



Fundamentals of Predictive Analytics with JMP[®]

Third Edition

Ron Klimberg

The correct bibliographic citation for this manual is as follows: Klimberg, Ron. 2023. *Fundamentals of Predictive Analytics with JMP®*, Third Edition. Cary, NC: SAS Institute Inc.

Fundamentals of Predictive Analytics with JMP®, Third Edition

Copyright © 2023, SAS Institute Inc., Cary, NC, USA

ISBN 978-1-68580-003-1 (Hardcover)

ISBN 978-1-68580-027-7 (Paperback)

ISBN 978-1-68580-000-0 (Web PDF)

ISBN 978-1-68580-001-7 (EPUB)

ISBN 978-1-68580-002-4 (Kindle)

All Rights Reserved. Produced in the United States of America.

For a hard copy book: No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

For a web download or e-book: Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

U.S. Government License Rights; Restricted Rights: The Software and its documentation is commercial computer software developed at private expense and is provided with RESTRICTED RIGHTS to the United States Government. Use, duplication, or disclosure of the Software by the United States Government is subject to the license terms of this Agreement pursuant to, as applicable, FAR 12.212, DFAR 227.7202-1(a), DFAR 227.7202-3(a), and DFAR 227.7202-4, and, to the extent required under U.S. federal law, the minimum restricted rights as set out in FAR 52.227-19 (DEC 2007). If FAR 52.227-19 is applicable, this provision serves as notice under clause (c) thereof and no other notice is required to be affixed to the Software or documentation. The Government's rights in Software and documentation shall be only those set forth in this Agreement.

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

April 2023

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

SAS software may be provided with certain third-party software, including but not limited to open-source software, which is licensed under its applicable third-party software license agreement. For license information about third-party software distributed with SAS software, refer to <https://support.sas.com/en/technical-support/license-assistance.html>.

Contents

About This Book	xiii
About The Author	xvii
Acknowledgments	xix
Dedication	xxi
Chapter 1: Introduction	1
Historical Perspective	1
Two Questions Organizations Need to Ask	1
Return on Investment	2
Cultural Change.....	2
Business Intelligence and Business Analytics	3
Introductory Statistics Courses	4
The Problem of Dirty Data	6
Added Complexities in Multivariate Analysis.....	6
Practical Statistical Study	6
Obtaining and Cleaning the Data	7
Understanding the Statistical Study as a Story	8
The Plan-Perform-Analyze-Reflect Cycle.....	8
Using Powerful Software.....	9
Framework and Chapter Sequence	10
Chapter 2: Statistics Review	13
Introduction.....	13
Fundamental Concepts 1 and 2	13
FC1: Always Take a Random and Representative Sample	14
FC2: Remember That Statistics Is Not an Exact Science.....	15
Fundamental Concept 3: Understand a Z-Score	16
Fundamental Concept 4	17
FC4: Understand the Central Limit Theorem	17
Learn from an Example.....	18
Fundamental Concept 5	22
Understand One-Sample Hypothesis Testing.....	22
Consider <i>p</i> -Values	23
Fundamental Concept 6	23
Understand That Few Approaches and Techniques Are Correct—Many Are Wrong.....	24

Ways JMP Can Access Data in Excel	24
Three Possible Outcomes When You Choose a Technique	33
Exercises	34
Chapter 3: Dirty Data	35
Introduction	35
Data Set	36
Error Detection	37
Outlier Detection	39
Approach 1	41
Approach 2	43
Missing Values	45
Statistical Assumptions of Patterns of Missing	46
Conventional Correction Methods.....	48
The JMP Approach	51
Example Using JMP	52
General First Steps on Receipt of a Data Set	58
Exercises	58
Chapter 4: Data Discovery with Multivariate Data	61
Introduction.....	61
Use Tables to Explore Multivariate Data.....	63
PivotTables	63
Tabulate in JMP.....	65
Use Graphs to Explore Multivariate Data	66
Graph Builder.....	66
Scatterplot	71
Explore a Larger Data Set.....	73
Trellis Chart	73
Bubble Plot	76
Explore a Real-World Data Set.....	79
Use Correlation Matrix and Scatterplot Matrix to Examine Relationships of	
Continuous Variables	79
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.....	85
Return to Graph Builder to Sort and Visualize a Larger Data Set	85
Exercises	87
Chapter 5: Regression and ANOVA.....	89
Introduction.....	89
Regression	89
Perform a Simple Regression and Examine Results	90
Understand and Perform Multiple Regression.....	92
Understand and Perform Regression with Categorical Data	104

Analysis of Variance.....	109
Perform a One-Way ANOVA.....	111
Evaluate the Model.....	111
Perform a Two-Way ANOVA.....	122
Exercises	130
Chapter 6: Logistic Regression	133
Introduction.....	133
Dependence Technique	133
The Linear Probability Model.....	134
The Logistic Function	135
A Straightforward Example Using JMP.....	137
Create a Dummy Variable	137
Use a Contingency Table to Determine the Odds Ratio	137
Calculate the Odds Ratio.....	140
Examine the Parameter Estimates	142
Compute Probabilities for Each Observation	147
Check the Model's Assumptions	148
A Realistic Logistic Regression Statistical Study	150
Understand the Model-Building Approach	151
Run Bivariate Analyses.....	154
Run the Initial Regression and Examine the Results	155
Convert a Continuous Variable to Discrete Variables.....	156
Producing Interaction Variables.....	158
Validate and Confusion Matrix.....	160
Exercises	161
Chapter 7: Principal Components Analysis	163
Introduction.....	163
Basic Steps in JMP.....	164
Produce the Correlations and Scatterplot Matrix	164
Create the Principal Components	164
Run a Regression of y on Prin1 and Excluding Prin2	167
Understand Eigenvalue Analysis	168
Conduct the Eigenvalue Analysis and the Bartlett Test.....	168
Verify Lack of Correlation.....	169
Dimension Reduction	169
Produce the Correlations and Scatterplot Matrix	169
Conduct the Principal Component Analysis.....	170
Determine the Number of Principal Components to Select	170
Compare Methods for Determining the Number of Components	172
Discovery of Structure in the Data.....	173
A Straightforward Example	173
An Example with Less Well-Defined Data	175
Exercises	177

Chapter 8: Least Absolute Shrinkage and Selection Operator and Elastic Net	179
Introduction.....	179
The Importance of the Bias-Variance Tradeoff	180
Ridge Regression.....	181
Least Absolute Shrinkage and Selection Operator.....	184
Perform the Technique	185
Examine the Results.....	185
Elastic Net.....	187
Perform the Technique	187
Compare with LASSO	187
Exercises	189
Chapter 9: Cluster Analysis	191
Introduction.....	191
Example Applications.....	191
An Example from the Credit Card Industry	192
The Need to Understand Statistics and the Business Problem	192
Hierarchical Clustering.....	193
Understand the Dendrogram.....	193
Understand the Methods for Calculating Distance between Clusters	193
Perform Hierarchical Clustering with Complete Linkage.....	194
Examine the Results.....	195
Consider a Scree Plot to Discern the Best Number of Clusters.....	196
Apply the Principles to a Small but Rich Data Set	198
Consider Adding Clusters in a Regression Analysis	201
k-Means Clustering.....	202
Understand the Benefits and Drawbacks of the Method	202
Choose k and Determine the Clusters.....	203
Perform k-Means Clustering	206
Change the Number of Clusters.....	206
Create a Profile of the Clusters with Parallel Coordinate Plots (Optional)	208
Perform Iterative Clustering.....	211
Score New Observations.....	213
k-Means Clustering versus Hierarchical Clustering	213
Exercises	214
Chapter 10: Decision Trees	217
Introduction.....	217
Benefits and Drawbacks.....	217
Definitions and an Example	218
Theoretical Questions.....	219
Classification Trees	220
Begin Tree and Observe Results.....	220
Use JMP to Choose the Split That Maximizes the LogWorth Statistic	222

Split the Root Node According to Rank of Variables	222
Split Second Node According to the College Variable	224
Examine Results and Predict the Variable for a Third Split	227
Examine Results and Predict the Variable for a Fourth Split	227
Examine Results and Continue Splitting to Gain Actionable Insights	228
Prune to Simplify Overgrown Trees	229
Examine Receiver Operator Characteristic and Lift Curves	229
Regression Trees	231
Understand How Regression Trees Work	231
Restart a Regression Driven by Practical Questions	235
Use Column Contributions and Leaf Reports for Large Data Sets	236
Exercises	237
Chapter 11: <i>k</i>-Nearest Neighbors	241
Introduction	241
Example—Age and Income as Correlates of Purchase	241
The Way That JMP Resolves Ties	243
The Need to Standardize Units of Measurement	243
<i>k</i> -Nearest Neighbors Analysis	244
Perform the Analysis	244
Make Predictions for New Data	245
<i>k</i> -Nearest Neighbor for Multiclass Problems	247
Understand the Variables	247
Perform the Analysis and Examine Results	248
The <i>k</i> -Nearest Neighbor Regression Models	250
Perform a Linear Regression as a Basis for Comparison	250
Apply the <i>k</i> -Nearest Neighbors Technique	250
Compare the Two Methods	250
Make Predictions for New Data	254
Limitations and Drawbacks of the Technique	254
Exercises	255
Chapter 12: Neural Networks	257
Introduction	257
Drawbacks and Benefits	257
A Simplified Representation	258
A More Realistic Representation	260
Understand Validation Methods	262
Holdback Validation	262
<i>k</i> -fold Cross Validation	263
Understand the Hidden Layer Structure	264
A Few Guidelines for Determining Number of Nodes	264
Practical Strategies for Determining Number of Nodes	265
The Method of Boosting	265

Understand Options for Improving the Fit of a Model.....	266
Complete the Data Preparation.....	267
Use JMP on an Example Data Set	269
Perform a Linear Regression as a Baseline.....	269
Perform the Neural Network Ten Times to Assess Default Performance	271
Boost the Default Model.....	272
Compare Transformation of Variables and Methods of Validation.....	273
Change the Architecture	276
Predict a Binary Dependent Variable.....	277
Exercises	279
Chapter 13: Bootstrap Forests and Boosted Trees	281
Introduction.....	281
Bootstrap Forests.....	282
Understand Bagged Trees	282
Perform a Bootstrap Forest.....	283
Perform a Bootstrap Forest for Regression Trees	288
Boosted Trees	289
Understand Boosting	289
Perform Boosting	289
Perform a Boosted Tree for Regression Trees.....	292
Use Validation and Training Samples	293
Exercises	298
Chapter 14: Model Comparison	299
Introduction.....	299
Perform a Model Comparison with Continuous Dependent Variable	300
Understand Absolute Measures	300
Understand Relative Measures	300
Understand Correlation between Variable and Prediction	301
Explore the Uses of the Different Measures	301
Perform a Model Comparison with Binary Dependent Variable	304
Understand the Confusion Matrix and Its Limitations	304
Understand True Positive Rate and False Positive Rate	305
Interpret Receiving Operator Characteristic Curves	306
Compare Two Example Models Predicting Churn.....	309
Perform a Model Comparison Using the Lift Chart.....	311
Train, Validate, and Test.....	313
Perform Stepwise Regression	313
Examine the Results of Stepwise Regression	316
Compute the MSE, MAE, and Correlation	316
Examine the Results for MSE, MAE, and Correlation.....	316
Understand Overfitting from a Coin-Flip Example	317
Use the Model Comparison Platform	318

Exercises	330
Chapter 15: Text Mining.....	333
Introduction.....	333
Historical Perspective.....	333
Unstructured Data	334
Developing the Document Term Matrix	335
Understand the Tokenizing Stage	335
Understand the Phrasing Stage	343
Understand the Terming Stage	344
Observe the Order of Operations	346
Developing the Document Term Matrix with a Larger Data Set	346
Generate a Word Cloud and Examine the Text	347
Examine and Group Terms	349
Add Frequent Phrases to List of Terms	350
Parse the List of Terms	350
Using Multivariate Techniques	350
Perform Latent Semantic Analysis	352
Perform Topic Analysis.....	357
Perform Cluster Analysis	358
Using Predictive Techniques	363
Perform Primary Analysis.....	364
Perform Logistic Regressions	365
Exercises	368
Chapter 16: Market Basket Analysis.....	371
Introduction.....	371
Association Analyses	371
Examples	372
Understand Support, Confidence, and Lift	372
Association Rules	373
Support	373
Confidence.....	373
Lift	374
Use JMP to Calculate Confidence and Lift	375
Use the A Priori Algorithm for More Complex Data Sets	375
Form Rules and Calculate Confidence and Lift	376
Analyze a Real Data Set	376
Perform Association Analysis with Default Settings.....	376
Reduce the Number of Rules and Sort Them.....	377
Examine Results	377

Target Results to Take Business Actions	378
Exercises	379
Chapter 17: Time Series Forecasting	381
Introduction	381
Discovery	382
Time Series Plot	383
Trend Analysis	385
Testing for Significant Linear Trend Component	386
Seasonal Component	388
Testing for Significant Seasonal Component	389
Cyclical Component	391
Autocorrelation	392
Lagging and Differencing	397
Lagging	397
Differencing	398
Decomposition	398
Stationarity	400
Randomness	401
Simple Moving Average and Simple Exponential Smoothing Models	409
Simple Moving Average	410
Simple Exponential Smoothing	414
Forecast Performance Measures	418
Autoregressive and Moving Average Models	421
ARIMA Models	423
ARIMA Modeling with Log Variable	425
ARIMA Modeling with Seasonality	426
Advanced Exponential Smoothing Models	430
State Space Smoothing Models	434
Holdback	437
Time Series Cross-Validation	438
Time Series Forecast	440
Exercises	445
Chapter 18: Statistical Storytelling	447
The Path from Multivariate Data to the Modeling Process	447
Early Applications of Data Mining	447
Numerous JMP Customer Stories of Modern Applications	448
Definitions of Data Mining	448
Data Mining	449
Predictive Analytics	449

A Framework for Predictive Analytics Techniques	450
The Goal, Tasks, and Phases of Predictive Analytics	451
The Difference between Statistics and Data Mining	453
SEMMA	454
References	457
Index	461

About This Book

What Does This Book Cover?

This book focuses on the business statistics intelligence component of business analytics. It covers processes to perform a statistical study that might include data mining or predictive analytics techniques. Some real-world business examples of using these techniques are as follows:

- target marketing
- customer relation management
- market basket analysis
- cross-selling
- forecasting
- market segmentation
- customer retention
- improved underwriting
- quality control
- competitive analysis
- fraud detection and management
- churn analysis

Specific applications can be found at https://www JMP.com/en_my/customer-stories/customer-listing/featured.html. The bottom line, as reported by the KDNuggets poll (2008), is this: The median return on investment for data mining projects is in the 125–150% range. (See <http://www.kdnuggets.com/polls/2008/roi-data-mining.htm>.)

This book is *not* an introductory statistics book, although it does introduce basic data analysis, data visualization, and analysis of multivariate data. For the most part, your introductory statistics course has not completely prepared you to move on to real-world statistical analysis. The primary objective of this book is, therefore, to provide a bridge from your introductory statistics course to practical statistical analysis. This book is also not a highly technical book that dives deeply into the theory or algorithms, but it will provide insight into the “black box” of the methods covered. Analytics techniques covered by this book include the following:

- regression
- ANOVA
- logistic regression
- principal component analysis

- LASSO and Elastic Net
- cluster analysis
- decision trees
- *k*-nearest neighbors
- neural networks
- bootstrap forests and boosted trees
- text mining
- time series forecasting
- association rules

Is This Book for You?

This book is designed for the student who wants to prepare for his or her professional career and who recognizes the need to understand both the concepts and the mechanics of predominant analytic modeling tools for solving real-world business problems. This book is designed also for the practitioner who wants to obtain a hands-on understanding of business analytics to make better decisions from data and models, and to apply these concepts and tools to business analytics projects.

This book is for you if you want to explore the use of analytics for making better business decisions and have been either intimidated by books that focus on the technical details, or discouraged by books that focus on the high-level importance of using data without including the how-to of the methods and analysis.

Although not required, your completion of a basic course in statistics will prove helpful. Experience with the book's software, JMP Pro 17, is not required.

What's New in This Edition?

This third edition includes one new chapter on time series forecasting. All the old chapters from the second edition are updated to JMP 17. In addition, about 60% more end-of-chapter exercises are provided.

What Should You Know about the Examples?

This book includes tutorials for you to follow to gain hands-on experience with JMP.

Software Used to Develop the Book's Content

JMP Pro 17 is the software used throughout this book.

Example Code and Data

You can access the example code and data for this book by linking to its author page at <http://support.sas.com/klimberg>. Some resources, such as instructor resources and add-ins used in the book, can be found on the JMP User Community file exchange at <https://community.jmp.com>.

Where Are the Exercise Solutions?

We strongly believe that for you to obtain maximum benefit from this book you need to complete the examples in each chapter. At the end of each chapter are suggested exercises so that you can practice what has been discussed in the chapter. Professors and instructors can obtain the exercise solutions by requesting them through the author's SAS Press webpage at <http://support.sas.com/klimberg>.

We Want to Hear from You

SAS Press books are written *by* SAS Users *for* SAS Users. We welcome your participation in their development and your feedback on SAS Press books that you are using. Please visit sas.com/books.

About The Author



Ron Klimberg, PhD, is a professor at the Haub School of Business at Saint Joseph's University in Philadelphia, PA. Before joining the faculty in 1997, he was a professor at Boston University, an operations research analyst at the U.S. Food and Drug Administration, and an independent consultant. His current primary interests include multiple criteria decision making, data envelopment analysis, data visualization, data mining, and modeling in general. Klimberg was the 2007 recipient of the Tengelman Award for excellence in scholarship, teaching, and research. He received his PhD from Johns Hopkins University and his MS from George Washington University.

Learn more about the author by visiting his author page, where you can download free book excerpts, access example code and data, read the latest reviews, get updates, and more: <http://support.sas.com/klimberg>.

Acknowledgments

I would like to thank Catherine Connolly and Suzanne Morgen of SAS Press for providing editorial project support from start to finish.

I want to thank Mia Stephens, Dan Obermiller, Adam Morris, Sue Walsh, and Sarah Mikol, as well as Russell Lavery, Majid Nabavi, and Donald N. Stengel, for their detailed reviews on previous editions, which improved the final product. I would like to thank Daniel Valente and Christopher Gotwalt of SAS Institute for their guidance and insight in writing the text mining chapter. Thank you also to Peng Liu and Jian Cao for their review of the time series forecasting chapter.

Dedication

This third edition of the book is dedicated to B. D. (Bruce) McCullough. Bruce and I wrote the first two editions of this book. Bruce sadly passed away September 2020 of complications from cancer. Bruce was a good colleague, friend, husband, and father.

Chapter 1: Introduction

Historical Perspective

In 1981, Bill Gates made his infamous statement that “640KB ought to be enough for anybody” (Lai, 2008).

Looking back even further, about 10 to 15 years before Bill Gates’s statement, we were in the middle of the Vietnam War era. State-of-the-art computer technology for both commercial and scientific areas at that time was the mainframe computer. A typical mainframe computer weighed tons, took an entire floor of a building, had to be air-conditioned, and cost about \$3 million. Mainframe memory was approximately 512 KB with disk space of about 352 MB and speed up to 1 MIPS (million instructions per second).

In 2016, only 45 years later, an iPhone 6 with 32-GB memory has about 9300% more memory than the mainframe and can fit in a hand. A laptop with the Intel Core i7 processor has speeds up to 238,310 MIPS, about 240,000 times faster than the old mainframe, and weighs less than 4 pounds. Further, an iPhone or a laptop cost significantly less than \$3 million. As Ray Kurzweil, an author, inventor, and futurist has stated (Lomas, 2008): “The computer in your cell phone today is a million times cheaper and a thousand times more powerful and about a hundred thousand times smaller (than the one computer at MIT in 1965) and so that’s a billion-fold increase in capability per dollar or per euro that we’ve actually seen in the last 40 years.” Technology has certainly changed!

Then in 2019, the Covid-19 pandemic turned our world upside down. The two major keys to many companies’ survival have been the ability to embrace technology and analytics, perhaps quicker than planned, and the ability to think outside the box. Before the Covid-19 pandemic, the statement was “we will see more change in the next five years than there have been in the last 50 years.” The pandemic has accelerated this change such that many of these changes will now occur in the next two to three years. Companies that take full advantage of new technology and analytics and find their distinct capability will have a competitive advantage to succeed.

Two Questions Organizations Need to Ask

Many organizations have realized or are just now starting to realize the importance of using analytics. One of the first strides an organization should take toward becoming an analytical competitor is to ask themselves the following two questions:

- With the huge investment in collecting data, do organizations get a decent return on investment (ROI)?
- What are your organization's two most important assets?

Return on Investment

With this new and ever-improving technology, most organizations (and even small organizations) are collecting an enormous amount of data. Each department has one or more computer systems. Many organizations are now integrating these department-level systems with organization systems, such as an enterprise resource planning (ERP) system. Newer systems are being deployed that store all these historical enterprise data in what is called a data warehouse. The IT budget for most organizations is a significant percentage of the organization's overall budget and is growing. The question is as follows:

With the huge investment in collecting this data, do organizations get a decent return on investment (ROI)?

The answer: mixed. No matter if the organization is large or small, only a limited number of organizations (yet growing in number) are using their data extensively. Meanwhile, most organizations are drowning in their data and struggling to gain some knowledge from it.

Cultural Change

How would managers respond to this question:

What are your organization's two most important assets?

Most managers would answer with their employees and the product or service that the organization provides (they might alternate which is first or second).

The follow-up question is more challenging: Given the first two most important assets of most organizations, what is the third most important asset of most organizations?

The actual answer is "the organization's data!" But to most managers, regardless of the size of their organizations, this answer would be a surprise. However, consider the vast amount of knowledge that's contained in customer or internal data. For many organizations, realizing and accepting that their data is the third most important asset would require a significant cultural change.

Rushing to the rescue in many organizations is the development of business intelligence (BI) and business analytics (BA) departments and initiatives. What is BI? What is BA? The answers seem to vary greatly depending on your background.

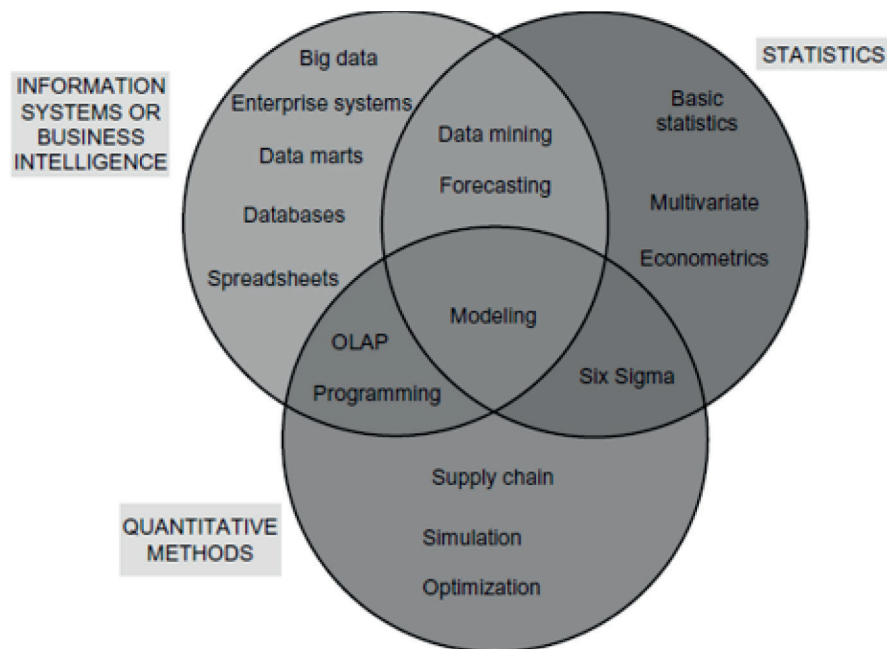
Business Intelligence and Business Analytics

Business intelligence (BI) and business analytics (BA) are considered by most people as providing information technology systems, such as dashboards and online analytical processing (OLAP) reports, to improve business decision-making. An expanded definition of BI is that it is a “broad category of applications and technologies for gathering, storing, analyzing, and providing access to data to help enterprise users make better business decisions. BI applications include the activities of decision support systems, query and reporting, online analytical processing (OLAP), statistical analysis, forecasting, and data mining” (Rahman, 2009).

The scope of BI and its growing applications have revitalized an old term: *business analytics* (BA). Davenport (Davenport and Harris, 2007) views BA as “the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions.” Davenport further elaborates that organizations should develop an analytics competency as a “distinctive business capability” that would provide the organization with a competitive advantage.

In 2007, BA was viewed as a subset of BI. However, in recent years, this view has changed. Today, BA is viewed as including BI’s core functions of reporting, OLAP and descriptive statistics, as well as the advanced analytics of data mining, forecasting, simulation, and optimization. Figure 1.1 presents a framework (adapted from Klimberg and Miori, 2010) that embraces this

Figure 1.1: A Framework of Business Analytics



expanded definition of BA (or simply analytics) and shows the relationship of its three disciplines (Information Systems/Business Intelligence, Statistics, and Operations Research) (Gorman and Klimberg, 2014). The Institute of Operations Research and Management Science (INFORMS), one of the largest professional and academic organizations in the field of analytics, breaks analytics into three categories:

- Descriptive analytics: provides insights into the past by using tools such as queries, reports, and descriptive statistics,
- Predictive analytics: understand the future by using predictive modeling, forecasting, and simulation,
- Prescriptive analytics: provide advice on future decisions using optimization.

The buzzword in this area of analytics for about the last 25 years has been *data mining*. Data mining is the process of finding patterns in data, usually using some advanced statistical techniques. The current buzzwords are *predictive analytics* and *predictive modeling*. What is the difference in these three terms? As discussed, with the many and evolving definitions of business intelligence, these terms seem to have many different yet quite similar definitions. Chapter 18 briefly discusses their different definitions. This text, however, generally will not distinguish between *data mining*, *predictive analytics*, and *predictive modeling* and will use them interchangeably to mean or imply the same thing.

Most of the terms mentioned here include the adjective *business* (as in *business intelligence* and *business analytics*). Even so, the application of the techniques and tools can be applied outside the business world and are used in the public and social sectors. In general, wherever data is collected, these tools and techniques can be applied.

Introductory Statistics Courses

Most introductory statistics courses (outside the mathematics department) cover the following topics:

- descriptive statistics
- probability
- probability distributions (discrete and continuous)
- sampling distribution of the mean
- confidence intervals
- one-sample hypothesis testing

They might also cover the following:

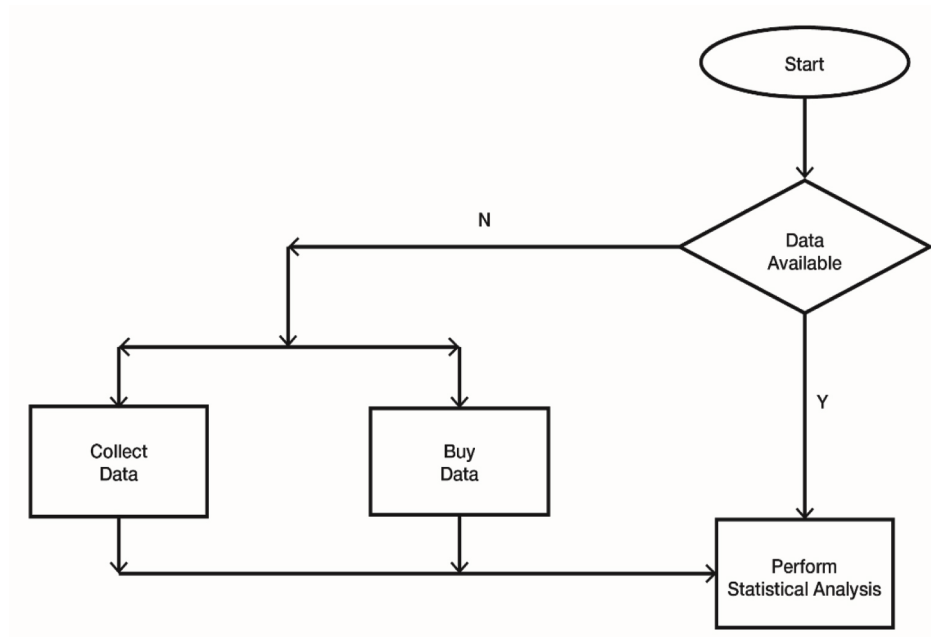
- two-sample hypothesis testing
- simple linear regression
- multiple linear regression
- analysis of variance (ANOVA)

Yes, multiple linear regression and ANOVA are multivariate techniques. But the complexity of the multivariate nature is for the most part not addressed in the introduction to statistics course. One main reason—not enough time!

Nearly all the topics, problems, and examples in the course are directed toward univariate (one variable) or bivariate (two variables) analysis. Univariate analysis includes techniques to summarize the variable and make statistical inferences from the data to a population parameter. Bivariate analysis examines the relationship between two variables (for example, the relationship between age and weight).

A typical student's understanding of the components of a statistical study is shown in Figure 1.2. If the data are not available, a survey is performed or the data are purchased. Once the data are obtained, all at one time, the statistical analyses are done—using Excel or a statistical package, drawing the appropriate graphs and tables, performing all the necessary statistical tests, and writing up or otherwise presenting the results. And then you are done. With such a perspective, many students simply look at this statistics course as another math course and might not realize the importance and consequences of the material.

Figure 1.2: A Student's View of a Statistical Study from a Basic Statistics Course



The Problem of Dirty Data

Although these first statistics courses provide a good foundation in introductory statistics, they provide a rather weak foundation for performing practical statistical studies. First, most real-world data are “dirty.” *Dirty data* are erroneous data, missing values, incomplete records, and the like. For example, suppose a data field or variable that represents gender is supposed to be coded as either M or F. If you find the letter N in the field or even a blank instead, then you have dirty data. Learning to identify dirty data and to determine corrective action are fundamental skills needed to analyze real-world data. Chapter 3 will discuss dirty data in detail.

Added Complexities in Multivariate Analysis

Second, most practical statistical studies have data sets that include more than two variables, called multivariate data. Multivariate analysis uses some of the same techniques and tools used in univariate and bivariate analysis as covered in the introductory statistics courses, but in an expanded and much more complex manner. Also, when performing multivariate analysis, you are exploring the relationships among several variables. There are several multivariate statistical techniques and tools to consider that are not covered in a basic applied statistics course.

Before jumping into multivariate techniques and tools, students need to learn the univariate and bivariate techniques and tools that are taught in the basic first statistics course. However, in some programs this basic introductory statistics class might be the last data analysis course required or offered. In many other programs that do offer or require a second statistics course, these courses are just a continuation of the first course, which might or might not cover ANOVA and multiple linear regression. (Although ANOVA and multiple linear regression are multivariate, this reference is to a second statistics course beyond these topics.) In either case, the students are ill-prepared to apply statistics tools to real-world multivariate data. Perhaps, with some minor adjustments, real-world statistical analysis can be introduced into these programs.

On the other hand, with the growing interest in BI, BA, and predictive analytics, more programs are offering and sometimes even requiring a subsequent statistics course in predictive analytics. So, most students jump from univariate/bivariate statistical analysis to statistical predictive analytics techniques, which include numerous variables and records. These statistical predictive analytics techniques require the student to understand the fundamental principles of multivariate statistical analysis and, more so, to understand the process of a statistical study. In this situation, many students are lost, which simply reinforces the students’ view that the course is just another math course.

Practical Statistical Study

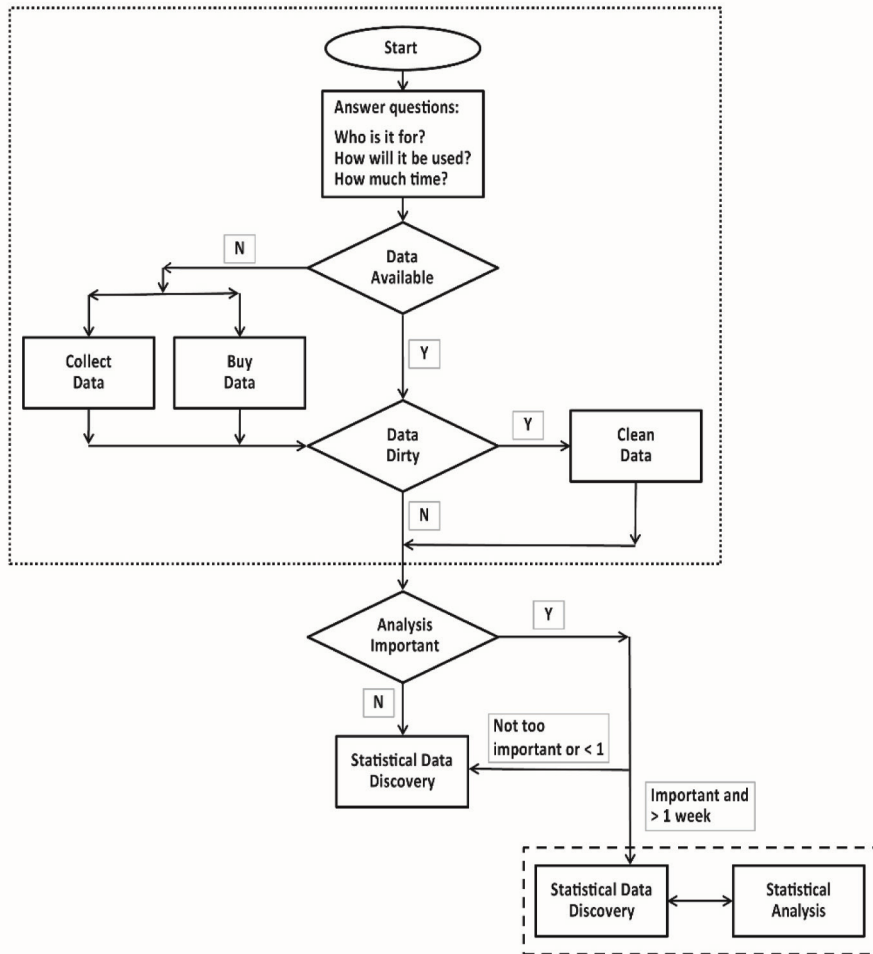
Even with these ill-prepared multivariate shortcomings, there is still a more significant concern to address: the idea that most students view statistical analysis as a straightforward exercise in

which you sit down once in front of your computer and just perform the necessary statistical techniques and tools, as in Figure 1.2. How boring! With such a viewpoint, this would be like telling someone that reading a book can simply be done by reading the book cover. The practical statistical study process of uncovering the *story* behind the data is what makes the work exciting.

Obtaining and Cleaning the Data

The prologue to a practical statistical study is determining the proper data needed, obtaining the data, and if necessary, cleaning the data (the dotted area in Figure 1.3). Answering the questions “Who is it for?” and “How will it be used?” will identify the suitable variables required and the

Figure 1.3: The Flow of a Real-World Statistical Study



appropriate level of detail. Who will use the results and how they will use them determine which variables are necessary and the level of granularity. If there is enough time and the essential data is not available, then the data might have to be obtained by a survey, purchasing it, through an experiment, compiled from different systems or databases, or other possible sources. Once the data is available, most likely the data will first have to be cleaned—in essence, eliminating erroneous data as much as possible. Various manipulations will prepare the data for analysis, such as creating new derived variables, data transformations, and changing the units of measuring. Also, the data might need to be aggregated or compiled in various ways. These preliminary steps account for about 75% of the time of a statistical study and are discussed further in Chapter 18.

As shown in Figure 1.3, the importance placed on the statistical study by the decision-makers/users and the amount of time allotted for the study will determine whether the study will be only a *statistical data discovery* or a more complete *statistical analysis*. *Statistical data discovery* is the discovery of significant and insignificant relationships among the variables and the observations in the data set.

Understanding the Statistical Study as a Story

The *statistical analysis* (the enclosed dashed-line area in Figure 1.3) should be read like a book—the data should tell a story. The first part of the story and continuing throughout the study is the *statistical data discovery*.

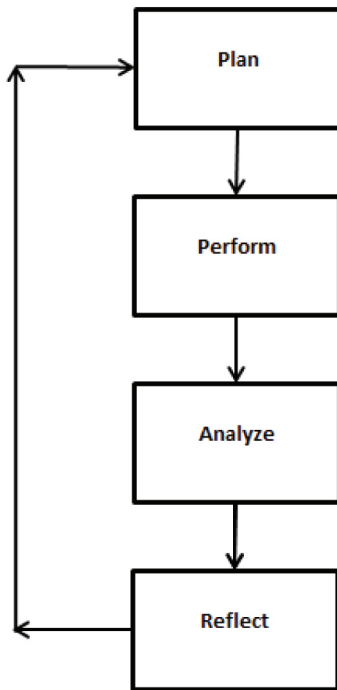
The story develops further as many different statistical techniques and tools are tried. Some will be helpful, some will not. With each iteration of applying the statistical techniques and tools, the story develops and is substantially further advanced when you relate the statistical results to the actual problem situation. As a result, your understanding of the problem and how it relates to the organization is improved. By doing the statistical analysis, you will make better decisions (most of the time). Furthermore, these decisions will be more informed so that you will be more confident in your decision. Finally, uncovering and telling this statistical story is fun!

The Plan-Perform-Analyze-Reflect Cycle

The development of the statistical story follows a process that is called here the *plan-perform-analyze-reflect* (PPAR) cycle, as shown in Figure 1.4. The PPAR cycle is an iterative progression.

The first step is to plan which statistical techniques or tools are to be applied. You are combining your statistical knowledge and your understanding of the business problem being addressed. You are asking pointed, directed questions to answer the business question by identifying a particular statistical tool or technique to use.

The second step is to perform the statistical analysis, using statistical software such as JMP.

Figure 1.4: The PPAR Cycle

The third step is to analyze the results using appropriate statistical tests and other relevant criteria to evaluate the results. The fourth step is to reflect on the statistical results. Ask questions like what do the statistical results mean in terms of the problem situation? What insights have I gained? Can you draw any conclusions? Sometimes the results are extremely useful, sometimes meaningless, and sometimes in the middle—a potential significant relationship.

Then, it is back to the first step to plan what to do next. Each progressive iteration provides a little more to the story of the problem situation. This cycle continues until you feel you have exhausted all possible statistical techniques or tools (visualization, univariate, bivariate, and multivariate statistical techniques) to apply, or you have results sufficient to consider the story completed.

Using Powerful Software

The software used in many initial statistics courses is Microsoft Excel, which is easily accessible and provides some basic statistical capabilities. However, as you advance through the course, because of Excel's statistical limitations, you might also use some nonprofessional, textbook-specific statistical software or perhaps some professional statistical software. Excel is not a professional statistics software application; it is a spreadsheet.