Categorical Data in a Designed Experiment Part 3: Logistic Regression

Authored by: Cory Natoli
Sarah Burke, PhD
Steve Oimoen, PhD
25 August 2020

The goal of the STAT COE is to assist in developing rigorous, defensible test strategies to more effectively quantify and characterize system performance and provide information that reduces risk. This and other COE products are available at www.afit.edu/STAT.

STAT Center of Excellence
2950 Hobson Way – Wright-Patterson AFB, OH 45433

Table of Contents

Introduction 3
Background 3
Categorical Variables 3
Binomial Distribution 4
Odds 4
Missile Example 4
Analysis 6
Generalized Linear Models (GLM) 6
Components – Logistic Regression 7
Logistic Regression 7
Model 7
Missile Example Revisited 8
Interpreting Logistic Regression Models Using the Odds Ratio 10
Assessing Logistic Regression Model 11
Confusion Matrix 11
Sensitivity and Specificity 12
Receiver Operating Characteristic (ROC) Curve 13
Coefficient of Determination – $\boldsymbol{R2}$ 14
Residuals on Model Fit 14
Considerations 15
Factor Assumptions 15
Separation 15
Fitting the Model 16
Extensions of Logistic Regression 17
Conclusion 17
References 18
Appendix A – JMP Tutorial – Missile Data 19
Curve Representation of Model of Probability of Hitting the Target 19
JMP Prediction Profiler 21
JMP Unit Odds Ratio 21
Confusion Matrix 21

ROC Curve 21

Introduction

Part 1 of this best practice series discussed how the response variable data type can influence an experiment's size and the associated analysis. Part 2 presented three methods typically used to approximate the required sample size when using a binary response variable in a designed experiment. Part 3 shows how generalized linear models (GLM), specifically logistic regression, can be used to analyze binary responses when there are one or more factors in the test.

We begin by providing some background information on categorical variables with a focus on binary responses, the binomial distribution, odds ratio, and linear regression assumptions. Then, we show an example having a binary response variable to explain why linear regression is not an appropriate analysis tool and the need to use logistic regression. Next, we detail the three components of the GLM and provide an interpretation of logistic regression. Then, we assess the logistic regression model and consider issues such as factor assumptions, separation, and fitting the model. We conclude with other types of logistic regression.

Keywords: generalized linear model, logistic regression, linear regression, binary responses, categorical factors

Background

We begin with categorical variables, specifically binary responses which only have two possible outcomes. We introduce the binomial distribution and then discuss how the odds ratio measures the association between a binary response and a factor based on the odds. We conclude this section by identifying the assumptions associated with linear regression and include an example showing why linear regression is not the analysis tool to use with binary responses.

Categorical Variables

Qualitative variables, also termed categorical, are non-numeric variable types. Categorical data falls into three data classifications:

- Binary: can only take on 2 possible values (e.g., pass/fail or yes/no)
- Ordinal: has an ordered scale; order of listing categories does matter (e.g., small/medium/large or disagree/neutral/agree)
- Nominal: there is no ordered scale; order of listing categories is irrelevant (e.g., aircraft type or color)

For this best practice, we concentrate on binary response variables. Examples of binary responses include whether a missile hits or misses a target, a sent message is received or not received, an item is either defective or not defective, or a mission is either a success or a failure.

Binomial Distribution

The binomial distribution is a discrete probability distribution defined by the probability of a success (p) and sample size (n). The binomial distribution is based on the following assumptions:

- Number of trials/runs (i.e. sample size, n) is a fixed value
- Only two possible (binary) outcomes, labelled as "success" or "failure"
- Probability of the outcome "success" is constant across the n trials
- Trials are independent; outcome of one trial does not affect the outcome of any other trial

The mean and variance of a binomial distribution are both defined by n and p:

$$mean = np$$

$$variance = np(1-p)$$

For more information on the binomial distribution, see Sigler (2018).

Odds

We often interpret continuous variables using summary statistics such as the mean and standard deviation. For binary responses, we often analyze the estimate proportion of success, the odds, and the odds ratio. We can determine odds using the estimated probabilities of success and failure. If we let p = probability of success, then 1 - p = probability of a failure and the odds of success is defined as the ratio of the probability of success and the probability of failure:

$$odds\ (succcess) = \frac{p}{1-p} \text{ and } odds\ (failure) = \frac{1-p}{p}$$

We can interpret odds as the likelihood that some event will occur. We discuss this as a proportion of the likelihood that it will occur by the likelihood that it will not occur. We discuss odds ratios and their interpretation later in this best practice.

Missile Example

An engineer studied the effect of air speed, in knots, with respect to the ability of a surface-to-air missile to hit the designated target (data source: Montgomery et al., 2012). The result of each test is measured as either a hit (response variable y = 1) or a miss (response variable y = 0). Table 1 shows the collected data.

Table 1: Missile Data

Test	Target Speed (knots)	y		Test	Target Speed (knots)	y
1	400	0		14	330	1
2	220	1		15	280	1
3	490	0		16	210	1
4	210	1		17	300	1
5	500	0		18	470	1
6	270	0		19	230	0
7	200	1		20	430	0
8	470	0		21	460	0
9	480	0		22	220	1
10	310	1		23	250	1
11	240	1		24	200	1
12	490	0		25	390	0
13	420	0				

Figure 1 shows a scatterplot to visually represent the missile data. The response $y = 1$ (hit) appears to be associated with slower target speeds. Intuitively, this hypothesis makes sense as a slower moving target should be easier to hit. Note that there are some overlapping results; the missile missed two slow moving targets and hit one fast target (target speed = 470 knots).

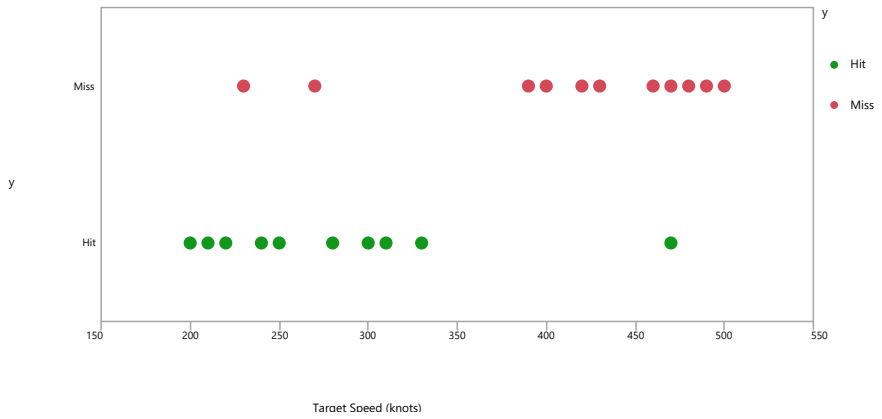


Figure 1: Scatterplot of Missile Data

We now want to create a statistical model to be able to predict the probability of hitting a target, given any target speed. We first incorrectly fit a linear regression model to the data to predict the probability of hitting the target at a given target speed.

Analysis

We first attempt to use linear regression to model the binary response (see Appendix A). Figure 2 shows the parameter estimates.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1.6016871	0.240255	6.67	<.0001*
Target Speed	-0.003193	0.000675	-4.73	<.0001*

Figure 2: Linear Regression Parameter Estimates

The linear regression model is therefore:

$$y = -0.003x + 1.60$$

The variable y represents the response (probability of a hit), and x represents the target speed in knots. At first glance, this model might seem reasonable. The model coefficient for target speed is negative, indicating that as target speed increases, the probability of hitting the target decreases. For example, we could estimate the probability of hitting the target to be 0.48 for a target speed of 350 knots.

However, there are critical issues when trying to use linear regression to model a binary response. First, since the response variable is binary, there are only two possible values for the error term (0 or 1). Therefore, the error terms cannot be normally distributed since the normal distribution is a continuous distribution that can take on an infinite number of values. We have already violated a core assumption of linear regression. In addition, the variance of a binomially distributed variable depends on the probability of success, which changes as a function of the target speed. We have now violated the constant variance assumption. Finally, using linear regression to model the probability of success does not guarantee the predicted values are between 0 and 1 (a requirement for a probability value). For example, for a target that is stationary (x = 0), we would predict that the probability of hitting the target is 1.6. Similarly, for a target moving at 525 knots, we would predict that the probability of hitting the target is -0.07. Neither of these values make any sense in the context of the problem. Instead of applying linear regression to model a binary response, we must use an alternative analysis method such as logistic regression.

Generalized Linear Models (GLM)

Logistic regression is a special case of a family of models called generalized linear models (GLMs). Linear regression is also a special case of a GLM. The GLM consists of three components (Agresti, 2017):

- Random component, which specifies the distribution of the response
- Systematic component, which specifies the factors
- Link function, which connects the random and systematic components together

Components – Logistic Regression

The random component for logistic regression is the binomial distribution, as we are dealing with binary responses. The systematic component specifies the form of the model in terms of the factors. This is a linear predictor used to model the response and takes on the form of the desired model. For example, a model with just main effects and k factors would have the following systematic component:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

This component could also include interactions and quadratic terms. The link function is used to connect the random and systematic components together so that it specifies a function that relates the mean of the response (from the random component) to the linear predictor (the systematic component). For logistic regression, the link function used is the logit function, shown below:

$$Logit\ (p) = ln\left(\frac{p}{1-p}\right)$$

The logit function provides nice properties for the resulting model parameters and interpretation as we discuss in the following section. In essence, we perform a log transformation from odds to log odds. Transforming attempts to remove the restriction of the probability range (zero to one). Essentially, we are mapping the probability range between zero and one to log odds ranging from negative infinity to positive infinity.

Logistic Regression

The following subsections present characteristics and interpretation of the logistic regression model, revisit the missile example using logistic regression, and discuss the odds ratio with respect to logistic regression.

Model

Logistic regression is a nonlinear model, but contains the linear predictor term (represented by the systematic component). The logistic function is an S-shaped curve whose shape depends on the direction and magnitude of the model parameters. The model predicts p, the probability of success at a given value of factor x. Because this model is nonlinear, the rate of change in p per unit increase in x depends on the value of x. This is in contrast to a linear regression model, where the magnitude of the model parameters is straightforward. Table 2 shows the equations for the log odds, odds of success, and the probability of success for a model with one factor.

Table 2: Logistic Regression Equations

Name	Equation
Log odds	$logit(p(x)) = ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$
Odds of success	$\frac{p(x)}{1-p(x)} = e^{(\beta_0+\beta_1 x)}$
Probability of success	$p(x) = \frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}$

To further illustrate, Figure 3 shows the logistic regression model with one factor and various values of the parameter β_1.

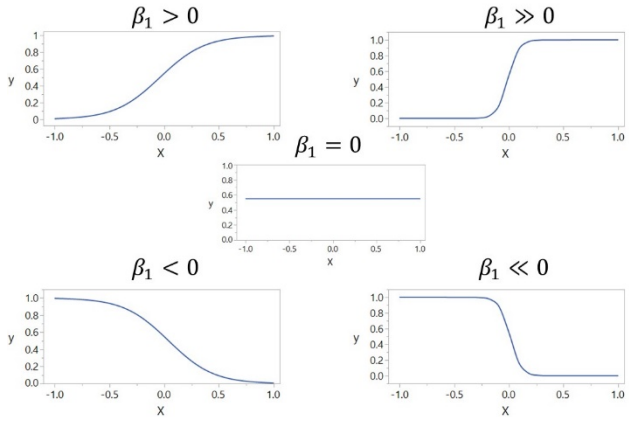


Figure 3: Logistic Regression with One Factor

The logistic regression model shown forms an S-shaped curve to model the effect of a single factor on the probability p. The value of β_1 determines the slope and direction of the curve. When β_1 is greater than zero, then p increases as x increases. As the magnitude of β_1 increases, the curve becomes steeper. When β_1 is equal to zero, then the factor x has no effect on p and the curve becomes a flat line. Note that at the extreme levels of the factor (x is near -1 or near 1), the effect of the factor on the response is not as large. In the end, we use the logistic regression model to predict the probability of success under specified factor levels.

Missile Example Revisited

The scatterplot displayed in Figure 1 showed that slower targets were more likely to be successfully hit. We now model the probability of hitting the target using logistic regression, shown in Figure 4. The curve suggests that targets at a low speed have a higher predicted probability of being hit.

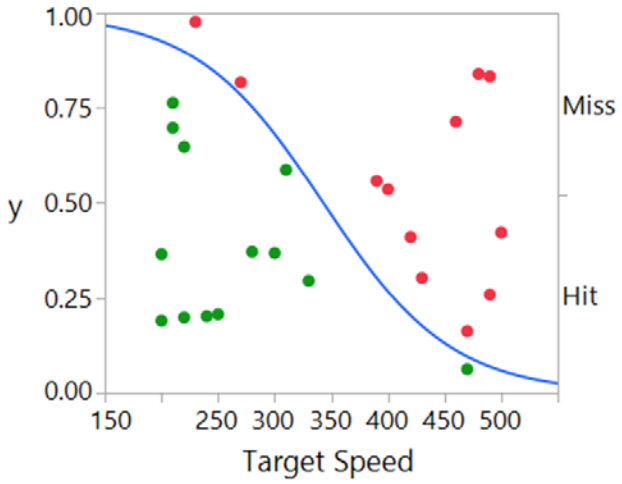


Figure 4: Curve Representation of Logistic Regression Model of Probability of Hitting the Target

Using logistic regression, the model estimates the probability of hitting the target over a range of target speeds. Using JMP, the estimated parameter values are $\hat{\beta}_0 = 6.07$ and $\hat{\beta}_1 = -0.0177$. We can now use these parameter estimates to estimate the probability of hitting a target, given a target speed. Using the probability of success equation from Table 2, we have

$$p(x) = \frac{e^{(6.07-0.0177x)}}{1 + e^{(6.07-0.0177x)}}$$

To calculate the probability of hitting a target at a speed of 350 knots, we have

$$p(x = 350) = \frac{e^{(6.07-0.0177(350))}}{1 + e^{(6.07-0.0177(350))}} = 0.469$$

To simplify matters, we turn to the prediction profiler in JMP to quickly compute the probability.

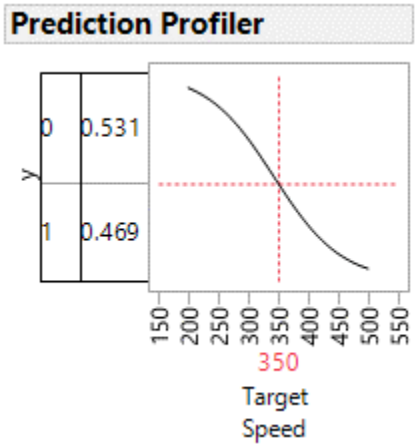


Figure 5: JMP Prediction Profiler for the missile data set

A predicted probability of hitting the target at x = 525 knots provides an estimate of 0.038. Compare this to the estimate using linear regression (-0.07). We now have a valid estimated probability using logistic regression.

Interpreting Logistic Regression Models Using the Odds Ratio

Interpreting the model coefficients of a logistic regression model is less straightforward compared to linear regression. The model parameter β_1 can be interpreted as the difference in the log-odds as factor x changes by one unit. This is not particularly insightful, so we often interpret the model in terms of the odds ratio (OR). Recall we defined odds earlier as the ratio of probability of success to probability of failure. The odds ratio takes odds one step further and is the ratio of two odds. In logistic regression, we compute the odds ratio for a factor at different values.

The odds ratio represents the constant effect a factor has on the likelihood that a specified outcome will happen. Using the logistic regression model, the odds is equal to:

$$odds = \frac{p(x)}{1-p(x)} = e^{(\hat{\beta}_0+\hat{\beta}_1 x)}$$

Therefore, the odds ratio for a one-unit increase in x is:

$$OR = \frac{odds_{x_i+1}}{odds_{x_i}} = e^{\hat{\beta}_1}$$

If β_1 is equal to zero (the factor has no effect on the response), then the odds ratio is equal to 1; i.e., the odds of a success do not change for different values of the factor. For the missile example, using the estimated value of $\hat{\beta}_1 = -0.0177$, the OR = $e^{-0.0177} = 0.982$. For every one-unit increase in the target speed, the odds of success decrease by 1.76% ((1-.982)*100%). You can interpret the odds in terms of missing the target by flipping the ratio. The odds ratio of missing the target to hitting the target is 1/0.982 = 1.018. For a one-unit increase in target speed, the odds of missing the target increase by 1.8%. Using JMP, we show the output in Figure 6.

Unit Odds Ratios

Per unit change in regressor

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
Target Speed	0.982451	0.970821	0.994221	1.0178624

Figure 6: JMP Unit Odds Ratio

Note that the default setting in JMP provides an odds ratio for a one unit increase in x which may or may not provide useful information because of the scaling of the factor. The odds ratio for a d-unit increase in x is:

$$OR = \frac{odds_{x_i+d}}{odds_{x_i}} = e^{\hat{\beta}_1 d}$$

Suppose a 10-knot increase is operationally relevant in the missile example. The odds ratio for a 10-unit increase in the target speed is $e^{-0.0177(10)} = 0.838$. For a 10-unit increase in the target speed, the odds of hitting the target decrease by 16.2%.

Assessing Logistic Regression Model

Next, we discuss several metrics to evaluate the logistic regression model for fit and accuracy.

Confusion Matrix

The confusion matric is a commonly used tool to evaluate how well the logistic regression model fits the data. We can also assess the accuracy of the model to predict the response outcome. The model provides a probability of success $p(x)$. Therefore, we typically use a threshold of $p(x)$, denoted p_0 to get a predicted value $\hat{y}(x)$. The most common threshold value is $p_0 = 0.5$; that is, if $\hat{p}(x)$ for the given factor levels is greater than or equal to 0.5, the predicted response is 1 (a success), otherwise, the predicted response is 0 (a failure). Other thresholds could be used depending on the dataset.

The confusion matrix is a two-way contingency table summarizing the actual responses versus the predicted responses from the model. Ideally, there would be few prediction errors. Table 3 shows a notional confusion matrix.

Table 3: Notional Confusion Matrix

Actual	**Predicted**		**Total**
y	**Success**	**Failure**	
Success	True Positive (TP)	False Negative (FN)	TP + FN
Failure	False Positive (FP)	True Negative (TN)	FP + TN
Total	TP + FP	FN + TN	TP + TN + FP + FN

We let TP, TN, FP, and FN each represent their respective number of responses (counts). The ideal confusion matrix would have zero false negatives and zero false positives so that the model perfectly predicts the response given the data. This will not occur in real life, but the goal is to have a model that minimizes these two errors. The total number of responses n would be the sum of TP, TN, FP, and FN. The confusion matrix produces the following metrics:

Table 4: Confusion Matrix Metrics

Metric	Formula
Accuracy	$\frac{TP+TN}{n}$
True Positive Rate (TPR) aka sensitivity	$\frac{TP}{TP+FN}$
False Positive Rate (FPR)	$\frac{FP}{FP+TN}$
True Negative Rate (TNR) aka specificity	$\frac{TN}{TN+FP}$
False Negative Rate (FNR)	$\frac{FN}{FN+TP}$
Precision	$\frac{TP}{TP+FP}$

Sensitivity and Specificity

Two metrics commonly used to assess the predictive capability of the logistic regression model are sensitivity and specificity. Sensitivity is the true positive rate and specificity is the true negative rate (both shown in Table 3). Sensitivity shows the predicted percentage of true positives while specificity tells the predicted percentage of true negatives. Ideally, these values are as close to one as possible. A value of one indicates there are no errors in the predictions. Depending on the experiment and their respective objectives, maximizing one metric may actually lower the other. We used JMP to create the confusion matrix (Figure 7) for the missile data.

Confusion Matrix

Training

Actual	Predicted Count	
y	1	0
1	12	1
0	2	10

Figure 7: Confusion Matrix – Missile Data

We can now calculate the sensitivity and specificity using the data in Figure 8 and the formulas in Table 3.

$$Sensitivity = \frac{TP}{TP + FN} = \frac{12}{12 + 1} \approx 0.9231$$

$$Specificity = \frac{TN}{TN + FP} = \frac{10}{10 + 2} \approx 0.8333$$

We now conclude the logistic regression model correctly predicted a hit 92.31% when the actual outcome was a hit and correctly predicted a miss 83.33% of the time when the true outcome was a miss. It is important to consider sensitivity and specificity versus just overall accuracy as a model may do very well for predicting successes, but not well for failures.

Receiver Operating Characteristic (ROC) Curve

The receiver operating characteristic (ROC) curve provides additional information and characterizes specificity and sensitivity over the possible threshold values p_0 used to predict the response (choice of p_0 is arbitrary). The plot of the curve has *sensitivity* (y-axis) against *1 – specificity* (x-axis). Ideally, the curve is a step function with (1 – specificity, sensitivity) = (0,1) and (1-specificity, sensitivity) = (1,1). When p_0 is close to zero, almost all predictions are $y = 1$ because we predict a success (y = 1) if $p(x)$ is greater than p_0 If p_0 is close to one, then nearly all predictions would be $y = 0$. ROC curves are beneficial when comparing multiple models. The area under the curve (AUC) is a metric that evaluates how close the model's curve is to this ideal curve. Ideally, the AUC should be close to one as this indicates a good model. If the AUC is approximately 0.5, the model is equivalent to a random guess. The ROC curve and associated AUC value (setting $y = 1$ to be the positive level) for the missile example is shown in Figure 8.

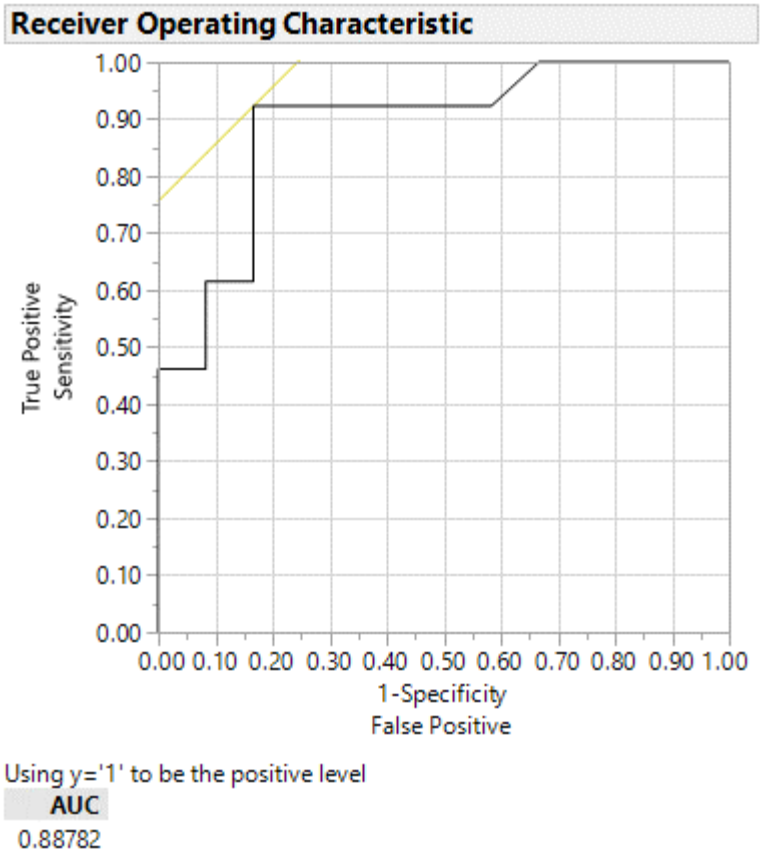


Figure 8: ROC Curve – Missile Data

The model's AUC value is 0.88782, which is close to one, indicating a good model fit.

Coefficient of Determination – R^2

The coefficient of determination, also called R^2, is a common model diagnostic for linear regression. There is a version of this for logistic regression called McFadden's pseudo R^2, which measures the proportion of the total uncertainty attributed to the model fit. A value close to one means that there is little uncertainty in the predicted probabilities. This is an uncommon outcome for logistic regression models, so this value is often low.

Residuals on Model Fit

Model checking is always important, whether using linear or logistic regression. Residuals are the differences between the observed values and the fitted values from the model. As previously stated, with logistic regression the variance is not constant. Therefore, we must make an adjustment since observations have different variances. We list two types of residuals common in logistic regression:

- Deviance residual: based on the deviance or likelihood ratio chi-squared statistic

- Pearson residual: computed difference between observed and fitted values and divided by an estimate of the standard deviation of the observed value

These residuals can be obtained in software and the typical residual analysis can be conducted for logistic regression models. See Burke (2017) for more information on residual analysis.

Considerations

Factor Assumptions

Multicollinearity is still a concern with logistic regression. Factors that are correlated with each other may negatively impact the model fit and the ability to identify which factors actually are statistically significant on the response. Ideally, the independent variables (factors) should be independent. One of the easiest approaches to assess multicollinearity is to review the pairwise scatterplots of the factors. If there are patterns between factors, then we conclude the factors are not independent (Figure 9a). We also include an orthogonal design (Figure 9b) showing uncorrelated estimates of regression coefficients; the estimates do not depend on other factors.

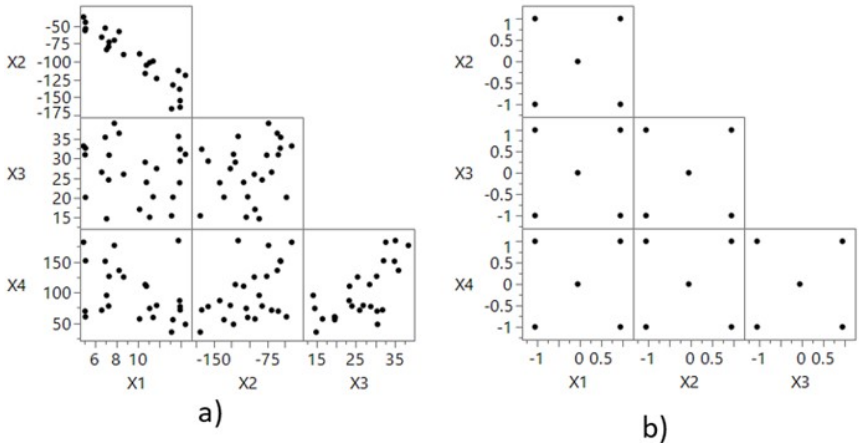


Figure 9: Multicollinearity

Multicollinearity impacts the correlated independent variables (factors), specifically the coefficients and p-values but does not influence the model's predictive capability.

Separation

Separation is a unique issue for binary responses that practitioners should be aware of. Complete separation occurs when a factor or a model term perfectly predicts the response. For example, Figure 10 shows a case where for all factor values of X1 ≤ 4, the response is always y = 0; when X1 > 4, the response is always y = 1. Because there is complete separation, the maximum likelihood estimates (MLEs) of the logistic regression model do not exist because an infinite number of solutions could fit this data.

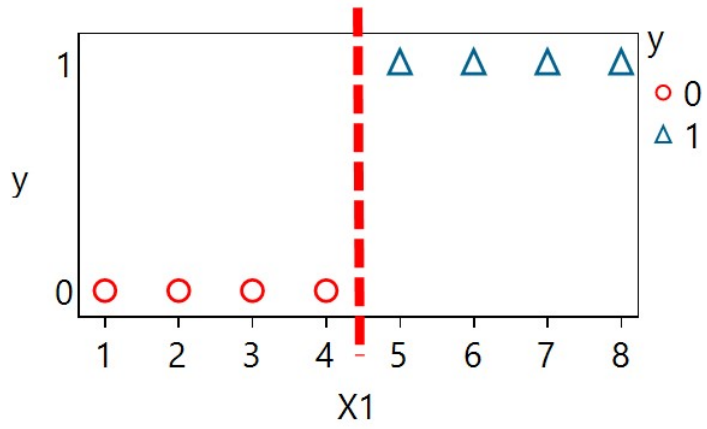


Figure 10: An Example of Complete Separation

Quasi-complete separation is when the break point for separation has both a "success" and "failure." In Figure 10, if an additional value of Y = 0 when X1 = 5 is added, then quasi-complete separation exists. Like with complete separation, MLEs do not exist for quasi-complete separation.

Detecting separation can be difficult, especially with two-factor interactions. When using JMP, look for "Unstable" next to the "Parameter Estimates" column and large values under the "Std Error" column (see Figure 11).

Parameter Estimates

Term		Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	Unstable	-65.478289	8166.8578	0.00	0.9936
X1	Unstable	2.44703409	1608.9752	0.00	0.9988
X2	Unstable	13.0488777	2398.0309	0.00	0.9957
(X1-4.44444)*(X2-3.27778)	Unstable	4.35679605	1645.3722	0.00	0.9979

For log odds of 1/0

Figure 11: Detecting Separation

Separation is likely to occur with small sample sizes, when categorical factors have more than three levels, or if there are a large number of model terms relative to the sample size. If separation exists, try to collect more data (larger sample size) and increase the number of levels for continuous factors. Two-level designs are often not the best choice for binary responses. Another approach is Firth's (1993) penalized method of maximum likelihood which provides bias-reduction for small sample sizes. Finally, you may need to use an alternative modeling technique like decision trees.

Fitting the Model

To ensure a good model fit, we recommend dividing your data into a training set and a test set. To obtain the test set, you can select about 10 to 20 percent of the full data set. You then fit an initial model to the remaining 80-90% of the data. Then use that model to predict the responses of the runs in the test set. This allows you to assess the predictive capability of the model using data that was not used

to fit the model. This is a common model validation technique when the sample size is sufficiently large and can also be used when fitting any other type of statistical model.

Extensions of Logistic Regression

We focused on logistic with one factor; however, we can extend these concepts to multiple logistic regression. Multiple logistic regression accounts for main effects of multiple factors, two-factor interactions, and quadratic effects. The multiple logistic regression model is:

$$E(y_i) = p_i = \frac{e^{(\eta_i)}}{1 + e^{(\eta_i)}} = \frac{1}{1 + e^{(-\eta_i)}}$$

$$where\ \eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$$

The systematic component could also include interaction or quadratic terms.

If the categorical response variable has more than two levels, we use nominal logistic regression or ordinal logistic regression. Nominal logistic regression allows for a nominal categorical response variable with three or more levels (e.g. Countermeasure A, Countermeasure B, Countermeasure C, and Countermeasure D). If there is a natural ordering of the response levels, we choose ordinal logistic regression to account for an ordinal response variable (e.g. Low Threat, Moderate Threat, and High Threat).

Conclusion

We discussed logistic regression as Part 3 of the Categorical Data in Designed Experiment series. We reviewed categorical variables focusing on binary responses, the binomial distribution, odds ratio and assumptions associated with linear regression. We used the missile example, which included a binary response, to show why linear regression is not the appropriate analysis tool to use and to provide the rationale to use logistic regression. We covered the three components of the generalized linear model and discussed methods to assess the logistic regression model, additional considerations, and other types of logistic regression. When dealing with a categorical response variable in your next designed experiment, you now know the appropriate statistical tool to employ.

References

Agresti, Alan. *Categorical Data Analysis*. 3rd ed., John Wiley & Sons, Inc., 2013.

Burke, Sarah. "Model Building Process Part I: Checking Model Assumptions V1.1." Scientific Test and Analysis Techniques Center of Excellence (STAT COE), 24 October 2017.

Firth D. (1993). Bias reduction of maximum likelihood estimates. Biometrika 80, 27–38 10.1093/biomet/80.1.27

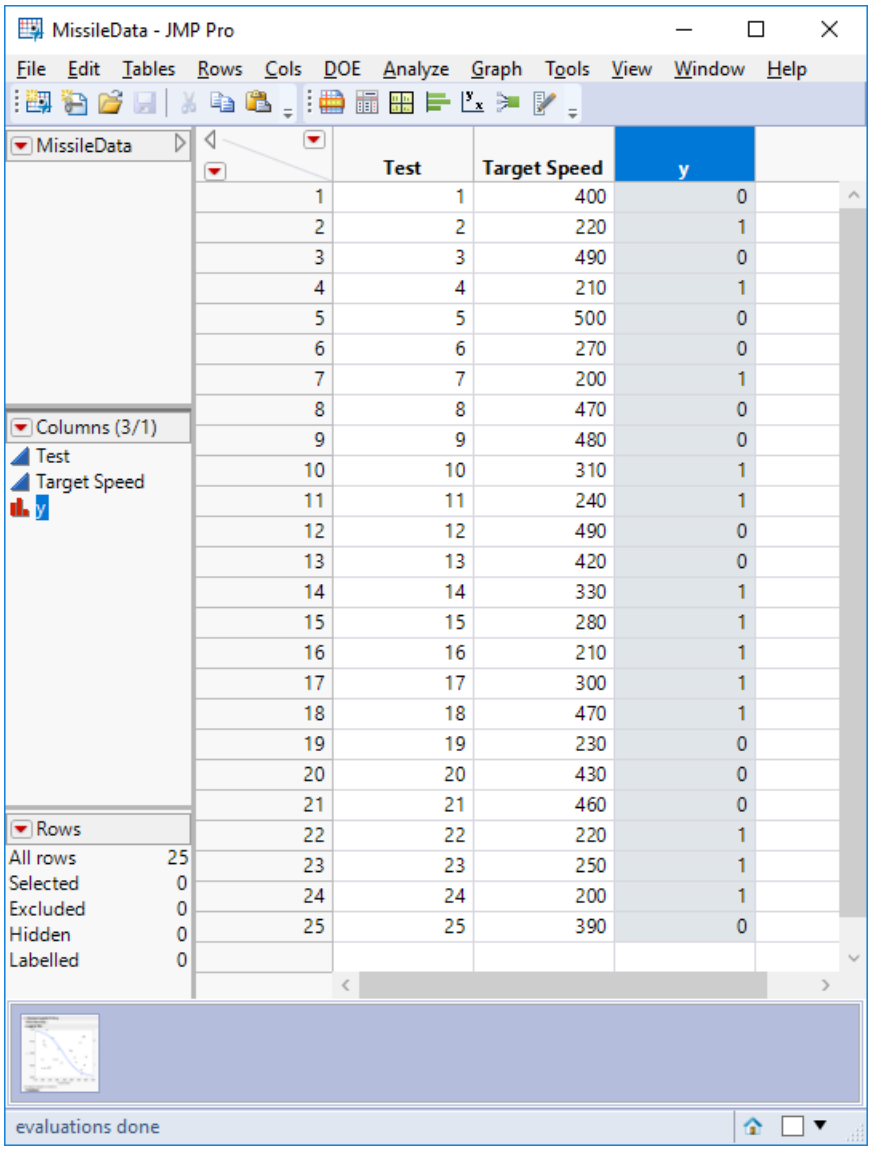
Montgomery, Douglas C., et al. *Introduction to Linear Regression Analysis*. 5th ed., John Wiley & Sons, Inc., 2012.

Sigler, Gina. "Statistics Reference Series Part 2: Probability." Scientific Test and Analysis Techniques Center of Excellence (STAT COE), 31 October 2018.

Appendix A – JMP Tutorial – Missile Data

Curve Representation of Model of Probability of Hitting the Target

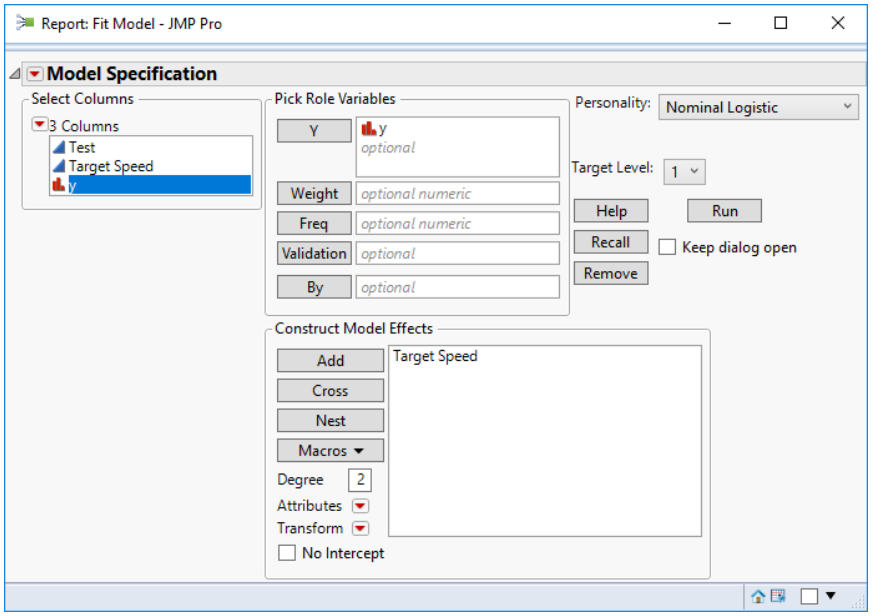
To create the logistic plot, we must first create the data table using the missile data. Notice, the response variable is now a nominal data type, denoted by the red bar chart next to the response name in the left panel.



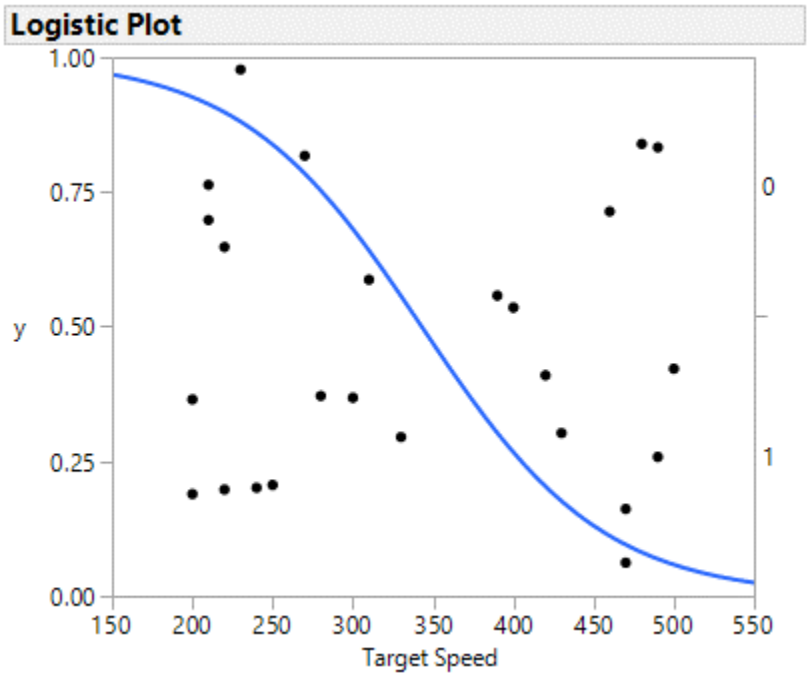
	Test	Target Speed	y
1	1	400	0
2	2	220	1
3	3	490	0
4	4	210	1
5	5	500	0
6	6	270	0
7	7	200	1
8	8	470	0
9	9	480	0
10	10	310	1
11	11	240	1
12	12	490	0
13	13	420	0
14	14	330	1
15	15	280	1
16	16	210	1
17	17	300	1
18	18	470	1
19	19	230	0
20	20	430	0
21	21	460	0
22	22	220	1
23	23	250	1
24	24	200	1
25	25	390	0

The next step is select “Analyze” then “Fit Model”. In this new window, select “y” under “Columns”, then “Y” under “Pick Role Variables”. Then, select “Target Speed”, then “Add” under “Construct Model Effects”. From the “Personality” drop down menu, select “Nominal Logistic”, set the “Target Level” to

"1", and then select "Run". Selecting the target level defines what constitutes a "success"; for this example, a value of 1 ensures the predicted probability is in terms of hitting the target.

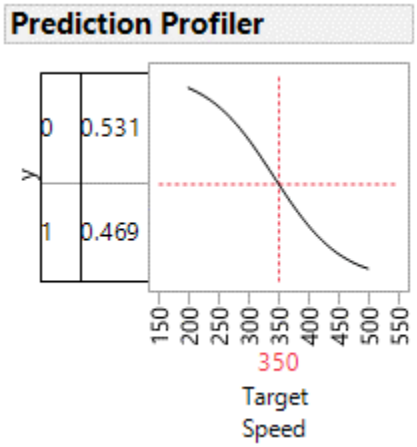


You can now select the "Logistic Plot" to show the output below.



JMP Prediction Profiler

Using the same output from above, select the red triangle to the left of "Nominal Logistic Fit for y", and select "Profiler". The Prediction Profiler will be displayed. Simply double click on the red value above the "Target Speed" to enter the value of 350.



JMP Unit Odds Ratio

Using the same output from above, select the red triangle to the left of "Nominal Logistic Fit for y", and select "Odds Ratio".

Unit Odds Ratios

Per unit change in regressor

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
Target Speed	0.982451	0.970821	0.994221	1.0178624

Confusion Matrix

Using the same output from above, select the red triangle to the left of "Nominal Logistic Fit for y", and select "Confusion Matrix".

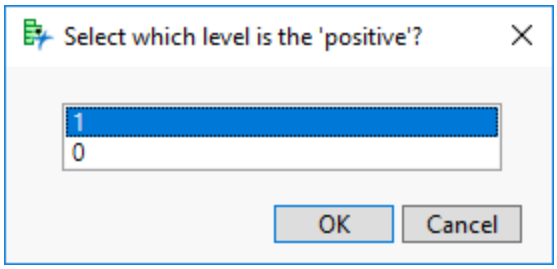
Confusion Matrix

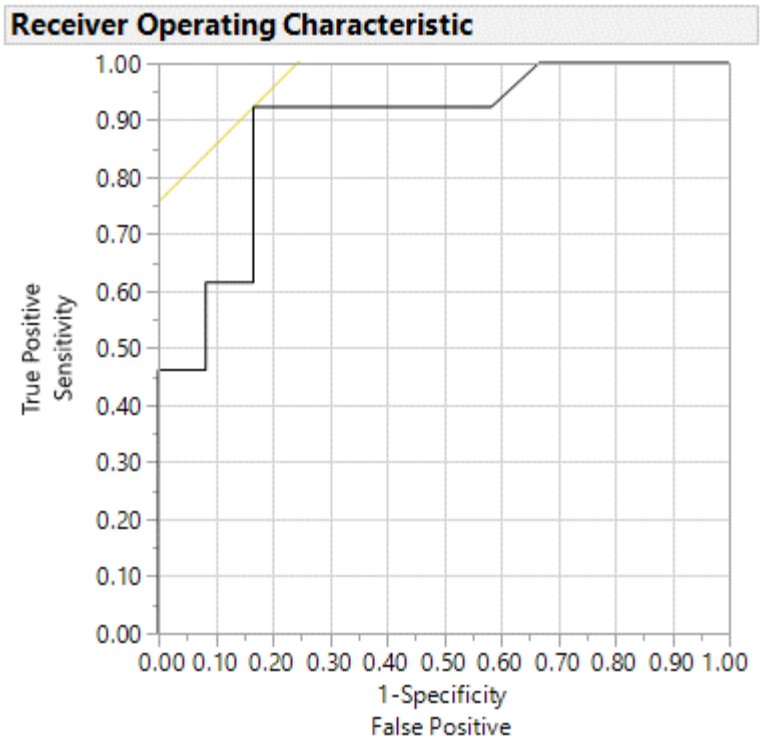
Training

Actual	Predicted Count	
y	1	0
1	12	1
0	2	10

ROC Curve

Using the same output from above, select the red triangle to the left of "Nominal Logistic Fit for y", and select "ROC Curve". Then select "1" as the positive level and then "OK".





Using y='1' to be the positive level

AUC
0.88782