Titanic Passengers
Logistic Regression

From *Building Better Models With JMP Pro*, Chapter 5, SAS Press (2015). Grayson, Gardner and Stephens.

Used with permission. For additional information, see community.jmp.com/docs/DOC-7562.

Titanic Passengers
Logistic Regression

Key ideas: Logistic regression, log odds and logit, odds, odds ratios.

Background

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1,502 of the 2,224 passengers and crew. This sensational tragedy shocked the international community and motivated the adoption of better maritime safety regulations.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others. ("Titanic: Machine Learning from Disaster." From a Kaggle competition. Available at http://bit.ly/1f2crzi, data accessed 08/2014.)

The Task

We use this rich and storied example to explore some questions of interest about Titanic survival rates. For example, were there any key characteristics shared by survivors? Were some passenger groups more likely to survive than others? Can we accurately predict survival?

We will fit a logistic regression model using the available data to explore these questions.

The Data Titanic Passengers BBM.jmp

This data table describes the survival status of 1,309 of the 1,324 individual passengers on the Titanic. Information on the 899 crew members is not included.

Name:	Passenger Name
Survived:	Yes or No
Passenger Class:	1, 2, or 3 corresponding to 1st, 2nd, or 3rd class
Sex:	Passenger sex
Age:	Passenger age
Siblings and Spouses	The number of siblings and spouses aboard
Parents and Children	The number of parents and children aboard
Fare	The passenger fare
Port	Port of embarkment (C=Cherbourg; Q=Queenstown; S=Southampton)
Home/Destination	The home or final intended destination of the passenger

Analysis

We begin by examining the data, one variable at a time, two at a time, and many at a time. We only show the distribution of the response **Survived** and the relationship between **Survived** and the other potential predictor variables; however, additional visualization and exploratory tools should also be used.

Since we're interested in understanding survival rates, we've applied the **Value Ordering** column property so that **Yes** (survived) appears first in graphics and analyses. This also allows us to model the probability of survival when we build our logistic regression model. (To apply value ordering, right-click on

the column name in the data table and select **Column Info**. Then, under **Column Properties**, select **Value Ordering**. Click on the category you'd like to appear first, and click **Move Up**, then click **OK**.

We use the **Distribution** platform to explore variables one at a time. The distribution of **Survived** is shown in Exhibit 1 (use **Analyze > Distribution**, select **Survived** as **Y, Columns**, and click **OK**. For a horizontal layout, select **Stack** from the top red triangle.)

About 38% of passengers in our data set survived the sinking of the Titanic.

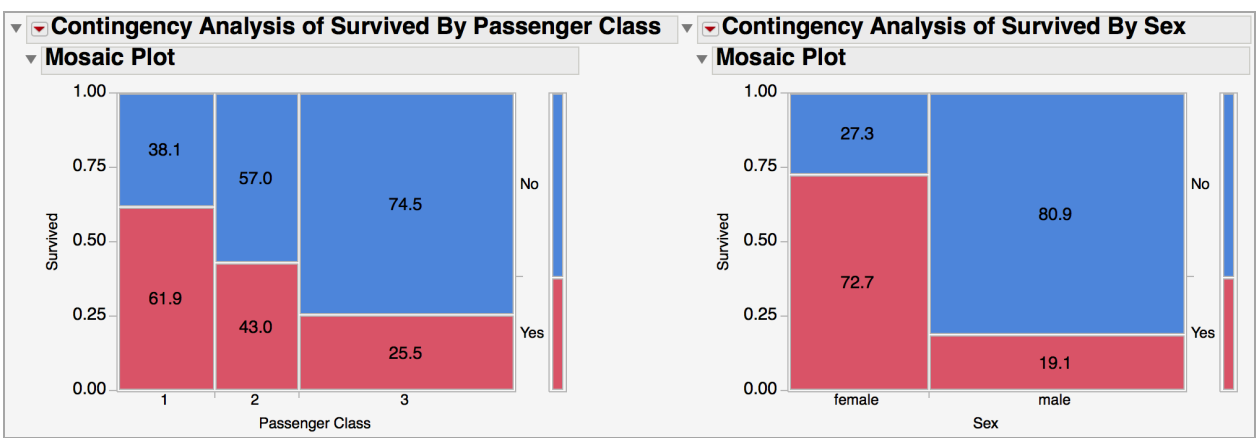
Exhibit 1 Distribution of Survived

A look at two-way relationships among the response and likely predictors shows many potential (and not altogether surprising) relationships. In the examples to follow, we use **Fit Y by X** with **Survived** as **Y, Response** and predictors as **X, Factors**.

In the mosaic plots in Exhibit 2, we see that first class passengers had a higher survival rate (61.9%) than second (43%) or third class passengers (25.5%), and females (72.2%) fared much better than males (19.1%). The contingency tables (not shown) display additional numeric summaries.

To label the cells with the row percentages, right-click on the graph and select **Set Colors** and **Cell Labeling > Show Percents**.

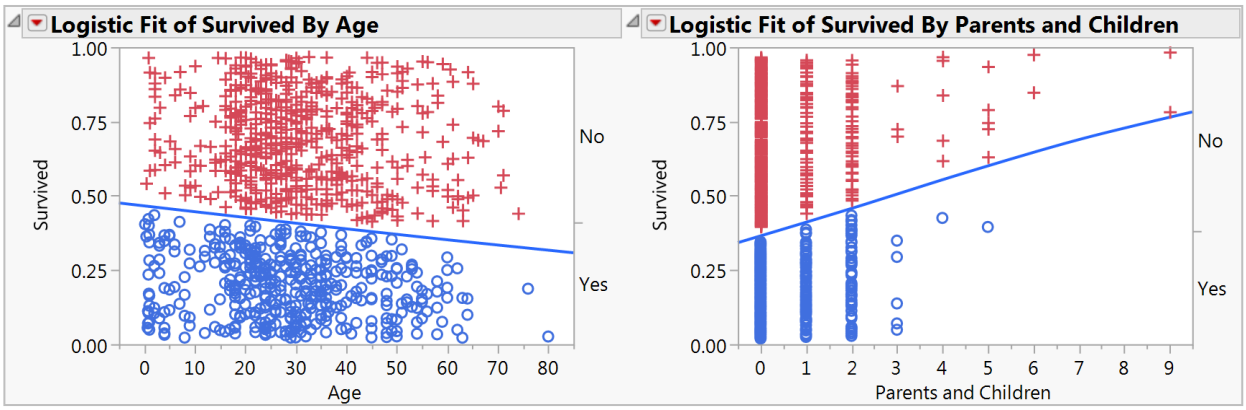
Exhibit 2 Survived versus Passenger Class and Sex



From the logistic plots in Exhibit 3, the survival rate at a particular value of the predictor is displayed as the area under the line. In each graph, the value on the *y*-axis is the probability of **Survived = Yes** (for a given value of the predictor). Ignoring other factors, it appears that the survival rate was higher for

younger passengers than for older passengers. It also appears that the survival rate is higher for passengers traveling with parents and children than for passengers traveling alone.

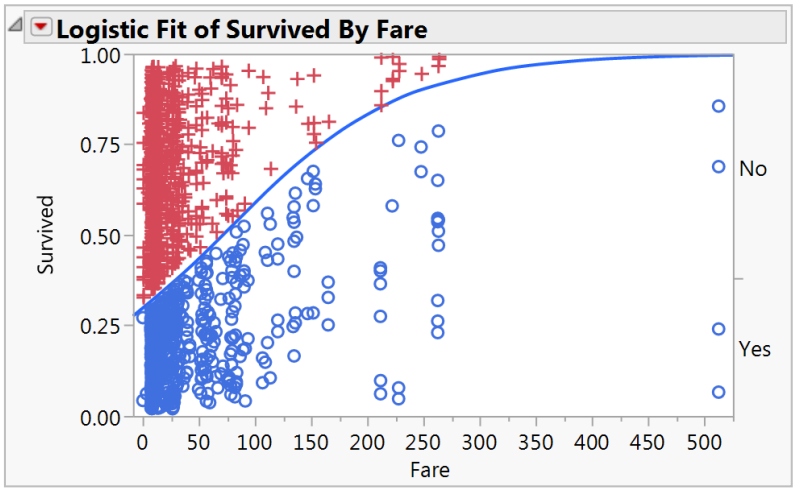
Exhibit 3 Survived versus Age and Parents and Children



In Exhibit 3, we use **Rows > Color or Mark by Column** to mark all rows where **Survived** is **Yes** with a "o" symbol, and Survived is No with an "+" symbol. These markers appear in all future exhibits.

Did the passengers who paid a higher fare have a better survival rate? Looking at the logistic fit for **Survived** versus **Fare** (Exhibit 4), we see that there is a strong relationship between the fare paid and the survival rate. Survival rates were higher for passengers who paid higher fares. However, note that there are four passengers who paid extremely high fares. Should we be concerned with this? Could these four observations be exerting a large influence on the model? The potential impact of these four points is examined as an exercise.

Exhibit 4 Survived versus Fare



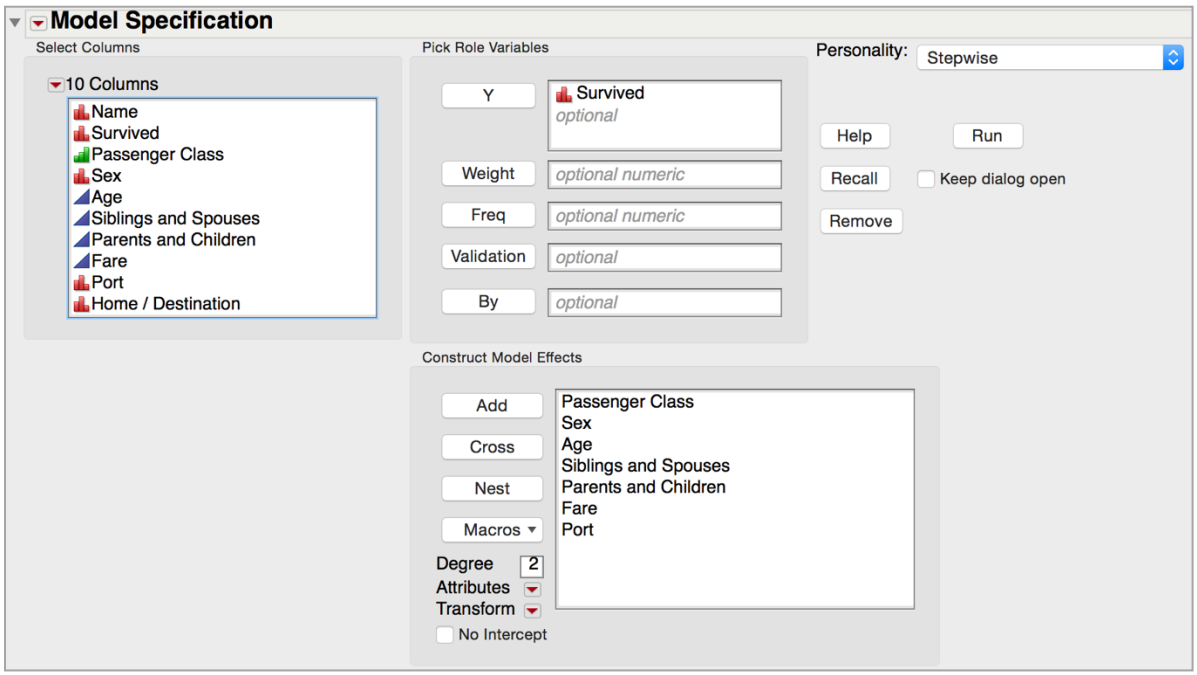
This exploratory work allows us to narrow down our list of potentially important predictor variables. We find that not all variables are important or even useful, such as **Name** and **Home/Destination**. We omit these variables from future analyses.

Modeling

We use **Analyze > Fit Model** to fit a nominal logistic regression model using possible predictors **Passenger Class**, **Sex**, **Age**, **Siblings, and Spouses**, **Parents and Children**, **Fare** and **Port**. Since we have many variables, we will use stepwise regression to aid in the selection of the final model (select the **Stepwise** personality in the Model Specification window). The completed **Fit Model** dialog is shown in Exhibit 5.

Click **Run** to launch the Stepwise platform.

Exhibit 5 Survived versus Fare



Stepwise Regression

Stepwise regression provides a number of stopping rules for selecting the best subset of variables for the model. The default stopping rule for model selection in the stepwise platform is **Minimum BIC**, or minimum *Bayesian Information Criterion*. Another stopping rule is **Minimum AICc** (*Akaike's Information Criterion*). The **Direction**, 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 or AICc statistic is selected.

In this situation, using either minimum AICc or BIC produces the same suggested model, but this is not always the case[1].

The **Rules** option in the **Stepwise Regression Control** panel relates to how stepwise regression handles categorical variables and interaction effects. We have entered two three-level categorical variables, **Passenger Class** and **Port**. For each of these variables, two parameters can be estimated (see the

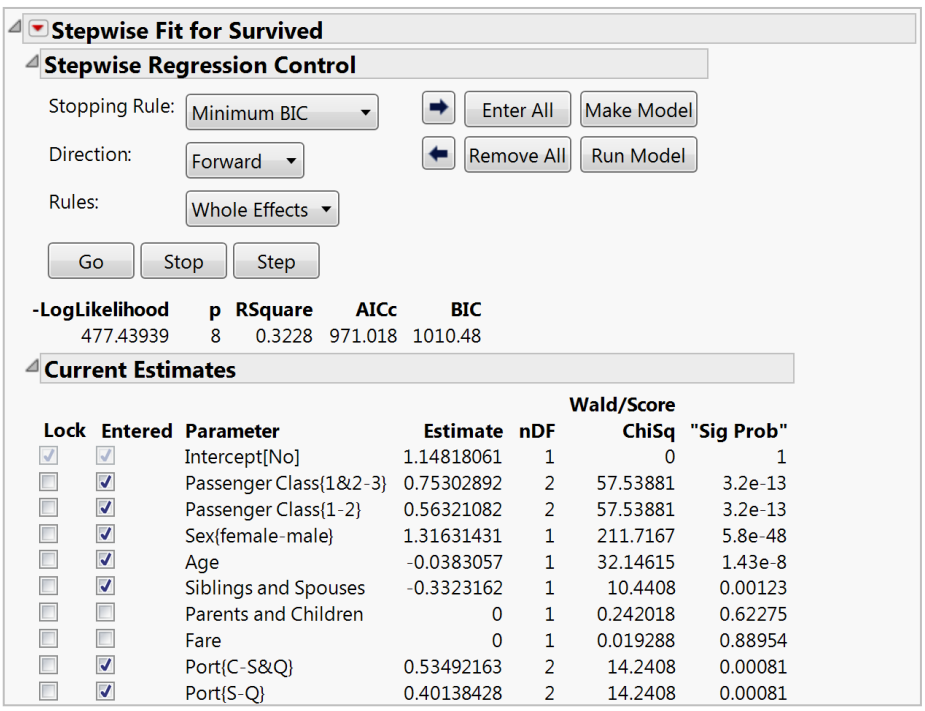
[1] See Building Better Models with JMP Pro, Chapter 4 for details on stopping rules, or search for Stopping Rule in the JMP Help.

Current Estimates section in Exhibit 6). In constructing these parameters, JMP codes factor levels in a hierarchical fashion. The first parameter for **Port**, **Port(Q&S-C)**, combines ports Q and S into one group with port C in a separate group. The split into these two groups results in the greatest difference in the probability of survival. JMP creates a total of k-1 of these parameters for each categorical variable, where k is the number of factor levels[2].

To simplify the model and interpretation of terms in the final model, we change the rule to **Whole Effects**. With this rule, if one parameter for a variable is entered into the model, JMP will enter all remaining parameters for that variable.

Stepwise results (after clicking **Go**) using **Minimum BIC** and **Whole Effects** are shown in Exhibit 6. Variables selected by stepwise regression are marked with a check under **Entered** in the **Current Estimates** panel (bottom, in Exhibit 6). The five whole effects estimated by stepwise regression are **Passenger Class**, **Sex**, **Age**, **Siblings and Spouses**, and **Port.** Note that these check boxes can be used to manually enter or remove terms from the model.

Exhibit 6 Stepwise Logistic Regression Results



To run the logistic model with the selected parameters, we click **Make Model** to return to the **Model Specification** window, and then click **Run**. The resulting summary of the model fit is shown in Exhibit 7.

The model is highly significant, with a **Prob>ChiSq** <0.0001. However, the misclassification rate indicates that the model misclassified 21.17% of the passengers. If the goal of this modeling process was to predict survival beforehand, this would certainly be undesirable. However, our goal is to uncover the important

[2] This feature is described in more detail in the *Specialized Models* JMP manual, found under the **Help > Books** menu in JMP.

factors related to survival. As such, in this analysis we are not overly concerned with misclassification rate.

Note that **Fare** was not included in the stepwise model fit (see Exhibit 6), although by itself it was strongly related to survival. Also, **Port** was included in the model, though it is reasonable to question whether that variable is truly predictive. We will explore possible reasons why predictors were or were not included in the reduced model in an exercise.

Exhibit 7 The Reduced Logistic Model for Survived

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	228.04248	7	456.085	<.0001*
Full	477.47308			
Reduced	705.51556			

RSquare (U)	0.3232
AICc	971.085
BIC	1010.55
Observations (or Sum Wgts)	1044

Measure	Training	Definition
Entropy RSquare	0.3232	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.4775	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.4573	∑ -Log(ρ[j])/n
RMSE	0.3821	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.2929	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.2117	∑ (ρ[j]≠ρMax)/n
N	1044	n

Looking at parameter estimates in Exhibit 8 allows us to gain insight into the effect of each predictor on survival rate. For example, the estimate for **Age** is negative, indicating that older passengers have a lower survival rate than younger passengers. For the two-level predictor **Sex**, only one of the parameter estimates is displayed—the estimate for females. The estimate for **Sex[Females]** is 1.316, so the estimate for **Sex[Males]** is -1.316. This tells us that females have a higher survival rate than males.

The note at the bottom of the Parameter Estimates table, **For log odds of Yes/No**, tells us that JMP predicts the probability that **Survival** = **Yes**. Recall that we changed the value ordering at the beginning of this example.

Exhibit 8 Parameter Estimates for Survived

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	2.28855656	0.3502206	42.70	<.0001*
Passenger Class[2-1]	-1.1270844	0.2437888	21.37	<.0001*
Passenger Class[3-2]	-0.9437064	0.2028802	21.64	<.0001*
Sex[female]	1.31659442	0.088189	222.88	<.0001*
Age	-0.0383599	0.0067077	32.70	<.0001*
Siblings and Spouses	-0.3323671	0.1030565	10.40	0.0013*
Port[C]	0.71326236	0.1882365	14.36	0.0002*
Port[Q]	-0.7578514	0.2757081	7.56	0.0060*

For log odds of Yes/No

What do the parameter estimates actually represent? And, how do we interpret parameter estimates for **Passenger Class**, a variable which has ordered categories and is coded as ordinal? To understand these estimates, we need to take a peek under the hood of logistic regression and provide a few technical details.

The Logistic Regression Model

For logistic regression, the probability of an event, *p*, is related to predictive factors ($X_1, X_2, \ldots, X_k$) by the mathematical relationship

$$\log(p/(1-p)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

where *p/(1-p)* is called the *odds* of the event and *log(p/(1-p))* is called the *log-odds* or the *logit*.

The right side of this equation looks a lot like our multiple linear regression model (without the error). When a logistic model is fit, JMP reports the coefficients (i.e., estimates of betas in the equation) in the parameter estimates table. The probability of an event can be estimated from these coefficients. Rearranging the formula above to solve for *p*, we have

$$p = 1/(1 + \mathrm{e}^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)})$$

These formulas can be seen when we save the probability formula to the data table (from the red triangle for the model select **Save > Save Probability Formula**). The resulting logit of the probability model, saved as **Lin[Yes]** in the data table, is shown in Exhibit 9. Here, the coefficients for **Passenger Class 2** and **Passenger Class 3** are the log odds of surviving compared to **Passenger Class 1**.

Exhibit 9 The Saved Logit Function, Lin[Yes]

```
2.28855656033793
+ Match( Passenger Class ) {
    1    ⇒ 0
    2    ⇒ -1.1270843854682
    3    ⇒ -2.0707907549958
    else ⇒ .
  }
+ Match( Sex ) {
    "female" ⇒ 1.31659442170986
    "male"   ⇒ -1.3165944217099
    else     ⇒ .
  }
+ -0.0383598691706 * Age
+ -0.3323671317032
  * Siblings and Spouses
+ Match( Port ) {
    "C" ⇒ 0.71326235623133
    "Q" ⇒ -0.7578513857347
    "S" ⇒ 0.04458902950341
    else ⇒ .
  }
```

The probability of survival, saved as **Prob[Yes]** in the data table, is then calculated from this formula (see Exhibit 10).

Exhibit 10 Formula for the Probability of a Survival, Prob[Yes]

$$\frac{1}{\left(1 + \text{Exp}\left(-\,Lin[Yes]\right)\right)}$$

Interpreting Regression Coefficients for Passenger Class

From the original data we see that the probability that a second class passenger survived was around 0.43. See the mosaic plot for **Survived** versus **Passenger Class** in Exhibit 2.

The odds of surviving for a second-class passenger can be calculated using the ratio *p/(1-p)*:

$$\left(\frac{\text{P(Survivel Second Class)}}{\text{P(Not Survivel Second Class)}}\right) = \left(\frac{\text{P(Survivel Second Class)}}{\text{1-P(Survivel Second Class)}}\right)$$

$$= \frac{0.43}{1-0.43} = 0.754$$

Similarly, we could calculate the odds of surviving for first class passengers. If we want to compare these two odds, we can calculate the *odds ratio*. If the odds ratio is close to 1, there is little difference in the two conditions with respect to helping to predict the response.

The parameter estimate for **Passenger Class[2-1]** is -1.127 (Exhibit 8). It turns out that this is an estimate of the *log of the odds ratio* (or *log odds*) for the odds of surviving in second class versus the odds of surviving in first class. This estimate indicates that the overall odds of surviving in second class was $e^{-1.127}$ = 0.324 lower than the odds of surviving in first class.

We see a similar result when we compare the odds of surviving in third class versus second class. The parameter estimate for **Passenger Class[3-2]** is -0.943, indicating that the odds of surviving in third class were $e^{-0.943}$ = 0.389 times lower than in second class.

To display these odds ratios, as well as odds ratios for all predictors in the model, select **Odds Ratios** from the red triangle in the **Nominal Logistic Fit** window. Odds ratios for **Passenger Class** are shown in Exhibit 11.

Exhibit 11 Odds Ratios for Passenger Class

Odds Ratios for Passenger Class

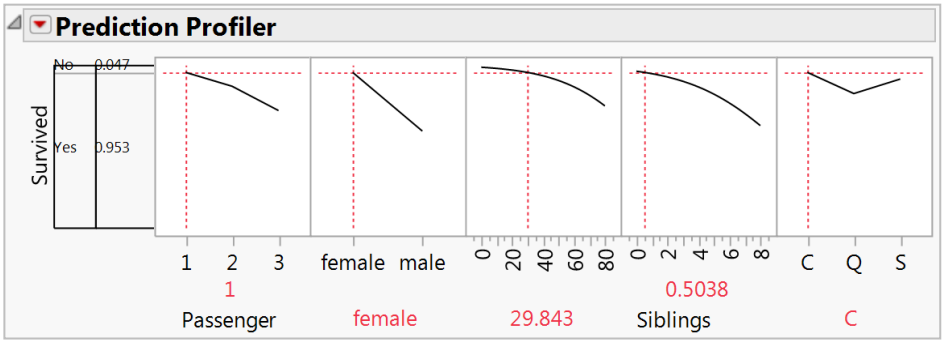
Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
2	1	0.3239765	<.0001*	0.1998471	0.5202023
3	1	0.126086	<.0001*	0.0783351	0.1999912
3	2	0.3891827	<.0001*	0.260867	0.5783584
1	2	3.0866439	<.0001*	1.922329	5.0038248
1	3	7.9310922	<.0001*	5.0002206	12.765669
2	3	2.5694873	<.0001*	1.7290317	3.8333716

We can use the **Prediction Profiler** (Exhibit 12) to help interpret parameter estimates and to explore predicted survival rates for different values of the predictor variables. (In the **Nominal Logistic Fit** window, select **Profiler** from the top red triangle).

From the Prediction Profiler in Exhibit 12, we can see that the probability of survival for a 30 year old first class female, traveling alone or with one sibling or spouse from Cherbourg, is roughly 0.95.

The slopes of the lines in the Profiler indicate how the predicted probability of survival changes for different predictor values. What happens to the probability of survival if we change the value of sex from female to male?

Exhibit 12 Prediction Profiler for Logistic Regression Model



Summary

Statistical Insights

Back to our original task: Determine the characteristics of the "survivors" and identify those groups of people that were most likely to survive. First-class and second-class females, particularly if they were young and traveling alone, had a very good chance of survival. The phrase "women and children first!" appeared to hold true for this group of passengers.

A final thought: while this model is highly significant, could a different model provide better predictions (and a lower misclassification rate)? We've fit a model involving only main effects. We did not include *interactions* between the predictors. If a two-way interaction is significant, then the effect of a factor on the response is dependent on the value of the other factor. What if we add interactions to the model? Would our model predictions improve? We revisit this question in an exercise.

Implications

The use of a validation column for stepwise procedures was not discussed in this case study. However, model validation is recommended, wherever possible, to aid in building and assessing models (see *Building Better Models with JMP Pro*, Chapters 6 and 8, for more information on model validation).

Although not discussed in this case study, misclassification is a key measure of how well a model predicts. The confusion matrix quantifies two types of misclassification (i.e., false positives and false negatives). Other ways to evaluate a classifier model, such as ROC and Lift Curves, should also be used to assess model performance (see *Building Better Models with JMP Pro*, Chapter 6 for more information on these measures).

JMP Features and Hints

Use exploratory graphical tools and platforms, such as **Distribution**, **Fit Y by X**, **Graph Builder**, and **Scatterplot Matrix** to get to know your data, explore potential relationships, identify potential problems (e.g., outliers), and narrow down the list of potential predictor variables to include in your initial model. Such exploratory work has been demonstrated only partially here.

For scenarios with a large number of possible predictors, use stepwise regression and explore different stopping rules to assist in selecting the best model.

By default, JMP will model the probability of the event that appears first in alphanumeric order. To change the target outcome category, you can use the value ordering property for the data table's response column. You can also change the labeling of response categories to make the labels more descriptive.

Exercises

Exercise 1: Use the Titanic Passengers BBM.jmp data set for this exercise.

a. Re-create the analysis and graphs for Exhibits 3 and 4, recording parameter estimates for each simple model.
b. Exclude points that appear to be extreme outliers or potentially influential, and refit the simple models.
c. Examine parameter estimates for the models fitted with and without extreme points. What can you conclude about the influence of these points on the models?
d. Repeat this examination of influential points on the full, fitted logistic regression model. Compare the resulting model to the model shown in Exhibit 8.

Exercise 2: Use the Titanic Passengers BBM.jmp data set for this exercise.

In the one-factor-at-a-time analysis of **Survival** versus each predictive factor, **Parents & Children** appeared to be strongly related to survival. However, when the stepwise regression procedure was used to choose the best model, this factor was not included.

Explore this data in an attempt to uncover reasons for this apparent paradox. In particular, look for relationships between potential predictors to see if this can provide a reasonable explanation.

Exercise 3: Use the Titanic Passengers BBM.jmp data set for this exercise.

a. Interpret the odds ratio for **Passenger Class 1** versus **Passenger Class 3** in Exhibit 11.
b. Use the output in Exhibit 8 to calculate the odds ratio for survival for **Passenger Class 3** versus **Passenger Class 1** (show your work). Interpret this odds ratio.

Exercise 4: Use the Titanic Passengers BBM.jmp data set for this exercise.

a. Fit a logistic regression model, but this time include two-way interactions between **Passenger Class**, **Age** and **Sex**. First, add all predictors to the model. Then, highlight these three variables in the column selection panel and select **Macros > Factorial to degree**. Interactions are represented like this: **Passenger Class*Age**. (Hint: To add interactions one at a time, select two variables in the column selection panel and choose **Cross**.)
b. Are any of the two-way interactions significant?
c. Use the **Effect Summary** table to reduce the model one term at a time. Note: If a term has a "^" next to the *p*-value in the **Effect Summary** table, it is involved in an interaction and should not be removed.
d. Use the **Prediction Profiler** to interpret the model. In particular, interpret the interaction between **Passenger Clas**s and **Age**.

Exercise 5: Use the Equity.jmp data from the **Sample Data Library** for this exercise.

A response (or Y) variable is **BAD**, which is coded as 0 (good credit risk) or 1 (bad credit risk). The other variables are:

LOAN	The amount of the loan requested
MORTDUE	How much the customer needs to pay on their mortgage
VALUE	Assessed valuation
REASON	Debt consolidation or home improvement (DebtCon or HomeImp)
JOB	Broad job category
YOJ	Years on the job
DEROG	Number of derogatory reports
DELINQ	The number of delinquent trade lines (or credit accounts)
CLAGE	Age of oldest trade line (oldest credit account)
NINQ	Number of recent credit inquiries
CLNO	Number of trade lines
DEBTINC	Debt to income ratio

a. Use the **Columns Viewer**, **Distribution** and **Graph Builder** to familiarize yourself with this data.

   1. Do any variables appear to be related to **BAD**? Explain.
   2. List any potential data quality issues that you observe.

b. Fit a logistic regression model for **BAD**, including all predictor variables. Do not address data quality issues first (i.e., proceed with the data in its current form).

   1. What is the *p*-value for the model?
   2. What is the misclassification rate?
   3. What are the two types of misclassification error that can occur in this example? How many misclassifications of each type were made?
   4. Use the **Effect Summary** table to slowly remove non-significant terms from the model. How many terms are in your final model?
   5. What is the misclassification rate for this reduced model?
   6. In the context of this example, define the two types of classification error: false positive and false negative. Which type of classification error occurred more often? Explain.
   7. What are estimates (coefficients) for **DEROG** and **CLAGE**? Open the **Prediction Profiler** to explore what happens to the predicted probability that **BAD=1** as you increase and decrease the values of these two variables.
   8. You need to explain to your manager what coefficients for **DEROG** and **CLAGE** represent. Interpret the coefficients for these two variables in non-technical terms.

c. This data set has some data quality issues that should be addressed before modeling.

   1. Determine how to handle missing values and other potential data quality issues (see "Common Problems with Data" in *Building Better Models*, Chapter 3 for details).
   2. Fit a logistic regression model for **BAD** using this data, and again reduce this model using the **Effect Summary** table.
   3. What are the terms in this reduced model?

4. How did you handle variables with missing values? Were the missing values informative? Did the missingness provide information as to whether a customer is a good or bad credit risk? Explain.
5. What is the misclassification rate?

d. Describe differences between the final reduced models produced in parts b and c above.

1. Were any predictors included in one model but not the other?
2. Did you gain any new information or knowledge from using “cleaned and prepared” data to build your model? Explain.

SAS Institute Inc. World Headquarters +1 919 677 8000

JMP is a software solution from SAS. To learn more about SAS, visit **www.sas.com**

For JMP sales in the US and Canada, call 877 594 6567 or go to **www.jmp.com**

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. S81971.1111