

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

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

Logistic Regression

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

<u>P</u>	<u>Logit(P)</u>	<u>P</u>	<u>Logit(P)</u>
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

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

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, \dots, 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

$$\mathbf{X} = (X_1, X_2, \dots, X_p)$$

$$\mathbf{B}_g = \begin{pmatrix} \beta_{g1} \\ \vdots \\ \beta_{gp} \end{pmatrix}$$

The logistic regression model is given by the G equations

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

$$= \ln\left(\frac{P_g}{P_1}\right) + \mathbf{XB}_g$$

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

$$p_g = \Pr(Y = g | \mathbf{X})$$

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

The quantities P_1, P_2, \dots, P_G represent the prior probabilities of outcome membership. If these prior probabilities are assumed equal, then the term $\ln(P_g / P_1)$ 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 $\beta_{11}, \beta_{12}, \dots, \beta_{1p}$ 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 β 's are population regression coefficients that are to be estimated from the data. Their estimates are represented by b 's. The β '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

$$p_g = \Pr(Y = g | \mathbf{X}) = \frac{e^{\mathbf{XB}_g}}{1 + e^{\mathbf{XB}_2} + e^{\mathbf{XB}_3} + \dots + e^{\mathbf{XB}_G}}$$

since $e^{\mathbf{XB}_1} = 1$ 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*.

Logistic Regression

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

$$\begin{aligned} e^{XB} &= e^{\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p} \\ &= e^{\beta_1 X_1} e^{\beta_2 X_2} \dots e^{\beta_p X_p} \end{aligned}$$

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

Solving the Likelihood Equations

To improve notation, let

$$\begin{aligned} \pi_{gj} &= \text{Prob}(Y = g | X_j) \\ &= \frac{e^{X_j B_g}}{e^{X_j B_1} + e^{X_j B_2} + \dots + e^{X_j B_G}} \\ &= \frac{e^{X_j B_g}}{\sum_{s=1}^G e^{X_j B_s}} \end{aligned}$$

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

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

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

$$\begin{aligned} L = \ln(l) &= \sum_{j=1}^N \sum_{g=1}^G y_{gj} \ln(\pi_{gj}) \\ &= \sum_{j=1}^N \sum_{g=1}^G y_{gj} \ln \left(\frac{e^{X_j B_g}}{\sum_{s=1}^G e^{X_j B_s}} \right) \\ &= \sum_{j=1}^N \left[\sum_{g=1}^G y_{gj} X_j B_g - \ln \left(\sum_{s=1}^G e^{X_j B_s} \right) \right] \end{aligned}$$

Maximum likelihood estimates of the β '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

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

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

Logistic Regression

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

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

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

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,

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

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

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

Logistic Regression

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

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

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

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

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.'

<u>Value of p</u>	<u>Odds of p</u>
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.

Logistic Regression

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, X_1 and X_2 , 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 X_1 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 X_1 , 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.

Logistic Regression

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

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

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

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

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

$$\begin{aligned} \Delta D_{X_1} &= D_{\text{without } X_1} - D_{\text{with } X_1} \\ &= -2[L_{\text{without } X_1} - L_{\text{Saturated}}] + 2[L_{\text{with } X_1} - L_{\text{Saturated}}] \\ &= -2[L_{\text{without } X_1} - L_{\text{with } X_1}] \end{aligned}$$

Logistic Regression

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, $V(\hat{\beta})$.

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

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

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

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

Logistic Regression

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.

<u>Y</u>	<u>X1</u>	<u>X2</u>	<u>X3</u>	<u>X4</u>
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 that 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 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.

Logistic Regression

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

$$\pi_{gj} = \text{Prob}(Y = g \mid X_j)$$

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 $n_j \pi_{gj} (1 - \pi_{gj})$, 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.

Logistic Regression

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

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

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

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

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

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

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

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

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

Logistic Regression

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

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

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

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

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

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

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.

Logistic Regression

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

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

$$\bar{C}_{gj} = \frac{\chi_j^2 h_{gj}}{(1 - h_{gj})}, \quad j = 1, 2, \dots, J$$

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

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

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

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,

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

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

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

Logistic Regression

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

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

and the corresponding confidence interval is given by

$$l_j \pm z_{\alpha/2} (X_j' V_g X_j)$$

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

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

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

where

$$\sigma_B = X_j' V_g X_j$$

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

$$l_j \pm z_{\alpha/2} (X_j' V_g X_j)$$

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.

Logistic Regression

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.

<u>Option</u>	<u>Value</u>
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 (<i>in the Toolbar</i>)	
Variable Labels	Column Names

3 Run the procedure

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

Logistic Regression

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.

Logistic Regression

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	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 R² 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: $\beta=0$	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: $\beta=0$

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.

Logistic Regression

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	Regression Coefficient	Standard Error	Lower 95% Confidence Limit	Upper 95% Confidence Limit	Odds Ratio
X	b(i)	Sb(i)	Limit	Limit	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 confidence 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.

Logistic Regression

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	Regression Coefficient	Odds Ratio	Lower 95% Confidence Limit	Upper 95% Confidence Limit
X	b(i)	Exp(b(i))		
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.

Logistic Regression

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 $\text{Logit}(Y) = XB$) for a specific Y outcome. To calculate the Y-value probabilities when there are only 2 outcomes, transform the logit using $\text{Prob}(Y = \text{outcome}) = 1/(1+\text{Exp}(-XB))$ or $\text{Prob}(Y \neq \text{outcome}) = \text{Exp}(-XB)/(1+\text{Exp}(-XB))$. 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 $\text{Logit}(Y) = XB$) for a specific Y outcome. To calculate the Y-value probabilities when there are only 2 outcomes, transform the logit using $\text{Prob}(Y = \text{outcome}) = 1/(1+\text{Exp}(-XB))$ or $\text{Prob}(Y \neq \text{outcome}) = \text{Exp}(-XB)/(1+\text{Exp}(-XB))$. 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.

Logistic Regression

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 $(G-1)DF_t$ where DF_t 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 (Chi²)

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 & R²

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	-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 R² 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.

Logistic Regression

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 $(G-1)DF_t$ where DF_t 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 that 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.

Logistic Regression

Classification Table

Classification Table

Actual	Estimated		Total
	0	1	
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

Logistic Regression

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 by 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.

Logistic Regression

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 $100(1 - \alpha)\%$ 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

$$V(\hat{\Psi}) = \frac{4}{(1 - \hat{A}^2)^2} V(\hat{A})$$

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

$$L, U = \hat{\Psi} \mp z_{1-\alpha/2} \sqrt{V(\hat{\Psi})}$$

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

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

Logistic Regression

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.

Logistic Regression

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 $n_j \pi_{gj}(1 - \pi_{gj})$, 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

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

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.

Logistic Regression

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

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

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

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

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.

Logistic Regression

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

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

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 Cook's 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 ."