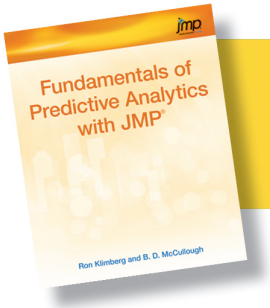
Fundamentals of Predictive Analytics with JMP®

Ron Klimberg and B. D. McCullough



From *Fundamentals of Predictive Analytics with JMP®*. Full book available for purchase here.

Contents

About This Book ix
About These Authors xviii
Acknowledgments xv

Chapter 1 Introduction 1

Two Questions Organizations Need to Ask 2
- **Return on Investment (ROI) 2**
- **Culture Change 3**

Business Intelligence 3
Clarification 5
Book Focus 5
Introductory Statistics Courses 5
Practical Statistical Study 8
Plan Perform, Analyze, Reflect (PPAR) Cycle 9
References 12

Chapter 2 Statistics Review 13

Always Take a Random and Representative Sample 14
Statistics Is Not an Exact Science 15
Understand a Z Score 17
Understand the Central Limit Theorem 18
Understand One-Sample Hypothesis Testing and p-Values 24
Many Approaches/Techniques Are Correct, and a Few Are Wrong 26

Chapter 3 Introduction to Multivariate Data 37

Multivariate Data and Multivariate Data Analysis 37
Using Tables to Explore Multivariate Data 40
Using Graphs to Explore Multivariate Data 45

Chapter 4 Regression and ANOVA Review 63

Regression 64
- Simple Regression 64
- Multiple Regression 67
- Regression with Categorical Data 76

ANOVA 82
- One-way ANOVA 83
- Testing Statistical Assumptions 85
- Testing for Differences 90
- Two-way ANOVA 97

References 102

Chapter 5 Logistic Regression 103

Dependence Technique: Logistic Regression 104
The Linear Probability Model (LPM) 105
The Logistic Function 106
Example: toylogistic.jmp 108
Odds Ratios in Logistic Regression 113
A Logistic Regression Statistical Study 122
References 133
Exercises 133

Chapter 6 Principal Components Analysis 135

Principal Component 140
Dimension Reduction 142
Discovering Structure in The Data 145
Exercises 149

Chapter 7 Cluster Analysis 151

Hierarchical Clustering 154
Using Clusters in Regression 164
K-means Clustering 164
K-means versus Hierarchical Clustering 177
References 177
Exercises 178

Chapter 8 Decision Trees 179

An Example of Classification Trees 182
An Example of a Regression Tree 192
References 199
Exercises 199

Chapter 9 Neural Networks 201

Validation Methods 206
Hidden Layer Structure 208
Fitting Options 211
Data Preparation 212
An Example 213
Summary 223
References 223
Exercises 224

Chapter 10 Model Comparison 225

Model Comparison with Continuous Dependent Variable 226
Model Comparison with Binary Dependent Variable 230
Model Comparison Using the Lift Chart 237
Train, Validate, and Test 240
References 246
Exercises 247

Chapter 11 Telling the Statistical Story 249

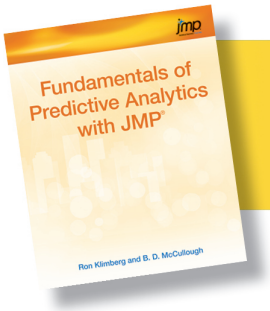
From Multivariate Data to the Modeling Process 249
What Is Data Mining? 251
A Framework for Predictive Analytics Techniques 252
The Goal, Tasks, and Phases of Predictive Analytics 253
References 257

Appendix Data Sets 259

Smaller Data Sets 260
Large Case Data Sets 263

Index 269

From *Fundamentals of Predictive Analytics with JMP®* by Ron Klimberg and B. D. McCullough.
Copyright © 2012, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.



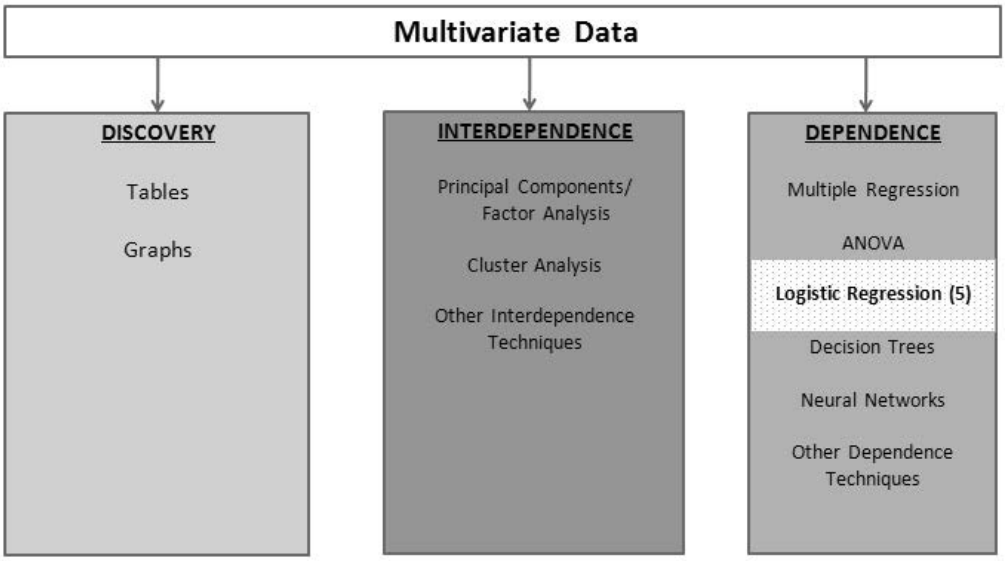
From *Fundamentals of Predictive Analytics with JMP®*. Full book available for purchase here.

Chapter 5

Logistic Regression

Dependence Technique: Logistic Regression 104
The Linear Probability Model (LPM) 105
The Logistic Function 106
Example: toylogistic.jmp 108
Odds Ratios in Logistic Regression 113
A Logistic Regression Statistical Study 122
References 133
Exercises 133

Figure 5.1 A Framework for Multivariate Analysis



Dependence Technique: Logistic Regression

Logistic regression, as shown in our multivariate analysis framework in Figure 5.1, is one of the dependence techniques in which the dependent variable is discrete and, more specifically, binary. That is, it takes on only two possible values. Here are some examples: Will a credit card applicant pay off a bill or not? Will a mortgage applicant default? Will someone who receives a direct mail solicitation respond to the solicitation? In each of these cases, the answer is either "yes" or "no." Such a categorical variable cannot directly be used as a dependent variable in a regression. But a simple transformation solves the problem: Let the dependent variable Y take on the value 1 for "yes" and 0 for "no."

Because Y takes on only the values 0 and 1, we know $E[Y_i] = 1*P[Y_i=1] + 0*P[Y_i=0] = P[Y_i=1]$. But from the theory of regression, we also know that $E[Y_i] = a + b*X_i$. (Here we use simple regression, but the same holds true for multiple regression.) Combining these two results, we have $P[Y_i=1] = a + b*X_i$. We can see that, in the case of a binary dependent variable, the regression may be interpreted as a probability. We then seek to use this regression to estimate the probability that Y takes on the value 1. If the estimated probability is high enough, say above 0.5, then we predict 1; conversely, if the estimated probability of a 1 is low enough, say below 0.5, then we predict 0.

The Linear Probability Model (LPM)

When linear regression is applied to a binary dependent variable, it is commonly called the Linear Probability Model (LPM). Traditional linear regression is designed for a continuous dependent variable, and is not well-suited to handling a binary dependent variable. Three primary difficulties arise in the LPM. First, the predictions from a linear regression do not necessarily fall between zero and one. What are we to make of a predicted probability greater than one? How do we interpret a negative probability? A model that is capable of producing such nonsensical results does not inspire confidence.

Second, for any given predicted value of y (denoted $\hat{y}$), the residual (resid= y - $\hat{y}$) can take only two values. For example, if $\hat{y}$ = 0.37, then the only possible values for the residual are resid= -0.37 or resid = 0.63 (= 1 – 0.37), because it has to be the case that $\hat{y}$ + resid equals zero or one. Clearly, the residuals will not be normal. Plotting a graph of $\hat{y}$ versus resid will produce not a nice scatter of points, but two parallel lines. The reader should verify this assertion by running such a regression and making the requisite scatterplot. A further implication of the fact that the residual can take on only two values for any $\hat{y}$ is that the residuals are heteroscedastic. This violates the linear regression assumption of homoscedasticity (constant variance). The estimates of the standard errors of the regression coefficients will not be stable and inference will be unreliable.

Third, the linearity assumption is likely to be invalid, especially at the extremes of the independent variable. Suppose we are modeling the probability that a consumer will pay back a $10,000 loan as a function of his/her income. The dependent variable is binary, 1 = the consumer pays back the loan, 0 = the consumer does not pay back the loan. The independent variable is income, measured in dollars. A consumer whose income is $50,000 might have a probability of 0.5 of paying back the loan. If the consumer's income is increased by $5,000, then the probability of paying back the loan might increase to 0.55, so that every $1,000 increase in income increases the probability of paying back the loan by 1%. A person with an income of $150,000 (who can pay the loan back very easily) might have a probability of 0.99 of paying back the loan. What happens to this probability when the consumer's income is increased by $5,000? Probability cannot increase by 5%, because then it would exceed 100%; yet according to the linearity assumption of linear regression, it must do so.

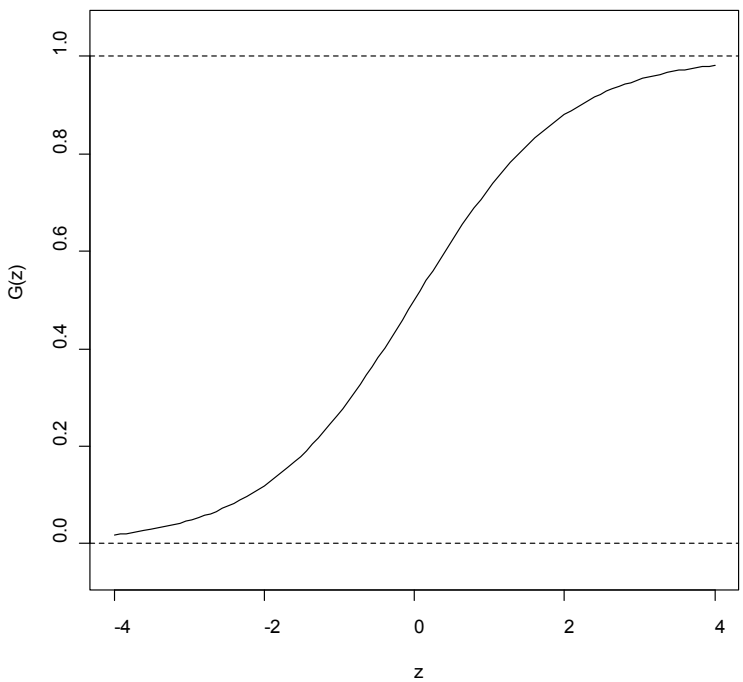
The Logistic Function

A better way to model $P[Y_i=1]$ would be to use a function that is not linear, one that increases slowly when $P[Y_i=1]$ is close to zero or one, and that increases more rapidly in between. It would have an "S" shape. One such function is the logistic function

$$G(z)=\frac{1}{1+e^{-z}}=\frac{e^{z}}{1+e^{z}}$$

whose cumulative distribution function is shown in Figure 5.2.

Figure 5.2 The Logistic Function



Another useful representation of the logistic function is

$$1-G(z)=\frac{e^{-z}}{1+e^{-z}}$$

Recognize that the y-axis, G(z), is a probability and let G(z) = π, the probability of the event occurring. We can form the odds ratio (the probability of the event occurring divided by the probability of the event not occurring) and do some simplifying:

$$\frac{\pi}{1-\pi} = \frac{G(z)}{1-G(z)} = \frac{\frac{1}{1+e^{-z}}}{\frac{e^{-z}}{1+e^{-z}}} = \frac{1}{e^{-z}} = e^{z}$$

Consider taking the natural logarithm of both sides. The left side will become log[$\pi/(1-\pi)$] and the log of the odds ratio is called the logit. The right side will become *z* (since log(e^{z}) = z) so that we have the relation

$$\log\left[\frac{\pi}{1-\pi}\right] = z$$

and this is called the logit transformation.

If we model the logit as a linear function of X (*i.e.*, let z = $\beta_0 + \beta_1 X$), then we have

$$\log\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 X$$

We could estimate this model by linear regression and obtain estimates b_0 of β_0 and b_1 of β_1 if only we knew the log of the odds ratio for each observation. Since we do not know the log of the odds ratio for each observation, we will use a form of nonlinear regression called logistic regression to estimate the model below:

$$E[Y_i] = \pi_i = G\left(\beta_0 + \beta_1 X_i\right) = \frac{1}{1+e^{-\beta_0-\beta_1 X_i}}$$

In so doing, we obtain the desired estimates b_0 of β_0 and b_1 of β_1. The estimated probability for an observation X_i will be

$$P\left[Y_i = 1\right] = \hat{\pi}_i = \frac{1}{1+e^{-b_0-b_1 X_i}}$$

and the corresponding estimated logit will be

$$\log\left[\frac{\hat{\pi}}{1-\hat{\pi}}\right] = b_0 + b_1 X$$

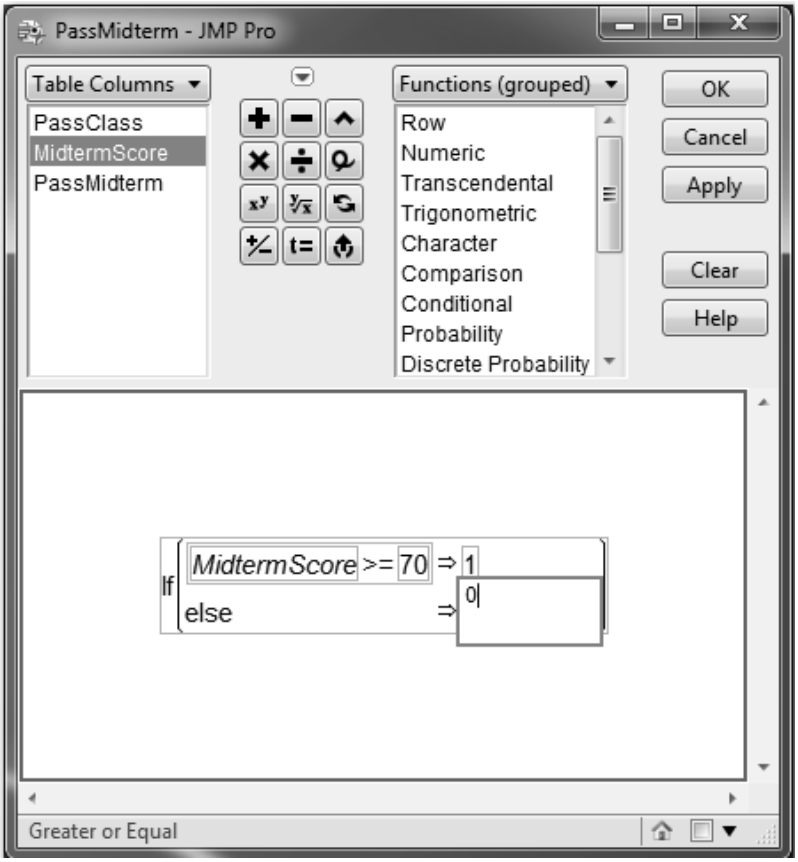
which leads to a natural interpretation of the estimated coefficient in a logistic regression: b_1 is the estimated change in the logit (log odds) for a one-unit change in X.

Example: toylogistic.jmp

To make these ideas concrete, suppose we open a small data set toylogistic.jmp, containing students' midterm exam scores (MidtermScore) and whether the student passed the class (PassClass=1 if pass, PassClass=0 if fail). A passing grade for the midterm is 70. The first thing to do is create a dummy variable to indicate whether the student passed the midterm: PassMidterm = 1 if MidtermScore ≥ 70 and PassMidterm = 0 otherwise:

Select **Cols→New Column** to open the New Column dialog box. In the Column Name text box, for our new dummy variable, type PassMidterm. Click the drop-down box for modeling type and change it to **Nominal**. Click the drop-down box for Column Properties and select **Formula**. The Formula dialog box appears. Under Functions, click **Conditional→If**. Under Table Columns, click **MidtermScore** so that it appears in the top box to the right of the If. Under Functions, click **Comparison Analyze→Distributions "a>=b"**. In the formula box to the right of >=, enter 70. Press the **Tab** key. Click in the box to the right of the ⇒, and enter the number 1. Similarly, enter 0 for the else clause. The Formula dialog box should look like Figure 5.3.

Figure 5.3 Formula Dialog Box

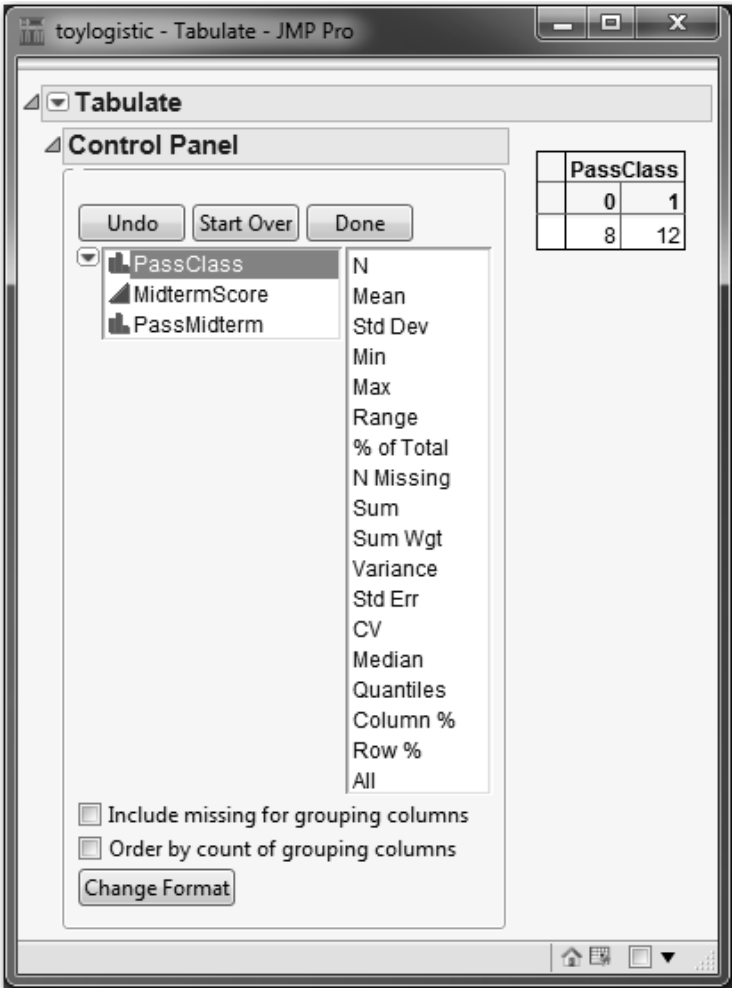


Select **OK→OK**.

First, let us use a traditional contingency table analysis to determine the odds ratio. Make sure that both PassClass and PassMidterm are classified as nominal variables. Right-click in the data grid of the column PassClass and select **Column Info**. Click the black triangle next to Modeling Type and select **Nominal→OK**. Do the same for PassMidterm.

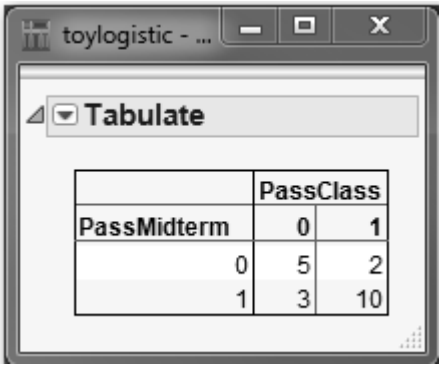
Select **Tables→Tabulate** to open the Control Panel. It shows the general layout for a table. Drag PassClass into the Drop zone for columns and select **Add Grouping Columns**. Now that data have been added, the words Drop zone for rows will no longer be visible, but the Drop zone for rows will still be in the lower left panel of the table. See Figure 5.4.

Figure 5.4 Control Panel for Tabulate



Drag PassMidterm to the panel immediately to the left of the 8 in the table. Select **Add Grouping Columns**. Click **Done**. A contingency table identical to Figure 5.5 will appear.

Figure 5.5 Contingency Table from toydataset.jmp



	PassClass	
PassMidterm	0	1
0	5	2
1	3	10

The probability of passing the class when you did not pass the midterm is

$$P(PassClass=1)|P(PassMidterm=0) = 2/7$$

The probability of not passing the class when you did not pass the midterm is

$$P(PassClass=0)|P(PassMidterm=0) = 5/7$$

(similar to row percentages). The odds of passing the class given that you have failed the midterm are

$$\frac{P(PassClass=1)\,|\,P(PassMidterm=0)}{P(PassClass=0)\,|\,P((PassMidterm=0)} = \frac{2/7}{5/7} = \frac{2}{5}$$

Similarly, we calculate the odds of passing the class given that you have passed the midterm as:

$$\frac{P(PassClass=1)\,|\,P(PassMidterm=1)}{P(PassClass=0)\,|\,P(PassMidterm=1)} = \frac{10/13}{3/13} = \frac{10}{3}$$

Of the students that did pass the midterm, the odds are the number of students that pass the class divided by the number of students that did not pass the class.

In the above paragraphs, we spoke only of odds. Now let us calculate an odds ratio. It is important to note that this can be done in two equivalent ways. Suppose we want to know the odds ratio of passing the class by comparing those who pass the midterm

(PassMidterm=1 in the numerator) to those who fail the midterm (PassMidterm=0 in the denominator). The usual calculation leads to:

$$\frac{\text{Odds of passing the class; given passed the Midterm}}{\text{Odds of passing the class; given failed the Midterm}} = \frac{10/3}{2/5} = \frac{50}{6} = 8.33.$$

which has the following interpretation: the odds of passing the class are 8.33 times the odds of failing the course if you pass the midterm. This odds ratio can be converted into a probability. We know that P(Y=1)/P(Y=0)=8.33; and by definition, P(Y=1)+P(Y=0)=1. So solving two equations in two unknowns yields P(Y=0) = (1/(1+8.33)) = (1/9.33)= 0.1072 and P(Y=1) = 0.8928. As a quick check, observe that 0.8928/0.1072=8.33. Note that the log-odds are ln(8.33) = 2.120. Of course, the user doesn't have to perform all these calculations by hand; JMP will do them automatically. When a logistic regression has been run, simply clicking the red triangle and selecting Odds Ratios will do the trick.

Equivalently, we could compare those who fail the midterm (PassMidterm=0 in the numerator) to those who pass the midterm (PassMidterm=1 in the denominator) and calculate:

$$\frac{\text{Odds of passing the class; given failed the Midterm}}{\text{Odds of passing the class; given passed the Midterm}} = \frac{2/5}{10/3} = \frac{6}{50} = \frac{1}{8.33} = 0.12.$$

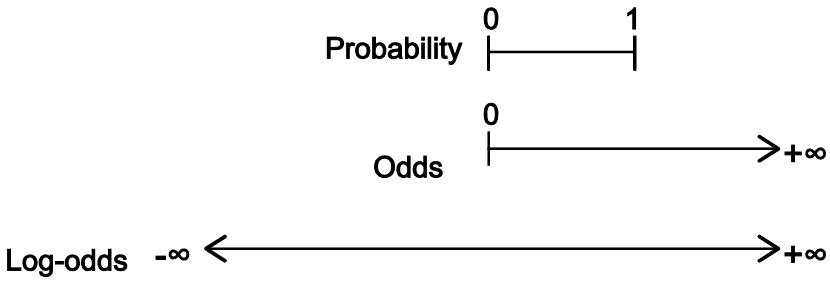
which tells us that the odds of failing the class are 0.12 times the odds of passing the class for a student who passes the midterm. Since P(Y = 0) = 1 - π (the probability of failing the midterm) is in the numerator of this odds ratio, we must interpret it in terms of the event failing the midterm. It is easier to interpret the odds ratio when it is less than 1 by using the following transformation: (OR – 1)*100%. Compared to a person who passes the midterm, a person who fails the midterm is 12% as likely to pass the class; or equivalently, a person who fails the midterm is 88% less likely, (OR – 1)*100% = (0.12 – 1)*100%= -88%, to pass the class than someone who passed the midterm. Note that the log-odds are ln(0.12) = -2.12.

The relationships between probabilities, odds (ratios), and log-odds (ratios) are straightforward. An event with a small probability has small odds, and also has small log-odds. An event with a large probability has large odds and also large log-odds. Probabilities are always between zero and unity; odds are bounded below by zero but can be arbitrarily large; log-odds can be positive or negative and are not bounded, as shown in Figure 5.6. In particular, if the odds ratio is 1 (so the probability of either event is 0.50), then the log-odds equal zero. Suppose π = 0.55, so the odds ratio 0.55/0.45 = 1.222. Then we say that the event in the numerator is (1.222-1) = 22.2% more likely to occur than the event in the denominator.

Odds Ratios in Logistic Regression

Different software applications adopt different conventions for handling the expression of odds ratios in logistic regression. By default, JMP = uses the "log odds of 0/1" convention, which puts the 0 in the numerator and the 1 in the denominator. This is a consequence of the sort order of the columns, which we will address shortly.

Figure 5.6 Ranges of Probabilities, Odds, and Log-odds



To see the practical importance of this, rather than compute a table and perform the above calculations, we can simply run a logistic regression. It is important to make sure that PassClass is nominal and that PassMidterm is continuous. If PassMidterm is nominal, JMP will fit a different but mathematically equivalent model that will give different (but mathematically equivalent) results. The scope of the reason for this is beyond this book, but, in JMP, interested readers can consult **Help→Books→Modeling and Multivariate Methods** and refer to Appendix A.

If you have been following along with the book, both variables ought to be classified as nominal, so PassMidterm needs to be changed to continuous. Right-click in the column PassMidterm in the data grid and select **Column Info**. Click the black triangle next to Modeling Type and select **Continuous**, and then click **OK**.

Now that the dependent and independent variables are correctly classified as Nominal and Continuous, respectively, let's run the logistic regression:

From the top menu, select **Analyze→Fit Model**. Select **PassClass→Y**. Select **PassMidterm→Add**. The Fit Model dialog box should now look like Figure 5.7. Click **Run**.

Figure 5.7 Fit Model Dialog Box

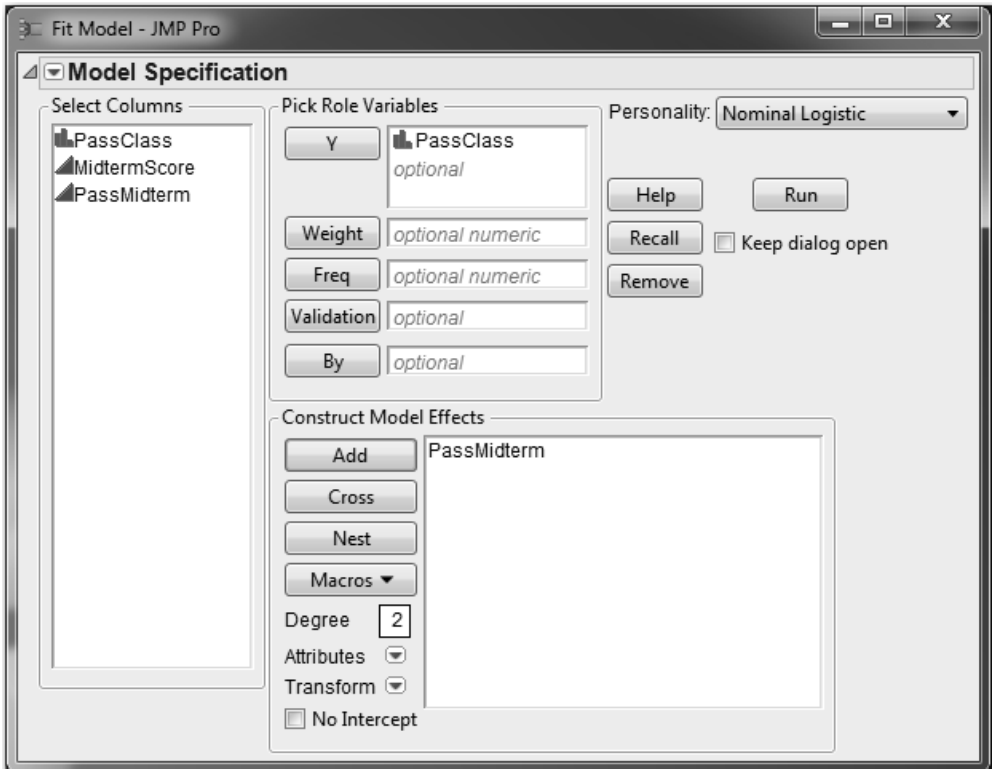


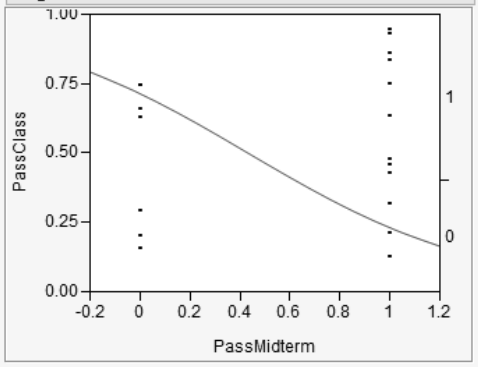
Figure 5.8 displays the logistic regression results.

Figure 5.8 Logistic Regression Results for toylogistic.jmp

toylogistic - Fit Nominal Logistic - JMP Pro

Nominal Logistic Fit for PassClass

Logistic Plot



Converged in Gradient, 4 iterations

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	2.249692	1	4.499385	0.0339*
Full	11.210541			
Reduced	13.460233			

RSquare (U)	0.1671
AICc	27.127
BIC	28.4125
Observations (or Sum Wgts)	20

Measure	Training	Definition
Entropy RSquare	0.1671	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.2723	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.5605	∑ -Log(ρ[j])/n
RMSE	0.4322	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.3736	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.2500	∑ (ρ[j]≠ρMax)/n
N	20	n

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	0.91629073	0.83666	1.20	0.2734
PassMidterm	-2.1202635	1.0645813	3.97	0.0464*

For log odds of 0/1

Covariance of Estimates

Effect Likelihood Ratio Tests

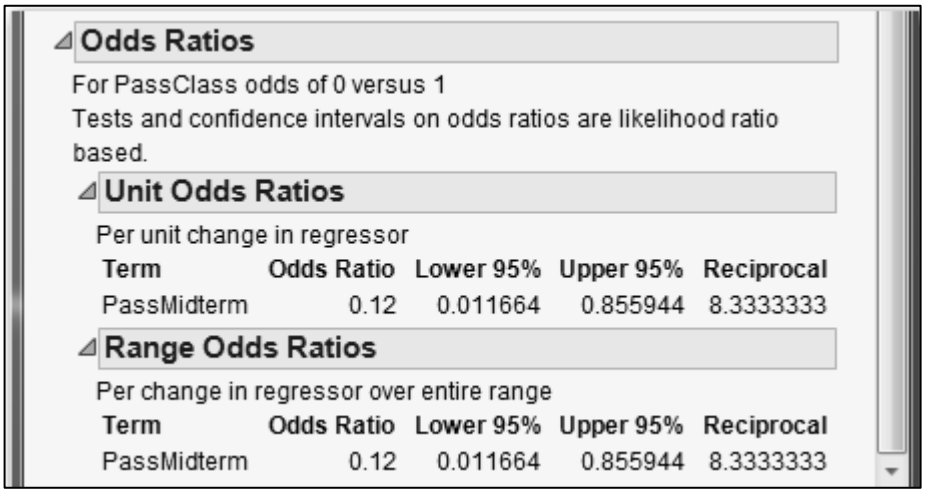
Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
PassMidterm	1	1	4.49938474	0.0339*

Examine the parameter estimates in Figure 5.8. The intercept is 0.91629073, and the slope is -2.1202635. The slope gives the expected change in the logit for a one-unit change in the independent variable (*i.e.*, the expected change on the log of the odds ratio). However, if we simply exponentiate the slope (*i.e.*, compute) $e^{-2.1202635} = 0.12$, then we get the 0/1 odds ratio.

There is no need for us to exponentiate the coefficient manually. JMP will do this for us:

Click the red triangle and click **Odds Ratios**. The Odds Ratios tables are added to the JMP output as shown in Figure 5.9.

Figure 5.9 Odds Ratios Tables Using the Nominal Independent Variable PassMidterm



Odds Ratios

For PassClass odds of 0 versus 1

Tests and confidence intervals on odds ratios are likelihood ratio based.

Unit Odds Ratios

Per unit change in regressor

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
PassMidterm	0.12	0.011664	0.855944	8.3333333

Range Odds Ratios

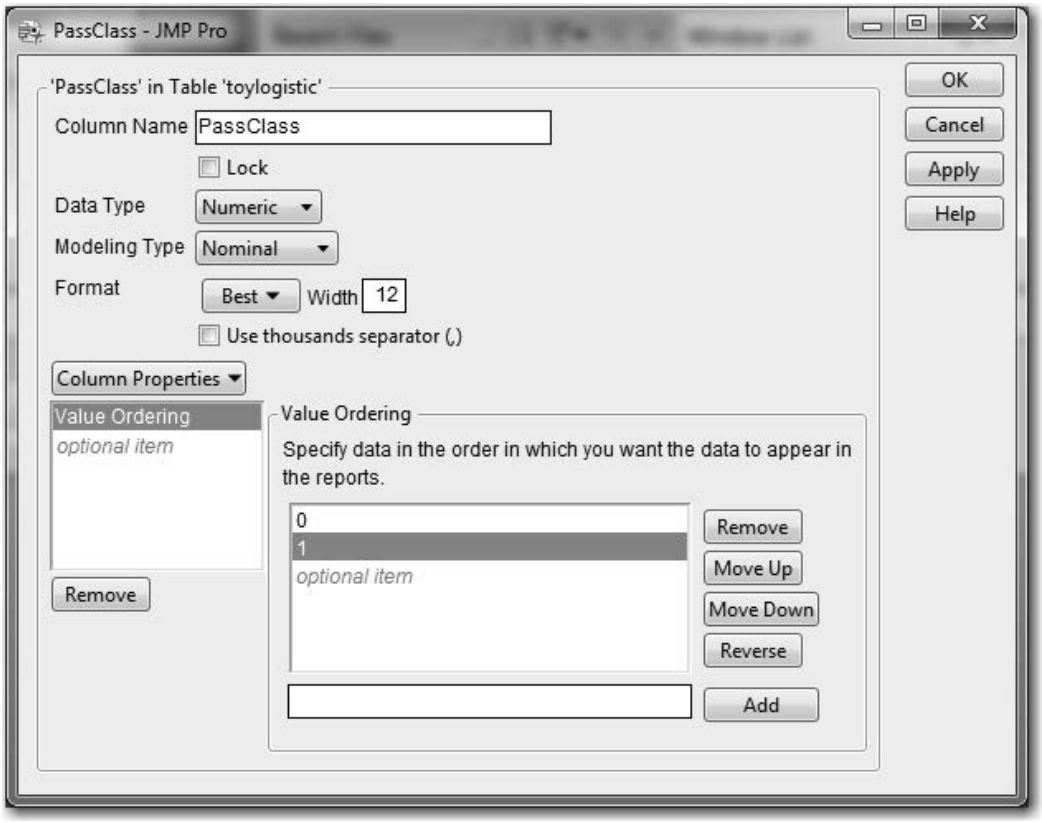
Per change in regressor over entire range

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
PassMidterm	0.12	0.011664	0.855944	8.3333333

Unit Odds Ratios refers to the expected change in the odds ratio for a one-unit change in the independent variable. Range Odds Ratios refers to the expected change in the odds ratio when the independent variable changes from its minimum to its maximum. Since the present independent variable is a binary 0-1 variable, these two definitions are the same. We get not only the odds ratio, but a confidence interval, too. Notice the right-skewed confidence interval; this is typical of confidence intervals for odds ratios.

To change from the default convention (log odds of 0/1, which puts the 0 in the numerator and the 1 in the denominator, in the data table), right-click to select the name of the PassClass column. Under Column Properties, select **Value Ordering**. Click on the value **1** and click **Move Up** as in Figure 5.10.

Figure 5.10 Changing the Value Order



Then, when you re-run the logistic regression, although the parameter estimates will not change, the odds ratios will change to reflect the fact that the 1 is now in the numerator and the 0 is in the denominator.

The independent variable is not limited to being only a nominal (or ordinal) dependent variable; it can be continuous. In particular, let's examine the results using the actual score on the midterm, with MidtermScore as an independent variable:

Select **Analyze→Fit Model**. Select **PassClass→Y** and then select **MidtermScore→Add.** Click **Run**.

This time the intercept is 25.6018754, and the slope is -0.3637609. So we expect the log-odds to decrease by 0.3637609 for every additional point scored on the midterm, as shown in Figure 5.11.

Figure 5.11 Parameter Estimates

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	25.6018754	11.184069	5.24	0.0221*
MidtermScore	-0.3637609	0.1581661	5.29	0.0215*

For log odds of 0/1

To view the effect on the odds ratio itself, as before click the red triangle and click **Odds Ratios**. Figure 5.12 displays the Odds Ratios tables.

Figure 5.12 Odds Ratios Tables Using the Continuous Independent Variable MidtermScore

Odds Ratios

For PassClass odds of 0 versus 1

Tests and confidence intervals on odds ratios are likelihood ratio based.

Unit Odds Ratios

Per unit change in regressor

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
MidtermScore	0.695057	0.451212	0.879069	1.4387302

Range Odds Ratios

Per change in regressor over entire range

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
MidtermScore	0.000335	2.491e-8	0.058681	2989.1384

For a one-unit increase in the midterm score, the new odds ratio will be 69.51% of the old odds ratio. Or, equivalently, we expect to see a 30.5% reduction in the odds ratio (0.695057 – 1)*(100%=-30.5%). For example, suppose a hypothetical student has a midterm score of 75%. The student's log odds of failing the class would be 25.6018754 – 0.3637609*75 = -1.680192. So the student's odds of failing the class would be exp(-1.680192) = 0.1863382. That is, the student is much more likely to pass than fail. Converting odds to probabilities (0.1863328/(1+0.1863328) = 0.157066212786159), we see that the student's probability of failing the class is 0.15707, and the probability of passing the class is 0.84293. Now, if the student's score increased by one point to 76, then the log odds of failing the class would be 25.6018754 – 0.3637609*76 = -2.043953. Thus, the student's odds of failing the class become exp(-2.043953)= 0.1295157. So, the probability of passing the class would rise to 0.885334, and the probability of failing the class would fall to 0.114666. With respect to the Unit Odds Ratio, which equals 0.695057, we see that a one-unit increase in the test score changes the odds ratio from 0.1863382 to 0.1295157. In accordance with the estimated coefficient for the logistic regression, the new odds ratio is 69.5% of the old odds ratio because 0.1295157/0.1863382 = 0.695057.

Finally, we can use the logistic regression to compute probabilities for each observation. As noted, the logistic regression will produce an estimated logit for each observation. These estimated logits can be used, in the obvious way, to compute probabilities for each observation. Consider a student whose midterm score is 70. The student's estimated logit is 25.6018754 – 0.3637609(70) = 0.1386124. Since exp(0.1386129) = 1.148679 = $\pi/(1-\pi)$, we can solve for π (the probability of failing) = 0.534597.

We can obtain the estimated logits and probabilities by clicking the red triangle on Normal Logistic Fit and selecting **Save Probability Formula**. Four columns will be added to the worksheet: Lin[0], Prob[0], Prob[1], and Most Likely PassClass. For each observation, these give the estimated logit, the probability of failing the class, and the probability of passing the class, respectively. Observe that the sixth student has a midterm score of 70. Look up this student's estimated probability of failing (Prob[0]); it is very close to what we just calculated above. See Figure 5.13. The difference is that the computer carries 16 digits through its calculations, but we carried only six.

Figure 5.13 Verifying Calculation of Probability of Failing

6/0	Pass Class	Midterm Score	Lin[0]	Prob[0]	Prob[1]	Most Likely PassClass
1	0	62	3.048697664	0.9547262676	0.0452737324	0
2	0	63	2.6849367335	0.936131922	0.063868078	0
3	0	64	2.3211758029	0.9106156911	0.0893843089	0
4	0	65	1.9574148724	0.8762529099	0.1237470901	0
5	0	66	1.5936539419	0.8311295662	0.1688704338	0
6	0	70	0.1386102197	0.5345971803	0.4654028197	0
7	0	72	-0.588911641	0.3568846133	0.6431153867	1
8	0	74	-1.316433502	0.2114122777	0.7885877223	1
9	1	68	0.8661320808	0.7039402276	0.2960597724	0
10	1	69	0.5023711503	0.6230163983	0.3769836017	0
11	1	71	-0.225150711	0.4439489049	0.5560510951	1
12	1	73	-0.952672572	0.2783476639	0.7216523361	1

The fourth column (Most Likely PassClass) classifies the observation as either 1 or 0, depending upon whether the probability is greater than or less than 50%. We can observe how well our model classifies all the observations (using this cut-off point of 50%) by producing a confusion matrix: Click the red triangle and click Confusion matrix. Figure 5.14 displays the confusion matrix for our example. The rows of the confusion matrix are the actual classification (that is, whether PassClass is 0 or 1). The columns are the predicted classification from the model (that is, the predicted 0/1 values from that last fourth column using our logistic model and a cutpoint of .50). Correct classifications are along the main diagonal from upper left to lower right. We see that the model has classified 6 students as not passing the class, and actually they did not pass the class. The model also classifies 10 students as passing the class when they actually did. The values on the other diagonal, both equal to 2, are misclassifications. The results of the confusion matrix will be examined in more detail when we discuss model comparison in Chapter 9.

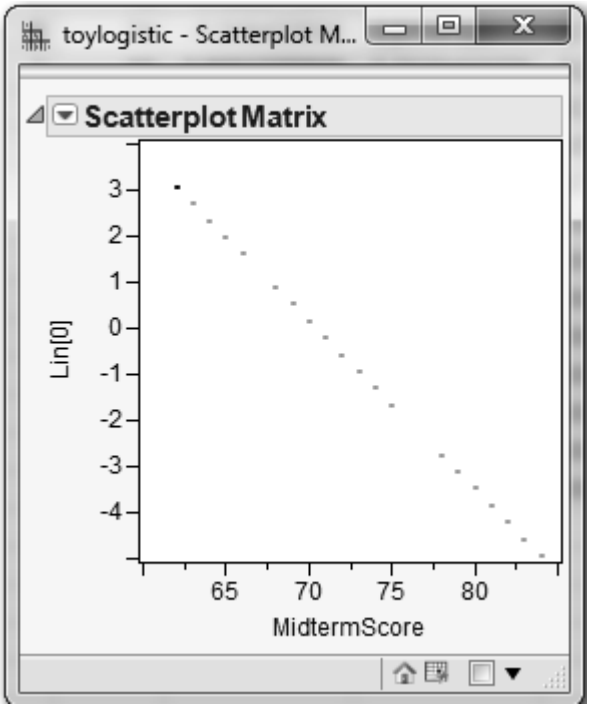
Figure 5.14 Confusion Matrix

Confusion Matrix

Actual		Predicted
Training	**0**	**1**
0	6	2
1	2	10

Of course, before we can use the model, we have to check the model's assumptions, etc. The first step is to verify the linearity of the logit. This can be done by plotting the estimated logit against PassClass. Select **Graph→Scatterplot Matrix.** Select **Lin[0]→Y, columns**. Select **MidtermScore→X**. Click **OK**. As shown in Figure 5.15, the linearity assumption appears to be perfectly satisfied.

Figure 5.15 Scatterplot of Lin[0] and MidtermScore



The analog to the ANOVA F-test for linear regression is found under the Whole Model Test, shown in Figure 5.16, in which the Full and Reduced models are compared. The null hypothesis for this test is that all the slope parameters are equal to zero. Since Prob>ChiSq is 0.0004, this null hypothesis is soundly rejected. For a discussion of other statistics found here, such as BIC and Entropy RSquare, see the **JMP Help.**

Figure 5.16 Whole Model Test for the Toylogistic Data Set

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	6.264486	1	12.52897	0.0004*
Full	7.195748			
Reduced	13.460233			

RSquare (U)	0.4654
AICc	19.0974
BIC	20.383
Observations (or Sum Wgts)	20

The next important part of model checking is the Lack of Fit test. See Figure 5.17. It compares the model actually fitted to the saturated model. The saturated model is a model generated by JMP that contains as many parameters as there are observations. So it fits the data very well. The null hypothesis for this test is that there is no difference between the estimated model and the saturated model. If this hypothesis is rejected, then more variables (such as cross-product or squared terms) need to be added to the model. In the present case, as can be seen, Prob>ChiSq=0.7032. We can therefore conclude that we do not need to add more terms to the model.

Figure 5.17 Lack of Fit Test for Current Model

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	18	7.1957477	14.3915
Saturated	19	0.0000000	Prob>ChiSq
Fitted	1	7.1957477	0.7032

A Logistic Regression Statistical Study

Let's turn now to a more realistic data set with several independent variables. During this discussion, we will also present briefly some of the issues that should be addressed and some of the thought processes during a statistical study.

Cellphone companies are very interested in determining which customers might switch to another company; this is called "churning." Predicting which customers might be about

to churn enables the company to make special offers to these customers, possibly stemming their defection. Churn.jmp contains data on 3333 cellphone customers, including the variable Churn (0 means the customer stayed with the company and 1 means the customer left the company).

Before we can begin constructing a model for customer churn, we need to discuss model building for logistic regression. Statistics and econometrics texts devote entire chapters to this concept. In several pages, we can only sketch the broad outline. The first thing to do is make sure that the data are loaded correctly. Observe that Churn is classified as Continuous; be sure to change it to Nominal. One way is to right-click in the Churn column in the data table, select **Column Info**, and under Modeling Type, click **Nominal**. Another way is to look at the list of variables on the left side of the data table, find Churn, click the blue triangle (which denotes a continuous variable), and change it to **Nominal** (the blue triangle then becomes a red histogram). Make sure that all binary variables are classified as Nominal. This includes Intl_Plan, VMail_Plan, E_VMAIL_PLAN, and D_VMAIL_PLAN. Should Area_Code be classified as Continuous or Nominal? (Nominal is the correct answer!) CustServ_Call, the number of calls to customer service, could be treated as either continuous or nominal/ordinal; we treat it as continuous.

When building a linear regression model and the number of variables is not so large that this cannot be done manually, one place to begin is by examining histograms and scatterplots of the continuous variables, and crosstabs of the categorical variables as discussed in Chapter 3. Another very useful device as discussed in Chapter 3 is the scatterplot/correlation matrix, which can, at a glance, suggest potentially *useful* independent variables that are correlated with the dependent variable. The scatterplot/correlation matrix approach cannot be used with logistic regression, which is nonlinear, but a method similar in spirit can be applied.

We are now faced with a similar situation that was discussed in Chapter 4. Our goal is to build a model that follows the principle of parsimony—that is, a model that explains as much as possible of the variation in Y and uses as few significant independent variables as possible. However, now with multiple logistic regression, we are in a nonlinear situation. We have four approaches that we could take. We briefly list and discuss each of these approaches and some of their advantages and disadvantages:

- **Include all the variables.** In this approach you just input all the independent variables into the model. An obvious advantage of this approach is that it is fast and easy. However, depending on the data set, most likely several independent variables will be insignificantly related to the dependent variable. Including variables that are not significant can cause severe problems—weakening the interpretation of the coefficients and lessening the prediction accuracy of the model. This approach definitely does not follow the principle of parsimony, and it can cause numerical problems for the nonlinear solver that may lead to a failure to obtain an answer.

- **Bivariate method.** In this approach, you search for independent variables that may have predictive value for the dependent variable by running a series of bivariate logistic regressions; *i.e.*, we run a logistic regression for each of the independent variables, searching for "significant" relationships. A major advantage of this approach is that it is the one most agreed upon by statisticians (Hosmer and Lemeshow, 2001). On the other hand, this approach is not automated, is very tedious, and is limited by the analyst's ability to run the regressions. That is, it is not practical with very large data sets. Further, it misses interaction terms, which, as we shall see, can be very important.
- **Stepwise.** In this approach, you would use the Fit Model platform, change the Personality to Stepwise and Direction to Mixed. The Mixed option is like Forward Stepwise, but variables can be dropped after they have been added. An advantage of this approach is that it is automated; so, it is fast and easy. The disadvantage of stepwise is that it could lead to possible interpretation and prediction errors depending on the data set. However, using the Mixed option, as opposed to the Forward or Backward Direction option, tends to lessen the magnitude and likelihood of these problems.
- **Decision Trees.** A Decision Tree is a data mining technique that can be used for variable selection and will be discussed in Chapter 8. The advantage of using the decision tree technique is that it is automated, fast, and easy to run. Further, it is a popular variable reduction approach taken by many data mining analysts (Pollack, 2008). However, somewhat like the stepwise approach, the decision tree approach could lead to some statistical issues. In this case, significant variables identified by a decision tree are very sample-dependent. These issues will be discussed further in Chapter 8.

No one approach is a clear cut winner. Nevertheless, we do not recommend using the "Include all the variables" approach. If the data set is too large and/or you do not have the time, we recommend that you run both the stepwise and decision trees models and compare the results. The data set churn.jmp is not too large, so we will apply the bivariate approach.

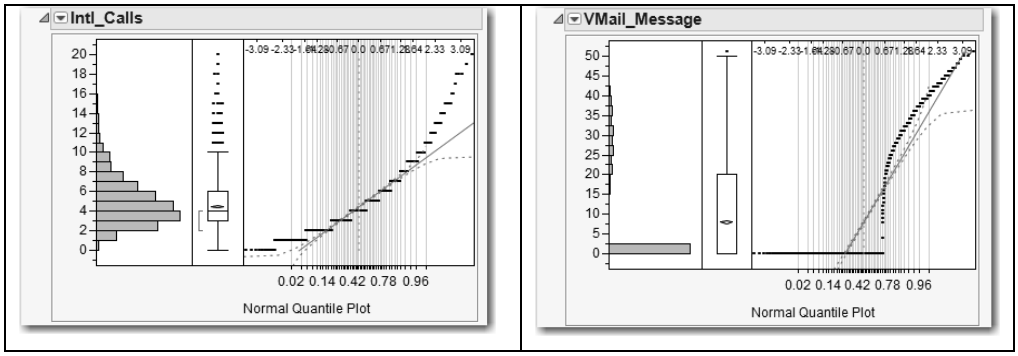
It is traditional to choose $\alpha = 0.05$. But in this preliminary stage, we adopt a more lax standard, $\alpha = 0.25$. The reason for this is that we want to include, if possible, a group of variables that individually are not significant but together are significant. Having identified an appropriate set of candidate variables, run a logistic regression including all of them. Compare the coefficient estimates from the multiple logistic regression with the estimates from the bivariate logistic regressions. Look for coefficients that have changed in sign or have dramatically changed in magnitude, as well as changes in significance. Such changes indicate the inadequacy of the simple bivariate models, and confirm the necessity of adding more variables to the model.

Three important ways to improve a model are as follows:

- If the logit appears to be nonlinear when plotted against some continuous variable, one resolution is to convert the continuous variable to a few dummies, say three, that cut the variable at its 25th, 50th, and 75th percentiles.
- If a histogram shows that a continuous variable has an excess of observations at zero (which can lead to nonlinearity in the logit), add a dummy variable that equals one if the continuous variable is zero and equals zero otherwise.
- Finally, a seemingly numeric variable that is actually discrete can be broken up into a handful of dummy variables (*e.g.*, ZIP codes).

Before we can begin modeling, we must first explore the data. With our churn data set, creating and examining the histograms of the continuous variables reveals nothing much of interest, except VMail_Message, which has an excess of zeros. (See the second point in the previous paragraph.) Figure 5.18 shows plots for Intl_Calls and VMail_Message. To produce such plots, select **Analyze→Distribution**, click **Intl_Calls**, and then **Y, Columns** and **OK**. To add the Normal Quantile Plot, click the red arrow next to Intl_Calls and select **Normal Quantile Plot**. Here it is obvious that Intl_Calls is skewed right. We note that a logarithmic transformation of this variable might be in order, but we will not pursue the idea.

Figure 5.18 Distribution of Intl_Calls and VMail_Message



A correlation matrix of the continuous variables (select **Graph→Scatterplot Matrix** and put the desired variables in **Y, Columns**) turns up a curious pattern. Day_Charge and Day_Mins, Eve_Charge and Eve_Mins, Night_Charge and Night_Mins, and Intl_Charge and Intl_Mins all are perfectly correlated. The charge is obviously a linear function of the number of minutes. Therefore, we can drop the Charge variables from our analysis. (We could also drop the "Mins" variables instead; it doesn't matter which one we drop.) If our data set had a very large number of variables, the scatterplot matrix would be too big to comprehend. In such a situation, we would choose groups of variables for which to make scatterplot matrices, and examine those.

A scatterplot matrix for the four binary variables turns up an interesting association. E_VMAIL_PLAN and D_VMAIL_PLAN are perfectly correlated; both have common 1s and where the former has -1, the latter has zero. It would be a mistake to include both of these variables in the same regression (try it and see what happens). Let's delete E_VMAIL_PLAN from the data set and also delete VMail_Plan because it agrees perfectly with E_VMAIL_PLAN: When the former has a "no," the latter has a "-1," and similarly for "yes" and "+1."

Phone is more or less unique to each observation. (We ignore the possibility that two phone numbers are the same but have different area codes.) Therefore, it should not be included in the analysis. So, we will drop Phone from the analysis.

A scatterplot matrix between the remaining continuous and binary variables turns up a curious pattern. D_VMAIL_PLAN and VMailMessage have a correlation of 0.96. They have zeros in common, and where the former has 1s, the latter has numbers. (See again point two in the above paragraph. We won't have to create a dummy variable to solve the problem because D_VMAIL_PLAN will do the job nicely.)

To summarize, we have dropped 7 of the original 23 variables from the data set (Phone, Day_Charge, Eve_Charge, Night_Charge, Intl_Charge, E_VMAIL_PLAN, and VMail_Plan). So there are now 16 variables left, one of which is the dependent variable, Churn. We have 15 possible independent variables to consider.

Next comes the time-consuming task of running several bivariate (two variables, one dependent and one independent) analyses, some of which will be logistic regressions (when the independent variable is continuous) and some of which will be contingency tables (when the independent variable is categorical). In total, we have 15 bivariate analyses to run. What about Area Code? JMP reads it as a continuous variable, but it's really nominal, so make sure to change it from continuous to nominal. Similarly, make sure that D_VMAIL_PLAN is set as a nominal variable, not continuous.

Do *not* try to keep track of the results in your head, or by referring to the 15 bivariate analyses that would fill your computer screen. Make a list of all 15 variables that need to be tested, and write down the test result (*e.g.*, the relevant p-value) and your conclusion (*e.g.*, "include" or "exclude"). This not only prevents simple errors; it is a useful record of your work should you have to come back to it later. There are few things more pointless than conducting an analysis that concludes with a 13-variable logistic regression, only to have some reason to rerun the analysis and now wind up with a 12-variable logistic regression. Unless you have documented your work, you will have no idea why the discrepancy exists or which is the correct regression.

Below we briefly show how to conduct both types of bivariate analyses, one for a nominal independent variable and one for a continuous independent variable. We leave the other 14 to the reader.

Make a contingency table of Churn versus State: Select **Analyze→Fit Y by X**, click Churn (which is nominal) and then click **Y, Response,** click **State** and then click **X, Factor**; and click **OK**. At the bottom of the table of results are the Likelihood Ratio and Pearson tests, both of which test the null hypothesis that State does not affect Churn, and both of which reject the null. The conclusion is that the variable State matters. On the other hand, perform a logistic regression of Churn on VMail_Message: select **Analyze→Fit Y by X**, click **Churn**, click **Y, Response**, and click **VMail_Message** and click **X**, Factor; and click **OK**. Under "Whole Model Test" that Prob>ChiSq, the p-value is less than 0.0001, so we conclude that VMail_message affects Churn. Remember that for all these tests, we are setting α (probability of Type I error) = 0.25.

In the end, we have 10 candidate variables for possible inclusion in our multiple logistic regression model:

State	Intl_Plan	D_VMAIL_PLAN
VMail_Message	Day_Mins	Eve_Mins
Night_Mins	Intl_Mins	Intl_Calls
CustServ_Call		

Remember that the first three of these variables (the first row) should be set to nominal, and the rest to continuous. (Of course, leave the dependent variable Churn as nominal!)

Let's run our initial multiple logistic regression with Churn as the dependent variable and the above 10 variables as independent variables:

Select **Analyze→Fit Model→Churn→Y**. Select the above 10 variables (to select variables that are not consecutive, click on each variable while holding down the **Ctrl** key), and click **Add**. Check the box next to Keep dialog open. Click **Run**.

The Whole Model Test lets us know that our included variables have an effect on the Churn and a p-value less than .0001, as shown in Figure 5.19.

Figure 5.19 Whole Model Test and Lack of Fit for the Churn Data Set

Nominal Logistic Fit for Churn

Converged in Gradient, 6 iterations

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	341.6995	59	683.3991	<.0001*
Full	1037.4471			
Reduced	1379.1467			

RSquare (U)	0.2478
AICc	2197.13
BIC	2561.59
Observations (or Sum Wgts)	3333

Measure	Training	Definition
Entropy RSquare	0.2478	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.3293	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.3113	∑ -Log(ρ[j])/n
RMSE	0.3070	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.1883	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1308	∑ (ρ[j]≠ρMax)/n
N	3333	n

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	3273	1037.4471	2074.894
Saturated	3332	0.0000	Prob>ChiSq
Fitted	59	1037.4471	1.0000

The lack-of-fit test tells us that we have done a good job explaining Churn. From the Lack of Fit, we see that –LogLikelihood for the Full model is 1037.4471. Now, linear regression minimizes the sum of squared residuals. So when you compare two linear regressions, the preferred one has the smaller sum of squared residuals. In the same way, the nonlinear optimization of the logistic regression minimizes the –LogLikelihood (which is equivalent to maximizing the LogLikelihood). So the model with the smaller –LogLikelihood is preferred to a model with a larger –LogLikelihood.

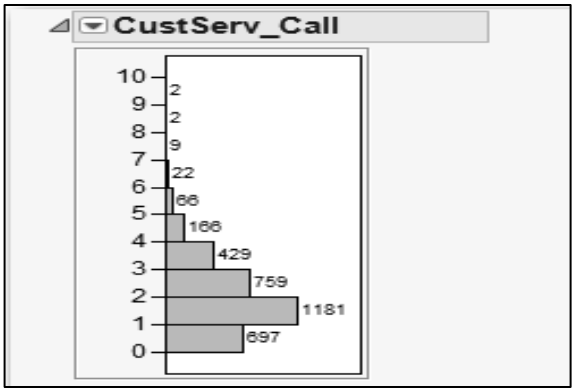
Examining the p-values of the independent variables in the Parameter Estimates, we find that a variable for which Prob>ChiSq is less than 0.05 is said to be significant. Otherwise, it is said to be insignificant, similar to what is practiced in linear regression. The regression output gives two sets of tests, one for the "Parameter Estimates" and another

for "Effect Likelihood Ratio Tests." We shall focus on the latter. To see why, consider the State variable, which is really not one variable but many dummy variables. We are not so much interested in whether any particular state is significant or not (which is what the Parameter Estimates tell us) but whether, overall, the collection of state dummy variables is significant. This is what the Effect Likelihood Ratio Tests tells us; the effect of all the state dummies is significant with a "Prob>ChiSq" of 0.0010. True, many of the State dummies are insignificant, but overall State is significant; we will keep this variable as it is. It may prove worthwhile to reduce the number of state dummies into a handful of significant states and small clusters of "other" states that are not significant, but we will not pursue this line of inquiry here.

We can see that all the variables in the model are significant. We may be able to derive some new variables that help improve the model. We will provide two examples of deriving new variables—(1) Converting a continuous variable into discrete variables; (2) Producing interaction variables.

Let us try to break up a continuous variable into a handful of discrete variables. An obvious candidate is CustServ_Call. Look at its distribution in Figure 5.20. Select **Analyze→Distribution**, select **CustServ_Call→Y, Columns**, and click **OK**. Click the red arrow next to CustServ_Call and uncheck **Outlier Box Plot**. Then choose **Histogram Options→Show Counts.**

Figure 5.20 Histogram of CustServ_Call

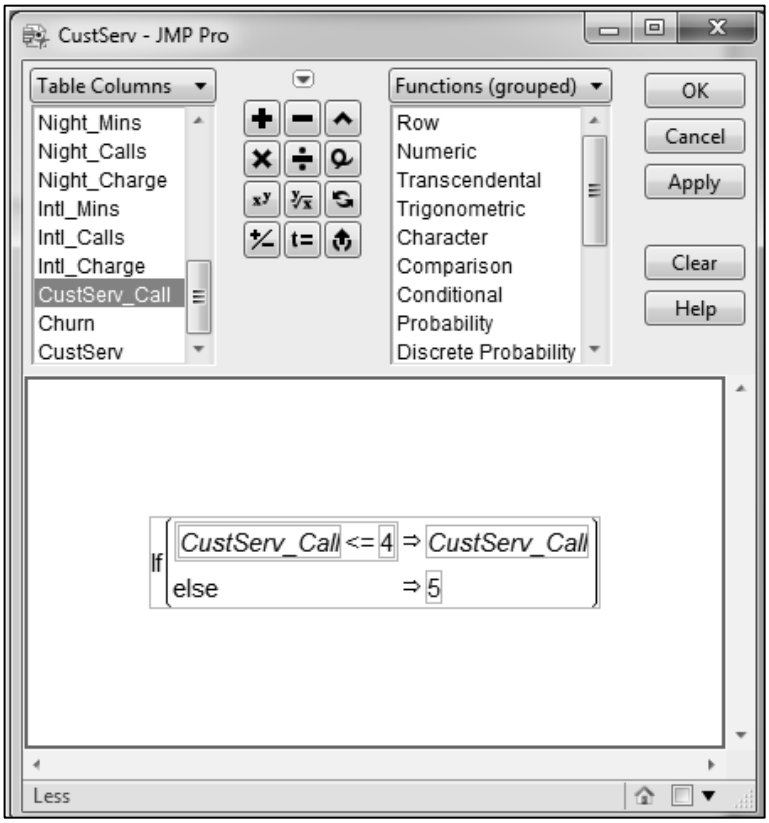


Let's create a new nominal variable called CustServ, so that all the counts for 5 and greater are collapsed into a single cell:

> Select **Cols→New Columns**. For column name type CustServ, for Modeling Type change it to **Nominal** and then click the drop-down arrow for Column Properties and click **Formula**. In the Formula dialog box, select **Conditional→If**. Then, in the top expr, click **CustServ_Call** and type <=4. In

the top then clause, click **CustServ_Call**. For the else clause, type **5**. See Figure 5.21. Click **OK** and click **OK**.

Figure 5.21 Creating the CustServ Variable



Now drop the CustServ_Call variable from the Logistic Regression and add the new CustServ nominal variable, which is equivalent to adding some dummy variables. Our new value of -LogLikelihood is 970.6171, which constitutes a very substantial improvement in the model.

Another possible important way to improve a model is to introduce interactions terms, that is, the product of two or more variables. Best practice would be to consult with subject-matter experts and seek their advice. Some thought is necessary to determine meaningful interactions, but it can pay off in substantially improved models. Thinking about what might make a cell phone customer want to switch to another carrier, we have all heard a friend complain about being charged an outrageous amount for making an international call. Based on this observation, we could conjecture that customers who

make international calls and who are not on the international calling plan might be more irritated and more likely to churn. A quick bivariate analysis shows that there are more than a few such persons in the data set. Select **Tables→Tabulate**, and drag Intl_Plan to Drop zone for columns. Drag Intl_Calls to Drop zone for rows. Click **Add Grouping Columns**. Observe that almost all customers make international calls, but most of them are not on the international plan (which gives cheaper rates for international calls). For example, for the customers who made no international call, all 18 of them were not on the international calling plan. For the customers who made 8 international calls, 106 were not on the international calling plan, and only 10 of them were. There is quite the potential for irritated customers here! This is confirmed by examining the output from the previous logistic regression. The parameter estimate for "Intl_Plan[no]" is positive and significant. This means that when a customer does not have an international plan, the probability is that the churn increases.

Customers who make international calls and don't get the cheap rates are perhaps more likely to churn than customers who make international calls and get cheap rates. Hence, the interaction term Int_Plan*Intl_Mins might be important. To create this interaction term, we have to create a new dummy variable for Intl_Plan, because the present variable is not numeric and cannot be multiplied by Intl_Mins:

> First, click on the Intl_Plan column in the data table to select it. Then select **Cols→Recode**. Under **New Value**, where it has **No**, type **0** and right below that where it has **Yes**, type **1**. From the **In Place** drop-down menu, select **New Column** and click OK. The new variable Intl_Plan2 is created. However, it is still nominal. Right-click on this column and under **Column Info**, change the Data Type to Numeric and the Modeling Type to Continuous. Click **OK**. (This variable has to be continuous so that we can use it in the interaction term, which is created by multiplication; nominal variables cannot be multiplied.)

To create the interaction term:

> Select **Cols→New Column** and call the new variable IntlPlanMins. Under Column Properties, click **Formula**. Click **Intl_Plan2**, click on the times sign (**x**) in the middle of the dialog box, click **Intl_Mins** and click **OK**. Click **OK** again.

Now add the variable IntlPlanMins as the 11th independent variable in multiple logistic regression that includes CustServ and run it. The variable IntlPlanMins is significant, and the –LogLikelihood has dropped to 947.1450, as shown in Figure 5.22. This is a substantial drop for adding one variable. Doubtless other useful interaction terms could be added to this model, but we will not further pursue this line of inquiry.

Figure 5.22 Logistic Regression Results with Interaction Term Added

Nominal Logistic Fit for Churn

Converged in Gradient, 6 iterations

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	432.0016	65	864.0033	<.0001*
Full	947.1450			
Reduced	1379.1467			

Now that we have built an acceptable model, it is time to validate the model. We have already checked the Lack of Fit, but now we have to check linearity of the logit. From the red arrow, click **Save Probability Formula**, which adds four variables to the data set: Lin[0] (which is the logit), Prob[0], Prob[1], and the predicted value of Churn, Most Likely Churn. Now we have to plot the logit against each of the continuous independent variables. The categorical independent variables do not offer much opportunity to reveal nonlinearity (plot some and see this for yourself). All the relationships of the continuous variables can be quickly viewed by generating a scatterplot matrix and then clicking the red triangle and **Fit Line**. Nearly all the red fitted lines are horizontal or near horizontal. For all of the logit vs. independent variable plots, there is no evidence of nonlinearity.

We can also see how well our model is predicting by examining the confusion matrix, which is shown in Figure 5.23.

Figure 5.23 Confusion Matrix

Confusion Matrix

Actual		Predicted
Training	**0**	**1**
0	2749	101
1	326	157

The actual number of churners in the data set is 326+157 = 483. The model predicted a total of 258 (=101+157) churners. The number of bad predictions made by the model is 326+101 = 427, which indicates that 326 that were predicted not to churn actually did churn, and 101 that were predicted to churn did not churn. Further, observe in the Prob[1] column of the data table that we have the probability that any customer will churn. Right-click on this column and select **Sort**. This will sort all the variables in the data set according to the probability of churning. Scroll to the top of the data set. Look at the

Churn column. It has mostly ones and some zeros here at the top, where the probabilities are all above 0.85. Scroll all the way to the bottom and see that the probabilities now are all below 0.01, and the values of Churn are all zero. We really have modeled the probability of churning.

Now that we have built a model for predicting churn, how might we use it? We could take the next month's data (when we do not yet know who has churned) and predict who is likely to churn. Then these customers can be offered special deals to keep them with the company, so that they do not churn.

References

Hosmer, D. W., and S. Lemeshow. (2001). *Applied Logistic Regression*. 2nd ed. New York: John Wiley & Sons.

Pollack, R. (2008). "Data Mining: Common Definitions, Applications, and Misunderstandings." *Data Mining Methods and Applications (Discrete Mathematics & Its Applications)*. Lawrence, K. D., S. Kudyba, and R. K. Klimberg (Eds.). Boca Raton, FL: Auerbach Publications.

Exercises

1. Consider the logistic regression for the toy data set, where π is the probability of passing the class:

$$\log\left[\frac{\hat{\pi}}{1-\hat{\pi}}\right] = 25.60188 - 0.363761\,MidtermScore$$

   Consider two students, one who scores 67% on the midterm and one who scores 73% on the midterm. What are the odds that each fails the class? What is the probability that each fails the class?

2. Consider the first logistic regression for the Churn data set, the one with 10 independent variables. Consider two customers, one with an international plan and one without. What are the odds that each churns? What is the probability that each churns?

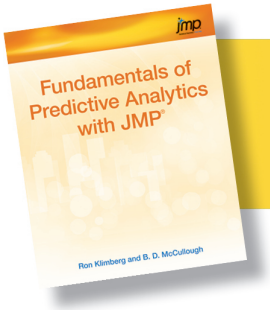
From *Fundamentals of Predictive Analytics with JMP®*. Full book available for purchase here.

3. We have already found that the interaction term IntlPlanMins significantly improves the model. Find another interaction term that does so.

4. Without deriving new variables such as CustServ or creating interaction terms such as IntlPlanMins, use a stepwise method to select variables for the Churn data set. Compare your results to the bivariate method used in the chapter; pay particular attention to the fit of the model and the confusion matrix.

5. Use the Freshmen1.jmp data set and build a logistic regression model to predict whether a student returns. Perhaps the continuous variables Miles from Home and Part Time Work Hours do not seem to have an effect. See whether turning them into discrete variables makes a difference. (*E.g.*, turn Miles from Home into some dummy variables, 0–20 miles, 21–100 miles, more than 100 miles.)

From *Fundamentals of Predictive Analytics with JMP®* by Ron Klimberg and B. D. McCullough.
Copyright © 2012, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.



From *Fundamentals of Predictive Analytics with JMP®*. Full book available for purchase here.

Index

A

absolute error measures 226, 227, 244
absolute penalty fit method 211
accuracy characteristic of prediction model 231
activation function, neural network 203–204
Adjusted R^2 75, 76
affinity grouping 252
agglomerative algorithm, clustering 154–163
AI (artificial intelligence) 250
AIC/AICc (Akaike information criterion) approach 76, 184, 195–196
Analyze command
  Fit Line 65
  Fit Model 67–69, 95
  Fit Stepwise 74
  Fit Y by X 31–33, 65, 83–84
ANOVA (analysis of variance)
  one-way/one-factor 83–96
  process 82–83
  two-way/two-factor 97–102
area under the curve (AUC) 235, 236, 237
artificial intelligence (AI) 250
association rules 252
association task, predictive analytics 254
AUC (area under the curve) 235, 236, 237
average linkage method, distance between clusters 154–155
Axis Titles, Histogram Data Analysis 23

B

BA (business analytics) 3–5, 250
bar chart 59–61
Bayesian information criterion (BIC) 76
bell-shaped distribution 18
BI (business intelligence) 3–5
bias term, neural network 203
BII (business information intelligence) 3–5
binary dependent variable 104, 221–222, 230–237
binary vs. multiway splits, decision tree 181
bivariate analysis 6, 31–36, 124–133
BMI (business modeling intelligence) 3–5
boosting option, neural network predictability 210, 216–217
BSI (business statistical intelligence) 3–5
bubble plot 53–55
business analytics (BA) 3–5, 250
business information intelligence (BII) 3–5
business intelligence (BI) 3–5
business modeling intelligence (BMI) 3–5
business statistics intelligence (BSI) 3–5

C

categorical variables
  See also ANOVA
  deciding on statistical technique 26, 28–29
  decision tree 180, 181–192
  graphs 45–46, 50–51, 52
  neural network 208, 212–213
  regression 76–82
  tables 42
causality 39
central limit theorem (CLT) 18–24
centroid method, distance between clusters 154–155, 164–166
chaining in single linkage method, distance between clusters 154
chi-square test of independence 185
churn analysis 122–133
classification task, predictive analytics 254
classification tree 181–192, 237, 239–240
cleaning data for practical study 8
CLT (central limit theorem) 18–24
cluster analysis
  credit card user example 152–153
  definition 152
  hierarchical clustering 154–163, 177
  k-means clustering 154, 164–177
  regression, using clusters in 164
Cluster command 156–157, 159
Clustering History, Cluster command 159
clustering task, predictive analytics 254
coefficient of determination (RSquare or R^2) 66
Color Clusters, Cluster command 157
Column Contributions, decision tree 198
complete linkage method, distance between clusters 154–155
confusion matrix
  binary dependent variable model comparison 230–231
  bivariate analysis contingency table 35

confusion matrix (*continued*)
logistic regression 120, 132
neural network 222–223
Connect Thru Missing, Overlay Plot command 166–167
constant variance assumption 105
contingency table 35
See also confusion matrix
continuous variables
See also ANOVA
deciding on statistical technique 26, 28
decision tree 180, 183, 192–199
logistic regression 129–130, 132
model comparison 226–230, 244
neural network 208
regression 76–77
contour graphs 55–56
cornerstone of statistics, CLT theorem 20
correlation coefficient 227
correlation matrix
logistic regression 125
multiple regression 67, 68
PCA 136–138, 142, 148–149
criterion function, decision tree 184–185
CRM (customer relation management) 250
cross-validation, neural network 207–208
customer relation management (CRM) 250

D

data, role of 2–3
data discovery 9, 11
See also graphs
See also tables
Data Filter, Graph Builder 56–61
data mining 4, 251, 254–255
See also predictive analytics
data warehouse 2
decision trees
classification tree 182–192, 237, 239–240
credit risk example 180–182
definition 180
pros and cons of using 124
regression tree 192–199
dendrogram 154, 157, 158, 159–160, 164
dependence, multivariate analysis framework 11
See also specific techniques
differences, testing for
one-way ANOVA 90–96
dimension reduction, PCA 136, 142–144
directed (supervised) predictive analytics techniques 252, 253, 254
"dirty data," problem of 6
discovery, multivariate analysis framework 9, 11
See also graphs
See also tables
discovery task, predictive analytics 254
Distribution command 30, 85–87, 125, 175
drop zones, Graph Builder 46–48
dummy variables 76–77, 79–82, 212
Dunnett's test 95
Durbin-Watson test 73
dynamic histogram, Excel 23
dynamic linking feature, JMP 58

E

Effect Likelihood Ratio Tests 129
eigenvalue analysis 141–144, 145, 147
eigenvalues-greater-than-1 method, PCA 144
elbow discovery method 144, 167
enterprise resource planning (ERP) 2
equal replication design, ANOVA with 97–102
error table 230
See also confusion matrix
estimation task, predictive analytics 254
Excel, Microsoft
measuring continuous variables 228–229
opening files in JMP 28
PivotTable 40–42
random sample generation 20–24
reasons for using 10–11
Exclude/Unexclude option, data table 147

F

factor analysis vs. PCA 140–141
See also Principal Component Analysis
factor loadings 145
false positive rate (FPR), prediction model 231–232
features, neural network 204–205
filtering data 56–61, 236–237
Fit Line, Analyze command 65
Fit Model, Analyze command 67–69, 95
Fit Stepwise, Analyze command 74
Fit Y by X, Analyze command 31–33, 65, 83–84
fitting to the model
ANOVA 83–84, 95
clusters 164

G^2 (goodness-of-fit) statistic, decision tree 184, 185–190
neural networks 206, 211, 215, 220
regression 65, 67–69, 71, 74, 122, 128
statistics review 31–33
train-validate-test paradigm for 240–246
Formula command 77–79
FPR (false positive rate), prediction model 231–232
fraud detection 250
frequency distribution, Excel Data Analysis Tool 22–23
F-test 65, 71–72, 83

G

G^2 (goodness-of-fit) statistic, decision tree 184, 185–190
Gaussian radial basis function 204
gradient boosting, neural network 210
Graph Builder 45–61
graphs
bar chart 59–61
bubble plot 53–55
contours 55–56
Graph Builder dialog box 45–48
line graphs 55–56
scatterplot matrix 48–51, 123, 126
trellis chart 51–53, 55–56, 58
Group X drop zone 46–47
Group Y drop zone 46–47

H

hidden layer, neural network 205, 208–210
hierarchical clustering 154–163, 177
high-variance procedure, decision tree as 198–199
Histogram, Excel Data Analysis Tool 21–22
holdback validation, neural network 206, 215
homocedasticity assumption 105
Hsu's MCB (multiple comparison with best) 94–95
hyperbolic tangent *(tanh)* 204
hypothesis testing 24–26

I

"include all variables" approach, logistic regression 123, 124
indicator variables 76–77, 79–82, 212
input layer, neural network 202–203
in-sample and out-of-sample data sets, measures to compare 82, 228–229, 244
interactions terms, introducing 130–132
interdependence, multivariate analysis framework 11
See also cluster analysis
See also Principal Component Analysis

J

JMP
See SAS JMP statistical software application
Johnson Sb transformation 211
Johnson Su transformation 211

K

k-fold cross-validation, neural network 207–208
k-means clustering 154, 164–177

L

Lack of Fit test, logistic regression 122, 128
Leaf Report, decision tree 198
learning rate for algorithm 210
least squares criterion 206, 211
least squares differences (LSD) 94
Levene test, ANOVA 89, 90, 102
lift chart 237–240
line graphs 55–56
Linear Probability Model (LPM) 105
linear regression
See also logistic regression
definition 65
LPM 105
multiple 67–76
simple 64–66
sum of squared residuals 128
linearity of logit, checking 132
loading plots 139, 145–146, 148–149
log odds of 0/1 convention 113
logistic function 106–112
logistic regression
bivariate method 124
decision tree method 124
lift curve 237–240
logistic function 106–112
LPM 105
odds ratios 109–111, 113–122
ROC curve 235–237
statistical study example 122–133
stepwise method 124
logit transformation 107
LogWorth statistic, decision tree 185, 186, 187–190

low- vs. high-variance procedures 198–199
LPM (Linear Probability Model) 105
LSD (least squares differences) 94
LSMeans Plot command 95, 97

M

Make into Data Table, ROC curve 236
Mark Clusters, Cluster command 157
market basket analysis 252
mean absolute error (MAE) measure 226, 244
mean square error (MSE) measure 226, 244
means comparison tests, ANOVA 90–95
Means/ANOVA command 88–89
model comparison
  binary dependent variable 230–237
  continuous dependent variable 226–230, 244
  introduction 225
  lift chart 237–240
  training-validation-test paradigm 240–246
Model Launch command, neural network 216
Mosaic plot 34–35
Move Up, Value Ordering 116–117
MSE (mean square error) measure 226, 244
multicollinearity of independent variables 73–74
multiple regression 67–76
Multivariate command 67, 142
multivariate data analysis
  and data sets 37–39
  as prerequisite to predictive modeling 249–250
  commonality for practical statistical study 7
  framework 9, 11
multiway splits in decision tree 181, 185

N

neural networks
  basic process 202–206
  data preparation 212–213
  fitting options for the model 206, 211, 215, 220
  hidden layer structure 205, 208–210
  prediction example 213–223
  purpose and application 201
  validation methods 206–208, 215–216
New Columns command 100
no penalty fit option 211
nominal data 26
nonlinear transformation 74, 204
normal (bell-shaped) distribution 18
Normal Quantile Plot, Distribution command 85–87, 125
Number of Models, neural network 216
Number of Tours, neural network model 216, 217

O

odds ratios, logistic regression 109–111, 113–122
Odds Ratios command 116, 118
one-sample hypothesis testing 24–25
one-way/one-factor ANOVA 83–96
online analytical processing (OLAP) 40–45
optimal classification, ROC curves 233–235, 236
ordinal data 26
outliers, scrubbing data of 212, 219
out-of-sample and in-sample data sets, measures to compare 82, 228–229, 244
output layer, neural network 202–203
overfitting the model/data
  clusters 164
  decision trees 191
  neural network 206–211, 216, 218
  train-validation-test paradigm to avoid 240–246
Overlap drop zone 46–47
Overlay Plot command 166–167

P

Pairwise Correlations, Multivariate command 142
parallel coordinate plots, *k*-means clustering 172–173
Parameter Estimates, Odds Ratios command 118
parsimony, principle of 74, 123
partition initial output, decision tree 183–184, 193
PCA
  See Principal Component Analysis
penalty fit method 211, 215, 220
PivotTable, Excel 40–42
Plot Residual by Predicted 72–73, 218–219
PPAR (plan, perform, analyze, reflect) cycle 9–11
practical statistical study 7, 8–9
prediction task, predictive analytics 254
predictive analytics
  availability of courses 7
  definition 4, 252
  framework 252–253
  goal 253–254
  model development and evaluation phase 255–256
  multivariate data analysis role in 249–250
  phases 254–256

specific applications 5
tasks of discovery 254
vs. statistics 254–255
predictive modeling
See predictive analytics
Principal Component Analysis (PCA)
dimension reduction 136, 142–144
eigenvalue analysis of weights 141–142
example 135–140
structure of data, insights into 145–149
vs. factor analysis 140–141
probabilities
estimating for logistic regression 119–120
relationship to odds 112
probability formula, saving 119
proportion of variation method, PCA 144, 148
pruning variables in decision tree 191, 195–196
p-values, hypothesis testing 25–26

R

random sample 14, 20–24
Range Odds Ratios, Odds Ratios command 116
Receiver Operating Characteristic (ROC) curve 191–192, 232–237
regression
See also logistic regression
categorical variables 76–82
clusters 164
continuous variables 76–77
fitting to the model 65, 67–69, 71, 74, 122, 128
linear 64–76, 105, 128
multiple 67–76
purposes 64
simple 64–66
stepwise 74–75, 124, 241–243
regression tree 192–199
relative absolute error 227
relative squared error 226
Remove Fit, neural network 215
repeated measures ANOVA 82
representative sample 14
residuals
ANOVA 85, 87
linear regression 128
multiple regression 72–73
neural network 218–219
return on investment (ROI) from data collection 2–3
robust fit method 211
ROC (Receiver Operating Characteristic) curve 191–192, 232–237
root mean square error (RMSE/s^e) measure 75, 76, 140, 192, 226
RSquare or R^2 (coefficient of determination) 66

S

sampling
in-sample and out-of-sample data sets 82, 228–229, 244
one-sample hypothesis testing 24–25
principles 14–15, 18–20
random sample generation 20–24
SAS JMP statistical software application
See also specific screen options and commands
as used in book 10, 11
deciding on best statistical technique 28–36
features to support predictive analytics 58, 254
opening files in Excel 28
saturated model, logistic regression 122
scales for standardizing data, neural network 212
scatterplot matrix 48–51, 123, 126
score plot 139, 145
scree plot
hierarchical clustering 160
PCA 142–143, 145, 146, 147
s^e (RMSE) 75, 76, 140, 192, 226
Selection button, copying output 44
SEMMA approach 256
sensitivity component of prediction model 231
Show Split Count, Display Options 188
Show Split Prob, Display Options 185
simple regression 64–66
single linkage method, distance between clusters 154–155
sorting data
Graph Builder 59–60
PCA 142, 145
specificity component of prediction model 231
Split command, decision tree variables 185–186
squared penalty fit method 211, 220
squaring distances, *k*-means clustering 173–174
SSBG (sum of squares between groups) 82, 83
SSE (sum of squares between groups [or error]) 82, 83, 166–167, 175
standard error 19
standardized beta coefficient (Std Beta) 69, 71

statistical assumptions, testing for
one-way ANOVA 85–89
statistics coursework
central limit theorem 18–24
coverage and real-world limitations 5–7
effective vs. ineffective approaches 26–36
one-sample hypothesis testing and p-values 24–26
sampling principles 14–15, 18–20
statistics as inexact science 14, 15–16
Z score/value 17, 24–25
statistics vs. predictive analytics 254–255
Std Beta, Fit Model command 69, 71
stepwise regression 74–75, 124, 241–243
Subset option, Table in Graph Builder 58–59
sum of squares between groups (or error) (SSE) 82, 83, 166–167, 175
sum of squares between groups (SSBG) 82, 83
Summary Statistics, Distribution command 175
supervised (directed) predictive analytics techniques 252, 253, 254

T

tables 40–45
Tabulate command 42–45
testing for differences, one-way ANOVA 90–96
testing statistical assumptions, one-way ANOVA 85–89
Tests that the Variances are Equal report 85
time series, Durbin-Watson test 73
total sum of squares (TSS) 82, 83
train-validate-test paradigm for model evaluation 240–246
Transform Covariates, neural network 212
trellis chart 51–53, 55–58
true positive rate (TPR) component of prediction model 232
TSS (total sum of squares) 82, 83
t-test 65, 71–72, 93
Tukey HSD test 93, 95
Tukey-Kramer HSD test 93, 95
2R (representative and random) sample 14, 16
two-way/two-factor ANOVA 97–102

U

unequal replication design, ANOVA with 97
Unequal Variances test, ANOVA 85, 86, 89
Unit Odds Ratios, Odds Ratios command 116
univariate analysis 6
unsupervised (undirected) predictive analytics techniques 252, 253, 254

V

validation
logistic regression 132
neural network 206–208, 215–216
train-validate-test paradigm 240–246
Validation variable 208
Value Ordering, Column properties 116–117
variables
See also categorical variables
See also continuous variables
automatic assignments for neural network 214
binary dependent variable 104, 221–222, 230–237
decision tree 182–191, 194–196
dummy 76–77, 79–82, 212
model building 123–124
multicollinearity 73–74
neural network 208
reclassifying 113, 123
weighting 141–142, 211, 215
variance inflation factor (VIF) 73, 74

W

Ward's method, distance between clusters 154
weak classifier, boosting option 210
weight decay penalty fit method 211
weighting of variables 141–142, 211, 215
Welch's Test 85, 86, 89, 90
Whole Model Test 121–122, 127–128
within-sample variability 82
without replication design, ANOVA 97
Wrap drop zone 46–47

Z

Z score/value 17, 24–25

From *Fundamentals of Predictive Analytics with JMP®* by Ron Klimberg and B. D. McCullough.
Copyright © 2012, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

About These Authors:

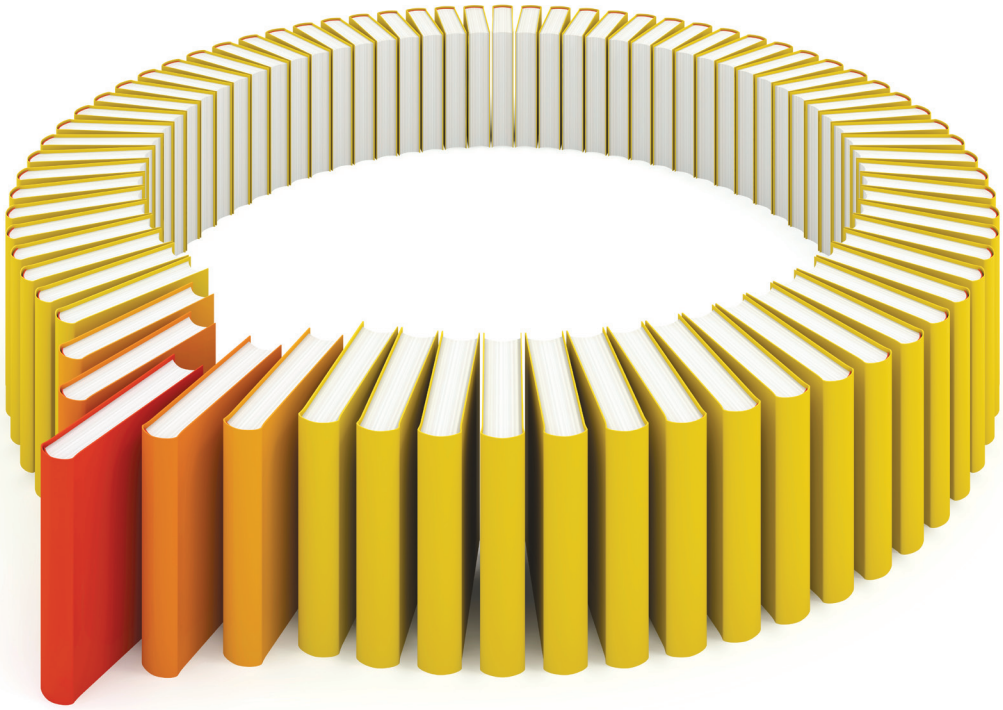
Ron Klimberg is professor at the Haub School of Business at Saint Joseph's University in Philadelphia, PA. Before joining the faculty in 1997, he was professor at Boston University, an operations research analyst for the Food and Drug Administration, and a consultant. His primary research interests lie in the areas of multiple criteria decision making, DEA, facility location, data visualization and data mining. Ron was the 2007 recipient of the Tengelmann Award for his excellence in scholarship, teaching, and research. He received his PhD from The Johns Hopkins University..

B.D. McCullough is professor at the LeBow College of Business at Drexel University in Philadelphia, PA. Prior to joining Drexel, he was a senior economist at the Federal Communications Commission and an assistant professor at Fordham University. His research fields include applied econometrics and time series, accuracy of statistical and econometrics software, replicability of research, and data mining. He received his PhD from the University of Texas at Austin.

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

http://support.sas.com/klimberg

http://support.sas.com/mccullough



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.

jmp

sas
THE POWER TO KNOW.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. © 2013 SAS Institute Inc. All rights reserved. S108082US.0613