JMP026: Lost Sales

Logistic Regression

Produced by:

Marlene Smith, University of Colorado Denver Business School

Lost Sales[1]

Logistic Regression

Background

In many industries throughout the world, suppliers compete for business by submitting quotes for work, services or products. A key criterion used to determine the winning quote is the dollar amount of the quote, but other factors include expected quality, estimated delivery time of the product, or quoted completion time of the work.

The focus of this case is a supplier of equipment to the automotive industry. The products of interest in this case are various precision metal components used in a range of automotive applications, such as braking systems, drive trains, and engines. Some of the products will be used in the manufacture or assembly of new automobiles (i.e. original equipment), while others will be used as replacement parts in automobiles already on the road (i.e. aftermarket).

The Task

The supplier wants to increase sales and expand its market position. Many of the quotes provided to prospective customers in the past haven't resulted in orders. Do the data provide any indication why? Are there certain situations that make it more or less likely that a customer will place an order?

The Data Lost Sales.jmp

The data set contains 550 records for quotes provided over a six month period. The variables in the data set are:

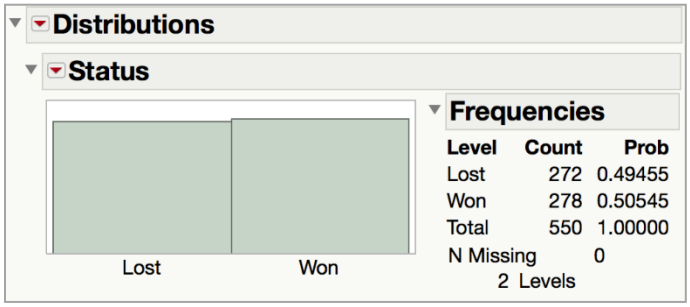
Quote	The quoted price, in dollars, for the order
Time to Delivery	The quoted number of calendar days within which the order is to be delivered
Part Type	OE = original equipment; AM = aftermarket
Status	Whether the quote resulted in a subsequent order within 30 days of receiving the quote: Lost = the order was not placed; Won = the order was placed.

Analysis

We begin by determining the current state of the company's ability to win orders. Exhibit 1 shows the bar chart and frequency distribution of Status. About half of the quotes don't result in subsequent orders within 30 days.

[1]Scenario and data provided by North Haven Group, LLC.

Exhibit 1 Bar Chart of Status

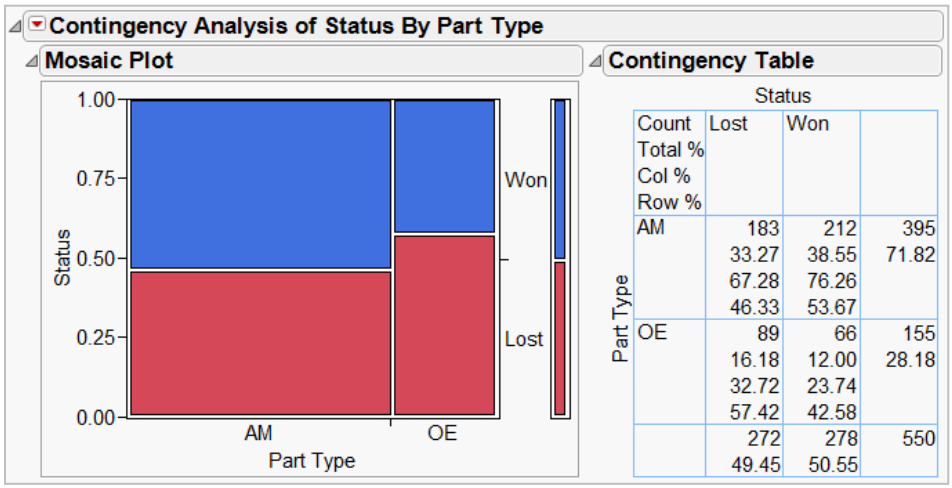


(Analyze > Distribution; Use Status as Y, Columns. Select Stack under the top red triangle for a horizontal layout.)

The company sells to both original equipment manufacturers (OE) and to aftermarket suppliers (AM). Sales managers expect that the "hit" rates for these two markets are vastly different.

Exhibit 2 looks at whether order-winning is associated with part type, OE or AM. There were 395 aftermarket quotes. Of these, 212, or 53.7%, resulted in an order. Of the 155 original equipment orders, only 42.6% led to an order.

Exhibit 2 Mosaic Plot and Cross-tabulation of Status and Part Type



(Analyze > Fit Y by X; Use Status as Y, Response and Part Type as X, Factor.)

The Likelihood Ratio and Pearson tests (Exhibit 3) indicate that the company's ability to win orders for original equipment and aftermarket is statistically different (at a significance level of 0.05). Significantly more of the original equipment orders are lost.

Exhibit 3 Likelihood Ratio and Pearson Tests

Tests

N	DF	-LogLike	RSquare (U)
550	1	2.7455572	0.0072

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	5.491	0.0191*
Pearson	5.477	0.0193*

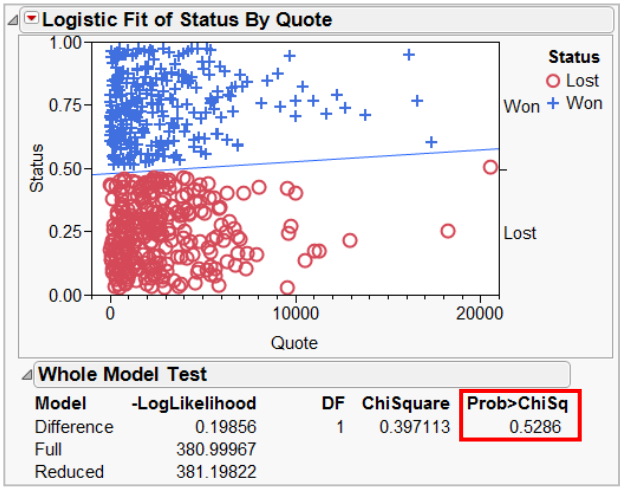
Fisher's Exact Test	Prob	Alternative Hypothesis
Left	0.0123*	Prob(Status=Won) is greater for Part Type=AM than OE
Right	0.9926	Prob(Status=Won) is greater for Part Type=OE than AM
2-Tail	0.0228*	Prob(Status=Won) is different across Part Type

(Results display by default at bottom of the previous analysis window.)

Many of the sales managers insist that a more important factor in lost sales opportunities is related to pricing. They believe that the company's prices are too high, and that customers are moving to lower cost providers.

To explore the relationship between Status and Quote we use logistic regression (Exhibit 4). The line in the graph represents the predicted probability that an order will be lost as the quoted price increases. Interestingly, the amount of the quote does not have a statistically significant influence on the probability that an order will be lost. This is evidenced by the nearly flat line, which remains near 0.5 regardless of the size of the quote, as well as the p-value (in the Whole Model Test, Prob>ChiSq = 0.5286).

Exhibit 4 Logistic Regression of Status by Quote



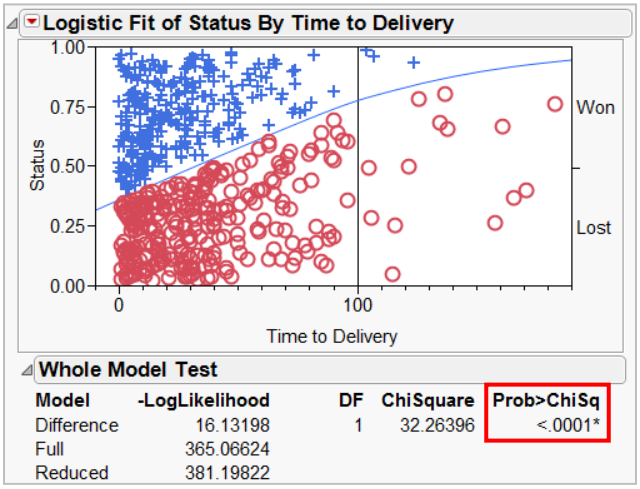
Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	0.19856	1	0.397113	0.5286
Full	380.99967			
Reduced	381.19822			

(Analyze > Fit Y by X; Use Status as Y, Response and Quote as X, Factor. To color the points by Status, right-click in the graph, select Row Legend, and select Status from the column list. To change the marker, select an option under Markers.)

The remaining factor is Time to Delivery – the quoted number of days before the order will be delivered. Again, we use logistic regression to explore the relationship between winning an order and delivery time (Exhibit 5), with the ultimate goal of predicting the probability of losing an order. Unlike Quote**,** Time to Delivery is an important predictor of whether or not the order is lost (the p-value is <0.0001). The line in the graph indicates that, as the quoted time to deliver increases, the probability of losing an order grows dramatically. For a quoted time of 10 days, the probability of losing an order is 0.4. At 100 days, that probability climbs to nearly 0.8!

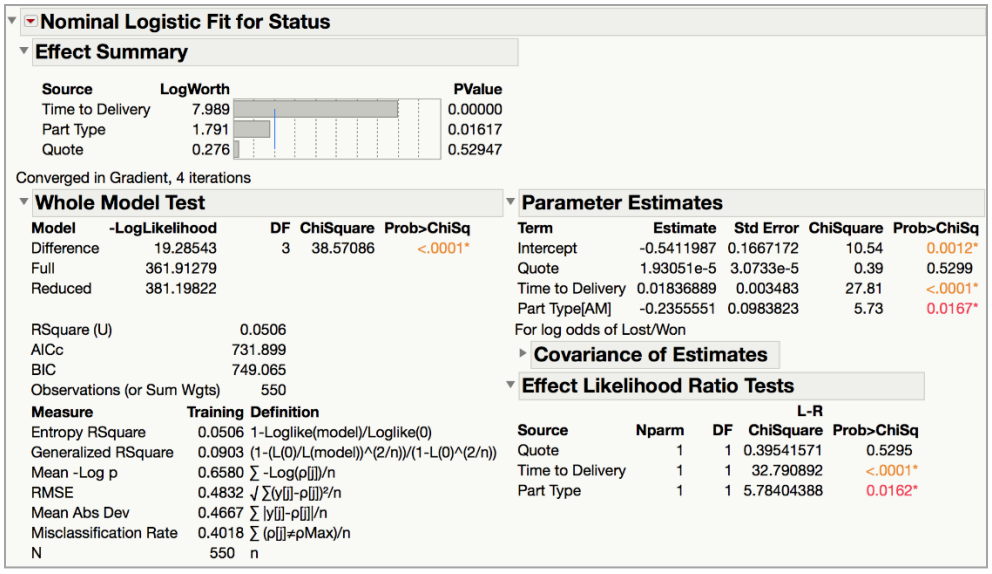
Exhibit 5 Logistic Regression of Status by Time to Delivery



(To add reference lines, double-click on the x-axis. At the bottom of the resulting window, under Reference lines, type a value, and click Add).

As with standard regression methods in which the response variable is continuous, we can also build multiple logistic regression models. We fit a model to estimate the probability of losing an order based on all three predictors: Part Type, Quote, and Time to Delivery (Exhibit 6).

Exhibit 6 Parameter Estimates for Multiple Logistic Model



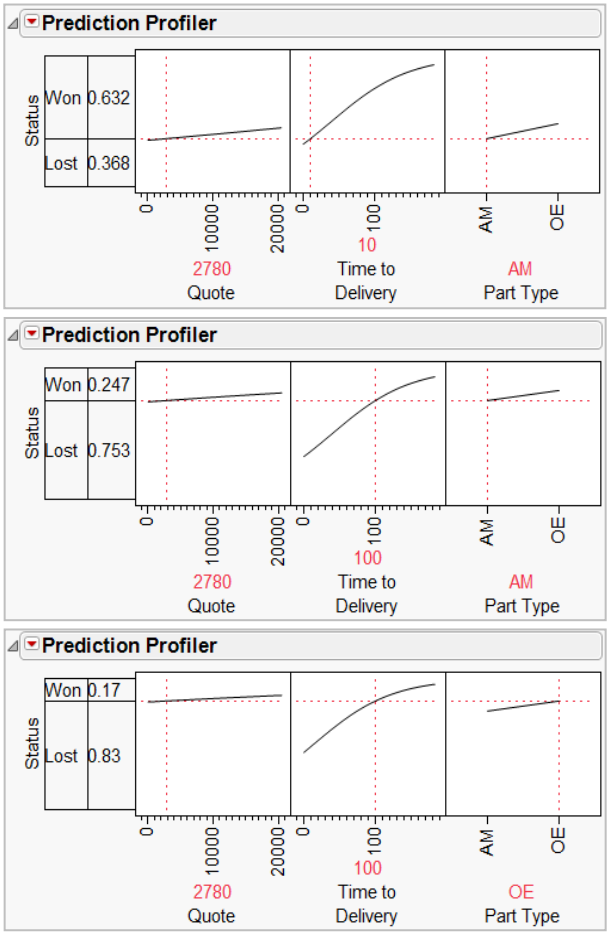
(Analyze > Fit Model; Use Status as Y, Response and the three predictors as model effects. By default, since Status is nominal, JMP will change the Personality to Nominal Logistic. Click Run to run the model. Note that the layout has been change to fit better on the page, and that not all default output is shown.)

As we discovered earlier, Quote is not statistically significant, while both Time to Delivery and Part Type are significant predictors of the probability of losing an order.

We explore the effects of these three predictors further using the Prediction Profiler (Exhibit 7). The profiler allows us to dynamically explore how the predicted probability of losing (or winning) an order changes as we change values of the predictors. The steep slope of Time to Delivery is a visual indication

of the significance of delivery time. For example, change the Time to Delivery to 10 days (as seen in the top panel in Exhibit 7), and the probability of losing an order drops to 0.368. Increase Time to Delivery to 100 days (middle panel, Exhibit 7), and the probability leaps to 0.753. (Note that these are slightly different from the earlier values, since we've included all three predictors in this model.)

Exhibit 7 Prediction Profiler for Status with all Predictors



(Click on the top red triangle next to Nominal Logistic Fit for Status, and select Profiler. To change a value for a predictor, click and drag the vertical red line for the predictor, or click on a value and type a new value.)

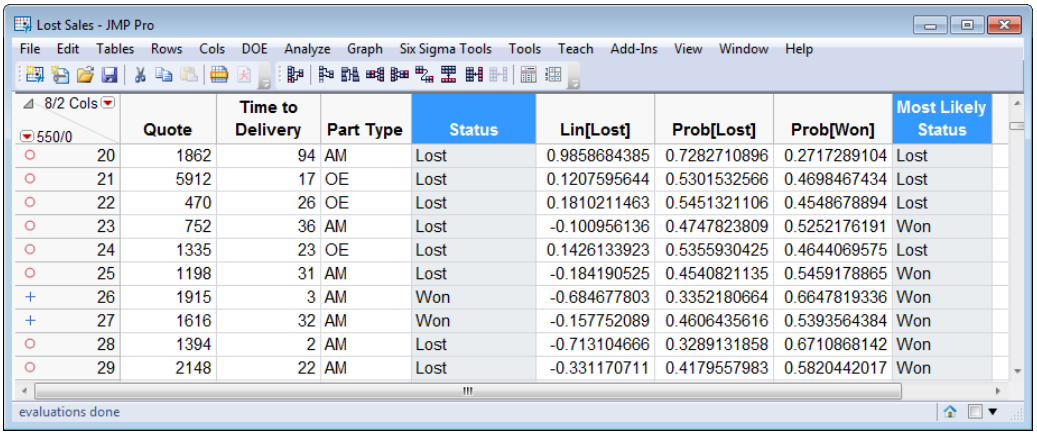
Exploring the profiler further, we see that the line for Quote is relatively flat. This is consistent with the p-value for Quote**;** changes in the quoted price do not have much of an impact on whether or not an order is lost. Meanwhile, changing Part Type from AM to OE increases the probability of losing an order by around 0.08, or 8% (bottom panel, Exhibit 7).

To summarize what we've learned thus far:

- Contrary to suspicions, quoted price is not a key driver in lost sales,
- As delivery time increases, so does the probability of losing an order, and
- For any given delivery time, an aftermarket order has a lower probability of being lost.

In a logistic regression model, as we have seen, a predicted value is actually a probability—in this case, the probability that a quote does not result in an order. The predicted probabilities can be saved to the JMP data table for further exploration. In Exhibit 8, the probability formulas have been saved to the data table. Rows 20 through 29 of the data table are displayed.

Exhibit 8 Lost Sales Data Table with Predicted Probabilities



	Quote	Time to Delivery	Part Type	Status	Lin[Lost]	Prob[Lost]	Prob[Won]	Most Likely Status
20	1862	94	AM	Lost	0.9858684385	0.7282710896	0.2717289104	Lost
21	5912	17	OE	Lost	0.1207595644	0.5301532566	0.4698467434	Lost
22	470	26	OE	Lost	0.1810211463	0.5451321106	0.4548678894	Lost
23	752	36	AM	Lost	-0.100956136	0.4747823809	0.5252176191	Won
24	1335	23	OE	Lost	0.1426133923	0.5355930425	0.4644069575	Lost
25	1198	31	AM	Lost	-0.184190525	0.4540821135	0.5459178865	Won
26	1915	3	AM	Won	-0.684677803	0.3352180664	0.6647819336	Won
27	1616	32	AM	Won	-0.157752089	0.4606435616	0.5393564384	Won
28	1394	2	AM	Lost	-0.713104666	0.3289131858	0.6710868142	Won
29	2148	22	AM	Lost	-0.331170711	0.4179557983	0.5820442017	Won

(Under the top red triangle, select Save Probability Formula.)

The predicted probability that a particular order was lost or won is displayed under the columns Prob[Lost] and Prob[Won]. The Most Likely Status column indicates the predicted Status, Lost or Won. For example, the model estimates that, for the quote in the 20th row, the probability of losing the order was 0.7283. Because this is higher than the probability of not losing the order (0.2717), the model predicts that it is most likely that the order was lost.

The Most Likely Status column can be used to explore the model misclassifications. Specifically, we might compare the outcome *predicted* by the model (Most Likely Status**)** to the *actual* outcome (Status**).** For the quote in the 20th row, for example, we see that the order was correctly classified; the order was lost, and the model correctly predicted it would be lost. The quote in the 23rd row, however, was not correctly classified.

A cross-tabulation, or *confusion matrix*, can be used to compare the actual Status versus the predicted Status for the entire data set. In Exhibit 9 we see the confusion matrix for all 550 quotes. For 136 of the quotes, the model predicted that an order would be lost and indeed it was; 193 times, the model estimated that an order would be won and it actually was won. Thus, the model correctly classified the actual outcomes 329 (136 + 193) times out of the 550 quotes in the data set.

However, the model *incorrectly* predicted 221 out of the 550 quotes. Specifically, 85 times the model predicted that an order would be lost when it was not, and 136 times the model estimated that an order would be won when it was actually lost. In both of these instances, the model misclassified the actual outcome. Thus, the model's *misclassification rate* is 221/550 = 40.2%. (This number is reported as Misclassification Rate under the Whole Model Test in Exhibit 6).

Exhibit 9 Confusion Matrix

Confusion Matrix

Actual		Predicted
Training	**Lost**	**Won**
Lost	136	136
Won	85	193

(Return to the Fit Model analysis window. Under the top red triangle, select Confusion Matrix.)

Summary

Statistical Insights

Logistic regression is a powerful statistical technique that can be used when a response variable is categorical. In this case, we have a *nominal* response variable (i.e., whether an order was placed or not). Logistic regression can also be used for *multinominal* response variables (with more than two outcomes) or for *ordinal* response variables.

Logistic regression differs from standard regression in that a predicted value from a logistic model is an estimated *probability* of something occurring—here, the probability that an order was received.

Like multiple regression, multiple logistic regression can serve different purposes. In this case, logistic regression was used to determine which variables are important and to find out which variables are not as critical in light of other variables. Logistic regression can also be used for predictive purposes – to forecast future outcomes or events for planning purposes.

Managerial Implications

This company should ascertain whether it can shorten delivery times, since the higher the quoted time to deliver the higher the probability that the order will be lost. Interestingly, the dollar amount of the quote does not influence the probability of losing an order. So, contrary to belief, pricing does not appear to be an issue. Thus, the company may wish to offer shorter delivery times as a premium service for its quoting purposes. All else the same, a higher percentage of original equipment orders are received than aftermarket. Management may want to explore the differences between these two markets.

The evidence suggests that other important factors might be influencing the winning or losing of orders, since the model predicts wrong about 40% of the time. Management might wish to determine whether there are other factors that might also be influencing lost sales opportunities.

JMP Features and Hints

In this case we used:

- The Distribution platform to understand the breakdown of the response variable,
- The Fit Y by X platform to explore the relationship between order Status and each predictor variable.
- The Fit Model platform for multiple logistic regression, and the Prediction Profiler to dynamically explore the logistic model and predicted probabilities, and
- The Confusion Matrix to assess misclassification rates.
- We also saved the probability formula to the data table. Note that a cross-tabulation (confusion matrix) can be created for the actual versus the most likely outcome using the Fit Y by X platform.

Exercise

Exercise 1

The model in Exhibit 6 includes Quote as a predictor variable. Create a logistic regression model without Quote. Then, build a second model with just Time to Delivery. Compare these models to the model involving all three predictors based on misclassification rates.

1. Are the misclassification rates for these models higher, lower, or about the same? Given the significance of the terms in the model, is this what you'd expect?
2. Based on misclassification rates, which model does the best job of predicting Status?
3. How could this model be improved? For example, are there any variables missing from this data set that might improve the model's predictive ability (i.e., lower the misclassification rate)?

Exercise 2

Open the data table Wine Quality BC.jmp (download from the Business Case Website). This data table contains information on quality ratings for 6,497 different wines, along with measures of wine properties.

1. Use Fit Y by X to explore the relationship between Quality 2 (Y, Response) and alcohol and density (X, Factors). Note that Quality 2 is an ordinal response, so ordinal logistic regression will be used by default for the analysis.
   a. Is alcohol (content) a significant predictor of Quality 2?
   b. Interpret the graph: In general, what happens to the probability that wine will be Good as the alcohol content increases? In other words, does wine with a higher alcohol content generally rate better or worse than wine with a lower alcohol content?
   c. Is density a significant predictor of wine quality? What do we learn from the graph in terms of the density level for Good wine?

2. Fit a multiple logistic regression model with Quality 2 as the response variable and all of the predictors (Fixed Acidity through Color).
   a. Which predictors are significant?
   b. Use the prediction profiler to explore the model. Change values of the predictors and note how the predicted probabilities of the Quality 2 outcomes change.
   c. In general, what are the characteristics of the Good quality wine? For example, does Good wine generally have lower or higher levels of residual sugar? Density? Chlorides?