Chapter 321

Logistic Regression

Introduction

Logistic regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name *logistic regression* is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. The name *multinomial logistic regression* is usually reserved for the case when the dependent variable has three or more unique values, such as Married, Single, Divorced, or Widowed. Although the type of data used for the dependent variable is different from that of multiple regression, the practical use of the procedure is similar.

Logistic regression competes with discriminant analysis as a method for analyzing categorical-response variables. Many statisticians feel that logistic regression is more versatile and better suited for modelling most situations than is discriminant analysis. This is because logistic regression does not assume that the independent variables are normally distributed, as discriminant analysis does.

This program computes binary logistic regression and multinomial logistic regression on both numeric and categorical independent variables. It reports on the regression equation as well as the goodness of fit, odds ratios, confidence limits, likelihood, and deviance. It performs a comprehensive residual analysis including diagnostic residual reports and plots. It can perform an independent variable subset selection search, looking for the best regression model with the fewest independent variables. It provides confidence intervals on predicted values and provides ROC curves to help determine the best cutoff point for classification. It allows you to validate your results by automatically classifying rows that are not used during the analysis.

The Logit and Logistic Transformations

In multiple regression, a mathematical model of a set of explanatory variables is used to predict the mean of a continuous dependent variable. In logistic regression, a mathematical model of a set of explanatory variables is used to predict a *logit* transformation of the dependent variable.

Suppose the numerical values of 0 and 1 are assigned to the two outcomes of a binary variable. Often, the 0 represents a negative response and the 1 represents a positive response. The mean of this variable will be the proportion of positive responses. If p is the proportion of observations with an outcome of 1, then 1-p is the probability of a outcome of 0. The ratio $p/(1-p)$ is called the *odds* and the *logit* is the logarithm of the odds, or just *log odds*. Mathematically, the logit transformation is written

$$l = \text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$$

The following table shows the logit for various values of *p*.

P	Logit(P)	P	Logit(P)
0.001	-6.907	0.999	6.907
0.01	-4.595	0.99	4.595
0.05	-2.944	0.95	2.944
0.10	-2.197	0.90	2.197
0.20	-1.386	0.80	1.386
0.30	-0.847	0.70	0.847
0.40	-0.405	0.60	0.405
0.50	0.000		

Note that while *p* ranges between zero and one, the logit ranges between minus and plus infinity. Also note that the zero logit occurs when *p* is 0.50.

The *logistic* transformation is the inverse of the logit transformation. It is written

$$p = \text{logistic}(l) = \frac{e^l}{1+e^l}$$

The Log Odds Ratio Transformation

The difference between two log odds can be used to compare two proportions, such as that of males versus females. Mathematically, this difference is written

$$\begin{aligned} l_1 - l_2 &= \text{logit}(p_1) - \text{logit}(p_2) \\ &= \ln\left(\frac{p_1}{1-p_1}\right) - \ln\left(\frac{p_2}{1-p_2}\right) \\ &= \ln\left(\frac{\left(\frac{p_1}{1-p_1}\right)}{\left(\frac{p_2}{1-p_2}\right)}\right) \\ &= \ln\left(\frac{p_1(1-p_2)}{p_2(1-p_1)}\right) \\ &= \ln(OR_{1,2}) \end{aligned}$$

This difference is often referred to as the *log odds ratio*. The odds ratio is often used to compare proportions across groups. Note that the logistic transformation is closely related to the odds ratio. The reverse relationship is

$$OR_{1,2} = \text{e}^{(l_1 - l_2)}$$

The Logistic Regression and Logit Models

In logistic regression, a categorical dependent variable *Y* having *G* (usually *G* = 2) unique values is regressed on a set of *p* independent variables $X_1, X_2, ..., X_p$. For example, *Y* may be presence or absence of a disease, condition after surgery, or marital status. Since the names of these partitions are arbitrary, we often refer to them by consecutive numbers. That is, in the discussion below, *Y* will take on the values 1, 2, … *G*. In fact, **NCSS** allows *Y* to have both numeric and text values, but the notation is much simpler if integers are used.

Let

$$\mathrm{X} = \left(X_1, X_2, \cdots, X_p\right)$$

$$B_g = \begin{pmatrix} \beta_{g1} \\ \vdots \\ \beta_{gp} \end{pmatrix}$$

The logistic regression model is given by the *G* equations

$$\ln\left(\frac{p_g}{p_1}\right) = \ln\left(\frac{P_g}{P_1}\right) + \beta_{g1}X_1 + \beta_{g2}X_2 + \cdots + \beta_{gp}X_p$$
$$= \ln\left(\frac{P_g}{P_1}\right) + \mathrm{XB}_g$$

Here, p_g is the probability that an individual with values $X_1, X_2, ..., X_p$ is in outcome *g*. That is,

$$p_g = \Pr\left(Y = g \mid \mathrm{X}\right)$$

Usually $X_1 \equiv 1$ (that is, an intercept is included), but this is not necessary.

The quantities $P_1, P_2, ..., P_G$ represent the prior probabilities of outcome membership. If these prior probabilities are assumed equal, then the term $\ln\left(P_g / P_1\right)$ becomes zero and drops out. If the priors are not assumed equal, they change the values of the intercepts in the logistic regression equation.

Outcome one is called the *reference value*. The regression coefficients $\beta_{11}, \beta_{12}, \cdots, \beta_{1p}$ for the reference value are set to zero. The choice of the reference value is arbitrary. Usually, it is the most frequent value or a control outcome to which the other outcomes are to be compared. This leaves *G*-1 logistic regression equations in the logistic model.

The $\beta's$ are population regression coefficients that are to be estimated from the data. Their estimates are represented by *b*'s. The $\beta's$ represents unknown parameters to be estimated, while the *b*'s are their estimates.

These equations are linear in the logits of *p*. However, in terms of the probabilities, they are nonlinear. The corresponding nonlinear equations are

$$\mathrm{p_g} = \mathrm{Prob}\left(Y = g \mid \mathrm{X}\right) = \frac{e^{\mathrm{XB}_g}}{1 + e^{\mathrm{XB}_2} + e^{\mathrm{XB}_3} + \cdots + e^{\mathrm{XB}_G}}$$

since $e^{\mathrm{XB}_1} = 1$ because all of its regression coefficients are zero.

A note on the names of the models. Often, all of these models are referred to as *logistic regression models*. However, when the independent variables are coded as ANOVA type models, they are sometimes called *logit models*.

A note about the interpretation of e^{XB} may be useful. Using the fact that $e^{a+b} = \left(e^{a}\right)\left(e^{b}\right)$, e^{XB} may be re-expressed as follows

$$\begin{aligned} e^{\mathrm{XB}} &= e^{\beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \cdots + \beta_p \mathrm{X}_p} \\ &= e^{\beta_1 \mathrm{X}_1} e^{\beta_2 \mathrm{X}_2} \cdots e^{\beta_p \mathrm{X}_p} \end{aligned}$$

This shows that the final value is the product of its individual terms.

Solving the Likelihood Equations

To improve notation, let

$$\begin{aligned} \pi_{gj} &= \mathrm{Prob}\left(Y = g \mid \mathrm{X}_j\right) \\ &= \frac{e^{\mathrm{X}_j \mathrm{B}_g}}{e^{\mathrm{X}_j \mathrm{B}_1} + e^{\mathrm{X}_j \mathrm{B}_2} + \cdots + e^{\mathrm{X}_j \mathrm{B}_G}} \\ &= \frac{e^{\mathrm{X}_j \mathrm{B}_g}}{\sum_{s=1}^{G} e^{\mathrm{X}_j \mathrm{B}_s}} \end{aligned}$$

The likelihood for a sample of *N* observations is then given by

$$l = \prod_{\mathrm{j}=1}^{\mathrm{N}} \prod_{g=1}^{G} \pi_{gj}^{\,y_{gj}}$$

where y_{gj} is one if the j^{th} observation is in outcome *g* and zero otherwise.

Using the fact that $\sum_{g=1}^{G} y_{gj} = 1$, the log likelihood, *L*, is given by

$$\begin{aligned} L = \ln(l) &= \sum_{j=1}^{\mathrm{N}} \sum_{g=1}^{G} \mathrm{y}_{gj} \ln\left(\pi_{gj}\right) \\ &= \sum_{j=1}^{\mathrm{N}} \sum_{g=1}^{G} \mathrm{y}_{gj} \ln\left(\frac{\mathrm{e}^{\mathrm{X}_j \mathrm{B}_g}}{\sum_{s=1}^{G} \mathrm{e}^{\mathrm{X}_j \mathrm{B}_s}}\right) \\ &= \sum_{\mathrm{j}=1}^{\mathrm{N}} \left[\sum_{\mathrm{g}=1}^{G} y_{gj} \mathrm{X}_j \mathrm{B}_g - \ln\left(\sum_{g=1}^{G} \mathrm{e}^{\mathrm{X}_j \mathrm{B}_g} \right) \right] \end{aligned}$$

Maximum likelihood estimates of the $\beta' s$ are those values that maximize this log likelihood equation. This is accomplished by calculating the partial derivatives and setting them to zero. The resulting likelihood equations are

$$\frac{\partial L}{\partial \beta_{ik}} = \sum_{j=1}^{\mathrm{N}} x_{kj}\left(y_{ig} - \pi_{ig}\right)$$

for *g* = 1, 2, …, *G* and *k* = 1, 2, …, *p*. Actually, since all coefficients are zero for *g* = 1, the effective range of *g* is from 2 to *G*.

Because of the nonlinear nature of the parameters, there is no closed-form solution to these equations, and they must be solved iteratively. The Newton-Raphson method as described in Albert and Harris (1987) is used to solve these equations. This method makes use of the information matrix, $I(\beta)$, which is formed from the matrix of second partial derivatives. The elements of the information matrix are given by

$$\frac{\partial^2 L}{\partial\beta_{ik}\partial\beta_{ik'}} = -\sum_{j=1}^{N} x_{kj}x_{k'j}\pi_{ig}\left(1-\pi_{ig}\right)$$

$$\frac{\partial^2 L}{\partial\beta_{ik}\partial\beta_{i'k'}} = \sum_{j=1}^{N} x_{kj}x_{k'j}\pi_{ig}\pi_{i'g}$$

The information matrix is used because the asymptotic covariance matrix of the maximum likelihood estimates is equal to the inverse of the information matrix. That is,

$$V\left(\hat{\beta}\right) = I(\beta)^{-1}$$

This covariance matrix is used in the calculation of confidence intervals for the regression coefficients, odds ratios, and predicted probabilities.

Interpretation of Regression Coefficients

The interpretation of the estimated regression coefficients is not as easy as in multiple regression. In logistic regression, not only is the relationship between *X* and *Y* nonlinear, but also, if the dependent variable has more than two unique values, there are several regression equations.

Consider the usual case of a binary dependent variable, *Y*, and a single independent variable, *X*. Assume that *Y* is coded so it takes on the values 0 and 1. In this case, the logistic regression equation is

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

Now consider impact of a unit increase in *X*. The logistic regression equation becomes

$$\begin{aligned}\ln\left(\frac{p'}{1-p'}\right) &= \beta_0 + \beta_1(X+1)\\ &= \beta_0 + \beta_1 X + \beta_1\end{aligned}$$

We can isolate the slope by taking the difference between these two equations. We have

$$\begin{aligned}\beta_1 &= \beta_0 + \beta_1(X+1) - (\beta_0 + \beta_1 X)\\ &= \ln\left(\frac{p'}{1-p'}\right) - \ln\left(\frac{p}{1-p}\right)\\ &= \ln\left(\frac{\frac{p'}{1-p'}}{\frac{p}{1-p}}\right)\\ &= \ln\left(\frac{odds'}{odds}\right)\end{aligned}$$

That is, β_1 is the log of the ratio of the odds at X+1 and X. Removing the logarithm by exponentiating both sides gives

$$e^{\beta_1} = \frac{odds'}{odds}$$

The regression coefficient β_1 is interpreted as the log of the odds ratio comparing the odds after a one unit increase in X to the original odds. Note that, unlike multiple regression, the interpretation of β_1 depends on the particular value of X since the probability values, the p's, will vary for different X.

Binary X

When X can take on only two values, say 0 and 1, the above interpretation becomes even simpler. Since there are only two possible values of X, there is a unique interpretation for β_1 given by the log of the odds ratio. In mathematical terms, the meaning of β_1 is then

$$\beta_1 = \ln\left(\frac{odds(X=1)}{odds(X=0)}\right)$$

To understand this equation further, consider first what the odds are. The odds is itself the ratio of two probabilities, p and 1-p. Consider the following table of odds values for various values of p. Note that 9:1 is read '9 to 1.'

Value of *p*	Odds of *p*
0.9	9:1
0.8	4:1
0.6	1.5:1
0.5	1:1
0.4	0.67:1
0.2	0.25:1
0.1	0.11:1

Now, using a simple example from horse racing, if one horse has 8:1 odds of winning and a second horse has 4:1 odds of winning, how do you compare these two horses? One obvious way is to look at the ratio of their odds. The first horse has twice the odds of winning as the second.

Consider a second example of two slow horses whose odds of winning are 0.1:1 and 0.05:1. Here again, their odds ratio is 2. The message here: the odds ratio gives a relative number. Even though the first horse is twice as likely to win as the first, it is still a long shot.

To completely interpret β_1, we must take the logarithm of the odds ratio. It is difficult to think in terms of logarithms. However, we can remember that the log of one is zero. So, a positive value of β_1 indicates that the odds of the numerator are larger, while a negative value indicates that the odds of the denominator are larger.

It is may easiest to think in terms of e^{β_1} rather than β_1, because e^{β_1} is the odds ratio while β_1 is the log of the odds ratio. Both quantities are displayed in the reports.

Multiple Independent Variables

When there are multiple independent variables, the interpretation of each regression coefficient becomes more difficult, especially if interaction terms are included in the model. In general, however, the regression coefficient is interpreted the same as above, except that the caveat 'holding all other independent variables constant' must be added. The question becomes, can the value of this independent variable be increased by one without changing any of the other variables. If it can, then the interpretation is as before. If not, then some type of conditional statement must be added that accounts for the values of the other variables.

Multinomial Dependent Variable

When the dependent variable has more than two values, there will be more than one regression equation. In fact, the number of regression equations is equal to one less than the number of outcomes. This makes interpretation more difficult because there are several regression coefficients associated with each independent variable. In this case, care must be taken to understand what each regression equation is predicting. Once this is understood, interpretation of each of the *G* - 1 regression coefficients for each variable can proceed as above.

Consider the following example in which there are two independent variables, X1 and X2, and the dependent variable has three groups: *A*, *B*, and *C*.

Row	Y	X1	X2	GA	GB	GC
1	A	3.2	5.8	1	0	0
2	A	4.7	6.1	1	0	0
3	B	2.8	3.5	0	1	0
4	B	3.3	4.6	0	1	0
5	B	3.9	5.2	0	1	0
6	C	4.2	3.7	0	0	1
7	C	7.3	4.4	0	0	1
8	C	5.3	5.1	0	0	1
9	C	6.8	4.5	0	0	1

Look at the three indicator variables: *GA*, *GB*, and *GC*. They are set to one or zero depending on whether *Y* takes on the corresponding value. Two regression equations will be generated corresponding to any two of these indicator variables. The value that is not used is called the *reference value*. Suppose the reference value is *C*. The two regression equations would be

$$\ln\left(\frac{p_A}{p_C}\right) = \beta_{A0} + \beta_{A1}X_1 + \beta_{A2}X_2$$

and

$$\ln\left(\frac{p_B}{p_C}\right) = \beta_{B0} + \beta_{B1}X_1 + \beta_{B2}X_2$$

The two coefficients for *X1* in these equations, β_{A1} and β_{B1}, give the change in the log odds of A versus C and B versus C for a one unit change in *X1*, respectively.

Statistical Tests and Confidence Intervals

Inferences about individual regression coefficients, groups of regression coefficients, goodness-of-fit, mean responses, and predictions of group membership of new observations are all of interest. These inference procedures can be treated by considering hypothesis tests and/or confidence intervals. The inference procedures in logistic regression rely on large sample sizes for accuracy.

Two procedures are available for testing the significance of one or more independent variables in a logistic regression: likelihood ratio tests and Wald tests. Simulation studies usually show that the likelihood ratio test performs better than the Wald test. However, the Wald test is still used to test the significance of individual regression coefficients because of its ease of calculation.

These two testing procedures will be described next.

Likelihood Ratio and Deviance

The *Likelihood Ratio* test statistic is -2 times the difference between the log likelihoods of two models, one of which is a subset of the other. The distribution of the LR statistic is closely approximated by the chi-square distribution for large sample sizes. The degrees of freedom (DF) of the approximating chi-square distribution is equal to the difference in the number of regression coefficients in the two models. The test is named as a ratio rather than a difference since the difference between two log likelihoods is equal to the log of the ratio of the two likelihoods. That is, if L_{full} is the log likelihood of the full model and L_{subset} is the log likelihood of a subset of the full model, the likelihood ratio is defined as

$$\begin{aligned} LR &= -2\left[L_{\text{subset}} - L_{\text{full}}\right] \\ &= -2\left[\ln\left(\frac{l_{\text{subset}}}{l_{\text{full}}}\right)\right] \end{aligned}$$

Note that the -2 adjusts *LR* so the chi-square distribution can be used to approximate its distribution.

The likelihood ratio test is the test of choice in logistic regression. Various simulation studies have shown that it is more accurate than the Wald test in situations with small to moderate sample sizes. In large samples, it performs about the same. Unfortunately, the likelihood ratio test requires more calculations than the Wald test, since it requires that two maximum-likelihood models must be fit.

Deviance

When the full model in the likelihood ratio test statistic is the saturated model, *LR* is referred to as the *deviance*. A saturated model is one which includes all possible terms (including interactions) so that the predicted values from the model equal the original data. The formula for the deviance is

$$D = -2\left[L_{\text{Reduced}} - L_{\text{Saturated}}\right]$$

The deviance may be calculated directly using the formula for the deviance residuals (discussed below). This formula is

$$D = 2\sum_{j=1}^{J}\sum_{g=1}^{G} w_{gj} \ln\left(\frac{w_{gj}}{n_j p_{gj}}\right)$$

This expression may be used to calculate the log likelihood of the saturated model without actually fitting a saturated model. The formula is

$$L_{\text{Saturated}} = L_{\text{Reduced}} + \frac{D}{2}$$

The deviance in logistic regression is analogous to the residual sum of squares in multiple regression. In fact, when the deviance is calculated in multiple regression, it is equal to the sum of the squared residuals. Deviance residuals, to be discussed later, may be squared and summed as an alternative way to calculate the deviance, *D*.

The change in deviance, ΔD, due to excluding (or including) one or more variables is used in logistic regression just as the partial *F* test is used in multiple regression. Many texts use the letter *G* to represent ΔD, but we have already used *G* to represent the number of groups in *Y*. Instead of using the *F* distribution, the distribution of the change in deviance is approximated by the chi-square distribution. Note that since the log likelihood for the saturated model is common to both deviance values, ΔD is calculated without actually estimating the saturated model. This fact becomes very important during subset selection. The formula for ΔD for testing the significance of the regression coefficient(s) associated with the independent variable *X*1 is

$$\begin{aligned} \Delta D_{X1} &= D_{\text{without } X1} - D_{\text{with } X1} \\ &= -2\left[L_{\text{without } X1} - L_{\text{Saturated}}\right] + 2\left[L_{\text{with } X1} - L_{\text{Saturated}}\right] \\ &= -2\left[L_{\text{without } X1} - L_{\text{with } X1}\right] \end{aligned}$$

Note that this formula looks identical to the likelihood ratio statistic. Because of the similarity between the change in deviance test and the likelihood ratio test, their names are often used interchangeably.

Wald Test

The Wald test will be familiar to those who use multiple regression. In multiple regression, the common *t*-test for testing the significance of a particular regression coefficient is a Wald test. In logistic regression, the Wald test is calculated in the same manner. The formula for the Wald statistic is

$$z_j = \frac{b_j}{s_{b_j}}$$

where s_{b_j} is an estimate of the standard error of b_j provided by the square root of the corresponding diagonal element of the covariance matrix, $V\left(\hat{\beta}\right)$.

With large sample sizes, the distribution of z_j is closely approximated by the normal distribution. With small and moderate sample sizes, the normal approximation is described as 'adequate.'

The Wald test is used in ***NCSS*** to test the statistical significance of individual regression coefficients.

Confidence Intervals

Confidence intervals for the regression coefficients are based on the Wald statistics. The formula for the limits of a $100(1-\alpha)\%$ two-sided confidence interval is

$$b_j \pm \left|z_{\alpha/2}\right| s_{b_j}$$

R-Squared

The following discussion summarizes the material on this subject in Hosmer and Lemeshow (1989). In multiple regression, R_M^2 represents the proportion of variation in the dependent variable accounted for by the independent variables. (The subscript "M" emphasizes that this statistic is for multiple regression.) It is the ratio of the regression sum of squares to the total sum of squares. When the residuals from the multiple regression can be assumed to be normally distributed, R_M^2 can be calculated as

$$R_M^2 = \frac{L_p - L_0}{L_0}$$

where L_0 is the log likelihood of the intercept-only model and L_p is the log likelihood of the model that includes the independent variables. Note that L_p varies from L_0 to 0. R_M^2 varies between zero and one.

This quantity has been proposed for use in logistic regression. Unfortunately, when R_L^2 (the R-squared for logistic regression) is calculated using the above formula, it does not necessarily range between zero and one. This is because the maximum value of L_p is not always 0 as it is in multiple regression. Instead, the maximum value of L_p is L_S, the log likelihood of the saturated model. To allow R_L^2 to vary from zero to one, it is calculated as follows

$$R_L^2 = \frac{L_p - L_0}{L_0 - L_S}$$

The introduction of L_S into this formula causes a degree of ambiguity with R_L^2 that does not exist with R_M^2. This ambiguity is due to the fact that the value of L_S depends on the configuration of independent variables. The following example will point out the problem.

Consider a logistic regression problem consisting of a binary dependent variable and a pool of four independent variables. The data for this example are given in the following table.

Y	X1	X2	X3	X4
0	1	1	2.3	5.9
0	1	1	3.6	4.8
1	1	1	4.1	5.6
0	1	2	5.3	4.1
0	1	2	2.8	3.1
1	1	2	1.9	3.7
1	1	2	2.5	5.4
1	2	1	2.3	2.6
1	2	1	3.9	4.6
0	2	1	5.6	4.9
0	2	2	4.2	5.9
0	2	2	3.8	5.7
0	2	2	3.1	4.5
1	2	2	3.2	5.5
1	2	2	4.5	5.2

Notice that if only X1 and X2 are included in the model, the dataset may be collapsed because of the number of repeats. In this case, the value of L_S will be less than zero. However, if X3 or X4 are used there are no repeats and the value of L_S will be zero. Hence, the denominator of R_L^2 depends on which of the independent variables is used. This is not the case for R_M^2. This ambiguity comes into play especially during subset selection. It means that as you enter and remove independent variables, the target value L_S can change.

Hosmer and Lemeshow (1989) recommend against the use R_L^2 as a goodness of fit measure. However, we have included it in our output because it does provide a comparative measure of the proportion of the log likelihood that is accounted for by the model. Just remember than an R_L^2 value of 1.0 indicates that the logistic regression model achieves the same log likelihood as the saturated model. However, this does not mean that it fits the data perfectly. Instead, it means that it fits the data as well as could be hoped for.

Residual Diagnostics

Residuals are the discrepancies between the data values and their predicted values from the fitted model. A residual analysis detects outliers, identifies influential observations, and diagnoses the appropriateness of the logistic model. An analysis of the residuals should be conducted before a regression model is used.

Unfortunately, the residuals are more difficult to define in logistic regression than in regular multiple regression because of the nonlinearity of the logistic model and because more than one regression equation is used. The discussion that follows provides an introduction to the residuals that are produced by the logistic regression procedure. Pregibon (1981) presented this material for the case of the two-outcome logistic regression. Extensions of Pregibon's results to the multiple-group case are provided in an article by Lesaffre and Albert (1989) and in the book by Hosmer and Lemeshow (1989). Lesaffre and Albert provide formulas for these extensions. On the other hand, Hosmer and Lemeshow recommend that individual logistic regressions be run in which the each group is treated separately. Hence, if you have three outcomes A, B, and C, you would run outcome A versus outcomes B and C, outcome B versus outcomes A and C, and outcome C versus outcomes and A and B. You would conduct a residual analysis for each of these regressions using Pregibon's two-outcome formulas. In ***NCSS***, we have adopted the approach of Hosmer and Lemeshow.

Data Configuration

When dealing with residuals, it is important to understand the data configuration. Often, residual formulations are presented for the case when each observation has a different combination of values of the independent variables. When some observations have identical independent variables or when you have specified a frequency variable, these observations are combined to form a single row of data. The *N* original observations are combined to form *J* unique rows. The response indicator variables y_{gj} for the original observations are replaced by two variables: w_{gj} and n_j. The variable n_j is the total number of observations with this independent variable configuration. The variable w_{gj} is the number of the n_j observations that are in outcome-group *g*.

NCSS automatically collapses the dataset of *N* observations into a combined dataset of *J* rows for analysis. The residuals are calculated using this last formula. However, the residuals are reported in the original observation order. Thus, if two identical observations have been combined, the residual is shown for each. If corrective action needs to be taken because a residual is too large, both observations must be deleted. Also, if you want to calculate the deviance or Pearson chi-square from the corresponding residuals, care must be taken that you use only the *J* collapsed rows, not the *N* original observations.

Simple Residuals

Each of the *G* logistic regression equations can be used to estimate the probabilities that an observation of independent variable values given by X_j belongs to the corresponding outcome-group. The actual values of these probabilities were defined earlier as

$$\pi_{gj} = \text{Prob}\left(Y = g \mid \mathrm{X}_j\right)$$

The estimated values of these probabilities are called p_{gj}. If the hat symbol is used to represent an estimated parameter, then

$$p_{gj} = \hat{\pi}_{gj}$$

These estimated probabilities can be compared to the actual probabilities occurring in the database by subtracting the two quantities, forming a residual. The actual values were defined as the indicator variables y_{gj}. Thus, simple residuals may be defined as

$$r_{gj} = y_{gj} - p_{gj}$$

Note that, unlike multiple regression, there are *g* residuals for each observation instead of just one. This makes residual analysis much more difficult. If the logistic regression model fits an observation closely, all of its residuals will be small. Hence, when y_{gj} is one, p_{gj} will be close to one and when y_{gj} is zero, p_{gj} will be close to zero.

Unfortunately, the simple residuals have unequal variance equal to $n_j\pi_{gj}\left(1-\pi_{gj}\right)$, where n_j is the number of observations with the same values of the independent variables as observation *j*. This unequal variance makes comparisons among the simple residuals difficult and alternative types of residuals are necessary.

Pearson Residuals

One popular alternative to the simple residuals are the *Pearson residuals* which are so named because they give the contribution of each observation to the Pearson chi-square goodness of fit statistic. When the values of the independent variables of each observation are unique, the formula this residual is

$$\chi'_j = \pm\sqrt{\sum_{g=1}^{G}\frac{\left(y_{gj}-p_{gj}\right)^2}{p_{gj}}}, \quad j=1,2,\cdots,N$$

The negative sign is used when $y_{gj}=0$ and the positive sign is used when $y_{gj}=1$.

When some of the observations are duplicates and the database has been collapsed (see Data Configuration above) the formula is

$$\chi_j = \pm\sqrt{\sum_{g=1}^{G}\frac{\left(w_{gj}-n_j p_{gj}\right)^2}{n_j p_{gj}}}, \quad j=1,2,\cdots,J$$

where the plus (minus) is used if w_{gj}/n_j is greater (less) than p_{gj}. Note that this is the formula used by ***NCSS***.

By definition, the sum of the squared Pearson residuals is the Pearson chi-square goodness of fit statistics. That is,

$$\chi^2 = \sum_{j=1}^{J}\chi_j^2$$

Deviance Residuals

Remember that the deviance is -2 times the difference between log likelihoods of a reduced model and the saturated model. The deviance is calculated using

$$\begin{aligned}
D &= -2\left[L_{\text{Reduced}} - L_{\text{Saturated}}\right] \\
&= -2\left[\sum_{j=1}^{N}\sum_{g=1}^{G} y_{gj}\ln\left(p_{gj}\right) - \sum_{j=1}^{N}\sum_{g=1}^{G} y_{gj}\ln\left(y_{gj}\right)\right] \\
&= -2\left[\sum_{j=1}^{N}\sum_{g=1}^{G} y_{gj}\ln\left(p_{gj}\right)\right] \\
&= \sum_{j=1}^{N}\left[2\sum_{g=1}^{G} y_{gj}\ln\left(\frac{1}{p_{gj}}\right)\right]
\end{aligned}$$

This formula uses the fact that the saturated model reproduces the original data exactly and that, in these sums, the value of 0 ln(0) is defined as 0 and that the ln(1) is also 0.

The deviance residuals are the square roots of the contribution of each observation to the overall deviance. Thus, the formula for the deviance residual is

$$d'_j = \pm\sqrt{2\sum_{g=1}^{G} y_{gj}\ln\left(\frac{1}{p_{gj}}\right)}, j=1,2,\cdots,N$$

The negative sign is used when $y_{gj}=0$ and the positive sign is used when $y_{gj}=1$.

When some of the observations are duplicates and the database has been collapsed (see Data Configuration above) the formula is

$$d_j = \pm\sqrt{2\sum_{g=1}^{G} w_{gj} \ln\left(\frac{w_{gj}}{n_j p_{gj}}\right)}, \quad j = 1,2,\cdots,J$$

where the plus (minus) is used if $w_{REF(g),j} / n_j$ is greater (less) than $p_{REF(g),j}$. Note that this is the formula used by **NCSS**.

By definition, the sum of the squared deviance residuals is the deviance. That is,

$$D = \sum_{j=1}^{J} d_j^2$$

Hat Matrix Diagonal

The diagonal elements of the hat matrix can be used to detect points that are extreme in the independent variable space. These are often called *leverage* design points. The larger the value of this statistic, the more the observation influences that estimates of the regression coefficients. An observation that is discrepant, but has low leverage, should not cause much concern. However, an observation with a large leverage and a large residual should be checked very carefully. The use of these hat diagonals is discussed further in the multiple regression chapter.

The formula for the hat diagonal associated with the *j*th observation and *g*th outcome is

$$h_{gj} = n_j p_{gj}\left(1 - p_{gj}\right)\sum_{i=1}^{p}\sum_{k=1}^{p} X_{ij} X_{kj} \hat{V}_{gik}, \quad j = 1,2,\cdots,J$$

where $\hat{V}_{gik}$ is the portion of the covariance matrix of the regression coefficients associated with the *g*th regression equation. The interpretation of this diagnostic is not as clear in logistic regression as in multiple regression because it involves the predicted values which in turn involve the dependent variable. In multiple regression, the hat diagonals only involve the independent variables.

Note that this formula matches Pregibon (1981) in the two-group case. In the multiple-group case, the two-group formula is applied to each group.

DFBETA

One way to study the impact of an observation on each regression coefficient is to determine how much that coefficient changes when the observation is deleted. The DFBETA statistic is the standardized difference between a regression coefficient before and after the removal of the *j*th observation.

The formula for DFBETA is approximated by

$$\text{DFBETA}_{gij} = \left(\frac{w_{gj} - n_j p_{gj}}{\left(1 - h_{gj}\right)\sqrt{\hat{V}_{gii}}}\right)\sum_{k=1}^{p} X_{kj} \hat{V}_{gik}, \quad j = 1,2,\cdots,J$$

where $\hat{V}_{gik}$ is the portion of the covariance matrix associated with the *g*th regression equation.

Note that this formula matches Pregibon (1981) in the two-outcome case, but is different from Lesaffre (1989) in the multi-group case.

Cooks Distance: C and Cbar

C and *Cbar* are extensions of Cooks distance for logistic regression. Quoting from Pregibon (1981), page 719:

"Cbar measures the overall change in fitted logits due to deleting the *l*th observation for all points excluding the one deleted. Conversely, *C* includes the deleted point. Although *C* will usually be the preferred diagnostic to measure overall coefficients changes, in the examples examined to date, the one-step approximations were more accurate for *Cbar* than *C*."

The formulas for *C* and *Cbar* are

$$C_{gj} = \frac{\chi_j^2 h_{gj}}{\left(1-h_{gj}\right)^2}, \quad j = 1,2,\cdots,J$$

$$\overline{C}_{gj} = \frac{\chi_j^2 h_{gj}}{\left(1-h_{gj}\right)}, \quad j = 1,2,\cdots,J$$

Note that this formula matches Pregibon (1981) in the two-group case. In the multiple-group case, the two-group formula is applied to each group.

DFDEV and DFCHI2

DFDEV and *DFCHI2* are statistics that measure the change in deviance and in Pearson's chi-square, respectively, that occurs when an observation is deleted from the dataset. Large values of these statistics indicate observations that have not been fitted well.

The formulas for these statistics are

$$DFDEV_{gj} = d_j^2 + \overline{C}_{gj}, \quad j = 1,2,\cdots,J$$

$$DFCHI2_{gj} = \frac{\overline{C}_{gj}}{h_{gj}}, \quad j = 1,2,\cdots,J$$

Note that this formula matches Pregibon (1981) in the two-group case. In the multiple-group case, the two-group formula is applied to each group.

Predicted Probabilities

This section describes how to calculate the predicted probabilities of outcome-group membership and associated confidence intervals. Recall that the regression equation is linear when expressed in logit form. That is,

$$\begin{aligned}\ln\left(\frac{p_g}{p_1}\right) &= \ln\left(\frac{P_g}{P_1}\right) + \beta_{g1}X_1 + \beta_{g2}X_2 + \cdots + \beta_{gp}X_p \\ &= \ln\left(\frac{P_g}{P_1}\right) + \mathrm{XB}_g\end{aligned}$$

The adjustment for the prior probabilities changes the value of the intercepts, so this expression may be simplified to

$$\begin{aligned}\ln\left(\frac{p_g}{p_1}\right) &= \beta_{g1}X_1 + \beta_{g2}X_2 + \cdots + \beta_{gp}X_p \\ &= \mathrm{XB}_g\end{aligned}$$

if we assume that the intercepts have been appropriately adjusted. Assuming that the estimated matrix of regression coefficients is distributed asymptotically as a multivariate normal, the point estimates of this quantity for a specific set of X values is given by

$$l_j = \ln\left(\frac{p_g}{p_1} \mid \mathrm{X}_j\right) = \mathrm{X}_j\hat{B}_g$$

and the corresponding confidence interval is given by

$$l_j \pm z_{\alpha/2}\left(\mathrm{X}_j' \mathrm{V}_g \mathrm{X}_j\right)$$

where V_g is that portion of the covariance matrix $V\left(\hat{B}\right)$ that deals with the gth regression equation.

When there are only two groups, these confidence limits can be inverted to give confidence limits on the predicted probabilities as

$$\frac{1}{1+e^{\mathrm{X}_j\hat{\mathrm{B}}\pm z_{\alpha/2}\sigma_B}} \text{ and } \frac{e^{\mathrm{X}_j\hat{\mathrm{B}}\pm z_{\alpha/2}\sigma_B}}{1+e^{\mathrm{X}_j\hat{\mathrm{B}}\pm z_{\alpha/2}\sigma_B}}$$

where

$$\sigma_B = \mathrm{X}_j' \mathrm{V}_g \mathrm{X}_j$$

When there are more than two groups, the confidence limits on the logits are still given by

$$l_j \pm z_{\alpha/2}\left(\mathrm{X}_j' \mathrm{V}_g \mathrm{X}_j\right)$$

However, this set of confidence limits of the logits cannot be inverted to give confidence limits for the predicted probabilities. We have found no presentation that gives an appropriate set of confidence limits. In order to provide an approximate answer, we provide approximate confidence limits by applying the inversion as if there were only two groups. This method ignores the correlation between the coefficients of the individual equations. However, we hope that it provides a useful approximation to the confidence intervals.

Subset Selection

Subset selection refers to the task of finding a small subset of the available independent variables that does a good job of predicting the dependent variable. Because logistic regression must be solved iteratively, the task of finding the best subset can be very time consuming. Hence, techniques that search all possible combinations of the independent variables are not feasible. Instead, algorithms that add or remove a variable at each step must be used. Two such searching algorithms are available in this module: forward selection and forward selection with switching.

Before discussing the details of these two algorithms, it is important to comment on a couple of issues that can come up. First of all, since there is more than one regression equation when there are more than two categories in the dependent variable, it is possible that a variable is important in one of the equations and not in the others. The algorithms presented here are based on the overall likelihood. This means that if an independent variable is important in at least one of the regression equations, it will be kept.

A second issue is what to do with the individual-degree of freedom variables that are generated for a categorical independent variable. If such a variable has six categories, five binary variables are generated. You can see that with two or three categorical variables, a large number of binary variables may result, which greatly increases the total number of variables that must be searched. To avoid this problem, the algorithms search on model terms rather than on the individual binary variables. Thus, the whole set of binary variables associated with a given term are considered together for inclusion in, or deletion from, the model. It is all or none. Because of the time consuming nature of the algorithm, this is the only feasible way to deal with categorical variables. If you want the subset algorithm to deal with them individually, you can generate the set of binary variables manually and designate them as Numeric Variables.

Hierarchical Models

A third issue is what to do with interactions. Usually, an interaction is not entered in the model unless the individual terms that make up that interaction are also in the model. For example, the interaction term A*B*C is not included unless the terms A, B, C, A*B, A*C, and B*C are already in the model. Such models are said to be *hierarchical*. You have the option during the search to force the algorithm to consider only hierarchical models during its search. Thus, if C is not in the model, interactions involving C are not even considered. Even though the option for non-hierarchical models is available, we recommend that you only consider hierarchical models.

Forward Selection

The method of forward selection proceeds as follows.

1. Begin with no terms in the model.
2. Find the term that, when added to the model, achieves the largest value of the log likelihood. Enter this term into the model.
3. Continue adding terms until a target value for the log-likelihood is achieved or until a preset limit on the maximum number of terms in the model is reached. Note that these terms can be limited to those keeping the model hierarchical.

This method is comparatively fast, but it does not guarantee that the best model is found except for the first step when it finds the best single term. You might use it when you have a large number of observations and terms so that other, more time consuming, methods are not feasible.

Forward Selection with Switching

This method is similar to the method of Forward Selection discussed above. However, at each step when a term is added, all terms in the model are switched one at a time with all candidate terms not in the model to determine if they increase the value of the log likelihood. If a switch can be found, it is made and the pool of terms is again searched to determine if another switch can be made. Note that this switching can be limited to those keeping the model hierarchical.

When the search for possible switches does not yield a candidate, the subset size is increased by one and a new search is begun. The algorithm is terminated when a target subset size is reached or all terms are included in the model.

Discussion

These algorithms usually require two runs. In the first run, you set the maximum subset size to a large value such as 10. By studying the Subset Selection reports from this run, you can quickly determine the optimum number of terms. You reset the maximum subset size to this number and make the second run. This two-step procedure works better than relying on some F-to-enter and F-to-remove tests whose properties are not well understood to begin with.

Data Structure

The data given below are the first few rows of a set of data about leukemia patients published in Lee (1980). The dependent variable is whether leukemia remission occurred (Remiss). The independent variables are cellularity of the marrow clot section (Cell), smear differential percentage of blasts (Smear), percentage of absolute marrow leukemia cell infiltrate (Infil), percentage labeling index of the bone marrow leukemia cells (LI), absolute number of blasts in the peripheral blood (Blast), and the highest temperature prior to start of treatment (Temp). This dataset is stored in the *Leukemia* dataset in the *Example Data* directory.

Leukemia dataset (subset)

Remiss	Cell	Smear	Infil	LI	Blast	Temp
1	80	83	66	190	11.6	996
1	90	36	32	140	4.5	992
0	80	88	70	80	0.5	982
0	100	87	87	70	10.3	986
1	90	75	68	130	2.3	980
0	100	65	65	60	2.3	982
1	95	97	92	100	16.0	992
0	95	87	83	190	21.6	1020

Missing Values

If missing values are found in any of the independent variables being used, the row is omitted. If only the dependent variable is missing, the row will not be used in the estimation of the regression coefficients, but a predicted value will be generated for that row.

Example 1 – Logistic Regression Analysis

This section presents an introductory example of how to run a logistic regression analysis. The data used are stored in the Leukemia dataset. In this analysis, a logistic regression will be run to determine the relationship between Cell, LI, and Temp on the binary dependent variable Remiss.

Setup

To run this example, complete the following steps:

1 Open the Leukemia example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Leukemia** and click **OK**.

2 Specify the Logistic Regression procedure options

- Find and open the **Logistic Regression** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables, Model Tab	
Y	**Remiss**
Numeric X's	**Cell, LI, Temp**
Reports Tab	
All Reports	**Checked**
Row Classification Report	**All Rows**
Row Classification Probs Report	**All Rows**
Simple Residuals Report	**All Rows**
Plots Tab	
All Plots	**Checked**
Report Options (*in the Toolbar*)	
Variable Labels	**Column Names**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Run Summary

Run Summary

Item	Value	Item	Value
Y Variable	Remiss	Rows Processed	29
Reference Value	0	Rows Used	27
Number of Y-Values	2	Rows for Validation	0
Frequency Variable	None	Rows X's Missing	0
Numeric X Variables	3	Rows Freq Miss. or 0	0
Categorical X Variables	0	Rows Prediction Only	2
Final Log Likelihood	-10.97669	Unique Rows (Y and X's)	26
Model R^2	0.36130	Sum of Frequencies	27
Actual Convergence	2.94261E-07	Likelihood Iterations	7
Target Convergence	1E-06	Maximum Iterations	20
Model D.F.	4	Completion Status	Normal Completion
Priors	Equal		

This report provides useful information about the reports to follow. It should be studied to make sure that the data were read in properly and that the logistic regression procedure terminated normally. We will only discuss those parameters that need special explanation.

Reference Value

The reference value is that category of the Y variable that is defined implicitly in terms of the other categories. This is the category that is skipped on much of the output. If you did not specify the reference value with the Y Variable, the reference value is chosen according to the 'Default Reference Value setting. This value is critical to interpretation of the rest of the output.

Number of Y-Values

This is the number of unique categories that were found for the Y variable. Check this count to make certain it agrees with what you anticipated.

Final Log Likelihood

This is the log likelihood of the model that is reported on here.

Model R^2

This is the R^2 that was achieved by your regression. Read the discussion of R^2 that was given earlier to better understand how to interpret R^2 in the case of logistic regression.

Actual and Target Convergence

The Target Convergence is the amount that is used to stop the iterative fitting of the maximum likelihood algorithm. If the Actual Convergence amount is larger than the Target amount, the algorithm ended before converging and care must be taken in using any of the results. If this happens, the usual remedy is to increase the maximum number of iterations. If this does not solve the problem, you will have to change the variables in the model.

Rows Processed, Used, etc.

These values record how many of each type of observation were encountered when the database was read. You should make sure that these amounts are what you expect.

Unique Rows (Y and X's)

This gives the number of unique patterns found in the variables. Both the dependent and independent variables are considered in forming this count.

Likelihood and Maximum Iterations

The Likelihood Iterations are the number of iterations necessary to solve the likelihood equations. Usually, fewer than ten iterations are necessary. If the number of Likelihood Iterations is equal to the Maximum Iterations, the maximum likelihood algorithm did not converge and you should take some remedial action such as increasing the Maximum Iterations or changing the regression model.

Completion Status

This is the message that was returned when the maximum likelihood algorithm ended. Unless the message "Normal Completion" is received, you should take appropriate corrective action.

Model D.F.

This is the number of degrees of freedom in the *G*-1 logistic regression models.

Y Variable Summary

Y Variable Summary

Y Remiss	Count	Unique Rows (Y and X's)	Y Proportion	Y Prior	R^2 (Y vs Pred. Probability)	Percent Correctly Classified
0	18	17	0.66667	0.50000	0.40318	83.333
1	9	9	0.33333	0.50000	0.40318	77.778
Total	27	26				81.481

This report describes the dependent variable. Use it to understand the dependent variable and how well the regression model approximates it.

Y

These are the unique values found for the dependent variable. Check to make sure that no unexpected outcomes were found.

Count

This is the sum of the frequencies (counts) for each outcome of the Y variable.

Unique Rows (Y and X's)

This is the number of unique rows in each outcome as determined by the values of the Y and X variables.

Y Proportion

This is the proportion of each outcome.

Prior

This is the prior probability of each Y-value as given by the user in the Prior Y-Value Probabilities section.

R2 (Y vs Pred. Probability)

This is the R^2 that is achieved when the indicator variable for this Y-value is regressed on the predicted probability of being in this category.

Percent Correctly Classified

This is the percent of the observations from this outcome that were correctly classified as such by the multinomial logistic regression model.

Coefficient Significance Tests

Coefficient Significance Tests

Independent Variable X	Regression Coefficient b(i)	Standard Error Sb(i)	Wald Z-Value H0: β=0	Wald P-Value	Odds Ratio Exp(b(i))
Intercept	68.32696	56.88604	1.201	0.22970	10000+
Cell	9.65213	7.75107	1.245	0.21303	10000+
LI	3.86710	1.77828	2.175	0.02966	47.80336
Temp	-82.07365	61.71233	-1.330	0.18354	0.00000

This report gives the estimated logistic regression equation and associated significance tests. The reference value of the dependent variable is shown in the title. If the dependent variable has more than two categories, the appropriate information is displayed for each of the *G*-1 equations.

Independent Variable X

This is the variable from the model that is displayed on this line. If the variable is continuous, it is displayed directly. If the variable is discrete, the definition of the binary variable that was generated is given. For example, suppose that a discrete independent GRADE variable has three values: A, B, and C. The name shown here would be something like *GRADE=B*. This refers to a binary variable that is one for those rows in which GRADE was B and zero otherwise.

Note that the placement of the name is controlled by the *Stagger label and output* option of the Report Options tab.

Regression Coefficient b(i)

This is the estimated value of the corresponding regression coefficient, sometimes referred to as B or Beta. The interpretation of the regression coefficients is difficult. We refer you to the discussion given at that beginning of this chapter for more details.

Standard Error Sb(i)

This is s_{b_j}, the large-sample estimate of the standard error of the regression coefficient. This is an estimate of the precision of the regression coefficient. It is used as the denominator of the Wald test.

Wald Z-Value H0: β=0

This is the *z* value of the Wald test used for testing the hypothesis that $\beta_{gj} = 0$ against the alternative $\beta_{gj} \neq 0$.

The Wald test is calculated using the formula

$$z_{gj} = \frac{b_{gj}}{s_{b_{gj}}}$$

The distribution of the Wald statistic is closely approximated by the normal distribution in large samples. However, in small samples, the normal approximation may be poor. For small samples, the deviance tests should be used instead to test significance since they perform better.

One problem that occurs in multiple-group logistic regression is that the test may be significant for the regression coefficient associated with one category, but not for the same coefficient associated with another category. In this case, we recommend that the independent variable be kept in the model if it is significant in at least one of the *G*-1 regression equations.

Wald P-Value

This is the significance level of the Wald test. If this value is less than some predefined alpha level, say 0.05, the variable is said to be statistically significant. Otherwise, the variable is not significant.

Odds Ratio Exp(b(i))

This is the estimated odds ratio associated with this regression coefficient. It is only useful for binary independent variables in which the two values are zero and one. These are the values that are generated for categorical independent variables. The formula used is

$$OR = e^b$$

Because of formatting limitations, the value is not displayed if it is larger than 10000.

Coefficient Confidence Intervals

Coefficient Confidence Intervals

Independent Variable X	Regression Coefficient b(i)	Standard Error Sb(i)	Lower 95% Confidence Limit	Upper 95% Confidence Limit	Odds Ratio Exp(b(i))
Intercept	68.32696	56.88604	-43.16763	179.82155	10000+
Cell	9.65213	7.75107	-5.53968	24.84394	10000+
LI	3.86710	1.77828	0.38174	7.35245	47.80336
Temp	-82.07365	61.71233	-203.02760	38.88029	0.00000

This report gives the estimated logistic regression equation and associated confidence limits. The reference value of the dependent variable is shown in parentheses in the page title. If the dependent variable has more than two outcomes, the information is displayed for each of the *G*-1 equations.

Independent Variable X

This is the independent variable that is displayed on this line. If the variable is continuous, it is displayed directly. If the variable is discrete, the definition of the binary variable that was generated is given. For example, suppose that a discrete independent GRADE variable has three values: A, B, and C. The name shown here would be something like *GRADE=B*. This refers to a binary variable that is one for those rows in which GRADE was B and zero otherwise.

Note that the placement of the name is controlled by the *Stagger label and output* option of the Report Options tab.

Regression Coefficient b(i)

This is the estimated value of the regression coefficient, sometimes referred to as B or Beta. The interpretation of the regression coefficients is difficult. We refer you to the discussion given at that beginning of this chapter for more details.

Standard Error Sb(i)

This is s_{b_j}, the large-sample estimate of the standard error of the regression coefficient. This is an estimate of the precision of the regression coefficient. It is used as the denominator of the Wald test.

Confidence Limits

These are the lower and upper confidences limits for β_{gj} based on the Wald statistic. These confidence limits are use the formula

$$b_{gj} \pm z_{1-\alpha/2} s_{b_{gj}}$$

Since they are based on the Wald test, they are only valid for large samples.

Odds Ratio Exp(b(i))

This is the estimated odds ratio associated with this regression coefficient. It is only useful for binary independent variables in which the two values are zero and one. These are the values that are generated for categorical independent variables. The formula used is

$$OR = e^{b}$$

Because of formatting limitations, the value is not displayed if it is larger than 10000.

Odds Ratios

Odds Ratios

Independent Variable X	Regression Coefficient b(i)	Odds Ratio Exp(b(i))	Lower 95% Confidence Limit	Upper 95% Confidence Limit
Intercept	68.32696	10000+	0.00000	10000+
Cell	9.65213	10000+	0.00393	10000+
LI	3.86710	47.80336	1.46483	1560.01770
Temp	-82.07365	0.00000	0.00000	10000+

This report presents estimates of the odds ratios and associated confidence limits associated with each variable in the model.

Independent Variable X

This is the independent variable that is displayed on this line. If the variable is continuous, it is displayed directly. If the variable is discrete, the definition of the binary variable that was generated is given. For example, suppose that a discrete independent GRADE variable has three values: A, B, and C. The name shown here would be something like *GRADE=B*. This refers to a binary variable that is one for those rows in which GRADE was B and zero otherwise.

Note that the placement of the name is controlled by the *Stagger label and output* option of the Report Options tab.

This is the estimated value of the corresponding regression coefficient, sometimes referred to as B or Beta. The interpretation of the regression coefficients is difficult. We refer you to the discussion given at that beginning of this chapter for more details.

Odds Ratio Exp(b(i))

This is the estimated odds ratio associated with this regression coefficient. It is only useful for binary independent variables in which the two values are zero and one. These are the values that are generated for categorical independent variables. The formula used is

$$OR = e^{b}$$

Because of formatting limitations, the value is not displayed if it is larger than 10000.

Confidence Limits

The lower and upper confidence limits yield an interval estimate of the odds ratio. The confidence coefficient is one minus alpha. Thus, when alpha is 0.05, the confidence coefficient is 0.95 or 95%. The formula used is

$$e^{(b_i \pm z_{1-\alpha/2} S_{b_i})}$$

Since these confidence limits are based on Wald statistics, they are only valid for large samples.

Estimated Logistic Regression Model(s)

Estimated Logistic Regression Model(s) in Reading Form

Model for Logit(Remiss) = XB when Remiss = 1
68.33 + 9.65 * Cell + 3.87 * LI - 82.07 * Temp

Each model estimates XB (where Logit(Y) = XB) for a specific Y outcome. To calculate the Y-value probabilities when there are only 2 outcomes, transform the logit using Prob(Y = outcome) = 1/(1+Exp(-XB)) or Prob(Y ≠ outcome) = Exp(-XB)/(1+Exp(-XB)). For the calculation formula to use when there are more than 2 outcomes, see the help documentation.

Estimated Logistic Regression Model(s) in Transformation Form

Model for Logit(Remiss) = XB when Remiss = 1
68.3269603055094 + 9.65212973758021*Cell + 3.86709587172725*LI -82.0736535775649*Temp

Each model estimates XB (where Logit(Y) = XB) for a specific Y outcome. To calculate the Y-value probabilities when there are only 2 outcomes, transform the logit using Prob(Y = outcome) = 1/(1+Exp(-XB)) or Prob(Y ≠ outcome) = Exp(-XB)/(1+Exp(-XB)). For the calculation formula to use when there are more than 2 outcomes, see the help documentation.

This report gives the logistic regression model in a regular text format that can be used as a transformation formula. A separate model is displayed for each of the *G*-1 categories of the dependent variable. The regression coefficients are displayed in double precision because a single-precision formula does not include the accuracy necessary to calculate the scores (logits) and predicted probabilities.

Note that a transformation must be less than 255 characters. Since these formulas are often greater than 255 characters in length, you must use the FILE(filename) transformation. To do so, copy the formula to a text file using Notepad, Windows Write, or Word to receive the model text. Be sure to save the file as an unformatted text (ASCII) file. The transformation is FILE(filename) where *filename* is the name of the text file, including directory information. When the transformation is executed, it will load the file and use the transformation stored there.

Analysis of Deviance

Analysis of Deviance

Term Omitted	DF	Deviance	Increase From Model Deviance (Chi²)	P-Value
All	3	34.37177	12.41839	0.00608
Cell	1	24.64782	2.69445	0.10070
LI	1	30.82856	8.87518	0.00289
Temp	1	24.34072	2.38734	0.12232
None(Model)	3	21.95337		

This report is the logistic regression analog of the analysis of variance table. It displays the results of a chi-square test used to test whether each of the individual terms in the regression are statistically significant after adjusting for all other terms in the model.

This report is not produced during a subset selection run.

Note that this report requires that a separate logistic regression be run for each line. Thus, if the running time is too long, you might consider omitting this report.

Term Omitted

This is the model term that is being tested. The test is formed by comparing the deviance statistic when the term is removed with the deviance of the complete model. Thus, the deviance when the term is left out of the model is shown.

The "All" line refers to the intercept-only model. This line tests the significance of the full model. The "None(Model)" refers to the complete model with no terms removed.

Note that it is usually not advisable to include an interaction term in a model when one of the associated main effects is missing—which is what happens here. However, in this case, we believe this to be a useful test.

Note that the name may become very long, especially for interaction terms. These long names may misalign the report. You can force the rest of the items to be printed on the next line by using the *Stagger label and output* option in the Report Options tab. This should create a better looking report when the names are extra long.

DF

This is the degrees of freedom of the chi-square test displayed on this line. DF is equal to $(G\text{-}1)DFt$ where DFt is the degrees of freedom of the term.

Deviance

The deviance is equal to minus two times the log likelihood achieved by the model being described on this line of the report. See the discussion given earlier in this chapter for a technical discussion of the deviance. A useful way to interpret the deviance is as the analog of the residual sum of squares in multiple regression. This value is used to create the difference in deviance that is used in the chi-square test.

Increase From Model Deviance (Chi^2)

This is the difference between the deviance for the model described on this line and the deviance of the complete model. This value follows the chi-square distribution in medium to large samples. See the discussion given earlier in this chapter for a technical discussion of this value. This value can be thought of as the analog of the residual sum of squares in multiple regression. Thus, you can think of this value as the increase in the residual sum of squares that occurs when this term is removed from the model.

Another way to interpret this test is as a redundancy test because it tests whether this term is redundant after considering all of the other terms in the model.

Note that the first line gives a test for the whole model.

P-Value

This is the significance level of the chi-square test. This is the probability that a chi-square value with degrees of freedom DF is equal to this value or greater. If this value is less than 0.05 (or other appropriate value), the term is said to be statistically significant.

Log Likelihood & R^2

Log Likelihood & R²

Term(s) Omitted	DF	Log Likelihood	R² of Remaining Term(s)	Reduction From Model R²	Reduction From Saturated R²
All	1	-17.18588	0.00000		
Cell	1	-12.32391	0.28290	0.07839	0.71710
LI	1	-15.41428	0.10308	0.25821	0.89692
Temp	1	-12.17036	0.29184	0.06946	0.70816
None(Model)	3	-10.97669	0.36130	0.00000	0.63870
None(Saturated)	28	0.00000	1.00000		0.00000

This report provides the log likelihoods and R^2 values of various models. This report is not produced during a subset selection run.

Note that this report requires that a separate logistic regression be run for each line. Thus, if the running time is too long, you might consider omitting this report.

Term Omitted

This is the term that is omitted from the model. The "All" line refers to the intercept-only model. The "None(Model)" refers to the complete model with no terms removed. The "None(Saturated)" line gives the results for the saturated model.

Note that the name may become very long, especially for interaction terms. These long names may misalign the report. You can force the rest of the items to be printed on the next line by using the *Stagger label and output* option in the Format tab. This should create a better looking report when the names are extra long.

DF

This is the degrees of freedom of the term displayed on this line. DF is equal to $(G\text{-}1)DFt$ where DFt is the degrees of freedom of the term.

Log Likelihood

This is the log likelihood of the model displayed on this line. Note that this is the log likelihood of the logistic regression without the term listed.

R^2 of Remaining Term(s)

This is the R^2 of the model displayed on this line, R_L^2. Note that the model does not include the term listed at the beginning of the line.

This R^2 is analogous to the R^2 in multiple regression, but it is not the same. This value is discussed in detail under the heading R^2 above. Refer to that section for more details about this statistic. We repeat the summary of the interpretation of R^2 in logistic regression.

Hosmer and Lemeshow (1989) recommend against the use R_L^2 as a goodness of fit measure. However, we have included it in our output because it does provide a comparative measure of the proportion of the log likelihood that is accounted for by the model. Just remember than an R_L^2 value of 1.0 indicates that the logistic regression model achieves the same log likelihood as the saturated model. However, this does not mean that it fits the data perfectly. Instead, it means that it fits the data as well as could be hoped for.

Reduction From Model R^2

This is amount that R^2 is reduced when the term is omitted from the regression model. This reduction is calculated from the R^2 achieved by the full model.

This quantity is used to determine if removing a term causes a large reduction in R^2. If it does not, then the term can be safely removed from the model.

Reduction From Saturated R^2

This is the amount that R^2 is reduced when the term is omitted from the regression model. This reduction is calculated from the R^2 achieved by the saturated model. This item is included because it shows how removal of this term impacts the best *R*-squared that is possible.

Classification Table

Classification Table

	Estimated		
Actual	**0**	**1**	**Total**
0	15	3	18
1	2	7	9
Total	17	10	27

Percent Correctly Classified = 81.5%

This table displays the results of classifying the data based on the logistic regression equations. The table presents the counts for each outcome.

The Percent Correctly Classified is also presented. This is the percent of the total count that fall on the diagonal of the table.

ROC Report

ROC Report for Value 0

Prob Cutoff	N(1|1) A	N(1|0) B	N(0|1) C	N(0|0) D	Sensitivity A/(A+C)	Specificity D/(B+D)	Sensitivity +Specificity	Proportion Correct
0.03333	18	9	0	0	1.00000	0.00000	1.00000	0.66667
0.06667	18	8	0	1	1.00000	0.11111	1.11111	0.70370
0.10000	17	8	1	1	0.94444	0.11111	1.05556	0.66667
0.13333	17	8	1	1	0.94444	0.11111	1.05556	0.66667
0.16667	17	6	1	3	0.94444	0.33333	1.27778	0.74074
0.20000	17	5	1	4	0.94444	0.44444	1.38889	0.77778
0.23333	17	4	1	5	0.94444	0.55556	1.50000	0.81481
0.26667	16	4	2	5	0.88889	0.55556	1.44444	0.77778
0.30000	15	3	3	6	0.83333	0.66667	1.50000	0.77778
0.33333	15	3	3	6	0.83333	0.66667	1.50000	0.77778
0.36667	15	3	3	6	0.83333	0.66667	1.50000	0.77778
0.40000	15	2	3	7	0.83333	0.77778	1.61111	0.81481
0.43333	15	2	3	7	0.83333	0.77778	1.61111	0.81481
0.46667	15	2	3	7	0.83333	0.77778	1.61111	0.81481
0.50000	15	2	3	7	0.83333	0.77778	1.61111	0.81481
0.53333	15	1	3	8	0.83333	0.88889	1.72222	0.85185
0.56667	13	1	5	8	0.72222	0.88889	1.61111	0.77778
0.60000	12	0	6	9	0.66667	1.00000	1.66667	0.77778
0.63333	12	0	6	9	0.66667	1.00000	1.66667	0.77778
0.66667	11	0	7	9	0.61111	1.00000	1.61111	0.74074
0.70000	11	0	7	9	0.61111	1.00000	1.61111	0.74074
0.73333	9	0	9	9	0.50000	1.00000	1.50000	0.66667
0.76667	9	0	9	9	0.50000	1.00000	1.50000	0.66667
0.80000	9	0	9	9	0.50000	1.00000	1.50000	0.66667
0.83333	8	0	10	9	0.44444	1.00000	1.44444	0.62963
0.86667	7	0	11	9	0.38889	1.00000	1.38889	0.59259
0.90000	7	0	11	9	0.38889	1.00000	1.38889	0.59259
0.93333	7	0	11	9	0.38889	1.00000	1.38889	0.59259
0.96667	7	0	11	9	0.38889	1.00000	1.38889	0.59259

Area Under ROC Curve = 0.88889 SE(AUC) = 0.07730 LCL(AUC) = 0.60103 UCL(AUC) = 0.97261

A separate ROC report is generated for each outcome. Only the report for outcome 0 is displayed here. ROC curves can be used to determine appropriate cutoff values for classification by letting you compare the sensitivity and specificity of various cutoff values. When classifying, you usually classify a row into that category that has the highest membership probability. However, this is not always the optimum strategy. This table shows you what happens when various cutoff values are selected.

Classifying an observation can have any one of four possible results. An observation from the outcome-group can be correctly classified as being from that outcome-group (state A) or incorrectly classified as being from another

outcome-group (state C). An observation from another outcome-group can be incorrectly classified as being from the outcome-group (state B) or correctly classified as being from another outcome-group (state D).

The number of observations in each state is computed for each cutoff value between zero and one. A number of measures can be calculated from these values. The measures used in ROC analysis are called *sensitivity* and *specificity*. Sensitivity is the proportion of those from this group that are correctly identified as such. In terms of the four states, sensitivity = *A/(A+C).* Specificity is the proportion of those from other groups that are correctly identified as such. In terms of four states, specificity = *D/(B+D).* Thus, the optimum cutoff value is that one for which the sum of sensitivity and specificity is the maximum. This may be found be investigating the report. In this example, the cutoff is between 0.40000 and 0.50000. An ROC plot is also generated for each report that gives a graphical display of this report.

An ROC analysis is most useful in the two-outcome case. In the multiple-outcome case, it is of only marginal usefulness, since a cutoff value is not specified. Rather, each observation is classified into that outcome-group which has the highest membership probability.

Prob Cutoff

This is the probability cutoff for classification into this outcome. If an observation's predicted probability for membership in this outcome is greater than this amount, the observation is classified in this outcome. Otherwise, it is classified as being in some other outcome.

A B C D

The counts for each of the four states. These counts are represented using the notation *N(i|j)* where *i* is the classified outcome and *j* is the actual outcome.

Sensitivity

Sensitivity is the proportion of those from this outcome that are correctly identified as such. In terms of the four states, sensitivity = *A/(A+C).*

Specificity

Specificity is the proportion of those from other outcomes that are correctly identified as such. In terms of four states, specificity = *D/(B+D).*

Sensitivity + Specificity

A common rule for selecting an appropriate cutoff value is to choose the cutoff with the largest total of sensitivity and specificity. This column allows you to do this very quickly.

Proportion Correct

Another rule for selecting an appropriate cutoff value is to choose that cutoff which maximizes the number of observations that are correctly classified. This column of the report allows you to quickly find the optimum cutoff value. Unfortunately, when one outcome has many more rows than the others, this rule may not be useful since it will lead you to classify everyone into the most prevalent outcome.

Area Under ROC Curve

The area under the ROC curve is a popular measure associated with ROC curves. When applied to classification in logistic regression, its maximum value of one occurs when all rows are correctly classified. Its minimum value of zero occurs when all rows are incorrectly classified. Thus, the nearer this value is to one, the better the classification.

The AUC's value depends on the number selected in the *Number Cutoffs* option on the Plots tab.

The calculation of the standard error of AUC and its confidence interval proceeds as given by Hanley and McNeil (1982). Let *AUC* denote the sample AUC value. For large samples, the distribution of AUC is approximately normal. Hence, a 100(1 – α)% confidence interval for AUC may be computed using the standard normal distribution as follows

$$AUC \pm z_{\alpha/2} SE(AUC)$$

The formula for SE(AUC) is

$$SE(AUC) = \sqrt{\frac{AUC(1-AUC) + (N_1 - 1)(Q_1 - AUC^2) + (N_2 - 1)(Q_2 - AUC^2)}{N_1 N_2}}$$

where

$$Q_1 = \frac{AUC}{2 - AUC}$$

$$Q_2 = \frac{2AUC^2}{1 + AUC}$$

Once estimates of AUC and SE(AUC) are calculated, hypothesis tests and confidence intervals can be calculated using standard methods. However, following the advice of Zhou et al. (2002) page 125, we use the following transformation which results in statistics that are closer to normality and ensures confidence limits that are inside the zero-one range. The transformation is

$$\hat{\Psi} = \ln\left(\frac{1 + \hat{A}}{1 - \hat{A}}\right)$$

The variance of $\hat{\psi}$ is estimated using

$$\mathrm{V}(\hat{\psi}) = \frac{4}{\left(1 - \hat{A}^2\right)^2} \mathrm{V}(\hat{A})$$

An $100(1-\alpha)\%$ confidence interval for ψ may then be constructed as

$$L, U = \hat{\psi} \mp \mathrm{z}_{1-\alpha/2} \sqrt{\mathrm{V}(\hat{\psi})}$$

Using the inverse transformation, the confidence interval for *A* is given by the two limits

$$\frac{1 - e^{-L}}{1 + e^{-L}} \quad \text{and} \quad \frac{1 - e^{-U}}{1 + e^{-U}}$$

Row Classification Report

Row Classification Report

Row	Actual Remiss	Estimated Remiss	Estimated Remiss Probability	Lower 95% Confidence Limit	Upper 95% Confidence Limit
1	1	1	0.83900	0.31617	0.98326
2	1	1	0.73317	0.48928	0.88739
3	0	0	0.81061	0.24565	0.98253
4	0	0	0.55936	0.24511	0.83230
5	1	1	0.83326	0.44347	0.96908
6	0	0	0.57370	0.21384	0.86943
7*	1	0	0.51337	0.32143	0.70145
8*	0	1	0.75562	0.21175	0.97267
9	0	0	0.71480	0.31903	0.93059
10	0	0	0.99687	0.19043	1.00000
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

This report displays the actual and predicted group and membership probability for each row of the report. It also provides confidence limits for the predicted group-membership probability.

Row

This is the row from the database. Rows that are starred are misclassified.

Actual Y

This is the outcome to which this row belongs (if known).

Estimated Y

This is the outcome with the largest membership probability.

Estimated Probability

This is the estimated probability that the row belongs to the outcome listed in the Estimated Y column.

These values allow you to determine how certain the classification is. When the value is near one (above 0.7), the logistic regression is convinced that the observation belongs in the designated group. When the value is near 0.5 or less, the classification was not as clear.

Lower and Upper Confidence Limits

These values provide a confidence interval for the estimated membership probability. Note that this confidence interval is only approximate in the multiple-outcome case. Formulas and technical details are given above in the section entitled Predicted Probabilities.

Row Classification Probabilities

Row Classification Probabilities

Row	Actual Remiss	Estimated Prob. in 0	Estimated Prob. in 1
1	1	0.16100	0.83900
2	1	0.26683	0.73317
3	0	0.81061	0.18939
4	0	0.55936	0.44064
5	1	0.16674	0.83326
6	0	0.57370	0.42630
7*	1	0.51337	0.48663
8*	0	0.24438	0.75562
9	0	0.71480	0.28520
10	0	0.99687	0.00313
.	.	.	.
.	.	.	.
.	.	.	.

This report displays the actual group and the membership probabilities for each group and each row. This allows you investigate how certain each classification is.

Row

This is the row from the database. Rows that are starred are misclassified.

Actual Y

This is the outcome to which this row belongs (if known).

Estimated Prob. In Y

This is the estimated probability that the row belongs in each outcome. These values allow you to determine how certain the classification is.

Simple Residual Report

Simple Residual Report

Row	Actual Remiss	Residual for Outcome 0	Residual for Outcome 1
1	1	-0.16100	0.16100
2	1	-0.26683	0.26683
3	0	0.18939	-0.18939
4	0	0.44064	-0.44064
5	1	-0.16674	0.16674
6	0	0.42630	-0.42630
7*	1	-0.51337	0.51337
8*	0	0.75562	-0.75562
9	0	0.28520	-0.28520
10	0	0.00313	-0.00313
.	.	.	.
.	.	.	.
.	.	.	.

This report displays the simple residuals for each group. Each of the g logistic regression equations can be used to estimate the probabilities that each observation belongs to the corresponding group.

Row

This is the row from the database. Rows that are starred are misclassified.

Actual Y

This is the outcome to which this row belongs (if known).

Residual for Group

These residuals are defined as

$$r_{gj} = y_{gj} - p_{gj}$$

where p_{gj} is the estimated membership probability and y_{gj} is an indicator variable that is one if the actual group is g and zero otherwise.

Note that, unlike multiple regression, there are g residuals for each observation instead of just one. This makes residual analysis much more difficult. If the logistic regression model fits an observation closely, all of its residuals will be small, but never zero.

Unfortunately, the simple residuals have unequal variance equal to $n_j \pi_{gj}\left(1-\pi_{gj}\right)$, where n_j is the number of observations with the same values of the independent variables as observation j. This unequal variance makes comparisons among the simple residuals difficult and alternative types of residuals are necessary.

Residual Report

Residual Report

Row	Actual Remiss	Pearson Residual		Deviance Residual		Maximum Hat Diagonal	
1	1	0.43806	|.....	0.59253	|||.....	0.20631	||||.....
2	1	0.60328	||.....	0.78789	|||||.....	0.05654	|.....
3	0	-0.48336	|.....	-0.64802	||||.....	0.26518	||||||.....
4	0	-1.25520	|||||.....	-1.52442	||||||||.....	0.23855	|||||.....
5	1	0.44733	|.....	0.60400	|||.....	0.12192	||.....
6	0	-0.86201	|||.....	-1.05417	|||||.....	0.16277	|||.....
7*	1	1.02710	||||.....	1.20021	||||||.....	0.04169	|.....
8*	0	-1.75843	|||||||.....	-1.67872	|||||||||.....	0.28695	||||||.....
9	0	-0.63166	||.....	-0.81945	|||||.....	0.14925	|||.....
10	0	-0.05607	|.....	-0.07923	|.....	0.04227	|.....
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

This report displays the Pearson residuals, the deviance residuals, and the hat diagonal for each row. These are the residuals that most textbooks on logistic regression recommend that you use.

Row

This is the row from the database. Rows that are starred are misclassified.

Actual Y

This is the outcome to which this row belongs (if known).

Pearson Residual

The *Pearson residuals* give the contribution of each row to the Pearson chi-square goodness of fit statistic. When the values of the independent variables of each observation are unique, the formula for this residual is

$$\chi_j = \pm\sqrt{\sum_{g=1}^{G}\frac{\left(w_{gj} - n_j p_{gj}\right)^2}{n_j p_{gj}}}, \quad j = 1,2,\cdots,J$$

where the plus (minus) is used if w_{gj} / n_j is greater (less) than p_{gj}. By definition, the sum of the squared Pearson residuals is the Pearson chi-square goodness of fit statistics.

Deviance Residuals

Remember that the deviance is -2 times the difference between log likelihoods of a reduced model and the saturated model. The formula for a deviance residual is

$$d_j = \pm\sqrt{2\sum_{g=1}^{G} w_{gj}\ln\left(\frac{w_{gj}}{n_j p_{gj}}\right)}, \quad j = 1,2,\cdots,J$$

where the plus (minus) is used if $w_{REF(g),j} / n_j$ is greater (less) than $p_{REF(g),j}$. By definition, the sum of the squared deviance residuals is the deviance.

Maximum Hat Diagonal

The diagonal elements of the hat matrix can be used to detect points that are extreme in the independent variable space. These are often called *leverage* design points. The larger the value of the hat diagonal, the more the observation influences estimates of the regression coefficients. There is a separate hat diagonal defined for each category. The value reported here is the maximum of all *G* of the hat diagonals for each row.

An observation that has a large residual, but has low leverage, does not cause much concern. However, an observation with a large leverage and a large residual should be checked very carefully. The formula for the hat diagonal associated with the *j*th observation and *g*th outcome is

$$h_{gj} = n_j p_{gj}\left(1 - p_{gj}\right)\sum_{i=1}^{p}\sum_{k=1}^{p} X_{ij} X_{kj} \hat{V}_{gik}, \quad j = 1,2,\cdots,J$$

where $\hat{V}_{gik}$ is the portion of the covariance matrix of the regression coefficients associated with the *g*th regression equation. The interpretation of this diagnostic is not as clear in logistic regression as in multiple regression because it involves the predicted values which in turn involve the dependent variable. In multiple regression, the hat diagonals only involve the independent variables.

Note that this formula matches Pregibon (1981) in the two-outcome case. In the multiple-outcome case, the two-outcome formula is applied to each outcome.

DFBetas Report

DFBetas Report For Remiss = 1

Row	Actual Remiss	DFBeta Intercept		DFBeta Cell		DFBeta LI	
1	1	0.05383	|.....	-0.11561	|.....	0.12403	|.....
2	1	0.06191	|.....	-0.03603	|.....	0.07986	|.....
3	0	0.03248	|.....	0.29680	|||.....	0.19367	||.....
4	0	-0.07853	|.....	-0.22408	||.....	0.36761	|||||.....
5	1	0.15954	||.....	0.02455	|.....	0.11640	|.....
6	0	-0.10146	|.....	-0.11173	|.....	0.16597	||.....
7*	1	-0.05201	|.....	-0.05264	|.....	-0.12518	|.....
8*	0	0.83713	||||||||||||||	0.10576	|.....	-0.19110	||.....
9	0	0.20605	|||.....	0.03081	|.....	0.23153	|||.....
10	0	0.01139	|.....	0.00613	|.....	0.01005	|.....
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

One way to study the impact of an observation on each regression coefficient is to determine how much that coefficient changes when the observation is deleted. The DFBETA statistic is the standardized difference between a regression coefficient before and after the removal of the *j*th observation.

Row

This is the row from the database. Rows that are starred are misclassified.

Actual Y

This is the outcome to which this row belongs (if known).

DFBeta

The DFBeta statistic is the standardized difference between a regression coefficient before and after the removal of the *j*th observation.

The formula for DFBeta is approximated by

$$\text{DFBeta}_{gij} = \left(\frac{w_{gj} - n_j p_{gj}}{\left(1 - h_{gj}\right)\sqrt{\hat{V}_{gii}}} \right) \sum_{k=1}^{p} X_{kj} \hat{V}_{gik}, \quad j = 1,2,\cdots,J$$

where $\hat{V}_{gik}$ is the portion of the covariance matrix associated with the *g*th regression equation. Note that this formula matches Pregibon (1981) in the two outcome case, but is different from Lesaffre (1989) in the multi-outcome case.

Influence Diagnostics Report

Influence Diagnostics Report For Remiss = 1

Row	Actual Remiss	Hat Diagonal		Cook's Distance (C)		Cook's Distance (CBar)	
1	1	0.20631	||||.....	0.06285	|.....	0.04988	|.....
2	1	0.05654	|.....	0.02312	|.....	0.02181	|.....
3	0	0.26518	||||||.....	0.11474	|.....	0.08432	|.....
4	0	0.23855	|||||.....	0.64822	|||||.....	0.49359	|||||.....
5	1	0.12192	||.....	0.03164	|.....	0.02778	|.....
6	0	0.16277	|||.....	0.17254	|.....	0.14446	|.....
7*	1	0.04169	|.....	0.04790	|.....	0.04590	|.....
8*	0	0.28695	||||||.....	1.74508	||||||||||||.	1.24433	||||||||||||||
9	0	0.14925	|||.....	0.08228	|.....	0.07000	|.....
10	0	0.04227	|.....	0.00014	|.....	0.00014	|.....
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

This report gives two distance measures similar to Cook's distance in multiple regression.

Row

This is the row from the database. Rows that are starred are misclassified.

Actual Y

This is the outcome to which this row belongs (if known).

Hat Diagonal

The diagonal elements of the hat matrix can be used to detect points that are extreme in the independent variable space. They are discussed in more detail in the Residual Report.

Cook's Distance (C) and (CBar)

C and *Cbar* are extensions of Cooks distance for logistic regression. Quoting from Pregibon (1981), page 719:

"Cbar measures the overall change in fitted logits due to deleting the *l*th observation for all points excluding the one deleted. Conversely, *C* includes the deleted point. Although *C* will usually be the preferred diagnostic to measure overall coefficients' changes, in the examples examined to date, the one-step approximations were more accurate for *Cbar* than *C*."

The formulas for *C* and *Cbar* are

$$C_{gj} = \frac{\chi_j^2 h_{gj}}{\left(1 - h_{gj}\right)^2}, \quad j = 1,2,\cdots,J$$

$$\overline{C}_{gj} = \frac{\chi_j^2 h_{gj}}{\left(1 - h_{gj}\right)}, \quad j = 1,2,\cdots,J$$

Note that this formula matches Pregibon (1981) in the two-outcome case. In the multiple-outcome case, the two-outcome formula is applied to each outcome.

Residual Diagnostics Report

Residual Diagnostics Report For Remiss = 1

Row	Actual Remiss	Hat Diagonal		Deviance Change (DFDev)		Chi-Square Change (DFChi2)	
1	1	0.20631	||||.....	0.40098	|.....	0.24178	|.....
2	1	0.05654	|.....	0.64257	|.....	0.38576	|.....
3	0	0.26518	||||||.....	0.50425	|.....	0.31795	|.....
4	0	0.23855	|||||.....	2.81743	||||||.....	2.06910	||.....
5	1	0.12192	||.....	0.39260	|.....	0.22789	|.....
6	0	0.16277	|||.....	1.25574	||.....	0.88752	|.....
7*	1	0.04169	|.....	1.48639	|||.....	1.10084	|.....
8*	0	0.28695	||||||.....	4.06243	|||||||.....	4.33640	|||.....
9	0	0.14925	|||.....	0.74150	|.....	0.46899	|.....
10	0	0.04227	|.....	0.00642	|.....	0.00328	|.....
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

This report gives statistics that help detect observations that have not been fitted well by the model.

Row

This is the row from the database. Rows that are starred are misclassified.

Actual Y

This is the outcome to which this row belongs (if known).

Hat Diagonal

The diagonal elements of the hat matrix can be used to detect points that are extreme in the independent variable space. They are discussed in more detail in the Residual Report.

Deviance Change (DFDev) and Chi-Square Change (DFChi2)

DFDEV and *DFCHI2* are statistics that measure the change in deviance and in Pearson's chi-square, respectively, that occurs when an observation is deleted from the dataset. Large values of these statistics indicate observations that have not been fitted well.

The formulas for these statistics are

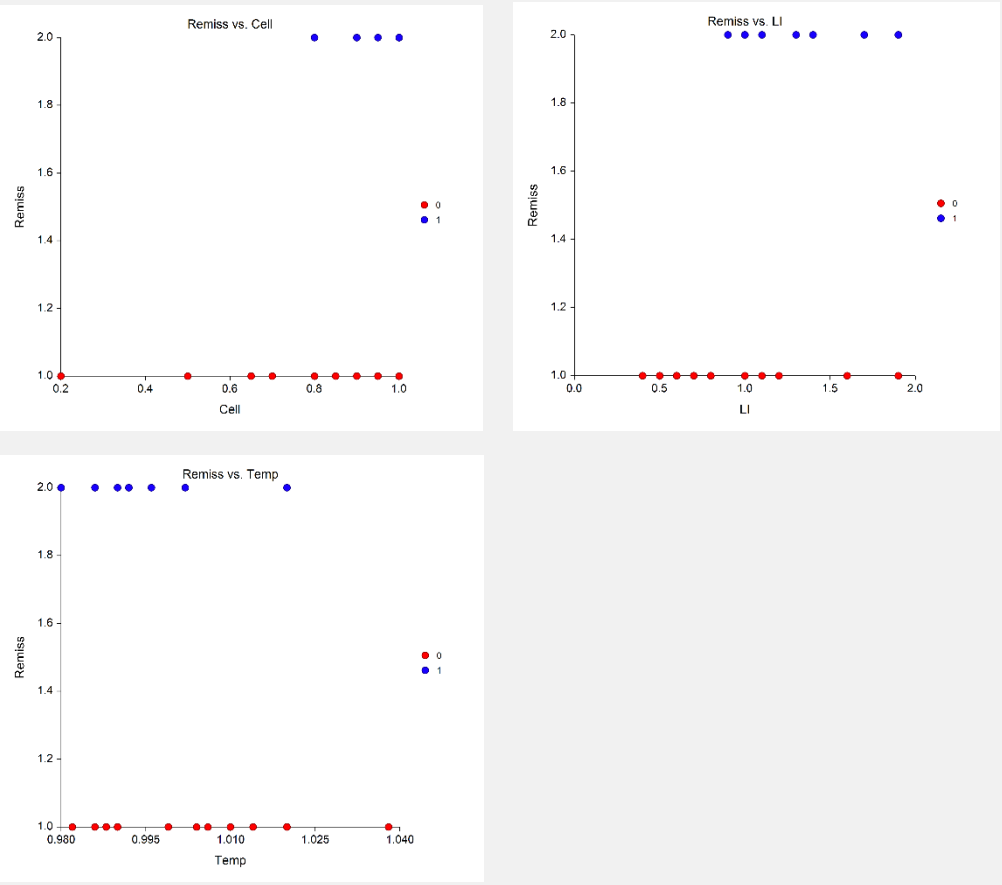
$$DFDEV_{gj} = d_j^2 + \overline{C}_{gj}, \quad j = 1,2,\cdots,J$$

$$DFCHI2_{gj} = \frac{\overline{C}_{gj}}{h_{gj}}, \quad j = 1,2,\cdots,J$$

Note that this formula matches Pregibon (1981) in the two-group case. In the multiple-group case, the two-group formula is applied to each group.

Y versus X Plots

Y vs X's Plots



This section shows scatter plots with the dependent variable on the vertical axis and each of the independent variables on the horizontal axis. The plot is useful for finding typos, outliers, and other anomalies in that data.

Vertical Axis

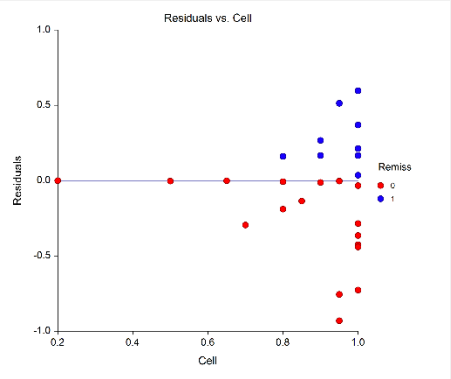
The categories of the dependent variable are shown on the vertical axis. Each category is assigned a whole number, beginning with the number one. The numbers are assigned in sorted order. Thus, if your dependent variable has values A, B, and C, it would be plotted on a numeric scale ranging from about 0.8 to 3.2. The groups would be plotted as the numbers 1, 2, and 3.

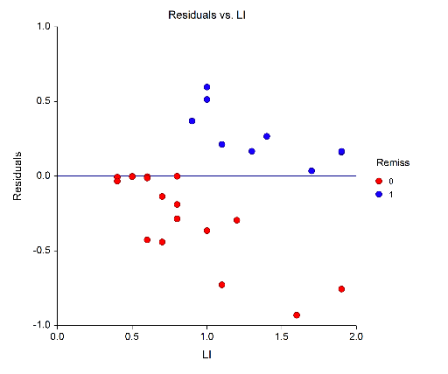
Horizontal Axis

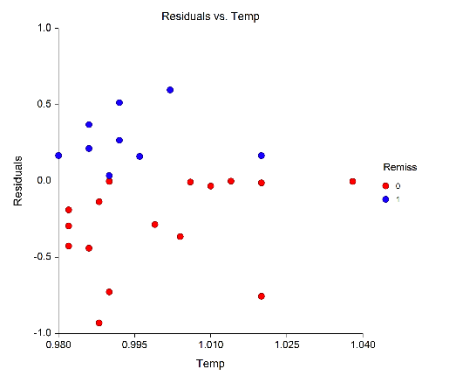
The independent variables are shown on the horizontal axis. When the independent variable is categorical, binary variables are generated for each of the categories and a separate scatter plot is generated for each binary variable.

Simple Residuals versus X Plots

Simple Residuals vs X's Plots







This section shows scatter plots with the simple residuals on the vertical axis and each of the independent variables on the horizontal axis. The plots are useful for finding outliers and other anomalies in the data.

Vertical Axis

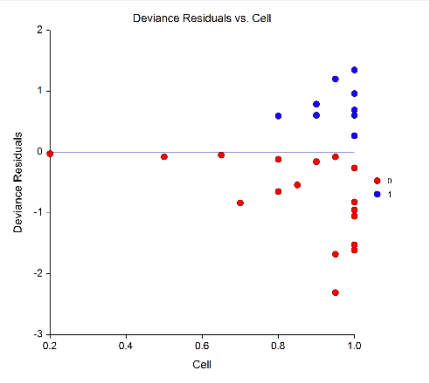
The residuals are displayed on the vertical axis. Note that the *G* residuals for each row corresponding to the simple residuals are displayed. Thus, if you have *N* rows, you will have *GN* points displayed on the plot.

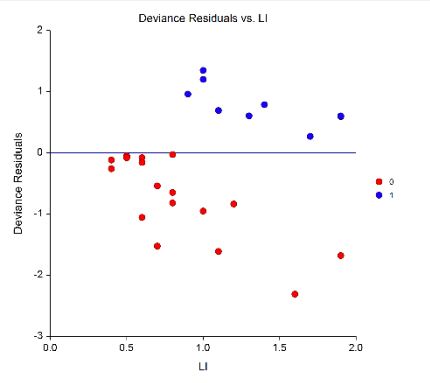
Horizontal Axis

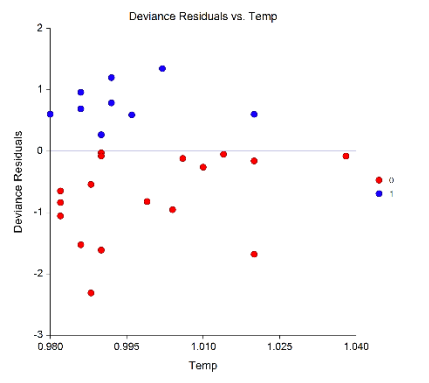
The independent variables are shown on the horizontal axis. When the independent variable is categorical, binary variables are generated for each of the categories and a separate scatter plot is generated for each binary variable.

Deviance Residuals versus X Plots

Deviance Residuals vs X's Plots







This section shows scatter plots with the deviance residuals on the vertical axis and each of the independent variables on the horizontal axis. The plots are useful for finding outliers and other anomalies in the data.

Vertical Axis

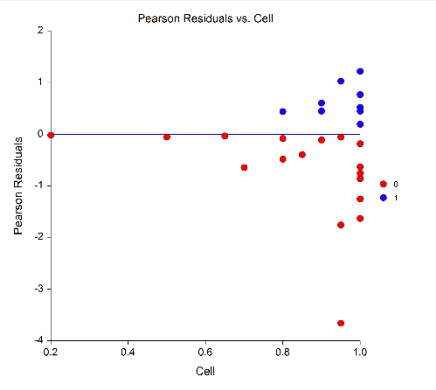
The deviance residuals are displayed on the vertical axis.

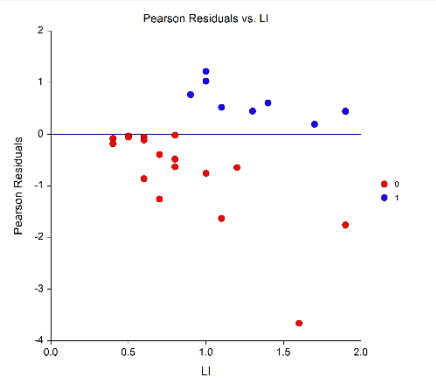
Horizontal Axis

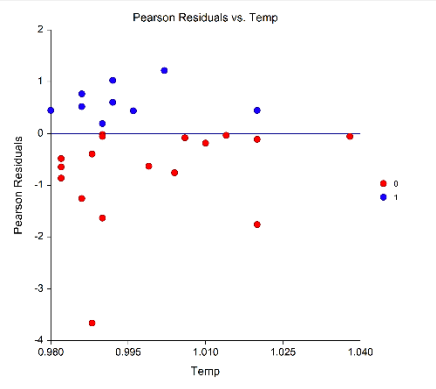
The independent variables are shown on the horizontal axis. When the independent variable is categorical, binary variables are generated for each of the categories and a separate scatter plot is generated for each binary variable.

Pearson Residuals versus X Plots

Pearson Residuals vs X's Plots







This section shows scatter plots with the Pearson residuals on the vertical axis and each of the independent variables on the horizontal axis. The plots are useful for finding outliers and other anomalies in the data.

Vertical Axis

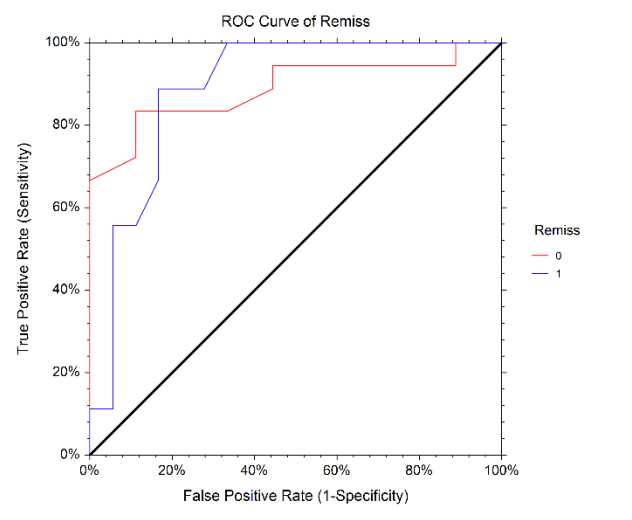
The Pearson residuals are displayed on the vertical axis.

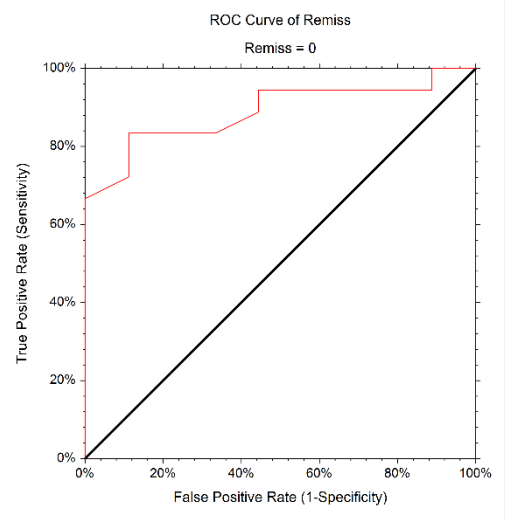
Horizontal Axis

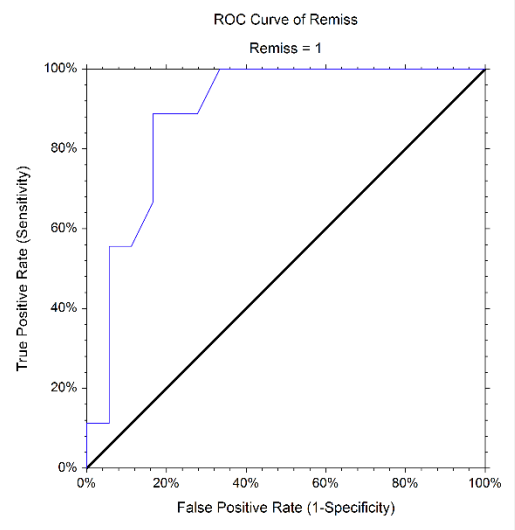
The independent variables are shown on the horizontal axis. When the independent variable is categorical, binary variables are generated for each of the categories and a separate scatter plot is generated for each binary variable.

ROC Curves - Combined and Separate

ROC Curves (Combined and Separate)







This section displays the ROC curves that can be used to help you find the best cutoff points to use for classification. The cutoff point nearest the top-left corner of the plot is the optimum cutoff. You will have to refer to the ROC Report to determine the exact value of the cutoff.

Vertical Axis

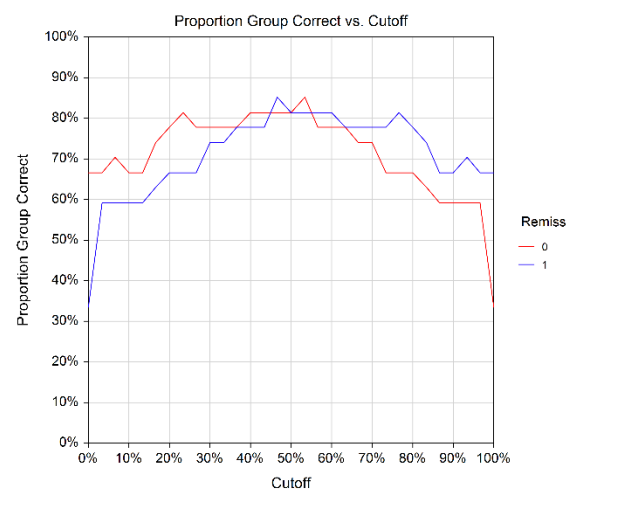
The sensitivity is displayed on the vertical axis.

Horizontal Axis

One minus the specificity is displayed on the horizontal axis.

Prob Correct versus Cutoff Plot

Prob Correct vs Cutoff Plot



This section displays a plot that shows the proportion correct versus the cutoff. It is useful to help determine the cutoff point used in classification. This plot may be difficult to use with three or more categories because of the ambiguity in the plot.

Vertical Axis

The proportion correctly classified for various cutoff values are displayed on the vertical axis.

Horizontal Axis

The cutoff values are displayed on the horizontal axis. These cutoff values are in terms of the estimated outcome-membership probabilities. Thus, a cutoff of 0.4 means that any rows with a outcome-membership probability of 0.4 or more are classified into this outcome.

Example 2 – Subset Selection

This section presents an example of how to conduct a subset selection. The data used are stored in the Leukemia dataset. This analysis will search for the best model from among a pool of the six numeric variables.

Setup

To run this example, complete the following steps:

1 Open the Leukemia example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Leukemia** and click **OK**.

2 Specify the Logistic Regression procedure options

- Find and open the **Logistic Regression** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 2a** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables, Model Tab	
Y	**Remiss**
Numeric X's	**Cell, Smear, Infil, LI, Blast, Temp**
Terms	**1-Way**
Subset Selection Tab	
Search for the Best Subset from the X's	**Checked**
Search Method	**Hierarchical Forward Selection**
Stop search when number of terms reaches	**6**
Reports Tab	
Run Summary	**Checked**
Subset Summary	**Checked**
Subset Detail	**Checked**
Coefficient Significance Tests	**Checked**
All Other Reports	**Unchecked**
Plots Tab	
All Plots	**Unchecked**
Report Options (*in the Toolbar*)	
Variable Labels	**Column Names**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Run Summary

Run Summary

Item	Value	Item	Value
Y Variable	Remiss	Rows Processed	29
Reference Value	0	Rows Used	27
Number of Y-Values	2	Rows for Validation	0
Frequency Variable	None	Rows X's Missing	2
Numeric X Variables	6	Rows Freq Miss. or 0	0
Categorical X Variables	0	Rows Prediction Only	0
Final Log Likelihood	-10.87752	Unique Rows (Y and X's)	27
Model R^2	0.36707	Sum of Frequencies	27
Actual Convergence	2.081623E-06	Likelihood Iterations	9
Target Convergence	1E-06	Maximum Iterations	20
Model D.F.	6	Completion Status	Quasi-Separation
Priors	Equal		
Subset Selection Method	Hierarchical Forward Selection		

******** WARNING ******** WARNING ******** WARNING ******** WARNING ******** WARNING ********
Your dataset had QUASI-COMPLETE SEPARATION which means that the maximum likelihood routine did NOT converge so the statistical tests are not valid. Although the prediction equations correctly classified much of your data, they may not do so for other observations. Quasi-Complete Separation often occurs because your sample size is too small.
******** WARNING ******** WARNING ******** WARNING ******** WARNING ******** WARNING ********

The first thing we notice is the warning message about quasi-separation. If quasi-separation occurs, the maximum likelihood estimates do not exist and all results are suspect. We note that 9 likelihood iterations occurred and the Actual Convergence is near the Target Convergence. We decide to rerun the analysis after resetting the Max Terms in Subset box from 6 to 5. Note that this error message often occurs when a small set of data is fit with a model with too many terms.

At this point, reset the value for **Stop search when number of terms reaches** (on the Subset Selection tab) to **5** manually or load the template **Example2b**. Now, rerun the analysis.

Run Summary

Run Summary

Item	Value	Item	Value
Y Variable	Remiss	Rows Processed	29
Reference Value	0	Rows Used	27
Number of Y-Values	2	Rows for Validation	0
Frequency Variable	None	Rows X's Missing	2
Numeric X Variables	6	Rows Freq Miss. or 0	0
Categorical X Variables	0	Rows Prediction Only	0
Final Log Likelihood	-10.92900	Unique Rows (Y and X's)	27
Model R^2	0.36407	Sum of Frequencies	27
Actual Convergence	7.136538E-07	Likelihood Iterations	7
Target Convergence	1E-06	Maximum Iterations	20
Model D.F.	5	Completion Status	Normal Completion
Priors	Equal		
Subset Selection Method	Hierarchical Forward Selection		

The warning message has disappeared and the algorithm finished normally.

Subset Selection Summary

Subset Selection Summary
Subset Selection Method = Hierarchical Forward Selection

No. Terms	No. X's	Log Likelihood	R^2 Value	R^2 Change
1	1	-17.18588	0.00000	0.00000
2	2	-13.03648	0.24144	0.24144
3	3	-12.17036	0.29184	0.05040
4	4	-10.97669	0.36130	0.06946
5	5	-10.92900	0.36407	0.00277

This report shows the best log-likelihood value for each subset size. In this example, it appears that four terms (the intercept and three variables) provides the best model. Note that adding the fifth variable does not increase the R-squared value very much.

No. Terms

The number of terms. Note that this includes the intercept.

No. X's

The number of *X*'s that were included in the model. Note that in this case, the number of terms matches the number of *X*'s. This would not be the case if some of the terms were categorical variables.

Log Likelihood

This is the value of the log likelihood function evaluated at the maximum likelihood estimates. Our goal is to find a subset size above which little is gained by adding more variables.

R^2 Value

This is the value of R^2 calculated using the formula

$$R_L^2 = \frac{L_p - L_0}{L_0 - L_S}$$

as discussed in the introduction. We are looking for the subset size at which this value does not increase by a meaningful amount.

R^2

This is the increase in R^2 that occurs when each new subset size is reached. Search for the subset size below which the R^2 value does not increase by more than 0.02 for small samples or 0.01 for large samples.

In this example, the optimum subset size appears to be four terms.

Subset Selection Detail

Subset Selection Detail
Subset Selection Method = Hierarchical Forward Selection

Step	Action	No. of Terms	No. of X's	Log Likelihood	Term Entered	Term Removed
1	Add	1	1	-17.18588	Intercept	
2	Add	2	2	-13.03648	LI	
3	Add	3	3	-12.17036	Cell	
4	Add	4	4	-10.97669	Temp	
5	Add	5	5	-10.92900	Smear	

This report shows the highest log likelihood for each subset size. In this example, it appears that four terms (the intercept and three variables) provide the best model. Note that adding the fifth variable does not increase the *R*-squared value very much.

Action

This item identifies the action that was taken at this step. A term was added, removed, or two were switched.

No. Terms

The number of terms. Note that this includes the intercept.

No. X's

The number of *X*'s that were included in the model. Note that in this case, the number of terms matches the number of *X*'s. This would not be the case if some of the terms were categorical variables.

Log Likelihood

This is the value of the log likelihood function after the completion of this step. Our goal is to find a subset size above which little is gained by adding more variables.

Terms Entered and Removed

These columns identify the terms added, removed, or switched.

Discussion of Example 2

After considering these reports, it was decided to include Cell, LI, and Temp in the final logistic regression model. Another run should now take place using only these independent variables. A complete residual analysis is necessary before the equation is finally adopted.

Example 3 – One Categorical X Variable

The independent variables in logistic regression may be categorical as well as numerical. This example is of the simplest categorical case of a binary response and a binary independent variable. More complicated examples will be shown below.

In this example, a simple yes-no question is asked of each member of two groups. The following two-by-two table presents the results. The analyst wants to understand the relationship between group membership and response to the question.

	Response		
Group	**Yes**	**No**	**Total**
A	91	9	100
B	93	27	120
Total	184	36	220

These data would normally be analyzed using the methods for comparing two proportions such as Fisher's exact test or the chi-square test for independence in a contingency table. The following table presents the results of this analysis.

Two Proportions Output

Counts and Proportions

	Response			
Group	**No Count**	**Yes Count**	**Total Count**	**Proportion***
A	9	91	100	p1 = 0.0900
B	27	93	120	p2 = 0.2250

* Proportion = No / Total

Proportions Analysis

Statistic	Value
Group 1 Event Rate (p1)	0.0900
Group 2 Event Rate (p2)	0.2250
Absolute Risk Difference |p1 - p2|	0.1350
Number Needed to Treat 1/|p1 - p2|	7.41
Relative Risk Reduction |p1 - p2|/p2	0.60
Relative Risk p1/p2	0.40
Odds Ratio o1/o2	0.34

Two-Sided Tests of the Difference (P1 - P2)

H0: P1 = P2 vs. Ha: P1 ≠ P2

Test Statistic Name	p1	p2	Difference p1 - p2	Test Statistic Value	Prob Level	Reject H0 at α = 0.05?
Wald Z	0.0900	0.2250	-0.1350	-2.695	0.0070	Yes
Fisher's Exact	0.0900	0.2250	-0.1350	0.010	0.0097	Yes

The conclusion of this analysis is to reject the null hypothesis that the two proportions are equal. The significance levels are 0.0097 using Fisher's exact test and 0.0070 using the normal approximation which is equivalent to the chi-square test for independence. Note that the odds ratio is 0.34.

We will now see how to analyze these data using logistic regression. The data must be entered into a database so that they can be processed. The following table shows how these data are rearranged and entered. These data have been entered into a database named 2BY2.

2By2 dataset (subset)

Group	Response	Count
A	No	9
A	Yes	91
B	No	27
B	Yes	93

Setup

To run this example, complete the following steps:

1 Open the 2By2 example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **2By2** and click **OK**.

2 Specify the Logistic Regression procedure options

- Find and open the **Logistic Regression** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 3** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables, Model Tab	
Y	**Response**
Categorical X's	**Group**
Default Recoding Scheme	**Binary**
Frequencies	**Count**
Priors	**Equal across Y Values**
Reports Tab	
Run Summary	**Checked**
Y Variable Summary	**Checked**
Coefficient Significance Tests	**Checked**
Odds Ratios	**Checked**
Analysis of Deviance	**Checked**
Log-Likelihood and R^2	**Checked**
All Other Reports	**Unchecked**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Logistic Regression Output

Run Summary

Item	Value	Item	Value
Y Variable	Response	Rows Processed	4
Reference Value	No	Rows Used	4
Number of Y-Values	2	Rows for Validation	0
Frequency Variable	Count	Rows X's Missing	0
Numeric X Variables	0	Rows Freq Miss. or 0	0
Categorical X Variables	1	Rows Prediction Only	0
Final Log Likelihood	-94.23344	Unique Rows (Y and X's)	4
Model R^2	0.06908	Sum of Frequencies	220
Actual Convergence	2.559022E-11	Likelihood Iterations	6
Target Convergence	1E-06	Maximum Iterations	20
Model D.F.	2	Completion Status	Normal Completion
Priors	Equal		

Y Variable Summary

Y Response	Count	Unique Rows (Y and X's)	Y Proportion	Y Prior	R^2 (Y vs Pred. Probability)	Percent Correctly Classified
No	36	2	0.16364	0.50000	0.03302	75.000
Yes	184	2	0.83636	0.50000	0.03302	49.457
Total	220	4				53.636

Coefficient Significance Tests

Independent Variable X	Regression Coefficient b(i)	Standard Error Sb(i)	Wald Z-Value H0: β=0	Wald P-Value	Odds Ratio Exp(b(i))
Intercept	0.68222	0.29814	2.288	0.02212	1.97826
(Group="B")	-1.07687	0.41218	-2.613	0.00898	0.34066

Odds Ratios

Independent Variable X	Regression Coefficient b(i)	Odds Ratio Exp(b(i))	Lower 95% Confidence Limit	Upper 95% Confidence Limit
Intercept	0.68222	1.97826	1.10282	3.54863
(Group="B")	-1.07687	0.34066	0.15187	0.76413

Analysis of Deviance

Term Omitted	DF	Deviance	Increase From Model Deviance (Chi²)	P-Value
All	1	196.08640	7.61951	0.00577
Group	1	196.08640	7.61951	0.00577
None(Model)	1	188.46689		

Log Likelihood & R^2

Term(s) Omitted	DF	Log Likelihood	R^2 of Remaining Term(s)	Reduction From Model R^2	Reduction From Saturated R^2
All	1	-98.04320	0.00000		
Group	1	-98.04320	0.00000	0.06908	1.00000
None(Model)	1	-94.23344	0.06908	0.00000	0.93092
None(Saturated)	4	-42.89226	1.00000		0.00000

Although a casual comparison between this report and that of the Two Proportion procedure shows little in common, a more detailed report shows many similarities. First of all, notice that the significance level of the test of GROUP in the Analysis of Deviance Section of 0.00577 compares very closely with the 0.007037 from the

chi-square test. Also notice that the odds ratios from both reports round to 0.34066. The confidence limits of these two reports are not exactly the same, but they are close.

To summarize the logistic regression analysis, we can conclude that there is a significant relationship between response and group.

This example has shown the similarities between these two approaches to the analysis of two proportions. Usually, you would analyze these data using the two proportions approach. However, that approach is not as easily extended to the case of several independent variables including a mixture of categorical and numeric.

Example 4 – Logit Model Validation with BMDP PR

This example will serve three purposes. First of all, it will be the first example of a dataset whose Y variable has more than two outcomes. Second, it will be an example of what the output looks like when all of the independent variables are categorical. And finally, it will validate the procedure by allowing the comparison of the **NCSS** output with that of the **BMDP PR** program which also performs multiple-group logistic regression. This example comes from the **BMDP** manual. The database containing the data used in this example is named NC Criminal

The NC Criminal dataset contains data that will be used to study the relationship between a cases verdict and three factors: race, county, and type of offense. The variables that are on the database are as follows.

Count contains the number of individuals with the characteristics specified on that row.

Verdict is the response variable. Three outcomes are given in the database: *G* for guilty, *NG* for not guilty, and *NP* for not prosecuted.

Race gives the race of the individual. It has two values: *A* and *B*.

County refers to county in North Carolina in which the offense was considered. The possible values are: *Durham* and *Orange*.

Offense contains the particular offense that the individual was accused of. These are *Drunk*, *Violence*, *Property*, *Major Traffic*, and *Speeding*.

You can view the data by loading the NC Criminal dataset, so they will not be displayed here.

Setup

To run this example, complete the following steps:

1 Open the NC Criminal example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **NC Criminal** and click **OK**.

2 Specify the Logistic Regression procedure options

- Find and open the **Logistic Regression** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 4** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables, Model Tab	
Y	**Verdict**
Reference Value	**NP**
Categorical X's	**Race(B;A) County(B;Durham) Offense(B;Drunk)**
Frequencies	**Count**
Priors	**Ni/N (Y-Value Proportions)**
Reports Tab	
Run Summary	**Checked**
Y Variable Summary	**Checked**
Coefficient Significance Tests	**Checked**
Analysis of Deviance	**Checked**
Log-Likelihood and R^2	**Checked**
All Other Reports	**Unchecked**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Logistic Regression Output

Run Summary

Item	Value	Item	Value
Y Variable	Verdict	Rows Processed	60
Reference Value	NP	Rows Used	57
Number of Y-Values	3	Rows for Validation	0
Frequency Variable	Count	Rows X's Missing	0
Numeric X Variables	0	Rows Freq Miss. or 0	3
Categorical X Variables	3	Rows Prediction Only	0
Final Log Likelihood	-408.29185	Unique Rows (Y and X's)	60
Model R^2	0.69779	Sum of Frequencies	615
Actual Convergence	4.751915E-11	Likelihood Iterations	6
Target Convergence	1E-06	Maximum Iterations	20
Model D.F.	14	Completion Status	Normal Completion
Priors	Ni/N		

Y Variable Summary

Y Verdict	Count	Unique Rows (Y and X's)	Y Proportion	Y Prior	R^2 (Y vs Pred. Probability)	Percent Correctly Classified
G	445	20	0.72358	0.72358	0.17107	93.933
NG	123	20	0.20000	0.20000	0.10397	20.325
NP	47	20	0.07642	0.07642	0.06628	0.000
Total	615	60				72.033

Coefficient Significance Tests

Independent Variable X	Regression Coefficient b(i)	Standard Error Sb(i)	Wald Z-Value H0: β=0	Wald P-Value	Odds Ratio Exp(b(i))
Intercept					
G	2.82983	0.44457	6.365	0.00000	16.94253
NG	1.24012	0.48781	2.542	0.01102	3.45604
(Race="B")					
G	0.26083	0.33984	0.767	0.44279	1.29800
NG	-0.10324	0.36248	-0.285	0.77579	0.90191
(County="Orange")					
G	-0.89593	0.33719	-2.657	0.00788	0.40823
NG	-0.12175	0.36036	-0.338	0.73547	0.88537
(Offense="MjTraffic")					
G	-0.21380	0.62893	-0.340	0.73390	0.80751
NG	0.48012	0.67038	0.716	0.47387	1.61627
(Offense="Property")					
G	-0.91853	0.57784	-1.590	0.11193	0.39911
NG	0.00928	0.61911	0.015	0.98804	1.00932
(Offense="Speed")					
G	0.49546	0.51245	0.967	0.33361	1.64126
NG	-0.26697	0.57599	-0.463	0.64301	0.76570
(Offense="Violence")					
G	-2.23014	0.51372	-4.341	0.00001	0.10751
NG	-0.57863	0.53748	-1.077	0.28168	0.56067

Logistic Regression

Analysis of Deviance

Term Omitted	DF	Deviance	Increase From Model Deviance (Chi²)	P-Value
All	12	925.59805	109.01434	0.00000
Race	2	819.21845	2.63475	0.26784
County	2	832.03780	15.45409	0.00044
Offense	8	898.18115	81.59744	0.00000
None(Model)	12	816.58371		

Log Likelihood & R²

Term(s) Omitted	DF	Log Likelihood	R² of Remaining Term(s)	Reduction From Model R²	Reduction From Saturated R²
All	2	-462.79903	0.00000		
Race	2	-409.60923	0.68093	0.01686	0.31907
County	2	-416.01890	0.59887	0.09892	0.40113
Offense	8	-449.09057	0.17549	0.52230	0.82451
None(Model)	12	-408.29185	0.69779	0.00000	0.30221
None(Saturated)	120	-384.68551	1.00000		0.00000

The output format is similar to previous examples. Notice in the analysis of deviance section that the variable *Race* is not significant. That is, in these data, the race of the defendant is not related to the verdict.

The *Coefficient Significance Tests* report combines the two logistic regression equations on one report. This makes it a bit more complicated to read, but it allows a quick comparison to be made of the corresponding regression coefficients. For each independent variable, the regression coefficient from each equation is shown. Thus, 2.82983 is the intercept for the *G* equation and 1.24012 is the intercept for the *NG* equation. No coefficient is shown for *NP* because it is the reference value.

Also note that the definition of the binary variables is as before. Thus the independent variable *County="Orange"* refers to a binary variable that was generated from the *County* variable. This binary variable is one when the county value is *Orange* and zero otherwise.

Validation

In order to validate this module, the estimated regression coefficients and the log likelihood generated by the ***BMDP*** (refer to page 1165 of version 7.0 of the ***BMDP*** manual) are displayed below.

Outcome: G	Coefficient	Std Error
1 RACE	0.2608	0.340
2 COUNTY	-0.8959	0.337
3 OFFENSE(1)	-2.230	0.514
4 OFFENSE(2)	-0.9185	0.578
5 OFFENSE(3)	-0.2138	0.629
6 OFFENSE(4)	0.4955	0.512
7 CONST1	2.830	0.445

Outcome: NG	Coefficient	Std Error
8 RACE	-0.1032	0.362
9 COUNTY	-0.1218	0.360
10 OFFENSE(1)	-0.5786	0.537
11 OFFENSE(2)	0.9281E-02	0.619
12 OFFENSE(3)	0.4801	0.670
13 OFFENSE(4)	-0.2670	0.576
14 CONST1	1.240	0.488

As you can see, these results match those displayed by **NCSS** exactly.

Example 5 – Logit Model with Interaction

This example continues with the analysis of the data given in Example 4. In that example, no interactions were included in the model. This example will include the two-way interactions in the model.

Setup

To run this example, complete the following steps:

1 Open the NC Criminal example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **NC Criminal** and click **OK**.

2 Specify the Logistic Regression procedure options

- Find and open the **Logistic Regression** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 5** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables, Model Tab	
Y	**Verdict**
Reference Value	**NP**
Categorical X's	**Race(B;A) County(B;Durham) Offense(B;Drunk)**
Frequencies	**Count**
Terms	**Up to 2-Way**
Priors	**Ni/N (Y-Value Proportions)**
Reports Tab	
Run Summary	**Checked**
Y Variable Summary	**Checked**
Coefficient Significance Tests	**Checked**
Analysis of Deviance	**Checked**
Log-Likelihood and R^2	**Checked**
All Other Reports	**Unchecked**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Logistic Regression Output

Coefficient Significance Tests

Independent Variable X	Regression Coefficient b(i)	Standard Error Sb(i)	Wald Z-Value H0: β=0	Wald P-Value	Odds Ratio Exp(b(i))
Intercept					
G	2.00583	0.50400	3.980	0.00007	7.43225
NG	0.72258	0.57465	1.257	0.20860	2.05975
(Race="B")					
G	1.44835	0.86924	1.666	0.09567	4.25608
NG	-1.10628	1.08369	-1.021	0.30733	0.33079
(County="Orange")					
G	0.14731	1.15368	0.128	0.89840	1.15871
NG	1.83395	1.18755	1.544	0.12251	6.25854
(Offense="MjTraffic")					
G	-0.30745	1.10221	-0.279	0.78029	0.73532
NG	-0.25450	1.23436	-0.206	0.83665	0.77531
(Offense="Property")					
G	-0.72178	0.83542	-0.864	0.38760	0.48589
NG	0.35757	0.89267	0.401	0.68874	1.42985
(Offense="Speed")					
G	1.93682	1.08041	1.793	0.07303	6.93666
NG	0.87254	1.19650	0.729	0.46586	2.39297
(Offense="Violence")					
G	-0.15836	0.87409	-0.181	0.85624	0.85354
NG	1.07460	0.91294	1.177	0.23916	2.92882
(Race="B")*(County="Orange")					
G	0.19528	0.81517	0.240	0.81067	1.21566
NG	0.83286	0.85899	0.970	0.33225	2.29990
(Race="B")*(Offense="MjTraffic")					
G	-1.17876	1.35078	-0.873	0.38285	0.30766
NG	1.16592	1.50638	0.774	0.43894	3.20886
(Race="B")*(Offense="Property")					
G	-0.83367	1.27452	-0.654	0.51305	0.43445
NG	1.35214	1.42888	0.946	0.34400	3.86569
(Race="B")*(Offense="Speed")					
G	-1.78987	1.25551	-1.426	0.15398	0.16698
NG	0.24862	1.45010	0.171	0.86387	1.28225
(Race="B")*(Offense="Violence")					
G	-2.31322	1.19041	-1.943	0.05199	0.09894
NG	0.51640	1.30133	0.397	0.69150	1.67598
(County="Orange")*(Offense="MjTraffic")					
G	0.45137	1.52019	0.297	0.76653	1.57046
NG	-0.53668	1.61710	-0.332	0.73998	0.58469
(County="Orange")*(Offense="Property")					
G	0.04871	1.41697	0.034	0.97258	1.04992
NG	-2.10279	1.47544	-1.425	0.15410	0.12212
(County="Orange")*(Offense="Speed")					
G	-1.39431	1.37573	-1.014	0.31082	0.24800
NG	-2.66093	1.48387	-1.793	0.07294	0.06988
(County="Orange")*(Offense="Violence")					
G	-2.42314	1.36627	-1.774	0.07614	0.08864
NG	-3.93664	1.38198	-2.849	0.00439	0.01951

Analysis of Deviance

Term Omitted	DF	Deviance	Increase From Model Deviance (Chi²)	P-Value
All	30	925.59805	146.82239	0.00000
Race	2	797.83870	19.06304	0.00007
County	2	788.31126	9.53560	0.00850
Offense	8	802.98614	24.21048	0.00211
Race*County	2	780.53878	1.76312	0.41414
Race*Offense	8	795.98619	17.21053	0.02799
County*Offense	8	798.81172	20.03607	0.01020
None(Model)	30	778.77566		

Logistic Regression

Log Likelihood & R²

Term(s) Omitted	DF	Log Likelihood	R² of Remaining Term(s)	Reduction From Model R²	Reduction From Saturated R²
All	2	-462.79903	0.00000		
Race	2	-398.91935	0.81778	0.12202	0.18222
County	2	-394.15563	0.87877	0.06104	0.12123
Offense	8	-401.49307	0.78483	0.15497	0.21517
Race*County	2	-390.26939	0.92852	0.01129	0.07148
Race*Offense	8	-397.99309	0.82964	0.11016	0.17036
County*Offense	8	-399.40586	0.81155	0.12825	0.18845
None(Model)	30	-389.38783	0.93980	0.00000	0.06020
None(Saturated)	120	-384.68554	1.00000		0.00000

Notice how the interactions are labeled. For example, the variable labeled *(Race="B")*(Offense="Violence")* is the interaction variable is generated by multiplying the binary variable defined by (Race="B") with the binary variable defined by (Offense="Violence"). The resulting variable is one if both of these conditions are true and zero otherwise.

Note that the R^2 is now 0.93980, so this model is almost as good as the saturated model.

Looking at the analysis of deviance table, we note that all terms are significant except for the Race*County interaction.

Example 6 – Odds Ratios for Categorical X's

Lachin (2000) pages 90, 91, and 257 presents an analysis of hypothetical data from an ulcer healing clinical trial conducted to study the effectiveness of a drug over a placebo. There were 100 patients assigned to the group receiving the drug and another 100 patients assigned to the group receiving the placebo. The ulcers were stratified into one of three types: 1. Acid-dependent, 2. Drug dependent, and 3. Intermediate. Each ulcer was followed for a period of time after which it was considered healed or not. The data for this experiment are given below. These data have been entered into a database named **Lachin91**.

Lachin91 dataset (subset)

Count	Ulcer	Drug	Healed
16	1	1	1
26	1	1	0
20	1	0	1
27	1	0	0
9	2	1	1
3	2	1	0
4	2	0	1
5	2	0	0
28	3	1	1
18	3	1	0
16	3	0	1
28	3	0	0

Setup

To run this example, complete the following steps:

1 Open the Lachin91 example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Lachin91** and click **OK**.

2 Specify the Logistic Regression procedure options

- Find and open the **Logistic Regression** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 6** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables, Model Tab	
Y	**Healed**
Categorical X's	**Ulcer Drug**
Frequencies	**Count**
Priors	**Equal across Y Values**
Reports Tab	
Run Summary	**Checked**
Coefficient Significance Tests	**Checked**
Odds Ratios	**Checked**
Analysis of Deviance	**Checked**
All Other Reports	**Unchecked**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Logistic Regression Output

Run Summary

Item	Value	Item	Value
Y Variable	Healed	Rows Processed	12
Reference Value	0	Rows Used	12
Number of Y-Values	2	Rows for Validation	0
Frequency Variable	Count	Rows X's Missing	0
Numeric X Variables	0	Rows Freq Miss. or 0	0
Categorical X Variables	2	Rows Prediction Only	0
Final Log Likelihood	-134.84531	Unique Rows (Y and X's)	12
Model R^2	0.54106	Sum of Frequencies	200
Actual Convergence	1.10275E-10	Likelihood Iterations	4
Target Convergence	1E-06	Maximum Iterations	20
Model D.F.	4	Completion Status	Normal Completion
Priors	Equal		

Coefficient Significance Tests

Independent Variable X	Regression Coefficient b(i)	Standard Error Sb(i)	Wald Z-Value H0: β=0	Wald P-Value	Odds Ratio Exp(b(i))
Intercept	-0.48951	0.21833	-2.242	0.02496	0.61293
(Ulcer=2)	0.83527	0.50247	1.662	0.09645	2.30543
(Ulcer=3)	0.32777	0.30424	1.077	0.28132	1.38787
(Drug=1)	0.50234	0.28845	1.742	0.08159	1.65259

Odds Ratios

Independent Variable X	Regression Coefficient b(i)	Odds Ratio Exp(b(i))	Lower 95% Confidence Limit	Upper 95% Confidence Limit
Intercept	-0.48951	0.61293	0.39955	0.94027
(Ulcer=2)	0.83527	2.30543	0.86109	6.17243
(Ulcer=3)	0.32777	1.38787	0.76451	2.51949
(Drug=1)	0.50234	1.65259	0.93894	2.90864

Analysis of Deviance

Term Omitted	DF	Deviance	Increase From Model Deviance (Chi²)	P-Value
All	3	276.27807	6.58746	0.08628
Ulcer	2	272.87155	3.18094	0.20383
Drug	1	272.74521	3.05460	0.08051
None(Model)	3	269.69061		

Note that neither Drug nor Ulcer is statistically significant at the 0.05 level using either the deviance tests in the *Analysis of Deviance* table or the Wald tests in the *Coefficient Significance Tests* section. From the *Odds Ratios* section, we see that the odds of healing are increased 1.65259 when the drug is administered.