Binary Logistic Regression – What, When, and How

JMP Discovery Conference 2016
Susan Walsh – SAS Institute

Abstract

Analysts in many application areas often have a response variable with only two possible levels, of which one is the desired outcome. Binary logistic regression will allow the analyst to predict the probability of the desired outcome, determine which input variables are most closely associated with that outcome, and produce odds ratios which provide a measure of the effect on the outcome. This paper provides an introduction to this type of analysis using binary logistic regression in the Fit Y by X and Fit Model platforms of JMP. It discusses the interpretation of the results including p-values, odds ratios, graphical displays and goodness of fit statistics.

Introduction

Just as linear regression can be appropriate when the response variable is continuous, logistic regression is often appropriate when the response variable is categorical. The focus of the analysis is to predict the probability of the levels of the categorical response (or outcome). Binary logistic regression is for the specific case when the response variable has only two possible values: yes or no, good or bad, 0 or 1. Generally, one of the two levels of the response is considered the level of interest.

This type of model has uses in almost any application area. For example, some questions that might be addressed with a binary logistic regression are:

- Will the consumer purchase my product?
- Will the student graduate in 4 years?
- Will the individual respond to my letter or call?
- Will the borrower default on the loan?
- Will my wireless user switch to another wireless provider?
- Will the customer be satisfied with the customer support service?

The predicted values from a logistic regression are probabilities. As such, they must be between zero and one. Because of these boundaries, a linear regression is not appropriate. The relationship between the probability of a particular level of the response and the predictor variable(s) is often best represented by an S shaped curve as shown in Figure 1.

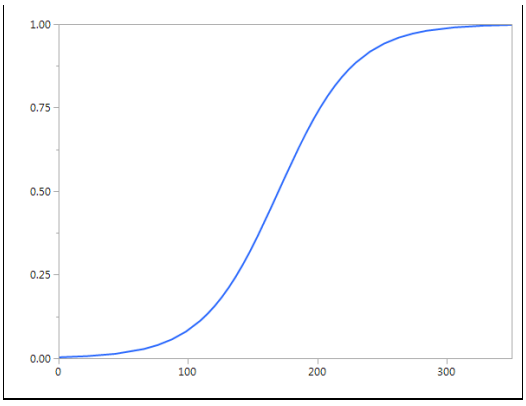

Figure 1

Logistic regression is performed using a logit transformation of the response. The assumption is that the logit transformation of the probabilities results in a linear relationship with the predictor variable(s). The simplest form of the model is then:

$$\log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 X_1 + \varepsilon_i$$

Therefore, in the case of this logistic regression model, the parameter estimates are related to the logit of the probability of the response variable. In that sense, while the significance of the parameters is of interest in determining which predictor variables are important, the actual values of the parameters are not generally of interest.

Odds and Odds Ratios

Rather than focusing on the values of the parameter estimates, focus for a logistic regression is often on odds and odds ratios. You often hear of odds in relation to horse racing; for example, the favorite is 3:2. To see how these odds are constructed (in a mathematical sense), consider two horses in a field of 6 or 8. Horse A has a 60% chance of winning the race. Stated another way, the probability of horse A winning the race is 0.6. Suppose, also, the probability of horse B winning the race is 0.2. The odds for each of these horses are calculated as the probability of winning divided by the probability of not winning as shown in Table 1

Horse	Probability of Winning	Odds of Winning
A	0.6	0.6 / (1 – 0.6) or 3/2
B	0.2	0.2 / (1 – 0.2) or 1/4

Table 1

Odds ratios, then, are the odds of one outcome (horse A winning) compared to the odds of a different outcome (horse B winning). In this example, the odds ratio comparing horse A to horse B is 1.5 divided by 0.25 or 6. Stated another way, the odds of horse A winning are 6 times the odds of horse B winning.

Since odds and odds ratios are constructed from probabilities, they can never be negative. Assuming we are comparing A to B, an odds ratio less than one is an indication that the odds for A are smaller than the odds for B. An odds ratio equal to 1 is an indication that the odds for A and B are not different. An odds ratio greater than 1 is an indication that the odds for A are larger than the odds for B, as in our horse example.

Sensitivity and Specificity

Measures often used to evaluate the worth of a logistic regression model are sensitivity and specificity. These measures have their roots in the health care arena. Consider a medical test that is used to determine if a user has a particular disease. Sensitivity is the ability of the test to correctly identify a patient with the disease. Specificity is the ability of the test to correctly identify a patient without the disease. A health care provider would like both of these measures to be high. However, in the real world, if a test has high sensitivity then what generally happens is some patients without the disease will be identified as having the disease. In other words, the specificity will suffer. Conversely, if the test has high specificity, there is a good chance that some patients with the disease will slip through undetected; the sensitivity suffers.

ROC curves provide a way to visualize sensitivity and specificity. An ROC curve actually plots the sensitivity versus (1 – specificity) across cut off values ranging from zero to 1. The cut-off value is applied to the predicted probabilities of the logistic regression and is the value above which an observation will be predicted to be a positive result. Based on this cut-off value, sensitivity and (1 – specificity) are graphed.

As this process is done for a series of cut-off values, the graph generally resembles a bow as shown in Figure 2.

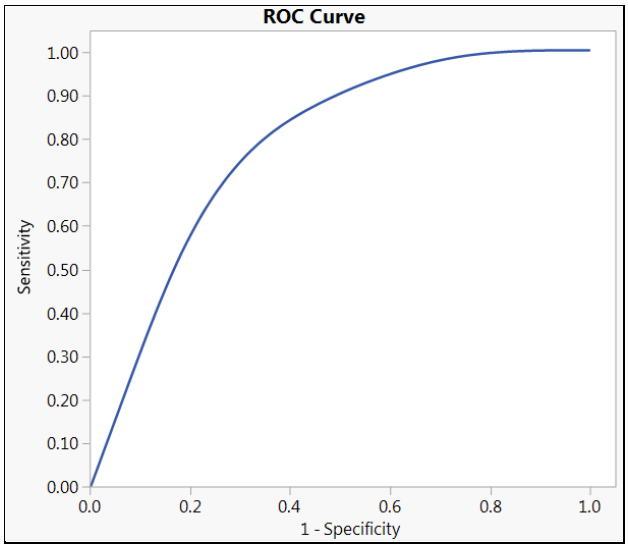

Figure 2

A model that does not predict well and is no better than flipping a coin to decide the value of the response variable would have an ROC curve along the diagonal from (0,0) to (1,1), as shown by the addition of the green dashed line in Figure 3.

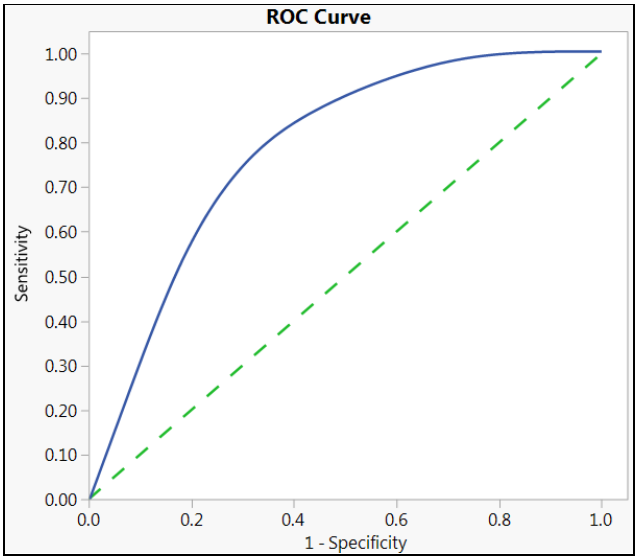

Figure 3

The faster the curve rises (without moving to the right), the better the model. A perfect model would have an ROC curve that goes from (0,0) to (0,1) to (1,1). In other words, whatever your choice of cut off value, all observations would be predicted perfectly giving a sensitivity of 1 before the specificity decreases.

A statistic often used with ROC curves is the area under the curve (or AUC). The area under the green dashed line is 0.5 and the total graphing area is 1, so the AUC should be between 0.5 and 1. The closer the AUC is to 1, the better the model.

Other Fit Statistics

Other statistics used to compare logistic regression models include entropy and generalized Rsquare, root mean square error (RMSE), mean absolute deviation, and misclassification rate. Which of these statistics is used for comparison is generally a matter of analyst preference.

Entropy RSquare first considers the difference between the negative log-likelihood for the reduced model and the negative log-likelihood for the full model. The ratio of this difference to the negative log-likelihood for the reduced model is then calculated. The value of Entropy RSquare ranges from zero for a model that does not improve over a model with just the intercept to one for a perfectly fitting model. The generalized RSquare is also based on a ratio between the likelihoods and is scaled to have a maximum value of one. "The Generalized RSquare measure simplifies to the traditional RSquare for continuous normal responses in the standard least squares setting." [1]

The root mean square error is calculated using the difference between the actual response of the observation and the predicted probability of that actual response. These differences are squared, summed, and divided by the sample size, after which the square root is taken. Smaller values indicate a better model fit. Rather than using a sum of squares, the mean absolute deviation sums the absolute values of the differences between the actual response and the predicted probability of that actual response. Again, smaller values indicate a better model fit.

The misclassification rate is the number of observations that are classified incorrectly given a cut off probability of 0.5. That is, each observation is predicted (or classified) to belong to the group for which it has the highest predicted probability. Those observations for which the predicted group is not the same as the actual group are misclassified.

Examples

The examples in this paper are based on data collected by SAS Technical Support indicating the satisfaction of the user with the service provided. The survey_data.jmp data table contains the results of surveys conducted in 2014 and 2015 for 13,289 different technical support tracks. The data table has only four of the original 95 columns. Those columns are shown in Table 2.

Column Name	Data Type	Modeling Type	Description
Satisfied	Numeric	Nominal	0 = No; 1 = Yes Value labels are used
Met All Response Goals	Numeric	Nominal	0 = No; 1 = Yes Value labels are used
Days to Resolution	Numeric	Continuous	Number of work days from opening of the track to closing
Grouped Days to Resolution	Character	Ordinal	Number of days grouped to 3 levels: 1; 2 to 5, and greater than 5.

Table 2

The goal is to determine if meeting the response goals and length of time to resolution of the question or problem impacts the user's sense of satisfaction.

Simple Logistic Regression – one continuous predictor

To begin, we will fit a model with the days to resolution as the single predictor variable. This model can be fit in the Fit Y by X platform.

1. Select Analyze → Fit Y by X.
2. Assign Satisfied to the Y role.
3. Assign Days to Resolution to the X role. Note you can choose the level of the response (or target) you want to model.
4. Select OK.

To gain a better understanding of the graph presented, color the points by the level of the response variable.

5. Select Rows → Color or Mark by Column
6. Choose Satisfied as the variable to color by, and change the colors if desired.
7. Select OK.

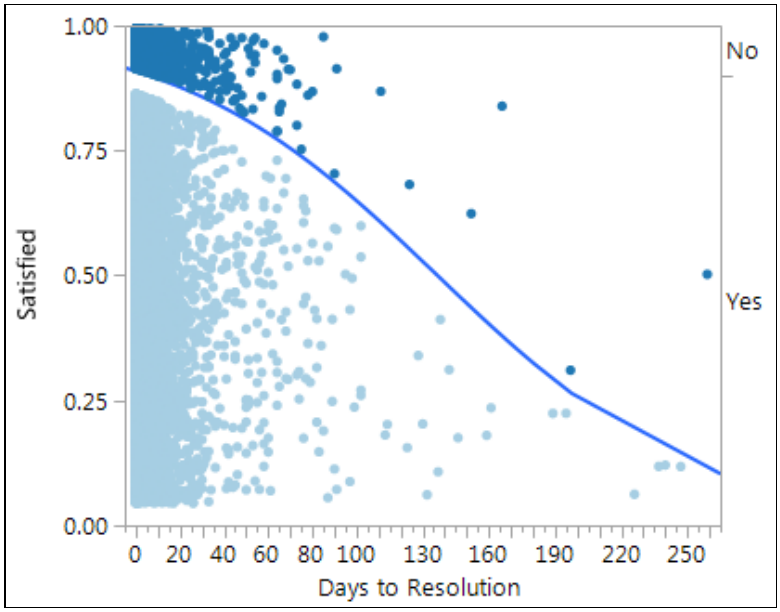

Figure 4

The light points are the customers who were satisfied. The darker points indicate those customers who were not satisfied. The curve is the logistic regression curve that has been fit to the data indicating the relationship between days to resolution and the probability of being satisfied with the assistance received. As one might expect, the longer the issue takes to resolve the lower the probability that the user will be satisfied.

The points on the graph are located from left to right according to the value of Days to Resolution for that point. The points are randomly placed in a vertical position either above or below the model curve. Those points for observations where the user is satisfied are below the curve, while those where the user was not satisfied are above the curve. The orientation of the points above and below the curve is based on the value chosen as the level of the outcome variable being modeled.

The Whole Model test, with a very small p-value, indicates this model is better than the model with only the intercept. That is, it is better than using the overall probability of satisfaction as the predicted probability for all respondents. In the Parameter Estimates table, shown in Figure 5, the estimate of the

slope for Days to Resolution is negative. This indicates that as the Days to Resolution increases, the probability of satisfaction decreases.

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	2.30200706	0.0345919	4428.6	<.0001*
Days to Resolution	-0.0168328	0.0018627	81.66	<.0001*

For log odds of Yes/No

Figure 5

To display the odds ratios for Days to Resolution, click on the red triangle at the top of the output and select Odds Ratios. The unit odds ratio and range odds ratio are added to the Parameter Estimates table as shown in Figure 6.

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq	Unit Odds Ratio	Odds Ratio
Intercept	2.30200706	0.0345919	4428.6	<.0001*	.	.
Days to Resolution	-0.0168328	0.0018627	81.66	<.0001*	0.9833081	0.01278235

For log odds of Yes/No

Figure 6

Note that the odds ratios are less than 1, indicating a decrease in the odds of satisfaction with an increase in the number of days to resolution. The unit odds ratio shows that for a 1 day increase in the number of days to resolution, the odds of being satisfied with the Technical Support assistance decrease by two percent. Across the entire range of the variable (259 days) the odds of being satisfied decrease by 98 percent.

To display the ROC curve, click on the red triangle and select ROC Curve. Choose Yes as the positive level, and then select OK.

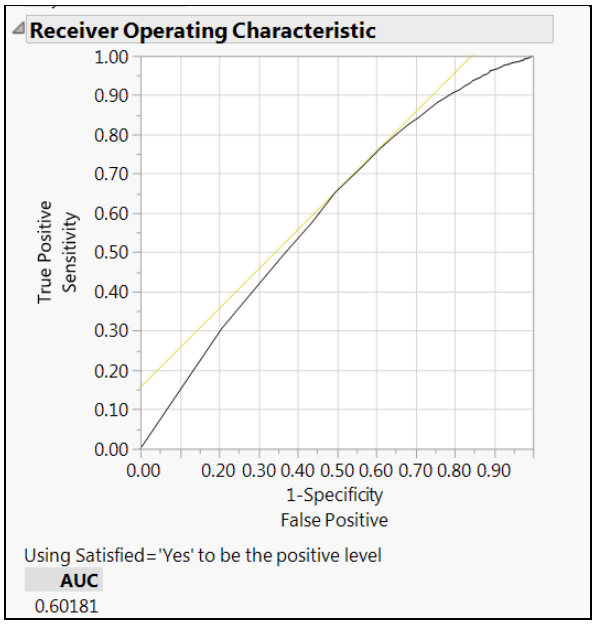

Figure 7

If desired, you can add the diagonal reference line to the graph.

1. Right-click on the graph and select Customize
2. Press the + button to add a script.
3. In the Properties field, add a script such as the following:

```
Pen Color("blue");
Pen Size(2);
Y Function(x, x);
```

4. Select Apply, then OK.

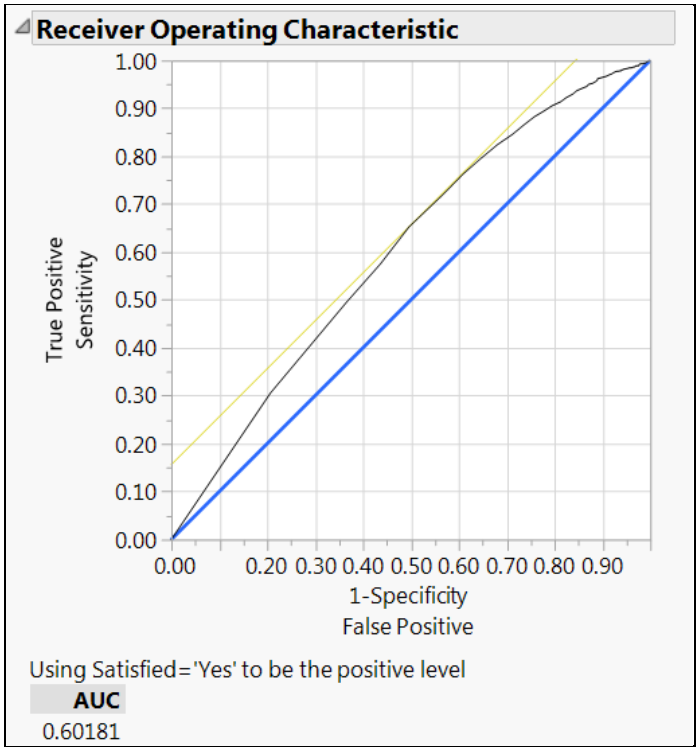

Figure 8

As you can see from the ROC Curve, the curve itself is not very far from the reference line. This is confirmed by the area under the curve (AUC) statistic of 0.60. While the model is better than no model at all, it is not much better.

You may also notice a light line drawn above the ROC curve. This line is a 45 degree line tangent to the point on the ROC curve where the sum of sensitivity and specificity is maximized. Assuming false positives and false negatives have the same cost, this point represents a good choice for the cut-off value.

Additional measures of fit for the model can be found in the Fit Details table in the results.

Simple Logistic Regression – one ordinal predictor

Occasionally, an analyst will opt to group a continuous variable and use it in a model as an ordinal variable. This may be especially true if the continuous variable is highly skewed. The survey data table contains such a variable (Grouped Days to Resolution). To fit a binary logistic regression model using a categorical variable as the predictor variable, the Fit Model platform must be used. If you use the Fit Y by X platform with categorical variables for both the X and Y role, a contingency table analysis will be performed rather than a logistic regression.

1. Select Analyze → Fit Model,
2. Assign Satisfied to the Y role.
3. Select Grouped Days to Resolution, and then select Add.
4. Make sure the Target Level is set to 1, to model the Yes level of the response variable.
5. Select Run.

Observe that just under the Effect Summary outline node, there is a note that the model converged. It is important to check to be sure of convergence before examining any of the other output provided. The whole model test indicates this model is significant and better than using the model with only the intercept term.

The parameter estimates table shows two parameters for the Grouped Days to Resolution variable. This variable has three levels and is therefore represented by two indicator variables. You can find information on the way JMP codes nominal and ordinal variables in the JMP Fitting Linear Models book. [1] Both of the model effects are significant and the parameter estimates are both negative. This is an indication that as you move from shorter to longer resolution times, the probability of a user being satisfied decreases.

To view the Odds Ratios for this model, click on the red triangle and select Odds Ratios. Notice the odds ratios are reported for both directions of comparison for your convenience.

Odds Ratios

For Satisfied odds of Yes versus No

Odds Ratios for Grouped Days to Resolution

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
between 2 and 5	1	0.8107955	0.0103*	0.6908127	0.9516172
more than 5	1	0.4477587	<.0001*	0.3879539	0.5167827
more than 5	between 2 and 5	0.5522462	<.0001*	0.4701308	0.6487043
1	between 2 and 5	1.2333567	0.0103*	1.0508427	1.4475704
1	more than 5	2.2333457	<.0001*	1.9350494	2.5776256
between 2 and 5	more than 5	1.8107865	<.0001*	1.5415345	2.1270675

Normal approximations used for ratio confidence limits

effects: Grouped Days to Resolution

Tests and confidence intervals on odds ratios are Wald based.

Figure 9

Looking at the odds ratios where Level1 is shorter than Level2:

- The odds of being satisfied if the question is resolved in one day is 1.23 times higher than if the question is resolved in between 2 and 5 days.
- The odds of being satisfied if the question is resolved in one day is 2.23 times higher than if the question takes more than 5 days to resolve.

Note that none of the confidence intervals around the odds ratios include one. Therefore, all of the odds ratios are significantly different than one. Select ROC curve from the red triangle menu. Choose Yes as the positive level, and then select OK.

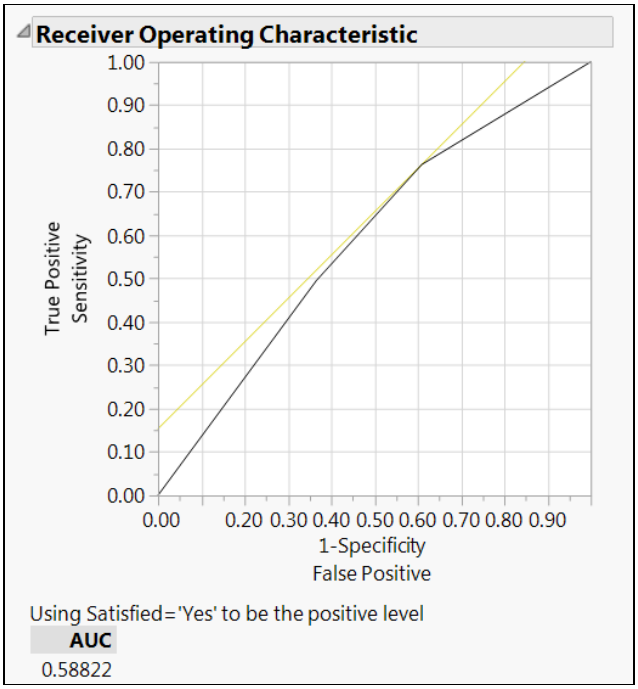

Figure 10

Examine the ROC curve shown in Figure 10. The curve appears to be just slightly above the diagonal line. This is verified by a relatively small AUC of 0.59.

The model fit statistics can be found under the Fit Details outline node. Click on the gray triangle to view them. It may be interesting to compare the model statistics from the two models built up to this point. These are shown in Table 3.

Statistic	Model with Continuous Predictor	Model with Categorical Predictor
AUC	0.602	0.588
Generalized RSquare	0.015	0.023
RMSE	0.300	0.299
Misclassification Rate	0.101	0.101

Table 3

The statistics for the two models are quite close, indicating there is not much difference between the two.

Multiple Logistic Regression – two predictors

Often more than one predictor variable affects the outcome. In that case, a multiple logistic regression model can be fit using the Fit Model platform. Consider, for example, the satisfaction of customers when both the number of days to resolution and whether or not all of the Technical Support response goals were met are in the model.

1. Select Analyze → Fit Model
2. Assign Satisfied to the Y role.
3. Select both Met All Response Goals and Days to Resolution.
4. Click Add to add both predictor variables to the model.
5. Select Run.

Again, we see convergence was achieved for this model. Additionally, the Whole Model Test indicates that the model has some value in predicting the probability of a user being satisfied with the support they received. However, in this case, the Lack of Fit test is also significant. This indicates that a more complex model might better fit this data. For example, perhaps additional variables that have not yet been considered or higher order terms are needed to adequately fit the data. Alternatively, perhaps the skewness in the continuous variable is causing issues. Refit the model using the grouped version of Days to Resolution.

1. Click on the red triangle in the results window and select Model Dialog.
2. In the Model Effects panel, select Days to Resolution and choose Remove.
3. Add Grouped Days to Resolution as a Model Effect.
4. Select Run.

Examining the output, the whole model test is significant and both model effects are significant. The Lack of Fit test shows a much larger p-value. One cannot conclude a significant lack of fit for this model. The parameter estimates for the Grouped Days to Resolution are negative, indicating the longer it takes to resolve the question the lower the predicted probability of the user being satisfied. The parameter estimate for meeting response goals is positive. This indicates if the technical support consultant responds to the customer within the times set by our policy, the predicted probability of the user being satisfied will be higher than if the response goals are not met.

Add the odds ratios to the output from the red triangle menu.

Odds Ratios

For Satisfied odds of Yes versus No

Odds Ratios for Met All Response Goals

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
No	Yes	0.7268987	0.0008*	0.6027776	0.8765784
Yes	No	1.3757074	0.0008*	1.1407993	1.6589867

Odds Ratios for Grouped Days to Resolution

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
between 2 and 5	1	0.7926737	0.0072*	0.6691649	0.9389788
more than 5	1	0.4659535	<.0001*	0.3959184	0.5483774
more than 5	between 2 and 5	0.5878251	<.0001*	0.4975426	0.69449
1	between 2 and 5	1.2615531	0.0072*	1.0649867	1.4944
1	more than 5	2.1461368	<.0001*	1.8235618	2.5257731
between 2 and 5	more than 5	1.7011863	<.0001*	1.4399055	2.0098783

Normal approximations used for ratio confidence limits
effects: Met All Response Goals Grouped Days to Resolution
Tests and confidence intervals on odds ratios are Wald based.

Figure 11

Examining the odds ratios for meeting response goals, the odds of a customer being satisfied are 1.38 times higher when all response goals are met than when all response goals are not met. The odds ratios for Grouped Days to Resolution are similar to those from the simpler model. The ROC Curve and Area Under the Curve for this model do not indicate much of an improvement over the simpler model.

Table 4 shows fit statistics for all three models.

Statistic	Model with Continuous Predictor	Model with Categorical Predictor	Model with two Predictors
AUC	0.602	0.588	0.602
Generalized RSquare	0.015	0.023	0.028
RMSE	0.300	0.299	0.299
Misclassification Rate	0.101	0.101	0.101

Table 4

While all of the models fit to the data are better than no model at all, none of them are clearly superior to the others. Perhaps the model that makes the most sense from the point of view of the domain expert would be the best model to use. Or, perhaps there is another variable that, if considered for inclusion, would make a marked improvement in the prediction of satisfaction.

Conclusion

Binary logistic regression is appropriate to use when attempting to predict the probability of an event occurring or not occurring. That is, it is appropriate when the response variable is a two level categorical variable and the interest is in predicting the probability that one of the two levels will occur.

When examining the results, focus is often on:

- model convergence
- lack of fit test
- overall model significance
- odds ratios
- ROC curve
- fit statistics

The resulting models can then be used to predict the probability of an event occurring (or not occurring) for new data and to understand the relationships between the predictors and the outcome variable.

[1] SAS Institute Inc. 2016. *JMP® 13 Fitting Linear Models*. Cary, NC: SAS Institute Inc.