Contents

Acknowledgments .................................................................................................................. ix
About This Book .................................................................................................................. xi
About These Authors .......................................................................................................... xiii

Part 1 Introduction .............................................................................................................. 1
Chapter 1 Introduction ........................................................................................................ 3
Overview .............................................................................................................................. 3
Analytics Is Hot! ................................................................................................................ 4
What You Will Learn ........................................................................................................... 6
Analytics and Data Mining ................................................................................................. 7
How the Book Is Organized ............................................................................................... 7
Let’s Get Started ................................................................................................................ 8
References .......................................................................................................................... 9

Chapter 2 An Overview of the Business Analytics Process ............................................. 11
Introduction ......................................................................................................................... 11
Commonly Used Process Models ....................................................................................... 12
The Business Analytics Process ........................................................................................ 13
  Define the Problem ........................................................................................................ 13
  Prepare for Modeling ..................................................................................................... 14
  Modeling ......................................................................................................................... 15
  Deploy Model ............................................................................................................... 15
  Monitor Performance ................................................................................................... 16
Conclusion .......................................................................................................................... 17
References .......................................................................................................................... 17
Part 2 Model Building ................................................................. 19

Chapter 3 Working with Data............................................................. 21

Introduction ....................................................................................... 21
JMP Basics ......................................................................................... 22
  Opening JMP and Getting Started ......................................................... 22
  JMP Data Tables .............................................................................. 23
  Examining and Understanding Your Data ............................................. 25
  Preparing Data for Modeling .............................................................. 41
  Summary and Getting Help in JMP .................................................... 51
Exercises ............................................................................................ 51
References .......................................................................................... 53

Part 3 Model Selection and Advanced Methods ................... 55

Chapter 4 Multiple Linear Regression ........................................... 57

In the News ......................................................................................... 57
Representative Business Problems ..................................................... 58
Preview of End Result ......................................................................... 58
Looking Inside the Black Box: How the Algorithm Works ................ 59
Example 1: Housing Prices ................................................................. 62
  Applying the Business Analytics Process ........................................... 62
  Summary .......................................................................................... 79
Example 2: Bank Revenues ................................................................. 80
  Applying the Business Analytics Process ........................................... 81
  Summary .......................................................................................... 96
Exercises ............................................................................................ 97
References .......................................................................................... 99

Chapter 5 Logistic Regression .......................................................... 101

In the News ......................................................................................... 101
Representative Business Problems ..................................................... 102
Preview of the End Result .................................................................... 102
Looking Inside the Black Box: How the Algorithm Works ................ 103
Example 1: Lost Sales Opportunities .................................................. 105
  Applying the Business Analytics Process ........................................... 105
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 2: Titanic Passengers</td>
<td>115</td>
</tr>
<tr>
<td>Applying the Business Analytics Process</td>
<td>115</td>
</tr>
<tr>
<td>Summary</td>
<td>124</td>
</tr>
<tr>
<td>Key Take-Aways and Additional Considerations</td>
<td>124</td>
</tr>
<tr>
<td>Exercises</td>
<td>125</td>
</tr>
<tr>
<td>References</td>
<td>129</td>
</tr>
<tr>
<td><strong>Chapter 6 Decision Trees</strong></td>
<td>131</td>
</tr>
<tr>
<td>In the News</td>
<td>132</td>
</tr>
<tr>
<td>Representative Business Problems</td>
<td>132</td>
</tr>
<tr>
<td>Preview of the End Result</td>
<td>133</td>
</tr>
<tr>
<td>Looking Inside the Black Box: How the Algorithm Works</td>
<td>134</td>
</tr>
<tr>
<td>Classification Tree for Status</td>
<td>135</td>
</tr>
<tr>
<td>Statistical Details Behind Classification Trees</td>
<td>137</td>
</tr>
<tr>
<td>Other General Modeling Considerations</td>
<td>144</td>
</tr>
<tr>
<td>Exploratory Modeling versus Predictive Modeling</td>
<td>144</td>
</tr>
<tr>
<td>Model Cross-Validation</td>
<td>145</td>
</tr>
<tr>
<td>Dealing with Missing Values</td>
<td>149</td>
</tr>
<tr>
<td>Decision Tree Modeling with Ordinal Predictors</td>
<td>149</td>
</tr>
<tr>
<td>Example 1: Credit Card Marketing</td>
<td>150</td>
</tr>
<tr>
<td>The Study</td>
<td>150</td>
</tr>
<tr>
<td>Applying the Business Analytics Process</td>
<td>151</td>
</tr>
<tr>
<td>Case Summary</td>
<td>166</td>
</tr>
<tr>
<td>Example 2: Printing Press Yield</td>
<td>166</td>
</tr>
<tr>
<td>The Study</td>
<td>167</td>
</tr>
<tr>
<td>Applying the Business Analytics Process</td>
<td>167</td>
</tr>
<tr>
<td>Case Summary</td>
<td>180</td>
</tr>
<tr>
<td>Summary</td>
<td>181</td>
</tr>
<tr>
<td>Exercises</td>
<td>181</td>
</tr>
<tr>
<td>References</td>
<td>183</td>
</tr>
<tr>
<td><strong>Chapter 7 Neural Networks</strong></td>
<td>185</td>
</tr>
<tr>
<td>In the News</td>
<td>186</td>
</tr>
<tr>
<td>Representative Business Problems</td>
<td>187</td>
</tr>
<tr>
<td>Measuring Success</td>
<td>187</td>
</tr>
<tr>
<td>Preview of the End Result</td>
<td>188</td>
</tr>
</tbody>
</table>
# Chapter 9 Advanced Methods

Overview

Concepts in Advanced Modeling
- Bagging
- Boosting
- Regularization

Advanced Partition Methods
- Bootstrap Forest
- Boosted Tree

Boosted Neural Network Models

Generalized Regression Models
- Maximum Likelihood Regression
- Ridge Regression
- Lasso Regression
- Elastic Net

Key Take-Aways

Exercises

References

# Chapter 10 Capstone and New Case Studies

Introduction

Case Study 1: Cell Classification
- Stage 1: Define the Problem
- Stage 2: Prepare for Modeling
- Stage 3: Modeling

Case Study 2: Blue Book for Bulldozers (Kaggle Contest)
- Getting to Know the Data
- Data Preparation
- Modeling
- Model Comparison
- Next Steps

Case Study 3: Default Credit Card, Presenting Results to Management
- Developing a Management Report

Case Study 4: Carvana (Kaggle Contest)

Exercises

References
Appendix ................................................................. 323
Index ............................................................................... 327
In the News

These days our entire lives revolve around predictions. Government departments project the cost of health exchanges, the rate of economic growth, next year’s crop yields, the future birth rate and the arms buildup of unfriendly countries. Websites and retailers anticipate what we want to find and buy; oil companies gauge the best sites for drilling; pharmaceutical companies assess the
probable efficacy of molecules on a disease; while, in the background, the bobble-heads on television incessantly spew out largely irrelevant and inaccurate forecasts. In the meantime, we busy ourselves with personal projections. How long will our commute take? When will the turkey be golden? How much will the price of a stock rise? What will the future value be of a law degree? (Michael Moritz. “Are We All Being Fooled by Big Data?” January 3, 2013. Accessed at http://linkd.in/1zFuuq2.)

Representative Business Problems

Multiple linear regression is perhaps the most widely used and well-known statistical modeling tool. A straight-forward extension of simple linear regression, multiple regression is used to predict the average response value based on values of multiple predictors, or factors. For example, multiple regression can be used for:

- Identifying and optimizing critical to quality characteristics, with the goal of developing low cost and high quality products
- Predicting customer spending based on demographic information and historical buying patterns
- Developing pricing strategies based on product mix and consumer characteristics
- Establishing housing prices
- Determining the optimal timing for traffic lights to minimize traffic delays
- Predicting future product success from the results of a pilot study or trial

Preview of End Result

Suppose a model is developed to predict the amount of money a patron will spend on food per day while attending a popular professional golf tournament. A study is conducted and a model is developed based on household income and the average cost of food items sold:

\[
\text{Golf Tournament Daily Spend ($/day)} = 25 + 0.08 \times \text{Annual Household Income (in thousands)} - 0.21 \times \text{Average Cost of Food Items}
\]

So, if the average cost of food items is $10 and a patron’s household income is $100,000, the predicted spending on food for that patron is:

\[
25 + (0.08 \times 100) + (-0.21 \times 10) = 30.90 \text{ per day.}
\]
Looking Inside the Black Box: How the Algorithm Works

Consider points on a scatterplot, where \( x \) is some predictor variable and \( y \) is some response. For example, the response in the previous section is money spent per day, and one of the \( x \) or predictor variables is household income. Now think about drawing a line that best represents or fits these data points. This line is our linear model fit to the data. The line provides a predicted value of \( y \) for each value of \( x \).

The equation for this line can be expressed as:

\[
y = b_0 + b_1 x
\]

where \( b_0 \) is the \( y \)-intercept and \( b_1 \) is the slope. The \( y \)-intercept is the predicted value of \( y \) if the value of \( x \) is zero, and slope represents the change in \( y \) for every unit increase in \( x \). Since we are fitting a model using data, this equation is generally expressed as:

\[
\hat{y} = b_0 + b_1 x
\]

where \( \hat{y} \) represents the fitted value of \( y \). This line doesn’t fit the data perfectly. There is generally some difference between each response value and the line that we have drawn. This difference is called a residual, and it tells us how far the predicted value is from the observed value. This is illustrated in Figure 4.1. Each point is an observed value (the actual Daily Spend) for a given income level, and the vertical line tells us how far off each point is from the Daily Spend predicted by our linear model.

In fitting a line to the data, we are attempting to model the true unknown relationship between our predictor and our response. That is, we are trying to model reality. Our line is actually an estimate for the model. It represents the true unknown relationship between \( y \) and \( x \), which is written as:

\[
y = \beta_0 + \beta_1 x + \epsilon
\]

Here, \( y \) is the response at a given value of \( x \), \( \beta_0 \) is the true \( y \)-intercept, \( \beta_1 \) is the true slope, and \( \epsilon \) represents random error.
We estimate this true unknown model by drawing a line through our sample data that best fits the data. How do we measure “best”? Intuitively, the line that produces the smallest residual values is the best fit to our data. Because some residuals are positive and some are negative, we use an algorithm that finds the line with the smallest sum of squared residuals. This algorithm is aptly named the Method of Least Squares, and the line with the lowest sum of squared residuals, also called sum of squared errors, is referred to as the least squares regression line.

A visual representation of squared residuals is shown in Figure 4.2. Larger residuals have larger squared residuals, represented by pink squares. Smaller residuals have smaller squares. Intuitively, the least squares regression line is the line that results in the smallest squares in terms of total area.
Simple linear regression, described above, involves one response and one predictor variable. In many situations, there are several variables that can be used to predict the response. In this case, we use the term multiple linear regression, and the formula is a straightforward extension of the simple regression model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

Each predictor has a corresponding slope, or coefficient, and the response is dependent upon the values of each predictor variable. In the original example, our model had two predictors, Annual Household Income (in thousands) and Average Cost of Food Items for Sale.

Note: Figures 4.1 and 4.2 were created with the Demonstrate Regression script for illustration purposes. The script is found in JMP (starting with JMP 12) under Help > Sample Data > Teaching Scripts > Interactive Teaching Modules. Click the Help button in the script for information on how to use the script to explore fitted lines, residuals, and sums of squares.
Example 1: Housing Prices

A real estate company that manages properties around a ski resort in the United States wishes to improve its method for pricing homes. Sample data is obtained on a number of measures, including size of the home and property, location, age of the house, and a strength-of-market indicator.

The Data  

The data set contains information on about 45 residential properties near a popular North American ski resort sold during a recent 12-month period. The data set is a representative sample of the full set of properties sold during that time period (example provided by Marlene Smith, University of Colorado at Denver). The variables in the data set are:

- **Price**: Selling price of the property (in thousands of dollars)
- **Beds**: Number of bedrooms in the house
- **Baths**: Number of bathrooms in the house
- **Square Feet**: Size of the house in square feet
- **Miles to Resort**: Miles from the property to the downtown resort area
- **Miles to Base**: Miles from the property to the base of the ski resort’s facing mountain
- **Acres**: Lot size in number of acres
- **Cars**: Number of cars that will fit into the garage
- **Years Old**: Age of the house at the time it was listed in years
- **DoM**: Number of days the house was on the market before it was sold

Applying the Business Analytics Process

**Define the Problem**

The real estate company wants to develop a model to predict the selling price of a home based on the data collected. The resulting pricing model will be used to determine initial asking prices for homes in the company’s portfolio.
Prepare for Analysis

We begin by getting to know our data. As we saw in Chapter 3, we explore the distributions for each of our variables using **Analyze > Distribution**. We investigate relationships between the response and potential predictor variables using **Analyze > Multivariate Methods > Multivariate**.

Note that in this example, and throughout the modeling chapters, our focus is on particular modeling techniques. In each example, we use only a handful of methods, discussed in Chapter 3 to provide you with some familiarity with the data. However, we recommend that you use the graphical and numeric tools for exploring variables that were introduced, follow the suggestions for data preparation, and have a good understanding of each data set prior to modeling.

Exploring One Variable at a Time

Distribution output for all of the variables is shown in Figure 4.3a and Figure 4.3b. For each of these continuous variables, we see a histogram, a box plot, and various summary statistics.

Notice that the homes in this sample range in price from $160,000 to $690,000. Many of the homes have three or four bedrooms, two or three baths, are under 2,000 square feet on average, and are within twenty miles of the resort.

**Figure 4.3a: A First Look at the Data – First 5 Housing Price Variables**
Figure 4.3b: A First Look at the Data – Last 5 Housing Price Variables

When we click on the $500-600,000 bin in the Price histogram, the values for these homes are also selected (shaded) in the other graphs. As one might expect, these more expensive homes tend to be on the larger side and are closer to the resort. For example, houses in this price range tend to have four or five bedrooms and three or four baths, are generally between 2000 and 3000 square feet, and are, for the most part, within twenty miles of the resort area.

Figure 4.4: Exploring Relationships

The histograms also tell us about the shapes of the distributions, and whether there are any patterns, unusual observations or potential outliers that may cause concern. Miles to Resort (in Figure 4.4) appears to be right-skewed, while the other four variables appear...
to be more symmetric. While the data set is small, there don’t appear to be any unusually large or small values for any of the variables in Figure 4.4.

Exploring Many Variables at a Time

The Multivariate platform in JMP can be used for a more formal exploration of the relationships between the predictor variable and potential response variables. The Correlations table (Figure 4.5) shows correlations between all of the variables. Strong positive correlations are shown in blue in JMP, strong negative correlations are shown in red, and weak correlations are gray.

We’re most interested in the correlation between Price and the other variables. We see strong positive correlations between Price and the number of beds, baths, and square feet, which are measures of house size, and negative correlations with miles from the base and miles from the resort, which are both distance measures.

We’re also interested in understanding potential relationships between the predictor variables. For example, we can see strong correlations between each of the size measures (Beds, Baths, and Square Feet) and between the two distance measures (Miles to Resort and Miles to Base).

Figure 4.5: Correlations between Variables

<table>
<thead>
<tr>
<th></th>
<th>Price</th>
<th>Beds</th>
<th>Baths</th>
<th>Square Feet</th>
<th>Miles to Resort</th>
<th>Miles to Base</th>
<th>Acres</th>
<th>Cars</th>
<th>Years Old</th>
<th>DoM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>1.0000</td>
<td>0.6753</td>
<td>0.8001</td>
<td>0.6970</td>
<td>-0.5391</td>
<td>-0.6332</td>
<td>0.0251</td>
<td>0.4523</td>
<td>-0.3551</td>
<td>0.2298</td>
</tr>
<tr>
<td>Beds</td>
<td>0.6753</td>
<td>1.0000</td>
<td>0.7322</td>
<td>0.7282</td>
<td>-0.3559</td>
<td>-0.4241</td>
<td>-0.1473</td>
<td>0.1045</td>
<td>-0.3403</td>
<td>-0.0971</td>
</tr>
<tr>
<td>Baths</td>
<td>0.8001</td>
<td>0.7322</td>
<td>1.0000</td>
<td>0.7901</td>
<td>-0.3745</td>
<td>-0.4680</td>
<td>-0.1930</td>
<td>0.4357</td>
<td>-0.3267</td>
<td>0.1205</td>
</tr>
<tr>
<td>Square Feet</td>
<td>0.6970</td>
<td>0.7282</td>
<td>0.7901</td>
<td>1.0000</td>
<td>-0.1959</td>
<td>-0.2972</td>
<td>-0.1456</td>
<td>0.3728</td>
<td>-0.3037</td>
<td>-0.0110</td>
</tr>
<tr>
<td>Miles to Resort</td>
<td>0.5391</td>
<td>-0.3609</td>
<td>-0.3715</td>
<td>-0.1895</td>
<td>1.0000</td>
<td>0.9480</td>
<td>0.2958</td>
<td>-0.1584</td>
<td>0.1082</td>
<td>-0.2219</td>
</tr>
<tr>
<td>Miles to Base</td>
<td>0.6302</td>
<td>-0.4241</td>
<td>-0.4860</td>
<td>-0.2972</td>
<td>0.9480</td>
<td>1.0000</td>
<td>0.2634</td>
<td>-0.2612</td>
<td>0.2211</td>
<td>-0.1841</td>
</tr>
<tr>
<td>Acres</td>
<td>0.0251</td>
<td>-0.1473</td>
<td>-0.1930</td>
<td>-0.1456</td>
<td>0.2958</td>
<td>0.2634</td>
<td>1.0000</td>
<td>0.1474</td>
<td>-0.0295</td>
<td>0.3288</td>
</tr>
<tr>
<td>Cars</td>
<td>0.4523</td>
<td>0.1045</td>
<td>0.4357</td>
<td>0.3728</td>
<td>-0.1584</td>
<td>-0.2612</td>
<td>0.1474</td>
<td>1.0000</td>
<td>-0.2714</td>
<td>0.3107</td>
</tr>
<tr>
<td>Years Old</td>
<td>0.3551</td>
<td>-0.3403</td>
<td>-0.3267</td>
<td>-0.3037</td>
<td>0.1082</td>
<td>0.2211</td>
<td>-0.0295</td>
<td>-0.2714</td>
<td>1.0000</td>
<td>0.0077</td>
</tr>
<tr>
<td>DoM</td>
<td>0.2298</td>
<td>-0.0971</td>
<td>0.1205</td>
<td>-0.0110</td>
<td>-0.2219</td>
<td>-0.1841</td>
<td>0.3288</td>
<td>0.3137</td>
<td>0.0077</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

These relationships can be examined visually with the scatterplot matrix, which displays all of the two-way scatterplots between each pair of variables. Figure 4.6 shows the correlations and scatterplot matrix for three of the variables: Price, Miles to Resort, and Miles to Base. In the first row of the matrix, the y-axis for each of the graphs is Price, and the x-axis corresponds to the variable on the diagonal. So, for example, the two scatterplots in the first row display the relationship between Price and Miles to Resort and between Price and Miles to Base.
In each scatterplot, the correlation is displayed graphically as a density ellipse. The tighter (less circular) the ellipse, the stronger the correlation. The direction of the ellipse indicates whether the correlation is positive (the ellipse slopes up) or negative (the ellipse slopes down).

Looking at the variables, individually and together, helps us understand our data and potential relationships. We are starting to get a sense of the data and the variables that might need to be included in the model.

Note that other graphical tools, such as Fit Y by X (under the Analyze menu) and Graph Builder (under the Graph menu) can also be useful in exploring potential bivariate and multivariate relationships. We urge you to explore this data set on your own using all of these tools.

Figure 4.6: Visually Examining Correlations with a Scatterplot Matrix

Build the Model
Our goal is to develop a model to predict the selling price of a home based on available data. Multiple linear regression is one of the core methods that can be used to develop a model to predict a continuous response from multiple predictor variables.
We begin by fitting a model using Price as the Y (response variable) and all of the potential factors as model effects using Analyze > Fit Model, as shown in Figure 4.7. Click Run to run the model.

**Figure 4.7: Fit Model Dialog Window**

The results are shown in Figure 4.8. The **Effect Summary** table shows each of the terms in the model, sorted in ascending order of the p-value.

The **Actual by Predicted Plot** provides a graphical indication of the overall significance of the model. The closer the data points are to the diagonal line, the better our model does at explaining the variation in the response. (The solid diagonal line on this plot is the line where the actual value equals the predicted value.)

The **Summary of Fit** table provides key statistics, such as **RSquare** and **RSquare Adj** (adjusted R Square). RSquare indicates the percent of variation in the data that is explained by our model (0.81, or 81%). Because RSquare can be inflated simply by adding additional predictors to the model, the adjusted RSquare is sometimes used instead of RSquare for comparing models with more than one predictor (the “adjustment” applied to the **RSquare Adj** is based on the number of terms in the model).
The Analysis of Variance table indicates that the overall model is statistically significant. The $p$-value, reported as $\text{Prob} > F$, is $< .0001$.

The Parameter Estimates table provides coefficients for our model, along with $p$-values for each of the terms in the model.

Other output (effects tests, a residual plot, and leverage plots) is also provided by default, and additional options are available under the top red triangle.

Figure 4.8: Fitted Model

In Figure 4.8, we see that only two of the predictors (Baths and Acres) are significant at the 0.05 level, given that all of the other variables are in the model. Does this mean that none of the other predictors are important in predicting housing prices? We’ll want to reduce this model to include only those variables that, in combination, do a good job of predicting the response. This is known as the principle of parsimony: the simplest model that can predict the response well is often the best model. Given that we have nine predictor variables, there can be many possible models to predict Price, from very simple models to more complex. In general, our goal is to find a concise model that makes
sense, fits the data, and predicts the response well (Hansen, 2001). But, since significance is dependent upon which other variables are in the model, it is difficult to determine which terms to keep in the model and which to remove.

For illustration, we click **Remove** at the bottom of the **Effect Summary** table to remove **Baths** from the model (top, in Figure 4.9). **Acres** is no longer significant at $\alpha = 0.05$, but **Square Feet** is now significant (bottom, in Figure 4.9).

**Figure 4.9: Effect Summary Table with and without Baths**

<table>
<thead>
<tr>
<th>Source</th>
<th>LogWorth</th>
<th>PValue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acres</td>
<td>1.880</td>
<td>0.01318</td>
</tr>
<tr>
<td>Baths</td>
<td>1.846</td>
<td>0.01424</td>
</tr>
<tr>
<td>Square Feet</td>
<td>0.983</td>
<td>0.10425</td>
</tr>
<tr>
<td>Miles to Base</td>
<td>0.319</td>
<td>0.48023</td>
</tr>
<tr>
<td>Miles to Resort</td>
<td>0.285</td>
<td>0.51915</td>
</tr>
<tr>
<td>Beds</td>
<td>0.265</td>
<td>0.54336</td>
</tr>
<tr>
<td>Cars</td>
<td>0.209</td>
<td>0.61769</td>
</tr>
<tr>
<td>Years Old</td>
<td>0.207</td>
<td>0.62093</td>
</tr>
<tr>
<td>DoM</td>
<td>0.092</td>
<td>0.80958</td>
</tr>
</tbody>
</table>

Part of the issue is that some of the variables are correlated with other variables in the model. Recall the correlation and scatterplot for **Miles to Resort** and **Miles to Base** (Figure 4.6). There is a very strong correlation between these two predictor variables. This means that they are somewhat redundant to one another. In fact, the resort and the base are in nearly the same geographic location.
A Bit About Multicollinearity

When two or more predictors are correlated with one another, the term *multicollinearity* is used. If multicollinearity is severe, then it is difficult to determine which of the correlated predictors are most important. In addition, the coefficients and standard errors for these coefficients may be inflated and the coefficients may have signs that don’t make sense.

A measure of multicollinearity is the *VIF* statistic, or *Variance Inflation Factor*. The VIF for a predictor, $VIF_j$, is calculated using the following formula:

$$VIF_j = \frac{1}{1 - R^2_{X_j}}$$

For each predictor, $X_j$, a regression model is fit using $X_j$ as the response and all of the other $X$ variables to predict $X_j$. The $R^2$ for that model fit ($R^2_{X_j}$) is calculated, and is then used to calculate the $VIF_j$. An $R^2_{X_j}$ of 0.9 results in a $VIF_j$ of 10, while an $R^2_{X_j}$ of 0.99 results in a $VIF_j$ of 100.

If the VIF is 1.0, then each of the predictor variables is completely independent of the other predictor variables. But if the VIF is large (say, greater than 10), then the multicollinearity is a problem that should be addressed (Neter, 1996). In some cases, this can be resolved by removing a redundant term from the model. In more severe cases, simply removing a term will not address the issue. In these cases, variable reduction techniques such as Principal Components Analysis (PCA), Partial Least Squares (PLS), tree-based methods (covered in Chapter 6), and generalized regression methods (Chapter 9) are recommended.

In Figure 4.10 (on the left), we see VIFs for the original model. To display VIFs, right-click on the **Parameter Estimates** table and select **Columns > VIF**. The VIFs for most of the predictors are relatively small (< 5), while the VIFs for **Miles to Resort** and **Miles to Base** are both over 10, indicating that multicollinearity is a problem. Since we’ve learned that these two variables are largely redundant to one another, it makes sense to re-fit the model with only one of these variables. Subject matter knowledge can be used to determine which variable to remove. On the right in Figure 4.9, we see the results after removing **Miles to Resort** from the model. Notice that all of the VIFs are now low. In

<table>
<thead>
<tr>
<th>Predictor</th>
<th>VIF Original</th>
<th>VIF After Removal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mtn. Elev.</td>
<td>1.5</td>
<td>1.2</td>
</tr>
<tr>
<td>Miles to Resort</td>
<td>10.5</td>
<td>1.2</td>
</tr>
<tr>
<td>Miles to Base</td>
<td>11.0</td>
<td>1.2</td>
</tr>
<tr>
<td>Subject Matter</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
addition, Miles to Base is now significant, and the coefficient and the standard error for Miles to Base have both changed substantially!

Figure 4.10: Variance Inflation Factor, VIF

<table>
<thead>
<tr>
<th>Term</th>
<th>Estimate</th>
<th>Std Error</th>
<th>t Ratio</th>
<th>Proba&gt;</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>179.45965</td>
<td>51.25743</td>
<td>3.50</td>
<td>0.0015</td>
<td></td>
</tr>
<tr>
<td>Beds</td>
<td>9.766937</td>
<td>15.9136</td>
<td>0.61</td>
<td>0.5434</td>
<td>3.5544519</td>
</tr>
<tr>
<td>Baths</td>
<td>53.275621</td>
<td>20.65062</td>
<td>2.58</td>
<td>0.0142*</td>
<td>4.3822145</td>
</tr>
<tr>
<td>Square Feet</td>
<td>0.0445899</td>
<td>0.026733</td>
<td>1.67</td>
<td>0.1043</td>
<td>3.4757465</td>
</tr>
<tr>
<td>Miles to Resort</td>
<td>-1.698122</td>
<td>2.60755</td>
<td>-0.65</td>
<td>0.5191</td>
<td>13.822225</td>
</tr>
<tr>
<td>Miles to Base</td>
<td>-1.878499</td>
<td>2.632567</td>
<td>-0.71</td>
<td>0.4802</td>
<td>14.510574</td>
</tr>
<tr>
<td>Acres</td>
<td>4.6547176</td>
<td>1.782149</td>
<td>2.61</td>
<td>0.0132*</td>
<td>1.2512784</td>
</tr>
<tr>
<td>Cars</td>
<td>3.805792</td>
<td>12.67613</td>
<td>0.50</td>
<td>0.6177</td>
<td>1.7853544</td>
</tr>
<tr>
<td>Years Old</td>
<td>-2.723892</td>
<td>0.547926</td>
<td>-0.50</td>
<td>0.6209</td>
<td>1.2901804</td>
</tr>
<tr>
<td>DoM</td>
<td>0.0324497</td>
<td>0.136351</td>
<td>0.24</td>
<td>0.9096</td>
<td>1.9627487</td>
</tr>
</tbody>
</table>

To further assist in refining our model, after removing Miles to Resort, we’ll rely on an automated variable selection approach, stepwise regression. We proceed with stepwise regression to identify the best subset of significant factors. Stepwise is a method, or a Personality, available in the Fit Model dialog (Figure 4.11). Note that stepwise regression does not address multicollinearity. If correlated terms are used as inputs, stepwise may not result in the “best” model because variable selection will be determined by which variables are selected first.
Figure 4.11: Fit Model Stepwise Dialog

Stepwise regression provides a number of stopping rules for selecting the best subset of variables for the model. The default rule is Minimum BIC, or minimum Bayesian Information Criterion. The Direction, which is set to Forward by default, indicates that variables will be added to the model one at a time. After you click Go, the model with the smallest BIC statistic is selected.

Another common rule, which works in a similar manner, is Minimum AICc (Akaike’s Information Criterion, with a correction for small sample sizes). Both of these rules attempt to explain the relationship between the predictors and the response, without building models that are overly complex in terms of the number of predictors. Since different criteria are used to determine when to stop adding terms to the model, these stopping rules may lead to different “best” models (Burnham, 2002).

We will develop a model using each criterion and then compare results. First, we use the default Minimum BIC criterion, and click Go to start the selection process. We identify four factors for the model: Baths, Square Feet, Acres, and Miles to Base (see Figure 4.12).
Figure 4.12: Stepwise Regression Variable Selection Using BIC

Next, we change the stopping rule to Minimum AICc, click Remove All to clear the estimates from the BIC model, and then click Go. In this example, BIC and AICc yield the same set of factors (AICc output not shown). This isn’t always the case.

We have identified a common set of factors for the model using two different stopping criteria. To run this regression model, select Make Model. Then in the Fit Model dialog, click Run (or, simply select Run Model from within the Stepwise platform).

Recall that our original model (Figure 4.8), with nine predictors, had only two significant terms and an adjusted R Square of 0.76. The results of fitting this reduced model are shown in Figure 4.13. As expected, this model is significant (Prob > F < .0001), and the four terms in the model are also significant. The adjusted R square is 0.782, which is slightly higher than our original model.
Before using this model, we need to check a few key assumptions. Namely, that our model errors are independent, have equal variance, and are normally distributed. Another key assumption is that the relationship between our response and the predictors is linear (i.e., that there isn’t an underlying non-linear relationship).

The variation in the residuals, which is another word for the errors, shows us the variation in the response that could not be explained by the model that we have fit. Plots of residuals can be used to check that our assumptions about the model errors were correct. The default residual plot in JMP shows the residuals for each point plotted against predicted values. If our model assumptions are met, the points should be randomly scattered about the center line (zero), with no obvious pattern (just a cloud of seemingly random points). Other residual plots are also available (under the red triangle > Row Diagnostics), and the residuals can be saved (using red triangle > Save Columns > Residuals) and evaluated using Distribution or the Graph Builder.

The Residual by Predicted plot (see Figure 4.14) shows a somewhat curved pattern. That is, the largest residuals are at the lower and higher predicted values, while the
The pattern that we see in the residuals may also be due to outliers or influential observations, which can tilt or warp the regression model.

**Figure 4.14: Examining Residuals**

![Residual by Predicted Plot](image)

A statistic that helps us determine whether particular points are influencing the model is *Cook’s D*, or *Cook’s Distance*. Cook’s D values for each observation can be saved to the data table, and then plotted using the **Analyze > Distribution** platform (see the plot of Cook’s D values in Figure 4.15). To save Cook’s D values from the Fit Model output window, click the red triangle and select **Save Columns > Cook’s D Influence**.

A high Cook’s D value for a particular observation indicates that the model predictions with and without that observation are different. What is considered high? A general rule of thumb is that any Cook’s D value >1 is worthy of investigation (Cook, 1982). Observation #7, with a Cook’s D value over 6, has a large influence on our model.
To illustrate how an influential point can impact model, see Figure 4.16. We use Analyze > Fit Y by X, with Price as Y, Response and Acres as X, Factor, and fit a line (select Fit Line from the red triangle). Then, we exclude observation 7, and again select Fit Line. The resulting regression lines are labeled (using the Annotate tool from the toolbar in JMP).

Note the difference in the slopes for fitted regression lines with and without observation 7 included in the model! Clearly, these two models will result in different predicted values, particularly for properties with higher acreage.

Upon investigation, we find that this property is actually a 40-acre farm rather than a residential property. The focus of this pricing study is residential properties. Since this
property is not of direct interest in this study and is influencing our predictions, we exclude and hide observation 7 (use Rows > Hide and Exclude) and proceed without it.

Since the data set is small and stepwise was performed using observation 7, the procedure may result in a different reduced model if run without this observation. We return to stepwise, and find that the same reduced model is produced without this observation (Figure 4.17).

Compared to the model in Figure 4.13, this model has a higher R Square Adjusted (0.84), and most of the p-values for the terms in the model are lower. In particular, the p-value for Acres has dropped from 0.0012 to <.0001 (and the coefficient is now much larger). In addition, the residuals now appear more randomly scattered about the center line, with no obvious patterns, evidence of curvature, or lack of constant variance.

**Figure 4.17: Model and Residuals after Removing Influential Point**

<table>
<thead>
<tr>
<th>Summary of Fit</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RSquare</td>
<td>0.854004</td>
</tr>
<tr>
<td>RSquare Adj</td>
<td>0.83903</td>
</tr>
<tr>
<td>Root Mean Square Error</td>
<td>52.88352</td>
</tr>
<tr>
<td>Mean of Response</td>
<td>394.8545</td>
</tr>
<tr>
<td>Observations (or Sum Wghts)</td>
<td>44</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Analysis of Variance</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>DF</td>
</tr>
<tr>
<td>Model</td>
<td>4</td>
</tr>
<tr>
<td>Error</td>
<td>39</td>
</tr>
<tr>
<td>C. Total</td>
<td>43</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter Estimates</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Term</td>
<td>Estimate</td>
</tr>
<tr>
<td>Intercept</td>
<td>175.46632</td>
</tr>
<tr>
<td>Baths</td>
<td>63.61719</td>
</tr>
<tr>
<td>Square Feet</td>
<td>0.0472992</td>
</tr>
<tr>
<td>Miles to Base</td>
<td>-3.025002</td>
</tr>
<tr>
<td>Acres</td>
<td>12.462947</td>
</tr>
</tbody>
</table>

Given that the houses are scattered around the geographic area surrounding the resort and there is no obvious clustering of points in the residual plot, we have some additional comfort that the independence assumption is also met.

For additional confirmation that the normality assumption has been met, we can save the residuals to the data table (click the red triangle and select Save Columns > Residuals), and then use a histogram and normal quantile plot to check normality (use **Distribution**, ...
and select **Normal Quantile Plot** from the red triangle). We leave it to the reader to confirm that the normality assumption has indeed been met.

Satisfied with our final model, we can now use this model to predict home prices. The terms in our model, and their coefficients, are given in the **Parameter Estimates** table (left, in Figure 4.18). Each estimate tells us how much the predicted selling price changes with a change in the value of the predictor.

The parameter estimates can be rewritten as a formula, or prediction expression (right, in Figure 4.18). To calculate the selling price of a home, we simply need to plug in the number of baths, the square feet, the miles to base, and the acres into the formula.

**Figure 4.18: Parameter Estimates and Prediction Expression**

| Term           | Estimate | Std Error | t Ratio | Prob>|t| |
|----------------|----------|-----------|---------|------|
| Intercept      | 175.46632| 30.1538   | 5.82    | <.0001*|
| Baths          | 63.81719 | 14.57999  | 4.36    | <.0001*|
| Square Feet    | 0.0472682| 0.019337  | 2.45    | 0.0191*|
| Miles to Base  | -3.025002| 0.894486  | -4.36   | <.0001*|
| Acres          | 12.462947| 2.237932  | 5.57    | <.0001*|

To display the prediction expression, click the red triangle and select **Estimates > Show Prediction Expression**. To save this formula to the data table, click the red triangle and select **Save Columns > Prediction Formula**.

The model can also be explored graphically using the **Prediction Profiler** (Figure 4.19). To access the profiler, select **Factor Profiling > Profiler** from the red triangle.

The profiler shows the predicted response (on the far left) at specified values of each of the predictor values (given at the bottom). The initial values for the predictors are predictor averages, and vertical red lines are drawn at these values. The starting value for the response is also the overall average (the mean **Price** in this example), and the bracketed values are the 95% confidence interval for the average. The confidence interval can be used to determine a margin of error for the prediction. The margin of error is the half-width of the confidence interval, or half the range of the interval. In this example, the margin of error is approximately \((410.984 - 378.732)/2 = $16.125K\).
Drag the vertical lines for a predictor to change the value for that predictor. The slopes of the lines for each predictor indicate whether predicted Price will increase or decrease if the predictor value increases, assuming that the other predictor values are held constant.

**Figure 4.19: Using Prediction Profiler**

Summary

In this example, we have created a regression model for home selling prices using historical data and have assessed model assumptions to ensure that the model makes sense. If we are satisfied with our model’s suitability and performance, we can put the model to use to predict selling prices of new homes entering the market in this geographic region. For example, our model tells us that the predicted selling price of a 2500 square foot home with 3 baths that is 13 miles to the base and sits on one acre is just over $457,704, with a margin of error of approximately $21.5K. (You can confirm this by entering these values into the Prediction Profiler.)

Of course, we might ask if this is the best possible model to predict selling price. Can we build a better model? Our margin of error is relatively large, and our standard deviation (reported as Root Mean Square Error in Figure 4.17) is just under $53K. Can we develop a model that provides more precise predictions? What if we built a model using information on a larger sample of houses? What if we include additional information for each of the houses sold, such as recent renovations, measures of home quality, or the time of year that the home was first put on the market? Would this lead to a better model?

We should also keep in mind that all models need to be updated periodically. Housing prices change over time, so the model to predict housing prices should be updated to stay current and reflect these changes.
Example 2: Bank Revenues

A bank wants to understand how customer banking habits contribute to revenues and profitability. The bank has the customer age and bank account information, such as whether the customer has a savings account, if the customer has received bank loans, and other indicators of account activity.

The Data BankRevenue.jmp

The data set contains information on 7420 bank customers:

- **Rev_Total**: Total revenue generated by the customer over a 6-month period.
- **Bal_Total**: Total of all account balances, across all accounts held by the customer.
- **Offer**: An indicator of whether the customer has received a special promotional offer in the previous one-month period. Offer=1 if the offer was received, Offer=0 if it was not.
- **AGE**: The customer’s age.
- **CHQ**: Indicator of debit card account activity. CHQ=0 is low (or zero) account activity, CHQ=1 is greater account activity.
- **CARD**: Indicator of credit card account activity. CARD=0 is low or zero account activity, CARD=1 is greater account activity.
- **SAV1**: Indicator of primary savings account activity. SAV1=0 is low or zero account activity, SAV1=1 is greater activity.
- **LOAN**: Indicator of personal loan account activity. LOAN=0 is low or zero account activity, LOAN=1 is greater activity.
- **MORT**: Indicator of mortgage account tier. MORT=0 is lower tier and less important to the bank’s portfolio. MORT=1 is higher tier and indicates the account is more important to the bank’s portfolio.
- **INSUR**: Indicator of insurance account activity. INSUR=0 is low or zero account activity, INSUR=1 is greater activity.
- **PENS**: Indicator or retirement savings (pension) account tier. PENS=0 is lower balance and less important to bank’s portfolio. PENS=1 is higher tier and of more importance to the bank’s portfolio.
Check: Indicator of checking account activity. Check=0 is low or zero account activity, Check=1 is greater activity.

CD: Indicator of certificate of deposit account tier. CD=0 is lower tier and of less importance to the bank’s portfolio. CD=1 is higher tier and of more importance to the bank’s portfolio.

MM: Indicator of money market account activity. MM=0 is low or zero account activity, MM=1 is greater activity.

Savings: Indicator of savings accounts (other than primary) activity. Savings=0 is low or zero account activity, Savings=1 is greater activity.

AccountAge: Number of years as a customer of the bank.

Applying the Business Analytics Process

Define the Problem
We want to build a model that allows the bank to predict profitability for a given customer. A surrogate for a customer’s profitability that is available in our data set is the Total Revenue a customer generates through their accounts and transactions. The resulting model will be used to forecast bank revenues and guide the bank in future marketing campaigns.

Prepare for Modeling
We begin by looking at the variable of interest, total revenue (Rev_Total) using Graph > Graph Builder. Rev_Total is highly skewed, which is fairly typical of financial data (Figure 4.20).

Note: To explore the underlying shape of the distribution, select the Grabber (hand) tool from your toolbar, click on the graph and drag up and down.
In regression situations, highly skewed data can result in a poorly fitting model. A transformation that can often be used to normalize highly skewed data, where all of the values are positive, is a log (natural logarithm) transformation (see Ramsey and Shafer, 2002, page 68).

We apply a log transformation to the Rev_Total variable directly in the Graph Builder and reexamine the distribution (Figure 4.21). To apply this transformation, right-click on the variable in the variable selection list, and select Transform > Log. Then, to save the transformation to the data table, right-click on Log(Rev_Total) and select Add to Data Table.

This transformation gives us a much less skewed and more symmetric distribution, so we use Log(Rev_Total) for the rest of our analysis.

A similar examination of the total account balance (Bal_Total), which also has a skewed distribution, leads to using the Log(Bal_Total) in our analyses.
Figure 4.21: Transformed Total Revenue Using Log Transformation

The relationship between the log total revenue and log total account balance is shown in the scatterplot in Figure 4.22. The relationship appears nearly linear at lower account balances—higher account balances generally have higher revenues. But the relationship seems to change at the higher account balances.

Figure 4.22: Relationship between Log(Rev_Total) and Log(Bal_Total)
Now we examine the other variables. We can see their distributions and also their relationship to $\log(\text{Rev\_Total})$. Many of the variables are categorical, with two levels. Higher revenue values are selected in Figure 4.23, and we can see this selection across the other variables in our data set. (The **Arrange In Rows** option under the red triangle was used to generate Figure 4.23; not all variables are displayed.) Other than total account balance, $\log(\text{Bal\_Total})$, there is no variable that stands out as being strongly related to revenue.

As we have discussed, other graphical and analytic tools can be used to understand the data and explore potential relationships, such as **Fit Y by X** and **Graph Builder**. In addition, the **Data Filter** (under the **Rows** menu) and **Column Switcher** (under the **red triangle > Scripts** in any output window) are dynamic tools that allow you to dive deeper into your data to explore variables of interest and potential relationships. Again, we encourage you to explore the data using these tools on your own. See Chapter 3 for discussion and illustration of different exploratory tools.

**Note:** Recall that within JMP there are a number of preferences that can be set (under **File > Preferences** or **JMP > Preferences** on a Mac), and all JMP output is customizable with your mouse and keystrokes. Going forward, we periodically resize graphs and change axis scaling to better fit content on the page, and change marker sizes or colors to improve interpretability. We also turn off shaded table headings in output to provide a cleaner display (within **Preferences, Styles > Report Tables**).
Figure 4.23: Relationships between Transformed Variables and Other Variables

Build the Model
We now build a regression model to predict Log(Rev_Total) using Fit Model and the 15 potential predictor variables. There are some immediate signs of trouble (Figure 4.24). At the top of the Fit Least Squares window, we see some unexpected output, Singularity Details. This means that there are linear dependencies between the predictor variables. The first row of this table, LOAN[0] = CD[0], indicates that JMP can’t tell the difference between these two variables, LOAN and CD. The second line indicates that JMP can’t tell the difference between INSUR, MM, and Savings.

The cause of this problem is illustrated in the Distribution output in Figure 4.25. The distributions of these three variables are identical. Every time LOAN = 1 (a customer has high loan activity), MM and Savings are also 1 (money market and savings activity are also high). The variables within each grouping are completely redundant to one another!

The result of this problem is seen in the parameter estimates table. JMP can’t estimate all of these coefficients, indicating that the estimates for LOAN and INSUR are biased, and the estimates for CD, MM, and Savings are zeroed. JMP can estimate some of the parameters for the redundant variables (these estimates are biased), but not all (these are zeroed). Whether the variables appear as biased or zeroed depends entirely on the order
in which they were entered into the model—those entered first into the model are displayed as biased.

**Figure 4.24: Fit Least Squares with Singularity**

![Table and Figure]

*Table:*

| Term       | Estimate | Std Error | t Ratio | Prob>|t| |
|------------|----------|-----------|---------|------|
| Intercept  | 0.3113431| 0.049687  | 6.27    | <.0001* |
| Bal_Total  | 2.3813e-5| 1.27e-6   | 18.74   | <.0001* |
| Offer[0]   | -0.106738| 0.027105  | -3.94   | <.0001* |
| AGE        | -0.002446| 0.000645  | -3.79   | 0.0002* |
| CHQ[0]     | -0.017754| 0.014937  | -1.19   | 0.2346  |
| CARD[0]    | -0.002709| 0.040624  | -0.07   | 0.9468  |
| SAV1[0]    | -0.010189| 0.017976  | -0.57   | 0.5709  |
| LOAN[0]    | -0.599056| 0.023225  | -25.79  | <.0001* |
| MORT[0]    | 0.0162212| 0.023909  | 0.68    | 0.4975  |
| INSUR[0]   | -0.091974| 0.019419  | -4.74   | <.0001* |
| PENS[0]    | -0.00232 | 0.013493  | -0.17   | 0.8635  |
| Check[0]   | -0.310457| 0.039308  | -7.90   | <.0001* |
| CD[0]      | 0        | 0         | .       | .      |
| MM[0]      | 0        | 0         | .       | .      |
| Savings[0] | 0        | 0         | .       | .      |
| AccountAge | -0.005122| 0.003674  | -1.39   | 0.1633  |
We refit the model without the redundant variables. In JMP 12, this can be done using the Remove button at the bottom of the Effect Summary table. We keep LOAN (and eliminate CD), and keep INSUR (eliminating MM and Savings). Note that this was an arbitrary decision: subject matter knowledge should guide the decision as to which redundant variables to remove (and which variables to keep in the model). As we remove each variable (or term), the Singularity Details table updates, along with all of the other statistical output. JMP is now able to estimate coefficients for each of the parameters (Figure 4.26).
A quick check of the VIFs indicates that multicollinearity is not a serious issue (Figure 4.26).

Since we have 11 remaining potential predictor variables, we use again stepwise regression to help with variable selection. We use the Stepwise Personality in the Fit Model platform, with Log(Rev_Total) as our Y and the other variables as model effects. For this example, we use the Minimum AICc stopping rule.

Stepwise selects six variables for the model. These are checked under Current Estimates in Figure 4.27. Note that when using AICc (or BIC), the resulting models may include terms that are not significant. This is because both AICc and BIC build models based on important effects (effects that explain the relationship between the response and the predictors) rather than searching for significant effects (see Burhnam, 2002). However, in this example, all six selected variables have low p-values.
We now run this model, and explore the results (Figure 4.28). As expected, the overall model is significant with a p-value < .0001, as are all of the terms in the model. The R-Square is 0.5986, indicating that our model explains nearly 60% of the variation in the response.
Before interpreting the results of the regression model, we check that the regression assumptions are met. Since the data are from over 7400 different customers, we have some assurance that the independence assumption is met. The default residual versus predicted value plot (Figure 4.29) shows some diagonal striations in the lower left corner.
Figure 4.29: Residual versus Predicteds

To explore these values, we use the lasso tool on the toolbar to select the observations, activate the data table, and then use the F7 function key to scroll through these selected observations in the data table. The first strip on the left corresponds to revenue $0.01, and the second is revenue $0.02. This is the result of the fact that there are many customers who generate little, if any, revenue for the bank.

Otherwise, points appear randomly scattered around the center line (zero), and the residual plot shows no obvious evidence of unusual patterns.

For further exploration of the regression assumptions, we save the residuals to the data table (under the red triangle, select Save Columns > Residuals), and use the Distribution platform to generate a histogram with a normal quantile plot (Figure 4.30). These plots provide evidence that the normality assumption has been met.
We also see (in Figure 4.30) that there are no serious outliers. A quick peek at Cook’s D values (Figure 4.31) confirms that there are no highly influential observations. No single point is exerting too much influence over our model.
After investigating residuals and looking at Cook’s D values, we have confidence that the regression assumptions have been satisfied. Our final model, shown in Figure 4.32, includes the following variables:

- The total account balance (Log(Bal_Total))
- Whether the customer received a promotional offer (Offer)
- Credit card activity (CARD)
- Personal loan account activity (LOAN),
- Insurance account activity (INSUR)
- Checking account activity (Check)

All of the significant variables except Log(Bal_Total) are binary categorical variables. For the continuous predictor, Log(Bal_Total), the coefficient in the parameter estimates table (top in Figure 4.32) indicates how the revenues change as the account balance changes. The positive coefficient indicates that revenues increase on average as account balances increase. The coefficient value itself is a little difficult to interpret because it reflects the transformation of Rev_Total to Log(Rev_Total).

For each of the two-level categorical predictors, the parameter estimates show how the average response changes at the low level of each predictor. For example, the coefficient for CARD[0] is negative 0.7964. This indicates that log revenues are 0.7964 lower on average if credit card activity is low, and 0.7964 higher on average if credit card activity is high. The coefficients for LOAN, Check, and INSUR are all positive, indicating that low activity in these three accounts leads to higher revenues.

**Note:** When fitting regression models in JMP, two-level categorical predictors are automatically transformed into coded indicator variables using a -1/+1 coding scheme. The parameter estimate is reported for the lowest level or value of the predictor. In this example, CARD is a nominal predictor with levels with 0 and 1. The term in the reduced model is represented as CARD[0], and the parameter estimate is -0.7964 (see Figure 4.32). The estimate for CARD[1], which is not reported, is +0.7964. To display both estimates, select **Expanded Estimates** from the **top red triangle > Estimates**.

Many statistical software packages require dummy coding of categorical predictors, using a 0/1 “dummy” or “indicator” coding scheme. This is done prior to fitting the model, and results in different parameter estimates and a different interpretation of the estimates. For example, the parameter estimate for CARD, using 0/1 dummy coding, is 1.5928 instead of -0.7964. The sign is different, and the estimate is exactly
twice the magnitude. To confirm this, change the modeling type for CARD to **Continuous** (to tell JMP to use dummy coding) and refit the reduced model shown in Figure 4.28. Note that, although the parameter estimates are different, the two coding schemes produce identical model predictions.

To view the indicator-coded version of the parameter estimates in the **Fit Least Squares** output, select **Indicator Parameterization Estimates** from the top red triangle > Estimates. Further details of how JMP transforms categorical factors can be found in the Statistical Details section of the book *Fitting Linear Models* (under Help > Books).

The prediction profiler (bottom of Figure 4.32) can help us see the impact of changes in values of the predictor variables on Log(Rev_Total).

**Figure 4.32: Exploring the Reduced Model with the Prediction Profiler**

<table>
<thead>
<tr>
<th>Parameter Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>Log(Bal_Total)</td>
</tr>
<tr>
<td>Offer[0]</td>
</tr>
<tr>
<td>CARD[0]</td>
</tr>
<tr>
<td>LOAN[0]</td>
</tr>
<tr>
<td>INSUR[0]</td>
</tr>
<tr>
<td>Check[0]</td>
</tr>
</tbody>
</table>

Clearly, Log(Bal_Total) has a large positive effect on the response. Three predictors, Offer, LOAN, and INSUR, while significant, have a relatively small effect on the response.

To show the predicted values for each bank customer, the prediction equation (the formula) can be saved to the data table (red triangle, **Save Columns > Prediction Formula**). Unfortunately, these are the log predicted values, which are difficult to interpret.
The inverse transformation (in this case the exponential, or Exp function) can be used to see the predicted values on the original scale. To apply this transformation, create a new column in the data table (we’ve named this column Pred Rev_Total). Then, right-click on the column and select Formula to open the Formula Editor, and use the Transcendental > Exp function from the Functions (grouped) list (see Figure 4.33).

Note that this formula can be created using a shortcut. Simply right-click on the saved prediction formula column, and select New Formula Column > Transform > Exp. JMP will create the new column with the stored formula shown in Figure 4.33.

Figure 4.33: Transforming Predicted Log(Rev_Total) to Predicted Rev_Total

Now, we can explore the distribution of these values using Distribution or Graph Builder (Figure 4.34).
Figure 4.34: Distribution of Predicted Rev_Total

We can also explore the formula itself using Graph > Profiler (Figure 4.35). Select the transformed prediction formula as the Y, Prediction Formula, and check the Expand Intermediate Formulas box to drill down to the original saved prediction formula. Now, we can readily see and explore the impact of changes to each of the variables on the predicted revenues in the original scale.

Figure 4.35: Prediction Profiler for Predicted Rev_Total

Summary

It is clear that high account balance customers, and those who use their credit card frequently, generate more revenue. What is curious is that high checking account usage seems to indicate lower revenue, and that customers with higher activity on the loan and insurance accounts have lower predicted revenue on average.

Was the promotional offer was successful? That is, did it lead to increased revenue? For a customer maintaining an account balance of $82,500, with low credit card, loan, insurance, and checking account activity, the promotional offer increased revenues from $11.08 to $12.75 on average. If this same customer had high credit card activity instead of
low, the predicted revenue increased from $54.5 to $62.7. However, this analysis does not determine return on investment. Further information would need to be gathered to determine the cost of the promotional offer program and to examine the increased revenue relative to that cost. All of these insights lead to more questions, with new business problems to solve.

**Exercises**

**Exercise 4.1:** In this exercise, we use the HousingPrices.jmp data. In this chapter, we built a predictive model for Price, but limited model terms to main effects (that is, the predictors themselves). This is due, in part, to the fact that our data set is very small. However, other possible model effects include interactions and squared terms (quadratic effects).

Consider a model with all of the original predictors, plus one two-factor interaction and one quadratic term. Pick one interaction and one quadratic term that you think might be significant in predicting house prices.

Build a model using all of the original model effects and these two new terms.

1. Add all of the terms to the model.
2. Add the interaction term. Select the two terms from the Select Columns list and click Cross.
3. Add the squared term. Select the term in both the Select Columns list and the Model Effects list and click Cross.

Using this model, repeat the analysis illustrated in Example 1.

Questions:

a. Why did you pick the particular interaction and quadratic effect?

b. Are either of these two new terms significant?

c. Do they improve our model predictions?

d. Can you think of other predictors or terms, either in the data set or not contained in the data, that might improve the ability of our model to predict house prices?
**Exercise 4.2:** In this exercise we use the BankRevenue.jmp data.

Fit a full model to \( \log(Rev\_Total) \) using \( \log(Bal\_Total) \) and the other variables as model effects (using main effects only). Note, you may need to re-create these columns. Use the Minimum BIC stopping rule and stepwise regression to build your model.

a. Compare your reduced model to that obtained using Minimum AICc in this chapter. Describe the differences in terms of the variables in the model and key statistics (adjusted R Square, RMSE, and other statistics provided).

b. Which is the “better” model? Why? Does one model do a better job of predicting the response than the other? Explain

**Exercise 4.3:** Continue with the BankRevenue.JMP data.

Instead of fitting a model using the transformed variables, fit a model using the original (untransformed) variables. Use Rev_Total as the response, and Bal_Total and the other variables as model effects. Use stepwise and your preferred stopping rule to build the model.

a. What are the model assumptions?

b. Use the tools covered in this chapter to check model assumptions. Which tools should you use to check these assumptions? Explain how each tool helps check assumptions.

c. Explain why the model assumptions are or are not met.

d. Does it make sense to use this model to make predictions? Why or why not?

**Exercise 4.4:** Use the BostonHousing.jmp data set from the Sample Data Directory for this exercise. The response of interest, mvalue, is the median value of homes for towns in the Boston area in the 1970s.

a. Use the tools introduced in Chapter 3 to explore the data and prepare for modeling. Are there any potential data quality issues (other than the fact that the data are from the 1970s)? Determine what actions, if any, should be taken to address data quality issues that you identify.

b. Fit a model to mvalue using only chas and rooms. Recall that rooms is the number of rooms (rooms) and chas is a dummy variable (chas=1 indicates the town tracks the Charles River).

   i. Write down the equation for this model.

   ii. Interpret the coefficients for chas[0] and rooms.
iii. What is the predicted mvalue for a home that tracks the Charles River and has 6 rooms?

c. Fit a model to mvalue using all of the other variables as model effects. Use the Minimum BIC stopping rule and stepwise regression to build your model. How many terms are in the final model? Which terms are not included in the model?


e. How would a realtor, selling homes in the Boston area (in the same time period), use this model? How would a potential home buyer use this model?

References


Index

A

AAE (Average Absolute Error) 227, 250
activation functions 189–195
Actual by Predicted Plot 67
Adaptive option 276
Advanced Controls 276
advanced methods
  about 256
  bagging 256–257
  boosting 217, 257–258
  concepts in 256–259
  exercises 285–287
  Generalized Regression models 273–285
  partition 259–273
  regularization 258–259
Akaike's Information Criterion (AICc) 72
algorithms
  for decision trees 134
  for logistic regression 103–105
  for multiple linear regression 59–61
  for neural networks 188–195
Alternate Cut-off Confusion Matrix Add-In 304
Analysis of Variance table 68
analytics
  about 4–5
  data mining and 7
Analytics Job Task Analysis 12
Analyze menu 27
Annotate tool 76
area under the curve (AUC) 164
Average Absolute Error (AAE) 227, 250

B

bagging 256–257
BAP
  See Business Analytics Process (BAP)
Bayesian Information Criterion (BIC) 72
Berry, M. Michael 7
  Data Mining Techniques, 3rd ed. 12
Beta Continuous Response models 284
Beta-Binomial Categorical Response models 284
biased variable 85–86
BIC (Bayesian Information Criterion) 72
binning continuous data 47–49
Binomial Categorical Response models 284
Blue Book for Bulldozers (Kaggle Contest) case study 308–314, 323–324
Boosted Neural Network models 258, 268–273
Boosted Tree models 144, 258, 264–268, 300–301
boosting 217, 257–258
Bootstrap Forest models 144, 257, 260–264, 299–300
Box, George 224
business analytics, categories of 6
Business Analytics Process (BAP)
  about 11, 13
  commonly used models 12
  defining the problem 13–14
  deploying models 15–16
  modeling 15
  monitoring performance 16
  preparing for modeling 14–15

C

CAP (Certified Analytics Professional) 12
carat (\(^\)) 47
Carvana (Kaggle Contest) case study 317, 325–326
case studies
  Blue Book for Bulldozers (Kaggle Contest) 308–314, 323–324
  Carvana (Kaggle Contest) 317, 325–326
  Cell Classification 290–308
  Default Credit Card, Presenting Results to Management 314–316
  exercises 318–321
Categorical Profiler 206, 214
Categorical Response models 284–285
categorical responses, neural networks with 195
Cauchy Continuous Response models 284
Cell Classification case study 290–308
cell labels, turning on 107
Certified Analytics Professional (CAP) 12
churn 196–207
classification trees
  defined 134
  statistical details behind 137–143
  for status 135–137
Cluster Variables option 49–50
Column Switcher 84, 171, 237
Columns Panel 24
Columns Viewer tool 26
Competing on Analytics (Davenport and Harris) 6
Confusion Matrix 111, 143, 160–161
Contingency Analysis 236
continuous data, binning 47–49
Continuous modeling type 25–26
Continuous Response models 284
Cook's D (Cook's Distance) 75–76, 92–93
copying and pasting output 30
Cross Industry Standard Process for Data Mining (CRISP-DM) 12
cross-validation
  about 224–228
  examples 232–251
  exercises 251–253
  fitting neural network models using 246–249
  K-fold 230–231
  partitioning data for 228–232
  using for model fitting in JMP Pro 231–232
  using to build decision tree models 244–246
  using to build linear regression models 239–240
cut points 138

D
data
  common problems with 42–43
  examining 25–41
  exercises 51–53
  partitioning for cross-validation 228–232
  preparing for modeling 41–50
  restructuring 50
  working with 21–53
Data Filter 84
data mining, analytics and 7
Data Mining Techniques, 3rd ed. (Linoff and Berry) 12
data sets 42–43
data tables, JMP 23–25
Davenport, Thomas
  Competing on Analytics 6
  The New World of Business Analytics 4
decision trees
  about 132
  algorithm for 134
  classification trees 135–137
  Example 1: Credit Card Marketing 150–166
  Example 2: Printing Press Yield 166–181
  exercises 181–183
  exploratory modeling versus predictive modeling 144–145
  missing values 149
  model cross-validation 145–149
  modeling with ordinal predictors 149
  preview of end result 133–134
  representative business problems 132–133
  using cross-validation to build 244–246
deep learning 186
Default Credit Card, Presenting Results to Management case study 314–316
defining the problem, in Business Analytics Process (BAP) 13–14
degrees of freedom 48
deploying models, in Business Analytics Process (BAP) 15–16
deriving new variables 45–47
descriptive analytics 6
dimension reduction 49
dirty data 42
Discovering JMP 22, 25
Distribution platform 91, 95, 124–125, 309
Distribution tool 26–27
E

Early Stopping option 276
Effect Summary 67, 87
Elastic Net 281–285
"ensemble" model 264
Entropy RSquare 158
Estimation Method 276
examples
  Boosted Tree model 265
  Bootstrap Forest model 260–261
cross-validation 232–251
decision trees 150–181
logistic regression 105–125
multiple linear regression 62–97
neural networks 196–217
exercises
  advanced methods 285–287
case studies 318–321
cross-validation 251–253
data 51–53
decision trees 181–183
logistic regression 125–129
multiple linear regression 97–99
neural networks 218–220
exploratory analysis 6–7
exploratory modeling 144–145
exponential (EXP function) 95
Exponential Continuous Response models 284

F

Fit Details report 143
Fit Least Squares platform 85, 276
Fit Model platform 85, 88, 259, 282
folds 230–231
Formula Editor 47, 136
freedom, degrees of 48

G

Gamma Continuous Response models 284
Gaussian model 189–195
Generalized Regression models

about 273–276
Elastic Net 281–285
Lasso Regression 279–281
Maximum Likelihood Regression 276–277
Ridge Regression 277–279
Goldbloom, Anthony 186–187
Grabber tool 81
Graph Builder platform 31–41, 66, 81–84, 95, 124–125, 237, 242, 309
Gravure Association of the Americas (website) 166
gray triangles 24

H

half-width 78
Harris, Jeanne G.  Competing on Analytics 6
help 51
Hinton, Geoffrey E. 186
hold out set 145
Holdback Validation Method 200, 211
hotspots 24
Hover help 51

I

IIA (International Institute for Analytics) 6
important effects 88
incomplete data 42
incorrectly formatted data 43
Informative Missing option 45, 149, 155, 168, 199
Institute for Operations Research and the
  Management Sciences (INFORMS) 12
interaction term 75
International Institute for Analytics (IIA) 6

J

JMP

about 8
basics of 22–51
data tables 23–25
help 51
modeling types in 25–26
opening 22–23
setting preferences in 28, 84
JMP Documentation Library 51
.jmp extension 24
JMP File Exchange 146, 202
JMP Home 23
JMP Pro 8, 231–232
JMP Pro Partition platform 144
JMP Starter 22

K

k levels 48
Kaggle (website) 317
K-fold cross-validation 230–231

L

Lasso Regression 279–281
lasso tool 91
Leaf Report 141–142
LeCun, Yann 186
lift 203
Lift Curve 164–166, 213
likelihood ratio chi-square statistic 137
Linear model 189–195
linear regression models, using cross-validation to build 239–240
Linoff, Gordon 7
Data Mining Techniques, 3rd ed. 12
logistic regression
about 101–102, 124–125
algorithm for 103–105
Example 1: Lost Sales Opportunities 105–115
Example 2: Titanic Passengers 115–124
exercises 125–129
preview of end result 102–103
representative business problems 102
logit (log-odds) 103, 122
LogWorth 137–138

M

Make Validation Column Modeling Utility 311
margin of error 78
Marginal Model Plots 215
Master of Science in Analytics (MSA) degree 4
Max option 27
Max Validation RSquare 240
Maximum Likelihood Regression 276–277
maximum value of the Validation RSquare statistic 146
Mean Abs Dev (mean absolute deviation) 187
messy data 42
Method of Least Squares 60
method of maximum likelihood 104
Min option 27
Minimum AICc (Akaike's Information Criterion) 72, 88, 118–119
Minimum BIC 118–119
Minimum Size Split 145
Misclassification Rate 110, 125, 143, 160
missing data 42
missing values
dealing with 149
working with 43–45
model building 57–99
Model Comparison platform 249–251, 282, 302, 312
model cross-validation 145–149
model fitting, using cross-validation for in JMP Pro 231–232
modeling
in Business Analytics Process (BAP) 15
exploratory 144–145
with ordinal predictors 149
predictive 144–145
preparing data for 41–50
types in JMP 25–26
modeling, preparing for, in Business Analytics Process (BAP) 14–15
models
comparing 249–251
cross-validation 224–253
decision trees 131–183
deploying, in Business Analytics Process (BAP) 15–16
logistic regression 101–129
neural networks 186–221
monitoring performance, in Business Analytics Process (BAP) 16
MSA (Master of Science in Analytics) degree  4
multicollinearity  70
multiple linear regression
  about  57–58
  algorithm for  59–61
  Example 1: Housing Prices  62–79
  Example 2: Bank Revenues  80–97
  exercises  97–99
  preview of end result  58
  representative business problems  58
Multivariate platform  65–66

N
Negative Binomial Categorical Response models  285
Neter, J.M.  104
Neural Model Launch  217, 248, 268–269
neural networks
  about  186–187, 301–302
  algorithm for  188–195
  with categorical responses  195
  Example 1: Churn  196–207
  Example 2: Credit Risk  208–217
  exercises  218–220
  fitting using cross-validation  246–249
  measuring success  187
  preview of end result  188
  representative business problems  187
Neural platform  195, 202, 211, 258
The New World of Business Analytics (Davenport)  4
Nominal modeling type  25–26
Normal Continuous Response models  284

O
Ordinal modeling type  25–26
ordinal predictors, modeling with  149
Ordinal Restricts Order  155
output, copying and pasting  30
over-fitting  144, 227

P
Parameter Estimates table  68, 78
"parameter penalization"  195
parsimony, principle of  68
Partition dialog window  155
partition graph  139
partition methods, advanced  259–273
partitioning data for cross-validation  228–232
PCA (Principal Components Analysis)  49
performance, monitoring, in Business Analytics Process (BAP)  16
Poisson Categorical Response models  284
Prediction Profiler  78–79, 113–114, 123, 215
predictions, making  244
predictive analytics  6
predictive modeling  144–145
preferences, setting in JMP  28, 84
preparing for modeling, in Business Analytics Process (BAP)  14–15, 41–50
prescriptive analytics  6
Principal Components Analysis (PCA)  49
principle of parsimony  68
Prob > T rule  162
probability  103
problem, defining, in Business Analytics Process (BAP)  13–14
Profiler  251, 264, 313
Profit Matrix Column Property  304
Purely Random method  233
p-values  77, 137

Q
quadratic term  75
Quantile Regression Continuous Response models  284
question mark  51

R
Random Seed Reset add-in  146, 202, 211, 229, 300
Random Validation Portion  228–229
RASE (Root Average Squared Error) 227, 250
Receiver Operating Characteristic (ROC) curve 162–164, 213
red triangles 24
regression
See specific types
regression models, choosing terms with stepwise regression for 240–243
regression tree 134
regularization 258–259
regularized regression 195
Residual by Predicted plot 74–75
residuals 74–75
restructuring data 50
Ridge Regression 277–279
right-click 24
RMSE (root mean square error) 187
ROC (Receiver Operating Characteristic) curve 162–164, 213
Root Average Squared Error (RASE) 227, 250
root mean square error (RSME) 187
Rows Panel 24
RSquare (Entropy RSquare) 67, 158, 225
RSquare Adj 67
RSquare Validation 242

S
Sample, Explore, Modify, Model, and Assess (SEMMA) 12
saving
scripts 30
your work 30–31
Scatterplot Matrix platform 124–125, 294
scripts, saving 30
SEMMA (Sample, Explore, Modify, Model, and Assess) 12
Sigmoid TanH function 301
significant effects 88
Singularity Details option 85, 87
Specialized Models JMP manual 119, 195
Split History report 147, 245
Squared Penalty Method 200
SSE (sum of the squared prediction errors) 225–226
SST (sum of squares "total" for the response) 225–226
statistical analysis 6–7
statistical details, behind classification trees 137–143
status, classification trees for 135–137
stepwise logistic regression 298–299
Stepwise Personality option 88
stepwise regression 71–73, 125, 240–243
Stepwise Regression Control 119
Stratified Random method 233
sum of squared residuals 60
sum of squares "total" for the response (SST) 225–226
sum of the squared prediction errors (SSE) 225–226
Summary of Fit table 67

T
Table Panel 23
Tabulate platform 31–41
TanH model 189–195
test subsets 230, 233–235
training sets 145, 226
training subsets 229, 233–235
transforming variables 45–47

U
UC Irvine Machine Learning Repository (website) 167
under-fit 144

V
validation data set 226
Validation Method 276
validation roles, specifying for each row 229–230
validation set 145
validation subsets 229, 233–239
variable clustering 295–296
Variable Importance option 215
Variable Importance Summary Report 215
variables
   biased 85–86
   deriving new 45–47
   exploring many at a time 65–66
   exploring one at a time 63–65
   transforming 45–47
Variance Inflation Factor (VIF statistic) 70

W

Whole Model Test 110–111

Z

`zeroed variable 85–86
Zero-Inflated Distributions Categorical Response
   models 285
About These Authors

Jim Grayson, PhD, is a Professor of Management Science and Operations Management in the Hull College of Business Administration at Georgia Regents University. He currently teaches undergraduate and MBA courses in operations management and business analytics. Previously, Jim held managerial positions at Texas Instruments in quality and reliability assurance, supplier and subcontractor management, and software quality. He has a PhD in management science with an information systems minor from the University of North Texas, an MBA in marketing from the University of North Texas, and a BS from the United States Military Academy at West Point.

Sam Gardner is a Senior Research Scientist at Elanco, a business division of Eli Lilly and Company. He currently works as a statistician supporting manufacturing development for a variety of animal health products. He is recognized by the American Statistical Association as an Accredited Professional Statistician. He has an MS in mathematics from Creighton University and an MS in statistics from the University of Kentucky. He graduated from Purdue University with BS degrees in mathematics and chemistry.

Gardner started his professional career as a military officer in the US Air Force, where he had roles that focused on modeling and simulation, operational flight test planning and analysis, and research and development. He also taught statistics at the Air Force Institute of Technology. After leaving the military, he began his work in the pharmaceutical industry as a statistician supporting the development and manufacturing of active pharmaceutical ingredients, and later transitioned to a role as a chemist in pharmaceutical manufacturing. He also worked as a statistician in pharmaceutical marketing with a focus on using statistical modeling to help solve business problems related to sales and marketing effectiveness. An avid user of statistical software, he spent several years working for JMP®, a division of SAS®, where he worked as a product expert and seminar speaker, with a focus on data visualization, applied statistics, and modern statistical modeling.
Mia L. Stephens is an Academic Ambassador with JMP®, a division of SAS®. Prior to joining SAS®, she split her time between industrial consulting and teaching statistics at the University of New Hampshire. A coauthor of *Visual Six Sigma: Making Data Analysis Lean* and *JMP® Start Statistics: A Guide to Statistics and Data Analysis Using JMP®, Fifth Edition*, she has developed courses and training materials, taught, and consulted within a variety of manufacturing and service industries. Stephens holds an MS in statistics from the University of New Hampshire and is located in York Harbor, Maine.

Learn more about these authors by visiting their author pages, where you can download free book excerpts, access example code and data, read the latest reviews, get updates, and more:

http://support.sas.com/grayson
http://support.sas.com/gardner
http://support.sas.com/stephens
Gain Greater Insight into Your JMP® Software with SAS Books.

Discover all that you need on your journey to knowledge and empowerment.

support.sas.com/bookstore
for additional books and resources.