Chapter 4

Introduction to Categorical Data Analysis Procedures

Chapter Contents

Chapter 4
Introduction to Categorical Data Analysis Procedures

Overview

Several procedures in SAS/STAT software can be used for the analysis of categorical data:

CATMOD — fits linear models to functions of categorical data, facilitating such analyses as regression, analysis of variance, linear modeling, log-linear modeling, logistic regression, and repeated measures analysis. Maximum likelihood estimation is used for the analysis of logits and generalized logits, and weighted least squares analysis is used for fitting models to other response functions. Iterative proportional fitting (IPF), which avoids the need for parameter estimation, is available for fitting hierarchical log-linear models when there is a single population.

CORRESP — performs simple and multiple correspondence analyses, using a contingency table, Burt table, binary table, or raw categorical data as input. For more on PROC CORRESP, see Chapter 5, "Introduction to Multivariate Procedures," and Chapter 24, "The CORRESP Procedure."

FREQ — builds frequency tables or contingency tables and can produce numerous statistics. For one-way frequency tables, it can perform tests for equal proportions, specified proportions, or the binomial proportion. For contingency tables, it can compute various tests and measures of association and agreement including chi-square statistics, odds ratios, correlation statistics, Fisher's exact test for any size two-way table, kappa, and trend tests. In addition, it performs stratified analysis, computing Cochran-Mantel-Haenszel statistics and estimates of the common relative risk. Exact p-values and confidence intervals are available for various test statistics and measures.

GENMOD — fits generalized linear models with maximum-likelihood methods. This family includes logistic, probit, and complementary log-log regression models for binomial data, Poisson and negative binomial regression models for count data, and multinomial models for ordinal response data. It performs likelihood ratio and Wald tests for type I, type III, and user-defined contrasts. It analyzes repeated measures data with generalized estimating equation (GEE) methods.

LOGISTIC fits linear logistic regression models for discrete response data with maximum-likelihood methods. It provides four variable selection methods and computes regression diagnostics. It can also perform stratified conditional logistic regression analysis for binary response data and exact conditional regression analysis for binary and nominal response data. The logit link function in the logistic regression models can be replaced by the probit function or the complementary log-log function.

PROBIT fits models with probit, logit, or complementary log-log links for quantal assay or other discrete event data. It is mainly designed for dose-response analysis with a natural response rate. It computes the fiducial limits for the dose variable and provides various graphical displays for the analysis.

Other procedures that perform analyses for categorical data are the TRANSREG and PRINQUAL procedures. PROC PRINQUAL is summarized in Chapter 5, "Introduction to Multivariate Procedures," and PROC TRANSREG is summarized in Chapter 2, "Introduction to Regression Procedures."

A *categorical variable* is defined as one that can assume only a limited number of discrete values. The measurement scale for such a variable is unrestricted. It can be *nominal*, which means that the observed levels are not ordered. It can be *ordinal*, which means that the observed levels are ordered in some way. Or it can be *interval*, which means that the observed levels are ordered and numeric and that any interval of one unit on the scale of measurement represents the same amount, regardless of its location on the scale. One example of a categorical variable is litter size; another is the number of times a subject has been married. A variable that lies on a nominal scale is sometimes called a *qualitative* or *classification variable*.

Categorical data result from observations on multiple subjects where one or more categorical variables are observed for each subject. If there is only one categorical variable, then the data are generally represented by a *frequency table*, which lists each observed value of the variable and its frequency of occurrence.

If there are two or more categorical variables, then a subject's *profile* is defined as the subject's observed values for each of the variables. Such categorical data can be represented by a frequency table that lists each observed profile and its frequency of occurrence.

If there are exactly two categorical variables, then the data are often represented by a two-dimensional *contingency table*, which has one row for each level of variable 1 and one column for each level of variable 2. The intersections of rows and columns, called *cells*, correspond to variable profiles, and each cell contains the frequency of occurrence of the corresponding profile.

If there are more than two categorical variables, then the data can be represented by a *multidimensional contingency table*. There are two commonly used methods for displaying such tables, and both require that the variables be divided into two sets.

In the first method, one set contains a row variable and a column variable for a two-dimensional contingency table, and the second set contains all of the other variables. The variables in the second set are used to form a set of profiles. Thus, the data are represented as a series of two-dimensional contingency tables, one for each profile. This is the data representation used by PROC FREQ. For example, if you request tables for RACE*SEX*AGE*INCOME, the FREQ procedure represents the data as a series of contingency tables: the row variable is AGE, the column variable is INCOME, and the combinations of levels of RACE and SEX form a set of profiles.

In the second method, one set contains the independent variables, and the other set contains the dependent variables. Profiles based on the independent variables are called *population profiles*, whereas those based on the dependent variables are called *response profiles*. A two-dimensional contingency table is then formed, with one row for each population profile and one column for each response profile. Since any subject can have only one population profile and one response profile, the contingency table is uniquely defined. This is the data representation used by PROC CATMOD.

Sampling Frameworks and Distribution Assumptions

This section discusses the sampling frameworks and distribution assumptions for the CATMOD and FREQ procedures.

Simple Random Sampling: One Population

Suppose you take a simple random sample of 100 people and ask each person the following question: Of the three colors red, blue, and green, which is your favorite? You then tabulate the results in a frequency table as shown in Table 4.1.

Table 4.1. One-Way Frequency Table

	Favorite Color			
	Red	Blue	Green	Total
Frequency	52	31	17	100
Proportion	0.52	0.31	0.17	1.00

In the population you are sampling, you assume there is an unknown probability that a population member, selected at random, would choose any given color. In order to estimate that probability, you use the sample proportion

$$p_j = \frac{n_j}{n}$$

where n_j is the frequency of the jth response and n is the total frequency.

Because of the random variation inherent in any random sample, the frequencies have a probability distribution representing their relative frequency of occurrence in a hypothetical series of samples. For a simple random sample, the distribution of

frequencies for a frequency table with three levels is as follows. The probability that the first frequency is n_1, the second frequency is n_2, and the third is $n_3 = n - n_1 - n_2$, is given by

$$\Pr(n_1, n_2, n_3) = \frac{n!}{n_1! n_2! n_3!} \pi_1^{n_1} \pi_2^{n_2} \pi_3^{n_3}$$

where π_j is the true probability of observing the jth response level in the population.

This distribution, called the *multinomial distribution*, can be generalized to any number of response levels. The special case of two response levels is called the *binomial distribution*.

Simple random sampling is the type of sampling required by PROC CATMOD when there is one population. PROC CATMOD uses the multinomial distribution to estimate a probability vector and its covariance matrix. If the sample size is sufficiently large, then the probability vector is approximately normally distributed as a result of central limit theory. PROC CATMOD uses this result to compute appropriate test statistics for the specified statistical model.

Stratified Simple Random Sampling: Multiple Populations

Suppose you take two simple random samples, 50 men and 50 women, and ask the same question as before. You are now sampling two different populations that may have different response probabilities. The data can be tabulated as shown in Table 4.2.

Table 4.2. Two-Way Contingency Table: Sex by Color

	Favorite Color			
Sex	Red	Blue	Green	Total
Male	30	10	10	50
Female	20	10	20	50
Total	50	20	30	100

Note that the row marginal totals (50, 50) of the contingency table are fixed by the sampling design, but the column marginal totals (50, 20, 30) are random. There are six probabilities of interest for this table, and they are estimated by the sample proportions

$$p_{ij} = \frac{n_{ij}}{n_i}$$

where n_{ij} denotes the frequency for the ith population and the jth response, and n_i is the total frequency for the ith population. For this contingency table, the sample proportions are shown in Table 4.3.

Table 4.3. Table of Sample Proportions by Sex

	Favorite Color			
Sex	Red	Blue	Green	Total
Male	0.60	0.20	0. 20	1.00
Female	0.40	0. 20	0.40	1.00

The probability distribution of the six frequencies is the *product multinomial distribution*

$$\Pr(n_{11}, n_{12}, n_{13}, n_{21}, n_{22}, n_{23}) = \frac{n_1! n_2! \pi_{11}^{n_{11}} \pi_{12}^{n_{12}} \pi_{13}^{n_{13}} \pi_{21}^{n_{21}} \pi_{22}^{n_{22}} \pi_{23}^{n_{23}}}{n_{11}! n_{12}! n_{13}! n_{21}! n_{22}! n_{23}!}$$

where π_{ij} is the true probability of observing the jth response level in the ith population. The product multinomial distribution is simply the product of two or more individual multinomial distributions since the populations are independent. This distribution can be generalized to any number of populations and response levels.

Stratified simple random sampling is the type of sampling required by PROC CATMOD when there is more than one population. PROC CATMOD uses the product multinomial distribution to estimate a probability vector and its covariance matrix. If the sample sizes are sufficiently large, then the probability vector is approximately normally distributed as a result of central limit theory, and PROC CATMOD uses this result to compute appropriate test statistics for the specified statistical model. The statistics are known as Wald statistics, and they are approximately distributed as chi-square when the null hypothesis is true.

Observational Data: Analyzing the Entire Population

Sometimes the observed data do not come from a random sample but instead represent a complete set of observations on some population. For example, suppose a class of 100 students is classified according to sex and favorite color. The results are shown in Table 4.4.

In this case, you could argue that all of the frequencies are fixed since the entire population is observed; therefore, there is no sampling error. On the other hand, you could hypothesize that the observed table has only fixed marginals and that the cell frequencies represent one realization of a conceptual process of assigning color preferences to individuals. The assignment process is open to hypothesis, which means that you can hypothesize restrictions on the joint probabilities.

Table 4.4. Two-Way Contingency Table: Sex by Color

	Favorite Color			
Sex	Red	Blue	Green	Total
Male	16	21	20	57
Female	12	20	11	43
Total	28	41	31	100

The usual hypothesis (sometimes called *randomness*) is that the distribution of the column variable (Favorite Color) does not depend on the row variable (Sex). This implies that, for each row of the table, the assignment process corresponds to a simple random sample (without replacement) from the finite population represented by the column marginal totals (or by the column marginal subtotals that remain after sampling other rows). The hypothesis of randomness induces a probability distribution on the frequencies in the table; it is called the *hypergeometric distribution*.

If the same row and column variables are observed for each of several populations, then the probability distribution of all the frequencies can be called the *multiple hypergeometric distribution.* Each population is called a *stratum*, and an analysis that draws information from each stratum and then summarizes across them is called a *stratified analysis* (or a *blocked analysis* or a *matched analysis*). PROC FREQ does such a stratified analysis, computing test statistics and measures of association.

In general, the populations are formed on the basis of cross-classifications of independent variables. Stratified analysis is a method of adjusting for the effect of these variables without being forced to estimate parameters for them.

The multiple hypergeometric distribution is the one used by PROC FREQ for the computation of Cochran-Mantel-Haenszel statistics. These statistics are in the class of *randomization model test statistics*, which require minimal assumptions for their validity. PROC FREQ uses the multiple hypergeometric distribution to compute the mean and the covariance matrix of a function vector in order to measure the deviation between the observed and expected frequencies with respect to a particular type of alternative hypothesis. If the cell frequencies are sufficiently large, then the function vector is approximately normally distributed as a result of central limit theory, and FREQ uses this result to compute a quadratic form that has a chi-square distribution when the null hypothesis is true.

Randomized Experiments

Consider a *randomized experiment* in which patients are assigned to one of two treatment groups according to a randomization process that allocates 50 patients to each group. After a specified period of time, each patient's status (cured or uncured) is recorded. Suppose the data shown in Table 4.5 give the results of the experiment. The null hypothesis is that the two treatments are equally effective. Under this hypothesis, treatment is a randomly assigned label that has no effect on the cure rate of the patients. But this implies that each row of the table represents a simple random sample from the finite population whose cure rate is described by the column marginal totals. Therefore, the column marginals (58, 42) are fixed under the hypothesis. Since the row marginals (50, 50) are fixed by the allocation process, the hypergeometric distribution is induced on the cell frequencies. Randomized experiments can also be specified in a stratified framework, and Cochran-Mantel-Haenszel statistics can be computed relative to the corresponding multiple hypergeometric distribution.

Table 4.5. Two-Way Contingency Table: Treatment by Status

	Status		
Treatment	Cured	Uncured	Total
1	36	14	50
2	22	28	50
Total	58	42	100

Relaxation of Sampling Assumptions

As indicated previously, the CATMOD procedure assumes that the data are from a stratified simple random sample, so it uses the product multinomial distribution. If the data are not from such a sample, then in many cases it is still possible to use PROC CATMOD by arguing that each row of the contingency table *does* represent a simple random sample from some hypothetical population. The extent to which the inferences are generalizable depends on the extent to which the hypothetical population is perceived to resemble the target population.

Similarly, the Cochran-Mantel-Haenszel statistics use the multiple hypergeometric distribution, which requires fixed row and column marginal totals in each contingency table. If the sampling process does not yield a table with fixed margins, then it is usually possible to fix the margins through conditioning arguments similar to the ones used by Fisher when he developed the Exact Test for 2×2 tables. In other words, if you want fixed marginal totals, you can generally make your analysis conditional on those observed totals.

For more information on sampling models for categorical data, see Bishop, Fienberg, and Holland (1975, Chapter 13).

Comparison of FREQ and CATMOD Procedures

PROC FREQ is used primarily to investigate the relationship between two variables; any confounding variables are taken into account by stratification rather than by parameter estimation. PROC CATMOD is used to investigate the relationship among many variables, all of which are integrated into a parametric model.

When PROC CATMOD estimates the covariance matrix of the frequencies, it assumes that the frequencies were obtained by a stratified simple random sampling procedure. However, PROC CATMOD can also analyze input data that consist of a function vector and a covariance matrix. Therefore, if the sampling procedure is different, you can estimate the covariance matrix of the frequencies in the appropriate manner before submitting the data to PROC CATMOD.

For the FREQ procedure, Fisher's Exact Test and Cochran-Mantel-Haenszel statistics are based on the hypergeometric distribution, which corresponds to fixed marginal totals. However, by conditioning arguments, these tests are generally applicable to a wide range of sampling procedures. Similarly, the Pearson and likelihood-ratio chi-square statistics can be derived under a variety of sampling situations.

PROC FREQ can do some traditional nonparametric analysis (such as the Kruskal-Wallis test and Spearman's correlation) since it can generate rank scores internally. Fisher's Exact Test and the Cochran-Mantel-Haenszel statistics are also inherently nonparametric. However, the main vehicle for nonparametric analyses in the SAS System is the NPAR1WAY procedure.

A large sample size is required for the validity of the chi-square distributions, the standard errors, and the covariance matrices for both PROC FREQ and PROC CATMOD. If sample size is a problem, then PROC FREQ has the advantage with its CMH statistics because it does not use any degrees of freedom to estimate parameters for confounding variables. In addition, PROC FREQ can compute exact p-values for any two-way table, provided that the sample size is sufficiently small in relation to the size of the table. It can also produce exact p-values for many tests, including the test of binomial proportions, the Cochran-Armitage test for trend, and the Jonckheere-Terpstra test for ordered differences among classes.

See the chapters on the FREQ and CATMOD procedures for more information. In addition, some well-known texts that deal with analyzing categorical data are listed in the "References" section of this chapter.

Comparison of CATMOD, GENMOD, LOGISTIC, and PROBIT Procedures

The CATMOD, GENMOD, LOGISTIC, and PROBIT procedures can all be used for statistical modeling of categorical data. The CATMOD procedure provides maximum likelihood estimation for logistic regression, including the analysis of logits for dichotomous outcomes and the analysis of generalized logits for polychotomous outcomes. It provides weighted least squares estimation of many other response functions, such as means, cumulative logits, and proportions, and you can also compute and analyze other response functions that can be formed from the proportions corresponding to the rows of a contingency table. In addition, a user can input and analyze a set of response functions and user-supplied covariance matrix with weighted least squares. With the CATMOD procedure, by default, all explanatory (independent) variables are treated as classification variables.

The GENMOD procedure is also a general statistical modeling tool which fits generalized linear models to data: it fits several useful models to categorical data including logistic regression, the proportional odds model, and Poisson regression. The GENMOD procedure also provides a facility for fitting generalized estimating equations to correlated response data that are categorical, such as repeated dichotomous outcomes. The GENMOD procedure fits models using maximum likelihood estimation, and you include classification variables in your models with a CLASS statement. PROC GENMOD can perform type I and type III tests, and it provides predicted values and residuals.

The LOGISTIC procedure is specifically designed for logistic regression. It performs the usual logistic regression analysis for dichotomous outcomes and it fits the proportional odds model and the generalized logit model for ordinal and nominal outcomes, respectively, by the method of maximum likelihood. With the CLASS statement, you

can include independent CLASS variables in the model. This procedure has capabilities for a variety of model-building techniques, including stepwise, forward, and backward selection. It computes predicted values, the receiver operating characteristics (ROC) curve and the area beneath the curve, and a number of regression diagnostics. It can create output data sets containing these values and other statistics. PROC LOGISTIC can perform a conditional logistic regression analysis (matched-set and case-controlled) for binary response data. For small data sets, PROC LOGISTIC can perform the exact conditional logistic analysis of Hirji, Mehta, and Patel (1987) and Mehta, Patel, and Senchaudhuri (1992).

The PROBIT procedure is designed for quantal assay or other discrete event data. In additional to performing the logistic regression analysis, it can estimate the threshold response rate. PROC PROBIT can also estimate the values of independent variables that yield a desired response. With the CLASS statement, you can include CLASS variables in the model. PROC PROBIT allows only the less-than-full-rank parameterization for the CLASS variables.

Stokes, Davis, and Koch (2000) provide substantial discussion of these procedures, particularly the use of the FREQ, LOGISTIC, GENMOD, and CATMOD procedures for statistical modeling.

Logistic Regression

Dichotomous Response

You have many choices of performing logistic regression in the SAS System. The CATMOD, GENMOD, LOGISTIC, and PROBIT procedures fit the usual logistic regression model.

PROC LOGISTIC provides the capability of model-building, and performs conditional logistic regression analysis for case-control studies and exact conditional logistic regression analysis. You may choose to use it for these reasons.

PROC CATMOD may not be efficient when there are continous independent variables with large numbers of different values. For a continuous variable with a very limited number of values, PROC CATMOD may be useful. You list the continuous variables in the DIRECT statement.

The LOGISTIC, GENMOD, and PROBIT procedures can analyze summarized data by enabling you to input the numbers of events and trials; the ratio of events to trials must be between 0 and 1. PROC PROBIT enables you to estimate the natural response rate and compute fiducial limits for the dose variable.

Ordinal Response

PROC LOGISTIC fits the proportional odds model to the ordinal response data by default. PROC PROBIT fits this model if you specify the logistic distribution, and PROC GENMOD fits the same model if you specify the CLOGIT link and the multinomial distribution.

Nominal Response

When the response variable is nominal, there is no concept of ordering of the response values. PROC CATMOD fits a logistic model to response functions called *generalized logits*. PROC LOGISTIC fits the generalized logit model if you specify the GLOGIT link.

Parameterization

There are some differences in the way that models are parameterized, which means that you might get different parameter estimates if you were to perform logistic regression in each of these procedures.

- Parameter estimates from the procedures may differ in sign, depending on the ordering of response levels, which you can change if you want.
- The parameter estimates associated with a categorical independent variable may differ among the procedures, since the estimates depend on the coding of the indicator variables in the design matrix. By default, the design matrix column produced by PROC CATMOD for a binary independent variable is coded using the values 1 and -1 . The same column produced by the CLASS statement of PROC PROBIT is coded using 1 and 0. PROC CATMOD uses the deviation from the mean coding, which is a full-rank parameterization, and PROC PROBIT uses the less-than-full-rank coding. As a result, the parameter estimate printed by PROC CATMOD is one-half of the estimate produced by PROC PROBIT. Both PROC GENMOD and PROC LOGISTIC allow either a full-rank parameterization or the less-than-full-rank parameterization. See the "Details" sections in the chapters on the CATMOD, GENMOD, LOGISTIC, and PROBIT procedures for more information on the generation of the design matrices used by these procedures.
- The maximum-likelihood algorithm used differs among the procedures. PROC LOGISTIC uses the Fisher's scoring method by default, while PROC PROBIT, PROC GENMOD, and PROC CATMOD use the Newton-Raphson method. The parameter estimates should be the same for all three procedures, and the standard errors should be the same for the logistic model. For the normal and extreme-value (Gompertz) distributions in PROC PROBIT, which correspond to the probit and cloglog links, respectively, in PROC GENMOD and PROC LOGISTIC, the standard errors may differ. In general, tests computed using the standard errors from the Newton-Raphson method will be more conservative.
- The LOGISTIC, GENMOD, and PROBIT procedures can be used to fit a cumulative regression model for ordinal response data using maximum-likelihood estimation. PROC LOGISTIC and PROC GENMOD use a different parameterization from that of PROC PROBIT, which results in different intercept parameters. Estimates of the slope parameters, however, should be the same for both procedures. The estimated standard errors of the slope estimates are slightly different between the two procedures because of the different computational algorithms used as default.

References

Agresti, A. (1984), *Analysis of Ordinal Categorical Data*, New York: John Wiley & Sons, Inc.

Agresti, A. (2002), *Categorical Data Analysis*, Second Edition, New York: John Wiley & Sons, Inc.

Bishop, Y., Fienberg, S.E., and Holland, P.W. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: MIT Press.

Collett, D. (1991), *Modelling Binary Data*, London: Chapman and Hall.

Cox, D.R. and Snell, E.J. (1989), *The Analysis of Binary Data*, Second Edition, London: Chapman and Hall.

Dobson, A. (1990), *An Introduction to Generalized Linear Models*, London: Chapman and Hall.

Fleiss, J.L. (1981), *Statistical Methods for Rates and Proportions*, Second Edition, New York: John Wiley & Sons, Inc.

Freeman, D.H., (1987), *Applied Categorical Data Analysis*, New York: Marcel-Dekker.

Grizzle, J.E., Starmer, C.F., and Koch, G.G. (1969), "Analysis of Categorical Data by Linear Models," *Biometrics*, 25, 489–504.

Hirji, K.F., Mehta, C.R., and Patel, N.R. (1987), "Computing Distributions for Exact Logistic Regression," *Journal of the American Statistical Association*, 82, 1110–1117.

Hosmer, D.W, Jr. and Lemeshow, S. (1989), *Applied Logistic Regression*, New York: John Wiley & Sons, Inc.

Mehta, C.R., Patel, N. and Senchaudhuri, P. (1992), "Exact Stratified Linear Rank Tests for Ordered Categorical and Binary Data," *Journal of Computational and Graphical Statistics*, 1, 21–40.

McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models*, London: Chapman and Hall.

Stokes, M.E., Davis, C.S., and Koch, G.G (2000), *Categorical Data Analysis Using the SAS System*, Cary NC: SAS Institute Inc.

Chapter 42
The LOGISTIC Procedure

Chapter Contents

Chapter 42
The LOGISTIC Procedure

Overview

Binary responses (for example, success and failure), ordinal responses (for example, normal, mild, and severe), and nominal responses (for example, major TV networks viewed at a certain hour) arise in many fields of study. Logistic regression analysis is often used to investigate the relationship between these discrete responses and a set of explanatory variables. Several texts that discuss logistic regression are Collett (1991), Agresti (1990), Cox and Snell (1989), Hosmer and Lemeshow (2000), and Stokes, Davis, and Koch (2000).

For binary response models, the response, Y, of an individual or an experimental unit can take on one of two possible values, denoted for convenience by 1 and 2 (for example, $Y = 1$ if a disease is present, otherwise $Y = 2$). Suppose $\mathbf{x}$ is a vector of explanatory variables and $\pi = \Pr(Y = 1 \mid \mathbf{x})$ is the response probability to be modeled. The linear logistic model has the form

$$\text{logit}(\pi) \equiv \log\left(\frac{\pi}{1-\pi}\right) = \alpha + \boldsymbol{\beta}'\mathbf{x}$$

where α is the intercept parameter and $\boldsymbol{\beta}$ is the vector of parameters. Notice that the LOGISTIC procedure, by default, models the probability of the *lower* response levels.

The logistic model shares a common feature with a more general class of linear models, that a function $g = g(\mu)$ of the mean of the response variable is assumed to be linearly related to the explanatory variables. Since the mean μ implicitly depends on the stochastic behavior of the response, and the explanatory variables are assumed to be fixed, the function g provides the link between the random (stochastic) component and the systematic (deterministic) component of the response variable Y. For this reason, Nelder and Wedderburn (1972) refer to $g(\mu)$ as a link function. One advantage of the logit function over other link functions is that differences on the logistic scale are interpretable regardless of whether the data are sampled prospectively or retrospectively (McCullagh and Nelder 1989, Chapter 4). Other link functions that are widely used in practice are the probit function and the complementary log-log function. The LOGISTIC procedure enables you to choose one of these link functions, resulting in fitting a broader class of binary response models of the form

$$g(\pi) = \alpha + \boldsymbol{\beta}'\mathbf{x}$$

For ordinal response models, the response, Y, of an individual or an experimental unit may be restricted to one of a (usually small) number, $k + 1 (k \geq 1)$, of ordinal values, denoted for convenience by $1, \ldots, k, k + 1$. For example, the severity of

coronary disease can be classified into three response categories as 1=no disease, 2=angina pectoris, and 3=myocardial infarction. The LOGISTIC procedure fits a common slopes cumulative model, which is a parallel lines regression model based on the cumulative probabilities of the response categories rather than on their individual probabilities. The cumulative model has the form

$$g(\Pr(Y \leq i \mid \mathbf{x})) = \alpha_i + \boldsymbol{\beta}'\mathbf{x}, \quad i = 1, \ldots, k$$

where $\alpha_1, \ldots, \alpha_k$ are k intercept parameters, and $\boldsymbol{\beta}$ is the vector of parameters. This model has been considered by many researchers. Aitchison and Silvey (1957) and Ashford (1959) employ a probit scale and provide a maximum likelihood analysis; Walker and Duncan (1967) and Cox and Snell (1989) discuss the use of the log-odds scale. For the log-odds scale, the cumulative logit model is often referred to as the *proportional odds* model.

For nominal response logistic models, where the $k + 1$ possible responses have no natural ordering, the logit model can also be extended to a *generalized logit* model, which has the form

$$\log\left(\frac{\Pr(Y = i \mid \mathbf{x})}{\Pr(Y = k + 1 \mid \mathbf{x})}\right) = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}, \quad i = 1, \ldots, k$$

where the $\alpha_1, \ldots, \alpha_k$ are k intercept parameters, and the $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k$ are k vectors of parameters. These models were introduced by McFadden (1974) as the *discrete choice* model, and they are also known as *multinomial* models.

The LOGISTIC procedure fits linear logistic regression models for discrete response data by the method of maximum likelihood. It can also perform conditional logistic regression for binary response data and exact conditional logistic regression for binary and nominal response data. The maximum likelihood estimation is carried out with either the Fisher-scoring algorithm or the Newton-Raphson algorithm. You can specify starting values for the parameter estimates. The logit link function in the logistic regression models can be replaced by the probit function, the complementary log-log function, or the generalized logit function.

The LOGISTIC procedure enables you to specify categorical variables (also known as CLASS variables) or continuous variables as explanatory variables. You can also specify more complex model terms such as interactions and nested terms in the same way as in the GLM procedure. Any term specified in the model is referred to as an *effect*, whether it is a continuous variable, a CLASS variable, an interaction, or a nested term.

The LOGISTIC procedure allows either a full-rank parameterization or a less-than-full-rank parameterization. The full-rank parameterization offers eight coding methods: effect, reference, ordinal, polynomial, and orthogonalizations of these. The effect coding is the same method that is used in the CATMOD procedure. The less-than-full-rank parameterization is the same coding as that used in the GLM procedure.

The LOGISTIC procedure provides four effect selection methods: forward selection, backward elimination, stepwise selection, and best subset selection. The best subset

selection is based on the likelihood score statistic. This method identifies a specified number of best models containing one, two, three effects, and so on, up to a single model containing effects for all the explanatory variables.

The LOGISTIC procedure has some additional options to control how to move effects in and out of a model with various model-building strategies such as forward selection, backward elimination, or stepwise selection. When there are no interaction terms, a main effect can enter or leave a model in a single step based on the p-value of the score or Wald statistic. When there are interaction terms, the selection process also depends on whether you want to preserve model hierarchy. These additional options enable you to specify whether model hierarchy is to be preserved, how model hierarchy is applied, and whether a single effect or multiple effects can be moved in a single step.

Odds ratio estimates are displayed along with parameter estimates. You can also specify the change in the explanatory variables for which odds ratio estimates are desired. Confidence intervals for the regression parameters and odds ratios can be computed based either on the profile likelihood function or on the asymptotic normality of the parameter estimators.

Various methods to correct for overdispersion are provided, including Williams' method for grouped binary response data. The adequacy of the fitted model can be evaluated by various goodness-of-fit tests, including the Hosmer-Lemeshow test for binary response data.

Like many procedures in SAS/STAT software that enable the specification of CLASS variables, the LOGISTIC procedure provides a CONTRAST statement for specifying customized hypothesis tests concerning the model parameters. The CONTRAST statement also provides estimation of individual rows of contrasts, which is particularly useful for obtaining odds ratio estimates for various levels of the CLASS variables.

You can perform a conditional logistic regression on binary response data by specifying the STRATA statement. This enables you to perform matched-set and case-control analyses. The number of events and nonevents can vary across the strata. Many of the features available with the unconditional analysis are also available with a conditional analysis.

The LOGISTIC procedure enables you to perform exact conditional logistic regression using the method of Hirji, Mehta, and Patel (1987) and Mehta, Patel, and Senchaudhuri (1992) by specifying one or more EXACT statements. You can test individual parameters or conduct a joint test for several parameters. The procedure computes two exact tests: the exact conditional score test and the exact conditional probability test. You can request exact estimation of specific parameters and corresponding odds ratios where appropriate. Both point estimates and confidence intervals are provided.

Further features of the LOGISTIC procedure enable you to

- control the ordering of the response categories
- compute a generalized R^2 measure for the fitted model

- reclassify binary response observations according to their predicted response probabilities
- test linear hypotheses about the regression parameters
- create a data set for producing a receiver operating characteristic curve for each fitted model
- create a data set containing the estimated response probabilities, residuals, and influence diagnostics
- score a data set using a previously fitted model

Experimental graphics are now available with the LOGISTIC procedure. For more information, see the "ODS Graphics" section on page 2388.

The remaining sections of this chapter describe how to use PROC LOGISTIC and discuss the underlying statistical methodology. The "Getting Started" section introduces PROC LOGISTIC with an example for binary response data. The "Syntax" section (page 2289) describes the syntax of the procedure. The "Details" section (page 2329) summarizes the statistical technique employed by PROC LOGISTIC. The "Examples" section (page 2391) illustrates the use of the LOGISTIC procedure with 10 applications.

For more examples and discussion on the use of PROC LOGISTIC, refer to Stokes, Davis, and Koch (2000), Allison (1999), and SAS Institute Inc. (1995).

Getting Started

The LOGISTIC procedure is similar in use to the other regression procedures in the SAS System. To demonstrate the similarity, suppose the response variable y is binary or ordinal, and x1 and x2 are two explanatory variables of interest. To fit a logistic regression model, you can use a MODEL statement similar to that used in the REG procedure:

```
proc logistic;
   model y=x1 x2;
run;
```

The response variable y can be either character or numeric. PROC LOGISTIC enumerates the total number of response categories and orders the response levels according to the response variable option ORDER= in the MODEL statement. The procedure also allows the input of binary response data that are grouped:

```
proc logistic;
   model r/n=x1 x2;
run;
```

Here, n represents the number of trials and r represents the number of events.

The following example illustrates the use of PROC LOGISTIC. The data, taken from Cox and Snell (1989, pp. 10–11), consist of the number, r, of ingots not ready for rolling, out of n tested, for a number of combinations of heating time and soaking time. The following invocation of PROC LOGISTIC fits the binary logit model to the grouped data:

```
data ingots;
   input Heat Soak r n @@;
   datalines;
7 1.0 0 10  14 1.0 0 31  27 1.0 1 56  51 1.0 3 13
7 1.7 0 17  14 1.7 0 43  27 1.7 4 44  51 1.7 0  1
7 2.2 0  7  14 2.2 2 33  27 2.2 0 21  51 2.2 0  1
7 2.8 0 12  14 2.8 0 31  27 2.8 1 22  51 4.0 0  1
7 4.0 0  9  14 4.0 0 19  27 4.0 1 16
;


proc logistic data=ingots;
   model r/n=Heat Soak;
run;
```

The results of this analysis are shown in the following tables.

The LOGISTIC Procedure

Model Information

Data Set	WORK.INGOTS
Response Variable (Events)	r
Response Variable (Trials)	n
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	19
Number of Observations Used	19
Sum of Frequencies Read	387
Sum of Frequencies Used	387

Figure 42.1. Binary Logit Model

PROC LOGISTIC first lists background information in Figure 42.1 about the fitting of the model. Included are the name of the input data set, the response variable(s) used, the number of observations used, and the link function used.

Response Profile

Ordered Value	Binary Outcome	Total Frequency
1	Event	12
2	Nonevent	375

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Figure 42.2. Response Profile with Events/Trials Syntax

The "Response Profile" table (Figure 42.2) lists the response categories (which are Event and Nonevent when grouped data are input), their ordered values, and their total frequencies for the given data.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	108.988	101.346
SC	112.947	113.221
-2 Log L	106.988	95.346

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	11.6428	2	0.0030
Score	15.1091	2	0.0005
Wald	13.0315	2	0.0015

Figure 42.3. Fit Statistics and Hypothesis Tests

The "Model Fit Statistics" table (Figure 42.3) contains the Akaike Information Criterion (AIC), the Schwarz Criterion (SC), and the negative of twice the log likelihood (-2 Log L) for the intercept-only model and the fitted model. AIC and SC can be used to compare different models, and the ones with smaller values are preferred. Results of the likelihood ratio test and the efficient score test for testing the joint significance of the explanatory variables (Soak and Heat) are included in the "Testing Global Null Hypothesis: BETA=0" table (Figure 42.3).

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.5592	1.1197	24.6503	<.0001
Heat	1	0.0820	0.0237	11.9454	0.0005
Soak	1	0.0568	0.3312	0.0294	0.8639

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Heat	1.085	1.036	1.137
Soak	1.058	0.553	2.026

Figure 42.4. Parameter Estimates and Odds Ratios

The "Analysis of Maximum Likelihood Estimates" table in Figure 42.4 lists the parameter estimates, their standard errors, and the results of the Wald test for individual parameters. The odds ratio for each effect parameter, estimated by exponentiating the corresponding parameter estimate, is shown in the "Odds Ratios Estimates" table (Figure 42.4), along with 95% Wald confidence intervals.

Using the parameter estimates, you can calculate the estimated logit of π as

$$-5.5592 + 0.082 \times \text{Heat} + 0.0568 \times \text{Soak}$$

If Heat=7 and Soak=1, then $\text{logit}(\hat{\pi}) = -4.9284$. Using this logit estimate, you can calculate $\hat{\pi}$ as follows:

$$\hat{\pi} = 1/(1 + \mathrm{e}^{4.9284}) = 0.0072$$

This gives the predicted probability of the event (ingot not ready for rolling) for Heat=7 and Soak=1. Note that PROC LOGISTIC can calculate these statistics for you; use the OUTPUT statement with the PREDICTED= option.

Association of Predicted Probabilities and Observed Responses

Percent Concordant	64.4	Somers' D	0.460
Percent Discordant	18.4	Gamma	0.555
Percent Tied	17.2	Tau-a	0.028
Pairs	4500	c	0.730

Figure 42.5. Association Table

Finally, the "Association of Predicted Probabilities and Observed Responses" table (Figure 42.5) contains four measures of association for assessing the predictive abil-

ity of a model. They are based on the number of pairs of observations with different response values, the number of concordant pairs, and the number of discordant pairs, which are also displayed. Formulas for these statistics are given in the "Rank Correlation of Observed Responses and Predicted Probabilities" section on page 2350.

To illustrate the use of an alternative form of input data, the following program creates the INGOTS data set with new variables NotReady and Freq instead of n and r. The variable NotReady represents the response of individual units; it has a value of 1 for units not ready for rolling (event) and a value of 0 for units ready for rolling (nonevent). The variable Freq represents the frequency of occurrence of each combination of Heat, Soak, and NotReady. Note that, compared to the previous data set, NotReady=1 implies Freq=r, and NotReady=0 implies Freq=n−r.

```
data ingots;
   input Heat Soak NotReady Freq @@;
   datalines;
7 1.0 0 10  14 1.0 0 31  14 4.0 0 19  27 2.2 0 21  51 1.0 1  3
7 1.7 0 17  14 1.7 0 43  27 1.0 1  1  27 2.8 1  1  51 1.0 0 10
7 2.2 0  7  14 2.2 1  2  27 1.0 0 55  27 2.8 0 21  51 1.7 0  1
7 2.8 0 12  14 2.2 0 31  27 1.7 1  4  27 4.0 1  1  51 2.2 0  1
7 4.0 0  9  14 2.8 0 31  27 1.7 0 40  27 4.0 0 15  51 4.0 0  1
;
```

The following SAS statements invoke PROC LOGISTIC to fit the same model using the alternative form of the input data set.

```
proc logistic data=ingots;
   model NotReady(event='1') = Soak Heat;
   freq Freq;
run;
```

Results of this analysis are the same as the previous one. The displayed output for the two runs are identical except for the background information of the model fit and the "Response Profile" table shown in Figure 42.6.

```
                    The LOGISTIC Procedure

                       Response Profile

                Ordered                       Total
                  Value     NotReady      Frequency

                      1            0            375
                      2            1             12

                Probability modeled is NotReady=1.
```

Figure 42.6. Response Profile with Single-Trial Syntax

By default, Ordered Values are assigned to the sorted response values in ascending order, and PROC LOGISTIC models the probability of the response level that corresponds to the Ordered Value 1. There are several methods to change these defaults; the preceding statements specify the response variable option EVENT= to model the probability of NotReady=1 as displayed in Figure 42.6. See the "Response Level Ordering" section on page 2329 for more details.

Syntax

The following statements are available in PROC LOGISTIC:

```
PROC LOGISTIC < options >;
   BY variables ;
   CLASS variable <(v-options)> <variable <(v-options)>... >
         < / v-options >;
   CONTRAST 'label' effect values <,... effect values >< / options >;
   EXACT < 'label' >< Intercept >< effects >< / options > ;
   FREQ variable ;
   MODEL events/trials = < effects >< / options >;
   MODEL variable < (variable_options) > = < effects >< / options >;
   OUTPUT < OUT=SAS-data-set >
          < keyword=name. . .keyword=name >< / option >;
   SCORE < options >;
   STRATA effects < / options >;
   < label: > TEST equation1 < , . . . , < equationk >>< / option >;
   UNITS independent1=list1 < . . . independentk=listk >< / option > ;
   WEIGHT variable < / option >;
```

The PROC LOGISTIC and MODEL statements are required; only one MODEL statement can be specified. The CLASS statement (if used) must precede the MODEL statement, and the CONTRAST, EXACT, and STRATA statements (if used) must follow the MODEL statement. The rest of this section provides detailed syntax information for each of the preceding statements, beginning with the PROC LOGISTIC statement. The remaining statements are covered in alphabetical order.

PROC LOGISTIC Statement

PROC LOGISTIC < *options* > **;**

The PROC LOGISTIC statement starts the LOGISTIC procedure and optionally identifies input and output data sets and suppresses the display of results.

ALPHA=α

specifies the level of significance α for $100(1-\alpha)\%$ confidence intervals. The value α must be between 0 and 1; the default value is 0.05, which results in 95% intervals. This value is used as the default confidence level for limits computed by the following options.

Statement	Options
CONTRAST	ESTIMATE=
EXACT	ESTIMATE=
MODEL	CLODDS= CLPARM=
OUTPUT	UCL= LCL=
SCORE	CLM

You can override the default in each of these cases by specifying the ALPHA= option for each statement individually.

COVOUT

adds the estimated covariance matrix to the OUTEST= data set. For the COVOUT option to have an effect, the OUTEST= option must be specified. See the section "OUTEST= Output Data Set" on page 2374 for more information.

DATA=*SAS-data-set*

names the SAS data set containing the data to be analyzed. If you omit the DATA= option, the procedure uses the most recently created SAS data set. The INMODEL= option cannot be specified with this option.

DESCENDING
DESC

reverses the sorting order for the levels of the response variable. If both the DESCENDING and ORDER= options are specified, PROC LOGISTIC orders the levels according to the ORDER= option and then reverses that order. This option has the same effect as the response variable option DESCENDING in the MODEL statement. See the "Response Level Ordering" section on page 2329 for more detail.

EXACTONLY

requests only the exact analyses. The asymptotic analysis that PROC LOGISTIC usually performs is suppressed.

EXACTOPTIONS(*options*)
specifies options that apply to every EXACT statement in the program. The following options are available:

ADDTOBS adds the observed sufficient statistic to the sampled exact distribution if the statistic was not sampled. This option has no effect unless the METHOD=NETWORKMC option is specified and the ESTIMATE option is specified in the EXACT statement. If the observed statistic has not been sampled, then the parameter estimate does not exist; by specifying this option, you can produce (biased) estimates.

MAXTIME=*seconds* specifies the maximum clock time (in seconds) that PROC LOGISTIC can use to calculate the exact distributions. If the limit is exceeded, the procedure halts all computations and prints a note to the LOG. The default maximum clock time is seven days.

METHOD=*keyword* specifies which exact conditional algorithm to use for every EXACT statement specified. You can specify one of the following *keywords*:

DIRECT invokes the multivariate shift algorithm of Hirji, Mehta, and Patel (1987). This method directly builds the exact distribution, but it may require an excessive amount of memory in its intermediate stages. METHOD=DIRECT is invoked by default when you are conditioning out at most the intercept, or when the LINK=GLOGIT option is specified in the MODEL statement.

NETWORK invokes an algorithm similar to that described in Mehta, Patel, and Senchaudhuri (1992). This method builds a network for each parameter that you are conditioning out, combines the networks, then uses the multivariate shift algorithm to create the exact distribution. The NETWORK method can be faster and require less memory than the DIRECT method. The NETWORK method is invoked by default for most analyses.

NETWORKMC invokes the hybrid network and Monte Carlo algorithm of Mehta, Patel, and Senchaudhuri (2000). This method creates a network then samples from that network; this method does not reject any of the samples at the cost of using a large amount of memory to create the network. METHOD=NETWORKMC is most useful for producing parameter estimates for problems that are too large for the DIRECT and NETWORK methods to handle and for which asymptotic methods are invalid; for example, for sparse data on a large grid.

N=*n* specifies the number of Monte Carlo samples to take when METHOD=NETWORKMC. By default $n = 10,000$. If the procedure cannot obtain n samples due to a lack of memory, then a note is printed in the LOG (the number of valid samples is also reported in the listing) and the analysis continues.

Note that the number of samples used to produce any particular statistic may be smaller than *n*. For example, let $X1$ and $X2$ be continuous variables, denote their joint distribution by $f(X1, X2)$, and let $f(X1|X2 = x2)$ denote the

marginal distribution of $X1$ conditioned on the observed value of $X2$. If you request the JOINT test of $X1$ and $X2$, then *n* samples are used to generate the estimate $\hat{f}(X1, X2)$ of $f(X1, X2)$, from which the test is computed. However, the parameter estimate for $X1$ is computed from the subset of $\hat{f}(X1, X2)$ having $X2 = x2$, and this subset need not contain *n* samples. Similarly, the distribution for each level of a classification variable is created by extracting the appropriate subset from the joint distribution for the CLASS variable. The sample sizes used to compute the statistics are written to the ODS OUTPUT data set of the tables.

In some cases, the marginal sample size may be too small to admit accurate estimation of a particular statistic; a note is printed in the LOG when a marginal sample size is less than 100. Increasing *n* will increase the number of samples used in a marginal distribution; however, if you want to control the sample size exactly, you can:

- Remove the JOINT option from the EXACT statement.
- Create dummy variables in a DATA step to represent the levels of a CLASS variable, and specify them as independent variables in the MODEL statement.

ONDISK uses disk-space instead of random access memory to build the exact conditional distribution. Use this option to handle larger problems at the cost of slower processing.

SEED=*n* specifies the initial seed for the random number generator used to take the Monte Carlo samples for METHOD=NETWORKMC. The value of the SEED= option must be an integer. If you do not specify a seed, or if you specify a value less than or equal to zero, then PROC LOGISTIC uses the time of day from the computer's clock to generate an initial seed. The seed is displayed in the "Model Information" table.

STATUSN=*n* prints a status line in the LOG after every n Monte Carlo samples for METHOD=NETWORKMC. The number of samples taken and the current exact p-value for testing the significance of the model are displayed. You can use this status line to track the progress of the computation of the exact conditional distributions.

STATUSTIME=*seconds* specifies the time interval (in seconds) for printing a status line in the LOG. You can use this status line to track the progress of the computation of the exact conditional distributions. The time interval you specify is approximate; the actual time interval will vary. By default, no status reports are produced.

INEST=*SAS-data-set*

names the SAS data set that contains initial estimates for all the parameters in the model. BY-group processing is allowed in setting up the INEST= data set. See the section "INEST= Input Data Set" on page 2376 for more information.

INMODEL=*SAS-data-set*

specifies the name of the SAS data set that contains the model information needed for scoring new data. This INMODEL= data set is the OUTMODEL= data set saved in a previous PROC LOGISTIC call. The DATA= option cannot be specified with this option; instead, specify the data sets to be scored in the SCORE statements.

When the INMODEL= data set is specified, FORMAT statements are not allowed; variables in the DATA= and PRIOR= data sets should be formatted within the data sets. If a SCORE statement is specified in the same run as fitting the model, FORMAT statements should be specified after the SCORE statement in order for the formats to apply to all the DATA= and PRIOR= data sets in the SCORE statement.

You can specify the BY statement provided the INMODEL= data set is created under the same BY-group processing.

The CLASS, EXACT, MODEL, OUTPUT, TEST, and UNIT statements are not available with the INMODEL= option.

NAMELEN=*n*

specifies the length of effect names in tables and output data sets to be n characters, where n is a value between 20 and 200. The default length is 20 characters.

NOCOV

specifies that the covariance matrix is not saved in the OUTMODEL= data set. The covariance matrix is needed for computing the confidence intervals for the posterior probabilities in the OUT= data set in the SCORE statement. Specifying this option will reduce the size of the OUTMODEL= data set.

NOPRINT

suppresses all displayed output. Note that this option temporarily disables the Output Delivery System (ODS); see Chapter 14, "Using the Output Delivery System," for more information.

ORDER=DATA | FORMATTED | FREQ | INTERNAL
RORDER=DATA | FORMATTED | INTERNAL

specifies the sorting order for the levels of the response variable. See the response variable option ORDER= in the MODEL statement for more information.

OUTDESIGN=*SAS-data-set*

specifies the name of the data set that contains design matrix for the model. The data set contains the same number of observations as the corresponding DATA= data set and includes the response variable (with the same format as in the input data), the FREQ variable, the WEIGHT variable, the OFFSET variable, and the design variables for the covariates, including the Intercept variable of constant value 1 unless the NOINT option in the MODEL statement is specified.

OUTDESIGNONLY

suppresses the model fitting and only creates the OUTDESIGN= data set. This option is ignored if the OUTDESIGN= option is not specified.

OUTEST= *SAS-data-set*
creates an output SAS data set that contains the final parameter estimates and, optionally, their estimated covariances (see the preceding COVOUT option). The output data set also includes a variable named _LNLIKE_, which contains the log likelihood.

See the section "OUTEST= Output Data Set" on page 2374 for more information.

OUTMODEL=*SAS-data-set*
specifies the name of the SAS data set that contains the information about the fitted model. This data set contains sufficient information to score new data without having to refit the model. It is solely used as the input to the INMODEL= option in a subsequent PROC LOGISTIC call. **Note:** information is stored in this data set in a very compact form, hence you should not modify it manually.

SIMPLE
displays simple descriptive statistics (mean, standard deviation, minimum and maximum) for each continuous explanatory variable; and for each CLASS variable involved in the modeling, the frequency counts of the classification levels are displayed. The SIMPLE option generates a breakdown of the simple descriptive statistics or frequency counts for the entire data set and also for individual response categories.

TRUNCATE
specifies that class levels should be determined using no more than the first 16 characters of the formatted values of CLASS, response, and strata variables. When formatted values are longer than 16 characters, you can use this option to revert to the levels as determined in releases previous to Version 9. This option invokes the same option in the CLASS statement.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC LOGISTIC to obtain separate analyses on observations in groups defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. The *variables* are one or more variables in the input data set.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data using the SORT procedure with a similar BY statement.
- Specify the BY statement option NOTSORTED or DESCENDING in the BY statement for the LOGISTIC procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables using the DATASETS procedure (in base SAS software).

If a SCORE statement is specified, then define the *primary data set* to be the DATA= or the INMODEL=data set in the PROC LOGISTIC statement, and define the *secondary data set* to be the DATA= data set and PRIOR= data set in the SCORE statement. The primary data set contains all of the BY variables, and the secondary data set must contain either all of them or none of them. If the secondary data set contains all the BY-variables, matching is carried out between the primary and secondary data sets. If the secondary data set does not contain any of the BY-variables, the entire secondary data set is used for every BY-group in the primary data set and the BY-variables are added to the output data sets specified in the SCORE statement.

Caution: The order of your response and classification variables is determined by combining data across all BY groups; however, the observed levels may change between BY groups. This may affect the value of the reference level for these variables, and hence your interpretation of the model and the parameters.

For more information on the BY statement, refer to the discussion in *SAS Language Reference: Concepts*. For more information on the DATASETS procedure, refer to the discussion in the *SAS Procedures Guide*.

CLASS Statement

CLASS *variable* <*(v-options)*><*variable* <*(v-options)*>... > < / *v-options* > **;**

The CLASS statement names the classification variables to be used in the analysis. The CLASS statement must precede the MODEL statement. You can specify various *v-options* for each variable by enclosing them in parentheses after the variable name. You can also specify global *v-options* for the CLASS statement by placing them after a slash (/). Global *v-options* are applied to all the variables specified in the CLASS statement. If you specify more than one CLASS statement, the global *v-options* specified on any one CLASS statement apply to all CLASS statements. However, individual CLASS variable *v-options* override the global *v-options*.

CPREFIX= *n*

specifies that, at most, the first *n* characters of a CLASS variable name be used in creating names for the corresponding design variables. The default is $32 - \min(32, \max(2, f))$, where f is the formatted length of the CLASS variable.

DESCENDING
DESC

reverses the sorting order of the classification variable. If both the DESCENDING and ORDER= options are specified, PROC LOGISTIC orders the categories according to the ORDER= option and then reverses that order.

LPREFIX= *n*

specifies that, at most, the first *n* characters of a CLASS variable label be used in creating labels for the corresponding design variables. The default is $256 - \min(256, \max(2, f))$, where f is the formatted length of the CLASS variable.

MISSING

allows missing value ('.' for a numeric variable and blanks for a character variables) as a valid value for the CLASS variable.

ORDER=DATA | FORMATTED | FREQ | INTERNAL

specifies the sorting order for the levels of classification variables. By default, ORDER=FORMATTED. For ORDER=FORMATTED and ORDER=INTERNAL, the sort order is machine dependent. When ORDER=FORMATTED is in effect for numeric variables for which you have supplied no explicit format, the levels are ordered by their internal values. This ordering determines which parameters in the model correspond to each level in the data, so the ORDER= option may be useful when you use the CONTRAST statement.

The following table shows how PROC LOGISTIC interprets values of the ORDER= option.

Value of ORDER=	Levels Sorted By
DATA	order of appearance in the input data set
FORMATTED	external formatted value, except for numeric variables with no explicit format, which are sorted by their unformatted (internal) value
FREQ	descending frequency count; levels with the most observations come first in the order
INTERNAL	unformatted value

For more information on sorting order, see the chapter on the SORT procedure in the *SAS Procedures Guide* and the discussion of BY-group processing in *SAS Language Reference: Concepts*.

PARAM=*keyword*

specifies the parameterization method for the classification variable or variables. Design matrix columns are created from CLASS variables according to the following coding schemes. The default is PARAM=EFFECT. If PARAM=ORTHPOLY or PARAM=POLY, and the CLASS levels are numeric, then the ORDER= option in the CLASS statement is ignored, and the internal, unformatted values are used. See the "CLASS Variable Parameterization" section on page 2331 for further details.

EFFECT — specifies effect coding

GLM — specifies less-than-full-rank reference cell coding; this option can only be used as a global option

ORDINAL — specifies the cumulative parameterization for an ordinal CLASS variable.

POLYNOMIAL
POLY — specifies polynomial coding

REFERENCE
REF — specifies reference cell coding

ORTHEFFECT	orthogonalizes PARAM=EFFECT
ORTHORDINAL	orthogonalizes PARAM=ORDINAL
ORTHPOLY	orthogonalizes PARAM=POLYNOMIAL
ORTHREF	orthogonalizes PARAM=REFERENCE

The EFFECT, POLYNOMIAL, REFERENCE, ORDINAL, and their orthogonal parameterizations are full rank. The REF= option in the CLASS statement determines the reference level for the EFFECT, REFERENCE, and their orthogonal parameterizations.

Parameter names for a CLASS predictor variable are constructed by concatenating the CLASS variable name with the CLASS levels. However, for the POLYNOMIAL and orthogonal parameterizations, parameter names are formed by concatenating the CLASS variable name and keywords that reflect the parameterization.

REF=*'level'* **|** *keyword*

specifies the reference level for PARAM=EFFECT, PARAM=REFERENCE, and their orthogonalizations. For an individual (but not a global) variable REF= *option*, you can specify the *level* of the variable to use as the reference level. For a global or individual variable REF= *option*, you can use one of the following *keywords*. The default is REF=LAST.

FIRST	designates the first ordered level as reference
LAST	designates the last ordered level as reference

TRUNCATE

specifies that class levels should be determined using no more than the first 16 characters of the formatted values of CLASS, response, and strata variables. When formatted values are longer than 16 characters, you can use this option to revert to the levels as determined in releases previous to Version 9. The TRUNCATE option is only available as a global option. This option invokes the same option in the PROC LOGISTIC statement.

CONTRAST Statement

CONTRAST *'label' row-description* <,*...row-description* >< / *options* > **;**

where a *row-description* is: *effect values* <,*...effect values*>

The CONTRAST statement provides a mechanism for obtaining customized hypothesis tests. It is similar to the CONTRAST and ESTIMATE statements in PROC GLM and PROC CATMOD, depending on the coding schemes used with any classification variables involved.

The CONTRAST statement enables you to specify a matrix, $\mathbf{L}$, for testing the hypothesis $\mathbf{L}\boldsymbol{\theta} = \mathbf{0}$, where $\boldsymbol{\theta}$ is the parameter vector. You must be familiar with the details of the model parameterization that PROC LOGISTIC uses (for more information, see the PARAM= option in the section "CLASS Statement" on page 2295).

Optionally, the CONTRAST statement enables you to estimate each row, $\boldsymbol{l}_i'\boldsymbol{\theta}$, of $\mathbf{L}\boldsymbol{\theta}$ and test the hypothesis $\boldsymbol{l}_i'\boldsymbol{\theta} = 0$. Computed statistics are based on the asymptotic chi-square distribution of the Wald statistic.

There is no limit to the number of CONTRAST statements that you can specify, but they must appear after the MODEL statement.

The following parameters are specified in the CONTRAST statement:

label identifies the contrast on the output. A label is required for every contrast specified, and it must be enclosed in quotes.

effect identifies an effect that appears in the MODEL statement. The name INTERCEPT can be used as an effect when one or more intercepts are included in the model. You do not need to include all effects that are included in the MODEL statement.

values are constants that are elements of the $\mathbf{L}$ matrix associated with the effect. To correctly specify your contrast, it is crucial to know the ordering of parameters within each effect and the variable levels associated with any parameter. The "Class Level Information" table shows the ordering of levels within variables. The E option, described later in this section, enables you to verify the proper correspondence of *values* to parameters.

The rows of $\mathbf{L}$ are specified in order and are separated by commas. Multiple degree-of-freedom hypotheses can be tested by specifying multiple *row-descriptions*. For any of the full-rank parameterizations, if an effect is not specified in the CONTRAST statement, all of its coefficients in the $\mathbf{L}$ matrix are set to 0. If too many values are specified for an effect, the extra ones are ignored. If too few values are specified, the remaining ones are set to 0.

When you use effect coding (by default or by specifying PARAM=EFFECT in the CLASS statement), all parameters are directly estimable (involve no other parameters). For example, suppose an effect coded CLASS variable A has four levels. Then there are three parameters $(\alpha_1, \alpha_2, \alpha_3)$ representing the first three levels, and the fourth parameter is represented by

$$-\alpha_1 - \alpha_2 - \alpha_3$$

To test the first versus the fourth level of A, you would test

$$\alpha_1 = -\alpha_1 - \alpha_2 - \alpha_3$$

or, equivalently,

$$2\alpha_1 + \alpha_2 + \alpha_3 = 0$$

which, in the form $\mathbf{L}\boldsymbol{\theta} = \mathbf{0}$, is

$$\begin{bmatrix} 2 & 1 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = 0$$

Therefore, you would use the following CONTRAST statement:

```
contrast '1 vs. 4' A 2 1 1;
```

To contrast the third level with the average of the first two levels, you would test

$$\frac{\alpha_1 + \alpha_2}{2} = \alpha_3$$

or, equivalently,

$$\alpha_1 + \alpha_2 - 2\alpha_3 = 0$$

Therefore, you would use the following CONTRAST statement:

```
contrast '1&2 vs. 3' A 1 1 -2;
```

Other CONTRAST statements are constructed similarly. For example,

```
contrast '1 vs. 2    '  A  1 -1  0;
contrast '1&2 vs. 4  '  A  3  3  2;
contrast '1&2 vs. 3&4'  A  2  2  0;
contrast 'Main Effect'  A  1  0  0,
                        A  0  1  0,
                        A  0  0  1;
```

When you use the less-than-full-rank parameterization (by specifying PARAM=GLM in the CLASS statement), each row is checked for estimability. If PROC LOGISTIC finds a contrast to be nonestimable, it displays missing values in corresponding rows in the results. PROC LOGISTIC handles missing level combinations of classification variables in the same manner as PROC GLM. Parameters corresponding to missing level combinations are not included in the model. This convention can affect the way in which you specify the **L** matrix in your CONTRAST statement. If the elements of **L** are not specified for an effect that contains a specified effect, then the elements of the specified effect are distributed over the levels of the higher-order effect just as the GLM procedure does for its CONTRAST and ESTIMATE statements. For example, suppose that the model contains effects A and B and their interaction A*B. If you specify a CONTRAST statement involving A alone, the **L** matrix contains nonzero terms for both A and A*B, since A*B contains A.

The degrees of freedom is the number of linearly independent constraints implied by the CONTRAST statement, that is, the rank of **L**.

You can specify the following options after a slash (/).

ALPHA=α
specifies the level of significance α for the $100(1-\alpha)\%$ confidence interval for each contrast when the ESTIMATE option is specified. The value α must be between 0 and 1. By default, α is equal to the value of the ALPHA= option in the PROC LOGISTIC statement, or 0.05 if that option is not specified.

E
displays the $\mathbf{L}$ matrix.

ESTIMATE=*keyword*
requests that each individual contrast (that is, each row, $\boldsymbol{l}_i'\boldsymbol{\theta}$, of $\mathbf{L}\boldsymbol{\theta}$) or exponentiated contrast ($e^{\boldsymbol{l}_i'\boldsymbol{\theta}}$) be estimated and tested. PROC LOGISTIC displays the point estimate, its standard error, a Wald confidence interval, and a Wald chi-square test for each contrast. The significance level of the confidence interval is controlled by the ALPHA= option. You can estimate the contrast or the exponentiated contrast ($e^{\boldsymbol{l}_i'\boldsymbol{\theta}}$), or both, by specifying one of the following *keywords*:

PARM	specifies that the contrast itself be estimated
EXP	specifies that the exponentiated contrast be estimated
BOTH	specifies that both the contrast and the exponentiated contrast be estimated

SINGULAR = *number*
tunes the estimability check. This option is ignored when the full-rank parameterization is used. If $\boldsymbol{v}$ is a vector, define ABS($\boldsymbol{v}$) to be the largest absolute value of the elements of $\boldsymbol{v}$. For a row vector $\boldsymbol{l}'$ of the contrast matrix $\mathbf{L}$, define c to be equal to ABS($\boldsymbol{l}$) if ABS($\boldsymbol{l}$) is greater than 0; otherwise, c equals 1. If ABS($\boldsymbol{l}' - \boldsymbol{l}'\boldsymbol{T}$) is greater than $c * number$, then $\boldsymbol{l}$ is declared nonestimable. The $\boldsymbol{T}$ matrix is the Hermite form matrix $\boldsymbol{I}_0^-\boldsymbol{I}_0$, where $\boldsymbol{I}_0^-$ represents a generalized inverse of the information matrix $\boldsymbol{I}_0$ of the null model. The value for $number$ must be between 0 and 1; the default value is 1E−4.

EXACT Statement

EXACT *<'label'>< Intercept >< effects >< / options >* **;**

The EXACT statement performs exact tests of the parameters for the specified effects and optionally estimates the parameters and outputs the exact conditional distributions. You can specify the keyword INTERCEPT and any effects in the MODEL statement. Inference on the parameters of the specified effects is performed by conditioning on the sufficient statistics of all the other model parameters (possibly including the intercept).

You can specify several EXACT statements, but they must follow the MODEL statement. Each statement can optionally include an identifying label. If several EXACT statements are specified, any statement without a label will be assigned a label of the form "Exact$n$", where "n" indicates the nth EXACT statement. The label is included in the headers of the displayed exact analysis tables.

If a STRATA statement is also specified, then a stratified exact conditional logistic regression is performed. The model contains a different intercept for each stratum, and these intercepts are conditioned out of the model along with any other nuisance parameters (essentially, any parameters specified in the MODEL statement which are not in the EXACT statement).

If the LINK=GLOGIT option is specified in the MODEL statement, then the EXACTOPTION option METHOD=DIRECT is invoked by default and a generalized logit model is fit. Since each effect specified in the MODEL statement adds k parameters to the model (where $k+1$ is the number of response levels), exact analysis of the generalized logit model using this method is limited to rather small problems.

The CONTRAST, OUTPUT, SCORE, TEST, and UNITS statements are not available with an exact analysis. Exact analyses are not performed when you specify a WEIGHT statement, a link other than LINK=LOGIT or LINK=GLOGIT, an offset variable, the NOFIT option, or a model-selection method. Exact estimation is not available for ordinal response models.

The following options can be specified in each EXACT statement after a slash (/):

ALPHA=α
specifies the level of significance α for $100(1-\alpha)$% confidence limits for the parameters or odds ratios. The value α must be between 0 and 1. By default, α is equal to the value of the ALPHA= option in the PROC LOGISTIC statement, or 0.05 if that option is not specified.

ESTIMATE $<$ **=***keyword* $>$
estimates the individual parameters (conditional on all other parameters) for the effects specified in the EXACT statement. For each parameter, a point estimate, a confidence interval, and a p-value for a two-sided test that the parameter is zero are displayed. Note that the two-sided p-value is twice the one-sided p-value. You can optionally specify one of the following keywords:

PARM	specifies that the parameters be estimated. This is the default.
ODDS	specifies that the odds ratios be estimated. For classification variables, use of the reference parameterization is recommended.
BOTH	specifies that the parameters and odds ratios be estimated

JOINT
performs the joint test that all of the parameters are simultaneously equal to zero, individual hypothesis tests for the parameter of each continuous variable, and joint tests for the parameters of each classification variable. The joint test is indicated in the "Conditional Exact Tests" table by the label "Joint."

JOINTONLY
performs only the joint test of the parameters. The test is indicated in the "Conditional Exact Tests" table by the label "Joint." When this option is specified, individual tests for the parameters of each continuous variable and joint tests for the parameters of the classification variables are not performed.

CLTYPE=EXACT | MIDP
requests either the exact or mid-p confidence intervals for the parameter estimates. By default, the exact intervals are produced. The confidence coefficient can be specified with the ALPHA= option. The mid-p interval can be modified with the MIDPFACTOR= option. See the "Inference for a Single Parameter" section on page 2373 for details.

MIDPFACTOR=δ_1 | (δ_1, δ_2)
sets the tie factors used to produce the mid-p hypothesis statistics and the mid-p confidence intervals. δ_1 modifies both the hypothesis tests and confidence intervals, while δ_2 affects only the hypothesis tests. By default, $\delta_1 = 0.5$ and $\delta_2 = 1.0$. See the "Hypothesis Tests" section on page 2371 and the "Inference for a Single Parameter" section on page 2373 for details.

ONESIDED
requests one-sided confidence intervals and p-values for the individual parameter estimates and odds ratios. The one-sided p-value is the smaller of the left and right tail probabilities for the observed sufficient statistic of the parameter under the null hypothesis that the parameter is zero. The two-sided p-values (default) are twice the one-sided p-values. See the "Inference for a Single Parameter" section on page 2373 for more details.

OUTDIST=*SAS-data-set*
names the SAS data set containing the exact conditional distributions. This data set contains all of the exact conditional distributions required to process the corresponding EXACT statement. The data set contains the possible sufficient statistics for the parameters of the effects specified in the EXACT statement, the counts, and, when hypothesis tests are performed on the parameters, the probability of occurrence and the score value for each sufficient statistic. When you request an OUTDIST= data set, the observed sufficient statistics are displayed in the "Sufficient Statistics" table. See the "OUTDIST= Output Data Set" section on page 2377 for more information.

EXACT Statement Examples

- In the following example, two exact tests are computed: one for x1 and the other for x2. The test for x1 is based on the exact conditional distribution of the sufficient statistic for the x1 parameter given the observed values of the sufficient statistics for the intercept, x2, and x3 parameters; likewise, the test for x2 is conditional on the observed sufficient statistics for the intercept, x1, and x3:

```
proc logistic;
   model y= x1 x2 x3;
   exact 'lab1' x1 x2;
run;
```

- You can specify multiple EXACT statements in the same PROC LOGISTIC invocation. PROC LOGISTIC determines, from all the EXACT statements, the distinct conditional distributions that need to be evaluated. For example, there is only one exact conditional distribution for the following two EXACT

statements, and it would be a waste of resources to compute the same exact conditional distribution twice:

```
exact 'One' x1 / estimate=parm;
exact 'Two' x1 / estimate=parm onesided;
```

- For each EXACT statement, individual tests for the parameters of the specified effects are computed unless the JOINTONLY option is specified. Consider the following EXACT statements:

```
exact 'E12' x1 x2 / estimate;
exact 'E1'  x1    / estimate;
exact 'E2'  x2    / estimate;
exact 'J12' x1 x2 / joint;
```

In the E12 statement, the parameters for x1 and x2 are estimated and tested separately. Specifying the E12 statement is equivalent to specifying both the E1 and E2 statements. In the J12 statement, the joint test for the parameters of x1 and x2 is computed as well as the individual tests for x1 and x2.

All exact conditional distributions for the tests and estimates computed in a single EXACT statement are output to the corresponding OUTDIST= data set. For example, consider the following EXACT statements:

```
exact 'O1'   x1    /           outdist=o1;
exact 'OJ12' x1 x2 / jointonly outdist=oj12;
exact 'OA12' x1 x2 / joint     outdist=oa12;
exact 'OE12' x1 x2 / estimate  outdist=oe12;
```

The O1 statement outputs a single exact conditional distribution. The OJ12 statement outputs only the joint distribution for x1 and x2. The OA12 statement outputs three conditional distributions: one for x1, one for x2, and one jointly for x1 and x2. The OE12 statement outputs two conditional distributions: one for x1 and the other for x2. Data set oe12 contains both the x1 and x2 variables; the distribution for x1 has missing values in the x2 column while the distribution for x2 has missing values in the x1 column.

See the "OUTDIST= Output Data Set" section on page 2377 for more information.

FREQ Statement

FREQ *variable* **;**

The *variable* in the FREQ statement identifies a variable that contains the frequency of occurrence of each observation. PROC LOGISTIC treats each observation as if it appears n times, where n is the value of the FREQ variable for the observation. If it is not an integer, the frequency value is truncated to an integer. If the frequency value is less than 1 or missing, the observation is not used in the model fitting. When the FREQ statement is not specified, each observation is assigned a frequency of 1.

If a SCORE statement is specified, then the FREQ variable is used for computing fit statistics and the ROC curve, but they are not required for scoring. If the DATA= data

set in the SCORE statement does not contain the FREQ variable, the frequency values are assumed to be 1 and a warning message is issued in the LOG. If you fit a model and perform the scoring in the same run, the same FREQ variable is used for fitting and scoring. If you fit a model in a previous run and input it with the INMODEL= option in the current run, then the FREQ variable can be different from the one used in the previous run; however, if a FREQ variable was not specified in the previous run you can still specify a FREQ variable in the current run.

MODEL Statement

MODEL *events/trials*= < *effects* >< / *options* > **;**

MODEL *variable* < *(variable_options)* >= < *effects* >< / *options* > **;**

The MODEL statement names the response variable and the explanatory effects, including covariates, main effects, interactions, and nested effects; see the section "Specification of Effects" on page 1784 of Chapter 32, "The GLM Procedure," for more information. If you omit the explanatory effects, the procedure fits an intercept-only model. Model options can be specified after a slash (/).

Two forms of the MODEL statement can be specified. The first form, referred to as *single-trial* syntax, is applicable to binary, ordinal, and nominal response data. The second form, referred to as *events/trials* syntax, is restricted to the case of binary response data. The *single-trial* syntax is used when each observation in the DATA= data set contains information on only a single trial, for instance, a single subject in an experiment. When each observation contains information on multiple binary-response trials, such as the counts of the number of subjects observed and the number responding, then *events/trials* syntax can be used.

In the *events/trials* syntax, you specify two variables that contain count data for a binomial experiment. These two variables are separated by a slash. The value of the first variable, *events*, is the number of positive responses (or events). The value of the second variable, *trials*, is the number of trials. The values of both *events* and (*trials*−*events*) must be nonnegative and the value of *trials* must be positive for the response to be valid.

In the *single-trial* syntax, you specify one variable (on the left side of the equal sign) as the response variable. This variable can be character or numeric. Options specific to the response variable can be specified immediately after the response variable with a pair of parentheses around them.

For both forms of the MODEL statement, explanatory *effects* follow the equal sign. Variables can be either continuous or classification variables. Classification variables can be character or numeric, and they must be declared in the CLASS statement. When an effect is a classification variable, the procedure enters a set of coded columns into the design matrix instead of directly entering a single column containing the values of the variable.

Response Variable Options

You can specify the following options by enclosing them in a pair of parentheses after the response variable.

DESCENDING | DESC
reverses the order of the response categories. If both the DESCENDING and ORDER= options are specified, PROC LOGISTIC orders the response categories according to the ORDER= option and then reverses that order. See the "Response Level Ordering" section on page 2329 for more detail.

EVENT='*category***' |** *keyword*
specifies the event category for the binary response model. PROC LOGISTIC models the probability of the event category. The EVENT= option has no effect when there are more than two response categories. You can specify the value (formatted if a format is applied) of the event category in quotes or you can specify one of the following keywords. The default is EVENT=FIRST.

FIRST	designates the first ordered category as the event
LAST	designates the last ordered category as the event

One of the most common sets of response levels is {0,1}, with 1 representing the event for which the probability is to be modeled. Consider the example where Y takes the values 1 and 0 for event and nonevent, respectively, and Exposure is the explanatory variable. To specify the value 1 as the event category, use the MODEL statement

```
model Y(event='1') = Exposure;
```

ORDER= DATA | FORMATTED | FREQ | INTERNAL
specifies the sorting order for the levels of the response variable. By default, ORDER=FORMATTED. For ORDER=FORMATTED and ORDER=INTERNAL, the sort order is machine dependent. When ORDER=FORMATTED is in effect for numeric variables for which you have supplied no explicit format, the levels are ordered by their internal values.

The following table shows the interpretation of the ORDER= values.

Value of ORDER=	Levels Sorted By
DATA	order of appearance in the input data set
FORMATTED	external formatted value, except for numeric variables with no explicit format, which are sorted by their unformatted (internal) value
FREQ	descending frequency count; levels with the most observations come first in the order
INTERNAL	unformatted value

For more information on sorting order, see the chapter on the SORT procedure in the *SAS Procedures Guide* and the discussion of BY-group processing in *SAS Language Reference: Concepts*.

REFERENCE='*category***' | *keyword***
REF='*category***' | *keyword***
specifies the reference category for the generalized logit model and the binary response model. For the generalized logit model, each nonreference category is contrasted with the reference category. For the binary response model, specifying one response category as the reference is the same as specifying the other response category as the event category. You can specify the value (formatted if a format is applied) of the reference category in quotes or you can specify one of the following keywords. The default is REF=LAST.

FIRST	designates the first ordered category as the reference
LAST	designates the last ordered category as the reference

Model Options

Table 42.1 summarizes the options available in the MODEL statement, which can be specified after a slash (/).

Table 42.1. Model Statement Options

Option	Description
Model Specification Options	
LINK=	specifies link function
NOINT	suppresses intercept
NOFIT	suppresses model fitting
OFFSET=	specifies offset variable
SELECTION=	specifies effect selection method
Effect Selection Options	
BEST=	controls the number of models displayed for SCORE selection
DETAILS	requests detailed results at each step
FAST	uses fast elimination method
HIERARCHY=	specifies whether and how hierarchy is maintained and whether a single effect or multiple effects are allowed to enter or leave the model per step
INCLUDE=	specifies number of effects included in every model
MAXSTEP=	specifies maximum number of steps for STEPWISE selection
SEQUENTIAL	adds or deletes effects in sequential order
SLENTRY=	specifies significance level for entering effects
SLSTAY=	specifies significance level for removing effects
START=	specifies number of variables in first model
STOP=	specifies number of variables in final model
STOPRES	adds or deletes variables by residual chi-square criterion
Model-Fitting Specification Options	
ABSFCONV=	specifies absolute function convergence criterion
FCONV=	specifies relative function convergence criterion
GCONV=	specifies relative gradient convergence criterion

Table 42.1. (continued)

Option	Description
XCONV=	specifies relative parameter convergence criterion
MAXFUNCTION=	specifies maximum number of function calls for the conditional analysis
MAXITER=	specifies maximum number of iterations
NOCHECK	suppresses checking for infinite parameters
RIDGING=	specifies the technique used to improve the log-likelihood function when its value is worse than that of the previous step
SINGULAR=	specifies tolerance for testing singularity
TECHNIQUE=	specifies iterative algorithm for maximization
Options for Confidence Intervals	
ALPHA=	specifies α for the $100(1-\alpha)\%$ confidence intervals
CLPARM=	computes confidence intervals for parameters
CLODDS=	computes confidence intervals for odds ratios
PLCONV=	specifies profile likelihood convergence criterion
Options for Classifying Observations	
CTABLE	displays classification table
PEVENT=	specifies prior event probabilities
PPROB=	specifies probability cutpoints for classification
Options for Overdispersion and Goodness-of-Fit Tests	
AGGREGATE=	determines subpopulations for Pearson chi-square and deviance
SCALE=	specifies method to correct overdispersion
LACKFIT	requests Hosmer and Lemeshow goodness-of-fit test
Options for ROC Curves	
OUTROC=	names the output data set
ROCEPS=	specifies probability grouping criterion
Options for Regression Diagnostics	
INFLUENCE	displays influence statistics
IPLOTS	requests index plots
Options for Display of Details	
CORRB	displays correlation matrix
COVB	displays covariance matrix
EXPB	displays exponentiated values of estimates
ITPRINT	displays iteration history
NODUMMYPRINT	suppresses “Class Level Information” table
PARMLABEL	displays parameter labels
RSQUARE	displays generalized R^2
STB	displays standardized estimates
Computational Options	
NOLOGSCALE	performs calculations using normal scaling

The following list describes these options.

ABSFCONV=*value*

specifies the absolute function convergence criterion. Convergence requires a small change in the log-likelihood function in subsequent iterations,

$$|l_i - l_{i-1}| < value$$

where l_i is the value of the log-likelihood function at iteration i. See the section "Convergence Criteria" on page 2338.

AGGREGATE

AGGREGATE= *(variable-list)*

specifies the subpopulations on which the Pearson chi-square test statistic and the likelihood ratio chi-square test statistic (deviance) are calculated. Observations with common values in the given list of variables are regarded as coming from the same subpopulation. Variables in the list can be any variables in the input data set. Specifying the AGGREGATE option is equivalent to specifying the AGGREGATE= option with a variable list that includes all explanatory variables in the MODEL statement. The deviance and Pearson goodness-of-fit statistics are calculated only when the SCALE= option is specified. Thus, the AGGREGATE (or AGGREGATE=) option has no effect if the SCALE= option is not specified. See the section "Rescaling the Covariance Matrix" on page 2354 for more detail.

ALPHA=α

sets the level of significance α for $100(1-\alpha)\%$ confidence intervals for regression parameters or odds ratios. The value α must be between 0 and 1. By default, α is equal to the value of the ALPHA= option in the PROC LOGISTIC statement, or 0.05 if the option is not specified. This option has no effect unless confidence limits for the parameters or odds ratios are requested.

BEST=*n*

specifies that n models with the highest score chi-square statistics are to be displayed for each model size. It is used exclusively with the SCORE model selection method. If the BEST= option is omitted and there are no more than ten explanatory variables, then all possible models are listed for each model size. If the option is omitted and there are more than ten explanatory variables, then the number of models selected for each model size is, at most, equal to the number of explanatory variables listed in the MODEL statement.

CLODDS=PL | WALD | BOTH

requests confidence intervals for the odds ratios. Computation of these confidence intervals is based on the profile likelihood (CLODDS=PL) or based on individual Wald tests (CLODDS=WALD). By specifying CLODDS=BOTH, the procedure computes two sets of confidence intervals for the odds ratios, one based on the profile likelihood and the other based on the Wald tests. The confidence coefficient can be specified with the ALPHA= option.

CLPARM=PL | WALD | BOTH

requests confidence intervals for the parameters. Computation of these confidence intervals is based on the profile likelihood (CLPARM=PL) or individual Wald tests (CLPARM=WALD). By specifying CLPARM=BOTH, the procedure computes two sets of confidence intervals for the parameters, one based on the profile likelihood and the other based on individual Wald tests. The confidence coefficient can be specified with the ALPHA= option. See the "Confidence Intervals for Parameters" section on page 2345 for more information.

CORRB

displays the correlation matrix of the parameter estimates.

COVB

displays the covariance matrix of the parameter estimates.

CTABLE

classifies the input binary response observations according to whether the predicted event probabilities are above or below some cutpoint value z in the range $(0, 1)$. An observation is predicted as an event if the predicted event probability exceeds z. You can supply a list of cutpoints other than the default list by using the PPROB= option (page 2315). The CTABLE option is ignored if the data have more than two response levels. Also, false positive and negative rates can be computed as posterior probabilities using Bayes' theorem. You can use the PEVENT= option to specify prior probabilities for computing these rates. For more information, see the "Classification Table" section on page 2352.

DETAILS

produces a summary of computational details for each step of the effect selection process. It produces the "Analysis of Effects Not in the Model" table before displaying the effect selected for entry for FORWARD or STEPWISE selection. For each model fitted, it produces the "Type 3 Analysis of Effects" table if the fitted model involves CLASS variables, the "Analysis of Maximum Likelihood Estimates" table, and measures of association between predicted probabilities and observed responses. For the statistics included in these tables, see the "Displayed Output" section on page 2381. The DETAILS option has no effect when SELECTION=NONE.

EXPB

EXPEST

displays the exponentiated values ($e^{\hat{\beta}_i}$) of the parameter estimates $\hat{\beta}_i$ in the "Analysis of Maximum Likelihood Estimates" table for the logit model. These exponentiated values are the estimated odds ratios for the parameters corresponding to the continuous explanatory variables.

FAST

uses a computational algorithm of Lawless and Singhal (1978) to compute a first-order approximation to the remaining slope estimates for each subsequent elimination of a variable from the model. Variables are removed from the model based on these approximate estimates. The FAST option is extremely efficient because the model is not refitted for every variable removed. The FAST option is used when SELECTION=BACKWARD and in the backward elimina-

tion steps when SELECTION=STEPWISE. The FAST option is ignored when SELECTION=FORWARD or SELECTION=NONE.

FCONV=*value*

specifies the relative function convergence criterion. Convergence requires a small relative change in the log-likelihood function in subsequent iterations,

$$\frac{|l_i - l_{i-1}|}{|l_{i-1}| + 1\text{E}{-}6} < value$$

where l_i is the value of the log likelihood at iteration i. See the section "Convergence Criteria" on page 2338.

GCONV=*value*

specifies the relative gradient convergence criterion. Convergence requires that the normalized prediction function reduction is small,

$$\frac{\mathbf{g}_i' \mathbf{I}_i \mathbf{g}_i}{|l_i| + 1\text{E}{-}6} < value$$

where l_i is the value of the log-likelihood function, $\mathbf{g}_i$ is the gradient vector, and $\mathbf{I}_i$ is the (expected) information matrix, all at iteration i. This is the default convergence criterion, and the default value is 1E−8. See the section "Convergence Criteria" on page 2338.

HIERARCHY=*keyword*
HIER=*keyword*

specifies whether and how the model hierarchy requirement is applied and whether a single effect or multiple effects are allowed to enter or leave the model in one step. You can specify that only CLASS effects, or both CLASS and interval effects, be subject to the hierarchy requirement. The HIERARCHY= option is ignored unless you also specify one of the following options: SELECTION=FORWARD, SELECTION=BACKWARD, or SELECTION=STEPWISE.

Model hierarchy refers to the requirement that, for any term to be in the model, all effects contained in the term must be present in the model. For example, in order for the interaction A*B to enter the model, the main effects A and B must be in the model. Likewise, neither effect A nor B can leave the model while the interaction A*B is in the model.

The keywords you can specify in the HIERARCHY= option are as follows:

NONE — Model hierarchy is not maintained. Any single effect can enter or leave the model at any given step of the selection process.

SINGLE — Only one effect can enter or leave the model at one time, subject to the model hierarchy requirement. For example, suppose that you specify the main effects A and B and the interaction A*B in the model. In the first step of the selection process, either A or B can enter the model. In the second step, the other main effect can enter

the model. The interaction effect can enter the model only when both main effects have already been entered. Also, before A or B can be removed from the model, the A*B interaction must first be removed. All effects (CLASS and interval) are subject to the hierarchy requirement.

SINGLECLASS This is the same as HIERARCHY=SINGLE except that only CLASS effects are subject to the hierarchy requirement.

MULTIPLE More than one effect can enter or leave the model at one time, subject to the model hierarchy requirement. In a forward selection step, a single main effect can enter the model, or an interaction can enter the model together with all the effects that are contained in the interaction. In a backward elimination step, an interaction itself, or the interaction together with all the effects that the interaction contains, can be removed. All effects (CLASS and interval) are subject to the hierarchy requirement.

MULTIPLECLASS This is the same as HIERARCHY=MULTIPLE except that only CLASS effects are subject to the hierarchy requirement.

The default value is HIERARCHY=SINGLE, which means that model hierarchy is to be maintained for all effects (that is, both CLASS and interval effects) and that only a single effect can enter or leave the model at each step.

INCLUDE=*n*

includes the first n effects in the MODEL statement in every model. By default, INCLUDE=0. The INCLUDE= option has no effect when SELECTION=NONE.

Note that the INCLUDE= and START= options perform different tasks: the INCLUDE= option includes the first n effects variables in every model, whereas the START= option only requires that the first n effects appear in the first model.

INFLUENCE

displays diagnostic measures for identifying influential observations in the case of a binary response model. It has no effect otherwise. For each observation, the INFLUENCE option displays the case number (which is the sequence number of the observation), the values of the explanatory variables included in the final model, and the regression diagnostic measures developed by Pregibon (1981). For a discussion of these diagnostic measures, see the "Regression Diagnostics" section on page 2359. When a STRATA statement is specified, the diagnostics are computed following Storer and Crowley (1985); see the "Regression Diagnostic Details" section on page 2367 for details.

IPLOTS

produces an index plot for each regression diagnostic statistic. An index plot is a scatterplot with the regression diagnostic statistic represented on the y-axis and the case number on the x-axis. See Example 42.6 on page 2422 for an illustration.

ITPRINT

displays the iteration history of the maximum-likelihood model fitting. The ITPRINT option also displays the last evaluation of the gradient vector and the final change in the -2 Log Likelihood.

LACKFIT

LACKFIT<(*n*)>

performs the Hosmer and Lemeshow goodness-of-fit test (Hosmer and Lemeshow 2000) for the case of a binary response model. The subjects are divided into approximately ten groups of roughly the same size based on the percentiles of the estimated probabilities. The discrepancies between the observed and expected number of observations in these groups are summarized by the Pearson chi-square statistic, which is then compared to a chi-square distribution with t degrees of freedom, where t is the number of groups minus n. By default, n=2. A small p-value suggests that the fitted model is not an adequate model. See the "The Hosmer-Lemeshow Goodness-of-Fit Test" section on page 2356 for more information.

LINK=*keyword*

L=*keyword*

specifies the link function linking the response probabilities to the linear predictors. You can specify one of the following keywords. The default is LINK=LOGIT.

CLOGLOG — the complementary log-log function. PROC LOGISTIC fits the binary complementary log-log model when there are two response categories and fits the cumulative complementary log-log model when there are more than two response categories. Aliases: CCLOGLOG, CCLL, CUMCLOGLOG.

GLOGIT — the generalized logit function. PROC LOGISTIC fits the generalized logit model where each nonreference category is contrasted with the reference category. You can use the response variable option REF= to specify the reference category.

LOGIT — the log odds function. PROC LOGISTIC fits the binary logit model when there are two response categories and fits the cumulative logit model when there are more than two response categories. Aliases: CLOGIT, CUMLOGIT.

PROBIT — the inverse standard normal distribution function. PROC LOGISTIC fits the binary probit model when there are two response categories and fits the cumulative probit model when there are more than two response categories. Aliases: NORMIT, CPROBIT, CUMPROBIT.

See the section "Link Functions and the Corresponding Distributions" on page 2334 for details.

MAXFUNCTION=*n*

specifies the maximum number of function calls to perform when maximizing the conditional likelihood. This option is only valid when a STRATA statement is specified. The default values are

- 125 when the number of parameters $p < 40$
- 500 when $40 \leq p < 400$
- 1000 when $p \geq 400$

Since the optimization is terminated only after completing a full iteration, the number of function calls that are actually performed can exceed n. If convergence is not attained, the displayed output and all output data sets created by the procedure contain results based on the last maximum likelihood iteration.

MAXITER=*n*

specifies the maximum number of iterations to perform. By default, MAXITER=25. If convergence is not attained in n iterations, the displayed output and all output data sets created by the procedure contain results that are based on the last maximum likelihood iteration.

MAXSTEP=*n*

specifies the maximum number of times any explanatory variable is added to or removed from the model when SELECTION=STEPWISE. The default number is twice the number of explanatory variables in the MODEL statement. When the MAXSTEP= limit is reached, the stepwise selection process is terminated. All statistics displayed by the procedure (and included in output data sets) are based on the last model fitted. The MAXSTEP= option has no effect when SELECTION=NONE, FORWARD, or BACKWARD.

NOCHECK

disables the checking process to determine whether maximum likelihood estimates of the regression parameters exist. If you are sure that the estimates are finite, this option can reduce the execution time if the estimation takes more than eight iterations. For more information, see the "Existence of Maximum Likelihood Estimates" section on page 2338.

NODUMMYPRINT
NODESIGNPRINT
NODP

suppresses the "Class Level Information" table, which shows how the design matrix columns for the CLASS variables are coded.

NOINT

suppresses the intercept for the binary response model, the first intercept for the ordinal response model (which forces all intercepts to be nonnegative), or all intercepts for the generalized logit model. This can be particularly useful in conditional logistic analysis; see Example 42.10 on page 2443.

NOFIT
performs the global score test without fitting the model. The global score test evaluates the joint significance of the effects in the MODEL statement. No further analyses are performed. If the NOFIT option is specified along with other MODEL statement options, NOFIT takes effect and all other options except LINK=, TECHNIQUE=, and OFFSET= are ignored.

NOLOGSCALE
specifies that computations for the conditional and exact conditional logistic model should be computed using normal scaling. Log-scaling can handle numerically larger problems than normal scaling; however, computations in the log-scale are slower than computations in normal-scale.

OFFSET= *name*
names the offset variable. The regression coefficient for this variable will be fixed at 1.

OUTROC=*SAS-data-set*
OUTR=*SAS-data-set*
creates, for binary response models, an output SAS data set that contains the data necessary to produce the receiver operating characteristic (ROC) curve. See the section "OUTROC= Output Data Set" on page 2378 for the list of variables in this data set.

PARMLABEL
displays the labels of the parameters in the "Analysis of Maximum Likelihood Estimates" table.

PEVENT= *value*
PEVENT= (*list* **)**
specifies one prior probability or a list of prior probabilities for the event of interest. The false positive and false negative rates are then computed as posterior probabilities by Bayes' theorem. The prior probability is also used in computing the rate of correct prediction. For each prior probability in the given list, a classification table of all observations is computed. By default, the prior probability is the total sample proportion of events. The PEVENT= option is useful for stratified samples. It has no effect if the CTABLE option is not specified. For more information, see the section "False Positive and Negative Rates Using Bayes' Theorem" on page 2353. Also see the PPROB= option for information on how the *list* is specified.

PLCL
is the same as specifying CLPARM=PL.

PLCONV= *value*
controls the convergence criterion for confidence intervals based on the profile likelihood function. The quantity *value* must be a positive number, with a default value of 1E−4. The PLCONV= option has no effect if profile likelihood confidence intervals (CLPARM=PL) are not requested.

PLRL
is the same as specifying CLODDS=PL.

PPROB=*value*
PPROB= (*list* **)**
specifies one critical probability value (or cutpoint) or a list of critical probability values for classifying observations with the CTABLE option. Each *value* must be between 0 and 1. A response that has a cross validated predicted probability greater than or equal to the current PPROB= value is classified as an event response. The PPROB= option is ignored if the CTABLE option is not specified.

A classification table for each of several cutpoints can be requested by specifying a list. For example,

```
pprob= (0.3, 0.5 to 0.8 by 0.1)
```

requests a classification of the observations for each of the cutpoints 0.3, 0.5, 0.6, 0.7, and 0.8. If the PPROB= option is not specified, the default is to display the classification for a range of probabilities from the smallest estimated probability (rounded down to the nearest 0.02) to the highest estimated probability (rounded up to the nearest 0.02) with 0.02 increments.

RIDGING=ABSOLUTE | RELATIVE | NONE
specifies the technique used to improve the log-likelihood function when its value in the current iteration is less than that in the previous iteration. If you specify the RIDGING=ABSOLUTE option, the diagonal elements of the negative (expected) Hessian are inflated by adding the ridge value. If you specify the RIDGING=RELATIVE option, the diagonal elements are inflated by a factor of 1 plus the ridge value. If you specify the RIDGING=NONE option, the crude line search method of taking half a step is used instead of ridging. By default, RIDGING=RELATIVE.

RISKLIMITS
RL
WALDRL
is the same as specifying CLODDS=WALD.

ROCEPS= *number*
specifies the criterion for grouping estimated event probabilities that are close to each other for the ROC curve. In each group, the difference between the largest and the smallest estimated event probabilities does not exceed the given value. The value for $number$ must be between 0 and 1; the default value is 1E$-$4. The smallest estimated probability in each group serves as a cutpoint for predicting an event response. The ROCEPS= option has no effect if the OUTROC= option is not specified.

RSQUARE
RSQ
requests a generalized R^2 measure for the fitted model. For more information, see the "Generalized Coefficient of Determination" section on page 2342.

SCALE= *scale*

enables you to supply the value of the dispersion parameter or to specify the method for estimating the dispersion parameter. It also enables you to display the "Deviance and Pearson Goodness-of-Fit Statistics" table. To correct for overdispersion or underdispersion, the covariance matrix is multiplied by the estimate of the dispersion parameter. Valid values for *scale* are as follows:

D | DEVIANCE	specifies that the dispersion parameter be estimated by the deviance divided by its degrees of freedom.
P | PEARSON	specifies that the dispersion parameter be estimated by the Pearson chi-square statistic divided by its degrees of freedom.
WILLIAMS <(*constant*)>	specifies that Williams' method be used to model overdispersion. This option can be used only with the *events/trials* syntax. An optional *constant* can be specified as the scale parameter; otherwise, a scale parameter is estimated under the full model. A set of weights is created based on this scale parameter estimate. These weights can then be used in fitting subsequent models of fewer terms than the full model. When fitting these submodels, specify the computed scale parameter as *constant*. See Example 42.9 on page 2438 for an illustration.
N | NONE	specifies that no correction is needed for the dispersion parameter; that is, the dispersion parameter remains as 1. This specification is used for requesting the deviance and the Pearson chi-square statistic without adjusting for overdispersion.
constant	sets the estimate of the dispersion parameter to be the square of the given *constant*. For example, SCALE=2 sets the dispersion parameter to 4. The value *constant* must be a positive number.

You can use the AGGREGATE (or AGGREGATE=) option to define the subpopulations for calculating the Pearson chi-square statistic and the deviance. In the absence of the AGGREGATE (or AGGREGATE=) option, each observation is regarded as coming from a different subpopulation. For the *events/trials* syntax, each observation consists of n Bernoulli trials, where n is the value of the *trials* variable. For *single-trial* syntax, each observation consists of a single response, and for this setting it is not appropriate to carry out the Pearson or deviance goodness-of-fit analysis. Thus, PROC LOGISTIC ignores specifications SCALE=P, SCALE=D, and SCALE=N when *single-trial* syntax is specified without the AGGREGATE (or AGGREGATE=) option.

The "Deviance and Pearson Goodness-of-Fit Statistics" table includes the Pearson chi-square statistic, the deviance, their degrees of freedom, the ratio of each statistic

divided by its degrees of freedom, and the corresponding p-value. For more information, see the "Overdispersion" section on page 2354.

SELECTION=BACKWARD | B
| FORWARD | F
| NONE | N
| STEPWISE | S
| SCORE

specifies the method used to select the variables in the model. BACKWARD requests backward elimination, FORWARD requests forward selection, NONE fits the complete model specified in the MODEL statement, and STEPWISE requests stepwise selection. SCORE requests best subset selection. By default, SELECTION=NONE. For more information, see the "Effect Selection Methods" section on page 2340.

SEQUENTIAL
SEQ

forces effects to be added to the model in the order specified in the MODEL statement or eliminated from the model in the reverse order specified in the MODEL statement. The model-building process continues until the next effect to be added has an insignificant adjusted chi-square statistic or until the next effect to be deleted has a significant Wald chi-square statistic. The SEQUENTIAL option has no effect when SELECTION=NONE.

SINGULAR=*value*

specifies the tolerance for testing the singularity of the Hessian matrix (Newton-Raphson algorithm) or the expected value of the Hessian matrix (Fisher-scoring algorithm). The Hessian matrix is the matrix of second partial derivatives of the log-likelihood function. The test requires that a pivot for sweeping this matrix be at least this number times a norm of the matrix. Values of the SINGULAR= option must be numeric. By default, *value* is the machine epsilon times 10^7, which is approximately 10^{-9} on most machines.

SLENTRY=*value*
SLE=*value*

specifies the significance level of the score chi-square for entering an effect into the model in the FORWARD or STEPWISE method. Values of the SLENTRY= option should be between 0 and 1, inclusive. By default, SLENTRY=0.05. The SLENTRY= option has no effect when SELECTION=NONE, SELECTION=BACKWARD, or SELECTION=SCORE.

SLSTAY=*value*
SLS=*value*

specifies the significance level of the Wald chi-square for an effect to stay in the model in a backward elimination step. Values of the SLSTAY= option should be between 0 and 1, inclusive. By default, SLSTAY=0.05. The SLSTAY= option has no effect when SELECTION=NONE, SELECTION=FORWARD, or SELECTION=SCORE.

START=*n*

begins the FORWARD, BACKWARD, or STEPWISE effect selection process with the first n effects listed in the MODEL statement. The value of n ranges from 0 to s, where s is the total number of effects in the MODEL statement. The default value of n is s for the BACKWARD method and 0 for the FORWARD and STEPWISE methods. Note that START=n specifies only that the first n effects appear in the first model, while INCLUDE=n requires that the first n effects be included in every model. For the SCORE method, START=n specifies that the smallest models contain n effects, where n ranges from 1 to s; the default value is 1. The START= option has no effect when SELECTION=NONE.

STB

displays the standardized estimates for the parameters for the continuous explanatory variables in the "Analysis of Maximum Likelihood Estimates" table. The standardized estimate of β_i is given by $\hat{\beta}_i/(s/s_i)$, where s_i is the total sample standard deviation for the ith explanatory variable and

$$s = \begin{cases} \pi/\sqrt{3} & \text{Logistic} \\ 1 & \text{Normal} \\ \pi/\sqrt{6} & \text{Extreme-value} \end{cases}$$

For the intercept parameters and parameters associated with a CLASS variable, the standardized estimates are set to missing.

STOP=*n*

specifies the maximum (FORWARD method) or minimum (BACKWARD method) number of effects to be included in the final model. The effect selection process is stopped when n effects are found. The value of n ranges from 0 to s, where s is the total number of effects in the MODEL statement. The default value of n is s for the FORWARD method and 0 for the BACKWARD method. For the SCORE method, STOP=n specifies that the largest models contain n effects, where n ranges from 1 to s; the default value of n is s. The STOP= option has no effect when SELECTION=NONE or STEPWISE.

STOPRES
SR

specifies that the removal or entry of effects be based on the value of the residual chi-square. If SELECTION=FORWARD, then the STOPRES option adds the effects into the model one at a time until the residual chi-square becomes insignificant (until the p-value of the residual chi-square exceeds the SLENTRY= *value*). If SELECTION=BACKWARD, then the STOPRES option removes effects from the model one at a time until the residual chi-square becomes significant (until the p-value of the residual chi-square becomes less than the SLSTAY= *value*). The STOPRES option has no effect when SELECTION=NONE or SELECTION=STEPWISE.

TECHNIQUE=FISHER | NEWTON
TECH=FISHER | NEWTON

specifies the optimization technique for estimating the regression parameters. NEWTON (or NR) is the Newton-Raphson algorithm and FISHER (or FS) is the

Fisher-scoring algorithm. Both techniques yield the same estimates, but the estimated covariance matrices are slightly different except for the case when the LOGIT link is specified for binary response data. The default is TECHNIQUE=FISHER. See the section "Iterative Algorithms for Model-Fitting" on page 2336 for details.

WALDCL
CL
is the same as specifying CLPARM=WALD.

XCONV=*value*
specifies the relative parameter convergence criterion. Convergence requires a small relative parameter change in subsequent iterations,

$$\max_j |\delta_j^{(i)}| < value$$

where

$$\delta_j^{(i)} = \begin{cases} \theta_j^{(i)} - \theta_j^{(i-1)} & |\theta_j^{(i-1)}| < 0.01 \\ \frac{\theta_j^{(i)} - \theta_j^{(i-1)}}{\theta_j^{(i-1)}} & \text{otherwise} \end{cases}$$

and $\theta_j^{(i)}$ is the estimate of the jth parameter at iteration i. See the section "Convergence Criteria" on page 2338.

OUTPUT Statement

OUTPUT < **OUT=***SAS-data-set* >< *options* > **;**

The OUTPUT statement creates a new SAS data set that contains all the variables in the input data set and, optionally, the estimated linear predictors and their standard error estimates, the estimates of the cumulative or individual response probabilities, and the confidence limits for the cumulative probabilities. Regression diagnostic statistics and estimates of cross validated response probabilities are also available for binary response models. Formulas for the statistics are given in the "Linear Predictor, Predicted Probability, and Confidence Limits" section on page 2350, the "Regression Diagnostics" section on page 2359, and, for conditional logistic regression, in the "Conditional Logistic Regression" section on page 2365.

If you use the *single-trial* syntax, the data set also contains a variable named _LEVEL_, which indicates the level of the response that the given row of output is referring to. For instance, the value of the cumulative probability variable is the probability that the response variable is as large as the corresponding value of _LEVEL_. For details, see the section "OUT= Output Data Set in the OUTPUT Statement" on page 2376.

The estimated linear predictor, its standard error estimate, all predicted probabilities, and the confidence limits for the cumulative probabilities are computed for all observations in which the explanatory variables have no missing values, even if the

response is missing. By adding observations with missing response values to the input data set, you can compute these statistics for new observations or for settings of the explanatory variables not present in the data without affecting the model fit.

OUT= *SAS-data-set*
names the output data set. If you omit the OUT= option, the output data set is created and given a default name using the DATAn convention.

The following sections explain options in the OUTPUT statement, divided into statistic options for any type of categorical responses, statistic options only for binary response, and other options. The statistic options specify the statistics to be included in the output data set and name the new variables that contain the statistics. If a STRATA statement is specified, only the PREDICTED=, DFBETAS=, and H= options are available; see the "Regression Diagnostic Details" section on page 2367 for details.

Statistic Options for Any Type of Categorical Response

LOWER=*name*
L=*name*
names the variable containing the lower confidence limits for π, where π is the probability of the event response if *events/trials* syntax or *single-trial* syntax with binary response is specified; for a cumulative model, π is cumulative probability (that is, the probability that the response is less than or equal to the value of _LEVEL_); for the generalized logit model, it is the individual probability (that is, the probability that the response category is represented by the value of _LEVEL_). See the ALPHA= option to set the confidence level.

PREDICTED=*name*
PRED=*name*
PROB=*name*
P=*name*
names the variable containing the predicted probabilities. For the *events/trials* syntax or *single-trial* syntax with binary response, it is the predicted event probability. For a cumulative model, it is the predicted cumulative probability (that is, the probability that the response variable is less than or equal to the value of _LEVEL_); and for the generalized logit model, it is the predicted individual probability (that is, the probability of the response category represented by the value of _LEVEL_).

PREDPROBS=(*keywords***)**
requests individual, cumulative, or cross validated predicted probabilities. Descriptions of the *keywords* are as follows.

INDIVIDUAL | I requests the predicted probability of each response level. For a response variable Y with three levels, 1, 2, and 3, the individual probabilities are Pr(Y=1), Pr(Y=2), and Pr(Y=3).

CUMULATIVE | C requests the cumulative predicted probability of each response level. For a response variable Y with three levels, 1,2, and 3, the cumulative probabilities are Pr(Y≤1), Pr(Y≤2), and Pr(Y≤3). The cumulative probability for the last response level always has the

constant value of 1. For generalized logit models, the cumulative predicted probabilities are not computed and are set to missing.

CROSSVALIDATE | XVALIDATE | X requests the cross validated individual predicted probability of each response level. These probabilities are derived from the leave-one-out principle; that is, dropping the data of one subject and reestimating the parameter estimates. PROC LOGISTIC uses a less expensive one-step approximation to compute the parameter estimates. This option is only valid for binary response models; for nominal and ordinal models, the cross validated probabilities are not computed and are set to missing.

See the "Details of the PREDPROBS= Option" section on page 2322 at the end of this section for further details.

STDXBETA=*name*
names the variable containing the standard error estimates of XBETA (the definition of which follows).

UPPER=*name*
U=*name*
names the variable containing the upper confidence limits for π, where π is the probability of the event response if *events/trials* syntax or *single-trial* syntax with binary response is specified; for a cumulative model, π is cumulative probability (that is, the probability that the response is less than or equal to the value of _LEVEL_); for the generalized logit model, it is the individual probability (that is, the probability that the response category is represented by the value of _LEVEL_). See the ALPHA= option to set the confidence level.

XBETA=*name*
names the variable containing the estimates of the linear predictor $\alpha_i + \boldsymbol{\beta}'\mathbf{x}$, where i is the corresponding ordered value of _LEVEL_.

Statistic Options Only for Binary Response

C=*name*
specifies the confidence interval displacement diagnostic that measures the influence of individual observations on the regression estimates.

CBAR=*name*
specifies the another confidence interval displacement diagnostic, which measures the overall change in the global regression estimates due to deleting an individual observation.

DFBETAS= _ALL_
DFBETAS=*var-list*
specifies the standardized differences in the regression estimates for assessing the effects of individual observations on the estimated regression parameters in the fitted model. You can specify a list of up to $s + 1$ variable names, where s is the number of explanatory variables in the MODEL statement, or you can specify just the

keyword _ALL_. In the former specification, the first variable contains the standardized differences in the intercept estimate, the second variable contains the standardized differences in the parameter estimate for the first explanatory variable in the MODEL statement, and so on. In the latter specification, the DFBETAS statistics are named DFBETA_xxx, where xxx is the name of the regression parameter. For example, if the model contains two variables X1 and X2, the specification DFBETAS=_ALL_ produces three DFBETAS statistics: DFBETA_Intercept, DFBETA_X1, and DFBETA_X2. If an explanatory variable is not included in the final model, the corresponding output variable named in DFBETAS=*var-list* contains missing values.

DIFCHISQ=*name*
specifies the change in the chi-square goodness-of-fit statistic attributable to deleting the individual observation.

DIFDEV=*name*
specifies the change in the deviance attributable to deleting the individual observation.

H=*name*
specifies the diagonal element of the hat matrix for detecting extreme points in the design space.

RESCHI=*name*
specifies the Pearson (Chi) residual for identifying observations that are poorly accounted for by the model.

RESDEV=*name*
specifies the deviance residual for identifying poorly fitted observations.

Other Options

You can specify the following option after a slash.

ALPHA=α
sets the level of significance α for $100(1-\alpha)$% confidence limits for the appropriate response probabilities. The value α must be between 0 and 1. By default, α is equal to the value of the ALPHA= option in the PROC LOGISTIC statement, or 0.05 if that option is not specified.

Details of the PREDPROBS= Option

You can request any of the three given types of predicted probabilities. For example, you can request both the individual predicted probabilities and the cross validated probabilities by specifying PREDPROBS=(I X).

When you specify the PREDPROBS= option, two automatic variables _FROM_ and _INTO_ are included for the *single-trial* syntax and only one variable, _INTO_, is included for the *events/trials* syntax. The _FROM_ variable contains the formatted value of the observed response. The variable _INTO_ contains the formatted value of the response level with the largest individual predicted probability.

If you specify PREDPROBS=INDIVIDUAL, the OUTPUT data set contains k additional variables representing the individual probabilities, one for each response level,

where k is the maximum number of response levels across all BY-groups. The names of these variables have the form IP_*xxx*, where *xxx* represents the particular level. The representation depends on the following situations.

- If you specify *events/trials* syntax, *xxx* is either 'Event' or 'Nonevent'. Thus, the variable containing the event probabilities is named IP_Event and the variable containing the nonevent probabilities is named IP_Nonevent.
- If you specify the *single-trial* syntax with more than one BY group, *xxx* is 1 for the first ordered level of the response, 2 for the second ordered level of the response, . . ., and so forth, as given in the "Response Profile" table. The variable containing the predicted probabilities Pr(Y=1) is named IP_1, where Y is the response variable. Similarly, IP_2 is the name of the variable containing the predicted probabilities Pr(Y=2), and so on.
- If you specify the *single-trial* syntax with no BY-group processing, *xxx* is the left-justified formatted value of the response level (the value may be truncated so that IP_*xxx* does not exceed 32 characters.) For example, if Y is the response variable with response levels 'None', 'Mild', and 'Severe', the variables representing individual probabilities Pr(Y='None'), P(Y='Mild'), and P(Y='Severe') are named IP_None, IP_Mild, and IP_Severe, respectively.

If you specify PREDPROBS=CUMULATIVE, the OUTPUT data set contains k additional variables representing the cumulative probabilities, one for each response level, where k is the maximum number of response levels across all BY-groups. The names of these variables have the form CP_*xxx*, where *xxx* represents the particular response level. The naming convention is similar to that given by PREDPROBS=INDIVIDUAL. The PREDPROBS=CUMULATIVE values are the same as those output by the PREDICT=keyword, but are arranged in variables on each output observation rather than in multiple output observations.

If you specify PREDPROBS=CROSSVALIDATE, the OUTPUT data set contains k additional variables representing the cross validated predicted probabilities of the k response levels, where k is the maximum number of response levels across all BY-groups. The names of these variables have the form XP_*xxx*, where *xxx* represents the particular level. The representation is the same as that given by PREDPROBS=INDIVIDUAL except that for the *events/trials* syntax there are four variables for the cross validated predicted probabilities instead of two:

XP_EVENT_R1E is the cross validated predicted probability of an event when a current event trial is removed.

XP_NONEVENT_R1E is the cross validated predicted probability of a nonevent when a current event trial is removed.

XP_EVENT_R1N is the cross validated predicted probability of an event when a current nonevent trial is removed.

XP_NONEVENT_R1N is the cross validated predicted probability of a nonevent when a current nonevent trial is removed.

The cross validated predicted probabilities are precisely those used in the CTABLE option. See the "Predicted Probability of an Event for Classification" section on page 2352 for details of the computation.

SCORE Statement

SCORE < *options* > **;**

The SCORE statement creates a data set that contains all the data in the DATA= data set together with posterior probabilities and, optionally, prediction confidence intervals. Fit statistics are displayed on request. If you have binary response data, the SCORE statement can be used to create the OUTROC= data set containing data for the ROC curve. You can specify several SCORE statements. FREQ, WEIGHT, and BY statements can be used with the SCORE statements.

See the "Scoring Data Sets" section on page 2362 for more information, and see Example 42.13 on page 2462 for an illustration of how to use this statement.

You can specify the following options:

ALPHA=α

specifies the significance level α for $100(1-\alpha)\%$ confidence intervals. By default, α is equal to the value of the ALPHA= option in the PROC LOGISTIC statement, or 0.05 if that option is not specified. This option has no effect unless the CLM option in the SCORE statement is requested.

CLM

outputs the Wald-test-based confidence limits for the predicted probabilities. This option is not available when the INMODEL= data set is created with the NOCOV option.

DATA=*SAS-data-set*

names the SAS data set that you want to score. If you omit the DATA= option in the SCORE statement, then scoring is performed on the DATA= input data set in the PROC LOGISTIC statement, if specified; otherwise, the DATA=_LAST_ data set is used.

It is not necessary for the DATA= data set in the SCORE statement to contain the response variable unless you are specifying the FITSTAT or OUTROC= option.

Only those variables involved in the fitted model effects are required in the DATA= data set in the SCORE statement. For example, the following code uses forward selection to select effects.

```
proc logistic data=Neuralgia outmodel=sasuser.Model;
   class Treatment Sex;
   model Pain(event='Yes')= Treatment|Sex Age
         / selection=forward sle=.01;
run;
```

Suppose Treatment and Age are the effects selected for the final model. You can score a data set which does not contain the variable Sex since the effect Sex is not in the model that the scoring is based on.

```
proc logistic inmodel=sasuser.Model;
   score data=Neuralgia(drop=Sex);
run;
```

FITSTAT
displays a table of fit statistics. Four statistics are computed: total frequency, total weight, log likelihood, and misclassification rate.

OUT=*SAS-data-set*
names the SAS data set that contains the predicted information. If you omit the OUT= option, the output data set is created and given a default name using the DATAn convention.

OUTROC=*SAS-data-set*
names the SAS data set that contains the ROC curve for the DATA= data set. The ROC curve is computed only for binary response data. See the section "OUTROC= Output Data Set" on page 2378 for the list of variables in this data set.

PRIOR=*SAS-data-set*
names the SAS data set that contains the priors of the response categories. The priors may be values proportional to the prior probabilities; thus, they do not necessarily sum to one. This data set should include a variable named _PRIOR_ that contains the prior probabilities. For events/trials MODEL syntax, this data set should also include an _OUTCOME_ variable that contains the values EVENT and NONEVENT; for single-trial MODEL syntax, this data set should include the response variable that contains the unformatted response categories. See Example 42.13 on page 2462 for an example.

PRIOREVENT=*value*
specifies the prior event probability for a binary response model. If both PRIOR= and PRIOREVENT= options are specified, the PRIOR= option takes precedence.

ROCEPS=*value*
specifies the criterion for grouping estimated event probabilities that are close to each other for the ROC curve. In each group, the difference between the largest and the smallest estimated event probability does not exceed the given value. The $value$ must be between 0 and 1; the default value is 1E$-$4. The smallest estimated probability in each group serves as a cutpoint for predicting an event response. The ROCEPS= option has no effect if the OUTROC= option is not specified.

STRATA Statement

STRATA *variable <(option)>< variable <(option)>...>< / options >* **;**

The STRATA statement names the *variables* that define *strata* or *matched sets* to use in a *stratified conditional logistic regression* of binary response data. Observations having the same variable levels are in the same matched set. At least one variable must be specified to invoke the stratified analysis, and the usual unconditional asymptotic analysis is not performed. The stratified logistic model has the form

$$\text{logit}(\pi_{hi}) = \alpha_h + \mathbf{x}'_{hi}\boldsymbol{\beta}$$

where π_{hi} is the event probability for the ith observation in stratum h having covariates $\mathbf{x}_{hi}$, and where the stratum-specific intercepts α_h are the nuisance parameters which are to be conditioned out.

STRATA variables can also be specified in the MODEL statement as classification or continuous covariates; however, the effects are nondegenerate only when crossed with a non-stratification variable. Specifying several STRATA statements is the same as specifying one STRATA statement containing all the strata variables. The STRATA variables can be either character or numeric, and the formatted values of the STRATA variables determine the levels. Thus, you can use also use formats to group values into levels. See the discussion of the FORMAT procedure in the *SAS Procedures Guide*.

If an EXACT statement is also specified, then a stratified *exact* conditional logistic regression is performed.

The SCORE and WEIGHT statements are not available with a STRATA statement. The following MODEL options are also not supported with a STRATA statement: CLPARM=PL, CLODDS=PL, CTABLE, LACKFIT, LINK=, NOFIT, OUTMODEL=, OUTROC=, and SCALE=.

The "Strata Summary" table is displayed by default; it displays the number of strata which have a specific number of events and nonevents. For example, if you are analyzing a 1:5 matched study, this table enables you to verify that every stratum in the analysis has exactly one event and five non-events. Strata containing only events or only non-events are reported in this table, but such strata are uninformative and are not used in the analysis. (Note that you can use the response variable option EVENT= to identify the events; otherwise, the first ordered response category is the event.)

The following option can be specified for a stratification variable by enclosing the option in parentheses after the variable name, or it can be specified globally for all STRATA variables after a slash (/).

MISSING
treats missing values ('.', '.A',...,'.Z' for numeric variables and blanks for character variables) as valid STRATA variable values.

The following strata options are also available after the slash.

NOSUMMARY
suppresses the display of the "Strata Summary" table.

INFO
displays the "Strata Information" table, which includes the stratum number, levels of the STRATA variables that define the stratum, the number of events, the number of nonevents, and the total frequency for each stratum. Since the number of strata can be very large, this table is only displayed on request.

TEST Statement

< *label:* > **TEST** *equation1* < , . . . , < *equationk* >>< / *option* > **;**

The TEST statement tests linear hypotheses about the regression coefficients. The Wald test is used to test jointly the null hypotheses (H_0: $\mathbf{L}\boldsymbol{\theta} = \mathbf{c}$) specified in a single TEST statement. When $\mathbf{c} = \mathbf{0}$ you should specify a CONTRAST statement instead.

Each *equation* specifies a linear hypothesis (a row of the $\mathbf{L}$ matrix and the corresponding element of the $\mathbf{c}$ vector); multiple *equations* are separated by commas. The label, which must be a valid SAS name, is used to identify the resulting output and should always be included. You can submit multiple TEST statements.

The form of an *equation* is as follows:

term < ±*term* . . . > < = ±*term* < ±*term* . . . >>

where *term* is a parameter of the model, or a constant, or a constant times a parameter. For a binary response model, the intercept parameter is named INTERCEPT; for an ordinal response model, the intercept parameters are named INTERCEPT, INTERCEPT2, INTERCEPT3, and so on. See the "Parameter Names in the OUTEST= Data Set" section on page 2375 for details on parameter naming conventions. When no equal sign appears, the expression is set to 0. The following code illustrates possible uses of the TEST statement:

```
proc logistic;
   model y= a1 a2 a3 a4;
   test1: test intercept + .5 * a2 = 0;
   test2: test intercept + .5 * a2;
   test3: test a1=a2=a3;
   test4: test a1=a2, a2=a3;
run;
```

Note that the first and second TEST statements are equivalent, as are the third and fourth TEST statements.

You can specify the following option in the TEST statement after a slash(/).

PRINT
displays intermediate calculations in the testing of the null hypothesis H_0: $\mathbf{L}\boldsymbol{\theta} = \mathbf{c}$. This includes $\mathbf{L}\widehat{\mathbf{V}}(\widehat{\boldsymbol{\theta}})\mathbf{L}'$ bordered by $(\mathbf{L}\widehat{\boldsymbol{\theta}} - \mathbf{c})$ and $[\mathbf{L}\widehat{\mathbf{V}}(\widehat{\boldsymbol{\theta}})\mathbf{L}']^{-1}$ bordered by $[\mathbf{L}\widehat{\mathbf{V}}(\widehat{\boldsymbol{\theta}})\mathbf{L}']^{-1}(\mathbf{L}\widehat{\boldsymbol{\theta}} - \mathbf{c})$, where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator of $\boldsymbol{\theta}$ and $\widehat{\mathbf{V}}(\widehat{\boldsymbol{\theta}})$ is the estimated covariance matrix of $\hat{\boldsymbol{\theta}}$.

For more information, see the "Testing Linear Hypotheses about the Regression Coefficients" section on page 2358.

UNITS Statement

UNITS *independent1 = list1* $<$ *. . . independentk = listk* $>$ $<$ */ option* $>$ **;**

The UNITS statement enables you to specify units of change for the continuous explanatory variables so that customized odds ratios can be estimated. An estimate of the corresponding odds ratio is produced for each unit of change specified for an explanatory variable. The UNITS statement is ignored for CLASS variables. If the CLODDS= option is specified in the MODEL statement, the corresponding confidence limits for the odds ratios are also displayed.

The term *independent* is the name of an explanatory variable and *list* represents a list of units of change, separated by spaces, that are of interest for that variable. Each unit of change in a list has one of the following forms:

- *number*
- SD or $-$SD
- *number* * SD

where *number* is any nonzero number, and SD is the sample standard deviation of the corresponding independent variable. For example, $X = -2$ requests an odds ratio that represents the change in the odds when the variable X is decreased by two units. $X = 2*$SD requests an estimate of the change in the odds when X is increased by two sample standard deviations.

You can specify the following option in the UNITS statement after a slash(/).

DEFAULT= *list*
gives a list of units of change for all explanatory variables that are not specified in the UNITS statement. Each unit of change can be in any of the forms described previously. If the DEFAULT= option is not specified, PROC LOGISTIC does not produce customized odds ratio estimates for any explanatory variable that is not listed in the UNITS statement.

For more information, see the "Odds Ratio Estimation" section on page 2347.

WEIGHT Statement

WEIGHT *variable* $<$ */ option* $>$ **;**

When a WEIGHT statement appears, each observation in the input data set is weighted by the value of the WEIGHT variable. The values of the WEIGHT variable can be nonintegral and are not truncated. Observations with negative, zero, or missing values for the WEIGHT variable are not used in the model fitting. When the WEIGHT statement is not specified, each observation is assigned a weight of 1.

If a SCORE statement is specified, then the WEIGHT variable is used for computing fit statistics and the ROC curve, but it is not required for scoring. If the DATA= data

set in the SCORE statement does not contain the WEIGHT variable, the weights are assumed to be 1 and a warning message is issued in the LOG. If you fit a model and perform the scoring in the same run, the same WEIGHT variable is used for fitting and scoring. If you fit a model in a previous run and input it with the INMODEL= option in the current run, then the WEIGHT variable can be different from the one used in the previous run; however, if a WEIGHT variable was not specified in the previous run you can still specify a WEIGHT variable in the current run.

The following option can be added to the WEIGHT statement after a slash (/).

NORMALIZE
NORM
causes the weights specified by the WEIGHT variable to be normalized so that they add up to the actual sample size. With this option, the estimated covariance matrix of the parameter estimators is invariant to the scale of the WEIGHT variable.

Details

Missing Values

Any observation with missing values for the response, offset, strata, or explanatory variables is excluded from the analysis; however, missing values are valid for variables specified with the MISSING option in the CLASS or STRATA statements. The estimated linear predictor and its standard error estimate, the fitted probabilities and confidence limits, and the regression diagnostic statistics are not computed for any observation with missing offset or explanatory variable values. However, if only the response value is missing, the linear predictor, its standard error, the fitted individual and cumulative probabilities, and confidence limits for the cumulative probabilities can be computed and output to a data set using the OUTPUT statement.

Response Level Ordering

Response level ordering is important because, by default, PROC LOGISTIC models the probability of response levels with *lower Ordered Value*. Ordered Values are assigned to response levels in ascending sorted order (that is, the lowest response level is assigned Ordered Value 1, the next lowest is assigned Ordered Value 2, and so on) and are displayed in the "Response Profiles" table. If your response variable Y takes values in $\{1, \ldots, k+1\}$, then, by default, the functions modeled with the cumulative model are

$$\text{logit}(\Pr(Y \leq i|\mathbf{x})), \quad i = 1, \ldots, k$$

and for the generalized logit model the functions modeled are

$$\log\left(\frac{\Pr(Y = i|\mathbf{x})}{\Pr(Y = k+1|\mathbf{x})}\right), \quad i = 1, \ldots, k$$

where the highest Ordered Value $Y = k+1$ is the reference level. You can change which probabilities are modeled by specifying the EVENT=, REF=, DESCENDING, or ORDER= response variable options in the MODEL statement.

For binary response data with event and nonevent categories, if your event category has a higher Ordered Value, then the nonevent is modeled and, since the default response function modeled is

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

where π is the probability of the response level assigned Ordered Value 1, and since

$$\text{logit}(\pi) = -\text{logit}(1-\pi)$$

the effect of reversing the order of the two response values is to change the signs of α and $\boldsymbol{\beta}$ in the model $\text{logit}(\pi) = \alpha + \boldsymbol{\beta}'\boldsymbol{x}$.

For example, suppose the binary response variable Y takes the values 1 and 0 for event and nonevent, respectively, and Exposure is the explanatory variable. By default, PROC LOGISTIC assigns Ordered Value 1 to response level Y=0, and Ordered Value 2 to response level Y=1. As a result, PROC LOGISTIC models the probability of the nonevent (Ordered Value=1) category. To model the event without changing the values of the variable Y, you can do the following:

- Explicitly state which response level is to be modeled using the response variable option EVENT= in the MODEL statement,

```
model Y(event='1') = Exposure;
```

- Specify the response variable option REF= in the MODEL statement as the nonevent category for the response variable. This option is most useful for generalized logit models.

```
model Y(ref='0') = Exposure;
```

- Specify the response variable option DESCENDING in the MODEL statement,

```
model Y(descending)=Exposure;
```

- Assign a format to Y such that the first formatted value (when the formatted values are put in sorted order) corresponds to the event. For this example, Y=1 is assigned formatted value 'event' and Y=0 is assigned formatted value 'nonevent'. Since ORDER=FORMATTED by default, Ordered Value 1 is assigned to response level Y=1 so the procedure models the event.

```
proc format;
   value Disease 1='event' 0='nonevent';
run;
proc logistic;
   format Y Disease.;
   model Y=Exposure;
run;
```

CLASS Variable Parameterization

Consider a model with one CLASS variable A with four levels, 1, 2, 5, and 7. Details of the possible choices for the PARAM= option follow.

EFFECT

Three columns are created to indicate group membership of the nonreference levels. For the reference level, all three design variables have a value of -1. For instance, if the reference level is 7 (REF='7'), the design matrix columns for A are as follows.

Effect Coding			
	Design Matrix		
A	**A1**	**A2**	**A5**
1	1	0	0
2	0	1	0
5	0	0	1
7	−1	−1	−1

Parameter estimates of CLASS main effects using the effect coding scheme estimate the difference in the effect of each nonreference level compared to the average effect over all 4 levels.

Caution: PROC LOGISTIC initially parameterizes the CLASS variables by looking at the levels of the variables across the complete data set. If you have an *unbalanced* replication of levels across variables, then the design matrix and the parameter interpretation may be different from what you expect. For instance, suppose that in addition to the four-level variable A discussed above, you have another variable B with two levels, where the fourth level of A only occurs with the first level of B. If your model contains the effect A(B), then the design for A within the second level of B will not be a differential effect. In particular, the design will look like the following.

Effect Coding							
		Design Matrix					
		A(B=1)			A(B=2)		
B	**A**	**A1**	**A2**	**A5**	**A1**	**A2**	**A5**
1	1	1	0	0	0	0	0
1	2	0	1	0	0	0	0
1	5	0	0	1	0	0	0
1	7	−1	−1	−1	0	0	0
2	1	0	0	0	1	0	0
2	2	0	0	0	0	1	0
2	5	0	0	0	0	0	1

PROC LOGISTIC will then detect linear dependency among the last three design variables and set the parameter for A5(B=2) to zero, resulting in an interpretation of these parameters as if they were reference- or dummy-coded. The GLM or REFERENCE parameterization may be more appropriate for such problems.

GLM

As in PROC GLM, four columns are created to indicate group membership. The design matrix columns for A are as follows.

GLM Coding				
	Design Matrix			
A	A1	A2	A5	A7
1	1	0	0	0
2	0	1	0	0
5	0	0	1	0
7	0	0	0	1

Parameter estimates of CLASS main effects using the GLM coding scheme estimate the difference in the effects of each level compared to the last level.

ORDINAL

Three columns are created to indicate group membership of the higher levels of the effect. For the first level of the effect (which for A is 1), all three design variables have a value of 0. The design matrix columns for A are as follows.

Ordinal Coding			
	Design Matrix		
A	A2	A5	A7
1	0	0	0
2	1	0	0
5	1	1	0
7	1	1	1

The first level of the effect is a control or baseline level. Parameter estimates of CLASS main effects using the ORDINAL coding scheme estimate the effect on the response as the ordinal factor is set to each succeeding level. When the parameters for an ordinal main effect have the same sign, the response effect is monotonic across the levels.

POLYNOMIAL

POLY Three columns are created. The first represents the linear term (x), the second represents the quadratic term (x^2), and the third represents the cubic term (x^3), where x is the level value. If the CLASS levels are not numeric, they are translated into 1, 2, 3, ... according to their sorting order. The design matrix columns for A are as follows.

Polynomial Coding			
	Design Matrix		
A	**APOLY1**	**APOLY2**	**APOLY3**
1	1	1	1
2	2	4	8
5	5	25	125
7	7	49	343

REFERENCE

REF Three columns are created to indicate group membership of the nonreference levels. For the reference level, all three design variables have a value of 0. For instance, if the reference level is 7 (REF='7'), the design matrix columns for A are as follows.

Reference Coding			
	Design Matrix		
A	**A1**	**A2**	**A5**
1	1	0	0
2	0	1	0
5	0	0	1
7	0	0	0

Parameter estimates of CLASS main effects using the reference coding scheme estimate the difference in the effect of each nonreference level compared to the effect of the reference level.

ORTHEFFECT The columns are obtained by applying the Gram-Schmidt orthogonalization to the columns for PARAM=EFFECT. The design matrix columns for A are as follows.

Orthogonal Effect Coding			
	Design Matrix		
A	**AOEFF1**	**AOEFF2**	**AOEFF3**
1	1.41421	−0.81650	−0.57735
2	0.00000	1.63299	−0.57735
5	0.00000	0.00000	1.73205
7	−1.41421	−0.81649	−0.57735

ORTHORDINAL The columns are obtained by applying the Gram-Schmidt orthogonalization to the columns for PARAM=ORDINAL. The design matrix columns for A are as follows.

Orthogonal Ordinal Coding			
	Design Matrix		
A	AOORD1	AOORD2	AOORD3
1	−1.73205	0.00000	0.00000
2	0.57735	−1.63299	0.00000
5	0.57735	0.81650	−1.41421
7	0.57735	0.81650	1.41421

ORTHPOLY The columns are obtained by applying the Gram-Schmidt orthogonalization to the columns for PARAM=POLY. The design matrix columns for A are as follows.

Orthogonal Polynomial Coding			
	Design Matrix		
A	AOPOLY1	AOPOLY2	AOPOLY5
1	−1.153	0.907	−0.921
2	−0.734	−0.540	1.473
5	0.524	−1.370	−0.921
7	1.363	1.004	0.368

ORTHREF The columns are obtained by applying the Gram-Schmidt orthogonalization to the columns for PARAM=REFERENCE. The design matrix columns for A are as follows.

Orthogonal Reference Coding			
	Design Matrix		
A	AOREF1	AOREF2	AOREF3
1	1.73205	0.00000	0.00000
2	−0.57735	1.63299	0.00000
5	−0.57735	−0.81650	1.41421
7	−0.57735	−0.81650	−1.41421

Link Functions and the Corresponding Distributions

Four link functions are available in the LOGISTIC procedure. The logit function is the default. To specify a different link function, use the LINK= option in the MODEL statement. The link functions and the corresponding distributions are as follows:

- The logit function

  $$g(p) = \log(p/(1-p))$$

  is the inverse of the cumulative logistic distribution function, which is

  $$F(x) = 1/(1+\exp(-x)) = \exp(x)/(1+\exp(x))$$

- The probit (or normit) function

  $$g(p) = \Phi^{-1}(p)$$

  is the inverse of the cumulative standard normal distribution function, which is

  $$F(x) = \Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} \exp(-z^2/2)dz$$

  Traditionally, the probit function contains the additive constant 5, but throughout PROC LOGISTIC, the terms probit and normit are used interchangeably.

- The complementary log-log function

  $$g(p) = \log(-\log(1-p))$$

  is the inverse of the cumulative extreme-value function (also called the Gompertz distribution), which is

  $$F(x) = 1 - \exp(-\exp(x))$$

- The generalized logit function extends the binary logit link to a vector of levels $(p_1, \ldots, p_{k+1})$ by contrasting each level with a fixed level

  $$g(p_i) = \log(p_i/p_{k+1}) \quad i = 1, \ldots, k$$

The variances of the normal, logistic, and extreme-value distributions are not the same. Their respective means and variances are

Distribution	Mean	Variance
Normal	0	1
Logistic	0	$\pi^2/3$
Extreme-value	$-\gamma$	$\pi^2/6$

where γ is the Euler constant. In comparing parameter estimates using different link functions, you need to take into account the different scalings of the corresponding distributions and, for the complementary log-log function, a possible shift in location. For example, if the fitted probabilities are in the neighborhood of 0.1 to 0.9, then the parameter estimates using the logit link function should be about $\pi/\sqrt{3}$ larger than the estimates from the probit link function.

Determining Observations for Likelihood Contributions

If you use *events/trials* MODEL syntax, each observation is split into two observations. One has response value 1 with a frequency equal to the frequency of the original observation (which is 1 if the FREQ statement is not used) times the value of the *events* variable. The other observation has response value 2 and a frequency equal to the frequency of the original observation times the value of (*trials* − *events*). These two observations will have the same explanatory variable values and the same FREQ and WEIGHT values as the original observation.

For either *single-trial* or *events/trials* syntax, let j index all observations. In other words, for *single-trial* syntax, j indexes the actual observations. And, for *events/trials* syntax, j indexes the observations after splitting (as described previously). If your data set has 30 observations and you use *single-trial* syntax, j has values from 1 to 30; if you use *events/trials* syntax, j has values from 1 to 60.

Suppose the response variable in a cumulative response model can take on the ordered values $1, \ldots, k, k+1$ where k is an integer ≥ 1. The likelihood for the jth observation with ordered response value y_j and explanatory variables vector $\mathbf{x}_j$ is given by

$$L_j = \begin{cases} F(\alpha_1 + \boldsymbol{\beta}'\mathbf{x}_j) & y_j = 1 \\ F(\alpha_i + \boldsymbol{\beta}'\mathbf{x}_j) - F(\alpha_{i-1} + \boldsymbol{\beta}'\mathbf{x}_j) & 1 < y_j = i \leq k \\ 1 - F(\alpha_k + \boldsymbol{\beta}'\mathbf{x}_j) & y_j = k+1 \end{cases}$$

where $F(\cdot)$ is the logistic, normal, or extreme-value distribution function, $\alpha_1, \ldots, \alpha_k$ are ordered intercept parameters, and $\boldsymbol{\beta}$ is the slope parameter vector.

For the generalized logit model, letting the $k+1$st level be the reference level, the intercepts $\alpha_1, \ldots, \alpha_k$ are unordered and the slope vector $\boldsymbol{\beta}_i$ varies with each logit. The likelihood for the jth observation with ordered response value y_j and explanatory variables vector $\mathbf{x}_j$ is given by

$$L_j = \Pr(Y = y_j | \mathbf{x}_j) = \begin{cases} \dfrac{e^{\alpha_i + \mathbf{x}_j'\boldsymbol{\beta}_i}}{1 + \sum_{m=1}^{k} e^{\alpha_m + \mathbf{x}_j'\boldsymbol{\beta}_m}} & 1 \leq y_j = i \leq k \\ \dfrac{1}{1 + \sum_{m=1}^{k} e^{\alpha_m + \mathbf{x}_j'\boldsymbol{\beta}_m}} & y_j = k+1 \end{cases}$$

Iterative Algorithms for Model-Fitting

Two iterative maximum likelihood algorithms are available in PROC LOGISTIC. The default is the Fisher-scoring method, which is equivalent to fitting by iteratively reweighted least squares. The alternative algorithm is the Newton-Raphson method. Both algorithms give the same parameter estimates; however, the estimated covariance matrix of the parameter estimators may differ slightly. This is due to the fact that the Fisher-scoring method is based on the expected information matrix while the Newton-Raphson method is based on the observed information matrix. In the case of a binary logit model, the observed and expected information matrices are identical, resulting in identical estimated covariance matrices for both algorithms. For a generalized logit model, only the Newton-Raphson technique is available. You can use the TECHNIQUE= option to select a fitting algorithm.

Iteratively Reweighted Least-Squares Algorithm (Fisher Scoring)

Consider the multinomial variable $\mathbf{Z}_j = (Z_{1j}, \ldots, Z_{kj})'$ such that

$$Z_{ij} = \begin{cases} 1 & \text{if } Y_j = i \\ 0 & \text{otherwise} \end{cases}$$

With π_{ij} denoting the probability that the *j*th observation has response value *i*, the expected value of $\mathbf{Z}_j$ is $\boldsymbol{\pi}_j = (\pi_{1j}, \ldots, \pi_{kj})'$, and $\pi_{(k+1)j} = 1 - \sum_{i=1}^{k} \pi_{ij}$. The covariance matrix of $\mathbf{Z}_j$ is $\mathbf{V}_j$, which is the covariance matrix of a multinomial random variable for one trial with parameter vector $\boldsymbol{\pi}_j$. Let $\boldsymbol{\theta}$ be the vector of regression parameters; in other words, $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_k, \boldsymbol{\beta}')'$. Let $\mathbf{D}_j$ be the matrix of partial derivatives of $\boldsymbol{\pi}_j$ with respect to $\boldsymbol{\theta}$. The estimating equation for the regression parameters is

$$\sum_j \mathbf{D}_j' \mathbf{W}_j (\mathbf{Z_j} - \boldsymbol{\pi}_\mathbf{j}) = \mathbf{0}$$

where $\mathbf{W}_j = w_j f_j \mathbf{V}_j^-$, w_j and f_j are the WEIGHT and FREQ values of the jth observation, and $\mathbf{V}_j^-$ is a generalized inverse of $\mathbf{V}_j$. PROC LOGISTIC chooses $\mathbf{V}_j^-$ as the inverse of the diagonal matrix with $\boldsymbol{\pi}_j$ as the diagonal.

With a starting value of $\boldsymbol{\theta}_0$, the maximum likelihood estimate of $\boldsymbol{\theta}$ is obtained iteratively as

$$\boldsymbol{\theta}_{m+1} = \boldsymbol{\theta}_m + (\sum_j \mathbf{D}_j' \mathbf{W_j} \mathbf{D}_j)^{-1} \sum_j \mathbf{D}_j' \mathbf{W}_j (\mathbf{Z_j} - \boldsymbol{\pi}_\mathbf{j})$$

where $\mathbf{D}_j$, $\mathbf{W}_j$, and $\boldsymbol{\pi}_j$ are evaluated at $\boldsymbol{\theta}_m$. The expression after the plus sign is the step size. If the likelihood evaluated at $\boldsymbol{\theta}_{m+1}$ is less than that evaluated at $\boldsymbol{\theta}_m$, then $\boldsymbol{\theta}_{m+1}$ is recomputed by step-halving or ridging. The iterative scheme continues until convergence is obtained, that is, until $\boldsymbol{\theta}_{m+1}$ is sufficiently close to $\boldsymbol{\theta}_m$. Then the maximum likelihood estimate of $\boldsymbol{\theta}$ is $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_{m+1}$.

The covariance matrix of $\widehat{\boldsymbol{\theta}}$ is estimated by

$$\widehat{\text{cov}}(\widehat{\boldsymbol{\theta}}) = (\sum_j \widehat{\mathbf{D}}_j' \widehat{\mathbf{W}}_j \widehat{\mathbf{D}}_j)^{-1}$$

where $\widehat{\mathbf{D}}_j$ and $\widehat{\mathbf{W}}_j$ are, respectively, $\mathbf{D}_j$ and $\mathbf{W}_j$ evaluated at $\widehat{\boldsymbol{\theta}}$.

By default, starting values are zero for the slope parameters, and for the intercept parameters, starting values are the observed cumulative logits (that is, logits of the observed cumulative proportions of response). Alternatively, the starting values may be specified with the INEST= option.

Newton-Raphson Algorithm

For cumulative models, let the parameter vector be $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_k, \boldsymbol{\beta}')'$, and for the generalized logit model denote $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_k, \boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_k')'$. The gradient vector and the Hessian matrix are given, respectively, by

$$\begin{aligned} \mathbf{g} &= \sum_j w_j f_j \frac{\partial l_j}{\partial \boldsymbol{\theta}} \\ \mathbf{H} &= \sum_j -w_j f_j \frac{\partial^2 l_j}{\partial \boldsymbol{\theta}^2} \end{aligned}$$

where $l_j = \log L_j$ is the log likelihood for the jth observation. With a starting value of $\boldsymbol{\theta}_0$, the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is obtained iteratively until convergence is obtained:

$$\boldsymbol{\theta}_{m+1} = \boldsymbol{\theta}_m + \mathbf{H}^{-1}\mathbf{g}$$

where $\mathbf{H}$ and $\mathbf{g}$ are evaluated at $\boldsymbol{\theta}_m$. If the likelihood evaluated at $\boldsymbol{\theta}_{m+1}$ is less than that evaluated at $\boldsymbol{\theta}_m$, then $\boldsymbol{\theta}_{m+1}$ is recomputed by step-halving or ridging.

The covariance matrix of $\widehat{\boldsymbol{\theta}}$ is estimated by

$$\widehat{\mathrm{cov}}(\widehat{\boldsymbol{\theta}}) = \widehat{\mathbf{H}}^{-1}$$

where $\widehat{\mathbf{H}}$ is $\mathbf{H}$ evaluated at $\widehat{\boldsymbol{\theta}}$.

Convergence Criteria

Four convergence criteria are allowed, namely, ABSFCONV=, FCONV=, GCONV=, and XCONV=. If you specify more than one convergence criterion, the optimization is terminated as soon as one of the criteria is satisfied. If none of the criteria is specified, the default is GCONV=1E−8.

If you specify a STRATA statement, then all unspecified (or non-default) criteria are also compared to zero. For example, only specifying the criterion XCONV=1e−8 but attaining FCONV=0 terminates the optimization even if the XCONV= criterion is not satisfied, because the log likelihood has reached its maximum.

Existence of Maximum Likelihood Estimates

The likelihood equation for a logistic regression model does not always have a finite solution. Sometimes there is a nonunique maximum on the boundary of the parameter space, at infinity. The existence, finiteness, and uniqueness of maximum likelihood estimates for the logistic regression model depend on the patterns of data points in the observation space (Albert and Anderson 1984; Santner and Duffy 1986). The existence checks are not performed for conditional logistic regression.

Consider a binary response model. Let Y_j be the response of the ith subject and let $\mathbf{x}_j$ be the vector of explanatory variables (including the constant 1 associated with the intercept). There are three mutually exclusive and exhaustive types of data configurations: complete separation, quasi-complete separation, and overlap.

Complete Separation

There is a complete separation of data points if there exists a vector $\mathbf{b}$ that correctly allocates all observations to their response groups; that is,

$$\begin{cases} \mathbf{b}'\mathbf{x}_j > 0 & Y_j = 1 \\ \mathbf{b}'\mathbf{x}_j < 0 & Y_j = 2 \end{cases}$$

This configuration gives nonunique infinite estimates. If the iterative process of maximizing the likelihood function is allowed to continue, the log likelihood diminishes to zero, and the dispersion matrix becomes unbounded.

Quasi-Complete Separation

The data are not completely separable but there is a vector $\mathbf{b}$ such that

$$\begin{cases} \mathbf{b}'\mathbf{x}_j \geq 0 & Y_j = 1 \\ \mathbf{b}'\mathbf{x}_j \leq 0 & Y_j = 2 \end{cases}$$

and equality holds for at least one subject in each response group. This configuration also yields nonunique infinite estimates. If the iterative process of maximizing the likelihood function is allowed to continue, the dispersion matrix becomes unbounded and the log likelihood diminishes to a nonzero constant.

Overlap

If neither complete nor quasi-complete separation exists in the sample points, there is an overlap of sample points. In this configuration, the maximum likelihood estimates exist and are unique.

Complete separation and quasi-complete separation are problems typically encountered with small data sets. Although complete separation can occur with any type of data, quasi-complete separation is not likely with truly continuous explanatory variables.

The LOGISTIC procedure uses a simple empirical approach to recognize the data configurations that lead to infinite parameter estimates. The basis of this approach is that any convergence method of maximizing the log likelihood must yield a solution giving complete separation, if such a solution exists. In maximizing the log likelihood, there is no checking for complete or quasi-complete separation if convergence is attained in eight or fewer iterations. Subsequent to the eighth iteration, the probability of the observed response is computed for each observation. If the probability of the observed response is one for all observations, there is a complete separation

of data points and the iteration process is stopped. If the complete separation of data has not been determined and an observation is identified to have an extremely large probability ($\geq$0.95) of the observed response, there are two possible situations. First, there is overlap in the data set, and the observation is an atypical observation of its own group. The iterative process, if allowed to continue, will stop when a maximum is reached. Second, there is quasi-complete separation in the data set, and the asymptotic dispersion matrix is unbounded. If any of the diagonal elements of the dispersion matrix for the standardized observations vectors (all explanatory variables standardized to zero mean and unit variance) exceeds 5000, quasi-complete separation is declared and the iterative process is stopped. If either complete separation or quasi-complete separation is detected, a warning message is displayed in the procedure output.

Checking for quasi-complete separation is less foolproof than checking for complete separation. The NOCHECK option in the MODEL statement turns off the process of checking for infinite parameter estimates. In cases of complete or quasi-complete separation, turning off the checking process typically results in the procedure failing to converge. The presence of a WEIGHT statement also turns off the checking process.

Effect Selection Methods

Five effect-selection methods are available. The simplest method (and the default) is SELECTION=NONE, for which PROC LOGISTIC fits the complete model as specified in the MODEL statement. The other four methods are FORWARD for forward selection, BACKWARD for backward elimination, STEPWISE for stepwise selection, and SCORE for best subsets selection. These methods are specified with the SELECTION= option in the MODEL statement. Intercept parameters are forced to stay in the model unless the NOINT option is specified.

When SELECTION=FORWARD, PROC LOGISTIC first estimates parameters for effects forced into the model. These effects are the intercepts and the first n explanatory effects in the MODEL statement, where n is the number specified by the START= or INCLUDE= option in the MODEL statement (n is zero by default). Next, the procedure computes the score chi-square statistic for each effect not in the model and examines the largest of these statistics. If it is significant at the SLENTRY= level, the corresponding effect is added to the model. Once an effect is entered in the model, it is never removed from the model. The process is repeated until none of the remaining effects meet the specified level for entry or until the STOP= value is reached.

When SELECTION=BACKWARD, parameters for the complete model as specified in the MODEL statement are estimated unless the START= option is specified. In that case, only the parameters for the intercepts and the first n explanatory effects in the MODEL statement are estimated, where n is the number specified by the START= option. Results of the Wald test for individual parameters are examined. The least significant effect that does not meet the SLSTAY= level for staying in the model is removed. Once an effect is removed from the model, it remains excluded. The process is repeated until no other effect in the model meets the specified level for removal or

until the STOP= value is reached. Backward selection is often less successful than forward or stepwise selection because the full model fit in the first step is the model most likely to result in a complete or quasi-complete separation of response values as described in the previous section.

The SELECTION=STEPWISE option is similar to the SELECTION=FORWARD option except that effects already in the model do not necessarily remain. Effects are entered into and removed from the model in such a way that each forward selection step may be followed by one or more backward elimination steps. The stepwise selection process terminates if no further effect can be added to the model or if the effect just entered into the model is the only effect removed in the subsequent backward elimination.

For SELECTION=SCORE, PROC LOGISTIC uses the branch and bound algorithm of Furnival and Wilson (1974) to find a specified number of models with the highest likelihood score (chi-square) statistic for all possible model sizes, from 1, 2, 3 effect models, and so on, up to the single model containing all of the explanatory effects. The number of models displayed for each model size is controlled by the BEST= option. You can use the START= option to impose a minimum model size, and you can use the STOP= option to impose a maximum model size. For instance, with BEST=3, START=2, and STOP=5, the SCORE selection method displays the best three models (that is, the three models with the highest score chi-squares) containing 2, 3, 4, and 5 effects. The SELECTION=SCORE option is not available for models with CLASS variables.

The options FAST, SEQUENTIAL, and STOPRES can alter the default criteria for entering or removing effects from the model when they are used with the FORWARD, BACKWARD, or STEPWISE selection methods.

Model Fitting Information

Suppose the model contains s explanatory effects. For the jth observation, let $\hat{\pi}_j$ be the estimated probability of the observed response. The three criteria displayed by the LOGISTIC procedure are calculated as follows:

- -2 Log Likelihood:

$$-2 \text{ Log L} = -2 \sum_j w_j f_j \log(\hat{\pi}_j)$$

  where w_j and f_j are the weight and frequency values of the jth observation. For binary response models using *events/trials* MODEL syntax, this is equivalent to

$$-2 \text{ Log L} = -2 \sum_j w_j f_j \{ r_j \log(\hat{\pi}_j) + (n_j - r_j) \log(1 - \hat{\pi}_j) \}$$

  where r_j is the number of events, n_j is the number of trials, and $\hat{\pi}_j$ is the estimated event probability.

- Akaike Information Criterion:

$$\text{AIC} = -2 \text{ Log L} + 2p$$

  where p is the number of parameters in the model. For cumulative response models, $p = k + s$ where k is the total number of response levels minus one, and s is the number of explanatory effects. For the generalized logit model, $p = k(s + 1)$.

- Schwarz Criterion:

$$\text{SC} = -2 \text{ Log L} + p \log(\sum_j f_j)$$

  where p is as defined previously.

The -2 Log Likelihood statistic has a chi-square distribution under the null hypothesis (that all the explanatory effects in the model are zero) and the procedure produces a p-value for this statistic. The AIC and SC statistics give two different ways of adjusting the -2 Log Likelihood statistic for the number of terms in the model and the number of observations used. These statistics should be used when comparing different models for the same data (for example, when you use the METHOD=STEPWISE option in the MODEL statement); lower values of the statistic indicate a more desirable model.

Generalized Coefficient of Determination

Cox and Snell (1989, pp. 208–209) propose the following generalization of the coefficient of determination to a more general linear model:

$$R^2 = 1 - \left\{ \frac{L(\mathbf{0})}{L(\widehat{\boldsymbol{\theta}})} \right\}^{\frac{2}{n}}$$

where $L(\mathbf{0})$ is the likelihood of the intercept-only model, $L(\widehat{\boldsymbol{\theta}})$ is the likelihood of the specified model, and n is the sample size. The quantity R^2 achieves a maximum of less than one for discrete models, where the maximum is given by

$$R^2_{\max} = 1 - \{L(\mathbf{0})\}^{\frac{2}{n}}$$

Nagelkerke (1991) proposes the following adjusted coefficient, which can achieve a maximum value of one:

$$\tilde{R}^2 = \frac{R^2}{R^2_{\max}}$$

Properties and interpretation of R^2 and $\tilde{R}^2$ are provided in Nagelkerke (1991). In the "Testing Global Null Hypothesis: BETA=0" table, R^2 is labeled as "RSquare" and $\tilde{R}^2$ is labeled as "Max-rescaled RSquare." Use the RSQUARE option to request R^2 and $\tilde{R}^2$.

Score Statistics and Tests

To understand the general form of the score statistics, let $\mathbf{U}(\boldsymbol{\theta})$ be the vector of first partial derivatives of the log likelihood with respect to the parameter vector $\boldsymbol{\theta}$, and let $\mathbf{H}(\boldsymbol{\theta})$ be the matrix of second partial derivatives of the log likelihood with respect to $\boldsymbol{\theta}$. That is, $\mathbf{U}(\boldsymbol{\theta})$ is the gradient vector, and $\mathbf{H}(\boldsymbol{\theta})$ is the Hessian matrix. Let $\mathbf{I}(\boldsymbol{\theta})$ be either $-\mathbf{H}(\boldsymbol{\theta})$ or the expected value of $-\mathbf{H}(\boldsymbol{\theta})$. Consider a null hypothesis H_0. Let $\hat{\boldsymbol{\theta}}_0$ be the MLE of $\boldsymbol{\theta}$ under H_0. The chi-square score statistic for testing H_0 is defined by

$$\mathbf{U}'(\hat{\boldsymbol{\theta}}_0)\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}_0)\mathbf{U}(\hat{\boldsymbol{\theta}}_0)$$

and it has an asymptotic χ^2 distribution with r degrees of freedom under H_0, where r is the number of restrictions imposed on $\boldsymbol{\theta}$ by H_0.

Residual Chi-Square

When you use SELECTION=FORWARD, BACKWARD, or STEPWISE, the procedure calculates a residual score chi-square score statistic and reports the statistic, its degrees of freedom, and the p-value. This section describes how the statistic is calculated.

Suppose there are s explanatory effects of interest. The full cumulative response model has a parameter vector

$$\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_s)'$$

where $\alpha_1, \ldots, \alpha_k$ are intercept parameters, and $\beta_1, \ldots, \beta_s$ are the common slope parameters for the explanatory effects, and the full generalized logit model has a parameter vector

$$\begin{aligned} \boldsymbol{\theta} &= (\alpha_1, \ldots, \alpha_k, \boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_k')' \quad \text{with} \\ \boldsymbol{\beta}_i' &= (\beta_{i1}, \ldots, \beta_{is}), \quad i = 1, \ldots, k \end{aligned}$$

where β_{ij} is the slope parameter for the jth effect in the ith logit.

Consider the null hypothesis H_0: $\beta_{t+1} = \ldots = \beta_s = 0$ where $t < s$ for the cumulative response model, and H_0: $\beta_{i,t+1} = \ldots = \beta_{is} = 0, t < s, i = 1, \ldots, k$ for the generalized logit model. For the reduced model with t explanatory effects, let $\hat{\alpha}_1, \ldots, \hat{\alpha}_k$ be the MLEs of the unknown intercept parameters, let $\hat{\beta}_1, \ldots, \hat{\beta}_t$ be the MLEs of the unknown slope parameters, and let $\hat{\boldsymbol{\beta}}_{i(t)}' = (\hat{\beta}_{i1}, \ldots, \hat{\beta}_{it}), i = 1, \ldots, k$ be those for the generalized logit model. The residual chi-square is the chi-square score statistic testing the null hypothesis H_0; that is, the residual chi-square is

$$\mathbf{U}'(\hat{\boldsymbol{\theta}}_0)\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}_0)\mathbf{U}(\hat{\boldsymbol{\theta}}_0)$$

where for the cumulative response model $\hat{\boldsymbol{\theta}}_0 = (\hat{\alpha}_1, \ldots, \hat{\alpha}_k, \hat{\beta}_1, \ldots, \hat{\beta}_t, 0, \ldots, 0)'$, and for the generalized logit model $\hat{\boldsymbol{\theta}}_0 = (\hat{\alpha}_1, \ldots, \hat{\alpha}_k, \hat{\boldsymbol{\beta}}'_{1(t)}, \mathbf{0}'_{(s-t)}, \ldots \hat{\boldsymbol{\beta}}'_{k(t)}, \mathbf{0}'_{(s-t)})'$. where $\mathbf{0}_{(s-t)}$ denote a vector of $s-t$ zeros.

The residual chi-square has an asymptotic chi-square distribution with $s-t$ degrees of freedom ($k(s-t)$ for the generalized logit model). A special case is the global score chi-square, where the reduced model consists of the k intercepts and no explanatory effects. The global score statistic is displayed in the "Testing Global Null Hypothesis: BETA=0" table. The table is not produced when the NOFIT option is used, but the global score statistic is displayed.

Testing Individual Effects Not in the Model

These tests are performed in the FORWARD or STEPWISE method, and are displayed when the DETAILS option is specified. In the displayed output, the tests are labeled "Score Chi-Square" in the "Analysis of Effects Not in the Model" table and in the "Summary of Stepwise (Forward) Selection" table. This section describes how the tests are calculated.

Suppose that k intercepts and t explanatory variables (say $v_1, \ldots, v_t$) have been fitted to a model and that v_{t+1} is another explanatory variable of interest. Consider a full model with the k intercepts and $t+1$ explanatory variables ($v_1, \ldots, v_t, v_{t+1}$) and a reduced model with v_{t+1} excluded. The significance of v_{t+1} adjusted for $v_1, \ldots, v_t$ can be determined by comparing the corresponding residual chi-square with a chi-square distribution with one degree of freedom (k degrees of freedom for the generalized logit model).

Testing the Parallel Lines Assumption

For an ordinal response, PROC LOGISTIC performs a test of the parallel lines assumption. In the displayed output, this test is labeled "Score Test for the Equal Slopes Assumption" when the LINK= option is NORMIT or CLOGLOG. When LINK=LOGIT, the test is labeled as "Score Test for the Proportional Odds Assumption" in the output. For small sample sizes, this test may be too liberal (Stokes, Davis, and Koch 2000). This section describes the methods used to calculate the test.

For this test the number of response levels, $k+1$, is assumed to be strictly greater than 2. Let Y be the response variable taking values $1, \ldots, k, k+1$. Suppose there are s explanatory variables. Consider the general cumulative model without making the parallel lines assumption

$$g(\Pr(Y \le i \mid \mathbf{x})) = (1, \mathbf{x}')\boldsymbol{\theta}_i, \quad 1 \le i \le k$$

where $g(\cdot)$ is the link function, and $\boldsymbol{\theta}_i = (\alpha_i, \beta_{i1}, \ldots, \beta_{is})'$ is a vector of unknown parameters consisting of an intercept α_i and s slope parameters $\beta_{i1}, \ldots, \beta_{is}$. The parameter vector for this general cumulative model is

$$\boldsymbol{\theta} = (\boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_k)'$$

Under the null hypothesis of parallelism H_0: $\beta_{1m} = \beta_{2m} = \cdots = \beta_{km}, 1 \leq m \leq s$, there is a single common slope parameter for each of the s explanatory variables. Let $\beta_1, \ldots, \beta_s$ be the common slope parameters. Let $\hat{\alpha}_1, \ldots, \hat{\alpha}_k$ and $\hat{\beta}_1, \ldots, \hat{\beta}_s$ be the MLEs of the intercept parameters and the common slope parameters. Then, under H_0, the MLE of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}}_0 = (\hat{\boldsymbol{\theta}}_1', \ldots, \hat{\boldsymbol{\theta}}_k')' \quad \text{with} \quad \hat{\boldsymbol{\theta}}_i = (\hat{\alpha}_i, \hat{\beta}_1, \ldots, \hat{\beta}_s)' \quad 1 \leq i \leq k$$

and the chi-squared score statistic $\mathbf{U}'(\hat{\boldsymbol{\theta}}_0)\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}_0)\mathbf{U}(\hat{\boldsymbol{\theta}}_0)$ has an asymptotic chi-square distribution with $s(k-1)$ degrees of freedom. This tests the parallel lines assumption by testing the equality of separate slope parameters simultaneously for all explanatory variables.

Confidence Intervals for Parameters

There are two methods of computing confidence intervals for the regression parameters. One is based on the profile likelihood function, and the other is based on the asymptotic normality of the parameter estimators. The latter is not as time-consuming as the former, since it does not involve an iterative scheme; however, it is not thought to be as accurate as the former, especially with small sample size. You use the CLPARM= option to request confidence intervals for the parameters.

Likelihood Ratio-Based Confidence Intervals

The likelihood ratio-based confidence interval is also known as the profile likelihood confidence interval. The construction of this interval is derived from the asymptotic χ^2 distribution of the generalized likelihood ratio test (Venzon and Moolgavkar 1988). Suppose that the parameter vector is $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_s)'$ and you want to compute a confidence interval for β_j. The profile likelihood function for $\beta_j = \gamma$ is defined as

$$l_j^*(\gamma) = \max_{\boldsymbol{\beta} \in \mathcal{B}_j(\gamma)} l(\boldsymbol{\beta})$$

where $\mathcal{B}_j(\gamma)$ is the set of all $\boldsymbol{\beta}$ with the jth element fixed at γ, and $l(\boldsymbol{\beta})$ is the log-likelihood function for $\boldsymbol{\beta}$. If $l_{\max} = l(\widehat{\boldsymbol{\beta}})$ is the log likelihood evaluated at the maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$, then $2(l_{\max} - l_j^*(\beta_j))$ has a limiting chi-square distribution with one degree of freedom if β_j is the true parameter value. Let $l_0 = l_{\max} - .5\chi_1^2(1-\alpha)$, where $\chi_1^2(1-\alpha)$ is the $100(1-\alpha)$ percentile of the chi-square distribution with one degree of freedom. A $100(1-\alpha)$% confidence interval for β_j is

$$\{\gamma : l_j^*(\gamma) \geq l_0\}$$

The endpoints of the confidence interval are found by solving numerically for values of β_j that satisfy equality in the preceding relation. To obtain an iterative algorithm

for computing the confidence limits, the log-likelihood function in a neighborhood of $\boldsymbol{\beta}$ is approximated by the quadratic function

$$\tilde{l}(\boldsymbol{\beta}+\boldsymbol{\delta}) = l(\boldsymbol{\beta}) + \boldsymbol{\delta}'\mathbf{g} + \frac{1}{2}\boldsymbol{\delta}'\mathbf{V}\boldsymbol{\delta}$$

where $\mathbf{g} = \mathbf{g}(\boldsymbol{\beta})$ is the gradient vector and $\mathbf{V} = \mathbf{V}(\boldsymbol{\beta})$ is the Hessian matrix. The increment $\boldsymbol{\delta}$ for the next iteration is obtained by solving the likelihood equations

$$\frac{d}{d\boldsymbol{\delta}}\{\tilde{l}(\boldsymbol{\beta}+\boldsymbol{\delta}) + \lambda(\mathbf{e}_j'\boldsymbol{\delta} - \gamma)\} = \mathbf{0}$$

where λ is the Lagrange multiplier, $\mathbf{e}_j$ is the jth unit vector, and γ is an unknown constant. The solution is

$$\boldsymbol{\delta} = -\mathbf{V}^{-1}(\mathbf{g} + \lambda\mathbf{e}_j)$$

By substituting this $\boldsymbol{\delta}$ into the equation $\tilde{l}(\boldsymbol{\beta}+\boldsymbol{\delta}) = l_0$, you can estimate λ as

$$\lambda = \pm\left(\frac{2(l_0 - l(\boldsymbol{\beta}) + \frac{1}{2}\mathbf{g}'\mathbf{V}^{-1}\mathbf{g})}{\mathbf{e}_j'\mathbf{V}^{-1}\mathbf{e}_j}\right)^{\frac{1}{2}}$$

The upper confidence limit for β_j is computed by starting at the maximum likelihood estimate of $\boldsymbol{\beta}$ and iterating with positive values of λ until convergence is attained. The process is repeated for the lower confidence limit using negative values of λ.

Convergence is controlled by value ϵ specified with the PLCONV= option in the MODEL statement (the default value of ϵ is 1E$-$4). Convergence is declared on the current iteration if the following two conditions are satisfied:

$$|l(\boldsymbol{\beta}) - l_0| \leq \epsilon$$

and

$$(\mathbf{g} + \lambda\mathbf{e}_j)'\mathbf{V}^{-1}(\mathbf{g} + \lambda\mathbf{e}_j) \leq \epsilon$$

Wald Confidence Intervals

Wald confidence intervals are sometimes called the normal confidence intervals. They are based on the asymptotic normality of the parameter estimators. The $100(1-\alpha)$% Wald confidence interval for β_j is given by

$$\widehat{\beta}_j \pm z_{1-\alpha/2}\widehat{\sigma}_j$$

where z_p is the $100p$th percentile of the standard normal distribution, $\widehat{\beta}_j$ is the maximum likelihood estimate of β_j, and $\widehat{\sigma}_j$ is the standard error estimate of $\widehat{\beta}_j$.

Odds Ratio Estimation

Consider a dichotomous response variable with outcomes *event* and *nonevent*. Consider a dichotomous risk factor variable X that takes the value 1 if the risk factor is present and 0 if the risk factor is absent. According to the logistic model, the log odds function, $g(X)$, is given by

$$g(X) \equiv \log\left(\frac{\Pr(\textit{event} \mid X)}{\Pr(\textit{nonevent} \mid X)}\right) = \beta_0 + \beta_1 X$$

The odds ratio ψ is defined as the ratio of the odds for those with the risk factor ($X = 1$) to the odds for those without the risk factor ($X = 0$). The log of the odds ratio is given by

$$\log(\psi) \equiv \log(\psi(X = 1, X = 0)) = g(X = 1) - g(X = 0) = \beta_1$$

The parameter, β_1, associated with X represents the change in the log odds from $X = 0$ to $X = 1$. So, the odds ratio is obtained by simply exponentiating the value of the parameter associated with the risk factor. The odds ratio indicates how the odds of *event* change as you change X from 0 to 1. For instance, $\psi = 2$ means that the odds of an event when $X = 1$ are twice the odds of an event when $X = 0$.

Suppose the values of the dichotomous risk factor are coded as constants a and b instead of 0 and 1. The odds when $X = a$ become $\exp(\beta_0 + a\beta_1)$, and the odds when $X = b$ become $\exp(\beta_0 + b\beta_1)$. The odds ratio corresponding to an increase in X from a to b is

$$\psi = \exp[(b - a)\beta_1] = [\exp(\beta_1)]^{b-a} \equiv [\exp(\beta_1)]^c$$

Note that for any a and b such that $c = b - a = 1, \psi = \exp(\beta_1)$. So the odds ratio can be interpreted as the change in the odds for any increase of one unit in the corresponding risk factor. However, the change in odds for some amount other than one unit is often of greater interest. For example, a change of one pound in body weight may be too small to be considered important, while a change of 10 pounds may be more meaningful. The odds ratio for a change in X from a to b is estimated by raising the odds ratio estimate for a unit change in X to the power of $c = b - a$ as shown previously.

For a polytomous risk factor, the computation of odds ratios depends on how the risk factor is parameterized. For illustration, suppose that Race is a risk factor with four categories: White, Black, Hispanic, and Other.

For the effect parameterization scheme (PARAM=EFFECT) with White as the reference group, the design variables for Race are as follows.

	Design Variables		
Race	X_1	X_2	X_3
Black	1	0	0
Hispanic	0	1	0
Other	0	0	1
White	−1	−1	−1

The log odds for Black is

$$\begin{aligned} g(\text{Black}) &= \beta_0 + \beta_1(X_1 = 1) + \beta_2(X_2 = 0) + \beta_3(X_3 = 0) \\ &= \beta_0 + \beta_1 \end{aligned}$$

The log odds for White is

$$\begin{aligned} g(\text{White}) &= \beta_0 + \beta_1(X_1 = -1) + \beta_2(X_2 = -1) + \beta_3(X_3 = -1)) \\ &= \beta_0 - \beta_1 - \beta_2 - \beta_3 \end{aligned}$$

Therefore, the log odds ratio of Black versus White becomes

$$\begin{aligned} \log(\psi(\text{Black}, \text{White})) &= g(\text{Black}) - g(\text{White}) \\ &= 2\beta_1 + \beta_2 + \beta_3 \end{aligned}$$

For the reference cell parameterization scheme (PARAM=REF) with White as the reference cell, the design variables for race are as follows.

	Design Variables		
Race	X_1	X_2	X_3
Black	1	0	0
Hispanic	0	1	0
Other	0	0	1
White	0	0	0

The log odds ratio of Black versus White is given by

$$\begin{aligned} &\log(\psi(\text{Black}, \text{White})) \\ &= g(\text{Black}) - g(\text{White}) \\ &= (\beta_0 + \beta_1(X_1 = 1) + \beta_2(X_2 = 0)) + \beta_3(X_3 = 0)) - \\ &\quad (\beta_0 + \beta_1(X_1 = 0) + \beta_2(X_2 = 0) + \beta_3(X_3 = 0)) \\ &= \beta_1 \end{aligned}$$

For the GLM parameterization scheme (PARAM=GLM), the design variables are as follows.

	Design Variables			
Race	X_1	X_2	X_3	X_4
Black	1	0	0	0
Hispanic	0	1	0	0
Other	0	0	1	0
White	0	0	0	1

The log odds ratio of Black versus White is

$$\begin{aligned}
&\log(\psi(\text{Black}, \text{White})) \\
&= g(\text{Black}) - g(\text{White}) \\
&= (\beta_0 + \beta_1(X_1 = 1) + \beta_2(X_2 = 0) + \beta_3(X_3 = 0) + \beta_4(X_4 = 0)) - \\
&\quad (\beta_0 + \beta_1(X_1 = 0) + \beta_2(X_2 = 0) + \beta_3(X_3 = 0) + \beta_4(X_4 = 1)) \\
&= \beta_1 - \beta_4
\end{aligned}$$

Consider the hypothetical example of heart disease among race in Hosmer and Lemeshow (2000, p 56). The entries in the following contingency table represent counts.

	Race			
Disease Status	White	Black	Hispanic	Other
Present	5	20	15	10
Absent	20	10	10	10

The computation of odds ratio of Black versus White for various parameterization schemes is tabulated in the following table.

Odds Ratio of Heart Disease Comparing Black to White					
	Parameter Estimates				
PARAM	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Odds Ratio Estimation
EFFECT	0.7651	0.4774	0.0719		$\exp(2 \times 0.7651 + 0.4774 + 0.0719) = 8$
REF	2.0794	1.7917	1.3863		$\exp(2.0794) = 8$
GLM	2.0794	1.7917	1.3863	0.0000	$\exp(2.0794) = 8$

Since the log odds ratio ($\log(\psi)$) is a linear function of the parameters, the Wald confidence interval for $\log(\psi)$ can be derived from the parameter estimates and the estimated covariance matrix. Confidence intervals for the odds ratios are obtained by exponentiating the corresponding confidence intervals for the log odd ratios. In the displayed output of PROC LOGISTIC, the "Odds Ratio Estimates" table contains the odds ratio estimates and the corresponding 95% Wald confidence intervals. For continuous explanatory variables, these odds ratios correspond to a unit increase in the risk factors.

To customize odds ratios for specific units of change for a continuous risk factor, you can use the UNITS statement to specify a list of relevant units for each explanatory variable in the model. Estimates of these customized odds ratios are given in a separate table. Let (L_j, U_j) be a confidence interval for $\log(\psi)$. The corresponding lower and upper confidence limits for the customized odds ratio $\exp(c\beta_j)$ are $\exp(cL_j)$ and $\exp(cU_j)$, respectively (for $c > 0$), or $\exp(cU_j)$ and $\exp(cL_j)$, respectively (for

$c < 0$). You use the CLODDS= option to request the confidence intervals for the odds ratios.

For a generalized logit model, odds ratios are computed similarly, except k odds ratios are computed for each effect, corresponding to the k logits in the model.

Rank Correlation of Observed Responses and Predicted Probabilities

The predicted mean score of an observation is the sum of the Ordered Values (shown in the Response Profile table) minus one, weighted by the corresponding predicted probabilities for that observation; that is, the predicted means score$= \sum_{i=1}^{k+1}(i-1)\hat{\pi}_i$, where $k+1$ is the number of response levels and $\hat{\pi}_i$ is the predicted probability of the ith (ordered) response.

A pair of observations with different observed responses is said to be *concordant* if the observation with the lower ordered response value has a lower predicted mean score than the observation with the higher ordered response value. If the observation with the lower ordered response value has a higher predicted mean score than the observation with the higher ordered response value, then the pair is *discordant*. If the pair is neither concordant nor discordant, it is a *tie*. Enumeration of the total numbers of concordant and discordant pairs is carried out by categorizing the predicted mean score into intervals of length $k/500$ and accumulating the corresponding frequencies of observations.

Let N be the sum of observation frequencies in the data. Suppose there is a total of t pairs with different responses, n_c of them are concordant, n_d of them are discordant, and $t - n_c - n_d$ of them are tied. PROC LOGISTIC computes the following four indices of rank correlation for assessing the predictive ability of a model:

$$
\begin{aligned}
c &= (n_c + 0.5(t - n_c - n_d))/t \\
\text{Somers' } D &= (n_c - n_d)/t \\
\text{Goodman-Kruskal Gamma} &= (n_c - n_d)/(n_c + n_d) \\
\text{Kendall's Tau-}a &= (n_c - n_d)/(0.5N(N-1))
\end{aligned}
$$

Note that c also gives an estimate of the area under the receiver operating characteristic (ROC) curve when the response is binary (Hanley and McNeil 1982).

For binary responses, the predicted mean score is equal to the predicted probability for Ordered Value 2. As such, the preceding definition of concordance is consistent with the definition used in previous releases for the binary response model.

Linear Predictor, Predicted Probability, and Confidence Limits

This section describes how predicted probabilities and confidence limits are calculated using the maximum likelihood estimates (MLEs) obtained from PROC LOGISTIC. For a specific example, see the "Getting Started" section on page 2284. Predicted probabilities and confidence limits can be output to a data set with the OUTPUT statement.

Cumulative Response Models

For a vector of explanatory variables $\mathbf{x}$, the linear predictor

$$\eta_i = g(\Pr(Y \le i \mid \mathbf{x})) = \alpha_i + \boldsymbol{\beta}'\mathbf{x} \quad 1 \le i \le k$$

is estimated by

$$\hat{\eta}_i = \hat{\alpha}_i + \hat{\boldsymbol{\beta}}'\mathbf{x}$$

where $\hat{\alpha}_i$ and $\hat{\boldsymbol{\beta}}$ are the MLEs of α_i and $\boldsymbol{\beta}$. The estimated standard error of η_i is $\hat{\sigma}(\hat{\eta}_i)$, which can be computed as the square root of the quadratic form $(1, \mathbf{x}')\hat{\mathbf{V}}_{\mathbf{b}}(1, \mathbf{x}')'$ where $\hat{\mathbf{V}}_{\mathbf{b}}$ is the estimated covariance matrix of the parameter estimates. The asymptotic $100(1-\alpha)\%$ confidence interval for η_i is given by

$$\hat{\eta}_i \pm z_{\alpha/2}\hat{\sigma}(\hat{\eta}_i)$$

where $z_{\alpha/2}$ is the $100(1-\alpha/2)$ percentile point of a standard normal distribution.

The predicted value and the $100(1-\alpha)\%$ confidence limits for $\Pr(Y \le i \mid \mathbf{x})$ are obtained by back-transforming the corresponding measures for the linear predictor.

Link	Predicted Probability	100(1-α) Confidence Limits
LOGIT	$1/(1+e^{-\hat{\eta}_i})$	$1/(1+\mathrm{e}^{-\hat{\eta}_i \pm z_{\alpha/2}\hat{\sigma}(\hat{\eta}_i)})$
PROBIT	$\Phi(\hat{\eta}_i)$	$\Phi(\hat{\eta}_i \pm z_{\alpha/2}\hat{\sigma}(\hat{\eta}_i))$
CLOGLOG	$1-\mathrm{e}^{-\mathrm{e}^{\hat{\eta}_i}}$	$1-\mathrm{e}^{-\mathrm{e}^{\hat{\eta}_i \pm z_{\alpha/2}\hat{\sigma}(\hat{\eta}_i)}}$

Generalized Logit Model

For a vector of explanatory variables $\mathbf{x}$, let π_i denote the probability of obtaining the response value i:

$$\pi_i = \begin{cases} \pi_{k+1}e^{\alpha_i + \mathbf{x}'\boldsymbol{\beta}_i} & 1 \le i \le k \\ \dfrac{1}{1+\sum_{i=1}^{k} e^{\alpha_i + \mathbf{x}'\boldsymbol{\beta}_i}} & i = k+1 \end{cases}$$

By the *delta method*,

$$\sigma^2(\pi_i) = \left(\frac{\partial \pi_i}{\partial \boldsymbol{\theta}}\right)' \mathbf{V}(\boldsymbol{\theta}) \frac{\partial \pi_i}{\partial \boldsymbol{\theta}}$$

A 100(1$-\alpha$)% confidence level for π_i is given by

$$\hat{\pi}_i \pm z_{\alpha/2}\hat{\sigma}(\hat{\pi}_i)$$

where $\hat{\pi}_i$ is the estimated expected probability of response i, and $\hat{\sigma}(\hat{\pi}_i)$ is obtained by evaluating $\sigma(\pi_i)$ at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

Classification Table

For binary response data, the response is either an *event* or a *nonevent*. In PROC LOGISTIC, the response with Ordered Value 1 is regarded as the *event*, and the response with Ordered Value 2 is the *nonevent*. PROC LOGISTIC models the probability of the *event*. From the fitted model, a predicted *event* probability can be computed for each observation. The method to compute a reduced-bias estimate of the predicted probability is given in the "Predicted Probability of an Event for Classification" section on page 2352, which follows. If the predicted *event* probability exceeds some cutpoint value $z \in [0, 1]$, the observation is predicted to be an *event* observation; otherwise, it is predicted as a *nonevent*. A 2×2 frequency table can be obtained by cross-classifying the observed and predicted responses. The CTABLE option produces this table, and the PPROB= option selects one or more cutpoints. Each cutpoint generates a classification table. If the PEVENT= option is also specified, a classification table is produced for each combination of PEVENT= and PPROB= values.

The accuracy of the classification is measured by its *sensitivity* (the ability to predict an *event* correctly) and specificity (the ability to predict a *nonevent* correctly). *Sensitivity* is the proportion of *event* responses that were predicted to be *events*. *Specificity* is the proportion of *nonevent* responses that were predicted to be *nonevents*. PROC LOGISTIC also computes three other conditional probabilities: *false positive rate*, *false negative rate*, and *rate of correct classification*. The *false positive rate* is the proportion of predicted *event* responses that were observed as *nonevents*. The *false negative rate* is the proportion of predicted *nonevent* responses that were observed as *events*. Given prior probabilities specified with the PEVENT= option, these conditional probabilities can be computed as posterior probabilities using Bayes' theorem.

Predicted Probability of an Event for Classification

When you classify a set of binary data, if the same observations used to fit the model are also used to estimate the classification error, the resulting error-count estimate is biased. One way of reducing the bias is to remove the binary observation to be classified from the data, reestimate the parameters of the model, and then classify the observation based on the new parameter estimates. However, it would be costly to fit the model leaving out each observation one at a time. The LOGISTIC procedure provides a less expensive one-step approximation to the preceding parameter estimates. Let $\mathbf{b}$ be the MLE of the parameter vector $(\alpha, \boldsymbol{\beta}')'$ based on all observations. Let $\mathbf{b}_j$ denote the MLE computed without the jth observation. The one-step estimate of $\mathbf{b}_j$ is given by

$$\mathbf{b}_j^1 = \mathbf{b} - \frac{w_j(y_j - \hat{\pi}_j)}{1 - h_{jj}} \hat{\mathbf{V}}_{\mathbf{b}} \begin{pmatrix} 1 \\ \mathbf{x}_j \end{pmatrix}$$

where

y_j is 1 for an event response and 0 otherwise

w_j is the WEIGHT value

$\hat{\pi}_j$ is the predicted event probability based on $\mathbf{b}$

h_{jj} is the hat diagonal element (defined on page 2359) with $n_j = 1$ and $r_j = y_j$

$\hat{\mathbf{V}}_{\mathbf{b}}$ is the estimated covariance matrix of $\mathbf{b}$

False Positive and Negative Rates Using Bayes' Theorem

Suppose n_1 of n individuals experience an event, for example, a disease. Let this group be denoted by $\mathcal{C}_1$, and let the group of the remaining $n_2 = n - n_1$ individuals who do not have the disease be denoted by $\mathcal{C}_2$. The jth individual is classified as giving a positive response if the predicted probability of disease ($\hat{\pi}_j^*$) is large. The probability $\hat{\pi}_j^*$ is the reduced-bias estimate based on a one-step approximation given in the preceding section. For a given cutpoint z, the jth individual is predicted to give a positive response if $\hat{\pi}_j^* \geq z$.

Let B denote the event that a subject has the disease and $\bar{B}$ denote the event of not having the disease. Let A denote the event that the subject responds positively, and let $\bar{A}$ denote the event of responding negatively. Results of the classification are represented by two conditional probabilities, $\Pr(A|B)$ and $\Pr(A|\bar{B})$, where $\Pr(A|B)$ is the sensitivity, and $\Pr(A|\bar{B})$ is one minus the specificity.

These probabilities are given by

$$\Pr(A|B) = \frac{\sum_{j\in\mathcal{C}_1} I(\hat{\pi}_j^* \geq z)}{n_1}$$
$$\Pr(A|\bar{B}) = \frac{\sum_{j\in\mathcal{C}_2} I(\hat{\pi}_j^* \geq z)}{n_2}$$

where $I(\cdot)$ is the indicator function.

Bayes' theorem is used to compute the error rates of the classification. For a given prior probability $\Pr(B)$ of the disease, the false positive rate P_{F+} and the false negative rate P_{F-} are given by Fleiss (1981, pp. 4–5) as follows:

$$P_{F+} = \Pr(\bar{B}|A) = \frac{\Pr(A|\bar{B})[1 - \Pr(B)]}{\Pr(A|\bar{B}) + \Pr(B)[\Pr(A|B) - \Pr(A|\bar{B})]}$$
$$P_{F-} = \Pr(B|\bar{A}) = \frac{[1 - \Pr(A|B)]\Pr(B)}{1 - \Pr(A|\bar{B}) - \Pr(B)[\Pr(A|B) - \Pr(A|\bar{B})]}$$

The prior probability $\Pr(B)$ can be specified by the PEVENT= option. If the PEVENT= option is not specified, the sample proportion of diseased individuals is used; that is, $\Pr(B) = n_1/n$. In such a case, the false positive rate and the false negative rate reduce to

$$P_{F+} = \frac{\sum_{j\in\mathcal{C}_2} I(\hat{\pi}_j^* \geq z)}{\sum_{j\in\mathcal{C}_1} I(\hat{\pi}_j^* \geq z) + \sum_{j\in\mathcal{C}_2} I(\hat{\pi}_j^* \geq z)}$$
$$P_{F-} = \frac{\sum_{j\in\mathcal{C}_1} I(\hat{\pi}_j^* < z)}{\sum_{j\in\mathcal{C}_1} I(\hat{\pi}_j^* < z) + \sum_{j\in\mathcal{C}_2} I(\hat{\pi}_j^* < z)}$$

Note that for a stratified sampling situation in which n_1 and n_2 are chosen a priori, n_1/n is not a desirable estimate of $\Pr(B)$. For such situations, the PEVENT= option should be specified.

Overdispersion

For a correctly specified model, the Pearson chi-square statistic and the deviance, divided by their degrees of freedom, should be approximately equal to one. When their values are much larger than one, the assumption of binomial variability may not be valid and the data are said to exhibit overdispersion. Underdispersion, which results in the ratios being less than one, occurs less often in practice.

When fitting a model, there are several problems that can cause the goodness-of-fit statistics to exceed their degrees of freedom. Among these are such problems as outliers in the data, using the wrong link function, omitting important terms from the model, and needing to transform some predictors. These problems should be eliminated before proceeding to use the following methods to correct for overdispersion.

Rescaling the Covariance Matrix

One way of correcting overdispersion is to multiply the covariance matrix by a dispersion parameter. This method assumes that the sample sizes in each subpopulation are approximately equal. You can supply the value of the dispersion parameter directly, or you can estimate the dispersion parameter based on either the Pearson chi-square statistic or the deviance for the fitted model.

The Pearson chi-square statistic χ^2_P and the deviance χ^2_D are given by

$$\chi^2_P = \sum_{i=1}^{m}\sum_{j=1}^{k+1}\frac{(r_{ij}-n_i\hat{\pi}_{ij})^2}{n_i\hat{\pi}_{ij}}$$

$$\chi^2_D = 2\sum_{i=1}^{m}\sum_{j=1}^{k+1} r_{ij}\log\left(\frac{r_{ij}}{n_i\hat{\pi}_{ij}}\right)$$

where m is the number of subpopulation profiles, $k+1$ is the number of response levels, r_{ij} is the total weight (sum of the product of the frequencies and the weights) associated with jth level responses in the ith profile, $n_i = \sum_{j=1}^{k+1} r_{ij}$, and $\hat{\pi}_{ij}$ is the fitted probability for the jth level at the ith profile. Each of these chi-square statistics has $mk-p$ degrees of freedom, where p is the number of parameters estimated. The dispersion parameter is estimated by

$$\widehat{\sigma}^2 = \begin{cases} \chi^2_P/(mk-p) & \text{SCALE=PEARSON} \\ \chi^2_D/(mk-p) & \text{SCALE=DEVIANCE} \\ (\textit{constant})^2 & \text{SCALE=}\textit{constant} \end{cases}$$

In order for the Pearson statistic and the deviance to be distributed as chi-square, there must be sufficient replication within the subpopulations. When this is not true, the data are sparse, and the p-values for these statistics are not valid and should be

ignored. Similarly, these statistics, divided by their degrees of freedom, cannot serve as indicators of overdispersion. A large difference between the Pearson statistic and the deviance provides some evidence that the data are too sparse to use either statistic.

You can use the AGGREGATE (or AGGREGATE=) option to define the subpopulation profiles. If you do not specify this option, each observation is regarded as coming from a separate subpopulation. For *events/trials* syntax, each observation represents n Bernoulli trials, where n is the value of the *trials* variable; for *single-trial* syntax, each observation represents a single trial. Without the AGGREGATE (or AGGREGATE=) option, the Pearson chi-square statistic and the deviance are calculated only for *events/trials* syntax.

Note that the parameter estimates are not changed by this method. However, their standard errors are adjusted for overdispersion, affecting their significance tests.

Williams' Method

Suppose that the data consist of n binomial observations. For the ith observation, let r_i/n_i be the observed proportion and let $\mathbf{x}_i$ be the associated vector of explanatory variables. Suppose that the response probability for the ith observation is a random variable P_i with mean and variance

$$E(P_i) = \pi_i \quad \text{and} \quad V(P_i) = \phi\pi_i(1-\pi_i)$$

where p_i is the probability of the event, and ϕ is a nonnegative but otherwise unknown scale parameter. Then the mean and variance of r_i are

$$E(r_i) = n_i\pi_i \quad \text{and} \quad V(r_i) = n_i\pi_i(1-\pi_i)[1+(n_i-1)\phi]$$

Williams (1982) estimates the unknown parameter ϕ by equating the value of Pearson's chi-square statistic for the full model to its approximate expected value. Suppose w_i^* is the weight associated with the ith observation. The Pearson chi-square statistic is given by

$$\chi^2 = \sum_{i=1}^{n} \frac{w_i^*(r_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}$$

Let $g'(\cdot)$ be the first derivative of the link function $g(\cdot)$. The approximate expected value of χ^2 is

$$E_{\chi^2} = \sum_{i=1}^{n} w_i^*(1 - w_i^* v_i d_i)[1+\phi(n_i-1)]$$

where $v_i = n_i/(\pi_i(1-\pi_i)[g'(\pi_i)]^2)$ and d_i is the variance of the linear predictor $\hat{\alpha}_i + \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$. The scale parameter ϕ is estimated by the following iterative procedure.

At the start, let $w_i^* = 1$ and let π_i be approximated by r_i/n_i, $i = 1, 2, \ldots, n$. If you apply these weights and approximated probabilities to χ^2 and E_{χ^2} and then equate them, an initial estimate of ϕ is therefore

$$\hat{\phi}_0 = \frac{\chi^2 - (n-p)}{\sum_i (n_i - 1)(1 - v_i d_i)}$$

where p is the total number of parameters. The initial estimates of the weights become $\hat{w}_{i0}^* = [1 + (n_i - 1)\hat{\phi}_0]^{-1}$. After a weighted fit of the model, $\widehat{\boldsymbol{\beta}}$ is recalculated, and so is χ^2. Then a revised estimate of ϕ is given by

$$\hat{\phi}_1 = \frac{\chi^2 - \sum_i w_i^*(1 - w_i^* v_i d_i)}{w_i^*(n_i - 1)(1 - w_i^* v_i d_i)}$$

The iterative procedure is repeated until χ^2 is very close to its degrees of freedom.

Once ϕ has been estimated by $\hat{\phi}$ under the full model, weights of $(1 + (n_i - 1)\hat{\phi})^{-1}$ can be used in fitting models that have fewer terms than the full model. See Example 42.9 on page 2438 for an illustration.

The Hosmer-Lemeshow Goodness-of-Fit Test

Sufficient replication within subpopulations is required to make the Pearson and deviance goodness-of-fit tests valid. When there are one or more continuous predictors in the model, the data are often too sparse to use these statistics. Hosmer and Lemeshow (2000) proposed a statistic that they show, through simulation, is distributed as chi-square when there is no replication in any of the subpopulations. This test is only available for binary response models.

First, the observations are sorted in increasing order of their estimated event probability. The event is the response level specified in the response variable option EVENT=, or the response level which is not specified in the REF= option, or, if neither of these options were specified, then the event is the response level identified in the "Response Profiles" table as "Ordered Value 1". The observations are then divided into approximately ten groups according to the following scheme. Let N be the total number of subjects. Let M be the target number of subjects for each group given by

$$M = [0.1 \times N + 0.5]$$

where $[x]$ represents the integral value of x. If the *single-trial* syntax is used, blocks of subjects are formed of observations with identical values of the explanatory variables. Blocks of subjects are not divided when being placed into groups.

Suppose there are n_1 subjects in the first block and n_2 subjects in the second block. The first block of subjects is placed in the first group. Subjects in the second block are added to the first group if

$$n_1 < M \quad \text{and} \quad n_1 + [0.5 \times n_2] \leq M$$

Otherwise, they are placed in the second group. In general, suppose subjects of the (j-1)th block have been placed in the kth group. Let c be the total number of subjects currently in the kth group. Subjects for the jth block (containing n_j subjects) are also placed in the kth group if

$$c < M \quad \text{and} \quad c + [0.5 \times n_j] \leq M$$

Otherwise, the n_j subjects are put into the next group. In addition, if the number of subjects in the last group does not exceed $[0.05 \times N]$ (half the target group size), the last two groups are collapsed to form only one group.

Note that the number of groups, g, may be smaller than 10 if there are fewer than 10 patterns of explanatory variables. There must be at least three groups in order for the Hosmer-Lemeshow statistic to be computed.

The Hosmer-Lemeshow goodness-of-fit statistic is obtained by calculating the Pearson chi-square statistic from the $2 \times g$ table of observed and expected frequencies, where g is the number of groups. The statistic is written

$$\chi^2_{HL} = \sum_{i=1}^{g} \frac{(O_i - N_i\bar{\pi}_i)^2}{N_i\bar{\pi}_i(1-\bar{\pi}_i)}$$

where N_i is the total frequency of subjects in the ith group, O_i is the total frequency of event outcomes in the ith group, and $\bar{\pi}_i$ is the average estimated predicted probability of an event outcome for the ith group. The Hosmer-Lemeshow statistic is then compared to a chi-square distribution with $(g - n)$ degrees of freedom, where the value of n can be specified in the LACKFIT option in the MODEL statement. The default is $n = 2$. Large values of χ^2_{HL} (and small p-values) indicate a lack of fit of the model.

Receiver Operating Characteristic Curves

In a sample of n individuals, suppose n_1 individuals are observed to have a certain condition or event. Let this group be denoted by $\mathcal{C}_1$, and let the group of the remaining $n_2 = n - n_1$ individuals who do not have the condition be denoted by $\mathcal{C}_2$. Risk factors are identified for the sample, and a logistic regression model is fitted to the data. For the jth individual, an estimated probability $\hat{\pi}_j$ of the event of interest is calculated. Note that the $\hat{\pi}_j$ are computed as shown in the "Linear Predictor, Predicted Probability, and Confidence Limits" section on page 2350 and are not the cross validated estimates discussed in the "Classification Table" section on page 2352.

Suppose the n individuals undergo a test for predicting the event and the test is based on the estimated probability of the event. Higher values of this estimated probability are assumed to be associated with the event. A receiver operating characteristic (ROC) curve can be constructed by varying the cutpoint that determines which estimated event probabilities are considered to predict the event. For each cutpoint z, the following measures can be output to a data set using the OUTROC= option in the

MODEL statement or the OUTROC= option in the SCORE statement:

$$
\begin{aligned}
\_\mathrm{POS}\_(z) &= \sum_{i \in \mathcal{C}_1} I(\hat{\pi}_i \geq z) \\
\_\mathrm{NEG}\_(z) &= \sum_{i \in \mathcal{C}_2} I(\hat{\pi}_i < z) \\
\_\mathrm{FALPOS}\_(z) &= \sum_{i \in \mathcal{C}_2} I(\hat{\pi}_i \geq z) \\
\_\mathrm{FALNEG}\_(z) &= \sum_{i \in \mathcal{C}_1} I(\hat{\pi}_i < z) \\
\_\mathrm{SENSIT}\_(z) &= \frac{\_\mathrm{POS}\_(z)}{n_1} \\
\_\mathrm{1MSPEC}\_(z) &= \frac{\_\mathrm{FALPOS}\_(z)}{n_2}
\end{aligned}
$$

where $I(\cdot)$ is the indicator function.

Note that _POS_(z) is the number of correctly predicted event responses, _NEG_(z) is the number of correctly predicted nonevent responses, _FALPOS_(z) is the number of falsely predicted event responses, _FALNEG_(z) is the number of falsely predicted nonevent responses, _SENSIT_(z) is the sensitivity of the test, and _1MSPEC_(z) is one minus the specificity of the test.

A plot of the ROC curve can be constructed by using the PLOT or GPLOT procedure with the OUTROC= data set and plotting sensitivity (_SENSIT_) against 1-specificity (_1MSPEC_); see Example 42.7 on page 2429 for an illustration. The area under the ROC curve, as determined by the trapezoidal rule, is estimated by the statistic *c* in the "Association of Predicted Probabilities and Observed Responses" table.

Testing Linear Hypotheses about the Regression Coefficients

Linear hypotheses for $\boldsymbol{\theta}$ are expressed in matrix form as

$$H_0\colon \mathbf{L}\boldsymbol{\theta} = \mathbf{c}$$

where $\mathbf{L}$ is a matrix of coefficients for the linear hypotheses, and $\mathbf{c}$ is a vector of constants. The vector of regression coefficients $\boldsymbol{\theta}$ includes slope parameters as well as intercept parameters. The Wald chi-square statistic for testing H_0 is computed as

$$\chi^2_W = (\mathbf{L}\widehat{\boldsymbol{\theta}} - \mathbf{c})'[\mathbf{L}\widehat{\mathbf{V}}(\widehat{\boldsymbol{\theta}})\mathbf{L}']^{-1}(\mathbf{L}\widehat{\boldsymbol{\theta}} - \mathbf{c})$$

where $\widehat{\mathbf{V}}(\widehat{\boldsymbol{\theta}})$ is the estimated covariance matrix. Under H_0, χ^2_W has an asymptotic chi-square distribution with r degrees of freedom, where r is the rank of $\mathbf{L}$.

Regression Diagnostics

For binary response data, regression diagnostics developed by Pregibon (1981) can be requested by specifying the INFLUENCE option. For diagnostics available with conditional logistic regression, see the "Regression Diagnostic Details" section on page 2367.

This section uses the following notation:

r_j, n_j	r_j is the number of event responses out of n_j trials for the jth observation. If *events/trials* syntax is used, r_j is the value of *events* and n_j is the value of *trials*. For *single-trial* syntax, $n_j = 1$, and $r_j = 1$ if the ordered response is 1, and $r_j = 0$ if the ordered response is 2.
w_j	is the weight of the jth observation.
π_j	is the probability of an event response for the jth observation given by $\pi_j = F(\alpha + \boldsymbol{\beta}'\mathbf{x}_j)$, where $F(\cdot)$ is the inverse link function defined on page 2335.
$\mathbf{b}$	is the maximum likelihood estimate (MLE) of $(\alpha, \boldsymbol{\beta}')'$.
$\hat{\mathbf{V}}_{\mathbf{b}}$	is the estimated covariance matrix of $\mathbf{b}$.
$\hat{p}_j, \hat{q}_j$	$\hat{p}_j$ is the estimate of π_j evaluated at $\mathbf{b}$, and $\hat{q}_j = 1 - \hat{p}_j$.

Pregibon suggests using the index plots of several diagnostic statistics to identify influential observations and to quantify the effects on various aspects of the maximum likelihood fit. In an index plot, the diagnostic statistic is plotted against the observation number. In general, the distributions of these diagnostic statistics are not known, so cutoff values cannot be given for determining when the values are large. However, the IPLOTS and INFLUENCE options provide displays of the diagnostic values allowing visual inspection and comparison of the values across observations. In these plots, if the model is correctly specified and fits all observations well, then no extreme points should appear.

The next five sections give formulas for these diagnostic statistics.

Hat Matrix Diagonal

The diagonal elements of the hat matrix are useful in detecting extreme points in the design space where they tend to have larger values. The jth diagonal element is

$$h_{jj} = \begin{cases} \widetilde{w}_j(1, \mathbf{x}_j')\hat{\mathbf{V}}_{\mathbf{b}}(1, \mathbf{x}_j')' & \text{Fisher-Scoring} \\ \widehat{w}_j(1, \mathbf{x}_j')\hat{\mathbf{V}}_{\mathbf{b}}(1, \mathbf{x}_j')' & \text{Newton-Raphson} \end{cases}$$

where

$$\begin{aligned} \widetilde{w}_j &= \frac{w_j n_j}{\hat{p}_j\hat{q}_j[g'(\hat{p}_j)]^2} \\ \widehat{w}_j &= \widetilde{w}_j + \frac{w_j(r_j - n_j\hat{p}_j)[\hat{p}_j\hat{q}_j g''(\hat{p}_j) + (\hat{q}_j - \hat{p}_j)g'(\hat{p}_j)]}{(\hat{p}_j\hat{q}_j)^2[g'(\hat{p}_j)]^3} \end{aligned}$$

and $g'(\cdot)$ and $g''(\cdot)$ are the first and second derivatives of the link function $g(\cdot)$, respectively.

For a binary response logit model, the hat matrix diagonal elements are

$$h_{jj} = w_j n_j \hat{p}_j \hat{q}_j (1, \mathbf{x}_j') \hat{\mathbf{V}}_{\mathbf{b}} \begin{pmatrix} 1 \\ \mathbf{x}_j \end{pmatrix}$$

If the estimated probability is extreme (less than 0.1 and greater than 0.9, approximately), then the hat diagonal may be greatly reduced in value. Consequently, when an observation has a very large or very small estimated probability, its hat diagonal value is not a good indicator of the observation's distance from the design space (Hosmer and Lemeshow 2000, p 171).

Pearson Residuals and Deviance Residuals

Pearson and Deviance residuals are useful in identifying observations that are not explained well by the model. Pearson residuals are components of the Pearson chi-square statistic and deviance residuals are components of the deviance. The Pearson residual for the jth observation is

$$\chi_j = \frac{\sqrt{w_j}(r_j - n_j \hat{p}_j)}{\sqrt{n_j \hat{p}_j \hat{q}_j}}$$

The Pearson chi-square statistic is the sum of squares of the Pearson residuals. The deviance residual for the jth observation is

$$d_j = \begin{cases} -\sqrt{-2w_j n_j \log(\hat{q}_j)} & \text{if } r_j = 0 \\ \pm\sqrt{2w_j[r_j \log(\frac{r_j}{n_j \hat{p}_j}) + (n_j - r_j)\log(\frac{n_j - r_j}{n_j \hat{q}_j})]} & \text{if } 0 < r_j < n_j \\ \sqrt{-2w_j n_j \log(\hat{p}_j)} & \text{if } r_j = n_j \end{cases}$$

where the plus (minus) in $\pm$ is used if r_j/n_j is greater (less) than $\hat{p}_j$. The deviance is the sum of squares of the deviance residuals.

DFBETAS

For each parameter estimate, the procedure calculates a DFBETAS diagnostic for each observation. The DFBETAS diagnostic for an observation is the standardized difference in the parameter estimate due to deleting the observation, and it can be used to assess the effect of an individual observation on each estimated parameter of the fitted model. Instead of re-estimating the parameter every time an observation is deleted, PROC LOGISTIC uses the one-step estimate. See the section "Predicted Probability of an Event for Classification" on page 2352. For the jth observation, the DFBETAS are given by

$$\text{DFBETAS}i_j = \mathbf{\Delta}_i \mathbf{b}_j^1 / \hat{\sigma}(b_i)$$

where $i = 0, 1, \ldots, s, \hat{\sigma}(b_i)$ is the standard error of the ith component of $\mathbf{b}$, and $\mathbf{\Delta}_i \mathbf{b}_j^1$ is the ith component of the one-step difference

$$\mathbf{\Delta b}_j^1 = \frac{w_j(r_j - n_j\hat{p}_j)}{1 - h_{jj}} \hat{\mathbf{V}}_{\mathbf{b}} \begin{pmatrix} 1 \\ \mathbf{x}_j \end{pmatrix}$$

$\mathbf{\Delta b}_j^1$ is the approximate change ($\mathbf{b} - \mathbf{b}_j^1$) in the vector of parameter estimates due to the omission of the jth observation. The DFBETAS are useful in detecting observations that are causing instability in the selected coefficients.

C and CBAR

C and CBAR are confidence interval displacement diagnostics that provide scalar measures of the influence of individual observations on $\mathbf{b}$. These diagnostics are based on the same idea as the Cook distance in linear regression theory, and by using the one-step estimate, C and CBAR for the jth observation are computed as

$$C_j = \chi_j^2 h_{jj}/(1 - h_{jj})^2$$

and

$$\overline{C}_j = \chi_j^2 h_{jj}/(1 - h_{jj})$$

respectively.

Typically, to use these statistics, you plot them against an index (as the IPLOT option does) and look for outliers.

DIFDEV and DIFCHISQ

DIFDEV and DIFCHISQ are diagnostics for detecting ill-fitted observations; in other words, observations that contribute heavily to the disagreement between the data and the predicted values of the fitted model. DIFDEV is the change in the deviance due to deleting an individual observation while DIFCHISQ is the change in the Pearson chi-square statistic for the same deletion. By using the one-step estimate, DIFDEV and DIFCHISQ for the jth observation are computed as

$$\text{DIFDEV} = d_j^2 + \overline{C}_j$$

and

$$\text{DIFCHISQ} = \overline{C}_j / h_{jj}$$

Scoring Data Sets

Scoring a data set, which is especially important for predictive modeling, means applying a previously fitted model to a new data set in order to compute the conditional, or *posterior*, probabilities of each response category given the values of the explanatory variables in each observation.

The SCORE statement enables you to score new data sets and output the scored values and, optionally, the corresponding confidence limits into a SAS data set. If the response variable is included in the new data set, then you can request fit statistics for the data, which is especially useful for test or validation data. If the response is binary, you can also create a SAS data set containing the *receiver operating characteristic* (ROC) curve. You can specify multiple SCORE statements in the same invocation of PROC LOGISTIC.

By default, the posterior probabilities are based on implicit prior probabilities that are proportional to the frequencies of the response categories in the *training data* (the data used to fit the model). Explicit prior probabilities should be specified when the sample proportions of the response categories in the training data differ substantially from the operational data to be scored. For example, to detect a rare category, it is common practice to use a training set in which the rare categories are over-represented; without prior probabilities that reflect the true incidence rate, the predicted posterior probabilities for the rare category will be too high. By specifying the correct priors, the posterior probabilities are adjusted appropriately.

The model fit to the DATA= data set in the PROC LOGISTIC statement is the default model used for the scoring. Alternatively, you can save a fit model on one run of PROC LOGISTIC and use it to score new data on a subsequent run. The OUTMODEL= option in the PROC LOGISTIC statement saves the model information in a SAS data set. Specifying this data set in the INMODEL= option of a new PROC LOGISTIC run will score the DATA= data set in the SCORE statement without refitting the model.

The rest of this section provides some computational details about the scoring.

Posterior Probabilities and Confidence Limits

Let F be the inverse link function. That is,

$$F(t) = \begin{cases} \frac{1}{1+\exp(-t)} & \text{logistic} \\ \Phi(t) & \text{normal} \\ 1-\exp(-\exp(t)) & \text{complementary log-log} \end{cases}$$

The first derivative of F is given by

$$F'(t) = \begin{cases} \frac{\exp(-t)}{(1+\exp(-t))^2} & \text{logistic} \\ \phi(t) & \text{normal} \\ \exp(t)\exp(-\exp(t)) & \text{complementary log-log} \end{cases}$$

Suppose there are $k+1$ response categories. Let Y be the response variable with levels $1, \ldots, k+1$. Let $\mathbf{x} = (x_0, x_1, \ldots, x_p)'$ be a $(p+1)$-vector of covariates, with $x_0 \equiv 1$. Let $\boldsymbol{\theta}$ be the vector of regression parameters.

Posterior probabilities are given by

$$\mathrm{P}_n(i) = \frac{\mathrm{P}_o(i)\frac{\widetilde{\mathrm{p}_n(i)}}{\mathrm{p}_o(i)}}{\sum_j \mathrm{P}_o(j)\frac{\widetilde{\mathrm{p}_n(j)}}{\mathrm{p}_o(j)}} \qquad i = 1, \ldots, k+1$$

where the old posterior probabilities (P_o) are the conditional probabilities of the response categories given $\mathbf{x}$, and the old priors (p_o) are the sample proportions of response categories of the training data. To simplify notation, absorb the old priors into the new priors; that is

$$\mathrm{p}_n(i) = \frac{\widetilde{\mathrm{p}_n(i)}}{\mathrm{p}_o(i)} \qquad i = 1, \ldots, k+1$$

The posterior probabilities are functions of $\boldsymbol{\theta}$ and their estimates are obtained by substituting $\boldsymbol{\theta}$ by its MLE $\widehat{\boldsymbol{\theta}}$. The variances of the estimated posterior probabilities are given by the *delta method* as follows:

$$Var(\widehat{\mathrm{P}_n(i)}) = \left[\frac{\partial \mathrm{P}_n(i)}{\partial \boldsymbol{\theta}}\right]' Var(\widehat{\boldsymbol{\theta}}) \left[\frac{\partial \mathrm{P}_n(i)}{\partial \boldsymbol{\theta}}\right]$$

where

$$\frac{\partial \mathrm{P}_n(i)}{\partial \boldsymbol{\theta}} = \frac{\frac{\partial \mathrm{P}_o(i)}{\partial \boldsymbol{\theta}}\mathrm{p}_n(i)}{\sum_j \mathrm{P}_o(j)\mathrm{p}_n(j)} - \frac{\mathrm{P}_o(i)\mathrm{p}_n(i)\sum_j \frac{\partial \mathrm{P}_o(j)}{\partial \boldsymbol{\theta}}\mathrm{p}_n(j)}{[\sum_j \mathrm{P}_o(j)\mathrm{p}_n(j)]^2}$$

A 100(1-α) percent confidence interval for $\mathrm{P}_n(i)$ is

$$\widehat{\mathrm{P}_n(i)} \pm z_{1-\alpha/2}\sqrt{\widehat{Var}(\widehat{\mathrm{P}_n(i)})}$$

where z_τ is the upper 100τ percentile of the standard normal distribution.

Cumulative Response Model

Let $\alpha_1, \ldots, \alpha_k$ be the intercept parameters and let $\boldsymbol{\beta}$ be the vector of slope parameters. Denote $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_k, \boldsymbol{\beta}')'$. Let

$$\eta_i = \eta_i(\boldsymbol{\theta}) = \alpha_i + \mathbf{x}'\boldsymbol{\beta}, \qquad i = 1, \ldots, k$$

Estimates of $\eta_1, \ldots, \eta_k$ are obtained by substituting the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$.

The predicted probabilities of the responses are

$$\widehat{\mathrm{P}_o}(i) = \widehat{\Pr}(Y = i) = \begin{cases} F(\hat{\eta}_1) & i = 1 \\ F(\hat{\eta}_i) - F(\hat{\eta}_{i-1}) & i = 2, \ldots, k \\ 1 - F(\hat{\eta}_k) & i = k+1 \end{cases}$$

For $i = 1, \ldots, k$, let $\delta_i(x)$ be a $(k+1)$ column vector with ith entry equal to 1, $k+1$th entry equal to x, and all other entries 0. The derivative of $\mathrm{P}_o(i)$ with respect to $\boldsymbol{\theta}$ are

$$\frac{\partial \mathrm{P}_o(i)}{\partial \boldsymbol{\theta}} = \begin{cases} F'(\alpha_1 + x'\beta)\delta_1(x) & i = 1 \\ F'(\alpha_i + x'\beta)\delta_i(x) - F'(\alpha_{i-1} + x'\beta)\delta_{i-1}(x) & i = 2, \ldots, k \\ -F'(\alpha_k + x'\beta)\delta_k(x) & i = k+1 \end{cases}$$

Generalized Logit Model

Consider the last response level (Y=k+1) as the reference. Let $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k$ be the parameter vectors for the first k logits, respectively. Denote $\boldsymbol{\theta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_k')'$. Let $\eta = (\eta_1, \ldots, \eta_k)'$ with

$$\eta_i = \eta_i(\boldsymbol{\theta}) = \mathbf{x}'\boldsymbol{\beta}_i \quad i = 1, \ldots, k$$

Estimates of $\eta_1, \ldots, \eta_k$ are obtained by substituting the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$.

The predicted probabilities are

$$\begin{aligned} \widehat{\mathrm{P}_o}(k+1) \equiv \Pr(Y = k+1|\mathbf{x}) &= \frac{1}{1 + \sum_{l=1}^{k} \exp(\hat{\eta}_l)} \\ \widehat{\mathrm{P}_o}(i) \equiv \Pr(Y = i|\mathbf{x}) &= \widehat{\mathrm{P}_o}(k+1)\exp(\eta_i), \quad i = 1, \ldots, k \end{aligned}$$

The derivative of $\mathrm{P}_o(i)$ with respect to $\boldsymbol{\theta}$ are

$$\begin{aligned} \frac{\partial \mathrm{P}_o(i)}{\partial \boldsymbol{\theta}} &= \frac{\partial \eta}{\partial \boldsymbol{\theta}} \frac{\partial \mathrm{P}_o(i)}{\partial \eta} \\ &= (I_k \otimes \mathbf{x}) \left(\frac{\partial \mathrm{P}_o(i)}{\partial \eta_1}, \cdots, \frac{\partial \mathrm{P}_o(i)}{\partial \eta_k} \right)' \end{aligned}$$

where

$$\frac{\partial \mathrm{P}_o(i)}{\partial \eta_j} = \begin{cases} \mathrm{P}_o(i)(1 - \mathrm{P}_o(i)) & j = i \\ -\mathrm{P}_o(i)\mathrm{P}_o(j) & \text{otherwise} \end{cases}$$

Special Case of Binary Response Model with No Priors

Let $\boldsymbol{\beta}$ be the vector of regression parameters. Let

$$\eta = \eta(\boldsymbol{\beta}) = \mathbf{x}'\boldsymbol{\beta}$$

The variance of $\hat{\eta}$ is given by

$$Var(\hat{\eta}) = \mathbf{x}'Var(\hat{\boldsymbol{\beta}})\mathbf{x}$$

A 100(1-α) percent confidence interval for η is

$$\hat{\eta} \pm z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{\eta})}$$

Estimates of $\mathrm{P}_o(1)$ and confidence intervals for the $\mathrm{P}_o(1)$ are obtained by back-transforming $\hat{\eta}$ and the confidence intervals for η, respectively. That is,

$$\widehat{\mathrm{P}}_o(1) = F(\hat{\eta})$$

and the confidence intervals are

$$F\left(\hat{\eta} \pm z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{\eta})}\right)$$

Conditional Logistic Regression

The method of maximum likelihood described in the preceding sections relies on large-sample asymptotic normality for the validity of estimates and especially of their standard errors. When you do not have a large sample size compared to the number of parameters, this approach may be inappropriate resulting in biased inferences. This situation typically arises when your data are stratified and you fit intercepts to each stratum so that the number of parameters is of the same order as the sample size. For example, in a 1:1 matched pairs study with n pairs and p covariates, you would estimate $n-1$ intercept parameters and p slope parameters. Taking the stratification into account by "conditioning out" (and not estimating) the stratum-specific intercepts gives consistent and asymptotically normal MLEs for the slope coefficients. See Breslow and Day (1980) and Stokes, Davis, and Koch (2000) for more information. If your nuisance parameters are not just stratum-specific intercepts, you can perform an exact conditional logistic regression.

Computational Details

For each stratum h, $h = 1, \ldots, H$, number the observations as $i = 1, \ldots, n_h$ so that hi indexes the ith observation in the hth stratum. Denote the p covariates for observation hi as $\mathbf{x}_{hi}$ and its binary response as y_{hi}, let $\mathbf{y} = (y_{11}, \ldots, y_{1n_1}, \ldots, y_{H1}, \ldots, y_{Hn_H})'$, $\mathbf{X}_h = (\mathbf{x}_{h1} \ldots \mathbf{x}_{hn_h})'$, and $\mathbf{X} = (\mathbf{X}_1' \ldots \mathbf{X}_H')'$. Let the dummy variables $z_h, h = 1, \ldots, H$, be indicator

functions for the strata ($z_h = 1$ if the observation is in stratum h), denote $\mathbf{z}_{hi} = (z_1, \ldots, z_H)$ for observation hi, $\mathbf{Z}_h = (\mathbf{z}_{h1} \ldots \mathbf{z}_{hn_h})'$, and $\mathbf{Z} = (\mathbf{Z}_1' \ldots \mathbf{Z}_H')'$. Denote $\mathbf{X}^* = (\mathbf{Z}||\mathbf{X})$ and $\mathbf{x}^*_{hi} = (\mathbf{z}'_{hi}||\mathbf{x}'_{hi})'$. Arrange the observations in each stratum h so that $y_{hi} = 1$ for $i = 1, \ldots, m_h$, and $y_{hi} = 0$ for $i = m_{h+1}, \ldots, n_h$. Suppose all observations have unit frequency.

Consider the binary logistic regression model on page 2405 written as

$$\text{logit}(\boldsymbol{\pi}) = \mathbf{X}^* \boldsymbol{\theta}$$

where the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$ consists of $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_H)'$, α_h is the intercept for stratum $h, h = 1, \ldots, H$, and $\boldsymbol{\beta}$ is the parameter vector for the p covariates.

From the "Determining Observations for Likelihood Contributions" section on page 2336, you can write the likelihood contribution of observation $hi, i = 1, \ldots, n_h, h = 1, \ldots, H$, as

$$L_{hi}(\boldsymbol{\theta}) = \frac{e^{y_{hi}\mathbf{x}^{*}_{hi}{}'\boldsymbol{\theta}}}{1 + e^{\mathbf{x}^{*}_{hi}{}'\boldsymbol{\theta}}}$$

where $y_{hi} = 1$ when the response takes Ordered Value 1, and $y_{hi} = 0$ otherwise.

The full likelihood is

$$L(\boldsymbol{\theta}) = \prod_{h=1}^{H}\prod_{i=1}^{n_h} L_{hi}(\boldsymbol{\theta}) = \frac{e^{\mathbf{y}'\mathbf{X}^*\boldsymbol{\theta}}}{\prod_{h=1}^{H}\prod_{i=1}^{n_h}\left(1 + e^{\mathbf{x}^{*}_{hi}{}'\boldsymbol{\theta}}\right)}$$

Unconditional likelihood inference is based on maximizing this likelihood function.

When your nuisance parameters are the stratum-specific intercepts $(\alpha_1, \ldots, \alpha_H)'$, and $\boldsymbol{\beta}$ are your parameters of interest, "conditioning out" the nuisance parameters produces the following conditional likelihood (Lachin 2000)

$$L(\boldsymbol{\beta}) = \prod_{h=1}^{H} L_h(\boldsymbol{\beta}) = \prod_{h=1}^{H} \frac{\prod_{i=1}^{m_h}\exp(\mathbf{x}'_{hi}\boldsymbol{\beta})}{\sum\prod_{j=j_1}^{j_{m_h}}\exp(\mathbf{x}'_{hj}\boldsymbol{\beta})}$$

where the summation is over all $\binom{n_h}{m_h}$ subsets $\{j_1, \ldots, j_{m_h}\}$ of m_h observations chosen from the n_h observations in stratum h. Note that the nuisance parameters have been factored out of this equation.

For conditional asymptotic inference, maximum likelihood estimates $\widehat{\boldsymbol{\beta}}$ of the regression parameters are obtained by maximizing the conditional likelihood, and asymptotic results are applied to the conditional likelihood function and the maximum likelihood estimators. A relatively fast method for computing this conditional likelihood and its derivatives is given by Gail, Lubin, and Rubinstein (1981) and Howard (1972). The default optimization techniques, which are the same as those implemented by the NLP procedure in SAS/OR software, are

- Newton-Raphson with ridging when the number of parameters $p < 40$
- quasi-Newton when $40 \leq p < 400$
- conjugate gradient when $p \geq 400$

Sometimes the log likelihood converges but the estimates diverge. This condition is flagged by having inordinately large standard errors for some of your parameter estimates, and can be monitored by specifying the ITPRINT option. Unfortunately, broad existence criteria such as those discussed in the "Existence of Maximum Likelihood Estimates" section on page 2338 do not exist for this model. It may be possible to circumvent such a problem by standardizing your independent variables before fitting the model.

Regression Diagnostic Details

Diagnostics are used to indicate observations that may have undue influence on the model fit, or which may be outliers. Further investigation should be performed before removing such an observation from the data set.

The derivations in this section follow Storer and Crowley's (1985) method of augmenting the logistic regression model, which provides an estimate of the "one-step" DFBETAS estimates advocated by Pregibon (1984). The method also provides estimates of conditional stratum-specific predicted values, residuals, and leverage for each observation.

Following Storer and Crowley (1985), the log-likelihood contribution can be written as

$$\begin{aligned} l_h &= \log(L_h) = \mathbf{y}_h'\boldsymbol{\gamma}_h - a(\boldsymbol{\gamma}_h) \quad \text{where} \\ a(\boldsymbol{\gamma}_h) &= \log\left[\sum \prod_{j=j_1}^{j_{m_h}} \exp(\gamma_{hj})\right] \end{aligned}$$

and the h subscript on matrices indicates the submatrix for the stratum, $\boldsymbol{\gamma}_h' = (\gamma_{h1}, \ldots, \gamma_{hn_h})$, and $\gamma_{hi} = \mathbf{x}_{hi}'\boldsymbol{\beta}$. Then the gradient and information matrix are

$$\begin{aligned} \mathbf{g}(\boldsymbol{\beta}) &= \left\{\frac{\partial l_h}{\partial \boldsymbol{\beta}}\right\}_{h=1}^{H} = \mathbf{X}'(\mathbf{y} - \boldsymbol{\pi}) \\ \boldsymbol{\Lambda}(\boldsymbol{\beta}) &= \left\{\frac{\partial^2 l_h}{\partial \boldsymbol{\beta}^2}\right\}_{h=1}^{H} = \mathbf{X}'\text{diag}(\mathbf{U}_1, \ldots, \mathbf{U}_H)\mathbf{X} \end{aligned}$$

where

$$\begin{aligned} \pi_{hi} &= \frac{\partial a(\boldsymbol{\gamma}_h)}{\partial \gamma_{hi}} = \frac{\sum_{j(i)} \prod_{j=j_1}^{j_{m_h}} \exp(\gamma_{hj})}{\sum \prod_{j=j_1}^{j_{m_h}} \exp(\gamma_{hj})} \\ \boldsymbol{\pi}_h &= (\pi_{h1}, \ldots, \pi_{hn_h}) \end{aligned}$$

$$\begin{aligned}
\mathbf{U}_h &= \frac{\partial^2 a(\boldsymbol{\gamma}_h)}{\partial \boldsymbol{\gamma}_h^2} = \left\{ \frac{\partial^2 \mathbf{a}(\boldsymbol{\gamma}_h)}{\partial \gamma_{hi} \partial \gamma_{hj}} \right\} = \{a_{ij}\} \\
a_{ij} &= \frac{\sum_{k(i,j)} \prod_{k=k_1}^{k_{m_h}} \exp(\gamma_{hk})}{\sum \prod_{k=k_1}^{k_{m_h}} \exp(\gamma_{hk})} - \frac{\partial a(\boldsymbol{\gamma}_h)}{\partial \gamma_{hi}} \frac{\partial a(\boldsymbol{\gamma}_h)}{\partial \gamma_{hj}} = \pi_{hij} - \pi_{hi}\pi_{hj}
\end{aligned}$$

where π_{hi} is the conditional stratum-specific probability that subject i in stratum h is a case, the summation on $j(i)$ is over all subsets from $\{1, \ldots, n_h\}$ of size m_h that contain the index i, and the summation on $k(i,j)$ is over all subsets from $\{1, \ldots, n_h\}$ of size m_h that contain the indices i and j.

To produce the true one-step estimate $\boldsymbol{\beta}_{hi}^1$, start at the MLE $\widehat{\boldsymbol{\beta}}$, delete the hith observation, and take one-step of the Newton-Raphson algorithm using the reduced data set. Note that if there is only one event or one nonevent in a stratum, deletion of that single observation is equivalent to deletion of the entire stratum. The augmentation method does not take this into account.

The augmented model is

$$\text{logit}(\Pr(y_{hi} = 1 | \mathbf{x}_{hi})) = \mathbf{x}_{hi}'\boldsymbol{\beta} + \mathbf{z}_{hi}'\gamma$$

where $\mathbf{z}_{hi} = (0, \ldots, 0, 1, 0, \ldots, 0)'$ has a 1 in the hith coordinate, and use $\boldsymbol{\beta}^0 = (\widehat{\boldsymbol{\beta}}, 0)'$ as the initial estimate for $(\boldsymbol{\beta}, \gamma)'$. The gradient and information matrix before the step are

$$\begin{aligned}
\mathbf{g}(\boldsymbol{\beta}^0) &= \begin{bmatrix} \mathbf{X}' \\ \mathbf{z}_{hi}' \end{bmatrix} (\mathbf{y} - \boldsymbol{\pi}) = \begin{bmatrix} \mathbf{0} \\ y_{hi} - \pi_{hi} \end{bmatrix} \\
\boldsymbol{\Lambda}(\boldsymbol{\beta}^0) &= \begin{bmatrix} \mathbf{X}' \\ \mathbf{z}_{hi}' \end{bmatrix} \mathbf{U} \begin{bmatrix} \mathbf{X} & \mathbf{z}_{hi} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Lambda}(\boldsymbol{\beta}) & \mathbf{X}'\mathbf{U}\mathbf{z}_{hi} \\ \mathbf{z}_{hi}'\mathbf{U}\mathbf{X} & \mathbf{z}_{hi}'\mathbf{U}\mathbf{z}_{hi} \end{bmatrix}
\end{aligned}$$

Inserting the $\boldsymbol{\beta}^0$ and $(\mathbf{X}', \mathbf{z}_{hi}')'$ into the Gail, Lubin, and Rubinstein (1981) algorithm provides the appropriate estimates of $\mathbf{g}(\boldsymbol{\beta}^0)$ and $\boldsymbol{\Lambda}(\boldsymbol{\beta}^0)$. Indicate these estimates with $\widehat{\boldsymbol{\pi}} = \boldsymbol{\pi}(\widehat{\boldsymbol{\beta}})$, $\widehat{\mathbf{U}} = \mathbf{U}(\widehat{\boldsymbol{\beta}})$, $\widehat{\mathbf{g}}$, and $\widehat{\boldsymbol{\Lambda}}$.

DFBETA is computed from the information matrix as

$$\begin{aligned}
\Delta_{hi}\boldsymbol{\beta} &= \boldsymbol{\beta}^0 - \boldsymbol{\beta}_{hi}^1 \\
&= -\widehat{\boldsymbol{\Lambda}}^{-1}(\boldsymbol{\beta}^0)\widehat{\mathbf{g}}(\boldsymbol{\beta}^0) \\
&= -\widehat{\boldsymbol{\Lambda}}^{-1}(\widehat{\boldsymbol{\beta}})(\mathbf{X}'\widehat{\mathbf{U}}\mathbf{z}_{hi})\mathbf{M}^{-1}\mathbf{z}_{hi}'(\mathbf{y} - \widehat{\boldsymbol{\pi}}) \quad \text{where} \\
\mathbf{M} &= (\mathbf{z}_{hi}'\widehat{\mathbf{U}}\mathbf{z}_{hi}) - (\mathbf{z}_{hi}'\widehat{\mathbf{U}}\mathbf{X})\widehat{\boldsymbol{\Lambda}}^{-1}(\widehat{\boldsymbol{\beta}})(\mathbf{X}'\widehat{\mathbf{U}}\mathbf{z}_{hi})
\end{aligned}$$

For each observation in the dataset, a DFBETA statistic is computed for each parameter β_j, $1 \le j \le p$, and standardized by the standard error of β_j from the full data set to produce the estimate of DFBETAS.

The estimated residuals $e_{hi} = y_{hi} - \widehat{\pi}_{hi}$ are obtained from $\widehat{\mathbf{g}}(\boldsymbol{\beta}^0)$, and the weights, or predicted probabilities, are then $\widehat{\pi}_{hi} = y_{hi} - e_{hi}$. The residuals are standardized and reported as (estimated) Pearson residuals:

$$\frac{r_{hi} - n_{hi}\widehat{\pi}_{hi}}{\sqrt{n_{hi}\widehat{\pi}_{hi}(1-\widehat{\pi}_{hi})}}$$

where r_{hi} is the number of events in the observation and n_{hi} is the number of trials.

The estimated leverage is defined as

$$\mathsf{h}_{hi} = \frac{\text{trace}\{(\mathbf{z}'_{hi}\widehat{\mathbf{U}}\mathbf{X})\widehat{\boldsymbol{\Lambda}}^{-1}(\widehat{\boldsymbol{\beta}})(\mathbf{X}'\widehat{\mathbf{U}}\mathbf{z}_{hi})\}}{\text{trace}\{\mathbf{z}'_{hi}\widehat{\mathbf{U}}\mathbf{z}_{hi}\}}$$

This definition of leverage produces different values from those defined by Pregibon (1984), Moolgavkar, Lustbader, and Venzon (1985), and Hosmer and Lemeshow (2000); however, it has the advantage that no extra computations beyond those for the DFBETAS are required.

For events/trials MODEL syntax, treat each observation as two observations (the first for the nonevents and the second for the events) with frequencies $f_{h,2i-1} = n_{hi} - r_{hi}$ and $f_{h,2i} = r_{hi}$, and augment the model with a matrix $\mathbf{Z}_{hi} = [\mathbf{z}_{h,2i-1}\mathbf{z}_{h,2i}]$ instead of a single $\mathbf{z}_{hi}$ vector. Writing $\gamma_{hi} = \mathbf{x}'_{hi}\boldsymbol{\beta} f_{hi}$ in the preceding section results in the following gradient and information matrix.

$$\begin{aligned}
\mathbf{g}(\boldsymbol{\beta}^0) &= \begin{bmatrix} \mathbf{0} \\ f_{h,2i-1}(y_{h,2i-1} - \pi_{h,2i-1}) \\ f_{h,2i}(y_{h,2i} - \pi_{h,2i}) \end{bmatrix} \\
\boldsymbol{\Lambda}(\boldsymbol{\beta}^0) &= \begin{bmatrix} \boldsymbol{\Lambda}(\boldsymbol{\beta}) & \mathbf{X}'\text{diag}(\mathbf{f})\mathbf{U}\text{diag}(\mathbf{f})\mathbf{Z}_{hi} \\ \mathbf{Z}'_{hi}\text{diag}(\mathbf{f})\mathbf{U}\text{diag}(\mathbf{f})\mathbf{X} & \mathbf{Z}'_{hi}\text{diag}(\mathbf{f})\mathbf{U}\text{diag}(\mathbf{f})\mathbf{Z}_{hi} \end{bmatrix}
\end{aligned}$$

The predicted probabilities are then $\widehat{\pi}_{hi} = y_{h,2i} - e_{h,2i}/r_{h,2i}$, while the leverage and the DFBETAs are produced from $\boldsymbol{\Lambda}(\boldsymbol{\beta}^0)$ in a similar fashion as for the preceding single-trial equations.

Exact Conditional Logistic Regression

The theory of exact conditional logistic regression analysis was originally laid out by Cox (1970), and the computational methods employed in PROC LOGISTIC are described in Hirji, Mehta, and Patel (1987), Hirji (1992), and Mehta, Patel, and Senchaudhuri (1992). Other useful references for the derivations include Cox and Snell (1989), Agresti (1990), and Mehta and Patel (1995).

Exact conditional inference is based on generating the conditional distribution for the sufficient statistics of the parameters of interest. This distribution is called the *permutation* or *exact conditional* distribution. Using the notation in the "Computational

Details" section on page 2365, follow Mehta and Patel (1995) and first note that the sufficient statistics $\mathbf{T} = (T_1, \ldots, T_p)$ for $\boldsymbol{\theta}$ are

$$T_j = \sum_{i=1}^{n} y_i x_{ij}, \quad j = 1, \ldots, p$$

Denote a vector of observable sufficient statistics as $\mathbf{t} = (t_1, \ldots, t_p)'$.

The probability density function (pdf) for $\mathbf{T}$ can be created by summing over all binary sequences $\mathbf{y}$ that generate an observable $\mathbf{t}$ and letting $C(\mathbf{t}) = ||\{\mathbf{y} : \mathbf{y}'\mathbf{X} = \mathbf{t}'\}||$ denote the number of sequences $\mathbf{y}$ that generate $\mathbf{t}$

$$\Pr(\mathbf{T} = \mathbf{t}) = \frac{C(\mathbf{t}) \exp(\mathbf{t}'\boldsymbol{\theta})}{\prod_{i=1}^{n}[1 + \exp(\mathbf{x}_i'\boldsymbol{\theta})]}$$

In order to condition out the stratum parameters, partition the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_0', \boldsymbol{\theta}_1')'$, where $\boldsymbol{\theta}_0$ is a $p_0 \times 1$ vector of the nuisance parameters, and $\boldsymbol{\theta}_1$ is the parameter vector for the remaining $p_1 = p - p_0$ parameters of interest. Likewise, partition $\mathbf{X}$ into $\mathbf{X}_0$ and $\mathbf{X}_1$, $\mathbf{T}$ into $\mathbf{T}_0$ and $\mathbf{T}_1$, and $\mathbf{t}$ into $\mathbf{t}_0$ and $\mathbf{t}_1$. The nuisance parameters can be removed from the analysis by conditioning on their sufficient statistics to create the conditional likelihood of $\mathbf{T}_1$ given $\mathbf{T}_0 = \mathbf{t}_0$

$$\begin{aligned}\Pr(\mathbf{T}_1 = \mathbf{t}_1 | \mathbf{T}_0 = \mathbf{t}_0) &= \frac{\Pr(\mathbf{T} = \mathbf{t})}{\Pr(\mathbf{T}_0 = \mathbf{t}_0)} \\ &= f_{\boldsymbol{\theta}_1}(\mathbf{t}_1 | \mathbf{t}_0) = \frac{C(\mathbf{t}_0, \mathbf{t}_1) \exp(\mathbf{t}_1'\boldsymbol{\theta}_1)}{\sum_u C(\mathbf{t}_0, \mathbf{u}) \exp(\mathbf{u}'\boldsymbol{\theta}_1)}\end{aligned}$$

where $C(\mathbf{t}_0, \mathbf{u})$ is the number of vectors $\mathbf{y}$ such that $\mathbf{y}'\mathbf{X}_0 = \mathbf{t}_0$ and $\mathbf{y}'\mathbf{X}_1 = \mathbf{u}$. Note that the nuisance parameters have factored out of this equation, and that $C(\mathbf{t}_0, \mathbf{t}_1)$ is a constant.

The goal of the exact conditional analysis is to determine how likely the observed response $\mathbf{y}_0$ is with respect to all 2^n possible responses $\mathbf{y} = (y_1, \ldots, y_n)'$. One way to proceed is to generate every $\mathbf{y}$ vector for which $\mathbf{y}'\mathbf{X}_0 = \mathbf{t}_0$, and count the number of vectors $\mathbf{y}$ for which $\mathbf{y}'\mathbf{X}_1$ is equal to each unique $\mathbf{t}_1$. Generating the conditional distribution from complete enumeration of the joint distribution is conceptually simple; however, this method becomes computationally infeasible very quickly. For example, if you had only 30 observations, you'd have to scan through 2^{30} different $\mathbf{y}$ vectors.

Several algorithms are available in PROC LOGISTIC to generate the exact distribution. All of the algorithms are based on the following observation. Given any $\mathbf{y} = (y_1, \ldots, y_n)'$ and a design $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$, let $\mathbf{y}_{(i)} = (y_1, \ldots, y_i)'$ and $\mathbf{X}_{(i)} = (\mathbf{x}_1, \ldots, \mathbf{x}_i)'$ be the first i rows of each matrix. Write the sufficient statistic based on these i rows as $\mathbf{t}_{(i)}' = \mathbf{y}_{(i)}'\mathbf{X}_{(i)}$. A recursion relation results: $\mathbf{t}_{(i+1)} = \mathbf{t}_{(i)} + y_{i+1}\mathbf{x}_{i+1}$.

The following methods are available.

- The *multivariate shift algorithm* developed by Hirji, Mehta, and Patel (1987) steps through the recursion relation by adding one observation at a time and building an intermediate distribution at each step. If it determines that $\mathbf{t}_{(i)}$ for the nuisance parameters could eventually equal $\mathbf{t}$, then $\mathbf{t}_{(i)}$ is added to the intermediate distribution.
- Hirji (1992) extends the multivariate shift algorithm to generalized logit models. Since the generalized logit model fits a new set of parameters to each logit, the number of parameters in the model can easily get too large for this algorithm to handle. Note for these models that the hypothesis tests for each effect are computed across the logit functions, while individual parameters are estimated for each logit function.
- A network algorithm described in Mehta, Patel, and Senchaudhuri (1992) builds a network for each parameter that you are conditioning out in order to identify feasible y_i for the $\mathbf{y}$ vector. These networks are combined and the set of feasible y_i is further reduced, then the multivariate shift algorithm uses this knowledge to build the exact distribution without adding as many intermediate $\mathbf{t}_{(i+1)}$ as the multivariate shift algorithm does.
- Mehta, Patel, and Senchaudhuri (2000) devised a hybrid Monte-Carlo and network algorithm that extends their 1992 algorithm by sampling from the combined network to build the exact distribution.

The bulk of the computation time and memory for these algorithms is consumed by the creation of the networks and the exact joint distribution. After the joint distribution for a set of effects is created, the computational effort required to produce hypothesis tests and parameter estimates for any subset of the effects is (relatively) trivial.

Hypothesis Tests

Consider testing the null hypothesis H_0: $\boldsymbol{\beta}_1 = \mathbf{0}$ against the alternative H_A: $\boldsymbol{\beta}_1 \neq \mathbf{0}$, conditional on $\mathbf{T}_0 = \mathbf{t}_0$. Under the null hypothesis, the test statistic for the *exact probability test* is just $f_{\beta_1=0}(\mathbf{t}_1|\mathbf{t}_0)$, while the corresponding p-value is the probability of getting a less likely (more extreme) statistic,

$$p(\mathbf{t}_1|\mathbf{t}_0) = \sum_{u \in \Omega_p} f_0(\mathbf{u}|\mathbf{t}_0)$$

where $\Omega_p = \{\mathbf{u}$: there exist $\mathbf{y}$ with $\mathbf{y}'\mathbf{X}_1 = \mathbf{u}$, $\mathbf{y}'\mathbf{X}_0 = \mathbf{t}_0$, and $f_0(\mathbf{u}|\mathbf{t}_0) \leq f_0(\mathbf{t}_1|\mathbf{t}_0)\}$.

For the *exact conditional scores test*, the conditional mean $\boldsymbol{\mu}_1$ and variance matrix $\boldsymbol{\Sigma}_1$ of the $\mathbf{T}_1$ (conditional on $\mathbf{T}_0 = \mathbf{t}_0$) are calculated, and the score statistic for the observed value,

$$s = (\mathbf{t}_1 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\mathbf{t}_1 - \boldsymbol{\mu}_1)$$

is compared to the score for each member of the distribution

$$S(\mathbf{T}_1) = (\mathbf{T}_1 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\mathbf{T}_1 - \boldsymbol{\mu}_1)$$

The resulting p-value is

$$p(\mathbf{t}_1|\mathbf{t}_0) = Pr(S \geq s) = \sum_{u \in \Omega_s} f_0(\mathbf{u}|\mathbf{t}_0)$$

where $\Omega_s = \{\mathbf{u}$: there exist $\mathbf{y}$ with $\mathbf{y}'\mathbf{X}_1 = \mathbf{u}$, $\mathbf{y}'\mathbf{X}_0 = \mathbf{t}_0$, and $S(\mathbf{u}) \geq s\}$.

The mid-p statistic, defined as

$$p(\mathbf{t}_1|\mathbf{t}_0) - \frac{1}{2} f_0(\mathbf{t}_1|\mathbf{t}_0)$$

was proposed by Lancaster (1961) to compensate for the discreteness of a distribution. Refer to Agresti (1992) for more information. However, to allow for more flexibility in handling ties, you can write the mid-p statistic as (based on a suggestion by LaMotte 2002 and generalizing Vollset, Hirji, and Afifi 1991)

$$\sum_{u \in \Omega_<} f_0(\mathbf{u}|\mathbf{t}_0) + \delta_1 f_0(\mathbf{t}_1|\mathbf{t}_0) + \delta_2 \sum_{u \in \Omega_=} f_0(\mathbf{u}|\mathbf{t}_0)$$

where, for $i \in \{p, s\}$, $\Omega_<$ is Ω_i using strict inequalities, and $\Omega_=$ is Ω_i using equalities with the added restriction that $\mathbf{u} \neq \mathbf{t}_1$. Letting $(\delta_1, \delta_2) = (0.5, 1.0)$ yields Lancaster's mid-p.

Caution: When the exact distribution has ties and METHOD=NETWORKMC is specified, the Monte Carlo algorithm estimates $p(\mathbf{t}|\mathbf{t}_0)$ with error, and hence it cannot determine precisely which values contribute to the reported p-values. For example, if the exact distribution has densities $\{0.2, 0.2, 0.2, 0.4\}$ and if the observed statistic has probability 0.2, then the exact probability p-value is exactly 0.6. Under Monte Carlo sampling, if the densities after N samples are $\{0.18, 0.21, 0.23, 0.38\}$ and the observed probability is 0.21, then the resulting p-value is 0.39. Therefore, the exact probability test p-value for this example fluctuates between 0.2, 0.4, and 0.6, and the reported p-values are actually lower bounds for the true p-values. If you need more precise values, you can specify the OUTDIST= option, determine appropriate cutoff values for the observed probability and score, then construct the true p-value estimates from the OUTDIST= data set using the following statements.

```
data _null_;
   set outdist end=end;
   retain pvalueProb 0 pvalueScore 0;
   if prob < ProbCutOff then pvalueProb+prob;
   if score > ScoreCutOff then pvalueScore+prob;
   if end then put pvalueProb pvalueScore;
run;
```

Inference for a Single Parameter

Exact parameter estimates are derived for a single parameter β_i by regarding all the other parameters $\boldsymbol{\beta}_0 = (\beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_{p+q})'$ as nuisance parameters. The appropriate sufficient statistics are $\mathbf{T}_1 = T_i$ and $\mathbf{T}_0 = (T_1, \ldots, T_{i-1}, T_{i+1}, \ldots, T_{p+q})'$, with their observed values denoted by the lowercase t. Hence, the conditional pdf used to create the parameter estimate for β_i is

$$f_{\beta_i}(t_i|\mathbf{t}_0) = \frac{C(\mathbf{t}_0, t_i)\exp(t_i\beta_i)}{\sum_{u\in\Omega} C(\mathbf{t}_0, u)\exp(u\beta_i)}$$

for $\Omega = \{u\text{: there exist } \mathbf{y} \text{ with } T_i = u \text{ and } \mathbf{T}_0 = \mathbf{t}_0\}$.

The maximum exact conditional likelihood estimate is the quantity $\widehat{\beta}_i$, which maximizes the conditional pdf. A Newton-Raphson algorithm is used to perform this search. However, if the observed t_i attains either its maximum or minimum value in the exact distribution (that is, either $t_i = \min\{u : u \in \Omega\}$ or $t_i = \max\{u : u \in \Omega\}$), then the conditional pdf is monotonically increasing in β_i and cannot be maximized. In this case, a median unbiased estimate (Hirji, Tsiatis, and Mehta 1989) $\widehat{\beta}_i$ is produced that satisfies $f_{\widehat{\beta}_i}(t_i|\mathbf{t}_0) = 0.5$, and a Newton-Raphson-type algorithm is used to perform the search.

Likelihood ratio tests based on the conditional pdf are used to test the null $H_0\text{: } \beta_i = 0$ against the alternative $H_A\text{: } \beta_i > 0$. The critical region for this UMP test consists of the upper tail of values for T_i in the exact distribution. Thus, the one-sided significance level $p_+(t_i; 0)$ is

$$p_+(t_i; 0) = \sum_{u\geq t_i} f_0(u|\mathbf{t}_0)$$

Similarly, the one-sided significance level $p_-(t_i; 0)$ against $H_A\text{: } \beta_i < 0$ is

$$p_-(t_i; 0) = \sum_{u\leq t_i} f_0(u|\mathbf{t}_0)$$

The two-sided significance level $p(t_i; 0)$ against $H_A\text{: } \beta_i \neq 0$ is calculated as

$$p(t_i; 0) = 2\min[p_-(t_i; 0), p_+(t_i; 0)]$$

An upper $100(1-2\epsilon)$% exact confidence limit for $\widehat{\beta}_i$ corresponding to the observed t_i is the solution $\beta_U(t_i)$ of $\epsilon = p_-(t_i, \beta_U(t_i))$, while the lower exact confidence limit is the solution $\beta_L(t_i)$ of $\epsilon = p_+(t_i, \beta_L(t_i))$. Again, a Newton-Raphson procedure is used to search for the solutions.

Specifying the ONESIDED option displays only one p-value and one confidence interval, because small values of $p_+(t_i; 0)$ and $p_-(t_i; 0)$ support different alternative hypotheses and only one of these p-values can be less than 0.50.

The mid-p confidence limits are the solutions to $\min\{p_-(t_i,\beta(t_i)),p_+(t_i,\beta(t_i))\} - (1-\delta_1)f_{\beta(t_i)}(u|\mathbf{t}_0) = \epsilon$ for $\epsilon = \alpha/2, 1-\alpha/2$ (Vollset, Hirji, and Afifi 1991). $\delta_1 = 1$ produces the usual exact (or *max-p*) confidence interval, $\delta_1 = 0.5$ yields the mid-p interval, and $\delta_1 = 0$ gives the *min-p* interval. The mean of the endpoints of the max-p and min-p intervals provides the *mean-p* interval as defined by Hirji, Mehta, and Patel (1988).

Estimates and confidence intervals for the odds-ratios are produced by exponentiating the estimates and interval endpoints for the parameters.

OUTEST= Output Data Set

The OUTEST= data set contains one observation for each BY group containing the maximum likelihood estimates of the regression coefficients. If you also use the COVOUT option in the PROC LOGISTIC statement, there are additional observations containing the rows of the estimated covariance matrix. If you use the FORWARD, BACKWARD, or STEPWISE selection method, only the estimates of the parameters and covariance matrix for the final model are output to the OUTEST= data set.

Variables in the OUTEST= Data Set

The OUTEST= data set contains the following variables:

- any BY variables specified
- _LINK_, a character variable of length 8 with four possible values: CLOGLOG for the complementary log-log function, LOGIT for the logit function, NORMIT for the probit (alias normit) function, and GLOGIT for the generalized logit function.
- _TYPE_, a character variable of length 8 with two possible values: PARMS for parameter estimates or COV for covariance estimates. If an EXACT statement is also specified, then two other values are possible: EPARMMLE for the exact maximum likelihood estimates and EPARMMUE for the exact median unbiased estimates.
- _NAME_, a character variable containing the name of the response variable when _TYPE_=PARMS, EPARMMLE, and EPARMMUE, or the name of a model parameter when _TYPE_=COV
- _STATUS_, a character variable that indicates whether the estimates have converged
- one variable for each intercept parameter
- one variable for each slope parameter and one variable for the offset variable if the OFFSET= option if specified. If an effect is not included in the final model in a model building process, the corresponding parameter estimates and covariances are set to missing values.
- _LNLIKE_, the log likelihood

Parameter Names in the OUTEST= Data Set

If there are only two response categories in the entire data set, the intercept parameter is named Intercept. If there are more than two response categories in the entire data set, the intercept parameters are named Intercept_xxx, where xxx is the value (formatted if a format is applied) of the corresponding response category.

For continuous explanatory variables, the names of the parameters are the same as the corresponding variables. For class variables, the parameter names are obtained by concatenating the corresponding CLASS variable name with the CLASS category; see the PARAM= option in the CLASS statement and the "CLASS Variable Parameterization" section on page 2331 for more details. For interaction and nested effects, the parameter names are created by concatenating the names of each effect.

For the generalized logit model, names of parameters corresponding to each nonreference category contain _xxx as the suffix, where xxx is the value (formatted if a format is applied) of the corresponding nonreference category. For example, suppose the variable Net3 represents the television network (ABC, CBS, and NBC) viewed at a certain time. The following code fits a generalized logit model with Age and Gender (a CLASS variable with values Female and Male) as explanatory variables.

```
proc logistic;
   class Gender;
   model Net3 = Age Gender / link=glogit;
run;
```

There are two logit functions, one contrasting ABC with NBC and the other contrasting CBS with NBC. For each logit, there are three parameters: an intercept parameter, a slope parameter for Age, and a slope parameter for Gender (since there are only two gender levels and the EFFECT parameterization is used by default). The names of the parameters and their descriptions are as follows.

Intercept_ABC	intercept parameter for the logit contrasting ABC with NBC
Intercept_CBS	intercept parameter for the logit contrasting CBS with NBC
Age_ABC	Age slope parameter for the logit contrasting ABC with NBC
Age_CBS	Age slope parameter for the logit contrasting CBS with NBC
GenderFemale_ABC	Gender=Female slope parameter for the logit contrasting ABC with NBC
GenderFemale_CBS	Gender=Female slope parameter for the logit contrasting CBS with NBC

INEST= Input Data Set

You can specify starting values for the iterative algorithm in the INEST= data set. The INEST= data set has the same structure as the OUTEST= data set but is not required to have all the variables or observations that appear in the OUTEST= data set.

The INEST= data set must contain the intercept variables (named Intercept for binary response models and Intercept, Intercept2, Intercept3, and so forth, for ordinal and nominal response models) and all explanatory variables in the MODEL statement. If BY processing is used, the INEST= data set should also include the BY variables, and there must be one observation for each BY group. If the INEST= data set also contains the _TYPE_ variable, only observations with _TYPE_ value 'PARMS' are used as starting values.

OUT= Output Data Set in the OUTPUT Statement

The OUT= data set in the OUTPUT statement contains all the variables in the input data set along with statistics you request using *keyword=name* options or the PREDPROBS= option in the OUTPUT statement. In addition, if you use the *single-trial* syntax and you request any of the XBETA=, STDXBETA=, PREDICTED=, LCL=, and UCL= options, the OUT= data set contains the automatic variable _LEVEL_. The value of _LEVEL_ identifies the response category upon which the computed values of XBETA=, STDXBETA=, PREDICTED=, LCL=, and UCL= are based.

When there are more than two response levels, only variables named by the XBETA=, STDXBETA=, PREDICTED=, LOWER=, and UPPER= options and the variables given by PREDPROBS=(INDIVIDUAL CUMULATIVE) have their values computed; the other variables have missing values. If you fit a generalized logit model, the cumulative predicted probabilities are not computed.

When there are only two response categories, each input observation produces one observation in the OUT= data set.

If there are more than two response categories and you only specify the PREDPROBS= option, then each input observation produces one observation in the OUT= data set. However, if you fit an ordinal (cumulative) model and specify options other than the PREDPROBS= options, each input observation generates as many output observations as one fewer than the number of response levels, and the predicted probabilities and their confidence limits correspond to the cumulative predicted probabilities. If you fit a generalized logit model and specify options other than the PREDPROBS= options, each input observation generates as many output observations as the number of response categories; the predicted probabilities and their confidence limits correspond to the probabilities of individual response categories.

For observations in which only the response variable is missing, values of the XBETA=, STDXBETA=, PREDICTED=, UPPER=, LOWER=, and the PREDPROBS= options are computed even though these observations do not affect the model fit. This enables, for instance, predicted probabilities to be computed for new observations.

OUT= Output Data Set in a SCORE Statement

The OUT= data set in a SCORE statement contains all the variables in the data set being scored. The data set being scored can be either the input DATA= data set in the PROC LOGISTIC statement or the DATA= data set in the SCORE statement. The DATA= data set in the SCORE statement may not contain a response variable.

If the data set being scored contains a response variable, then denote the *normalized* levels (left justified formatted values of 16 characters or less) of your response variable Y by $Y_1, \ldots, Y_{k+1}$. For each response level, the OUT= data set also contains:

- F_*Y*, the normalized levels of the response variable Y in the data set being scored. If the *events*/*trials* syntax is used, the F_*Y* variable is not created.
- I_*Y*, the normalized levels that the observations are classified into. Note that an observation is classified into the level with the largest probability. If the *events*/*trials* syntax is used, the _INTO_ variable is created instead and it contains the values EVENT and NONEVENT.
- P_Y_i, the posterior probabilities of the normalized response level Y_i.
- If the CLM option is specified in the SCORE statement, the OUT= data set also includes:
  - LCL_Y_i, the lower 100(1-α)% confidence limits for P_Y_i
  - UCL_Y_i, the upper 100(1-α)% confidence limits for P_Y_i

OUTDIST= Output Data Set

The OUTDIST= data set contains every exact conditional distribution necessary to process the EXACT statement. For example, the following statements create one distribution for the x1 parameter and another for the x2 parameters, and produces the data set dist shown in Figure 42.7:

```
proc logistic;
   class x2 / param=ref;
   model y=x1 x2;
   exact x1 x2/ outdist=dist;
proc print data=dist;
run;
```

Obs	x1	x20	x21	Count	Score	Prob
1	.	0	0	3	5.81151	0.03333
2	.	0	1	15	1.66031	0.16667
3	.	0	2	9	3.12728	0.10000
4	.	1	0	15	1.46523	0.16667
5	.	1	1	18	0.21675	0.20000
6	.	1	2	6	4.58644	0.06667
7	.	2	0	19	1.61869	0.21111
8	.	2	1	2	3.27293	0.02222
9	.	3	0	3	6.27189	0.03333
10	2	.	.	6	3.03030	0.12000
11	3	.	.	12	0.75758	0.24000
12	4	.	.	11	0.00000	0.22000
13	5	.	.	18	0.75758	0.36000
14	6	.	.	3	3.03030	0.06000

Figure 42.7. OUTDIST

The first nine observations in the dist data set contain a exact distribution for the parameters of the x2 effect (hence the values for the x1 parameter are missing), and the remaining five observations are for the x1 parameter. If a joint distribution was created, there would be observations with values for both the x1 and x2 parameters. For CLASS variables, the corresponding parameters in the dist data set are identified by concatenating the variable name with the appropriate classification level.

The data set contains the possible sufficient statistics of the parameters for the effects specified in the EXACT statement, and the Count variable contains the number of different responses that yield these statistics. For example, there were 6 possible response vectors $\mathbf{y}$ for which the dot product $\mathbf{y}'\mathbf{x1}$ was equal to 2, and for which $\mathbf{y}'\mathbf{x20}$, $\mathbf{y}'\mathbf{x21}$, and $\mathbf{y}'\mathbf{1}$ were equal to their actual observed values (displayed in the "Sufficient Statistics" table). When hypothesis tests are performed on the parameters, the Prob variable contains the probability of obtaining that statistic (which is just the count divided by the total count), and the Score variable contains the score for that statistic. For more information, see the section "EXACT Statement Examples" on page 2302.

OUTROC= Output Data Set

The OUTROC= data set contains data necessary for producing the ROC curve, and can be created by specifying the OUTROC= option in the MODEL statement or the OUTROC= option in the SCORE statement: It has the following variables:

- any BY variables specified
- _STEP_, the model step number. This variable is not included if model selection is not requested.
- _PROB_, the estimated probability of an event. These estimated probabilities serve as cutpoints for predicting the response. Any observation with an estimated event probability that exceeds or equals _PROB_ is predicted to be an event; otherwise, it is predicted to be a nonevent. Predicted probabilities

that are close to each other are grouped together, with the maximum allowable difference between the largest and smallest values less than a constant that is specified by the ROCEPS= option. The smallest estimated probability is used to represent the group.

- _POS_, the number of correctly predicted event responses
- _NEG_, the number of correctly predicted nonevent responses
- _FALPOS_, the number of falsely predicted event responses
- _FALNEG_, the number of falsely predicted nonevent responses
- _SENSIT_, the sensitivity, which is the proportion of event observations that were predicted to have an event response
- _1MSPEC_, one minus specificity, which is the proportion of nonevent observations that were predicted to have an event response

Note that none of these statistics are affected by the bias-correction method discussed in the "Classification Table" section on page 2352. An ROC curve is obtained by plotting _SENSIT_ against _1MSPEC_. For more information, see the section "Receiver Operating Characteristic Curves" on page 2357.

Computational Resources

The memory needed to fit an unconditional model is approximately $24(p+2)^2$ bytes, where p is the number of parameters estimated. For cumulative response models with more than two response levels, a test of the parallel lines assumption requires an additional memory of approximately $4k^2(m+1)^2+24(m+2)^2$ bytes, where k is the number of response levels and m is the number of slope parameters. However, if this additional memory is not available, the procedure skips the test and finishes the other computations. You may need more memory if you use the SELECTION= option for model building.

The data that consist of relevant variables (including the design variables for model effects) and observations for fitting the model are stored in the utility file. If sufficient memory is available, such data will also be kept in memory; otherwise, the data are reread from the utility file for each evaluation of the likelihood function and its derivatives, with the resulting execution time of the procedure substantially increased.

If a conditional logistic regression is performed, then approximately $4(m^2+m+4)\max_h(m_h)+(8s_H+36)H+12s_H$ additional bytes of memory are needed, where m_h is the number of events in stratum h, H is the total number of strata, and s_H is the number of variables used to define the strata.

Computational Resources for Exact Conditional Logistic Regression

Many problems require a prohibitive amount of time and memory for exact computations, depending on the speed and memory available on your computer. For such problems, consider whether exact methods are really necessary. Stokes, Davis, and Koch (2000) suggest looking at exact p-values when the sample size is small and the approximate p-values from the unconditional analysis are less than 0.10, and they provide *rules of thumb* for determining when various models are valid.

A formula does not exist that can predict the amount of time and memory necessary to generate the exact conditional distributions for a particular problem. The time and memory required depends on several factors, including the total sample size, the number of parameters of interest, the number of nuisance parameters, and the order in which the parameters are processed. To provide a feel for how these factors affect performance, 19 data sets containing Nobs $\in \{10, \ldots, 500\}$ observations consisting of up to 10 independent uniform binary covariates (X1,. . .,*XN*) and a binary response variable (Y), are generated and exact conditional distributions are created for X1 conditional on the other covariates using the default METHOD=NETWORK. Figure 42.8 displays results obtained on a 400Mhz PC with 768MB RAM running Microsoft Windows NT.

```
data one;
   do obs=1 to HalfNobs;
      do Y=0 to 1;
         X1=round(ranuni(0));
         ...
         XN=round(ranuni(0));
         output;
      end;
   end;
options fullstimer;
proc logistic exactonly exactoptions(method=network maxtime=1200);
   class X1 ... XN / param=ref;
   model Y=X1 ... XN;
   exact X1 / outdist=dist;
run;
```

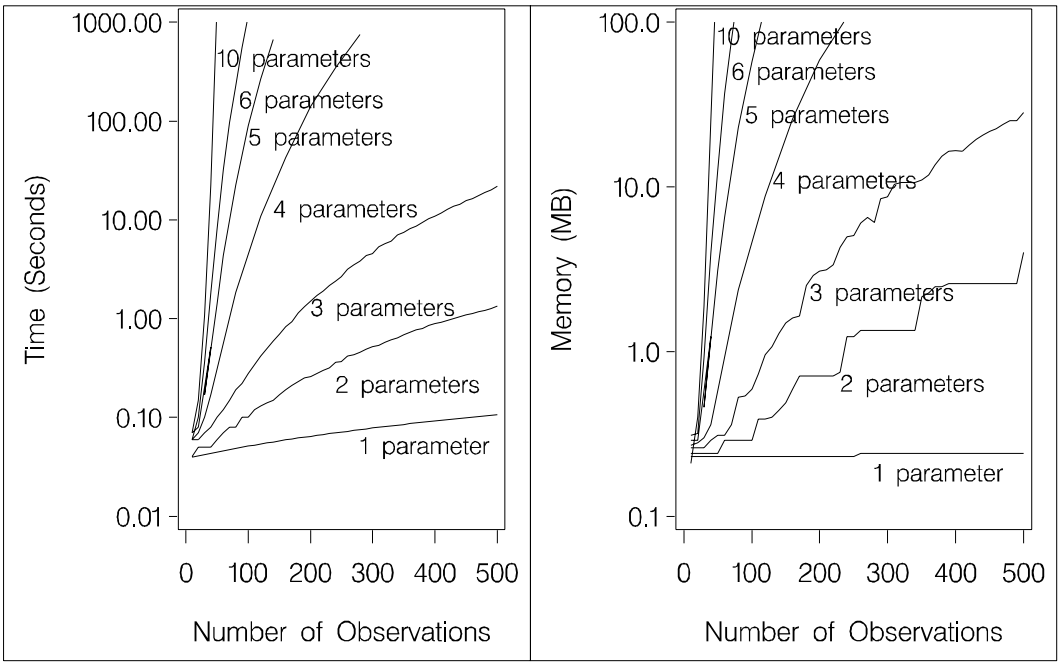


Figure 42.8. Mean Time and Memory Required

At any time while PROC LOGISTIC is deriving the distributions, you can terminate the computations by pressing the system interrupt key sequence (refer to the SAS Companion for your system) and choosing to stop computations. If you run out of memory, refer to the SAS Companion for your system to see how to allocate more.

You can use the EXACTOPTIONS option MAXTIME= to limit the total amount of time PROC LOGISTIC uses to derive all of the exact distributions. If PROC LOGISTIC does not finish within that time, the procedure terminates.

Calculation of frequencies are performed in the log-scale by default. This reduces the need to check for excessively large frequencies but can be slower than not scaling. You can turn off the log-scaling by specifying the NOLOGSCALE option in the MODEL statement. If a frequency in the exact distribution is larger than the largest integer that can be held in double-precision, a warning is printed to the LOG, but since inaccuracies due to adding small numbers to these large frequencies may have little-or-no effect on the statistics, the exact computations continue.

You can monitor the progress of the procedure by submitting your program with the EXACTOPTIONS option STATUSTIME=. If the procedure is too slow, you can try another method by specifying the EXACTOPTIONS option METHOD=, you can try reordering the variables in the MODEL statement (note that CLASS variables are always processed before continuous covariates), or you can try reparameterizing your classification variables, for example:

```
class  class-variables / param=ref ref=first order=freq;
```

Displayed Output

If you use the NOPRINT option in the PROC LOGISTIC statement, the procedure does not display any output. Otherwise, the displayed output of the LOGISTIC procedure includes the following:

- "Model Information" table, which gives
  - name of the input Data Set
  - name and label of the Response Variable, if the *single-trial* syntax is used
  - number of Response Levels, if the *single-trial* syntax is used
  - name of the Events Variable, if the *events/trials* syntax is used
  - name of the Trials Variable, if the *events/trials* syntax is used
  - name of the Offset Variable, if the OFFSET= option is specified
  - name of the Frequency Variable, if the FREQ statement is specified
  - name of the Weight Variable, if the WEIGHT statement is specified
  - Number of Strata, if the STRATA statement is specified
  - Number of Strata Ignored and the total Frequency Ignored, if the STRATA statement is specified and at least one stratum has no events or no nonevents
  - Link Function

  - Optimization Technique
  - seed, if METHOD=NETWORKMC is specified

- "Number of Observations" table, which gives

  - Number of Observations read from the input data set
  - Number of Observations used in the analysis
  - Sum of Frequencies of all the observations read from the input data set
  - Sum of Frequencies of all the observations used in the analysis
  - Sum of Weights of all the observations read from the input data set
  - Sum of Weights of all the observations used in the analysis
  - Normalized Sum of Weights of all the observations used in the analysis, if the SCALE=WILLIAMS option is specified in the MODEL statement or the NORMALIZE option is specified in the WEIGHT statement.

  An ODS OUTPUT data set created from this table contains all of the information in every row.

- "Response Profile" table, which gives, for each response level, the ordered value (an integer between one and the number of response levels, inclusive); the value of the response variable if the *single-trial* syntax is used or the values "Event" and "Nonevent" if the *events/trials* syntax is used; the count or frequency; and the sum of weights if the WEIGHT statement is specified
- "Class Level Information" table, which gives the level and the design variables for each CLASS explanatory variable
- "Descriptive Statistics for Continuous Explanatory Variables" table for continuous explanatory variables, the "Frequency Distribution of Class Variables," and the "Weight Distribution of Class Variables" tables (if the WEIGHT statement is specified), if you specify the SIMPLE option in the PROC LOGISTIC statement. The "Descriptive Statistics for Continuous Explanatory Variables" table contains the mean, standard deviation, maximum and minimum of each continuous variable specified in the MODEL statement.
- "Maximum Likelihood Iterative Phase" table, if you use the ITPRINT option in the MODEL statement. This table gives the iteration number, the step size (in the scale of 1.0, .5, .25, and so on) or the ridge value, -2 log likelihood, and parameter estimates for each iteration. Also displayed are the last evaluation of the gradient vector and the last change in the -2 log likelihood.
- Pearson and deviance goodness-of-fit statistics, if you use the SCALE= option in the MODEL statement
- score test result for testing the parallel lines assumption, if an ordinal response model is fitted. If LINK=CLOGLOG or LINK=PROBIT, this test is labeled "Score Test for the Parallel Slopes Assumption." The proportion odds assumption is a special case of the parallel lines assumption when LINK=LOGIT. In this case, the test is labeled "Score Test for the Proportional Odds Assumption".
- "Model Fit Statistics" and "Testing Global Null Hypothesis: BETA=0" tables, which give the various criteria (-2 Log L, AIC, SC) based on the likelihood

for fitting a model with intercepts only and for fitting a model with intercepts and explanatory variables. If you specify the NOINT option, these statistics are calculated without considering the intercept parameters. The third column of the table gives the chi-square statistics and p-values for the -2 Log L statistic and for the Score statistic. These test the joint effect of the explanatory variables included in the model. The Score criterion is always missing for the models identified by the first two columns of the table. Note also that the first two rows of the Chi-Square column are always missing, since tests cannot be performed for AIC and SC.

- generalized R^2 measures for the fitted model, if you specify the RSQUARE option in the MODEL statement
- "Type 3 Analysis of Effects" table, if the model contains an effect involving a CLASS variable. This table gives the Wald Chi-square statistic, the degrees of freedom, and the p-value for each effect in the model
- "Analysis of Maximum Likelihood Estimates" table, which includes
  - parameter name, which also identifies the CLASS variable level and, for generalized logit models, a response variable column to identify the corresponding logit by displaying the nonreference level of the logit
  - maximum likelihood estimate of the parameter
  - estimated standard error of the parameter estimate, computed as the square root of the corresponding diagonal element of the estimated covariance matrix
  - Wald chi-square statistic, computed by squaring the ratio of the parameter estimate divided by its standard error estimate
  - p-value of the Wald chi-square statistic with respect to a chi-square distribution with one degree of freedom
  - standardized estimate for the slope parameter, if you specify the STB option in the MODEL statement. This estimate is given by $\hat{\beta}_i/(s/s_i)$, where s_i is the total sample standard deviation for the ith explanatory variable and

    $$s = \begin{cases} \pi/\sqrt{3} & \text{Logistic} \\ 1 & \text{Normal} \\ \pi/\sqrt{6} & \text{Extreme-value} \end{cases}$$

    Standardized estimates of the intercept parameters are set to missing.
  - $e^{\hat{\beta}_i}$ for each slope parameter β_i, if you specify the EXPB option in the MODEL statement. For continuous variables, this is equivalent to the estimated odds ratio for a 1 unit change.
  - label of the variable, if you specify the PARMLABEL option in the MODEL statement and if space permits. Due to constraints on the line size, the variable label may be suppressed in order to display the table in one panel. Use the SAS system option LINESIZE= to specify a larger line size to accommodate variable labels. A shorter line size can break the table into two panels allowing labels to be displayed.

- "Odds Ratio Estimates" table, which contains the odds ratio estimates and the corresponding 95% Wald confidence intervals. For continuous explanatory variables, these odds ratios correspond to a unit increase in the risk factors.
- "Association of Predicted Probabilities and Observed Responses" table, which includes a breakdown of the number of pairs with different responses, and four rank correlation indexes: Somers' *D*, Goodman-Kruskal Gamma, and Kendall's Tau-*a*, and *c*
- confidence intervals for all the parameters, if you use the CLPARM= option in the MODEL statement
- confidence intervals for all the odds ratios, if you use the CLODDS= option in the MODEL statement
- a summary of the model-building process, if you use a FORWARD, BACKWARD, or STEPWISE selection method. This summary gives the step number, the explanatory variables entered or removed at each step, the chi-square statistic, and the corresponding p-value on which the entry or removal of the variable is based (the score chi-square is used to determine entry; the Wald chi-square is used to determine removal)
- "Analysis of Variables Removed by Fast Backward Elimination" table, if you specify the FAST option in the MODEL statement. This table gives the approximate chi-square statistic for the variable removed, the corresponding p-value with respect to a chi-square distribution with one degree of freedom, the residual chi-square statistic for testing the joint significance of the variable and the preceding ones, the degrees of freedom, and the p-value of the residual chi-square with respect to a chi-square distribution with the corresponding degrees of freedom
- "Analysis of Effects not in the Model" table, if you specify the DETAILS option in the MODEL statement. This table gives the score chi-square statistic for testing the significance of each variable not in the model after adjusting for the variables already in the model, and the p-value of the chi-square statistic with respect to a chi-square distribution with one degree of freedom
- classification table, if you use the CTABLE option in the MODEL statement. For each prior event probability (labeled "Prob Event") specified by the PEVENT= option and each cutpoint specified in the PPROB= option, the table gives the four entries of the 2×2 table of observed and predicted responses and the percentages of correct classification, sensitivity, specificity, false positive, and false negative. The columns labeled "Correct" give the number of correctly classified events and nonevents. "Incorrect Event" gives the number of nonevents incorrectly classified as events. "Incorrect Nonevent" gives the number of nonevents incorrectly classified as events.
- estimated covariance matrix of the parameter estimates, if you use the COVB option in the MODEL statement
- estimated correlation matrix of the parameter estimates, if you use the CORRB option in the MODEL statement
- "Contrast Test Results" table, if you specify a CONTRAST statement. This table gives the result of the Wald test for each CONTRAST specified. If you

specify the E option in the CONTRAST statement, then the contrast matrix is displayed. If you specify the ESTIMATE= option in the CONTRAST statement, then estimates and Wald tests for each contrast (row of the contrast matrix) or exponentiated contrast are produced.

- "Linear Hypothesis Testing" table, if you specify a TEST statement. This table gives the result of the Wald test for each TEST statement specified. If you specify the PRINT option in the TEST statement, then matrices used in the intermediate calculations are also displayed.
- results of the Hosmer and Lemeshow test for the goodness of fit of the fitted model, if you use the LACKFIT option in the MODEL statement
- "Regression Diagnostics" table, if you use the INFLUENCE option in the MODEL statement. This table gives, for each observation, the case number (which is the observation number), the values of the explanatory variables included in the model, the Pearson residual, the deviance residual, the diagonal element of the hat matrix, the standardized difference in the estimate for each parameter (*name* DFBETA, where *name* is either Intercept or the name of an explanatory variable), two confidence interval displacement diagnostics (C and CBAR), the change in the Pearson chi-square statistic (DIFCHISQ), and the change in the deviance (DIFDEV)

  If you also specify the STRATA statement, then this table contains the case number (which is the observation number), the values of the explanatory variables included in the model, the estimated one-step Pearson residual, the estimated one-step diagonal element of the hat matrix, and the estimated one-step standardized difference in the estimate for each parameter.
- index plots of regression diagnostics, if you specify the IPLOTS option in the MODEL statement. These include plots of

  – Pearson residuals
  – deviance residuals
  – diagonal elements of the hat matrix
  – standardized differences in parameter estimates, DFBETA0 for the intercept estimate, DFBETA1 for the slope estimate of the first explanatory variable in the MODEL statement, and so on
  – confidence interval displacement diagnostics C
  – confidence interval displacement diagnostics CBAR
  – changes in the Pearson chi-square statistic
  – changes in the deviance

- if you specify a STRATA statement

  – "Strata Summary" table, which displays a pattern of the number of events and the number of non-events in a stratum, the number of strata having that pattern, and the total number of observations contained in those strata
  – "Strata Information" table, if you specify the INFO option on the STRATA statement. This table displays each stratum, its frequency, and the number of events and non-events in that stratum.

- if you specify an EXACT statement
  - "Sufficient Statistics" table, if you request an OUTDIST= data set. This table is displayed before printing any of the exact analysis results and lists the parameters and their observed sufficient statistics.
  - "Conditional Exact Tests" table, which provides two tests for the null hypothesis that the parameters for the specified effects are zero: the Exact Probability Test and the Exact Conditional Scores test. For each test, the test statistic, an exact p-value (the probability of obtaining a more extreme statistic than the observed, assuming the null hypothesis), and a mid p-value (which adjusts for the discreteness of the distribution) are displayed.
  - "Exact Parameter Estimates" table, if you specify the ESTIMATE, ESTIMATE=PARM, or ESTIMATE=BOTH options. This table gives individual parameter estimates for each variable (conditional on the values of all the other parameters in the model), confidence limits, and a two-sided p-value (twice the one-sided p-value) for testing that the parameter is zero.
  - "Exact Odds Ratios" table, if you specify the ESTIMATE=ODDS or ESTIMATE=BOTH options. This table gives odds ratio estimates for the individual parameters, confidence limits, and a two-sided p-value for testing that the odds ratio is 1.

ODS Table Names

PROC LOGISTIC assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in the following table. For more information on ODS, see Chapter 14, "Using the Output Delivery System."

Table 42.2. ODS Tables Produced in PROC LOGISTIC

ODS Table Name	Description	Statement	Option
Association	Association of predicted probabilities and observed responses	MODEL	default
BestSubsets	Best subset selection	MODEL	SELECTION=SCORE
ClassFreq	Frequency breakdown of CLASS variables	PROC	Simple (with CLASS vars)
ClassLevelInfo	CLASS variable levels and design variables	MODEL	default (with CLASS vars)
Classification	Classification table	MODEL	CTABLE
ClassWgt	Weight breakdown of CLASS variables	PROC, WEIGHT	Simple (with CLASS vars)
CLOddsPL	Profile likelihood confidence limits for odds ratios	MODEL	CLODDS=PL
CLOddsWald	Wald's confidence limits for odds ratios	MODEL	CLODDS=WALD

Table 42.2. (continued)

ODS Table Name	Description	Statement	Option
CLParmPL	Profile likelihood confidence limits for parameters	MODEL	CLPARM=PL
CLParmWald	Wald's confidence limits for parameters	MODEL	CLPARM=WALD
ContrastCoeff	L matrix from CONTRAST	CONTRAST	E
ContrastEstimate	Estimates from CONTRAST	CONTRAST	ESTIMATE=
ContrastTest	Wald test for CONTRAST	CONTRAST	default
ConvergenceStatus	Convergence status	MODEL	default
CorrB	Estimated correlation matrix of parameter estimators	MODEL	CORRB
CovB	Estimated covariance matrix of parameter estimators	MODEL	COVB
CumulativeModelTest	Test of the cumulative model assumption	MODEL	(ordinal response)
EffectNotInModel	Test for effects not in model	MODEL	SELECTION=S/F
ExactOddsRatio	Exact Odds Ratios	EXACT	ESTIMATE=ODDS, ESTIMATE=BOTH
ExactParmEst	Parameter Estimates	EXACT	ESTIMATE, ESTIMATE=PARM, ESTIMATE=BOTH
ExactTests	Conditional Exact Tests	EXACT	default
FastElimination	Fast backward elimination	MODEL	SELECTION=B,FAST
FitStatistics	Model fit statistics	MODEL	default
GlobalScore	Global score test	MODEL	NOFIT
GlobalTests	Test for global null hypothesis	MODEL	default
GoodnessOfFit	Pearson and deviance goodness-of-fit tests	MODEL	SCALE
IndexPlots	Batch capture of the index plots	MODEL	IPLOTS
Influence	Regression diagnostics	MODEL	INFLUENCE
IterHistory	Iteration history	MODEL	ITPRINT
LackFitChiSq	Hosmer-Lemeshow chi-square test results	MODEL	LACKFIT
LackFitPartition	Partition for the Hosmer-Lemeshow test	MODEL	LACKFIT
LastGradient	Last evaluation of gradient	MODEL	ITPRINT
LogLikeChange	Final change in the log likelihood	MODEL	ITPRINT
ModelBuildingSummary	Summary of model building	MODEL	SELECTION=B/F/S
ModelInfo	Model information	PROC	default
NObs	Number of Observations	PROC	default
OddsRatios	Odds ratios	MODEL	default

Table 42.2. (continued)

ODS Table Name	Description	Statement	Option
ParameterEstimates	Maximum likelihood estimates of model parameters	MODEL	default
RSquare	R-square	MODEL	RSQUARE
ResidualChiSq	Residual chi-square	MODEL	SELECTION=F/B
ResponseProfile	Response profile	PROC	default
SimpleStatistics	Summary statistics for explanatory variables	PROC	SIMPLE
StrataSummary	Number of strata with specific response frequencies	STRATA	default
StrataInfo	Event and non-event frequencies for each stratum	STRATA	INFO
SuffStats	Sufficient Statistics	EXACT	OUTDIST=
TestPrint1	**L**[cov(**b**)]**L**' and **Lb-c**	TEST	PRINT
TestPrint2	Ginv(**L**[cov(**b**)]**L**') and Ginv(**L**[cov(**b**)]**L**')(**Lb-c**)	TEST	PRINT
TestStmts	Linear hypotheses testing results	TEST	default
Type3	Type 3 tests of effects	MODEL	default (with CLASS variables)

ODS Graphics (Experimental)

This section describes the use of ODS for creating graphics with the LOGISTIC procedure. These graphics are experimental in this release, meaning that both the graphical results and the syntax for specifying them are subject to change in a future release.

To request these graphs you must specify the ODS GRAPHICS statement in addition to options on the MODEL or GRAPHICS statement as described in the following sections. For more information on the ODS GRAPHICS statement, see Chapter 15, "Statistical Graphics Using ODS."

MODEL Statement Options

If the INFLUENCE or IPLOTS option is specified in the MODEL statement, then the lineprinter plots are suppressed and ODS GRAPHICS versions of the plots are produced.

If you specify the OUTROC= option, and if ROCEPS= is not specified, then ROC curves are produced. If you also specify a SELECTION= method then an overlaid plot of all the ROC curves for each step of the selection process is displayed.

GRAPHICS Statement and Options

GRAPHICS *options* **;**

The GRAPHICS statement provides options for requesting and modifying certain graphical displays. This statement has no effect unless ODS GRAPHICS ON has been specified. The functionality of this statement may be replaced by alternative syntax in a future release.

The following options are available.

DFBETAS displays the DFBETAS versus Case Number plots. This acts like DFBETAS=_ALL_ in the OUTPUT statement. These plots are produced by default when the GRAPHICS statement is specified.

HATDIAG displays plots of DIFCHISQ, DIFDEV, and DFBETAS (when the DFBETAS option is specified) versus the hat diagonals.

INFLUENCE | INDEX displays the INFLUENCE plots with no DFBETAS. These plots are produced by default when the GRAPHICS statement is specified.

PHAT displays plots of DIFCHISQ, DIFDEV, and DFBETAS (when the DFBETAS option is specified) versus the predicted event probability.

ALL invokes the DFBETAS, HATDIAG, INFLUENCE, and PHAT options.

NOINFLUENCE suppresses the default INFLUENCE and DFBETAS plots.

NOPANELS unpanels the graphical displays and produces a series of plots which form the panelled display.

ROC displays the ROC curve. If the ROCEPS= option is specified on the MODEL statement then it must be equal to zero, otherwise no ROC curve is produced. If you also specify a SELECTION= method then an overlaid plot of all the ROC curves for each step of the selection process is displayed.

ESTPROB(*fit-options***)** displays the fit curves for the model when only one continuous covariate is specified in the model. If you use events/trials syntax, then this displays the estimated event probability and the prediction limits versus the covariate with the observed proportions overlaid on the graph. If you use single-trial syntax, this displays the estimated event probability and the prediction limits versus the covariate with the observed responses overlaid on the graph. If you specify a polytomous logit model, then the estimated probabilities for each possible response level are graphed. If you have an OFFSET= variable with more than one value, then the prediction curves are replaced with error bars and the estimated probabilities are displayed at the observed covariate values.

The following *fit-options* are available with the ESTPROB option.

ALPHA=α specifies the size of the prediction interval. The ALPHA= value specified on the PROC statement is the default. If neither ALPHA= value is specified, then ALPHA=0.05 by default.

GRIDSIZE=n specifies the number of equally-spaced points at which the fit curve is computed. By default, GRIDSIZE=50.

OBSERVE specifies that the fit curve should be computed at the observed values only.

See Example 42.6 on page 2422 and Example 42.7 on page 2429 for examples of the ODS graphical displays.

ODS Graph Names

PROC LOGISTIC assigns a name to each graph it creates using ODS. You can use these names to reference the graphs when using ODS. The names are listed in Table 42.3.

To request these graphs you must specify the ODS GRAPHICS statement in addition to the options indicated in Table 42.3. For more information on the ODS GRAPHICS statement, see Chapter 15, "Statistical Graphics Using ODS."

Table 42.3. ODS Graphics Produced by PROC LOGISTIC

ODS Graph Name	Plot Description	Statement	Option
InfluencePlots	Panel of influence statistics vs. case number	GRAPHICS or MODEL	INFLUENCE INFLUENCE or IPLOTS
PearsonChisquarePlot	Pearson chi-square residual vs. case number	GRAPHICS	INFLUENCE NOPANELS
DevianceResidualPlot	Deviance residual vs. case number	GRAPHICS	INFLUENCE NOPANELS
HatPlot	Hat diagonal vs. case number	GRAPHICS	INFLUENCE NOPANELS
CPlot	CI displacement C vs. case number	GRAPHICS	INFLUENCE NOPANELS
CBarPlot	CI displacement Cbar vs. case number	GRAPHICS	INFLUENCE NOPANELS
DeltaChisqPlot	Difchisq vs. case number	GRAPHICS	INFLUENCE NOPANELS
DeltaDeviancePlot	Difdev vs. case number	GRAPHICS	INFLUENCE NOPANELS
DFBetasPlot	DFBetas vs. case number	GRAPHICS	DFBETAS NOPANELS
EstProbPlots	Panel of estimated probability vs. influence	GRAPHICS	PHAT
PhatDifChisqPlot	Estimated probability vs. difchisq	GRAPHICS	PHAT NOPANELS
PhatDifDevPlot	Estimated probability vs. difdev	GRAPHICS	PHAT NOPANELS
PhatDFBetasPlot	Estimated probability vs. dfbetas	GRAPHICS	PHAT NOPANELS
HatDiagPlots	Panel of hat diagonals vs. influence statistics	GRAPHICS	HATDIAG

Table 42.3. (continued)

ODS Graph Name	Plot Description	Statement	Option
HatDiagDifChisqPlot	Hat diagonals vs. difchisq	GRAPHICS	HATDIAG NOPANELS
HatDiagDifDevPlot	Hat diagonals vs. difdev	GRAPHICS	HATDIAG NOPANELS
HatDiagDFBetasPlot	Hat diagonals vs. dfbetas	GRAPHICS	HATDIAG NOPANELS
ROCCurve	Receiver operating characteristics curve	GRAPHICS or MODEL	ROC OUTROC=
ROCOverlay	ROC curves for model selection steps	GRAPHICS and MODEL	ROC SELECTION=
FitCurve	Estimated probability vs. one continuous covariate	GRAPHICS	ESTPROB

Examples

Example 42.1. Stepwise Logistic Regression and Predicted Values

Consider a study on cancer remission (Lee 1974). The data, consisting of patient characteristics and whether or not cancer remission occurred, are saved in the data set Remission.

```
data Remission;
   input remiss cell smear infil li blast temp;
   label remiss='Complete Remission';
   datalines;
1   .8   .83  .66  1.9  1.1     .996
1   .9   .36  .32  1.4   .74    .992
0   .8   .88  .7    .8   .176   .982
0  1     .87  .87   .7  1.053   .986
1   .9   .75  .68  1.3   .519   .98
0  1     .65  .65   .6   .519   .982
1   .95  .97  .92  1    1.23    .992
0   .95  .87  .83  1.9  1.354  1.02
0  1     .45  .45   .8   .322   .999
0   .95  .36  .34   .5  0      1.038
0   .85  .39  .33   .7   .279   .988
0   .7   .76  .53  1.2   .146   .982
0   .8   .46  .37   .4   .38   1.006
0   .2   .39  .08   .8   .114   .99
0  1     .9   .9   1.1  1.037   .99
1  1     .84  .84  1.9  2.064  1.02
0   .65  .42  .27   .5   .114  1.014
0  1     .75  .75  1    1.322  1.004
0   .5   .44  .22   .6   .114   .99
1  1     .63  .63  1.1  1.072   .986
0  1     .33  .33   .4   .176  1.01
0   .9   .93  .84   .6  1.591  1.02
1  1     .58  .58  1     .531  1.002
0   .95  .32  .3   1.6   .886   .988
1  1     .6   .6   1.7   .964   .99
```

```
1  1     .69  .69   .9   .398   .986
0  1     .73  .73   .7   .398   .986
;
```

The data set Remission contains seven variables. The variable remiss is the cancer remission indicator variable with a value of 1 for remission and a value of 0 for nonremission. The other six variables are the risk factors thought to be related to cancer remission.

The following invocation of PROC LOGISTIC illustrates the use of stepwise selection to identify the prognostic factors for cancer remission. A significance level of 0.3 (SLENTRY=0.3) is required to allow a variable into the model, and a significance level of 0.35 (SLSTAY=0.35) is required for a variable to stay in the model. A detailed account of the variable selection process is requested by specifying the DETAILS option. The Hosmer and Lemeshow goodness-of-fit test for the final selected model is requested by specifying the LACKFIT option. The OUTEST= and COVOUT options in the PROC LOGISTIC statement create a data set that contains parameter estimates and their covariances for the final selected model. The response variable option EVENT= sets remiss=1 (remission) to be Ordered Value 1 so that the probability of remission is modeled. The OUTPUT statement creates a data set that contains the cumulative predicted probabilities and the corresponding confidence limits, and the individual and cross validated predicted probabilities for each observation.

```
title 'Stepwise Regression on Cancer Remission Data';
proc logistic data=Remission outest=betas covout;
   model remiss(event='1')=cell smear infil li blast temp
                / selection=stepwise
                  slentry=0.3
                  slstay=0.35
                  details
                  lackfit;
   output out=pred p=phat lower=lcl upper=ucl
          predprob=(individual crossvalidate);
run;

proc print data=betas;
   title2 'Parameter Estimates and Covariance Matrix';
run;

proc print data=pred;
   title2 'Predicted Probabilities and 95% Confidence Limits';
run;
```

In stepwise selection, an attempt is made to remove any insignificant variables from the model before adding a significant variable to the model. Each addition or deletion of a variable to or from a model is listed as a separate step in the displayed output, and at each step a new model is fitted. Details of the model selection steps are shown in Output 42.1.1 –Output 42.1.5.

Output 42.1.1. Startup Model

Stepwise Regression on Cancer Remission Data

The LOGISTIC Procedure

Model Information

Data Set	WORK.REMISSION	
Response Variable	remiss	Complete Remission
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	27
Number of Observations Used	27

Response Profile

Ordered Value	remiss	Total Frequency
1	0	18
2	1	9

Probability modeled is remiss=1.

Stepwise Selection Procedure

Step 0. Intercept entered:

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.6931	0.4082	2.8827	0.0895

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
9.4609	6	0.1493

Analysis of Effects Eligible for Entry

Effect	DF	Score Chi-Square	Pr > ChiSq
cell	1	1.8893	0.1693
smear	1	1.0745	0.2999
infil	1	1.8817	0.1701
li	1	7.9311	0.0049
blast	1	3.5258	0.0604
temp	1	0.6591	0.4169

Output 42.1.2. Step 1 of the Stepwise Analysis

Step 1. Effect li entered:

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	36.372	30.073
SC	37.668	32.665
-2 Log L	34.372	26.073

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	8.2988	1	0.0040
Score	7.9311	1	0.0049
Wald	5.9594	1	0.0146

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.7771	1.3786	7.5064	0.0061
li	1	2.8973	1.1868	5.9594	0.0146

Association of Predicted Probabilities and Observed Responses

Percent Concordant	84.0	Somers' D	0.710
Percent Discordant	13.0	Gamma	0.732
Percent Tied	3.1	Tau-a	0.328
Pairs	162	c	0.855

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
3.1174	5	0.6819

NOTE: No effects for the model in Step 1 are removed.

Analysis of Effects Eligible for Entry

Effect	DF	Score Chi-Square	Pr > ChiSq
cell	1	1.1183	0.2903
smear	1	0.1369	0.7114
infil	1	0.5715	0.4497
blast	1	0.0932	0.7601
temp	1	1.2591	0.2618

Output 42.1.3. Step 2 of the Stepwise Analysis

Step 2. Effect temp entered:

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	36.372	30.648
SC	37.668	34.535
-2 Log L	34.372	24.648

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	9.7239	2	0.0077
Score	8.3648	2	0.0153
Wald	5.9052	2	0.0522

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	47.8448	46.4381	1.0615	0.3029
li	1	3.3017	1.3593	5.9002	0.0151
temp	1	-52.4214	47.4897	1.2185	0.2697

Association of Predicted Probabilities and Observed Responses

Percent Concordant	87.0	Somers' D	0.747
Percent Discordant	12.3	Gamma	0.752
Percent Tied	0.6	Tau-a	0.345
Pairs	162	c	0.873

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
2.1429	4	0.7095

NOTE: No effects for the model in Step 2 are removed.

Analysis of Effects Eligible for Entry

Effect	DF	Score Chi-Square	Pr > ChiSq
cell	1	1.4700	0.2254
smear	1	0.1730	0.6775
infil	1	0.8274	0.3630
blast	1	1.1013	0.2940

Output 42.1.4. Step 3 of the Stepwise Analysis

Step 3. Effect cell entered:

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	36.372	29.953
SC	37.668	35.137
-2 Log L	34.372	21.953

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	12.4184	3	0.0061
Score	9.2502	3	0.0261
Wald	4.8281	3	0.1848

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	67.6339	56.8875	1.4135	0.2345
cell	1	9.6521	7.7511	1.5507	0.2130
li	1	3.8671	1.7783	4.7290	0.0297
temp	1	-82.0737	61.7124	1.7687	0.1835

Association of Predicted Probabilities and Observed Responses

Percent Concordant	88.9	Somers' D	0.778
Percent Discordant	11.1	Gamma	0.778
Percent Tied	0.0	Tau-a	0.359
Pairs	162	c	0.889

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
0.1831	3	0.9803

NOTE: No effects for the model in Step 3 are removed.

Analysis of Effects Eligible for Entry

Effect	DF	Score Chi-Square	Pr > ChiSq
smear	1	0.0956	0.7572
infil	1	0.0844	0.7714
blast	1	0.0208	0.8852

NOTE: No (additional) effects met the 0.3 significance level for entry into the model.

Output 42.1.5. Summary of the Stepwise Selection

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	li		1	1	7.9311		0.0049
2	temp		1	2	1.2591		0.2618
3	cell		1	3	1.4700		0.2254

Prior to the first step, the intercept-only model is fitted and individual score statistics for the potential variables are evaluated (Output 42.1.1). In Step 1 (Output 42.1.2), variable li is selected into the model since it is the most significant variable among those to be chosen ($p = 0.0049 < 0.3$). The intermediate model that contains an intercept and li is then fitted. li remains significant ($p = 0.0146 < 0.35$) and is not removed. In Step 2 (Output 42.1.3), variable temp is added to the model. The model then contains an intercept and variables li and temp. Both li and temp remain significant at 0.035 level; therefore, neither li nor temp is removed from the model. In Step 4 (Output 42.1.4), variable cell is added to the model. The model then contains an intercept and variables li, temp, and cell. None of these variables are removed from the model since all are significant at the 0.35 level. Finally, none of the remaining variables outside the model meet the entry criterion, and the stepwise selection is terminated. A summary of the stepwise selection is displayed in Output 42.1.5.

Output 42.1.6. Display of the LACKFIT Option

Partition for the Hosmer and Lemeshow Test

Group	Total	remiss = 1 Observed	remiss = 1 Expected	remiss = 0 Observed	remiss = 0 Expected
1	3	0	0.00	3	3.00
2	3	0	0.01	3	2.99
3	3	0	0.19	3	2.81
4	3	0	0.56	3	2.44
5	4	1	1.09	3	2.91
6	3	2	1.35	1	1.65
7	3	2	1.84	1	1.16
8	3	3	2.15	0	0.85
9	2	1	1.80	1	0.20

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
6.2983	7	0.5054

Results of the Hosmer and Lemeshow test are shown in Output 42.1.6. There is no evidence of a lack of fit in the selected model ($p = 0.5054$).

Output 42.1.7. Data Set of Estimates and Covariances

Stepwise Regression on Cancer Remission Data
Parameter Estimates and Covariance Matrix

Obs	_LINK_	_TYPE_	_STATUS_	_NAME_	Intercept	cell
1	LOGIT	PARMS	0 Converged	remiss	67.63	9.652
2	LOGIT	COV	0 Converged	Intercept	3236.19	157.097
3	LOGIT	COV	0 Converged	cell	157.10	60.079
4	LOGIT	COV	0 Converged	smear	.	.
5	LOGIT	COV	0 Converged	infil	.	.
6	LOGIT	COV	0 Converged	li	64.57	6.945
7	LOGIT	COV	0 Converged	blast	.	.
8	LOGIT	COV	0 Converged	temp	-3483.23	-223.669

Obs	smear	infil	li	blast	temp	_LNLIKE_
1	.	.	3.8671	.	-82.07	-10.9767
2	.	.	64.5726	.	-3483.23	-10.9767
3	.	.	6.9454	.	-223.67	-10.9767
4	.	.	.	.	.	-10.9767
5	.	.	.	.	.	-10.9767
6	.	.	3.1623	.	-75.35	-10.9767
7	.	.	.	.	.	-10.9767
8	.	.	-75.3513	.	3808.42	-10.9767

The data set betas created by the OUTEST= and COVOUT options is displayed in Output 42.1.7. The data set contains parameter estimates and the covariance matrix for the final selected model. Note that all explanatory variables listed in the MODEL statement are included in this data set; however, variables that are not included in the final model have all missing values.

Output 42.1.8. Predicted Probabilities and Confidence Intervals

Stepwise Regression on Cancer Remission Data
Predicted Probabilities and 95% Confidence Limits

Obs	remiss	cell	smear	infil	li	blast	temp	_FROM_	_INTO_	IP_0	IP_1	XP_0	XP_1	_LEVEL_	phat	lcl	ucl
1	1	0.80	0.83	0.66	1.9	1.100	0.996	1	1	0.27735	0.72265	0.43873	0.56127	1	0.72265	0.16892	0.97093
2	1	0.90	0.36	0.32	1.4	0.740	0.992	1	1	0.42126	0.57874	0.47461	0.52539	1	0.57874	0.26788	0.83762
3	0	0.80	0.88	0.70	0.8	0.176	0.982	0	0	0.89540	0.10460	0.87060	0.12940	1	0.10460	0.00781	0.63419
4	0	1.00	0.87	0.87	0.7	1.053	0.986	0	0	0.71742	0.28258	0.67259	0.32741	1	0.28258	0.07498	0.65683
5	1	0.90	0.75	0.68	1.3	0.519	0.980	1	1	0.28582	0.71418	0.36901	0.63099	1	0.71418	0.25218	0.94876
6	0	1.00	0.65	0.65	0.6	0.519	0.982	0	0	0.72911	0.27089	0.67269	0.32731	1	0.27089	0.05852	0.68951
7	1	0.95	0.97	0.92	1.0	1.230	0.992	1	0	0.67844	0.32156	0.72923	0.27077	1	0.32156	0.13255	0.59516
8	0	0.95	0.87	0.83	1.9	1.354	1.020	0	1	0.39277	0.60723	0.09906	0.90094	1	0.60723	0.10572	0.95287
9	0	1.00	0.45	0.45	0.8	0.322	0.999	0	0	0.83368	0.16632	0.80864	0.19136	1	0.16632	0.03018	0.56123
10	0	0.95	0.36	0.34	0.5	0.000	1.038	0	0	0.99843	0.00157	0.99840	0.00160	1	0.00157	0.00000	0.68962
11	0	0.85	0.39	0.33	0.7	0.279	0.988	0	0	0.92715	0.07285	0.91723	0.08277	1	0.07285	0.00614	0.49982
12	0	0.70	0.76	0.53	1.2	0.146	0.982	0	0	0.82714	0.17286	0.63838	0.36162	1	0.17286	0.00637	0.87206
13	0	0.80	0.46	0.37	0.4	0.380	1.006	0	0	0.99654	0.00346	0.99644	0.00356	1	0.00346	0.00001	0.46530
14	0	0.20	0.39	0.08	0.8	0.114	0.990	0	0	0.99982	0.00018	0.99981	0.00019	1	0.00018	0.00000	0.96482
15	0	1.00	0.90	0.90	1.1	1.037	0.990	0	1	0.42878	0.57122	0.35354	0.64646	1	0.57122	0.25303	0.83973
16	1	1.00	0.84	0.84	1.9	2.064	1.020	1	1	0.28530	0.71470	0.47213	0.52787	1	0.71470	0.15362	0.97189
17	0	0.65	0.42	0.27	0.5	0.114	1.014	0	0	0.99938	0.00062	0.99937	0.00063	1	0.00062	0.00000	0.62665
18	0	1.00	0.75	0.75	1.0	1.322	1.004	0	0	0.77711	0.22289	0.73612	0.26388	1	0.22289	0.04483	0.63670
19	0	0.50	0.44	0.22	0.6	0.114	0.990	0	0	0.99846	0.00154	0.99842	0.00158	1	0.00154	0.00000	0.79644
20	1	1.00	0.63	0.63	1.1	1.072	0.986	1	1	0.35089	0.64911	0.42053	0.57947	1	0.64911	0.26305	0.90555
21	0	1.00	0.33	0.33	0.4	0.176	1.010	0	0	0.98307	0.01693	0.98170	0.01830	1	0.01693	0.00029	0.50475
22	0	0.90	0.93	0.84	0.6	1.591	1.020	0	0	0.99378	0.00622	0.99348	0.00652	1	0.00622	0.00003	0.56062
23	1	1.00	0.58	0.58	1.0	0.531	1.002	1	0	0.74739	0.25261	0.84423	0.15577	1	0.25261	0.06137	0.63597
24	0	0.95	0.32	0.30	1.6	0.886	0.988	0	1	0.12989	0.87011	0.03637	0.96363	1	0.87011	0.40910	0.98481
25	1	1.00	0.60	0.60	1.7	0.964	0.990	1	1	0.06868	0.93132	0.08017	0.91983	1	0.93132	0.44114	0.99573
26	1	1.00	0.69	0.69	0.9	0.398	0.986	1	0	0.53949	0.46051	0.62312	0.37688	1	0.46051	0.16612	0.78529
27	0	1.00	0.73	0.73	0.7	0.398	0.986	0	0	0.71742	0.28258	0.67259	0.32741	1	0.28258	0.07498	0.65683

The data set pred created by the OUTPUT statement is displayed in Output 42.1.8. It contains all the variables in the input data set, the variable phat for the (cumulative) predicted probability, the variables lcl and ucl for the lower and upper confidence limits for the probability, and four other variables (viz., IP_1, IP_0, XP_1, and XP_0) for the PREDPROBS= option. The data set also contains the variable _LEVEL_, indicating the response value to which phat, lcl, and ucl refer. For instance, for the first row of the OUTPUT data set, the values of _LEVEL_ and phat, lcl, and ucl are 1, 0.72265, 0.16892 and 0.97093, respectively; this means that the estimated probability that remiss$\leq$1 is 0.723 for the given explanatory variable values, and the corresponding 95% confidence interval is (0.16892, 0.97093). The variables IP_1 and IP_0 contain the predicted probabilities that remiss=1 and remiss=0, respectively. Note that values of phat and IP_1 are identical since they both contain the probabilities that remiss=1. The variables XP_1 and XP_0 contain the cross validated predicted probabilities that remiss=1 and remiss=0, respectively.

Next, a different variable selection method is used to select prognostic factors for cancer remission, and an efficient algorithm is employed to eliminate insignificant variables from a model. The following SAS statements invoke PROC LOGISTIC to perform the backward elimination analysis.

```
title 'Backward Elimination on Cancer Remission Data';
proc logistic data=Remission;
   model remiss(event='1')=temp cell li smear blast
         / selection=backward fast slstay=0.2 ctable;
run;
```

The backward elimination analysis (SELECTION=BACKWARD) starts with a model that contains all explanatory variables given in the MODEL statement. By specifying the FAST option, PROC LOGISTIC eliminates insignificant variables without refitting the model repeatedly. This analysis uses a significance level of 0.2 (SLSTAY=0.2) to retain variables in the model, which is different from the previous stepwise analysis where SLSTAY=.35. The CTABLE option is specified to produce classifications of input observations based on the final selected model.

Output 42.1.9. Initial Step in Backward Elimination

```
                Backward Elimination on Cancer Remission Data

                           The LOGISTIC Procedure

                             Model Information

    Data Set                      WORK.REMISSION
    Response Variable             remiss                Complete Remission
    Number of Response Levels     2
    Model                         binary logit
    Optimization Technique        Fisher's scoring


                  Number of Observations Read          27
                  Number of Observations Used          27


                             Response Profile

                     Ordered                        Total
                       Value       remiss     Frequency

                           1            0            18
                           2            1             9

                     Probability modeled is remiss=1.


                       Backward Elimination Procedure

Step  0. The following effects were entered:

Intercept  temp  cell  li  smear  blast


                           Model Convergence Status

                Convergence criterion (GCONV=1E-8) satisfied.


                           Model Fit Statistics

                                                Intercept
                                   Intercept            and
                    Criterion          Only     Covariates

                    AIC              36.372         33.857
                    SC               37.668         41.632
                    -2 Log L         34.372         21.857


                  Testing Global Null Hypothesis: BETA=0

            Test                 Chi-Square       DF     Pr > ChiSq

            Likelihood Ratio        12.5146        5         0.0284
            Score                    9.3295        5         0.0966
            Wald                     4.7284        5         0.4499
```

Output 42.1.10. Fast Elimination Step

Step 1. Fast Backward Elimination:

Analysis of Effects Removed by Fast Backward Elimination

Effect Removed	Chi-Square	DF	Pr > ChiSq	Residual Chi-Square	DF	Pr > Residual ChiSq
blast	0.0008	1	0.9768	0.0008	1	0.9768
smear	0.0951	1	0.7578	0.0959	2	0.9532
cell	1.5134	1	0.2186	1.6094	3	0.6573
temp	0.6535	1	0.4189	2.2628	4	0.6875

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	36.372	30.073
SC	37.668	32.665
-2 Log L	34.372	26.073

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	8.2988	1	0.0040
Score	7.9311	1	0.0049
Wald	5.9594	1	0.0146

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
2.8530	4	0.5827

Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	blast	1	4	0.0008	0.9768
1	smear	1	3	0.0951	0.7578
1	cell	1	2	1.5134	0.2186
1	temp	1	1	0.6535	0.4189

Output 42.1.10. (continued)

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.7771	1.3786	7.5064	0.0061
li	1	2.8973	1.1868	5.9594	0.0146

Association of Predicted Probabilities and Observed Responses

Percent Concordant	84.0	Somers' D	0.710
Percent Discordant	13.0	Gamma	0.732
Percent Tied	3.1	Tau-a	0.328
Pairs	162	c	0.855

Results of the fast elimination analysis are shown in Output 42.1.9 and Output 42.1.10. Initially, a full model containing all six risk factors is fit to the data (Output 42.1.9). In the next step (Output 42.1.10), PROC LOGISTIC removes blast, smear, cell, and temp from the model all at once. This leaves li and the intercept as the only variables in the final model. Note that in this analysis, only parameter estimates for the final model are displayed because the DETAILS option has not been specified.

Note that you can also use the FAST option when SELECTION=STEPWISE. However, the FAST option operates only on backward elimination steps. In this example, the stepwise process only adds variables, so the FAST option would not be useful.

Output 42.1.11. Classifying Input Observations

Classification Table

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.060	9	0	18	0	33.3	100.0	0.0	66.7	.
0.080	9	2	16	0	40.7	100.0	11.1	64.0	0.0
0.100	9	4	14	0	48.1	100.0	22.2	60.9	0.0
0.120	9	4	14	0	48.1	100.0	22.2	60.9	0.0
0.140	9	7	11	0	59.3	100.0	38.9	55.0	0.0
0.160	9	10	8	0	70.4	100.0	55.6	47.1	0.0
0.180	9	10	8	0	70.4	100.0	55.6	47.1	0.0
0.200	8	13	5	1	77.8	88.9	72.2	38.5	7.1
0.220	8	13	5	1	77.8	88.9	72.2	38.5	7.1
0.240	8	13	5	1	77.8	88.9	72.2	38.5	7.1
0.260	6	13	5	3	70.4	66.7	72.2	45.5	18.8
0.280	6	13	5	3	70.4	66.7	72.2	45.5	18.8
0.300	6	13	5	3	70.4	66.7	72.2	45.5	18.8
0.320	6	14	4	3	74.1	66.7	77.8	40.0	17.6
0.340	5	14	4	4	70.4	55.6	77.8	44.4	22.2
0.360	5	14	4	4	70.4	55.6	77.8	44.4	22.2
0.380	5	15	3	4	74.1	55.6	83.3	37.5	21.1
0.400	5	15	3	4	74.1	55.6	83.3	37.5	21.1
0.420	5	15	3	4	74.1	55.6	83.3	37.5	21.1
0.440	5	15	3	4	74.1	55.6	83.3	37.5	21.1
0.460	4	16	2	5	74.1	44.4	88.9	33.3	23.8
0.480	4	16	2	5	74.1	44.4	88.9	33.3	23.8
0.500	4	16	2	5	74.1	44.4	88.9	33.3	23.8
0.520	4	16	2	5	74.1	44.4	88.9	33.3	23.8
0.540	3	16	2	6	70.4	33.3	88.9	40.0	27.3
0.560	3	16	2	6	70.4	33.3	88.9	40.0	27.3
0.580	3	16	2	6	70.4	33.3	88.9	40.0	27.3
0.600	3	16	2	6	70.4	33.3	88.9	40.0	27.3
0.620	3	16	2	6	70.4	33.3	88.9	40.0	27.3
0.640	3	16	2	6	70.4	33.3	88.9	40.0	27.3
0.660	3	16	2	6	70.4	33.3	88.9	40.0	27.3
0.680	3	16	2	6	70.4	33.3	88.9	40.0	27.3
0.700	3	16	2	6	70.4	33.3	88.9	40.0	27.3
0.720	2	16	2	7	66.7	22.2	88.9	50.0	30.4
0.740	2	16	2	7	66.7	22.2	88.9	50.0	30.4
0.760	2	16	2	7	66.7	22.2	88.9	50.0	30.4
0.780	2	16	2	7	66.7	22.2	88.9	50.0	30.4
0.800	2	17	1	7	70.4	22.2	94.4	33.3	29.2
0.820	2	17	1	7	70.4	22.2	94.4	33.3	29.2
0.840	0	17	1	9	63.0	0.0	94.4	100.0	34.6
0.860	0	17	1	9	63.0	0.0	94.4	100.0	34.6
0.880	0	17	1	9	63.0	0.0	94.4	100.0	34.6
0.900	0	17	1	9	63.0	0.0	94.4	100.0	34.6
0.920	0	17	1	9	63.0	0.0	94.4	100.0	34.6
0.940	0	17	1	9	63.0	0.0	94.4	100.0	34.6
0.960	0	18	0	9	66.7	0.0	100.0	.	33.3

Results of the CTABLE option are shown in Output 42.1.11. Each row of the "Classification Table" corresponds to a cutpoint applied to the predicted probabilities, which is given in the Prob Level column. The 2×2 frequency tables of observed and predicted responses are given by the next four columns. For example, with a cutpoint of 0.5, 4 events and 16 nonevents were classified correctly. On the other hand, 2 nonevents were incorrectly classified as events and 5 events were incorrectly classi-

fied as nonevents. For this cutpoint, the correct classification rate is 20/27 (=74.1%), which is given in the sixth column. Accuracy of the classification is summarized by the sensitivity, specificity, and false positive and negative rates, which are displayed in the last four columns. You can control the number of cutpoints used, and their values, by using the PPROB= option.

Example 42.2. Logistic Modeling with Categorical Predictors

Consider a study of the analgesic effects of treatments on elderly patients with neuralgia. Two test treatments and a placebo are compared. The response variable is whether the patient reported pain or not. Researchers recorded age and gender of the patients and the duration of complaint before the treatment began. The data, consisting of 60 patients, are contained in the data set Neuralgia.

```
Data Neuralgia;
   input Treatment $ Sex $ Age Duration Pain $ @@;
   datalines;
P  F  68   1  No   B  M  74  16  No  P  F  67  30  No
P  M  66  26  Yes  B  F  67  28  No  B  F  77  16  No
A  F  71  12  No   B  F  72  50  No  B  F  76   9  Yes
A  M  71  17  Yes  A  F  63  27  No  A  F  69  18  Yes
B  F  66  12  No   A  M  62  42  No  P  F  64   1  Yes
A  F  64  17  No   P  M  74   4  No  A  F  72  25  No
P  M  70   1  Yes  B  M  66  19  No  B  M  59  29  No
A  F  64  30  No   A  M  70  28  No  A  M  69   1  No
B  F  78   1  No   P  M  83   1  Yes B  F  69  42  No
B  M  75  30  Yes  P  M  77  29  Yes P  F  79  20  Yes
A  M  70  12  No   A  F  69  12  No  B  F  65  14  No
B  M  70   1  No   B  M  67  23  No  A  M  76  25  Yes
P  M  78  12  Yes  B  M  77   1  Yes B  F  69  24  No
P  M  66   4  Yes  P  F  65  29  No  P  M  60  26  Yes
A  M  78  15  Yes  B  M  75  21  Yes A  F  67  11  No
P  F  72  27  No   P  F  70  13  Yes A  M  75   6  Yes
B  F  65   7  No   P  F  68  27  Yes P  M  68  11  Yes
P  M  67  17  Yes  B  M  70  22  No  A  M  65  15  No
P  F  67   1  Yes  A  M  67  10  No  P  F  72  11  Yes
A  F  74   1  No   B  M  80  21  Yes A  F  69   3  No
;
```

The data set Neuralgia contains five variables: Treatment, Sex, Age, Duration, and Pain. The last variable, Pain, is the response variable. A specification of Pain=Yes indicates there was pain, and Pain=No indicates no pain. The variable Treatment is a categorical variable with three levels: A and B represent the two test treatments, and P represents the placebo treatment. The gender of the patients is given by the categorical variable Sex. The variable Age is the age of the patients, in years, when treatment began. The duration of complaint, in months, before the treatment began is given by the variable Duration. The following statements use the LOGISTIC procedure to fit a two-way logit with interaction model for the effect of Treatment and Sex, with Age and Duration as covariates. The categorical variables Treatment and Sex are declared in the CLASS statement.

```
proc logistic data=Neuralgia;
   class Treatment Sex;
   model Pain= Treatment Sex Treatment*Sex Age Duration / expb;
run;
```

In this analysis, PROC LOGISTIC models the probability of no pain (Pain=No). By default, effect coding is used to represent the CLASS variables. Two design variables are created for Treatment and one for Sex, as shown in Output 42.2.1.

Output 42.2.1. Effect Coding of CLASS Variables

The LOGISTIC Procedure

Class Level Information

Class	Value	Design Variables	
Treatment	A	1	0
	B	0	1
	P	-1	-1
Sex	F	1	
	M	-1	

PROC LOGISTIC displays a table of the Type 3 analysis of effects based on the Wald test (Output 42.2.2). Note that the Treatment*Sex interaction and the duration of complaint are not statistically significant (p =0.9318 and p =0.8752, respectively). This indicates that there is no evidence that the treatments affect pain differently in men and women, and no evidence that the pain outcome is related to the duration of pain.

Output 42.2.2. Wald Tests of Individual Effects

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Treatment	2	11.9886	0.0025
Sex	1	5.3104	0.0212
Treatment*Sex	2	0.1412	0.9318
Age	1	7.2744	0.0070
Duration	1	0.0247	0.8752

Parameter estimates are displayed in Output 42.2.3. The Exp(Est) column contains the exponentiated parameter estimates requested with the EXPB option. These values may, but do not necessarily, represent odds ratios for the corresponding variables. For continuous explanatory variables, the Exp(Est) value corresponds to the odds ratio for a unit increase of the corresponding variable. For CLASS variables using the effect coding, the Exp(Est) values have no direct interpretation as a comparison of levels.

However, when the reference coding is used, the Exp(Est) values represent the odds ratio between the corresponding level and the last level. Following the parameter estimates table, PROC LOGISTIC displays the odds ratio estimates for those variables that are not involved in any interaction terms. If the variable is a CLASS variable, the odds ratio estimate comparing each level with the last level is computed regardless of the coding scheme. In this analysis, since the model contains the Treatment*Sex interaction term, the odds ratios for Treatment and Sex were not computed. The odds ratio estimates for Age and Duration are precisely the values given in the Exp(Est) column in the parameter estimates table.

Output 42.2.3. Parameter Estimates with Effect Coding

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept		1	19.2236	7.1315	7.2661	0.0070	2.232E8
Treatment	A	1	0.8483	0.5502	2.3773	0.1231	2.336
Treatment	B	1	1.4949	0.6622	5.0956	0.0240	4.459
Sex	F	1	0.9173	0.3981	5.3104	0.0212	2.503
Treatment*Sex	A F	1	-0.2010	0.5568	0.1304	0.7180	0.818
Treatment*Sex	B F	1	0.0487	0.5563	0.0077	0.9302	1.050
Age		1	-0.2688	0.0996	7.2744	0.0070	0.764
Duration		1	0.00523	0.0333	0.0247	0.8752	1.005

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Age	0.764	0.629	0.929
Duration	1.005	0.942	1.073

The following PROC LOGISTIC statements illustrate the use of forward selection on the data set Neuralgia to identify the effects that differentiate the two Pain responses. The option SELECTION=FORWARD is specified to carry out the forward selection. The term Treatment|Sex@2 illustrates another way to specify main effects and two-way interaction as is available in other procedures such as PROC GLM. (Note that, in this case, the "@2" is unnecessary because no interactions besides the two-way interaction are possible).

```
proc logistic data=Neuralgia;
   class Treatment Sex;
   model Pain=Treatment|Sex@2 Age Duration
         /selection=forward expb;
run;
```

Results of the forward selection process are summarized in Output 42.2.4. The variable Treatment is selected first, followed by Age and then Sex. The results are consistent with the previous analysis (Output 42.2.2) in which the Treatment*Sex interaction and Duration are not statistically significant.

Output 42.2.4. Effects Selected into the Model

The LOGISTIC Procedure

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	Treatment	2	1	13.7143	0.0011
2	Age	1	2	10.6038	0.0011
3	Sex	1	3	5.9959	0.0143

Output 42.2.5 shows the Type 3 analysis of effects, the parameter estimates, and the odds ratio estimates for the selected model. All three variables, Treatment, Age, and Sex, are statistically significant at the 0.05 level (p =0.0011, p =0.0011, and p =0.0143, respectively). Since the selected model does not contain the Treatment*Sex interaction, odds ratios for Treatment and Sex are computed. The estimated odds ratio is 24.022 for treatment A versus placebo, 41.528 for Treatment B versus placebo, and 6.194 for female patients versus male patients. Note that these odds ratio estimates are not the same as the corresponding values in the Exp(Est) column in the parameter estimates table because effect coding was used. From Output 42.2.5, it is evident that both Treatment A and Treatment B are better than the placebo in reducing pain; females tend to have better improvement than males; and younger patients are faring better than older patients.

Output 42.2.5. Type 3 Effects and Parameter Estimates with Effect Coding

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Treatment	2	12.6928	0.0018
Sex	1	5.3013	0.0213
Age	1	7.6314	0.0057

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept		1	19.0804	6.7882	7.9007	0.0049	1.9343E8
Treatment	A	1	0.8772	0.5274	2.7662	0.0963	2.404
Treatment	B	1	1.4246	0.6036	5.5711	0.0183	4.156
Sex	F	1	0.9118	0.3960	5.3013	0.0213	2.489
Age		1	-0.2650	0.0959	7.6314	0.0057	0.767

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
Treatment	A vs P	24.022	3.295	175.121
Treatment	B vs P	41.528	4.500	383.262
Sex	F vs M	6.194	1.312	29.248
Age		0.767	0.636	0.926

Finally, PROC LOGISTIC is invoked to refit the previously selected model using reference coding for the CLASS variables. Two CONTRAST statements are specified. The one labeled 'Pairwise' specifies three rows in the contrast matrix, L, for all the pairwise comparisons between the three levels of Treatment. The contrast labeled 'Female vs Male' compares female to male patients. The option ESTIMATE=EXP is specified in both CONTRAST statements to exponentiate the estimates of $\mathbf{L}'\boldsymbol{\beta}$. With the given specification of contrast coefficients, the first row of the 'Pairwise' CONTRAST statement corresponds to the odds ratio of A versus P, the second row corresponds to B versus P, and the third row corresponds to A versus B. There is only one row in the 'Female vs Male' CONTRAST statement, and it corresponds to the odds ratio comparing female to male patients.

```
proc logistic data=Neuralgia;
   class Treatment Sex /param=ref;
   model Pain= Treatment Sex age;
   contrast 'Pairwise' Treatment 1  0,
                       Treatment 0  1,
                       Treatment 1 -1 / estimate=exp;
   contrast 'Female vs Male' Sex 1 / estimate=exp;
run;
```

Output 42.2.6. Reference Coding of CLASS Variables

The LOGISTIC Procedure

Class Level Information

Class	Value	Design Variables	
Treatment	A	1	0
	B	0	1
	P	0	0
Sex	F	1	
	M	0	

The reference coding is shown in Output 42.2.6. The Type 3 analysis of effects, the parameter estimates for the reference coding, and the odds ratio estimates are displayed in Output 42.2.7. Although the parameter estimates are different (because of the different parameterizations), the "Type 3 Analysis of Effects" table and the "Odds Ratio" table remain the same as in Output 42.2.5. With effect coding, the treatment A parameter estimate (0.8772) estimates the effect of treatment A compared to the average effect of treatments A, B, and placebo. The treatment A estimate (3.1790) under the reference coding estimates the difference in effect of treatment A and the placebo treatment.

Output 42.2.7. Type 3 Effects and Parameter Estimates with Reference Coding

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Treatment	2	12.6928	0.0018
Sex	1	5.3013	0.0213
Age	1	7.6314	0.0057

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	15.8669	6.4056	6.1357	0.0132
Treatment	A	1	3.1790	1.0135	9.8375	0.0017
Treatment	B	1	3.7264	1.1339	10.8006	0.0010
Sex	F	1	1.8235	0.7920	5.3013	0.0213
Age		1	-0.2650	0.0959	7.6314	0.0057

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Treatment A vs P	24.022	3.295	175.121
Treatment B vs P	41.528	4.500	383.262
Sex F vs M	6.194	1.312	29.248
Age	0.767	0.636	0.926

Output 42.2.8 contains two tables: the “Contrast Test Results” table and the “Contrast Rows Estimation and Testing Results” table. The former contains the overall Wald test for each CONTRAST statement. Although three rows are specified in the ’Pairwise’ CONTRAST statement, there are only two degrees of freedom, and the Wald test result is identical to the Type 3 analysis of Treatment in Output 42.2.7. The latter table contains estimates and tests of individual contrast rows. The estimates for the first two rows of the ’Pairwise’ CONTRAST statement are the same as those given in the “Odds Ratio Estimates” table (in Output 42.2.7). Both treatments A and B are highly effective over placebo in reducing pain. The third row estimates the odds ratio comparing A to B. The 95% confidence interval for this odds ratio is (0.0932, 3.5889), indicating that the pain reduction effects of these two test treatments are not that different. Again, the ’Female vs Male’ contrast shows that female patients fared better in obtaining relief from pain than male patients.

Output 42.2.8. Results of CONTRAST Statements

Contrast Test Results

Contrast	DF	Wald Chi-Square	Pr > ChiSq
Pairwise	2	12.6928	0.0018
Female vs Male	1	5.3013	0.0213

Contrast Rows Estimation and Testing Results

Contrast	Type	Row	Estimate	Standard Error	Alpha	Confidence Limits	
Pairwise	EXP	1	24.0218	24.3473	0.05	3.2951	175.1
Pairwise	EXP	2	41.5284	47.0877	0.05	4.4998	383.3
Pairwise	EXP	3	0.5784	0.5387	0.05	0.0932	3.5889
Female vs Male	EXP	1	6.1937	4.9053	0.05	1.3116	29.2476

Contrast Rows Estimation and Testing Results

Contrast	Type	Row	Wald Chi-Square	Pr > ChiSq
Pairwise	EXP	1	9.8375	0.0017
Pairwise	EXP	2	10.8006	0.0010
Pairwise	EXP	3	0.3455	0.5567
Female vs Male	EXP	1	5.3013	0.0213

Example 42.3. Ordinal Logistic Regression

Consider a study of the effects on taste of various cheese additives. Researchers tested four cheese additives and obtained 52 response ratings for each additive. Each response was measured on a scale of nine categories ranging from strong dislike (1) to excellent taste (9). The data, given in McCullagh and Nelder (1989, p. 175) in the form of a two-way frequency table of additive by rating, are saved in the data set Cheese.

```
data Cheese;
   do Additive = 1 to 4;
      do y = 1 to 9;
         input freq @@;
         output;
      end;
   end;
   label y='Taste Rating';
   datalines;
0  0  1  7  8  8 19  8  1
6  9 12 11  7  6  1  0  0
1  1  6  8 23  7  5  1  0
0  0  0  1  3  7 14 16 11
;
```

The data set Cheese contains the variables y, Additive, and freq. The variable y contains the response rating. The variable Additive specifies the cheese additive (1, 2, 3, or 4). The variable freq gives the frequency with which each additive received each rating.

The response variable y is ordinally scaled. A cumulative logit model is used to investigate the effects of the cheese additives on taste. The following SAS statements invoke PROC LOGISTIC to fit this model with y as the response variable and three indicator variables as explanatory variables, with the fourth additive as the reference level. With this parameterization, each Additive parameter compares an additive to the fourth additive. The COVB option produces the estimated covariance matrix.

```
proc logistic data=Cheese;
   freq freq;
   class Additive (param=ref ref='4');
   model y=Additive / covb;
   title1 'Multiple Response Cheese Tasting Experiment';
run;
```

Results of the analysis are shown in Output 42.3.1, and the estimated covariance matrix is displayed in Output 42.3.2.

Since the strong dislike (y=1) end of the rating scale is associated with lower Ordered Values in the Response Profile table, the probability of disliking the additives is modeled.

The score chi-square for testing the proportional odds assumption is 17.287, which is not significant with respect to a chi-square distribution with 21 degrees of freedom ($p = 0.694$). This indicates that the proportional odds model adequately fits the data. The positive value (1.6128) for the parameter estimate for Additive1 indicates a tendency towards the lower-numbered categories of the first cheese additive relative to the fourth. In other words, the fourth additive is better in taste than the first additive. Each of the second and the third additives is less favorable than the fourth additive. The relative magnitudes of these slope estimates imply the preference ordering: fourth, first, third, second.

Output 42.3.1. Proportional Odds Model Regression Analysis

Multiple Response Cheese Tasting Experiment

The LOGISTIC Procedure

Model Information

Data Set	WORK.CHEESE	
Response Variable	y	Taste Rating
Number of Response Levels	9	
Frequency Variable	freq	
Model	cumulative logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	36
Number of Observations Used	28
Sum of Frequencies Read	208
Sum of Frequencies Used	208

Response Profile

Ordered Value	y	Total Frequency
1	1	7
2	2	10
3	3	19
4	4	27
5	5	41
6	6	28
7	7	39
8	8	25
9	9	12

Probabilities modeled are cumulated over the lower Ordered Values.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Score Test for the Proportional Odds Assumption

Chi-Square	DF	Pr > ChiSq
17.2866	21	0.6936

Output 42.3.1. (continued)

Multiple Response Cheese Tasting Experiment

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	875.802	733.348
SC	902.502	770.061
-2 Log L	859.802	711.348

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	148.4539	3	<.0001
Score	111.2670	3	<.0001
Wald	115.1504	3	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept 1	1	-7.0801	0.5624	158.4851	<.0001
Intercept 2	1	-6.0249	0.4755	160.5500	<.0001
Intercept 3	1	-4.9254	0.4272	132.9484	<.0001
Intercept 4	1	-3.8568	0.3902	97.7087	<.0001
Intercept 5	1	-2.5205	0.3431	53.9704	<.0001
Intercept 6	1	-1.5685	0.3086	25.8374	<.0001
Intercept 7	1	-0.0669	0.2658	0.0633	0.8013
Intercept 8	1	1.4930	0.3310	20.3439	<.0001
Additive 1	1	1.6128	0.3778	18.2265	<.0001
Additive 2	1	4.9645	0.4741	109.6427	<.0001
Additive 3	1	3.3227	0.4251	61.0931	<.0001

Association of Predicted Probabilities and Observed Responses

Percent Concordant	67.6	Somers' D	0.578
Percent Discordant	9.8	Gamma	0.746
Percent Tied	22.6	Tau-a	0.500
Pairs	18635	c	0.789

Output 42.3.2. Estimated Covariance Matrix

Multiple Response Cheese Tasting Experiment

Estimated Covariance Matrix

Parameter	Intercept_ 1	Intercept_ 2	Intercept_ 3	Intercept_ 4	Intercept_ 5
Intercept_1	0.316291	0.219581	0.176278	0.147694	0.114024
Intercept_2	0.219581	0.226095	0.177806	0.147933	0.11403
Intercept_3	0.176278	0.177806	0.182473	0.148844	0.114092
Intercept_4	0.147694	0.147933	0.148844	0.152235	0.114512
Intercept_5	0.114024	0.11403	0.114092	0.114512	0.117713
Intercept_6	0.091085	0.091081	0.091074	0.091109	0.091821
Intercept_7	0.057814	0.057813	0.057807	0.05778	0.057721
Intercept_8	0.041304	0.041304	0.0413	0.041277	0.041162
Additive1	-0.09419	-0.09421	-0.09427	-0.09428	-0.09246
Additive2	-0.18686	-0.18161	-0.1687	-0.14717	-0.11415
Additive3	-0.13565	-0.13569	-0.1352	-0.13118	-0.11207

Estimated Covariance Matrix

Parameter	Intercept_ 6	Intercept_ 7	Intercept_ 8	Additive1	Additive2	Additive3
Intercept_1	0.091085	0.057814	0.041304	-0.09419	-0.18686	-0.13565
Intercept_2	0.091081	0.057813	0.041304	-0.09421	-0.18161	-0.13569
Intercept_3	0.091074	0.057807	0.0413	-0.09427	-0.1687	-0.1352
Intercept_4	0.091109	0.05778	0.041277	-0.09428	-0.14717	-0.13118
Intercept_5	0.091821	0.057721	0.041162	-0.09246	-0.11415	-0.11207
Intercept_6	0.09522	0.058312	0.041324	-0.08521	-0.09113	-0.09122
Intercept_7	0.058312	0.07064	0.04878	-0.06041	-0.05781	-0.05802
Intercept_8	0.041324	0.04878	0.109562	-0.04436	-0.0413	-0.04143
Additive1	-0.08521	-0.06041	-0.04436	0.142715	0.094072	0.092128
Additive2	-0.09113	-0.05781	-0.0413	0.094072	0.22479	0.132877
Additive3	-0.09122	-0.05802	-0.04143	0.092128	0.132877	0.180709

Example 42.4. Nominal Response Data: Generalized Logits Model

Over the course of one school year, third graders from three different schools are exposed to three different styles of mathematics instruction: a self-paced computer-learning style, a team approach, and a traditional class approach. The students are asked which style they prefer and their responses, classified by the type of program they are in (a regular school day versus a regular day supplemented with an afternoon school program) are displayed in Table 42.4. The data set is from Stokes, Davis, and Koch (2000), and is also analyzed in the "Generalized Logits Model" section on page 824 of Chapter 22, "The CATMOD Procedure."

Table 42.4. School Program Data

School	Program	Learning Style Preference		
		Self	Team	Class
1	Regular	10	17	26
1	Afternoon	5	12	50
2	Regular	21	17	26
2	Afternoon	16	12	36
3	Regular	15	15	16
3	Afternoon	12	12	20

The levels of the response variable (self, team, and class) have no essential ordering, so a logistic regression is performed on the generalized logits. The model to be fit is

$$\log\left(\frac{\pi_{hij}}{\pi_{hir}}\right) = \alpha_j + \mathbf{x}'_{hi}\boldsymbol{\beta}_j$$

where π_{hij} is the probability that a student in school h and program i prefers teaching style j, $j \neq r$, and style r is the baseline style (in this case, class). There are separate sets of intercept parameters α_j and regression parameters $\boldsymbol{\beta}_j$ for each logit, and the matrix $\mathbf{x}_{hi}$ is the set of explanatory variables for the hith population. Thus, two logits are modeled for each school and program combination: the logit comparing self to class and the logit comparing team to class.

The following statements create the data set school and request the analysis. The LINK=GLOGIT option forms the generalized logits. The response variable option ORDER=DATA means that the response variable levels are ordered as they exist in the data set: self, team, and class; thus, the logits are formed by comparing self to class and by comparing team to class. The ODS statement suppresses the display of the maximum likelihood estimates. The results of this analysis are shown in Output 42.4.1 through Output 42.4.4.

```
data school;
   length Program $ 9;
   input School Program $ Style $ Count @@;
   datalines;
1 regular   self 10  1 regular   team 17  1 regular   class 26
1 afternoon self  5  1 afternoon team 12  1 afternoon class 50
2 regular   self 21  2 regular   team 17  2 regular   class 26
2 afternoon self 16  2 afternoon team 12  2 afternoon class 36
3 regular   self 15  3 regular   team 15  3 regular   class 16
3 afternoon self 12  3 afternoon team 12  3 afternoon class 20
;


proc logistic data=school;
   freq Count;
   class School Program(ref=first);
   model Style(order=data)=School Program School*Program
         / link=glogit;
run;
```

Output 42.4.1. Analysis of Saturated Model

The LOGISTIC Procedure

Model Information

Data Set	WORK.SCHOOL
Response Variable	Style
Number of Response Levels	3
Frequency Variable	Count
Model	generalized logit
Optimization Technique	Fisher’s scoring

Number of Observations Read	18
Number of Observations Used	18
Sum of Frequencies Read	338
Sum of Frequencies Used	338

Response Profile

Ordered Value	Style	Total Frequency
1	self	79
2	team	85
3	class	174

Logits modeled use Style=’class’ as the reference category.

Class Level Information

Class	Value	Design Variables	
School	1	1	0
	2	0	1
	3	-1	-1
Program	afternoon	-1	
	regular	1	

Output 42.4.2. Fit Statistics

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	699.404	689.156
SC	707.050	735.033
-2 Log L	695.404	665.156

Output 42.4.3. Tests

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	30.2480	10	0.0008
Score	28.3738	10	0.0016
Wald	25.6828	10	0.0042

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
School	4	14.5522	0.0057
Program	2	10.4815	0.0053
School*Program	4	1.7439	0.7827

Output 42.4.4. Estimates

Analysis of Maximum Likelihood Estimates

Parameter			Style	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept			self	1	-0.8097	0.1488	29.5989	<.0001
Intercept			team	1	-0.6585	0.1366	23.2449	<.0001
School	1		self	1	-0.8194	0.2281	12.9066	0.0003
School	1		team	1	-0.2675	0.1881	2.0233	0.1549
School	2		self	1	0.2974	0.1919	2.4007	0.1213
School	2		team	1	-0.1033	0.1898	0.2961	0.5863
Program	regular		self	1	0.3985	0.1488	7.1684	0.0074
Program	regular		team	1	0.3537	0.1366	6.7071	0.0096
School*Program	1	regular	self	1	0.2751	0.2281	1.4547	0.2278
School*Program	1	regular	team	1	0.1474	0.1881	0.6143	0.4332
School*Program	2	regular	self	1	-0.0998	0.1919	0.2702	0.6032
School*Program	2	regular	team	1	-0.0168	0.1898	0.0079	0.9293

The "Type 3 Analysis of Effects" table in Output 42.4.3 shows that the interaction effect is clearly nonsignificant, so a main effects model is fit with the following statements.

```
proc logistic data=school;
   freq Count;
   class School Program(ref=first);
   model Style(order=data)=School Program / link=glogit;
run;
```

Output 42.4.5. Analysis of Main Effects Model

The LOGISTIC Procedure

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	699.404	682.934
SC	707.050	713.518
-2 Log L	695.404	666.934

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	28.4704	6	<.0001
Score	27.1190	6	0.0001
Wald	25.5881	6	0.0003

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
School	4	14.8424	0.0050
Program	2	10.9160	0.0043

All of the global fit tests in Output 42.4.5 suggest the model is significant, and the Type 3 tests show that the school and program effects are also significant.

Output 42.4.6. Estimates

Analysis of Maximum Likelihood Estimates

Parameter		Style	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		self	1	-0.7978	0.1465	29.6502	<.0001
Intercept		team	1	-0.6589	0.1367	23.2300	<.0001
School	1	self	1	-0.7992	0.2198	13.2241	0.0003
School	1	team	1	-0.2786	0.1867	2.2269	0.1356
School	2	self	1	0.2836	0.1899	2.2316	0.1352
School	2	team	1	-0.0985	0.1892	0.2708	0.6028
Program	regular	self	1	0.3737	0.1410	7.0272	0.0080
Program	regular	team	1	0.3713	0.1353	7.5332	0.0061

Odds Ratio Estimates

Effect			Style	Point Estimate	95% Wald Confidence Limits	
School	1 vs 3		self	0.269	0.127	0.570
School	1 vs 3		team	0.519	0.267	1.010
School	2 vs 3		self	0.793	0.413	1.522
School	2 vs 3		team	0.622	0.317	1.219
Program	regular	vs afternoon	self	2.112	1.215	3.670
Program	regular	vs afternoon	team	2.101	1.237	3.571

The parameter estimates, tests for individual parameters, and odds ratios are displayed in Output 42.4.6. The Program variable has nearly the same effect on both logits, while School=1 has the largest effect of the schools.

Example 42.5. Stratified Sampling

Consider the hypothetical example in Fleiss (1981, pp. 6–7) in which a test is applied to a sample of 1,000 people known to have a disease and to another sample of 1,000 people known not to have the same disease. In the diseased sample, 950 test positive; in the nondiseased sample, only 10 test positive. If the true disease rate in the population is 1 in 100, specifying PEVENT=0.01 results in the correct false positive and negative rates for the stratified sampling scheme. Omitting the PEVENT= option is equivalent to using the overall sample disease rate (1000/2000 = 0.5) as the value of the PEVENT= option, which would ignore the stratified sampling.

The SAS code is as follows:

```
data Screen;
   do Disease='Present','Absent';
      do Test=1,0;
         input Count @@;
         output;
      end;
   end;
   datalines;
950  50
 10 990
;
```

```
proc logistic data=Screen;
   freq Count;
   model Disease(event='Present')=Test
         / pevent=.5 .01 ctable pprob=.5;
run;
```

The response variable option EVENT= indicates that Disease='Present' is the event. The CTABLE option is specified to produce a classification table. Specifying PPROB=0.5 indicates a cutoff probability of 0.5. A list of two probabilities, 0.5 and 0.01, is specified for the PEVENT= option; 0.5 corresponds to the overall sample disease rate, and 0.01 corresponds to a true disease rate of 1 in 100.

The classification table is shown in Output 42.5.1.

Output 42.5.1. False Positive and False Negative Rates

The LOGISTIC Procedure

Classification Table

		Correct		Incorrect		Percentages				
Prob Event	Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.500	0.500	950	990	10	50	97.0	95.0	99.0	1.0	4.8
0.010	0.500	950	990	10	50	99.0	95.0	99.0	51.0	0.1

In the classification table, the column "Prob Level" represents the cutoff values (the settings of the PPROB= option) for predicting whether an observation is an event. The "Correct" columns list the numbers of subjects that are correctly predicted as events and nonevents, respectively, and the "Incorrect" columns list the number of nonevents incorrectly predicted as events and the number of events incorrectly predicted as nonevents, respectively. For PEVENT=0.5, the false positive rate is 1% and the false negative rate is 4.8%. These results ignore the fact that the samples were stratified and incorrectly assume that the overall sample proportion of disease (which is 0.5) estimates the true disease rate. For a true disease rate of 0.01, the false positive rate and the false negative rate are 51% and 0.1%, respectively, as shown on the second line of the classification table.

Example 42.6. Logistic Regression Diagnostics

In a controlled experiment to study the effect of the rate and volume of air inspired on a transient reflex vaso-constriction in the skin of the digits, 39 tests under various combinations of rate and volume of air inspired were obtained (Finney 1947). The end point of each test is whether or not vaso-constriction occurred. Pregibon (1981) uses this set of data to illustrate the diagnostic measures he proposes for detecting influential observations and to quantify their effects on various aspects of the maximum likelihood fit.

The vaso-constriction data are saved in the data set vaso:

```
data vaso;
   length Response $12;
   input Volume Rate Response @@;
   LogVolume=log(Volume);
   LogRate=log(Rate);
   datalines;
3.70  0.825  constrict       3.50  1.09   constrict
1.25  2.50   constrict       0.75  1.50   constrict
0.80  3.20   constrict       0.70  3.50   constrict
0.60  0.75   no_constrict    1.10  1.70   no_constrict
0.90  0.75   no_constrict    0.90  0.45   no_constrict
0.80  0.57   no_constrict    0.55  2.75   no_constrict
0.60  3.00   no_constrict    1.40  2.33   constrict
0.75  3.75   constrict       2.30  1.64   constrict
3.20  1.60   constrict       0.85  1.415  constrict
1.70  1.06   no_constrict    1.80  1.80   constrict
0.40  2.00   no_constrict    0.95  1.36   no_constrict
1.35  1.35   no_constrict    1.50  1.36   no_constrict
1.60  1.78   constrict       0.60  1.50   no_constrict
1.80  1.50   constrict       0.95  1.90   no_constrict
1.90  0.95   constrict       1.60  0.40   no_constrict
2.70  0.75   constrict       2.35  0.03   no_constrict
1.10  1.83   no_constrict    1.10  2.20   constrict
1.20  2.00   constrict       0.80  3.33   constrict
0.95  1.90   no_constrict    0.75  1.90   no_constrict
1.30  1.625  constrict
;
```

In the data set vaso, the variable Response represents the outcome of a test. The variable LogVolume represents the log of the volume of air intake, and the variable LogRate represents the log of the rate of air intake.

The following SAS statements invoke PROC LOGISTIC to fit a logistic regression model to the vaso-constriction data, where Response is the response variable, and LogRate and LogVolume are the explanatory variables. The INFLUENCE option and the IPLOTS option are specified to display the regression diagnostics and the index plots.

```
ods html;
ods graphics on;

title 'Occurrence of Vaso-Constriction';
proc logistic data=vaso;
   model Response=LogRate LogVolume/influence iplots;
run;

ods graphics off;
ods html close;
```

Results of the model fit are shown in Output 42.6.1. Both LogRate and LogVolume are statistically significant to the occurrence of vaso-constriction ($p = 0.0131$ and $p = 0.0055$, respectively). Their positive parameter estimates indicate that a higher

inspiration rate or a larger volume of air intake is likely to increase the probability of vaso-constriction.

Output 42.6.1. Logistic Regression Analysis for Vaso-Constriction Data

```
                Occurrence of Vaso-Constriction

                     The LOGISTIC Procedure

                       Model Information

Data Set                          WORK.VASO
Response Variable                 Response
Number of Response Levels         2
Model                             binary logit
Optimization Technique            Fisher's scoring


   Number of Observations Read          39
   Number of Observations Used          39


                     Response Profile

    Ordered                          Total
      Value     Response         Frequency

          1     constrict               20
          2     no_constrict            19

 Probability modeled is Response='constrict'.


                  Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.
```

Output 42.6.1. (continued)

Occurrence of Vaso-Constriction

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	56.040	35.227
SC	57.703	40.218
-2 Log L	54.040	29.227

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	24.8125	2	<.0001
Score	16.6324	2	0.0002
Wald	7.8876	2	0.0194

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.8754	1.3208	4.7395	0.0295
LogRate	1	4.5617	1.8380	6.1597	0.0131
LogVolume	1	5.1793	1.8648	7.7136	0.0055

Association of Predicted Probabilities and Observed Responses

Percent Concordant	93.7	Somers' D	0.874
Percent Discordant	6.3	Gamma	0.874
Percent Tied	0.0	Tau-a	0.448
Pairs	380	c	0.937

The INFLUENCE option displays the values of the explanatory variables (LogRate and LogVolume) for each observation, a column for each diagnostic produced, and the *case number* which represents the sequence number of the observation (Output 42.6.2). Also produced (but not shown here) is a lineprinter plot where the vertical axis represents the case number and the horizontal axis represents the value of the diagnostic statistic.

The index plots produced by the IPLOTS option are essentially the same lineprinter plots as those produced by the INFLUENCE option with a 90-degree rotation and perhaps on a more refined scale. This version of the plots are not displayed here. The vertical axis of an index plot represents the value of the diagnostic and the horizontal axis represents the sequence (case number) of the observation. The index plots are useful for identification of extreme values.

Since the experimental ODS GRAPHICS statement is also specified, the lineprinter plots from the INFLUENCE and IPLOTS options are suppressed and graphical displays are produced as shown in Output 42.6.3 through Output 42.6.5. For general

information about ODS graphics, see Chapter 15, "Statistical Graphics Using ODS." For specific information about the graphics available in the LOGISTIC procedure, see the "ODS Graphics" section on page 2388.

Output 42.6.2. Regression Diagnostics from the INFLUENCE Option (Experimental)

The LOGISTIC Procedure

Regression Diagnostics

Case Number	Covariates LogRate	Log Volume	Pearson Residual	Deviance Residual	Hat Matrix Diagonal	Intercept DfBeta	LogRate DfBeta
1	-0.1924	1.3083	0.2205	0.3082	0.0927	-0.0165	0.0193
2	0.0862	1.2528	0.1349	0.1899	0.0429	-0.0134	0.0151
3	0.9163	0.2231	0.2923	0.4049	0.0612	-0.0492	0.0660
4	0.4055	-0.2877	3.5181	2.2775	0.0867	1.0734	-0.9302
5	1.1632	-0.2231	0.5287	0.7021	0.1158	-0.0832	0.1411
6	1.2528	-0.3567	0.6090	0.7943	0.1524	-0.0922	0.1710
7	-0.2877	-0.5108	-0.0328	-0.0464	0.00761	-0.00280	0.00274
8	0.5306	0.0953	-1.0196	-1.1939	0.0559	-0.1444	0.0613
9	-0.2877	-0.1054	-0.0938	-0.1323	0.0342	-0.0178	0.0173
10	-0.7985	-0.1054	-0.0293	-0.0414	0.00721	-0.00245	0.00246
11	-0.5621	-0.2231	-0.0370	-0.0523	0.00969	-0.00361	0.00358
12	1.0116	-0.5978	-0.5073	-0.6768	0.1481	-0.1173	0.0647
13	1.0986	-0.5108	-0.7751	-0.9700	0.1628	-0.0931	-0.00946
14	0.8459	0.3365	0.2559	0.3562	0.0551	-0.0414	0.0538
15	1.3218	-0.2877	0.4352	0.5890	0.1336	-0.0940	0.1408
16	0.4947	0.8329	0.1576	0.2215	0.0402	-0.0198	0.0234
17	0.4700	1.1632	0.0709	0.1001	0.0172	-0.00630	0.00701
18	0.3471	-0.1625	2.9062	2.1192	0.0954	0.9595	-0.8279
19	0.0583	0.5306	-1.0718	-1.2368	0.1315	-0.2591	0.2024
20	0.5878	0.5878	0.2405	0.3353	0.0525	-0.0331	0.0421
21	0.6931	-0.9163	-0.1076	-0.1517	0.0373	-0.0180	0.0158
22	0.3075	-0.0513	-0.4193	-0.5691	0.1015	-0.1449	0.1237
23	0.3001	0.3001	-1.0242	-1.1978	0.0761	-0.1961	0.1275
24	0.3075	0.4055	-1.3684	-1.4527	0.0717	-0.1281	0.0410
25	0.5766	0.4700	0.3347	0.4608	0.0587	-0.0403	0.0570
26	0.4055	-0.5108	-0.1595	-0.2241	0.0548	-0.0366	0.0329
27	0.4055	0.5878	0.3645	0.4995	0.0661	-0.0327	0.0496
28	0.6419	-0.0513	-0.8989	-1.0883	0.0647	-0.1423	0.0617
29	-0.0513	0.6419	0.8981	1.0876	0.1682	0.2367	-0.1950
30	-0.9163	0.4700	-0.0992	-0.1400	0.0507	-0.0224	0.0227
31	-0.2877	0.9933	0.6198	0.8064	0.2459	0.1165	-0.0996
32	-3.5066	0.8544	-0.00073	-0.00103	0.000022	-3.22E-6	3.405E-6
33	0.6043	0.0953	-1.2062	-1.3402	0.0510	-0.0882	-0.0137
34	0.7885	0.0953	0.5447	0.7209	0.0601	-0.0425	0.0877
35	0.6931	0.1823	0.5404	0.7159	0.0552	-0.0340	0.0755
36	1.2030	-0.2231	0.4828	0.6473	0.1177	-0.0867	0.1381
37	0.6419	-0.0513	-0.8989	-1.0883	0.0647	-0.1423	0.0617
38	0.6419	-0.2877	-0.4874	-0.6529	0.1000	-0.1395	0.1032
39	0.4855	0.2624	0.7053	0.8987	0.0531	0.0326	0.0190

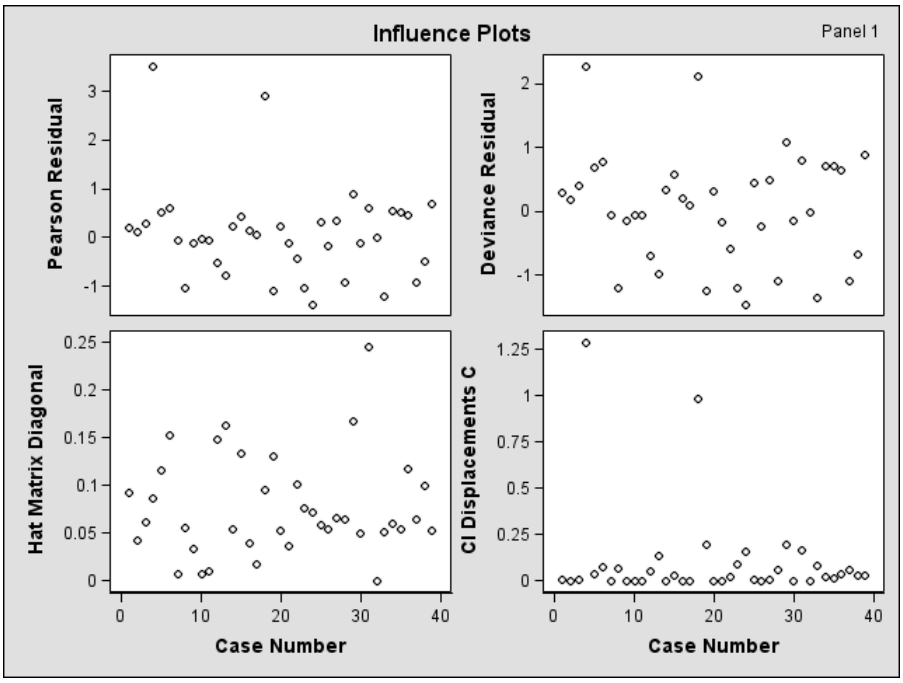
Output 42.6.2. (continued)

The LOGISTIC Procedure

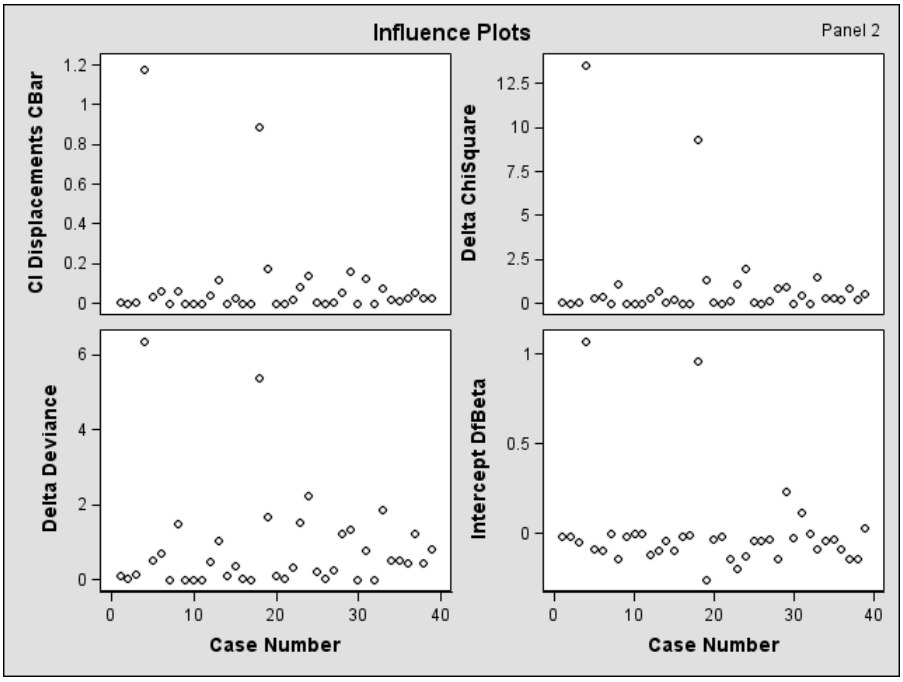
Regression Diagnostics

Case Number	Log Volume DfBeta	Confidence Interval Displacement C	Confidence Interval Displacement CBar	Delta Deviance	Delta Chi-Square
1	0.0556	0.00548	0.00497	0.1000	0.0536
2	0.0261	0.000853	0.000816	0.0369	0.0190
3	0.0589	0.00593	0.00557	0.1695	0.0910
4	-1.0180	1.2873	1.1756	6.3626	13.5523
5	0.0583	0.0414	0.0366	0.5296	0.3161
6	0.0381	0.0787	0.0667	0.6976	0.4376
7	0.00265	8.321E-6	8.258E-6	0.00216	0.00109
8	0.0570	0.0652	0.0616	1.4870	1.1011
9	0.0153	0.000322	0.000311	0.0178	0.00911
10	0.00211	6.256E-6	6.211E-6	0.00172	0.000862
11	0.00319	0.000014	0.000013	0.00274	0.00138
12	0.1651	0.0525	0.0447	0.5028	0.3021
13	0.1775	0.1395	0.1168	1.0577	0.7175
14	0.0527	0.00404	0.00382	0.1307	0.0693
15	0.0643	0.0337	0.0292	0.3761	0.2186
16	0.0307	0.00108	0.00104	0.0501	0.0259
17	0.00914	0.000089	0.000088	0.0101	0.00511
18	-0.8477	0.9845	0.8906	5.3817	9.3363
19	-0.00488	0.2003	0.1740	1.7037	1.3227
20	0.0518	0.00338	0.00320	0.1156	0.0610
21	0.0208	0.000465	0.000448	0.0235	0.0120
22	0.1179	0.0221	0.0199	0.3437	0.1956
23	0.0357	0.0935	0.0864	1.5212	1.1355
24	-0.1004	0.1558	0.1447	2.2550	2.0171
25	0.0708	0.00741	0.00698	0.2193	0.1190
26	0.0373	0.00156	0.00147	0.0517	0.0269
27	0.0788	0.0101	0.00941	0.2589	0.1423
28	0.1025	0.0597	0.0559	1.2404	0.8639
29	0.0286	0.1961	0.1631	1.3460	0.9697
30	0.0159	0.000554	0.000526	0.0201	0.0104
31	0.1322	0.1661	0.1253	0.7755	0.5095
32	2.48E-6	1.18E-11	1.18E-11	1.065E-6	5.324E-7
33	-0.00216	0.0824	0.0782	1.8744	1.5331
34	0.0671	0.0202	0.0190	0.5387	0.3157
35	0.0711	0.0180	0.0170	0.5295	0.3091
36	0.0631	0.0352	0.0311	0.4501	0.2641
37	0.1025	0.0597	0.0559	1.2404	0.8639
38	0.1397	0.0293	0.0264	0.4526	0.2639
39	0.0489	0.0295	0.0279	0.8355	0.5254

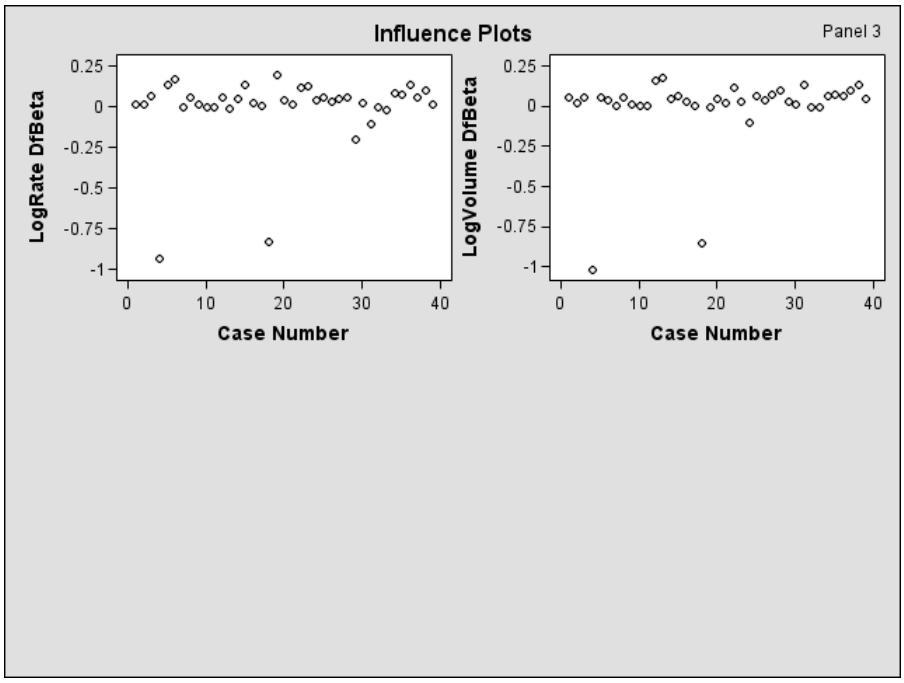
Output 42.6.3. Residuals, Hat Matrix, and CI Displacement C (Experimental)



Output 42.6.4. CI Displacement CBar, Change in Deviance and Pearson χ^2, and DFBETAS for the Intercept (Experimental)



Output 42.6.5. DFBETAS for LogRate and LogVolume (Experimental)



The index plots of the Pearson residuals and the deviance residuals (Output 42.6.3) indicate that case 4 and case 18 are poorly accounted for by the model. The index plot of the diagonal elements of the hat matrix (Output 42.6.3) suggests that case 31 is an extreme point in the design space. The index plots of DFBETAS (Output 42.6.4 and Output 42.6.5) indicate that case 4 and case 18 are causing instability in all three parameter estimates. The other four index plots in Output 42.6.3 and Output 42.6.4 also point to these two cases as having a large impact on the coefficients and goodness of fit.

Example 42.7. ROC Curve, Customized Odds Ratios, Goodness-of-Fit Statistics, R-Square, and Confidence Limits

This example plots an ROC curve, estimates a customized odds ratio, produces the traditional goodness-of-fit analysis, displays the generalized R^2 measures for the fitted model, calculates the normal confidence intervals for the regression parameters, and produces an experimental display of the probability function and prediction curves for the fitted model. The data consist of three variables: n (number of subjects in a sample), disease (number of diseased subjects in the sample), and age (age for the sample). A linear logistic regression model is used to study the effect of age on the probability of contracting the disease.

The SAS statements are as follows:

```
data Data1;
   input disease n age;
   datalines;
 0 14 25
 0 20 35
 0 19 45
 7 18 55
 6 12 65
17 17 75
;


ods html;
ods graphics on;

proc logistic data=Data1;
   model disease/n=age / scale=none
                         clparm=wald
                         clodds=pl
                         rsquare
                         outroc=roc1;
   units age=10;
run;

ods graphics off;
ods html close;
```

The option SCALE=NONE is specified to produce the deviance and Pearson goodness-of-fit analysis without adjusting for overdispersion. The RSQUARE option is specified to produce generalized R^2 measures of the fitted model. The CLPARM=WALD option is specified to produce the Wald confidence intervals for the regression parameters. The UNITS statement is specified to produce customized odds ratio estimates for a change of 10 years in the age variable, and the CLODDS=PL option is specified to produce profile likelihood confidence limits for the odds ratio. The OUTROC= option outputs the data for the ROC curve to the SAS data set, roc1.

Results are shown in Output 42.7.1 and Output 42.7.2.

Output 42.7.1. Deviance and Pearson Goodness-of-Fit Analysis

The LOGISTIC Procedure

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	7.7756	4	1.9439	0.1002
Pearson	6.6020	4	1.6505	0.1585

Number of events/trials observations: 6

Output 42.7.2. R-Square, Confidence Intervals, and Customized Odds Ratio

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	124.173	52.468
SC	126.778	57.678
-2 Log L	122.173	48.468

R-Square 0.5215 Max-rescaled R-Square 0.7394

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	73.7048	1	<.0001
Score	55.3274	1	<.0001
Wald	23.3475	1	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-12.5016	2.5555	23.9317	<.0001
age	1	0.2066	0.0428	23.3475	<.0001

Association of Predicted Probabilities and Observed Responses

Percent Concordant	92.6	Somers' D	0.906
Percent Discordant	2.0	Gamma	0.958
Percent Tied	5.4	Tau-a	0.384
Pairs	2100	c	0.953

Wald Confidence Interval for Parameters

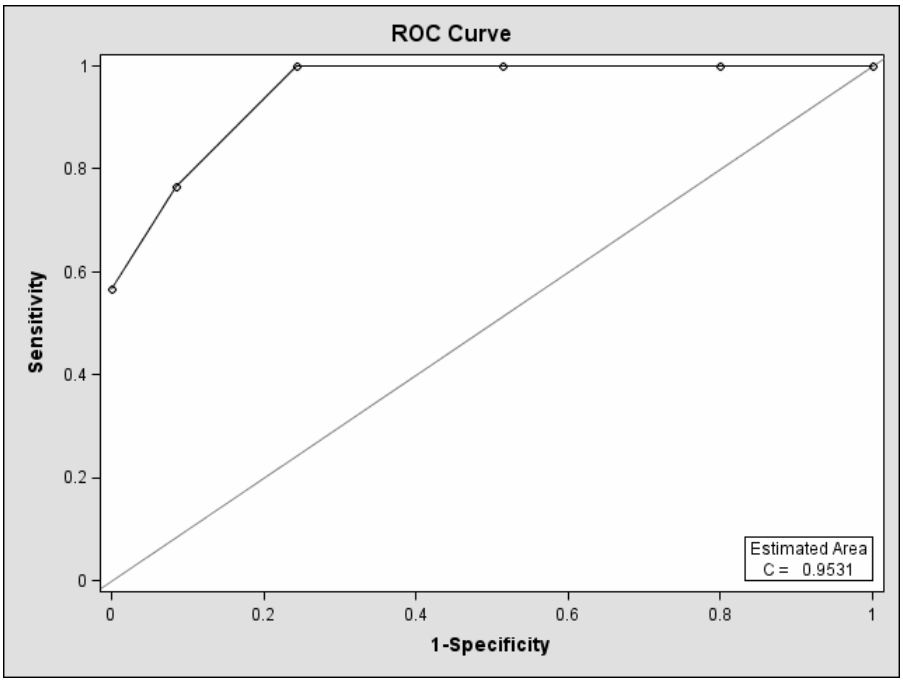
Parameter	Estimate	95% Confidence Limits	
Intercept	-12.5016	-17.5104	-7.4929
age	0.2066	0.1228	0.2904

Profile Likelihood Confidence Interval for Adjusted Odds Ratios

Effect	Unit	Estimate	95% Confidence Limits	
age	10.0000	7.892	3.881	21.406

Since the experimental ODS GRAPHICS statement is specified, a graphical display of the ROC curve is produced as shown in Output 42.7.3. For general information about ODS graphics, see Chapter 15, "Statistical Graphics Using ODS." For specific information about the graphics available in the LOGISTIC procedure, see the "ODS Graphics" section on page 2388.

Output 42.7.3. Receiver Operating Characteristic Curve (Experimental)



Note that the area under the ROC curve is given by the statistic *c* in the "Association of Predicted Probabilities and Observed Responses" table. In this example, the area under the ROC curve is 0.953.

The ROC curve may also be displayed with the GPLOT procedure by using the following code.

```
symbol1 i=join v=none c=black;
proc gplot data=roc1;
   title 'ROC Curve';
   plot _sensit_*_1mspec_=1 / vaxis=0 to 1 by .1 cframe=white;
run;
```

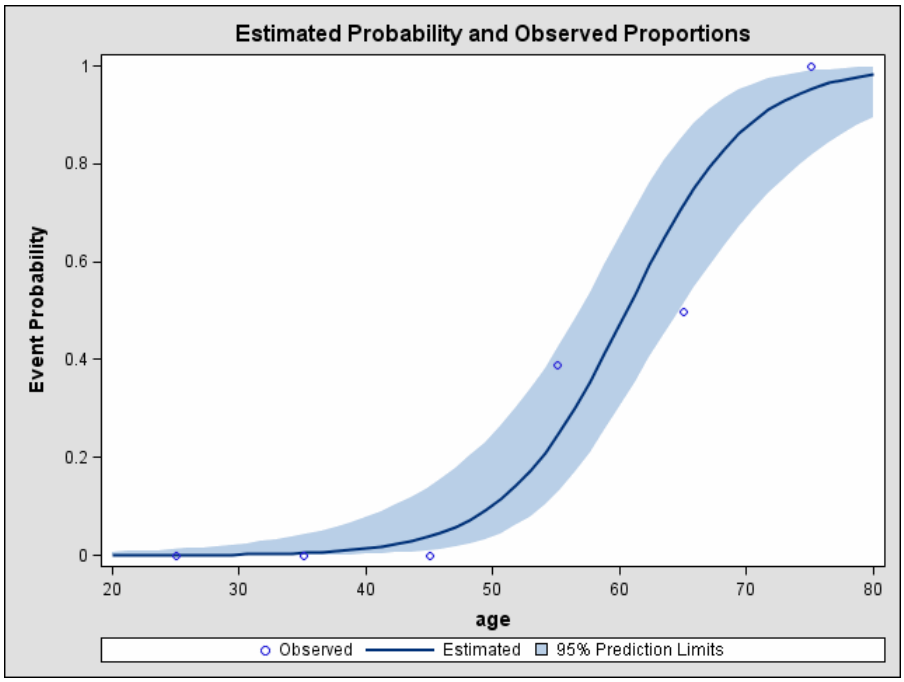
Because there is only one continuous covariate, if the experimental ODS GRAPHICS statement and the experimental GRAPHICS option ESTPROB are specified, then a graphical display of the estimated probability curve with bounding 95% prediction limits is displayed as shown in Output 42.7.4.

```
ods html;
ods graphics on;

proc logistic data=Data1;
   model disease/n=age / scale=none
                         clparm=wald
                         clodds=pl
                         rsquare
                         outroc=roc1;
   units age=10;
   graphics estprob;
run;

ods graphics off;
ods html close;
```

Output 42.7.4. Estimated Probability and 95% Prediction Limits (Experimental)



Example 42.8. Goodness-of-Fit Tests and Subpopulations

A study is done to investigate the effects of two binary factors, A and B, on a binary response, Y. Subjects are randomly selected from subpopulations defined by the four possible combinations of levels of A and B. The number of subjects responding with each level of Y is recorded and entered into data set A.

```
data a;
   do A=0,1;
      do B=0,1;
         do Y=1,2;
            input F @@;
            output;
         end;
      end;
   end;
   datalines;
23 63 31 70 67 100 70 104
;
```

A full model is fit to examine the main effects of A and B as well as the interaction effect of A and B.

```
proc logistic data=a;
   freq F;
   model Y=A B A*B;
run;
```

Output 42.8.1. Full Model Fit

```
                    The LOGISTIC Procedure

                      Model Information

          Data Set                      WORK.A
          Response Variable             Y
          Number of Response Levels     2
          Frequency Variable            F
          Model                         binary logit
          Optimization Technique        Fisher's scoring


             Number of Observations Read           8
             Number of Observations Used           8
             Sum of Frequencies Read             528
             Sum of Frequencies Used             528


                      Response Profile

            Ordered                       Total
              Value            Y     Frequency

                1              1           191
                2              2           337

             Probability modeled is Y=1.


                  Model Convergence Status

          Convergence criterion (GCONV=1E-8) satisfied.


                  Model Fit Statistics

                                       Intercept
                          Intercept            and
          Criterion            Only     Covariates

          AIC               693.061        691.914
          SC                697.330        708.990
          -2 Log L          691.061        683.914


             Testing Global Null Hypothesis: BETA=0

     Test                 Chi-Square       DF     Pr > ChiSq

     Likelihood Ratio         7.1478        3         0.0673
     Score                    6.9921        3         0.0721
     Wald                     6.9118        3         0.0748
```

Output 42.8.1. (continued)

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.0074	0.2436	17.1015	<.0001
A	1	0.6069	0.2903	4.3714	0.0365
B	1	0.1929	0.3254	0.3515	0.5533
A*B	1	-0.1883	0.3933	0.2293	0.6321

Association of Predicted Probabilities and Observed Responses

Percent Concordant	42.2	Somers' D	0.118
Percent Discordant	30.4	Gamma	0.162
Percent Tied	27.3	Tau-a	0.054
Pairs	64367	c	0.559

Pearson and Deviance goodness-of-fit tests cannot be obtained for this model since a full model containing four parameters is fit, leaving no residual degrees of freedom. For a binary response model, the goodness-of-fit tests have $m-q$ degrees of freedom, where m is the number of subpopulations and q is the number of model parameters. In the preceding model, $m = q = 4$, resulting in zero degrees of freedom for the tests.

Results of the model fit are shown in Output 42.8.1. Notice that neither the A*B interaction nor the B main effect is significant. If a reduced model containing only the A effect is fit, two degrees of freedom become available for testing goodness of fit. Specifying the SCALE=NONE option requests the Pearson and deviance statistics. With *single-trial* syntax, the AGGREGATE= option is needed to define the subpopulations in the study. Specifying AGGREGATE=(A B) creates subpopulations of the four combinations of levels of A and B. Although the B effect is being dropped from the model, it is still needed to define the original subpopulations in the study. If AGGREGATE=(A) were specified, only two subpopulations would be created from the levels of A, resulting in $m = q = 2$ and zero degrees of freedom for the tests.

```
proc logistic data=a;
   freq F;
   model Y=A / scale=none aggregate=(A B);
run;
```

Output 42.8.2. Reduced Model Fit

The LOGISTIC Procedure

Model Information

Data Set	WORK.A
Response Variable	Y
Number of Response Levels	2
Frequency Variable	F
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	8
Number of Observations Used	8
Sum of Frequencies Read	528
Sum of Frequencies Used	528

Response Profile

Ordered Value	Y	Total Frequency
1	1	191
2	2	337

Probability modeled is Y=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	0.3541	2	0.1770	0.8377
Pearson	0.3531	2	0.1765	0.8382

Number of unique profiles: 4

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	693.061	688.268
SC	697.330	696.806
-2 Log L	691.061	684.268

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	6.7937	1	0.0091
Score	6.6779	1	0.0098
Wald	6.6210	1	0.0101

Output 42.8.2. (continued)

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.9013	0.1614	31.2001	<.0001
A	1	0.5032	0.1955	6.6210	0.0101

Association of Predicted Probabilities and Observed Responses

Percent Concordant	28.3	Somers' D	0.112
Percent Discordant	17.1	Gamma	0.246
Percent Tied	54.6	Tau-a	0.052
Pairs	64367	c	0.556

The goodness-of-fit tests (Output 42.8.2) show that dropping the B main effect and the A*B interaction simultaneously does not result in significant lack of fit of the model. The tests' large p-values indicate insufficient evidence for rejecting the null hypothesis that the model fits.

Example 42.9. Overdispersion

In a seed germination test, seeds of two cultivars were planted in pots of two soil conditions. The following SAS statements create the data set seeds, which contains the observed proportion of seeds that germinated for various combinations of cultivar and soil condition. Variable n represents the number of seeds planted in a pot, and variable r represents the number germinated. The indicator variables cult and soil represent the cultivar and soil condition, respectively.

```
data seeds;
   input pot n r cult soil;
   datalines;
 1 16     8      0       0
 2 51    26      0       0
 3 45    23      0       0
 4 39    10      0       0
 5 36     9      0       0
 6 81    23      1       0
 7 30    10      1       0
 8 39    17      1       0
 9 28     8      1       0
10 62    23      1       0
11 51    32      0       1
12 72    55      0       1
13 41    22      0       1
14 12     3      0       1
15 13    10      0       1
16 79    46      1       1
17 30    15      1       1
```

```
18 51    32      1       1
19 74    53      1       1
20 56    12      1       1
;
```

PROC LOGISTIC is used to fit a logit model to the data, with cult, soil, and cult × soil interaction as explanatory variables. The option SCALE=NONE is specified to display goodness-of-fit statistics.

```
proc logistic data=seeds;
   model r/n=cult soil cult*soil/scale=none;
   title 'Full Model With SCALE=NONE';
run;
```

Output 42.9.1. Results of the Model Fit for the Two-Way Layout

```
                    Full Model With SCALE=NONE

                      The LOGISTIC Procedure

          Deviance and Pearson Goodness-of-Fit Statistics

   Criterion          Value       DF     Value/DF     Pr > ChiSq

   Deviance         68.3465       16       4.2717         <.0001
   Pearson          66.7617       16       4.1726         <.0001

              Number of events/trials observations: 20


                       Model Fit Statistics

                                          Intercept
                              Intercept           and
                Criterion          Only    Covariates

                AIC            1256.852      1213.003
                SC             1261.661      1232.240
                -2 Log L       1254.852      1205.003


              Testing Global Null Hypothesis: BETA=0

      Test                 Chi-Square       DF     Pr > ChiSq

      Likelihood Ratio        49.8488        3         <.0001
      Score                   49.1682        3         <.0001
      Wald                    47.7623        3         <.0001


             Analysis of Maximum Likelihood Estimates

                                Standard          Wald
Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept     1     -0.3788      0.1489        6.4730        0.0110
cult          1     -0.2956      0.2020        2.1412        0.1434
soil          1      0.9781      0.2128       21.1234        <.0001
cult*soil     1     -0.1239      0.2790        0.1973        0.6569
```

Results of fitting the full factorial model are shown in Output 42.9.1. Both Pearson χ^2 and deviance are highly significant ($p < 0.0001$), suggesting that the model does not fit well. If the link function and the model specification are correct and if there are no outliers, then the lack of fit may be due to overdispersion. Without adjusting for the overdispersion, the standard errors are likely to be underestimated, causing the Wald tests to be too sensitive. In PROC LOGISTIC, there are three SCALE= options to accommodate overdispersion. With unequal sample sizes for the observations, SCALE=WILLIAMS is preferred. The Williams model estimates a scale parameter ϕ by equating the value of Pearson χ^2 for the full model to its approximate expected value. The full model considered here is the model with cultivar, soil condition, and their interaction. Using a full model reduces the risk of contaminating ϕ with lack of fit due to incorrect model specification.

```
proc logistic data=seeds;
   model r/n=cult soil cult*soil / scale=williams;
   title 'Full Model With SCALE=WILLIAMS';
run;
```

Output 42.9.2. Williams' Model for Overdispersion

Full Model With SCALE=WILLIAMS

The LOGISTIC Procedure

Model Information

Data Set	WORK.SEEDS
Response Variable (Events)	r
Response Variable (Trials)	n
Weight Variable	1 / (1 + 0.075941 * (n - 1))
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	20
Number of Observations Used	20
Sum of Frequencies Read	906
Sum of Frequencies Used	906
Sum of Weights Read	198.3216
Sum of Weights Used	198.3216

Response Profile

Ordered Value	Binary Outcome	Total Frequency	Total Weight
1	Event	437	92.95346
2	Nonevent	469	105.36819

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	16.4402	16	1.0275	0.4227
Pearson	16.0000	16	1.0000	0.4530

Number of events/trials observations: 20

NOTE: Since the Williams method was used to accommodate overdispersion, the Pearson chi-squared statistic and the deviance can no longer be used to assess the goodness of fit of the model.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	276.155	273.586
SC	280.964	292.822
-2 Log L	274.155	265.586

Output 42.9.2. (continued)

Full Model With SCALE=WILLIAMS

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	8.5687	3	0.0356
Score	8.4856	3	0.0370
Wald	8.3069	3	0.0401

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.3926	0.2932	1.7932	0.1805
cult	1	-0.2618	0.4160	0.3963	0.5290
soil	1	0.8309	0.4223	3.8704	0.0491
cult*soil	1	-0.0532	0.5835	0.0083	0.9274

Association of Predicted Probabilities and Observed Responses

Percent Concordant	50.6	Somers' D	0.258
Percent Discordant	24.8	Gamma	0.343
Percent Tied	24.6	Tau-a	0.129
Pairs	204953	c	0.629

Results using Williams' method are shown in Output 42.9.2. The estimate of ϕ is 0.075941 and is given in the formula for the Weight Variable at the beginning of the displayed output. Since neither cult nor cult $\times$ soil is statistically significant ($p = 0.5290$ and $p = 0.9274$, respectively), a reduced model that contains only the soil condition factor is fitted, with the observations weighted by $1/(1+0.075941(N-1))$. This can be done conveniently in PROC LOGISTIC by including the scale estimate in the SCALE=WILLIAMS option as follows:

```
proc logistic data=seeds;
   model r/n=soil / scale=williams(0.075941);
   title 'Reduced Model With SCALE=WILLIAMS(0.075941)';
run;
```

Output 42.9.3. Reduced Model with Overdispersion Controlled

Reduced Model With SCALE=WILLIAMS(0.075941)

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.5249	0.2076	6.3949	0.0114
soil	1	0.7910	0.2902	7.4284	0.0064

Results of the reduced model fit are shown in Output 42.9.3. Soil condition remains a significant factor ($p = 0.0064$) for the seed germination.

Example 42.10. Conditional Logistic Regression for Matched Pairs Data

In matched pairs, or *case-control*, studies, conditional logistic regression is used to investigate the relationship between an outcome of being an event (case) or a nonevent (control) and a set of prognostic factors.

The data in this example are a subset of the data from the Los Angeles Study of the Endometrial Cancer Data in Breslow and Day (1980). There are 63 matched pairs, each consisting of a case of endometrial cancer (Outcome=1) and a control (Outcome=0). The case and corresponding control have the same ID. Two prognostic factors are included: Gall (an indicator variable for gall bladder disease) and Hyper (an indicator variable for hypertension). The goal of the case-control analysis is to determine the relative risk for gall bladder disease, controlling for the effect of hypertension.

```
data Data1;
  do ID=1 to 63;
    do Outcome = 1 to 0 by -1;
      input Gall Hyper @@;
      output;
    end;
  end;
  datalines;
0 0  0 0    0 0  0 0    0 1  0 1    0 0  1 0    1 0  0 1
0 1  0 0    1 0  0 0    1 1  0 1    0 0  0 0    0 0  0 0
1 0  0 0    0 0  0 1    1 0  0 1    1 0  1 0    1 0  0 1
0 1  0 0    0 0  1 1    0 0  1 1    0 0  0 1    0 1  0 0
0 0  1 1    0 1  0 1    0 1  0 0    0 0  0 0    0 0  0 0
0 0  0 1    1 0  0 1    0 0  0 1    1 0  0 0    0 1  0 0
0 1  0 0    0 1  0 0    0 1  0 0    0 0  0 0    1 1  1 1
0 0  0 1    0 1  0 0    0 1  0 1    0 1  0 1    0 1  0 0
0 0  0 0    0 1  1 0    0 0  0 1    0 0  0 0    1 0  0 0
0 0  0 0    1 1  0 0    0 1  0 0    0 0  0 0    0 1  0 1
0 0  0 0    0 1  0 1    0 1  0 0    0 1  0 0    1 0  0 0
```

```
0 0  0 0    1 1  1 0    0 0  0 0    0 0  0 0    1 1  0 0
1 0  1 0    0 1  0 0    1 0  0 0
;
```

There are several ways to approach this problem with PROC LOGISTIC.

- Specify the STRATA statement to perform a conditional logistic regression.
- Specify EXACT and STRATA statements to perform an exact conditional logistic regression on the original data set, if you believe the data set is too small or too sparse for the usual asymptotics to hold.
- Transform each matched pair into a single observation then specify a PROC LOGISTIC statement on this transformed data without a STRATA statement; this also performs a conditional logistic regression and produces essentially the same results.
- Specify an EXACT statement on the transformed data.

SAS statements and selected results for these four approaches are given in the remainder of this example.

Conditional Analysis Using the STRATA Statement

In the following SAS statements, PROC LOGISTIC is invoked with the ID variable declared in the STRATA statement to obtain the conditional logistic model estimates. Two models are fitted. The first model contains Gall as the only predictor variable, and the second model contains both Gall and Hyper as predictor variables. Because the option CLODDS=Wald is specified, PROC LOGISTIC computes a 95% Wald confidence interval for the odds ratio for each predictor variable.

```
proc logistic data=Data1;
   strata ID;
   model outcome(event='1')=Gall / clodds=Wald;
run;


proc logistic data=Data1;
   strata ID;
   model outcome(event='1')=Gall Hyper /clodds=Wald;
run;
```

Results from the two conditional logistic analyses are shown in Output 42.10.1 and Output 42.10.2. Note that there is only one response level listed in the "Response Profile" tables, and there is no intercept term in the "Analysis of Maximum Likelihood Estimates" tables.

Output 42.10.1. Conditional Logistic Regression (Gall as Risk Factor)

```
                    The LOGISTIC Procedure

                     Conditional Analysis

                      Model Information

Data Set                          WORK.DATA1
Response Variable                 Outcome
Number of Response Levels         2
Number of Strata                  63
Model                             binary logit
Optimization Technique            Newton-Raphson ridge


     Number of Observations Read         126
     Number of Observations Used         126


                    Response Profile

        Ordered                        Total
          Value      Outcome         Frequency

              1            0              63
              2            1              63

        Probability modeled is Outcome=1.


                    Strata Summary

                Outcome
  Response      -------     Number of
   Pattern      0     1       Strata     Frequency

         1      1     1           63           126
```

Output 42.10.1. (continued)

Conditional Analysis

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
AIC	87.337	85.654
SC	87.337	88.490
-2 Log L	87.337	83.654

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	3.6830	1	0.0550
Score	3.5556	1	0.0593
Wald	3.2970	1	0.0694

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Gall	1	0.9555	0.5262	3.2970	0.0694

Wald Confidence Interval for Adjusted Odds Ratios

Effect	Unit	Estimate	95% Confidence Limits	
Gall	1.0000	2.600	0.927	7.293

Output 42.10.2. Conditional Logistic Regression (Gall and Hyper as Risk Factors)

```
                    The LOGISTIC Procedure

                     Conditional Analysis

                      Model Information

Data Set                          WORK.DATA1
Response Variable                 Outcome
Number of Response Levels         2
Number of Strata                  63
Model                             binary logit
Optimization Technique            Newton-Raphson ridge


     Number of Observations Read         126
     Number of Observations Used         126


                   Response Profile

          Ordered                       Total
            Value      Outcome     Frequency

                1            0            63
                2            1            63

          Probability modeled is Outcome=1.


                      Strata Summary

                 Outcome
  Response       -------    Number of
   Pattern       0     1       Strata    Frequency

         1       1     1           63          126
```

Output 42.10.2. (continued)

Conditional Analysis

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
AIC	87.337	86.788
SC	87.337	92.460
-2 Log L	87.337	82.788

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	4.5487	2	0.1029
Score	4.3620	2	0.1129
Wald	4.0060	2	0.1349

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Gall	1	0.9704	0.5307	3.3432	0.0675
Hyper	1	0.3481	0.3770	0.8526	0.3558

Wald Confidence Interval for Adjusted Odds Ratios

Effect	Unit	Estimate	95% Confidence Limits	
Gall	1.0000	2.639	0.933	7.468
Hyper	1.0000	1.416	0.677	2.965

In the first model, where Gall is the only predictor variable (Output 42.10.1), the odds ratio estimate for Gall is 2.60, which is marginally significant (p=0.0694) and which is an estimate of the relative risk for gall bladder disease. A 95% confidence interval for this relative risk is (0.927, 7.293).

In the second model, where both Gall and Hyper are present (Output 42.10.2), the odds ratio estimate for Gall is 2.639, which is an estimate of the relative risk for gall bladder disease adjusted for the effects of hypertension. A 95% confidence interval for this adjusted relative risk is (0.933, 7.468). Note that the adjusted values (accounting for hypertension) for gall bladder disease are not very different from the unadjusted values (ignoring hypertension). This is not surprising since the prognostic factor Hyper is highly statistically insignificant. The 95% Wald confidence interval for the odds ratio for Hyper is (0.677, 2.965), which contains unity with a p-value greater than 0.3.

Exact Analysis Using the STRATA Statement

When you believe there is not enough data or that the data are too sparse, you can perform a stratified exact conditional logistic regression. The following statements perform stratified exact conditional logistic regressions on the original data set by specifying both the STRATA and EXACT statements.

```
proc logistic data=Data1 exactonly;
   strata ID;
   model outcome(event='1')=Gall;
   exact Gall / estimate=both;
run;


proc logistic data=Data1 exactonly;
   strata ID;
   model outcome(event='1')=Gall Hyper;
   exact Gall Hyper / jointonly estimate=both;
run;
```

Output 42.10.3. Exact Conditional Logistic Regression (Gall as Risk Factor)

The LOGISTIC Procedure

Exact Conditional Analysis

Conditional Exact Tests

Effect	Test	Statistic	--- p-Value --- Exact	Mid
Gall	Score	3.5556	0.0963	0.0799
	Probability	0.0327	0.0963	0.0799

Exact Parameter Estimates

Parameter	Estimate	95% Confidence Limits		p-Value
Gall	0.9555	-0.1394	2.2316	0.0963

Exact Odds Ratios

Parameter	Estimate	95% Confidence Limits		p-Value
Gall	2.600	0.870	9.315	0.0963

Output 42.10.4. Exact Conditional Logistic Regression (Gall and Hyper as Risk Factors)

The LOGISTIC Procedure

Exact Conditional Analysis

Conditional Exact Tests

Effect	Test	Statistic	--- p-Value --- Exact	Mid
Joint	Score	4.3620	0.1150	0.1134
	Probability	0.00316	0.1150	0.1134

Exact Parameter Estimates

Parameter	Estimate	95% Confidence Limits		p-Value
Gall	0.9530	-0.1407	2.2292	0.0969
Hyper	0.3425	-0.4486	1.1657	0.4622

Exact Odds Ratios

Parameter	Estimate	95% Confidence Limits		p-Value
Gall	2.593	0.869	9.293	0.0969
Hyper	1.408	0.639	3.208	0.4622

Note that the score statistics in the "Conditional Exact Tests" tables in Output 42.10.3 and Output 42.10.4 are identical to the score statistics in the conditional analyses in Output 42.10.1 and Output 42.10.2, respectively. The exact odds ratio confidence intervals are much wider than their conditional analysis counterparts, but the parameter estimates are similar. The exact analyses confirm the marginal significance of Gall and the insignificance of Hyper as predictor variables.

Conditional Analysis Using Transformed Data

When each matched set consists of one event and one nonevent, the conditional likelihood is given by

$$\prod_i (1 + \exp(-\boldsymbol{\beta}'(\mathbf{x}_{i1} - \mathbf{x}_{i0}))^{-1}$$

where $\mathbf{x}_{i1}$ and $\mathbf{x}_{i0}$ are vectors representing the prognostic factors for the event and nonevent, respectively, of the ith matched set. This likelihood is identical to the likelihood of fitting a logistic regression model to a set of data with constant response, where the model contains no intercept term and has explanatory variables given by $\mathbf{d}_i = \mathbf{x}_{i1} - \mathbf{x}_{i0}$ (Breslow 1982).

To apply this method, each matched pair is transformed into a single observation, where the variables Gall and Hyper contain the differences between the corresponding values for the case and the control (case − control). The variable Outcome, which will be used as the response variable in the logistic regression model, is given a constant value of 0 (which is the Outcome value for the control, although any constant, numeric or character, will do).

```
data Data2;
   set Data1;
   drop id1 gall1 hyper1;
   retain id1 gall1 hyper1 0;
   if (ID = id1) then do;
      Gall=gall1-Gall; Hyper=hyper1-Hyper;
      output;
   end;
   else do;
      id1=ID; gall1=Gall; hyper1=Hyper;
   end;
run;
```

Note that there are 63 observations in the data set, one for each matched pair. The variable Outcome has a constant value of 0.

In the following SAS statements, PROC LOGISTIC is invoked with the NOINT option to obtain the conditional logistic model estimates. Because the option CLODDS=PL is specified, PROC LOGISTIC computes a 95% profile likelihood confidence interval for the odds ratio for each predictor variable; note that profile likelihood confidence intervals are not currently available when a STRATA statement is specified.

```
proc logistic data=Data2;
   model outcome=Gall / noint clodds=PL;
run;


proc logistic data=Data2;
   model outcome=Gall Hyper / noint clodds=PL;
run;
```

The results are not displayed here.

Exact Analysis Using Transformed Data

Sometimes the original data set in a matched-pairs study may be too large for the exact methods to handle. In such cases it may be possible to use the transformed data set. The following code performs exact conditional logistic regressions on the transformed data set. The results are not displayed here.

```
proc logistic data=Data2 exactonly;
   model outcome=Gall / noint;
   exact Gall / estimate=both;
run;
proc logistic data=Data2 exactonly;
   model outcome=Gall Hyper / noint;
   exact Gall Hyper / jointonly estimate=both;
run;
```

Example 42.11. Complementary Log-Log Model for Infection Rates

Antibodies produced in response to an infectious disease like malaria remain in the body after the individual has recovered from the disease. A serological test detects the presence or absence of such antibodies. An individual with such antibodies is termed seropositive. In areas where the disease is endemic, the inhabitants are at fairly constant risk of infection. The probability of an individual never having been infected in Y years is $\exp(-\mu Y)$, where μ is the mean number of infections per year (refer to the appendix of Draper, Voller, and Carpenter 1972). Rather than estimating the unknown μ, it is of interest to epidemiologists to estimate the probability of a person living in the area being infected in one year. This infection rate γ is given by

$$\gamma = 1 - e^{-\mu}$$

The following statements create the data set sero, which contains the results of a serological survey of malarial infection. Individuals of nine age groups (Group) were tested. Variable A represents the midpoint of the age range for each age group. Variable N represents the number of individuals tested in each age group, and variable R represents the number of individuals that are seropositive.

```
data sero;
   input Group A N R;
   X=log(A);
   label X='Log of Midpoint of Age Range';
   datalines;
1  1.5  123  8
2  4.0  132  6
3  7.5  182 18
4 12.5  140 14
5 17.5  138 20
6 25.0  161 39
7 35.0  133 19
8 47.0   92 25
9 60.0   74 44
;
```

For the ith group with age midpoint A_i, the probability of being seropositive is $p_i = 1 - \exp(-\mu A_i)$. It follows that

$$\log(-\log(1 - p_i)) = \log(\mu) + \log(A_i)$$

By fitting a binomial model with a complementary log-log link function and by using X=log(A) as an offset term, you can estimate $\beta_0 = \log(\mu)$ as an intercept parameter. The following SAS statements invoke PROC LOGISTIC to compute the maximum likelihood estimate of β_0. The LINK=CLOGLOG option is specified to request the complementary log-log link function. Also specified is the CLPARM=PL option, which requests the profile likelihood confidence limits for β_0.

```
proc logistic data=sero;
   model R/N= / offset=X
                link=cloglog
                clparm=pl
                scale=none;
title 'Constant Risk of Infection';
run;
```

Output 42.11.1. Modeling Constant Risk of Infection

```
                        Constant Risk of Infection

                          The LOGISTIC Procedure

                            Model Information

Data Set                        WORK.SERO
Response Variable (Events)      R
Response Variable (Trials)      N
Offset Variable                 X                   Log of Midpoint of Age Range
Model                           binary cloglog
Optimization Technique          Fisher's scoring


                  Number of Observations Read           9
                  Number of Observations Used           9
                  Sum of Frequencies Read            1175
                  Sum of Frequencies Used            1175


                             Response Profile

                    Ordered     Binary          Total
                      Value     Outcome       Frequency

                          1     Event               193
                          2     Nonevent            982


                  Intercept-Only Model Convergence Status

              Convergence criterion (GCONV=1E-8) satisfied.

                           -2 Log L = 967.1158
```

Output 42.11.1. (continued)

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	41.5032	8	5.1879	<.0001
Pearson	50.6883	8	6.3360	<.0001

Number of events/trials observations: 9

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.6605	0.0725	4133.5626	<.0001
X	1	1.0000	0	.	.

Profile Likelihood Confidence Interval for Parameters

Parameter	Estimate	95% Confidence Limits	
Intercept	-4.6605	-4.8057	-4.5219

Results of fitting this constant risk model are shown in Output 42.11.1. The maximum likelihood estimate of $\beta_0 = \log(\mu)$ and its estimated standard error are $\widehat{\beta}_0 = -4.6605$ and $\widehat{\sigma}_{\widehat{\beta}_0} = 0.0725$, respectively. The infection rate is estimated as

$$\widehat{\gamma} = 1 - e^{-\widehat{\mu}} = 1 - e^{-e^{\widehat{\beta}_0}} = 1 - e^{-e^{-4.6605}} = 0.00942$$

The 95% confidence interval for γ, obtained by back-transforming the 95% confidence interval for β_0, is (0.0082, 0.0108); that is, there is a 95% chance that, in repeated sampling, the interval of 8 to 11 infections per thousand individuals contains the true infection rate.

The goodness of fit statistics for the constant risk model are statistically significant ($p < 0.0001$), indicating that the assumption of constant risk of infection is not correct. You can fit a more extensive model by allowing a separate risk of infection for each age group. Suppose μ_i is the mean number of infections per year for the ith age group. The probability of seropositive for the ith group with age midpoint A_i is $p_i = 1 - \exp(-\mu_i A_i)$, so that

$$\log(-\log(1 - p_i)) = \log(\mu_i) + \log(A_i)$$

In the following statements, a complementary log-log model is fit containing Group as an explanatory classification variable with the GLM coding (so that a dummy variable is created for each age group), no intercept term, and X=log(A) as an offset

term. The ODS OUTPUT statement saves the estimates and their 95% profile likelihood confidence limits to ClparmPL data set. Note that $\log(\mu_i)$ is the regression parameter associated with Group$= i$.

```
proc logistic data=sero;
   ods output ClparmPL=ClparmPL;
   class Group / param=glm;
   model R/N=Group / noint
                     offset=X
                     link=cloglog
                     clparm=pl;
   title 'Infectious Rates and 95% Confidence Intervals';
run;
```

Output 42.11.2. Modeling Separate Risk of Infection

Infectious Rates and 95% Confidence Intervals

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Group	1	1	-3.1048	0.3536	77.0877	<.0001
Group	2	1	-4.4542	0.4083	119.0164	<.0001
Group	3	1	-4.2769	0.2358	328.9593	<.0001
Group	4	1	-4.7761	0.2674	319.0600	<.0001
Group	5	1	-4.7165	0.2238	443.9920	<.0001
Group	6	1	-4.5012	0.1606	785.1350	<.0001
Group	7	1	-5.4252	0.2296	558.1114	<.0001
Group	8	1	-4.9987	0.2008	619.4666	<.0001
Group	9	1	-4.1965	0.1559	724.3157	<.0001
X		1	1.0000	0	.	.

Profile Likelihood Confidence Interval for Parameters

Parameter		Estimate	95% Confidence Limits	
Group	1	-3.1048	-3.8880	-2.4833
Group	2	-4.4542	-5.3769	-3.7478
Group	3	-4.2769	-4.7775	-3.8477
Group	4	-4.7761	-5.3501	-4.2940
Group	5	-4.7165	-5.1896	-4.3075
Group	6	-4.5012	-4.8333	-4.2019
Group	7	-5.4252	-5.9116	-5.0063
Group	8	-4.9987	-5.4195	-4.6289
Group	9	-4.1965	-4.5164	-3.9037

Results of fitting the model with a separate risk of infection are shown in Output 42.11.2. For the first age group (Group=1), the point estimate of $\log(\mu_1)$ is -3.1048, which transforms into an infection rate of $1 - \exp(-\exp(-3.1048)) = 0.0438$. A 95% confidence interval for this infection rate is obtained by transforming the 95% confidence interval for $\log(\mu_1)$. For the first age group, the lower and upper confidence limits are $1-\exp(-\exp(-3.8880) = 0.0203$ and $1-\exp(-\exp(-2.4833)) = 0.0801$, respectively; that is, there is a 95% chance that, in repeated sampling, the interval of 20 to 80 infections per thousand individuals contains the true infection rate.

The following statements perform this transformation on the estimates and confidence limits saved in the ClparmPL data set; the resulting estimated infection rates in one year's time for each age group are displayed in Table 42.5. Note that the infection rate for the first age group is high compared to the other age groups.

```
data ClparmPL;
   set ClparmPL;
   Estimate=round( 1000*( 1-exp(-exp(Estimate)) ) );
   LowerCL =round( 1000*( 1-exp(-exp(LowerCL )) ) );
   UpperCL =round( 1000*( 1-exp(-exp(UpperCL )) ) );
run;
```

Table 42.5. Infection Rate in One Year

	Number Infected per 1,000 People		
Age Group	Point Estimate	95% Confidence Limits Lower	Upper
1	44	20	80
2	12	5	23
3	14	8	21
4	8	5	14
5	9	6	13
6	11	8	15
7	4	3	7
8	7	4	10
9	15	11	20

Example 42.12. Complementary Log-Log Model for Interval-Censored Survival Times

Often survival times are not observed more precisely than the interval (for instance, a day) within which the event occurred. Survival data of this form are known as grouped or interval-censored data. A discrete analogue of the continuous proportional hazards model (Prentice and Gloeckler 1978; Allison 1982) is used to investigate the relationship between these survival times and a set of explanatory variables.

Suppose T_i is the discrete survival time variable of the ith subject with covariates $\mathbf{x}_i$. The discrete-time hazard rate λ_{it} is defined as

$$\lambda_{it} = \Pr(T_i = t \mid T_i \geq t, \mathbf{x}_i), \quad t = 1, 2, \ldots$$

Using elementary properties of conditional probabilities, it can be shown that

$$\Pr(T_i = t) = \lambda_{it} \prod_{j=1}^{t-1} (1 - \lambda_{ij}) \quad \text{and} \quad \Pr(T_i > t) = \prod_{j=1}^{t} (1 - \lambda_{ij})$$

Suppose t_i is the observed survival time of the ith subject. Suppose $\delta_i = 1$ if $T_i = t_i$ is an event time and 0 otherwise. The likelihood for the grouped survival data is given by

$$\begin{aligned} L &= \prod_i [\Pr(T_i = t_i)]^{\delta_i} [\Pr(T_i > t_i)]^{1-\delta_i} \\ &= \prod_i \left(\frac{\lambda_{it_i}}{1-\lambda_{it_i}} \right)^{\delta_i} \prod_{j=1}^{t_i} (1-\lambda_{ij}) \\ &= \prod_i \prod_{j=1}^{t_i} \left(\frac{\lambda_{ij}}{1-\lambda_{ij}} \right)^{y_{ij}} (1-\lambda_{ij}) \end{aligned}$$

where $y_{ij} = 1$ if the ith subject experienced an event at time $T_i = j$ and 0 otherwise.

Note that the likelihood L for the grouped survival data is the same as the likelihood of a binary response model with event probabilities λ_{ij}. If the data are generated by a continuous-time proportional hazards model, Prentice and Gloeckler (1978) have shown that

$$\lambda_{ij} = 1 - \exp(-\exp(\alpha_j + \boldsymbol{\beta}' \mathbf{x}_i))$$

where the coefficient vector $\boldsymbol{\beta}$ is identical to that of the continuous-time proportional hazards model, and α_j is a constant related to the conditional survival probability in the interval defined by $T_i = j$ at $\mathbf{x}_i = \mathbf{0}$. The grouped data survival model is therefore equivalent to the binary response model with complementary log-log link function. To fit the grouped survival model using PROC LOGISTIC, you must treat each discrete time unit for each subject as a separate observation. For each of these observations, the response is dichotomous, corresponding to whether or not the subject died in the time unit.

Consider a study of the effect of insecticide on flour-beetles. Four different concentrations of an insecticide were sprayed on separate groups of flour-beetles. The numbers of male and female flour-beetles dying in successive intervals were saved in the data set beetles.

```
data beetles(keep=time sex conc freq);
   input time m20 f20 m32 f32 m50 f50 m80 f80;
   conc=.20;
   freq= m20; sex=1; output;
   freq= f20; sex=2; output;
   conc=.32;
   freq= m32; sex=1; output;
   freq= f32; sex=2; output;
   conc=.50;
   freq= m50; sex=1; output;
   freq= f50; sex=2; output;
   conc=.80;
   freq= m80; sex=1; output;
   freq= f80; sex=2; output;
```

```
  datalines;
 1   3   0  7  1  5  0  4  2
 2  11   2 10  5  8  4 10  7
 3  10   4 11 11 11  6  8 15
 4   7   8 16 10 15  6 14  9
 5   4   9  3  5  4  3  8  3
 6   3   3  2  1  2  1  2  4
 7   2   0  1  0  1  1  1  1
 8   1   0  0  1  1  4  0  1
 9   0   0  1  1  0  0  0  0
10   0   0  0  0  0  0  1  1
11   0   0  0  0  1  1  0  0
12   1   0  0  0  0  1  0  0
13   1   0  0  0  0  1  0  0
14 101 126 19 47  7 17  2  4
;
```

The data set beetles contains four variables: time, sex, conc, and freq. time represents the interval death time; for example, time=2 is the interval between day 1 and day 2. Insects surviving the duration (13 days) of the experiment are given a time value of 14. The variable sex represents the sex of the insects (1=male, 2=female), conc represents the concentration of the insecticide (mg/cm^2), and freq represents the frequency of the observations.

To use PROC LOGISTIC with the grouped survival data, you must expand the data so that each beetle has a separate record for each day of survival. A beetle that died in the third day (time=3) would contribute three observations to the analysis, one for each day it was alive at the beginning of the day. A beetle that survives the 13-day duration of the experiment (time=14) would contribute 13 observations.

A new data set days that contains the beetle-day observations is created from the data set beetles. In addition to the variables sex, conc and freq, the data set contains an outcome variable y and 13 indicator variables day1, day2, . . ., day13. y has a value of 1 if the observation corresponds to the day that the beetle died and has a value of 0 otherwise. An observation for the first day will have a value of 1 for day1 and a value of 0 for day2–day13; an observation for the second day will have a value of 1 for day2 and a value of 0 for day1 and day2–day13. For instance, Output 42.12.1 shows an observation in the beetles data set with time=3, and Output 42.12.2 shows the corresponding beetle-day observations in the data set days.

```
data days;
   retain day1-day13 0;
   array dd[13] day1-day13;
   set beetles;
   if time = 14 then do day=1 to 13;
      y=0; dd[day]=1;
      output;
      dd[day]=0;
   end;
   else do day=1 to time;
      if day=time then y=1;
      else y=0;
      dd[day]=1;
      output;
      dd[day]=0;
   end;
```

Output 42.12.1. An Observation with Time=3 in Data Set Beetles

Obs	time	conc	freq	sex
17	3	0.2	10	1

Output 42.12.2. Corresponding Beetle-day Observations in Days

Obs	time	conc	freq	sex	day1	day2	day3	day4	day5	day6	day7	day8	day9	day10	day11	day12	day13
25	3	0.2	10	1	1	0	0	0	0	0	0	0	0	0	0	0	0
26	3	0.2	10	1	2	0	1	0	0	0	0	0	0	0	0	0	0
27	3	0.2	10	1	3	1	0	0	1	0	0	0	0	0	0	0	0

The following SAS statements invoke PROC LOGISTIC to fit a complementary log-log model for binary data with response variable Y and explanatory variables day1–day13, sex, and conc. Specifying the EVENT= option ensures that the event (y=1) probability is modeled. The coefficients of day1–day13 can be used to estimate the baseline survival function. The NOINT option is specified to prevent any redundancy in estimating the coefficients of day1–day13. The Newton-Raphson algorithm is used for the maximum likelihood estimation of the parameters.

```
proc logistic data=days outest=est1;
   model y(event='1')= day1-day13 sex conc
         / noint link=cloglog technique=newton;
   freq freq;
run;
```

Output 42.12.3. Parameter Estimates for the Grouped Proportional Hazards Model

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
day1	1	-3.9314	0.2934	179.5602	<.0001
day2	1	-2.8751	0.2412	142.0596	<.0001
day3	1	-2.3985	0.2299	108.8833	<.0001
day4	1	-1.9953	0.2239	79.3960	<.0001
day5	1	-2.4920	0.2515	98.1470	<.0001
day6	1	-3.1060	0.3037	104.5799	<.0001
day7	1	-3.9704	0.4230	88.1107	<.0001
day8	1	-3.7917	0.4007	89.5233	<.0001
day9	1	-5.1540	0.7316	49.6329	<.0001
day10	1	-5.1350	0.7315	49.2805	<.0001
day11	1	-5.1131	0.7313	48.8834	<.0001
day12	1	-5.1029	0.7313	48.6920	<.0001
day13	1	-5.0951	0.7313	48.5467	<.0001
sex	1	-0.5651	0.1141	24.5477	<.0001
conc	1	3.0918	0.2288	182.5665	<.0001

Results of the model fit are given in Output 42.12.3. Both sex and conc are statistically significant for the survival of beetles sprayed by the insecticide. Female beetles are more resilient to the chemical than male beetles, and increased concentration increases the effectiveness of the insecticide.

The coefficients of day1–day13 are the maximum likelihood estimates of $\alpha_1, \ldots, \alpha_{13}$, respectively. The baseline survivor function $S_0(t)$ is estimated by

$$\hat{S}_0(t) = \widehat{\Pr}(T > t) = \prod_{j \le t} \exp(-\exp(\hat{\alpha_j}))$$

and the survivor function for a given covariate pattern (sex=x_1 and conc=x_2) is estimated by

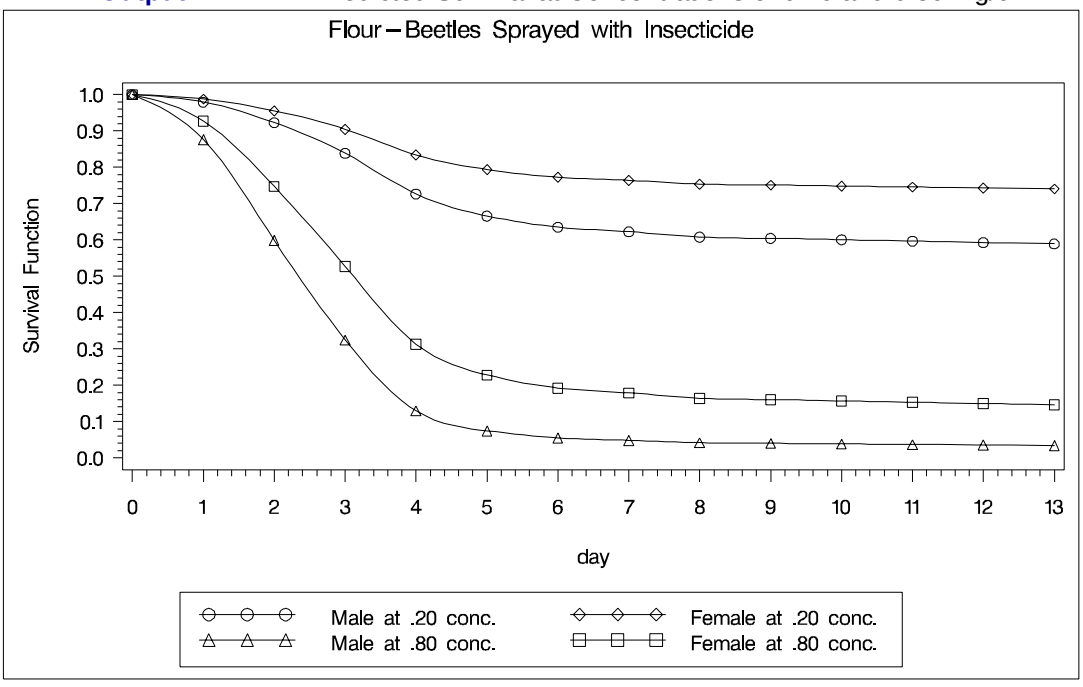
$$y\hat{S}(t) = [\hat{S}_0(t)]^{\exp(-0.5651x_1 + 3.0918x_2)}$$

The following statements compute the survivor curves for male and female flour-beetles exposed to the insecticide of concentrations 0.20 mg/cm^2 and 0.80 mg/cm^2. The GPLOT procedure in SAS/GRAPH software is used to plot the survival curves. Instead of plotting them as step functions, the SPLINE option is used to smooth the curves. These smoothed survival curves are displayed in Output 42.12.4.

```
legend1 label=none frame cframe=white cborder=black position=center
        value=(justify=center);
run;
axis1 label=(angle=90 'Survival Function');
proc gplot data=one;
   plot (s_m20 s_f20 s_m80 s_f80) * day
        / overlay legend=legend1 vaxis=axis1;
   symbol1 v=circle   i=spline c=black height=.8;
   symbol2 v=diamond  i=spline c=black height=.8;
   symbol3 v=triangle i=spline c=black height=.8;
   symbol4 v=square   i=spline c=black height=.8;
run;
```

The probability of survival is displayed on the vertical axis. Notice that most of the insecticide effect occurs by day 6 for both the high and low concentrations.

Output 42.12.4. Predicted Survival at Concentrations of 0.20 and 0.80 mg/cm^2



Example 42.13. Scoring Data Sets with the SCORE Statement

This example first illustrates the syntax used for scoring data sets, then uses a previously scored data set to score a new data set. A generalized logit model is fit to the remote-sensing data set used in Example 25.4 on page 1231 of Chapter 25, "The DISCRIM Procedure," to illustrate discrimination and classification methods. The response variable is Crop and the prognostic factors are x1 through x4.

```
data Crops;
   length Crop $ 10;
   infile datalines truncover;
   input Crop $ @@;
   do i=1 to 3;
     input x1-x4 @@;
     if (x1 ^= .) then output;
   end;
   input;
   datalines;
Corn       16 27 31 33  15 23 30 30  16 27 27 26
Corn       18 20 25 23  15 15 31 32  15 32 32 15
Corn       12 15 16 73
Soybeans   20 23 23 25  24 24 25 32  21 25 23 24
Soybeans   27 45 24 12  12 13 15 42  22 32 31 43
Cotton     31 32 33 34  29 24 26 28  34 32 28 45
Cotton     26 25 23 24  53 48 75 26  34 35 25 78
Sugarbeets 22 23 25 42  25 25 24 26  34 25 16 52
Sugarbeets 54 23 21 54  25 43 32 15  26 54  2 54
Clover     12 45 32 54  24 58 25 34  87 54 61 21
Clover     51 31 31 16  96 48 54 62  31 31 11 11
Clover     56 13 13 71  32 13 27 32  36 26 54 32
Clover     53 08 06 54  32 32 62 16
;
```

You can specify a SCORE statement to score the Crops data using the fitted model. The data together with the predicted values are saved into the data set Score1.

```
proc logistic data=Crops;
   model Crop=x1-x4 / link=glogit;
   score out=Score1;
run;
```

The OUTMODEL= option saves the fitted model information in a data set. In the following statements, the model is again fit, the data and the predicted values are saved into the data set Score2, and the model information is saved in the permanent SAS data set sasuser.CropModel.

```
proc logistic data=Crops outmodel=sasuser.CropModel;
   model Crop=x1-x4 / link=glogit;
   score data=Crops out=Score2;
run;
```

To score data without refitting the model, specify the INMODEL= option to identify a previously saved SAS data set of model information. In the following statements, the model is read from the sasuser.CropModel data set, and the data and the predicted values are saved into the data set Score3.

```
proc logistic inmodel=sasuser.CropModel;
   score data=Crops out=Score3;
run;
```

To set prior probabilities on the responses, specify the PRIOR= option to identify a SAS data set containing the response levels and their priors. In the following statements, the Prior data set contains the values of the response variable (because this example uses single-trial MODEL syntax) and a _PRIOR_ variable containing values proportional to the default priors. The model is fit, then the data and the predicted values are saved into the data set Score4.

```
data Prior;
   input Crop $ 1-10 _PRIOR_;
   datalines;
Clover     11
Corn        7
Cotton      6
Soybeans    6
Sugarbeets  6
;


proc logistic inmodel=sasuser.CropModel;
   score data=Crops prior=prior out=Score4;
run;
```

The data sets Score1, Score2, Score3, and Score4 are identical.

The following statements display the results of scoring the Crops data set in Output 42.13.1.

```
proc freq data=Score1;
   table F_Crop*I_Crop / nocol nocum nopercent;
run;
```

Output 42.13.1. Classification of Data used for Scoring

The FREQ Procedure

Table of F_Crop by I_Crop

F_Crop(From: Crop) / Frequency / Row Pct	I_Crop(Into: Crop) Clover	Corn	Cotton	Soybeans	Sugarbeets	Total
Clover	6	0	2	2	1	11
	54.55	0.00	18.18	18.18	9.09	
Corn	0	7	0	0	0	7
	0.00	100.00	0.00	0.00	0.00	
Cotton	4	0	1	1	0	6
	66.67	0.00	16.67	16.67	0.00	
Soybeans	1	1	1	3	0	6
	16.67	16.67	16.67	50.00	0.00	
Sugarbeets	2	0	0	2	2	6
	33.33	0.00	0.00	33.33	33.33	
Total	13	8	4	8	3	36

Now the previously fit data set sasuser.CropModel is used to score the new observations in the Test data set. The following statements save the results of scoring the test data in the ScoredTest data set and produces Output 42.13.2.

```
data Test;
   input Crop $ 1-10 x1-x4;
   datalines;
Corn       16 27 31 33
Soybeans   21 25 23 24
Cotton     29 24 26 28
Sugarbeets 54 23 21 54
Clover     32 32 62 16
;


proc logistic noprint inmodel=sasuser.CropModel;
   score data=Test out=ScoredTest;
proc print data=ScoredTest label noobs;
   var F_Crop I_Crop P_Clover P_Corn P_Cotton P_Soybeans P_Sugarbeets;
run;
```

Output 42.13.2. Classification of Test Data

From: Crop	Into: Crop	Predicted Probability: Crop=Clover	Predicted Probability: Crop=Corn	Predicted Probability: Crop=Cotton	Predicted Probability: Crop=Soybeans	Predicted Probability: Crop=Sugarbeets
Corn	Corn	0.00342	0.90067	0.00500	0.08675	0.00416
Soybeans	Soybeans	0.04801	0.03157	0.02865	0.82933	0.06243
Cotton	Clover	0.43180	0.00015	0.21267	0.07623	0.27914
Sugarbeets	Clover	0.66681	0.00000	0.17364	0.00000	0.15955
Clover	Cotton	0.41301	0.13386	0.43649	0.00033	0.01631

References

Agresti, A. (1984), *Analysis of Ordinal Categorical Data*, New York: John Wiley & Sons, Inc.

Agresti, A. (1990), *Categorical Data Analysis*, New York: John Wiley & Sons, Inc.

Agresti, A. (2002), *Categorical Data Analysis*, Second Edition, New York: John Wiley & Sons, Inc.

Agresti, A. (1992), "A Survey of Exact Inference for Contingency Tables," *Statistical Science*, 7, 131–177

Aitchison, J. and Silvey, S.D. (1957), "The Generalization of Probit Analysis to the Case of Multiple Responses," *Biometrika*, 44, 131–40.

Albert, A. and Anderson, J.A. (1984), "On the Existence of Maximum Likelihood Estimates in Logistic Regression Models," *Biometrika*, 71, 1–10.

Allison, P.D. (1982), "Discrete-Time Methods for the Analysis of Event Histories," in *Sociological Methods and Research*, 15, ed S. Leinhardt, San Francisco: Jossey-Bass, 61–98.

Allison, P.D. (1999), *Logistic Regression Using the SAS System: Theory and Application*, Cary, NC: SAS Institute Inc.

Ashford, J.R. (1959), "An Approach to the Analysis of Data for Semi-Quantal Responses in Biology Response," *Biometrics*, 15, 573–81.

Bartolucci, A.A. and Fraser, M.D. (1977), "Comparative Step-Up and Composite Test for Selecting Prognostic Indicator Associated with Survival," *Biometrical Journal*, 19, 437–448.

Breslow, N.E. (1982), "Covariance Adjustment of Relative-Risk Estimates in Matched Studies," *Biometrics*, 38, 661–672.

Breslow, N.E. and Day W. (1980), *Statistical Methods in Cancer Research, Volume 1–The Analysis of Case-Control Studies*, Lyon: IARC Scientific Publication No. 32.

Collett, D. (1991), *Modelling Binary Data*, London: Chapman and Hall.

Cook, R.D. and Weisberg, S. (1982), *Residuals and Influence in Regression*, New York: Chapman and Hall.

Cox, D.R. (1970), *Analysis of Binary Data*, New York: Chapman and Hall.

Cox, D.R. and Snell, E.J. (1989), *The Analysis of Binary Data*, Second Edition, London: Chapman and Hall.

DeLong, E.R., DeLong, D.M., and Clarke-Pearson, D.L. (1988), "Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: a Nonparametric Approach," *Biometrics*, 44, 837–845.

Draper, C.C., Voller, A., and Carpenter, R.G. (1972), "The Epidemiologic Interpretation of Serologic Data in Malaria," *American Journal of Tropical Medicine and Hygiene*, 21, 696–703.

Finney, D.J. (1947), "The Estimation from Individual Records of the Relationship between Dose and Quantal Response," *Biometrika*, 34, 320–334.

Fleiss, J.L. (1981), *Statistical Methods for Rates and Proportions*, Second Edition, New York: John Wiley & Sons, Inc.

Freeman, D.H., Jr. (1987), *Applied Categorical Data Analysis*, New York: Marcel Dekker, Inc.

Furnival, G.M. and Wilson, R.W. (1974), "Regressions by Leaps and Bounds," *Technometrics*, 16, 499–511.

Gail, M.H., Lubin, J.H., and Rubinstein, L.V. (1981), "Likelihood Calculations for Matched Case-Control Studies and Survival Studies with Tied Death Times," *Biometrika*, 68, 703–707.

Hanley, J.A. and McNeil, B.J. (1982), "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve," *Radiology*, 143 29–36.

Harrell, F.E. (1986), "The LOGIST Procedure," *SUGI Supplemental Library Guide, Version 5 Edition*, Cary, NC: SAS Institute Inc.

Hirji, K.F. (1992), "Computing Exact Distributions for Polytomous Response Data," *Journal of the American Statistical Association*, 87, 487–492.

Hirji, K.F., Mehta, C.R., and Patel, N.R. (1987), "Computing Distributions for Exact Logistic Regression," *Journal of the American Statistical Association*, 82, 1110–1117.

Hirji, K.F., Tsiatis, A.A., and Mehta, C.R. (1989), "Median Unbiased Estimation for Binary Data," *American Statistician*, 43, 7–11.

Hosmer, D.W, Jr. and Lemeshow, S. (2000), *Applied Logistic Regression*, Second Edition, New York: John Wiley & Sons, Inc.

Howard, S. in the discussion of Cox, D.R. (1972), "Regression Models and Life Tables," *Journal of the Royal Statistical Society, Series B*, 34, 187–220.

Lachin, J. M., (2000), *Biostatistical Methods: The Assessment of Relative Risks*, New York: John Wiley & Sons, Inc.

Lancaster, H. O., (1961), "Significance Tests in Discrete Distributions," *Journal of the American Statistical Association*, 56, 223–234.

LaMotte, L.R., (2002), Personal communication, June 2002 e-mail.

Lawless, J.F. and Singhal, K. (1978), "Efficient Screening of Nonnormal Regression Models," *Biometrics*, 34, 318–327.

Lee, E.T. (1974), "A Computer Program for Linear Logistic Regression Analysis," *Computer Programs in Biomedicine*, 80–92.

McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models*, London: Chapman Hall.

McFadden, D. (1974), "Conditional Logit Analysis of Qualitative Choice Behaviour" in *Frontiers in Econometrics*, edited by P. Zarembka, New York: Academic Press.

Mehta, C.R. and Patel, N.R. (1995), "Exact Logistic Regression: Theory and Examples," *Statistics in Medicine*, 14, 2143–2160.

Mehta, C.R., Patel, N. and Senchaudhuri, P. (1992), "Exact Stratified Linear Rank Tests for Ordered Categorical and Binary Data," *Journal of Computational and Graphical Statistics*, 1, 21–40.

Mehta, C.R., Patel, N. and Senchaudhuri, P. (2000), "Efficient Monte Carlo Methods for Conditional Logistic Regression," *Journal of the American Statistical Association*, 95, 99–108.

Moolgavkar, S.H., Lustbader, E.D., and Venzon, D.J. (1985), "Assessing the Adequacy of the Logistic Regression Model for Matched Case-Control Studies," *Statistics in Medicine*, 4, 425–435.

Naessens, J.M., Offord, K.P., Scott, W.F., and Daood, S.L., (1986), "The MCSTRAT Procedure," in *SUGI Supplemental Library User's Guide, Version 5 Edition*, Cary, NC., SAS Institute Inc. 307–328.

Nagelkerke, N.J.D. (1991), "A Note on a General Definition of the Coefficient of Determination," *Biometrika*, 78, 691–692.

Nelder, J.A. and Wedderburn, R.W.M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society, Series A*, 135, 761–768.

Pregibon, D. (1981), "Logistic Regression Diagnostics," *Annals of Statistics*, 9, 705–724.

Pregibon, D. (1984), "Data Analytic Methods for Matched Case-Control Studies," *Biometrics*, 40, 639–651.

Prentice, P.L. and Gloeckler, L.A. (1978), "Regression Analysis of Grouped Survival Data with Applications to Breast Cancer Data," *Biometrics*, 34, 57–67.

Press, S.J. and Wilson, S. (1978), "Choosing Between Logistic Regression and Discriminant Analysis," *Journal of the American Statistical Association*, 73, 699–705.

Santner, T.J. and Duffy, E.D. (1986), "A Note on A. Albert and J.A. Anderson's Conditions for the Existence of Maximum Likelihood Estimates in Logistic Regression Models," *Biometrika*, 73, 755–758.

SAS Institute Inc. (1995), *Logistic Regression Examples Using the SAS System*, Cary, NC: SAS Institute Inc.

Stokes, M.E., Davis, C.S., and Koch, G.G. (2000), *Categorical Data Analysis Using the SAS System, Second Edition*, Cary, NC: SAS Institute Inc.

Storer, B.E. and Crowley, J. (1985), "A Diagnostic for Cox Regression and General Conditional Likelihoods," *Journal of the American Statistical Association*, 80, 139–147.

Venzon, D.J. and Moolgavkar, S.H. (1988), "A Method for Computing Profile-Likelihood Based Confidence Intervals," *Applied Statistics*, 37, 87–94.

Vollset, S.E., Hirji, K.F., and Afifi, A.A. (1991), "Evaluation of Exact and Asymptotic Interval Estimators in Logistic Analysis of Matched Case-Control Studies," *Biometrics*, 47, 1311–1325.

Walker, S.H. and Duncan, D.B. (1967), "Estimation of the Probability of an Event as a Function of Several Independent Variables," *Biometrika*, 54, 167–179.

Williams, D.A. (1982), "Extra-Binomial Variation in Logistic Linear Models," *Applied Statistics*, 31, 144–148.