Contents at a Glance

Introduction ................................................................. 1
1 Components of Decision Analytics .............................. 5
2 Logistic Regression ...................................................... 21
3 Univariate Analysis of Variance (ANOVA) ...................... 53
4 Multivariate Analysis of Variance (MANOVA) .................. 89
5 Discriminant Function Analysis: The Basics............... 119
6 Discriminant Function Analysis: Further Issues ............ 139
7 Principal Components .................................................. 173
8 Cluster Analysis: The Basics ....................................... 209
9 Cluster Analysis: Further Issues .................................. 235
Index ........................................................................... 259

Decision Analytics: Microsoft® Excel®

Conrad Carlberg
Table of Contents

Introduction ............................................................... 1
  What's in the Book .................................................... 1
  Why Use Excel? ....................................................... 3

1 Components of Decision Analytics .................................. 5
  Classifying According to Existing Categories ........................ 5
    Using a Two-Step Approach ......................................... 6
    Multiple Regression and Decision Analytics ...................... 6
    Access to a Reference Sample ..................................... 8
    Multivariate Analysis of Variance ................................ 9
    Discriminant Function Analysis .................................. 10
    Logistic Regression .............................................. 12
  Classifying According to Naturally Occurring Clusters .......... 13
    Principal Components Analysis .................................. 13
    Cluster Analysis .................................................. 14
  Some Terminology Problems ......................................... 16
    The Design Sets the Terms ....................................... 17
    Causation Versus Prediction ...................................... 18
    Why the Terms Matter ........................................... 18

2 Logistic Regression .................................................. 21
  The Rationale for Logistic Regression ................................ 22
    The Scaling Problem ............................................... 24
    About Underlying Assumptions ................................... 25
    Equal Spread ..................................................... 25
    Equal Variances with Dichotomies ................................ 27
    Equal Spread and the Range ..................................... 28
  The Distribution of the Residuals .................................. 29
    Calculating the Residuals ......................................... 30
    The Residuals of a Dichotomy .................................... 30
  Using Logistic Regression ........................................... 31
    Using Odds Rather Than Probabilities ............................ 32
    Using Log Odds .................................................... 33
    Using Maximum Likelihood Instead of Least Squares .......... 34
  Maximizing the Log Likelihood ..................................... 35
    Setting Up the Data ............................................... 35
    Setting Up the Logistic Regression Equation ..................... 36
    Getting the Odds .................................................. 38
    Getting the Probabilities ........................................ 39
    Calculating the Log Likelihood .................................. 40
    Finding and Installing Solver .................................... 41
    Running Solver .................................................... 41

About Underlying Assumptions ........................................ 25

Discriminant Function Analysis ...................................... 10

Access to a Reference Sample ....................................... 8

Cluster Analysis ..................................................... 14

Causation Versus Prediction ......................................... 18

Finding and Installing Solver ....................................... 41

Equal Variances with Dichotomies .................................. 27

Equal Spread ......................................................... 25

About Underlying Assumptions ....................................... 24

Logistic Regression ................................................... 12

Maximizing the Log Likelihood ...................................... 35

Multivariate Analysis of Variance .................................. 9

Using Odds Rather Than Probabilities ................................ 32
3 Univariate Analysis of Variance (ANOVA) ........................................... 53
   The Rationale for Log Likelihood .................................................. 43
   The Probability of a Correct Classification .................................. 44
   Using the Log Likelihood ......................................................... 45
   The Statistical Significance of the Log Likelihood ......................... 48
   Setting Up the Reduced Model .................................................. 50
   Setting Up the Full Model ........................................................ 51
   Single Factor ANOVA .............................................................. 66
      Adopting an Error Rate ....................................................... 66
      Computing the Statistics ..................................................... 67
      Deriving the Standard Error of the Mean .................................. 70
   Using the Data Analysis Add-In ................................................ 72
      Installing the Data Analysis Add-In ........................................ 73
      Using the ANOVA: Single Factor Tool .................................... 73
   Understanding the ANOVA Output .............................................. 75
      Using the Descriptive Statistics ............................................ 75
      Using the Inferential Statistics ............................................. 76
   The Regression Approach .......................................................... 79
      Using Effect Coding ............................................................ 80
      The LINEST() Formula ....................................................... 82
      The LINEST() Results ....................................................... 83
      LINEST() Inferential Statistics ............................................. 85
   Expected Values of Variances (Within Groups) ................................ 56
   Expected Values of Variances (Between Groups) ............................. 58
   The F-Ratio ............................................................................. 61
   The Noncentral F Distribution .................................................... 64
   Using the Log Likelihood ............................................................ 45
   Setting Up the Reduced Model .................................................. 50
   Setting Up the Full Model ........................................................ 51

4 Multivariate Analysis of Variance (MANOVA) ................................. 89
   The Rationale for MANOVA ....................................................... 89
   Correlated Variables ................................................................. 90
   Correlated Variables in ANOVA ................................................ 91
   Visualizing Multivariate ANOVA ................................................ 92
   Univariate ANOVA Results ...................................................... 93
   Multivariate ANOVA Results ................................................... 93
   Means and Centroids ............................................................... 95
   From ANOVA to MANOVA ....................................................... 96
   Using SSCP Instead of SS ......................................................... 98
   Getting the Among and the Within SSCP Matrices ......................... 102
   Sums of Squares and SSCP Matrices ........................................ 104
   Getting to a Multivariate F-Ratio ............................................... 105
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilks' Lambda and the F-Ratio</td>
<td>107</td>
</tr>
<tr>
<td>Converting Wilks' Lambda to an F Value</td>
<td>108</td>
</tr>
<tr>
<td>Running a MANOVA in Excel</td>
<td>110</td>
</tr>
<tr>
<td>Laying Out the Data</td>
<td>110</td>
</tr>
<tr>
<td>Running the MANOVA Code</td>
<td>111</td>
</tr>
<tr>
<td>Descriptive Statistics</td>
<td>112</td>
</tr>
<tr>
<td>Equality of the Dispersion Matrices</td>
<td>113</td>
</tr>
<tr>
<td>The Univariate and Multivariate F-Tests</td>
<td>115</td>
</tr>
<tr>
<td>After the Multivariate Test</td>
<td>116</td>
</tr>
</tbody>
</table>

### 5 Discriminant Function Analysis: The Basics  

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treating a Category as a Number</td>
<td>120</td>
</tr>
<tr>
<td>The Rationale for Discriminant Analysis</td>
<td>122</td>
</tr>
<tr>
<td>Multiple Regression and Discriminant Analysis</td>
<td>122</td>
</tr>
<tr>
<td>Adjusting Your Viewpoint</td>
<td>123</td>
</tr>
<tr>
<td>Discriminant Analysis and Multiple Regression</td>
<td>125</td>
</tr>
<tr>
<td>Regression, Discriminant Analysis, and Canonical Correlation</td>
<td>125</td>
</tr>
<tr>
<td>Coding and Multiple Regression</td>
<td>127</td>
</tr>
<tr>
<td>The Discriminant Function and the Regression Equation</td>
<td>129</td>
</tr>
<tr>
<td>From Discriminant Weights to Regression Coefficients</td>
<td>130</td>
</tr>
<tr>
<td>Eigenstructures from Regression and Discriminant Analysis</td>
<td>133</td>
</tr>
<tr>
<td>Structure Coefficients Can Mislead</td>
<td>136</td>
</tr>
<tr>
<td>Wrapping It Up</td>
<td>137</td>
</tr>
</tbody>
</table>

### 6 Discriminant Function Analysis: Further Issues  

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using the Discriminant Workbook</td>
<td>139</td>
</tr>
<tr>
<td>Opening the Discriminant Workbook</td>
<td>140</td>
</tr>
<tr>
<td>Using the Discriminant Dialog Box</td>
<td>141</td>
</tr>
<tr>
<td>Why Run a Discriminant Analysis on Irises?</td>
<td>144</td>
</tr>
<tr>
<td>Evaluating the Original Measures</td>
<td>144</td>
</tr>
<tr>
<td>Discriminant Analysis and Investment</td>
<td>145</td>
</tr>
<tr>
<td>Benchmarking with R</td>
<td>147</td>
</tr>
<tr>
<td>Downloading R</td>
<td>147</td>
</tr>
<tr>
<td>Arranging the Data File</td>
<td>148</td>
</tr>
<tr>
<td>Running the Analysis</td>
<td>149</td>
</tr>
<tr>
<td>The Results of the Discrim Add-In</td>
<td>152</td>
</tr>
<tr>
<td>The Discriminant Results</td>
<td>153</td>
</tr>
<tr>
<td>Interpreting the Structure Coefficients</td>
<td>155</td>
</tr>
<tr>
<td>Eigenstructures and Coefficients</td>
<td>156</td>
</tr>
<tr>
<td>Other Uses for the Coefficients</td>
<td>159</td>
</tr>
<tr>
<td>Classifying the Cases</td>
<td>162</td>
</tr>
<tr>
<td>Distance from the Centroids</td>
<td>163</td>
</tr>
<tr>
<td>Correcting for the Means</td>
<td>164</td>
</tr>
</tbody>
</table>
7 Principal Components Analysis .......................................................... 173
   Estimating a Conceptual Framework for Principal Components Analysis. ........................................ 174
   Principal Components and Tests ........................................... 174
   PCA’s Ground Rules .......................................................... 175
   Correlation and Oblique Factor Rotation ................................. 176
   Using the Principal Components Add-In ................................ 177
   The Correlation Matrix ....................................................... 179
   The Inverse of the R Matrix .............................................. 179
   The Sphericity Test ............................................................. 179
   Counting Eigenvalues, Calculating Coefficients and Understanding Communalities ............................. 183
   How Many Components? ..................................................... 184
   Factor Score Coefficients .................................................. 186
   Communalities ................................................................. 186
   Relationships Between the Individual Results ....................... 187
   Using the Eigenvalues and Eigenvectors ............................... 187
   Eigenvalues, Eigenvectors, and Loadings ................................ 188
   Eigenvalues, Eigenvectors, and Factor Coefficients .................. 190
   Getting the Eigenvalues Directly from the Factor Scores ........... 191
   Getting the Eigenvalues and Eigenvectors ............................ 192
   Iteration and Exhaustion .................................................... 193
   Rotating Factors to a Meaningful Solution ............................ 196
   Identifying the Factors ...................................................... 197
   The Varimax Rotation ......................................................... 200
   Classification Examples ................................................... 202
   State Crime Rates ............................................................... 202
   Physical Measurements of Aphids ........................................ 206

8 Cluster Analysis: The Basics .......................................................... 209
   Cluster Analysis, Discriminant Analysis, and Logistic Regression ...................................................... 209
   Euclidean Distance .............................................................. 211
   Mahalanobis’ D² and Cluster Analysis ................................. 214
   Finding Clusters: The Single Linkage Method ....................... 215
   The Self-Selecting Nature of Cluster Analysis ....................... 220
   Finding Clusters: The Complete Linkage Method ................. 223
   Complete Linkage: An Example ........................................... 224
   Other Linkage Methods ....................................................... 227
Contents

Finding Clusters: The K-means Method .......................................................... 228
Characteristics of K-means Analysis ................................................................. 228
A K-means Example ......................................................................................... 229
Benchmarking K-means with R ......................................................................... 233

9 Cluster Analysis: Further Issues ................................................................. 235

Using the K-means Workbook ....................................................................... 235
Deciding on the Number of Clusters ............................................................. 237
The Cluster Members Worksheet ................................................................. 239
The Cluster Centroids Worksheet ................................................................. 241
The Cluster Variances Worksheet ................................................................. 242
The F-Ratios Worksheet ................................................................................. 244
Reporting Process Statistics ....................................................................... 247

Cluster Analysis Using Principal Components ......................................... 248
Principal Components Revisited .................................................................. 249
Clustering Wines ......................................................................................... 253
Cross-Validating the Results ....................................................................... 256

Index ........................................................................................................... 259
About the Author

Conrad Carlberg lives near San Diego with his wife, not too far from the beach, but high enough that the rise in the sea level is unlikely to convert their home to waterfront property. Two cats round out the indoor menagerie; the three rabbits are required to stay outside.
Dedication

Once again, to my beloved and beautiful Toni, who laughs with me.

Acknowledgments

My thanks once again to Loretta Yates for skillfully directing this book through the acquisition process, and for her steady hand and even temper. Michael Turner, of the University of Colorado's REM lab, provided a fine technical edit, one that kept me from embarrassing myself in print and that got me back on track when necessary. Anne Jones...what can I say? Her brilliant cover artwork for Statistical Analysis: Microsoft Excel 2010 and for Predictive Analytics: Microsoft Excel surely drove as many sales as the contents did, and she's done it again. Geneil Breeze's copy edit gently reminded me to leave the pedantry behind, in the first draft where it belongs. Elaine Wiley, for shepherding the project through all its incarnations. My heartfelt thanks to each of you.
We Want to Hear from You!

As the reader of this book, you are our most important critic and commentator. We value your opinion and want to know what we’re doing right, what we could do better, what areas you’d like to see us publish in, and any other words of wisdom you’re willing to pass our way.

We welcome your comments. You can email or write to let us know what you did or didn’t like about this book—as well as what we can do to make our books better.

Please note that we cannot help you with technical problems related to the topic of this book.

When you write, please be sure to include this book’s title and author as well as your name and email address. We will carefully review your comments and share them with the author and editors who worked on the book.

Email: feedback@quepublishing.com
Mail: Que Publishing
      ATTN: Reader Feedback
      800 East 96th Street
      Indianapolis, IN 46240 USA

Reader Services

Visit our website and register this book at quepublishing.com/register for convenient access to any updates, downloads, or errata that might be available for this book.
Introduction

I’d like to get something off my chest right away: Excel workbooks are waiting for you to download them from the publisher’s website, quepublishing.com/title/9780789751683. The download links are sometimes a little tough to distinguish from regular text, but both the links and the workbooks are there. Starting with Chapter 2, “Logistic Regression,” each chapter in this book has its own Excel workbook, and each figure in each chapter appears as a separate worksheet. A few additional workbooks carry out cluster analysis, discriminant function analysis, and other analytics procedures that don’t have their own worksheet functions.

Okay, with that item out of the way:

This is not a book about acquiring, storing, and partitioning so-called big data. This book is about how to make sense of the numbers, be they big data or small beer.

We’ve all been confronted by situations in which we have, say, 30 variables to deal with, each potentially important and each with a slightly different slant on the really interesting phenomenon, whether that’s 12-month survival, or probability of profiting from an investment, or seeing a new hire make good. Regardless of whether you have 200 or 200,000 records at your disposal, the real question is how to handle those 30 variables. How to combine or discard them so as to make the right decisions about the effects of a medication, about whether to extend the financing, or about which candidate to hire.

What’s in the Book

That’s what Decision Analytics: Microsoft Excel is about: finding the best mix of the variables you have at hand, so that your decision is as informed as you can make it.
That's an exercise in using quantitative classification techniques, and there are several, as follows.

*Discriminant function analysis* has a long and generally honorable history. It's been used for a broad array of purposes, from identifying the party affiliations of nineteenth century politicians on the basis of their legislative records, to flagging possibly fraudulent Form 1040s based on patterns of deductions and adjustments. Chapter 5, “Discriminant Function Analysis: The Basics,” and Chapter 6, “Discriminant Function Analysis: Further Issues,” of this book walk you through this sort of analysis and explore the data reduction techniques involved. They show you, in the context of both worksheets and charts, how discriminant function analysis goes about its business.

Because discriminant analysis depends on a multivariate approach to handling the continuous variables, I have included Chapter 7, “Principal Components Analysis.” This helps you get your arms around concepts such as eigenvalues and eigenvectors as they pertain to correlation matrixes—again, in the familiar context of Excel worksheets and charts.

There's also a workbook for you to download with VBA code that runs a complete discriminant function analysis and outputs the significance tests, the function coefficients, canonical correlations, and other bells and whistles, which are each explained in the text and further illustrated in the chapter's workbooks.

The best way to approach a discriminant analysis is by way of multivariate analysis of variance, or *MANOVA*. As you'll see, MANOVA helps you decide whether it even makes sense to carry out a discriminant function analysis—whether the correlations between the dependent variables, and their ability to distinguish between different groupings of people and actions, support further analysis at all. Therefore, Chapter 4, “Multivariate Analysis of Variance (MANOVA),” discusses MANOVA, and you can download a separate workbook that runs a one-factor MANOVA with multiple dependent variables.

If it's been a while since you gave any thought to either ANOVA or MANOVA, you might want to run through Chapter 3, “Univariate Analysis of Variance (ANOVA).” As background to MANOVA, it's helpful to see, in the context of a worksheet, what's going on with the way ANOVA manages variability.

Besides discriminant function analysis, another method of classifying people or market activities (or politicians, or houseplants) is by way of *logistic regression*. That's a useful method, and it avoids some of the pitfalls that discriminant analysis might put in your way. For example, logistic regression doesn't make all the assumptions about how the data is distributed that discriminant analysis does. So if you're concerned that your data violates those assumptions (and honestly, it doesn't necessarily invalidate your analysis even if the assumptions get trampled), you can often use logistic regression instead, as the analytic basis for your decisions.

On the other hand, those assumptions are what give discriminant analysis its statistical power—its ability to successfully and reliably distinguish between different groupings of subjects. Other things being equal, discriminant analysis is a more sensitive guide to classification than is logistic regression.
I give two chapters to logistic regression in my previous book *Predictive Analytics: Microsoft Excel*. I cover it here in Chapter 2, “Logistic Regression,” more as a review than as a full discussion.

Chapter 8, “Cluster Analysis: The Basics,” and Chapter 9, “Cluster Analysis: Further Issues,” go into yet another approach to decision analytics. In logistic regression and discriminant analysis, you know going in what your groups are. You have a sample of data, be it large or small, with observations that include group membership (survives/doesn’t, makes a profit/doesn’t, wins/loses) and variables that you hope will position you to make good decisions (demographic data, financials, purchasing history).

But in cluster analysis you don’t know what your groups are. You have a set of, for example, demographic variables and you’d like to know how they classify people. You turn one of the variations of cluster analysis loose on your data set, hoping that it will cluster those people such that differences on demographics will be relatively small within clusters, and relatively large between clusters.

Leland Wilkinson gave an apt description of this sort of decision analysis way back in 1986, when he wrote, “In rough terms, it is like doing a one-way analysis of variance where the groups are unknown and the largest \( F \)-value is sought by reassigning members to each group.” (See page 1 of the Cluster section of the SYSTAT manual.)

### Why Use Excel?

I write books and people buy them, thanks be. But I’m also a consultant and I like my clients to have an understanding of what I’ve done with the numbers they’ve handed me. I think that’s one of the main reasons that I’m still in this dodge after 20 years.

I don’t like to hand clients a pile of R printouts, whether literally or electronically. Nothing against R. It’s a fine set of statistical procedures even if its documentation is impenetrable and its results look like they were laid out by archaic FORTRAN. I often use R to benchmark work that I’ve done in Excel.

SAS, SPSS, Stata, and similar packages are much better documented than is R, and the analytic results are laid out in a much more straightforward fashion. They are costly, though. And, like R, they take a fair amount of study to learn the proper way to navigate the user interface and to handle the command syntax correctly.

In contrast, most of my clients are perfectly comfortable in the familiar Excel environment and appreciate how easy it is to view a set of numbers in an Excel chart. And of course you’d be hard put to find a Windows box in a corporate or educational environment that doesn’t already have some version of Excel up and running.

But is Excel capable of handling the complicated data reduction methods required by analytics in general and decision analytics in particular? Obviously, I believe it is. It’s true that Excel is a general numeric analysis package, not built from the ground up to offer specialized statistical functionality. There’s no WILKS() worksheet function.
Excel does offer an MDETERM() worksheet function, and if you point it at a Within matrix and a Total matrix then you’ve got your Wilks’ lambda. Suppose that you’re a budding analyst. Or suppose that you’re a corporate suit who wants to know why someone thinks that Wilks’ lambda says to stay out of a given line of business. I contend that either way you’re much better off knowing why it’s telling you something than just knowing what the numeric value is.

And Excel, used properly, can position you to know those things. Sometimes all you need is the built-in worksheet functions; it’s entirely possible to do a multivariate analysis of variance directly on the worksheet, without resorting to any add-ins. I think it’s a good idea to do that once or twice because it helps to cement the concepts in place.

But sometimes you need an assist from a tool such as Excel’s Solver, an add-in that comes with the Excel application software. Chapter 2 shows you how to use Solver to help you complete a logistic regression (and by the way, the purely statistical packages use the same optimization algorithms—they’re just tucked out of the way so you don’t see them).

And there are some processes, such as finding the eigenvalues of a large correlation matrix, that are just so complex and loop-dependent that it’s crazy to try to do them without coded procedures. But you’ll find those procedures, coded in VBA and some open for your inspection, in the Excel files available with this book.

To underscore my point: In early 2013 I was approached to help a company build a model that would evaluate prospective investments. The client had almost 100,000 records to use in developing the model. The nature of the data was such that a logistic regression analysis was called for, and the client supplied the data in a text file that Excel could read easily. Using Excel, I tried running a logistic regression on that data set and my formulas resulted in arithmetic underflows.

Then, using R, I put the data through a logistic regression routine that’s part of R’s library. I got underflows again. Too many cases, and intermediate results too small for either Excel or R to handle accurately.

Now, this didn’t pose any real problem. I wanted to save some of the data for cross-validation purposes, so I ran a random half of the data set through R’s logistic regression routine, got my results, and validated them on the remaining half of the data. Then I confirmed the results using Excel. The client got both the R and the Excel results, of course, but I noticed that subsequent work that the client did with the model took place in Excel, where the results of the formulas were much more transparent.

Now, this homely little story is not only anecdotal, it’s also a sample of 1. Nevertheless, it’s true, and it’s typical of my own experience with using Excel as an analytical engine. If you try out the methods I describe in this book, I’m confident you’ll come to the same conclusion.

Enough said. I suggest that you pour yourself a serving of your beverage of choice, fire up the laptop, and move on to Chapter 1, “Components of Decision Analytics.”
Components of Decision Analytics

Regardless of what line of work we’re in, we make decisions about people, medical treatments, marketing programs, soil amendments, and so on. If we’re to make informed, sensible decisions, we need to understand how to find clusters of people who are likely or unlikely to succeed in a job; how to classify medications according to their efficacy; how to classify mass mailings according to their likelihood of driving revenue; how to divide fertilizers into those that will work well with our crops and those that won’t.

The key is to find ways of classifying into categories that make sense and that stand up in more than just one sample. Decision analytics comprises several types of analysis that help you make that sort of classification. The techniques have been around for decades, but it’s only with the emergence of the term analytics that the ways that those techniques can work together have gained real currency.

This initial chapter provides a brief overview of each of the techniques discussed in the book’s remaining chapters, along with an introduction to the conditions that might guide you toward selecting a particular technique.

Classifying According to Existing Categories

Several techniques used in decision analytics are intended for sets of data where you already know the correct classification of the records. The idea of classifying records into known categories might seem pointless at first, but bear in mind that this is usually a preliminary analysis. You typically intend
to apply what you learn from such a pilot study to other records—and you don’t yet know which categories those other records belong to.

**Using a Two-Step Approach**

A classification procedure that informs your decision making often involves two steps. For example, suppose you develop a new antibiotic that shows promise of preventing or curing new bacterial infections that have so far proven drug-resistant. You test your antibiotic in a double-blind experiment that employs random selection and assignment, with a comparison arm getting a traditional antibiotic and an experimental arm getting your new medication. You get mixed results: Your medication stops the infection in about one third of the patients in the experimental arm, but it’s relatively ineffective in the remaining patients.

You would like to determine whether there are any patient characteristics among those who received your new medication that tend either to enable or to block its effects. You know your classification categories—those in whom the infection was stopped, and those in whom the infection was unaffected. You can now test whether other patient characteristics, such as age, sex, infection history, blood tests and so on, can reliably distinguish the two classification categories. Several types of analysis, each discussed in this book, are available to help you make those tests: Multivariate analysis of variance and discriminant function analysis are two such analyses. If those latter tests are successful, you can classify future patients into a group that’s likely to be helped by your medication and a group that’s unlikely to be helped.

Notice the sequence in the previous example. You start with a group whose category memberships are known—those who received your medication and were helped and those who weren’t. Pending a successful test of existing patient characteristics and their response to your medication, you might now be in a position to classify new patients into a group that your medication is likely to help, and a group that isn’t. Health care providers can now make more informed decisions about prescribing your medication.

**Multiple Regression and Decision Analytics**

The previous section discusses the issue of classifying and decision making purely from the standpoint of design. Let’s take another look from the point of view of analysis rather than design—and, not incidentally, in terms of multiple regression, which employs ideas that underlie many of the more advanced techniques described in this book.

You’re probably familiar to some degree with the technique of multiple regression. That technique seeks to develop an equation that looks something like this one:

\[ Y = a_1 X_1 + a_2 X_2 + b \]

In that equation, \( Y \) is a variable such as weight that you’d like to predict, \( X_1 \) is a variable such as height, and \( X_2 \) is another variable such as age. You’d like to use your knowledge of people’s heights and ages to predict their weight.
You locate a sample of, say, 50 people, weigh them, measure each person’s height, and record their ages. Then you push that data through an application that calculates multiple regression statistics and in that way learn the values of the remaining three items in the equation:

- $a_1$, a coefficient you multiply by a person’s height
- $a_2$, a coefficient you multiply by a person’s age
- $b$, a constant that you add to adjust the scale of the results

You can now find another person whose weight you don’t know. Get his height and age and plug them into your multiple regression equation. If your sample of 50 people is reasonably representative, and if height and age are reliably related to weight, you can expect to predict this new person’s weight with fair accuracy.

You have established the numeric relationships between two predictor variables, height and age, and a predicted variable, weight. You did so using a sample in which weight—which you want to predict—is known. You expect to use that information with people whose weight you don’t know.

At root, those concepts are the same as the ones that underlie several of the decision analytics techniques that this book discusses. You start out with a sample of records (for example, people, plants, or objects) whose categories you already know (for example, their recent purchase behaviors with respect to your products, whether they produce crops in relatively arid conditions, whether they shatter when you subject them to abnormal temperature ranges). You take the necessary measures on those records and run the numbers through one or more of the techniques described in this book.

Then you apply the resulting equations to a new sample of people (or plants or objects) whose purchasing behavior, or ability to produce crops, or resistance to unusual temperatures is unknown. If your original sample was a representative one, and if there are useful relationships between the variables you measured and the ones you want to predict, you’re in business. You can decide whether John Jones is likely or unlikely to buy your product, whether your new breed of corn will flourish or wither if it’s planted just east of Tucson, or whether pistons made from a new alloy will shatter in high temperature driving.

I slipped something in on you in the last two paragraphs. The first example in this section concerns the prediction of a continuous variable, weight. Ordinary, least-squares multiple regression is well suited to that sort of situation. But the example in the previous section uses categories, nominal classifications, as a predicted variable: cures infection versus doesn’t cure it. As the values of a predicted variable, categories present problems that multiple regression has difficulty overcoming. When the predictor variables are categories, there’s no problem. In fact, the traditional approach to analysis of variance (ANOVA) and the regression approach to ANOVA are both designed specifically to handle that sort of situation. The problem arises when it’s the predicted variable rather than the predictor variables that is measured on a nominal rather than a continuous scale.
But that's precisely the sort of situation you're confronted with when you have to make a choice between one of two or more alternatives. Will this new product succeed or fail? Will this new medicine prolong longevity or shorten it due to side effects? Based solely on their voting records, which political party did these two congressional representatives from the nineteenth century belong to?

So, to answer that sort of question, you need analysis techniques—decision analytics—designed specifically for situations in which the outcome or predicted variable is measured on a nominal scale, in terms of categories. That, of course, is the focus of this book: analysis techniques that enable you to use numeric variables to classify records into groups, and thereby make decisions about the records on the basis of the group you project them into. To anticipate some of the examples I use in subsequent chapters:

- How can you classify potential borrowers into those who are likely to repay loans in accordance with the loan schedules, and those who are unlikely to do so?
- How can you accurately classify apparently identical plants and animals into different species according to physical characteristics such as petal width or length of femur?
- Which people in this database are so likely to purchase our resort properties in the Bahamas that we should fly them there and house them for a weeklong sales pitch?

**Access to a Reference Sample**

In the examples I just cited, it's best if you have a reference sample: a sample of records that are representative of the records that you want to classify and that are already correctly classified. (Such samples are often termed *supervised* or *training samples.*) The second example outlined in this chapter, regarding weight, height, and age, discussed the development of an equation to predict weight using a sample in which weight was known. Later on you could use the equation with people whose weight is not known.

Similarly, if your purpose is to classify loan applicants into Approved versus Declined, it's best if you can start with a representative reference sample of applicants, perhaps culled from your company's historical records, along with variables such as default status, income, credit rating, and state of residence. You could develop an equation that classifies applicants into your Approved and Declined categories.

Multiple regression is not an ideal technique for this sort of decision analysis because, as I noted earlier, the predicted variable is not a continuous one such as weight but is a dichotomy. However, multiple regression shares many concepts and treatments with techniques that in fact are suited to classifying records into categories. So you're ahead of the game if you've had occasion to study or use multiple regression in the past. If not, don't be concerned; this book doesn't assume that you're a multiple regression maven.

Multiple regression does require that you have access to a reference sample, one in which the variable that is eventually to be predicted is known. That information is used to develop the prediction equation, which in turn is used with data sets in which the predicted variable
is as yet unknown. Other analytic techniques, designed for use with categorical outcome variables, and which also must make use of reference samples, include those I discuss in the next few sections.

**Multivariate Analysis of Variance**

Multivariate analysis of variance, or MANOVA, extends the purpose of ordinary ANOVA to multiple dependent variables. (Statistical jargon tends to use the term *multivariate* only when there is more than one predicted or outcome or dependent variable; however, even this distinction breaks down when you consider discriminant analysis.) Using ordinary univariate ANOVA, you might investigate whether people who pay back loans according to the agreed terms have, on average at the time the loan is made, different credit ratings than people who subsequently default. (I review the concepts and procedures used in ANOVA in Chapter 3, “Univariate Analysis of Variance (ANOVA).”) Here, the predictor variable is whether the borrower pays back the loan, and the predicted variable is the borrower's credit rating.

But you might be interested in more than just those people's credit ratings. Do the two groups differ in average age of the borrower? In the size of the loans they apply for? In the average term of the loan? If you want to answer all those questions, not just one, you typically start out with MANOVA, the multivariate version of ANOVA. Notice that if you want MANOVA to analyze group differences in average credit ratings, average age of borrower, average size of loan, and average term of loan, you need to work with multiple predicted variables, not solely the single predicted variable you would analyze using univariate ANOVA.

MANOVA is not a classification procedure in the sense I used the phrase earlier. You do not employ MANOVA to help determine whether some combination of credit rating, borrower's age, size of loan, and term of loan accurately classifies applicants according to whether they can be expected to repay the loan or default. Instead, MANOVA helps you decide whether those who repay their loans differ from those who don't on any one of, or a combination of, the outcome variables—credit rating, age, and so on.

You don't use one univariate ANOVA after another to make those inferences because the outcome variables are likely correlated with one another. Those correlations have an effect, which cannot be quantified, on the probability estimate of each univariate ANOVA. In other words, you might think that each of your univariate F-tests is operating at an alpha level of .05. But because of the correlations the F-tests are not independent of one another and the actual alpha level for one test might be .12, for another test .08, and so on. MANOVA helps to protect you against this kind of problem by taking all the outcome variables into account simultaneously. See Chapter 4, “Multivariate Analysis of Variance (MANOVA),” for a discussion of the methods used in MANOVA.

It surprises some multivariate analysts to learn that you can carry out an entire MANOVA using Excel's worksheet functions only. But it's true that by deploying Excel's matrix
functions properly—`MDETERM()`, `MINVERSE()`, `MMULT()`, `TRANSPOSE()` and so on—you can go from raw data to a complete MANOVA including Wilks’ Lambda and a multivariate F-test in just a few steps. Nevertheless, among the files you can download from the publisher’s website is a MANOVA workbook with subroutines that automate the process for you. Apart from learning what’s involved, there’s little point to doing it by hand if you can turn things over to code.

But MANOVA, despite its advantages in this sort of situation, still doesn’t classify records for you. The reason I’ve gone on about MANOVA is explained in the next section.

**Discriminant Function Analysis**

Discriminant function analysis is a technique developed by Sir Ronald Fisher during the 1930s. It is sometimes referred to as linear discriminant analysis or LDA, or as multiple discriminant analysis, both in writings and in the names conferred by statistical applications such as R. Like MANOVA, discriminant analysis is considered a true multivariate technique because its approach is to simultaneously analyze multiple continuous variables, even though they are treated as predictors rather than predicted or outcome variables.

Discriminant analysis is typically used as a followup to a MANOVA. If the MANOVA returns a multivariate F-ratio that is not significant at the alpha level selected by the researcher, there is no point to proceeding further. If the categories do not differ significantly as to their mean values on any of the continuous variables, then the reverse is also true. The continuous variables cannot reliably classify the records into the categories of interest.

But if the MANOVA returns a significant multivariate F-ratio, it makes sense to continue with a discriminant analysis, which, in effect, turns the MANOVA around. Instead of asking whether the categories differ in their mean values of the continuous variables, as does MANOVA, discriminant analysis asks how the continuous variables combine to separate the records into different categories.

The viewpoint adopted by discriminant analysis brings about two important outcomes:

- It enables you to look more closely than does MANOVA at how the continuous variables work together to distinguish the category membership.
- It provides you with an equation called a discriminant function that, when used like a multiple regression equation, assigns individual records to categories such as Repays versus Defaults or Buys versus Doesn’t Buy.

Chapter 5, “Discriminant Function Analysis: The Basics,” and Chapter 6, “Discriminant Function Analysis: Further Issues,” show you how to obtain the discriminant function, and what use you can make of it, using Excel as the platform. An associated Excel workbook automates a discriminant analysis using the results of a preliminary MANOVA.

Both MANOVA and discriminant analysis are legitimately thought of as multivariate techniques, particularly when you consider that they look at the same phenomena, but from
different ends of the telescope. They are also parametric techniques: Their statistical tests make use of theoretical distributions such as the F-ratio and Wilks' lambda. Therefore these parametric techniques are able to return to you information about, say, the likelihood of getting an F-ratio as large as the one you observed in your sample if the population means were actually identical.

Those parametric properties invest the tests with statistical power. Compared to other, nonparametric tests, techniques such as discriminant analysis are better (sometimes much better) able to inform you that an outcome is a reliable one. With a reliable finding, you have every right to expect that you would get the same results from a replication sample, constructed similarly.

But that added statistical power comes with a cost: You have to make some assumptions (which of course you can test). In the case of MANOVA and discriminant analysis, for example, you assume that the distribution of the continuous variables is “multivariate normal.” That assumption implies that you should check scattercharts of each pair of continuous variables, across all your groups, looking for nonlinear relationships between the variables. You should also arrange for histograms of each variable, again by group, to see whether the variable’s distribution appears skewed.

Excel can make your life a little easier here, although admittedly not by much. It’s easy enough to create a scatterchart in Excel. (Begin by going to the Insert tab on the Ribbon in Excel 2007 or 2010 or 2013. Click the Chart Wizard button in an earlier version.) But if you have 3 categories and 7 continuous variables, that’s 3 × 7 × 6 or a tedious 126 scattercharts to create. Pivot charts would make things a little quicker, but pivot charts do not offer a scatterchart type.

Excel has a worksheet function, SKEW(), which returns the skewness of a distribution of values. The function does not return perhaps the most popular version of skewness, the average cubed z-score. Instead, SKEW() uses this formula:

\[
N \sum_{i=1}^{N} z_i^3 / (N - 1)(N - 2)
\]

With a small number of records, Excel’s value of skewness can be easily half again as large as the average cubed z-score (which of course does not depend on the number of records). Still, using SKEW() is undoubtedly faster than creating histograms. (The Data Analysis add-in has a Histograms tool that can speed the process considerably.)

As another example, MANOVA assumes that the variance-covariance matrix is equivalent in the different categories. All that means is that if you assembled a matrix of your variables, showing each variable’s variance and its covariance with the other continuous variables in your design, the values in that matrix would be equivalent for the Repayers and for the Defaulters, for the Buyers and the Non-Buyers. Notice that I used the word “equivalent,” not “equal.” The issue is whether the variance-covariance matrices are equal.
in the population, not necessarily in the sample (where they’ll never be equal). Again, you can test whether your data meets this assumption. Bartlett’s test is the usual method and the MANOVA workbook, which you can download from the publisher’s website, carries that test out for you.

If these assumptions are met, you’ll have a more powerful test available than if they are not met. When the assumptions are not met, you can fall back on what’s typically a somewhat less powerful technique: logistic regression.

**Logistic Regression**

Logistic regression differs from ordinary least squares regression in a fundamental way. Least squares regression depends on correlations, which in turn depend on the calculation of the sums of squared deviations, and regression works to minimize those sums—hence the term “least squares.”

In contrast, logistic regression depends not on correlations but on odds ratios (or, less formally, odds). The process of logistic regression is not a straightforward computation as it is in simple or multiple regression. Logistic regression uses *maximum likelihood* techniques to arrive at the coefficients for its equation: for example, the values for \( a_i \) and \( a_j \) that I mentioned at the beginning of this chapter. Conceptually there's nothing magical about maximum likelihood. It's a matter of trial and error: the educated and automated process of trying out different values for the coefficients until they provide an optimal result. I discuss how to convert the probabilities to odds, the odds to a special formulation called the *logit*, how to get your maximum likelihood estimates using Excel's Solver—and how to find your way back to probabilities, in Chapter 2, “Logistic Regression.”

Largely because the logistic regression process does *not* rely on reference distributions such as the F distribution to help evaluate the sums of squares, logistic regression cannot be considered a parametric test. One important consequence is that logistic regression does not involve the assumptions that other techniques such as MANOVA and discriminant analysis employ. That means you can use logistic regression with some data sets when you might not be able to use parametric tests.

For example, in logistic regression there is no assumption that the continuous variables are normally distributed. There is no assumption that the continuous variables are related in a linear rather than curvilinear fashion. There is no assumption that their variances and covariances are equivalent across groups.

So, logistic regression positions you to classify cases using continuous variables that might well fail to behave as required by MANOVA and discriminant analysis. It extends the number of data sets that you can classify.

But the same tradeoff is in play. Although you can get away with violations in logistic regression that might cause grief in MANOVA, nothing’s free. You pay for discarding assumptions with a loss of statistical power. Logistic regression simply is not as sensitive to small changes in the data set as is discriminant analysis.
Classifying According to Naturally Occurring Clusters

In contrast to the techniques that I’ve briefly outlined in the preceding sections, the second set of analytic techniques that this book discusses does not necessarily start out with reference samples that include actual category membership. Instead, it’s hoped that the categories (often termed clusters in these methods) will emerge from combinations of the measured variables, combinations established and assessed by the software you use.

Principal Components Analysis

The book moves back to genuinely multivariate analysis in Chapter 7, “Principal Components Analysis.” This approach dates back to the beginning of the twentieth century and is therefore the oldest of the analytic techniques discussed in this book. Principal components analysis is the precursor to factor analysis, both historically and procedurally. Factor analysis began to gain currency in the 1930s, roughly 25 years after Karl Pearson was doing early work on principal components. And factor analysis nearly always starts with a principal components analysis, although the numeric inputs might be slightly different and the purposes of the two techniques aren’t quite the same.

The idea behind principal components analysis is that it’s possible to combine several measured variables into a single principal component (or, equivalently, many measured variables into just a few principal components) without losing a significant amount of meaningful information. The result is a much simpler data set that the researcher can investigate more easily. The same is true for factor analysis, but there the emphasis is on understanding how a factor, or component, which is not directly observable or measurable is expressed in variables that, in contrast, can be measured.

So far as I can tell, principal components analysis was used in just this way until the mid-1960s, when researchers in various fields began to notice that principal components analysis might also be useful for the purpose of classification. It can happen—and Chapter 7 details a couple of examples—that when you extract a few principal components from a larger number of measured variables, the records cluster together in the space that’s defined by the principal components. Along with the Excel workbook for Chapter 7 you’ll find a file named Factor.xls, which runs a principal components analysis for you and performs a Varimax rotation of the components’ axes.

If those clusters make sense, you might well be onto a finding that’s really compelling. But great care is needed: Methods that work with data that includes no information about actual membership are particularly susceptible to nonsense results. And when the composition of the clusters conforms to what you expect, it’s even more important to have another data sample at hand so that you can validate your results.

If the first analysis results in ridiculous cluster assignments, you’re probably not tempted to think that you’re on to something. If the cluster assignments make sense, it’s incumbent on you to replicate the finding. (You don’t want to claim that you’ve achieved cold fusion on your kitchen table if you can’t do it more than once.)
But principal components analysis has another role to play in the classification of records. Among the characteristics of the components you extract from a raw data set are the following:

- The components are uncorrelated. This characteristic has several advantages. One relatively minor advantage is that the components are orthogonal—their axes are at right angles to one another, which can make the interpretation less ambiguous.
- Another, more important advantage is that because two uncorrelated components cannot share variance, it’s possible to set aside as irrelevant components that are extracted later in the process. Components have characteristics called eigenvalues that can tell you how much of the overall variance in the data set each component accounts for.

You can often remove components with low eigenvalues from subsequent analysis. Doing so tends to discard junk variation: measurement error, sampling error, variation that’s specific to a particular variable and therefore contrary to the notion of a “component.”

All this is consistent with the basic notion of principal components: to find combinations of variables that bring important, meaningful variation along with them and that leave irrelevant, misleading variation behind. You also wind up with (usually) many fewer components than there were measured variables, and that eases the task of making sense of the results.

When you get rid of irrelevant, nuisance variation, you can work instead with relatively clean components. That can make the groups subsequently derived by cluster analysis much clearer.

Unfortunately, even if the groups are clear, well separated, and distinct, it doesn’t necessarily follow that they mean anything. Again, the better the results look at first, the more important it is to replicate them, validate them, and verify them.

Among the techniques that are particularly improved by the use of principal components instead of the raw data are those that are collectively termed cluster analysis.

### Cluster Analysis

Cluster analysis is distinguished from other approaches to classification by the fact that it does not require you to know up front what category—that is, what cluster—each record in your data set belongs to. True, principal components analysis doesn’t either, but although it’s useful as a tool in classification, it’s not intended as a classification technique.

---

NOTE

There are, however, certain equivalencies in the math that underlies principal components analysis and cluster analysis. It appears that the approach taken by principal components analysis is considerably closer to that taken by cluster analysis than had been previously recognized.
The two different branches of cluster analysis, the linkage methods and the centroid-based methods, are both designed to cause clusters to emerge from the variables (or, as I indicated in the previous section, the principal components) that you supply.

Methods such as discriminant analysis and logistic regression require you to supply a reference sample, one that separates a sample of records into the categories you’re interested in. You develop your classification equation using that data set. Later on you apply the equation to records that are as yet unclassified.

Cluster analysis works differently. The algorithms look for records that are similar to one another, and assign (and reassign) those records to the same clusters. Typically, the algorithms keep plugging away until no records are reassigned during an iteration through the loop.

One problem that arises early when you’re deciding on a clustering method is the question of what constitutes similarity. There are various ways to define similarity, and perhaps the most popular is by way of a distance measure: how far a given record is from another record, or from the center of a cluster.

Cluster analysis, along with MANOVA and discriminant analysis, provides for the use of multiple continuous variables in cluster or category formation. With multiple continuous variables, the center of a cluster is not a simple mean value, but a vector of averages. That vector is called a centroid, a term that you’ll come across often in this book.

It complicates matters that there are several types of distance. Ordinary Euclidean distance is one: The width of this sepal is 3.5 centimeters, and the width of that sepal is 3.75 centimeters. For various reasons, the Euclidean distance is usually squared: The squared Euclidean distance between the width of the two sepals is \((3.75 - 3.5)^2\). After some further calculations, though, some analyses take the square root of the results to return to a simple Euclidean metric.

Another type of distance measure is the familiar Pearson correlation (although the larger the correlation, the closer the records). Using Pearson correlation as a distance measure forces the standardization of the measures, and so the effect of the scale of the original measurement is removed.

Yet another measure is Mahalanobis’ \(D^2\), which incorporates the variables’ covariances into the squared Euclidean distance, so that the variables are not necessarily and automatically regarded as orthogonal to one another.

Measures of similarity are not limited to distances, though. Agreement measures are sometimes used: “Both these people have studied calculus.” But the math involved in handling agreement measures is not as straightforward as it is for distance measures.
Furthermore, with agreement measures you have to watch what you’re using to indicate similarity. If two people both own 1920 Duesenberg autos, that’s likely a more meaningful index of similarity than if two people don’t own Duesenbergs.

Because squared Euclidean distances represent something of a standard method of measurement in cluster analysis, I stick with them as measures of similarity in this book’s examples (see Chapter 8, “Cluster Analysis: The Basics,” and Chapter 9, “Cluster Analysis: Further Issues”).

The two branches of cluster analysis that I mentioned earlier, linkage methods and centroid-based methods, frequently return different results. Because cluster analysis is at root an exploratory method, one that seeks to establish undefined clusters, you might want to consider running both methods to determine which, if either, results in clusters that make sense to you.

The linkage methods work from the ground up, starting with the two records that have the smallest distance between them and calling those two records a cluster. Then the record with the smallest distance to that cluster’s centroid is found, and that distance is compared to the smallest distance between that record and another single record. If the distance to the existing cluster is smaller, the third record is assigned to that cluster. But if the distance to another individual record is smaller, a new cluster is created. (Several different definitions of “smallest” distance exist in the linkage methods, including single linkage, complete linkage, and average linkage.)

The most popular of the centroid-based methods is *k*-means, where “k” simply refers to the number of continuous variables involved. K-means is more of a top-down approach than are the linkage methods. Further, you are expected to specify the number of clusters to establish. The centroids of each cluster are established randomly at the outset, and the process continues by assigning and reassigning individual records to their nearest cluster, updating the centroid values accordingly, until a journey through the loop causes no records to be reassigned to a different cluster.

The Excel workbooks available for download from the publisher’s website include a workbook with VBA code that performs k-means analysis for you.

**Some Terminology Problems**

The first main section of this chapter focuses on situations in which you start out knowing which categories some of your records belong to, and you want to know how to classify other records whose categories aren’t yet known. The second main section introduces other techniques that ignore known classifications at first. They seek instead to determine whether classifications emerge from how individual records cluster together due to their similarities on measured, continuous variables. Those techniques include principal component analysis (PCA) and cluster analysis.

Before moving on to Chapter 2, “Logistic Regression,” I want to address some problems with terminology that characterize inferential statistics in general but multivariate statistics
in particular, because of the way that variables can switch roles before you’re through with them. These problems can create particular confusion in the exploratory context that often characterizes decision analytics.

**The Design Sets the Terms**

For a variety of reasons, it’s important to distinguish between an analytic technique (such as univariate ANOVA) and your reason for using it. Suppose you have three treatments you want to test: a new drug, a traditional drug and a placebo. You plan to test whether the treatments have different effects on adult females’ cholesterol levels. You intend to randomly sample your subjects from a population of women whose cholesterol levels are abnormally high, and to randomly assign them to one of your three treatments.

In this context it’s typical and meaningful to refer to the cholesterol measures as a dependent variable. Your hypothesis is that the subjects’ cholesterol levels depend on the treatment to which they are assigned. There is a causal relationship between the dependent variable and the treatment.

For some reason that’s apparently lost to history, the three treatments you have in mind, taken together, are termed an independent variable. There’s nothing independent about it. As the experimenter, you decide what values it takes on (here, new drug, traditional drug, and placebo). You decide (here, randomly) who to assign to which drug. The only rationale for terming it an independent variable seems to be to distinguish it from a dependent variable.

Still, that’s a relatively benign problem. Statistical jargon has many more misleading terms than “independent variable.” But it’s necessary to remember that independent variable and dependent variable belong to your design, not to the statistical procedure. As you conceive of and carry out your experiment, the differences in average cholesterol level among your three treatment groups depend on the three treatments, not on extraneous sources of variation such as subject self-selection or regression toward the mean. Over time, the terms have come to connote the nature of the design: an independent variable, which is under the experimenter’s control, and a dependent variable, which responds in a cause-and-effect fashion to differences in the independent variable.

Now suppose that you decide not to use traditional ANOVA math on your data. The design and management of your experiment is the same as before. But instead of accumulating squared deviations between and within groups, you use one of the available coding methods to represent group membership and pump your data through a multiple regression application.

The results—the sums of squares, the F-ratio, the p value—remain the same as with the traditional ANOVA. More important to this discussion, it’s still appropriate to use the terms dependent variable and independent variable. Your experimental design has not changed—just the way that you do the arithmetic. The differences in the group means on the dependent variable are still caused by the differential effects of the treatments and, to a degree that’s under your control, to the effects of chance.
Causation Versus Prediction

Now, completely alter the rationale for and the design of the research. Instead of researching the causal and differential effects of drugs on cholesterol levels, you're interested in determining whether a relationship exists between the Dow Jones Industrial Average (DJIA) and other indexes of market valuation, such as the advance-decline (A/D) ratio and the total volume on the New York Stock Exchange. The researcher could easily consult any of hundreds of online sources of historical information regarding the DJIA and associated statistics, such as trading volume and A/D ratios, to pick up tens of thousands of data points.

This is different. Here, the researcher is not in a position to manipulate the values of one or more independent variables, as is the case in a true experiment. The researcher cannot by fiat state that, “The advance-decline ratio shall be 1.5 on September 30,” and observe a resulting change in the DJIA as though there were a causal relationship. Nor is the researcher able to randomly select and assign subjects to one group or another: Membership in the companies that make up the DJIA is largely fixed and certainly beyond the researcher's control.

There's nothing intrinsically wrong with this sort of situation, although it's often referred to, a little insultingly, as a “grab sample.” It's well suited to making predictions, just not to explaining causation. The researcher can't directly manipulate the actual values of the predictor variables, but instead can ask, “What value of the DJIA would we expect to see if the trading volume increased by 10%?”

It's best to avoid terms such as independent variable and dependent variable with data acquired in this way. They imply that the researcher controls the independent variables, and that there is a causal relationship between an independent variable and the dependent variable. The relationship might indeed be causal, but the researcher is not in a position to control an independent variable so as to demonstrate the causality.

To acknowledge that “independent” and “dependent” might not be accurate terms without a randomized study with direct manipulation of an independent variable, many writers have adopted the terms predictor variable to refer, for example, to A/D ratios and trading volume, and predicted variable to refer to the variable they want to predict, such as the DJIA. (You also see terms such as outcome and criterion in place of dependent or predicted, but they just tend to beg the question.)

Why the Terms Matter

Two fundamental reasons explain why I have spent so much space here on what must seem the pedantry of terminology.

One reason is that most of the techniques of decision analytics are used in an exploratory way. You're looking for combinations of numerically measured variables that, taken together, might explain differences between categories. It's typical to use data that already exists, often in companies' operational databases, to search for those relationships and formulate hypotheses accordingly. Only then might you set up a true experiment in which
you randomly select and assign subjects to groups and manipulate directly the nature of treatments applied to each group. In this way you hope to confirm your hypotheses, and only then it might be appropriate to imply causation using terms such as independent and dependent.

The second reason is that in at least a major subset of decision analytics work, the variables change horses midstream. As described in previous sections of this chapter on MANOVA and discriminant analysis, you might start out with two or more groups that act as predictor variables and two or more continuous variables that act as predicted variables. MANOVA asks whether any groups differ significantly on one or more predicted variables or on some combination of the predicted variables.

If you get a significant result from the MANOVA, you generally proceed to discriminant analysis, where you seek to determine which continuous variables, alone or in combination, distinguish the groups. In effect, you turn the design end for end. The categories that were the predictors in the MANOVA now constitute the predicted variable in the discriminant analysis. The continuous variables that were the predicted variables in the MANOVA are now the predictor variables in the discriminant analysis.

The situation is clearly impossible if you begin by calling the categories an independent variable in a MANOVA and wind up calling them a dependent variable in a discriminant analysis. It's just a little confusing, not impossible, if you think of the categories as predictors in the MANOVA and as predicted variables in the subsequent discriminant analysis.

Therefore, I have tried in this book to use predictor variable and predicted variable unless the context makes it clear that an example assumes a true randomized experiment.

Okay, let's move on to the meat of decision analytics, starting with logistic regression.
Index

Symbols

#DIV/0! error, avoiding, 108
$ (dollar sign) in absolute references, 165

A

access to reference samples, 8-9
accuracy of logistic regression, assessing
  with cross-validation, 43-48
  with statistical inference, 48-52
agreement measures, 15-16
alpha levels, 66-67
Among SSCP matrix
  basis for, 105
  calculating in Excel, 102-104
analysis of variance. See ANOVA
  (univariate analysis of variance);
  MANOVA (multivariate analysis of variance)
Analysis ToolPak. See Data Analysis add-in
ANOVA (univariate analysis of variance)
  ANOVA: Single Factor tool, 73-75
    descriptive statistics, 75-76
    group names, 75
    inferential statistics, 76-79
  cholesterol treatment example, 66-72
  error rate adoption, 66-67
  results, 93
  standard error of the mean, 70-72
  statistical calculations, 67-70
comparison with MANOVA, 9-10, 53-54, 89-92, 96-105
Among and Within SSCP matrices calculations, 102-104
SS (sums of squares) versus SSCP (sum of squares and cross products), 98-102, 104-105
as follow-up to MANOVA, 116-117
multiple regression analysis
categories as numbers, 120-122
effect coding, 80-82
general linear model versus, 79-80
LINEST function, 82-87
variance
defined, 54
expected values between groups, 58-61
expected values within groups, 56-58
F-ratio, 61-64
noncentral F distribution, 64-66
partitioning, 55-56
uses for, 54-55
ANOVA: Single Factor tool, 73-75
descriptive statistics, 75-76
group names, 75
inferential statistics, 76-79
df (degrees of freedom), 77-78
F (F-ratio), 78-79
F Crit (critical value), 79
MS (mean squares), 78
p-value (probability of F-ratio), 79
source of variation, 76
SS (sums of squares), 76-77
aphid measurements example (principal components analysis), 206-207
arithmetic underflow, 45
array formulas, entering, 82-83, 101
assumptions in regression analysis, 25

AVERAGE function, 68
average linkage method, 227-228
axes, reversing, 204

benchmarking
discriminant function analysis results, 147-152
coefficient signs, 152
coefficients and constants returned, 151-152
downloading R, 147
running analysis, 149-152
saving CSV file in Excel, 148-149
k-means method results, 233-234
Berkeley admissions example (cluster analysis), 221-223

canonical correlation, 125-127
capitalization on chance, 171
categories
classification into known, 5-12
discriminant function analysis, 10-12, 162-171. See also
discriminant function analysis
logistic regression, 12. See also
logistic regression
MANOVA (multivariate analysis of variance), 9-10. See also
MANOVA (multivariate analysis of variance)
multiple regression analysis, 6-8. See also multiple regression
analysis
reference sample access, 8-9
two-step approach, 6
classification into naturally occurring, 13-16
cluster analysis, 14-16. See also cluster analysis
principal components analysis, 13-14. See also principal components analysis
determining number in k-means cluster analysis, 237-239
as numbers, 120-122, 127-129
viewpoint adjustments, 123-124
causation, prediction versus, 18
central F distribution, 105
centroid-based methods (cluster analysis), 16
centroids
ANOVA versus MANOVA results, 95-96
Cluster Centroids worksheet, 241-242
defined, 15, 92
distance from, 163
CHISQ.DIST.RT function, 52
cholesterol treatment example
multivariate ANOVA, visualizing results, 92-96
ANOVA results, 93
MANOVA results, 93-95
single factor ANOVA, 66-72
error rate adoption, 66-67
standard error of the mean, 70-72
statistical calculations, 67-70
classification
into known categories, 5-12
discriminant function analysis, 10-12, 162-171. See also discriminant function analysis
logistic regression, 12. See also logistic regression
MANOVA (multivariate analysis of variance), 9-10. See also MANOVA (multivariate analysis of variance)
multiple regression analysis, 6-8. See also multiple regression analysis
reference sample access, 8-9
two-step approach, 6
into naturally occurring clusters, 13-16
cluster analysis, 14-16. See also cluster analysis
principal components analysis, 13-14. See also principal components analysis
classification table, creating, 170-171
closing Discriminant workbook, 143
cluster analysis
average linkage method, 227-228
complete linkage method, 223-227
discriminant function analysis versus, 209-211
Euclidean distance, calculating in Excel, 211-214
explained, 14-16
k-means method, 228-233
benchmarking with R software, 233-234
characteristics of, 228-229
comparison with principal components analysis, 248-249
cross-validation, 256-258
example, 229-233
Kmeans Cluster Analysis workbook, 235-248
on state crime rates example (principal components analysis), 249-253
on wine data example, 253-256
logistic regression versus, 209-211
Mahalanobis’ D2 (distance from centroids) in, 214-215
purpose of, 202
record identifiers, importance of, 253
self-selecting nature of, 220-223
single linkage method, 215-220
Cluster Centroids worksheet, 241-242
Cluster Members worksheet, 239-241
Cluster Variances worksheet, 242-244
clusters. See categories coding techniques in multiple regression, 127-129
coefficients
in discriminant function analysis, 129-130
   Eigenstructures and Coefficients worksheet, 156-159
   relationship with regression coefficients, 130-133
   returned by R software, 151-152
   uses for, 159-162
factor score coefficients, 186
eigenvalues and eigenvectors and, 190-191
in multiple regression, 129-130
   relationship with discriminant coefficients, 130-133
structure coefficients
interpreting, 155-156
misleading results from, 136-137
positive versus negative signs, 152
communality of variables, 155, 186
complete linkage method, 223-227
correlation matrix in principal components analysis, 179, 187-192
eigenstructures, calculating in Excel, 192-196
eigenvalues and eigenvectors
data set variance and, 191-192
factor coefficients and, 190-191
loadings and, 188-190
role of, 187-188
critical value, 64
in ANOVA: Single Factor output, 79
cross-validation
in cluster analysis, 256-258
in discriminant function analysis, 171-172
in logistic regression, 43-48
CSV files, saving in Excel, 148-149
curvilinear relationships, 22-24

D

D2 (distance from centroids), 163
in cluster analysis, 214-215
variance-covariance matrix
   adjustments in discriminant function analysis, 167-169
Data Analysis add-in, 72
   ANOVA: Single Factor tool, 73-75
descriptive statistics, 75-76
   group names, 75
   inferential statistics, 76-79
   installing, 73
degrees of freedom in ANOVA: Single Factor output, 77-78
dependent variables
causation versus prediction, 18
terminology of, 17
descriptive statistics
in ANOVA: Single Factor output, 75-76
in MANOVA workbook output, 112-113
design of experiment, terminology defined by, 17
determinants
defined, 106-107
linear dependencies and, 180
deviations from mean in discriminant function analysis, 164-167
DEVSQ function, 55-56, 68, 76-77, 97
diagonal matrices, 180
dichotomies
equal variances with, 27-28
residual distribution of, 29-31
Discrim add-in. See Discriminant workbook
Discriminant dialog box, 141-143
discriminant function analysis
benchmarking with R software, 147-152
coefficient signs, 152
coefficients and constants returned, 151-152
downloading R, 147
running analysis, 149-152
saving CSV file in Excel, 148-149
classification into categories, 162-171
classification assignments, 169-170
classification table creation, 170-171
distance from centroids, 163
mean-corrected values, 164-167
variance-covariance matrix adjustments, 167-169
cluster analysis versus, 209-211
comparison with multiple regression analysis, 122-123, 125-129
canonical correlation, 125-127
coding techniques, 127-129
eigenstructures, 133-135
relationship between coefficients, 130-133
return values, 129-130
structure coefficients, 136-137
Discriminant workbook, 139-143
closing, 143
coefficient usages, 159-162
Discriminant dialog box, 141-143
Discriminant Results dialog box, 153-154
Eigenstructures and Coefficients worksheet, 156-159
Excel version compatibility, 153
interpreting structure coefficients, 155-156
opening, 140-141
explained, 10-12, 119-120
as follow-up to MANOVA, 116-117
multiple continuous variables, reasons for using, 144-147
purpose of, 202
trained versus untrained analysis, 171-172
viewpoint adjustments in, 123-124
Discriminant Results worksheet, 153-154
Discriminant workbook, 139-143
closing, 143
coefficient usages, 159-162
Discriminant dialog box, 141-143
Discriminant Results worksheet, 153-154
Eigenstructures and Coefficients worksheet, 156-159
Excel version compatibility, 153
interpreting structure coefficients, 155-156
opening, 140-141
distance from centroids, 163
  in cluster analysis, 214-215
  variance-covariance matrix
  adjustments in discriminant function analysis, 167-169

distance measures, 15
  Euclidean distance, 211-214
  Mahalanobis’ D2, 163
  in cluster analysis, 214-215
  in discriminant function analysis, 167-169

distributions of residuals in least squares regression, 29-31

dollar sign ($) in absolute references, 165
downloading R software, 147

effect coding, 80-82, 127-128
eigenstructures
  in discriminant function analysis, 133-135, 156-159
  in multiple regression, 133-135
  in principal components analysis, calculating in Excel, 192-196
Eigenstructures and Coefficients worksheet, 156-159
eigenvalues
  data set variance and, 191-192
  determining number of principal components, 184
  factor coefficients and, 190-191
  iteration and exhaustion process, 193-196
  loadings and, 188-190
  multicollinearity, 187-188
eigenvectors
  factor coefficients and, 190-191
  iteration and exhaustion process, 193-196
  loadings and, 188-190
  multicollinearity, 187-188
equal variances
  assumption of, 25-27
  with dichotomies, 27-28
  range/standard deviations and, 28-29
error rate adoption, 66-67
eta-squared, 116, 154
Euclidean distance
  calculating in Excel, 211-214
  Mahalanobis’ D2 (distance from centroids) versus, 214-215
EXP function, 34, 39
expected values of variances
  defined, 57
  within groups, 56-58
  between groups, 58-61
  within groups, 67-70
  between groups, 67-70
experiment design, terminology
  defined by, 17

factor analysis
  identifying factors, 197-199
  oblique factor rotation, 176-177, 201
  orthogonal factor rotation, 201
  Quartimax factor rotation, 196
  relationship with principal components analysis, 13, 187
  simple structure, 185
  Varimax rotation, 200-202
factor rotation. See factor analysis
factor score coefficients, 186, 190-191
Factor.xls workbook. See Principal Components add-in
F.DIST.RT function, 63
finding Solver, 41
F.INV.RT function, 64, 67
formulas
  array formulas, entering, 82-83, 101
  multiple in multiple cells, 177
  storing as values, 51
F-ratio, 61-64, 187
in ANOVA: Single Factor output, 78-79
central F distribution, 105
converting Wilks' lambda to, 107-110
F-Ratios worksheet, 244-247
in MANOVA workbook output, 115-116
multivariate ANOVA results, 94
multivariate F-ratio, calculating in Excel, 105-110
noncentral F distribution, 64-66
relationship with Wilks' lambda, 115-116
univariate ANOVA results, 93
F-Ratios worksheet, 244-247
F-test, 64
functions
AVERAGE, 68
CHISQ.DIST.RT, 52
DEVSQ, 55-56, 68, 76-77, 97
EXP, 34, 39
F.DIST.RT, 63
F.INV.RT, 64, 67
LINEST, 82-87, 93, 180-181
LN, 34
MATCH, 169-170
MDETERM, 106
MINVERSE, 180
MMULT, 101-102, 158
RAND, 177, 256-258
SKEW, 11
SQRT, 211-214
STDEV.P, 69
STDEV.S, 69
SUMPRODUCT, 101, 159
SUMXMY2, 213-214, 241
T.DIST, 85
TRANSPOSE, 101-102, 161
VAR.P, 55-56
VAR.S, 55-56, 69
G

general linear model, multiple regression analysis versus, 79-80
group names in ANOVA: Single Factor output, 75

I

identity matrices, 180
independence of observations, 45-46
independent variables
causation versus prediction, 18
terminology of, 17
inferential statistics
in ANOVA: Single Factor output, 76-79
df (degrees of freedom), 77-78
F (F-ratio), 78-79
F Crit (critical value), 79
MS (mean squares), 78
p-value (probability of F-ratio), 79
source of variation, 76
SS (sums of squares), 76-77
LINEST function, 85-87
installing
Data Analysis add-in, 73
Solver, 41
intercepts, 83-85
inverse of R matrix in principal components analysis, 179-182
Iris Study, 122, 144-147
iteration and exhaustion process, 193-196

K

Kmeans Cluster Analysis workbook, 235-248
  Cluster Centroids worksheet, 241-242
  Cluster Members worksheet, 239-241
Cluster Variances worksheet, 242-244
F-Ratios worksheet, 244-247
number of clusters, 237-239
Process Summary worksheet, 247-248
**k-means method (cluster analysis), 228-233**
benchmarking with R software, 233-234
characteristics of, 228-229
comparison with principal components analysis, 248-249
example, 229-233
**Kmeans Cluster Analysis workbook, 235-248**
Cluster Centroids worksheet, 241-242
Cluster Members worksheet, 239-241
Cluster Variances worksheet, 242-244
F-Ratios worksheet, 244-247
number of clusters, 237-239
Process Summary worksheet, 247-248
on state crime rates example (principal components analysis), 249-253
on wine data example, 253-256
cross-validation, 256-258
**known categories, classification into, 5-12**
discriminant function analysis, 10-12, 162-171. See also discriminant function analysis
logistic regression, 12. See also logistic regression
**MANOVA (multivariate analysis of variance), 9-10. See also MANOVA (multivariate analysis of variance)**
multiple regression analysis, 6-8. See also multiple regression analysis
reference sample access, 8-9
two-step approach, 6

LDA (linear discriminant analysis). See discriminant function analysis
**least squares regression**
logistic regression versus, 22-29
equal variances, 25-27
equal variances and range/standard deviations, 28-29
equal variances with dichotomies, 27-28
fitting regression lines to data, 22-24
scaling problems, 24-25
underlying assumptions, 25
residual distribution, 29-31
**linear dependencies**
determinants and, 180
multicollinearity, 187-188
**linear discriminant analysis. See discriminant function analysis**
**linear regression, logistic regression versus, 22-29**
equal variances, 25-27
with dichotomies, 27-28
and range/standard deviations, 28-29
fitting regression lines to data, 22-24
scaling problems, 24-25
underlying assumptions, 25
**LINEST function, 82-87, 93, 180-181**
**linkage methods (cluster analysis), 16**
average linkage, 227-228
complete linkage, 223-227
single linkage, 215-220
**LN function, 34**
loadings. See also structure coefficients
communality of variables, 186
eigenvalues and eigenvectors and, 188-190
loan repayment example (maximizing log likelihood), 35-42
  equation setup in Excel, 36-38
  log likelihood calculation in Excel, 40-41
  odds calculation in Excel, 38-39
  probabilities calculation in Excel, 39-40
log likelihood
  assessing logistic regression accuracy with cross-validation, 43-48
  with statistical inference, 48-52
  calculating in Excel, 40-41
log odds, 33-34
logistic regression
  assessing accuracy
  with cross-validation, 43-48
  with statistical inference, 48-52
  cluster analysis versus, 209-211
  explained, 12
  linear regression versus, 22-29
  equal variances, 25-27
  equal variances and range/standard deviations, 28-29
  equal variances with dichotomies, 27-28
  fitting regression lines to data, 22-24
  scaling problems, 24-25
  underlying assumptions, 25
  log odds, 33-34
  maximum likelihood, 34-35
  calculating in Excel, 40-41
  equation setup in Excel, 36-38
  odds calculation in Excel, 38-39
  online loan example, 35-42
  probabilities calculation in Excel, 39-40
  odds ratios, 32-33
  origin of terminology, 21-22
logit
  defined, 34
  odds calculation in Excel, 38-39

Mahalanobis’ D2 (distance from centroids), 163
  in cluster analysis, 214-215
  in discriminant function analysis, 167-169
MANOVA (multivariate analysis of variance)
  comparison with ANOVA, 53-54, 89-92, 96-105
  Among and Within SSCP matrices calculations, 102-104
  SS (sums of squares) versus SSCP (sum of squares and cross products), 98-102, 104-105
  comparison with discriminant function analysis, 10-12, 146
  defined, 89
  explained, 9-10
  follow-up tests, 116-117
  F-ratio, calculating in Excel, 105-110
MANOVA workbook, 110-116
  data input requirements, 110-111
  descriptive statistics, 112-113
  running analysis, 111-112
  univariate and multivariate F-tests, 115-116
  variance-covariance matrix equivalence, 113-115
  visualizing, 92-96
  ANOVA results, 93
  MANOVA results, 93-95
  means and centroids, 95-96
MANOVA workbook, 110-116
  data input requirements, 110-111
  running analysis, 111-113
univariate and multivariate F-tests, 115-116
variance-covariance matrix equivalence, 113-115
MASS (Modern Applied Statistics with S), 147
MATCH function, 169-170
matrix multiplication, 102
maximum likelihood in logistic regression, 34-35
calculating in Excel, 40-41
equation setup in Excel, 36-38
odds calculation in Excel, 38-39
online loan example, 35-42
probabilities calculation in Excel, 39-40
MDA (multiple discriminant analysis).
See discriminant function analysis
MDETERM function, 106
mean square
in ANOVA: Single Factor output, 78
between groups, 58-61, 67-70
defined, 58
F-Ratios worksheet, 245
SS (sums of squares) versus SSCP (sum of squares and cross products), 104-105
within groups, 67-70
mean-corrected values in discriminant function analysis, 164-167
means, ANOVA versus MANOVA results, 95-96
medical research example (single factor ANOVA), 66-72
error rate adoption, 66-67
standard error of the mean, 70-72
statistical calculations, 67-70
MINVERSE function, 180
MMULT function, 101-102, 158
Modern Applied Statistics with S (MASS), 147
multicollinearity, 91, 187-188

multiple discriminant analysis.
See discriminant function analysis

multiple regression analysis
categories as numbers, 120-122
comparison with discriminant function analysis, 122-123, 125-129
canonical correlation, 125-127
coding techniques, 127-129
eigenstructures, 133-135
relationship between coefficients, 130-133
return values, 129-130
structure coefficients, 136-137
effect coding, 80-82
explained, 6-8
general linear model versus, 79-80
LINEST function, 82-87
multivariate analysis of variance.
See MANOVA (multivariate analysis of variance)

N

named cells, 109
natural log, 34
naturally occurring clusters, classification into, 13-16
cluster analysis, 14-16. See also cluster analysis
principal components analysis, 13-14.
See also principal components analysis
negative coefficients, positive coefficients versus, 152
nominal variables, coding techniques, 127-129
noncentral F distribution, 64-66, 105
normal distributions of residuals in least squares regression, 29-31
null hypothesis, error rate adoption, 66-67
numbers, categories as, 120-122, 127-129.
See also discriminant function analysis
oblique factor rotation, 176-177, 201
observations, independence of, 45-46
odds
calculating in Excel, 38-39
converting to/from probabilities, 32-33
log odds, 33-34
online loan example (maximizing log likelihood), 35-42
equation setup in Excel, 36-38
log likelihood calculation in Excel, 40-41
odds calculation in Excel, 38-39
probabilities calculation in Excel, 39-40
opening
  Discriminant workbook, 140-141
  Principal Components add-in, 177-179
orthogonal factor rotation, 201

parametric techniques, 11
partitioning variance, 55-56
PCA. See principal components analysis
planned orthogonal contrasts, 127-128
positive coefficients, negative coefficients versus, 152
predicted variables, 18
prediction, causation versus, 18
predictor variables, 18
principal components
comparison with SAT scores, 174-175
defined, 174
determining number of, 183-185
principal components analysis
  correlation matrix, 179
  factor identification, 197-199
  inverse of R matrix, 179-182
  opening, 177-179
  sphericity test, 182-183
  Varimax rotation, 200-202
principal components add-in, 177-183
correlation matrix, 179
factor identification, 197-199
inverse of R matrix, 179-182
opening, 177-179
sphericity test, 182-183
Varimax rotation, 200-202
purpose of, 202
relationships between results, 187-192
data set variance and eigenvalues, 191-192
eigenvalues and eigenvectors, role of, 187-188
factor coefficients, eigenvalues and eigenvectors, 190-191
loadings, eigenvalues and eigenvectors, 188-190
state crime rates example, 202-206
cluster analysis on, 249-253
principal components analysis

variability in, 175-176
wine data example
  cluster analysis on, 253-256
  cross-validation, 256-258
probabilities
  assessing equation accuracy, 44-45
  calculating in Excel, 39-40
  converting to/from odds, 32-33
Process Summary worksheet, 247-248
p-value in ANOVA: Single Factor output, 79

Q
QR decomposition, 180
Quartimax factor rotation, 196

R
R matrix, inverse of, 179-182
R software, benchmarking with
discriminant function analysis results, 147-152
  coefficient signs, 152
  coefficients and constants returned, 151-152
downloading R, 147
running analysis, 149-152
saving CSV file in Excel, 148-149
k-means method results, 233-234

R2, 187
RAND function, 177, 256-258
randomly splitting data sets, 256-258
range, effect of sample size on, 28-29
raw coefficients, 159
record identifiers
  importance of, 253
  in k-means method (cluster analysis), 237-236
reference samples, access to, 8-9

regression analysis. See multiple regression analysis
regression coefficients, 83-85
regression lines. See also linear regression; logistic regression
  fitting to data, 22-24
  scaling problem, 24-25
  underlying assumptions, 25
relationships between results in principal components analysis, 187-192
eigenvalues, data set variance and, 191-192
eigenvalues and eigenvectors
  factor coefficients and, 190-191
  loadings and, 188-190
  role of, 187-188
repetition. See cross-validation
residual distribution in least squares regression, 29-31
return values
  in discriminant function analysis, 129-130
  in multiple regression, 129-130
reversing axes, 204
rotating factors. See factor analysis
running
  MANOVA workbook code, 111-112
  R software data analysis, 149-152
  Solver, 41-42

S
S programming language, 147
sample size, effect on range and standard deviation, 28-29
SAT scores, comparison with principal components, 174-175
saving CSV files in Excel, 148-149
scaling problems in regression lines, 24-25
self-selecting nature of cluster analysis, 220-223
sex bias in Berkeley admissions example (cluster analysis), 221-223
shrinkage, 171
simple structure in factor analysis, 185
single linkage method, 215-220
SKEW function, 11
Solver
finding and installing, 41
running, 41-42
source of variation in ANOVA: Single Factor output, 76
sphericity test in principal components analysis, 182-183
splitting data sets randomly, 256-258
SQRT function, 211-214
square matrices, inverse of, 180
SS (sums of squares)
in ANOVA: Single Factor output, 76-77
SSCP (sum of squares and cross products) versus, 98-102, 104-105
SSCP (sum of squares and cross products) Among and Within SSCP matrices calculations, 102-104
SS (sums of squares) versus, 98-102, 104-105
Wilks’ lambda calculations, 105-106
SSE (sum of squared errors), 238
standard deviation, effect of sample size on, 28-29
standard error of estimate, 86, 187
standard error of the mean, 70-72
standardized coefficients, 159
state crime rates example (principle components analysis), 202-206, 249-253
statistical inference, 48-52
STDEV.P function, 69
STDEV.S function, 69
structure coefficients. See also loadings
interpreting, 155-156
misleading results from, 136-137
positive versus negative signs, 152
sum of squared errors (SSE), 238
sum of squares and cross products (SSCP)
Among and Within SSCP matrices calculations, 102-104
sums of squares (SS) versus, 98-102, 104-105
Wilks’ lambda calculations, 105-106
SUMPRODUCT function, 101, 159
sums of squares (SS)
in ANOVA: Single Factor output, 76-77
sum of squares and cross products (SSCP) versus, 98-102, 104-105
SUMXMY2 function, 213-214, 241
T
T.DIST function, 85
terminology
importance of, 18-19
of logistic regression, origin of, 21-22
problems with, 16-19
causation versus prediction, 18
independent and dependent variables, 17
Total SSCP matrix
basis for, 105
Wilks’ lambda calculations, 105-106
trained analysis, untrained analysis versus, 171-172
TRANSPOSE function, 101-102, 161
tridiagonalization, 192-193
two-step approach to classification, 6
underlying assumptions in regression analysis, 25
univariate analysis of variance. See ANOVA (univariate analysis of variance)
untrained analysis, trained analysis versus, 171-172
values
expected values of variances
cholesterol treatment example, 67-70
within groups, 56-58
between groups, 58-61
mean-corrected values in discriminant
function analysis, 164-167
storing formulas as, 51
variability in principal components
analysis, 175-176
variance
Cluster Variances worksheet, 242-244
defined, 54
eigenvalues and, 191-192
expected values between groups, 58-61, 67-70
expected values within groups, 56-58, 67-70
F-ratio, 61-64
noncentral F distribution, 64-66
partitioning, 55-56
uses for, 54-55
variance error of the mean, defined, 70
variance-covariance matrix
adjustments in discriminant function
analysis, 167-169
equivalence in MANOVA workbook
output, 113-115
Varimax rotation, 200-202
VAR.P function, 55-56
VAR.S function, 55-56, 69
viewpoint adjustments in discriminant
function analysis, 123-124
visualizing MANOVA (multivariate analysis
of variance), 92-96
ANOVA results, 93
MANOVA results, 93-95
means and centroids, 95-96
Wilks' lambda
calculating in Excel, 105-106
converting to F-ratio, 107-110
relationship with F-ratio, 115-116
wine data example
cluster analysis on, 253-256
cross-validation, 256-258
Within SSCP matrix
basis for, 105
calculating in Excel, 102-104
Wilks' lambda calculations, 105-106