202
Create a Profile of the Clusters with Parallel Coordinate Plots	204
Perform Iterative Clustering	206
Score New Observations	207
K-Means Clustering versus Hierarchical Clustering	208
Exercises	209
Chapter 10: Decision Trees.....	211
Introduction	211
Benefits and Drawbacks.....	212
Definitions and an Example.....	213
Theoretical Questions	213
Classification Trees	214
Begin Tree and Observe Results	215
Use JMP to Choose the Split That Maximizes the LogWorth Statistic.....	217
Split the Root Node According to Rank of Variables.....	218
Split Second Node According to the College Variable	219
Examine Results and Predict the Variable for a Third Split	221
Examine Results and Predict the Variable for a Fourth Split.....	221
Examine Results and Continue Splitting to Gain Actionable Insights	222
Prune to Simplify Overgrown Trees.....	222
Examine Receiver Operator Characteristic and Lift Curves.....	223
Regression Trees	224
Understand How Regression Trees Work	225
Restart a Regression Driven by Practical Questions	228
Use Column Contributions and Leaf Reports for Large Data Sets	229
Exercises.....	230
Chapter 11: k-Nearest Neighbors	231
Introduction	231
Example—Age and Income as Correlates of Purchase	232
The Way That JMP Resolves Ties.....	234
The Need to Standardize Units of Measurement.....	234
k-Nearest Neighbors Analysis	235
Perform the Analysis	235
Make Predictions for New Data	236
k-Nearest Neighbor for Multiclass Problems	237
Understand the Variables	237
Perform the Analysis and Examine Results.....	238

The k-Nearest Neighbor Regression Models	239
Perform a Linear Regression as a Basis for Comparison	239
Apply the k-Nearest Neighbors Technique	239
Compare the Two Methods	240
Make Predictions for New Data	242
Limitations and Drawbacks of the Technique	242
Exercises	243
Chapter 12: Neural Networks	245
Introduction	245
Drawbacks and Benefits	246
A Simplified Representation	246
A More Realistic Representation	248
Understand Validation Methods	250
Holdback Validation	250
k-fold Cross-Validation	251
Understand the Hidden Layer Structure	252
A Few Guidelines for Determining Number of Nodes	252
Practical Strategies for Determining Number of Nodes	252
The Method of Boosting	253
Understand Options for Improving the Fit of a Model	254
Complete the Data Preparation	255
Use JMP on an Example Data Set	256
Perform a Linear Regression as a Baseline	256
Perform the Neural Network Ten Times to Assess Default Performance	258
Boost the Default Model	259
Compare Transformation of Variables and Methods of Validation	260
Exercises	266
Chapter 13: Bootstrap Forests and Boosted Trees	267
Introduction	267
Bootstrap Forests	268
Understand Bagged Trees	269
Perform a Bootstrap Forest	269
Perform a Bootstrap Forest for Regression Trees	274
Boosted Trees	275
Understand Boosting	275
Perform Boosting	275

Perform a Boosted Tree for Regression Trees	278
Use Validation and Training Samples.....	278
Exercises.....	284
Chapter 14: Model Comparison	285
Introduction	285
Perform a Model Comparison with Continuous Dependent Variable	286
Understand Absolute Measures	286
Understand Relative Measures.....	287
Understand Correlation between Variable and Prediction	287
Explore the Uses of the Different Measures.....	288
Perform a Model Comparison with Binary Dependent Variable	290
Understand the Confusion Matrix and Its Limitations.....	290
Understand True Positive Rate and False Positive Rate.....	291
Interpret Receiving Operator Characteristic Curves.....	291
Compare Two Example Models Predicting Churn.....	294
Perform a Model Comparison Using the Lift Chart.....	296
Train, Validate, and Test.....	299
Perform Stepwise Regression	299
Examine the Results of Stepwise Regression.....	302
Compute the MSE, MAE, and Correlation	302
Examine the Results for MSE, MAE, and Correlation.....	303
Understand Overfitting from a Coin-Flip Example.....	303
Use the Model Comparison Platform	305
Exercises.....	317
Chapter 15: Text Mining.....	319
Introduction	319
Historical Perspective	320
Unstructured Data.....	320
Developing the Document Term Matrix.....	321
Understand the Tokenizing Stage	323
Understand the Phrasing Stage.....	329
Understand the Terming Stage.....	330
Observe the Order of Operations	331
Developing the Document Term Matrix with a Larger Data Set	332
Generate a Word Cloud and Examine the Text.....	333
Examine and Group Terms.....	334
Add Frequent Phrases to List of Terms	335
Parse the List of Terms.....	336

Using Multivariate Techniques	338
Perform Latent Semantic Analysis	338
Perform Topic Analysis	342
Perform Cluster Analysis	342
Using Predictive Techniques	347
Perform Primary Analysis	348
Perform Logistic Regressions.....	349
Exercises.....	353
Chapter 16: Market Basket Analysis.....	355
Introduction	355
Association Analyses	355
Examples	356
Understand Support, Confidence, and Lift.....	357
Association Rules	357
Support	357
Confidence	357
Lift	358
Use JMP to Calculate Confidence and Lift	358
Use the A Priori Algorithm for More Complex Data Sets	359
Form Rules and Calculate Confidence and Lift.....	359
Analyze a Real Data Set	360
Perform Association Analysis with Default Settings	360
Reduce the Number of Rules and Sort Them.....	360
Examine Results	361
Target Results to Take Business Actions	362
Exercises.....	363
Chapter 17: Statistical Storytelling.....	365
The Path from Multivariate Data to the Modeling Process	365
Early Applications of Data Mining.....	365
Numerous JMP Customer Stories of Modern Applications.....	366
Definitions of Data Mining.....	366
Data Mining	367
Predictive Analytics	367
A Framework for Predictive Analytics Techniques	367
The Goal, Tasks, and Phases of Predictive Analytics	369
The Difference between Statistics and Data Mining	370
SEMMA	371

References	373
Index.....	377

From Fundamentals of Predictive Analytics with JMP®, Second Edition, by Ron Klimberg and B. D. McCullough. Copyright © 2016, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.