Chapter 39
The LOGISTIC Procedure

Chapter Table of Contents

Chapter 39
The LOGISTIC Procedure

Overview

Binary responses (for example, success and failure) and ordinal responses (for example, normal, mild, and severe) 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), and Hosmer and Lemeshow (1989).

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 $\boldsymbol{x}$ is a vector of explanatory variables and $p = \Pr(\mathrm{Y} = 1 \mid \boldsymbol{x})$ is the response probability to be modeled. The linear logistic model has the form

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

where α is the intercept parameter and $\boldsymbol{\beta}$ is the vector of slope 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

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

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

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

where $\alpha_1, \ldots, \alpha_k$ are k intercept parameters, and $\boldsymbol{\beta}$ is the vector of slope 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.

The LOGISTIC procedure fits linear logistic regression models for binary or ordinal response data by the method of maximum likelihood. 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 or the complementary log-log function.

The LOGISTIC procedure provides four variable 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 variables and so on, up to a single model containing all the explanatory variables.

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.

The LOGISTIC procedure enables you to specify categorical variables (also known as CLASS variables) as explanatory variables. It also enables you to specify interaction terms in the same way as in the GLM procedure.

The LOGISTIC procedure allows either a full-rank parameterization or a less than full-rank parameterization. The full-rank parameterization offers four coding methods: effect, reference, polynomial, and orthogonal polynomial. 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 and GENMOD procedures.

The LOGISTIC procedure has some additional options to control how to move effects (either variables or interactions) 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.

Like many procedures in SAS/STAT software that allow 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.

Further features of the LOGISTIC procedure enable you to

- control the ordering of the response levels
- 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

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 1910) describes the syntax of the procedure. The "Details" section (page 1939) summarizes the statistical technique employed by PROC LOGISTIC. The "Examples" section (page 1974) 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 (1995) and to *Logistic Regression Examples Using the SAS System*.

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 ORDER= option in the PROC LOGISTIC 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 SAS System

                      The LOGISTIC Procedure

                        Model Information

          Data Set                          WORK.INGOTS
          Response Variable (Events)        r
          Response Variable (Trials)        n
          Number of Observations            19
          Link Function                     Logit
          Optimization Technique            Fisher's scoring
```

PROC LOGISTIC first lists background information 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.

```
                      The LOGISTIC Procedure

                         Response Profile

              Ordered      Binary          Total
                Value      Outcome       Frequency

                    1      Event              12
                    2      Nonevent          375


                     Model Convergence Status

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

The "Response Profile" table lists the response categories (which are EVENT and NO EVENT when grouped data are input), their ordered values, and their total frequencies for the given data.

```
                      The LOGISTIC Procedure

                      Model Fit Statistics

                                           Intercept
                                Intercept        and
                Criterion        Only       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
```

The "Model Fit Statistics" table 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.

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	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

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

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

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

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

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

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 P= option.

```
                    The LOGISTIC Procedure

  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
```

Finally, the "Association of Predicted Probabilities and Observed Responses" table contains four measures of association for assessing the predictive ability 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 1955.

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 descending;
   model NotReady = 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.

PROC LOGISTIC models the probability of the response level that corresponds to the Ordered Value 1 as displayed in the "Response Profile" table. By default, Ordered Values are assigned to the sorted response values in ascending order.

The DESCENDING option reverses the default ordering of the response values so that NotReady=1 corresponds to the Ordered Value 1 and NotReady=0 corresponds to the Ordered Value 2, as shown in the following table:

The LOGISTIC Procedure

Response Profile

Ordered Value	NotReady	Total Frequency
1	1	12
2	0	375

If the ORDER= option and the DESCENDING option are specified together, the response levels are ordered according to the ORDER= option and then reversed. You should always check the "Response Profile" table to ensure that the outcome of interest has been assigned Ordered Value 1. See the "Response Level Ordering" section on page 1939 for more detail.

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* >**;**
FREQ *variable* **;**
MODEL *response* = < *effects* >< */ options* >**;**
MODEL *events/trials* = < *effects* >< */ options* >**;**
OUTPUT < **OUT=***SAS-data-set* >
< *keyword=name...keyword=name* > */* < *option* >**;**
< *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 statement (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, controls the ordering of the response levels, and suppresses the display of results.

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 1966 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.

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. See the "Response Level Ordering" section on page 1939 for more detail.

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= Data Set" on page 1967 for more information.

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.

NOPRINT
suppresses all displayed output. Note that this option temporarily disables the Output Delivery System (ODS); see Chapter 15, "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. When ORDER=FORMATTED (the default) for numeric variables for which you have supplied no explicit format (that is, for which there is no corresponding FORMAT statement in the current PROC LOGISTIC run or in the DATA step that created the data set), the levels are ordered by their internal (numeric) value. Note that this represents a change from previous releases for how class levels are ordered. In releases previous to Version 8, numeric class levels with no explicit format were ordered by their BEST12. formatted values, and in order to revert to the previous ordering you can specify this format explicitly for the affected classification variables. The change was implemented because the former default behavior for ORDER=FORMATTED

often resulted in levels not being ordered numerically and usually required the user to intervene with an explicit format or ORDER=INTERNAL to get the more natural ordering. 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

By default, ORDER=FORMATTED. For FORMATTED and INTERNAL, the sort order is machine dependent. 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*.

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 names of the variables in this data set are the same as those of the explanatory variables in the MODEL statement plus the name Intercept for the intercept parameter in the case of a binary response model. For an ordinal response model with more than two response categories, the parameters are named Intercept, Intercept2, Intercept3, and so on. The output data set also includes a variable named _LNLIKE_, which contains the log likelihood.

See the section "OUTEST= Output Data Set" on page 1966 for more information.

SIMPLE

displays simple descriptive statistics (mean, standard deviation, minimum and maximum) for each explanatory variable in the MODEL statement. The SIMPLE option generates a breakdown of the simple descriptive statistics for the entire data set and also for individual response levels. The NOSIMPLE option suppresses this output and is the default.

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).

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. 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 dummy 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.

LPREFIX= *n*
specifies that, at most, the first *n* characters of a CLASS variable label be used in creating labels for the corresponding dummy variables.

ORDER=DATA | FORMATTED | FREQ | INTERNAL
specifies the sorting order for the levels of classification variables. 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. When ORDER=FORMATTED (the default) for numeric variables for which you have supplied no explicit format (that is, for which there is no corresponding FORMAT statement

in the current PROC LOGISTIC run or in the DATA step that created the data set), the levels are ordered by their internal (numeric) value. Note that this represents a change from previous releases for how class levels are ordered. In releases previous to Version 8, numeric class levels with no explicit format were ordered by their BEST12. formatted values, and in order to revert to the previous ordering you can specify this format explicitly for the affected classification variables. The change was implemented because the former default behavior for ORDER=FORMATTED often resulted in levels not being ordered numerically and usually required the user to intervene with an explicit format or ORDER=INTERNAL to get the more natural ordering. 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

By default, ORDER=FORMATTED. For FORMATTED and INTERNAL, the sort order is machine dependent. 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.

EFFECT	specifies effect coding
GLM	specifies less than full rank, reference cell coding; this option can only be used as a global option
ORTHPOLY	specifies orthogonal polynomial coding
POLYNOMIAL | POLY	specifies polynomial coding
REFERENCE | REF	specifies reference cell coding

The EFFECT, POLYNOMIAL, REFERENCE, and ORTHPOLY parameterizations are full rank. For the EFFECT and REFERENCE parameterizations, the REF= option in the CLASS statement determines the reference level.

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 dummy 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			
A	**Design Matrix**		
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.

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

GLM Coding				
A	**Design Matrix**			
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.

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			
A	**Design Matrix**		
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

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			
A	**Design Matrix**		
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 dummy 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			
A	**Design Matrix**		
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.

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

specifies the reference level for PARAM=EFFECT or PARAM=REFERENCE. 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

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 statement 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 $\mathbf{L}\boldsymbol{\beta} = \mathbf{0}$. 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 1913). Optionally, the CONTRAST statement enables you to estimate each row, $l_i'\boldsymbol{\beta}$, of $\mathbf{L}\boldsymbol{\beta}$ and test the hypothesis $l_i'\boldsymbol{\beta} = 0$. 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. Labels can contain up to 256 characters.

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 **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 **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 **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 $\mathbf{L}\boldsymbol{\beta} = \mathbf{0}$, 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 $\mathbf{L}$ matrix in your CONTRAST statement. If the elements of $\mathbf{L}$ 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 $\mathbf{L}$ 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 $\mathbf{L}$.

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

ALPHA= *value*
specifies the significance level of the confidence interval for each contrast when the ESTIMATE option is specified. The default is ALPHA=.05, resulting in a 95% confidence interval for each contrast.

E
requests that the $\mathbf{L}$ matrix be displayed.

ESTIMATE=*keyword*
requests that each individual contrast (that is, each row, $l_i'\boldsymbol{\beta}$, of $\mathbf{L}\boldsymbol{\beta}$) or exponentiated contrast ($\mathrm{e}^{l_i'\boldsymbol{\beta}}$) 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 ($\mathrm{e}^{l_i'\boldsymbol{\beta}}$), 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 $\mathbf{v}$ is a vector, define ABS($\mathbf{v}$) to be the absolute value of the element of $\mathbf{v}$ with the largest absolute value. Define C to be equal to ABS($\mathbf{K}'$) if ABS($\mathbf{K}'$) is greater than 0; otherwise, C equals 1 for a row $\mathbf{K}'$ in the contrast. If ABS($\mathbf{K}' - \mathbf{K}'\mathbf{T}$) is greater than C$*$*number*, then $\mathbf{K}$ is declared nonestimable. The $\mathbf{T}$ matrix is the Hermite form matrix $(\mathbf{X}'\mathbf{X})^{-}(\mathbf{X}'\mathbf{X})$, and $(\mathbf{X}'\mathbf{X})^{-}$ represents a generalized inverse of the matrix $\mathbf{X}'\mathbf{X}$. The value for $number$ must be between 0 and 1; the default value is 1E$-$4.

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.

MODEL Statement

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

The MODEL statement names the response variable and the explanatory effects, including covariates, main effects, interactions, and nested effects. If you omit the explanatory variables, the procedure fits an intercept-only model.

Two forms of the MODEL statement can be specified. The first form, referred to as *single-trial* syntax, is applicable to both binary response data and ordinal 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 *single-trial* syntax, you specify one variable (preceding the equal sign) as the response variable. This variable can be character or numeric. Values of this variable are sorted by the ORDER= option (and the DESCENDING option, if specified) in the PROC LOGISTIC statement.

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.

For both forms of the MODEL statement, explanatory *effects* follow the equal sign. The 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. See the section "Specification of Effects" on page 1517 of Chapter 30, "The GLM Procedure."

Table 39.1 summarizes the options available in the MODEL statement.

Table 39.1. Model Statement Options

Option	Description
Model Specification Options	
LINK=	specifies link function
NOINT	suppresses intercept
NOFIT	suppresses model fitting
OFFSET=	specifies offset variable

Table 39.1. (continued)

Option	Description
SELECTION=	specifies variable selection method
Variable 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 variables included in every model
MAXSTEP=	specifies maximum number of steps for STEPWISE selection
SEQUENTIAL	adds or deletes variables in sequential order
SLENTRY=	specifies significance level for entering variables
SLSTAY=	specifies significance level for removing variables
START=	specifies the number of variables in first model
STOP=	specifies the number of variables in final model
STOPRES	adds or deletes variables by residual chi-square criterion
Model-Fitting Specification Options	
ABSFCONV=	specifies the absolute function convergence criterion
FCONV=	specifies the relative function convergence criterion
GCONV=	specifies the relative gradient convergence criterion
XCONV=	specifies the relative parameter convergence criterion
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	

Table 39.1. (continued)

Option	Description
INFLUENCE	displays influence statistics
IPLOTS	requests index plots
Options for Display of Details	
CORRB	displays correlation matrix
COVB	displays covariance matrix
EXPB	displays the exponentiated values of estimates
ITPRINT	displays iteration history
NODUMMYPRINT	suppresses the "Class Level Information" table
PARMLABEL	displays the parameter labels
RSQUARE	displays generalized R^2
STB	displays the standardized estimates

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 1944.

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 1959 for more detail.

ALPHA=*value*

sets the significance level for the confidence intervals for regression parameters or odds ratios. The value must be between 0 and 1. The default value of 0.05 results in the calculation of a 95% confidence interval. 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 CLPARM=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 1950 for more information.

CONVERGE=*value*

is the same as specifying the XCONV= option.

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 1928). 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 1956.

DETAILS

produces a summary of computational details for each step of the variable 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 III 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 1969. 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 elimination 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 1944.

GCONV=*value*

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

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

where l_i is value of the log-likelihood function, $\mathbf{g}_i$ is the gradient vector, and $\mathbf{H}_i$ is the negative (expected) Hessian matrix, all at iteration i. This is the default convergence criterion, and the default value is $1\text{E}{-}8$. See the section "Convergence Criteria" on page 1944.

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 described 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 of 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 1963.

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 39.4 on page 1998 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 1989) 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.

LINK=CLOGLOG | LOGIT | PROBIT

L=CLOGLOG | LOGIT | PROBIT

specifies the link function for the response probabilities. CLOGLOG is the complementary log-log function, LOGIT is the log odds function, and PROBIT (or NORMIT) is the inverse standard normal distribution function. By default, LINK=LOGIT. See the section "Link Functions and the Corresponding Distributions" on page 1940 for details.

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 1944.

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 or the first intercept for the ordinal response model. This can be particularly useful in conditional logistic analysis; see Example 39.9 on page 2026.

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.

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= Data Set" on page 1968 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 1957. 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 crossvalidated 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 below to the nearest 0.02) to the highest estimated probability (rounded above 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 default 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 1948.

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 39.8 on page 2021 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 1958.

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 1945.

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. 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, SINGULAR=1E−12.

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, START=n specifies that the smallest 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 1942 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,

$$\max_j |\delta_i^{(j)}| < value$$

where

$$\delta_i^{(j)} = \begin{cases} \theta_i^{(j)} - \theta_{i-1}^{(j)} & |\theta_{i-1}^{(j)}| < 0.01 \\ \frac{\theta_i^{(j)} - \theta_{i-1}^{(j)}}{\theta_{i-1}^{(j)}} & \text{otherwise} \end{cases}$$

and $\theta_i^{(j)}$ is the estimate of the jth parameter at iteration i. See the section "Convergence Criteria" on page 1944.

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 crossvalidated 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 1955 and the "Regression Diagnostics" section on page 1963.

If you use the *single-trial* syntax, the data set may also contain 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" on page 1967.

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 response variable, 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.

Statistic Options Valid When the Response is Binary or Ordinal

LOWER=*name*

L=*name*

specifies the lower confidence limit for the probability of an event response if *events/trials* syntax is specified, or the lower confidence limit for the probability that the response is less than or equal to the value of _LEVEL_ if *single-trial* syntax is specified. See the ALPHA= option , which follows.

PREDICTED=*name*

PRED=*name*

PROB=*name*

P=*name*

specifies the predicted probability of an event response if *events/trials* syntax is specified, or the predicted probability that the response variable is less than or equal to the value of _LEVEL_ if *single-trial* syntax is specified (in other words, $\Pr(Y \leq \_LEVEL\_)$, where Y is the response variable).

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 response levels, 1,2, and 3, the cumulative probabilities are Pr(Y$\leq$1), Pr(Y$\leq$2), and Pr(Y$\leq$3). The cumulative probability for the last response level always has the constant value of 1.

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. Note that, for ordinal models, the cross validated probabilities are not computed and are set to missing.

See the end of this section for further details regarding the PREDPROBS= option.

STDXBETA=*name*
specifies the standard error estimate of XBETA (the definition of which follows).

UPPER=*name*
U=*name*
specifies the upper confidence limit for the probability of an event response if *events/trials model* is specified, or the upper confidence limit for the probability that the response is less than or equal to the value of _LEVEL_ if *single-trial* syntax is specified. See the ALPHA=option mentioned previously.

XBETA=*name*
specifies the estimate of the linear predictor $\alpha_i + \boldsymbol{\beta}'\boldsymbol{x}$, where i is the corresponding ordered value of _LEVEL_.

Statistic Options Valid Only When the Response is Binary

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 named 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

ALPHA=*value*
sets the confidence level used for the confidence limits for the appropriate response probabilities. The quantity *value* must be between 0 and 1. By default, ALPHA=0.05, which results in the calculation of a 95% confidence interval.

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 predicated probability of an event when a current event trial is removed.

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

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

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

The cross-validated predicted probabilities are precisely those used in the CTABLE option. Refer to the "Predicted Probability of an Event for Classification" section on page 1957 for details of the computation.

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: $\mathbf{L}\boldsymbol{\beta} = \mathbf{c}$) specified in a single TEST statement.

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. 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: $\mathbf{L}\boldsymbol{\beta} = \mathbf{c}$. This includes $\mathbf{L}\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}})\mathbf{L}'$ bordered by $(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{c})$ and $[\mathbf{L}\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}})\mathbf{L}']^{-1}$ bordered by $[\mathbf{L}\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}})\mathbf{L}']^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{c})$, where $\widehat{\boldsymbol{\beta}}$ is the maximum likelihood estimator of $\boldsymbol{\beta}$ and $\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}})$ is the estimated covariance matrix of $\widehat{\boldsymbol{\beta}}$.

For more information, see the "Testing Linear Hypotheses about the Regression Coefficients" section on page 1963.

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 1952.

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.

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, or explanatory variables is excluded from the analysis. 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

For binary response data, the default response function modeled is

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

where p is the probability of the response level identified in the "Response Profiles" table in the displayed output as "Ordered Value 1." Since

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

the effect of reversing the order of the two values of the response is to change the signs of α and $\boldsymbol{\beta}$ in the model $\text{logit}(p) = \alpha + \boldsymbol{\beta}'\boldsymbol{x}$. Response level ordering is important because PROC LOGISTIC always models the probability of response levels with *lower* Ordered Value.

By default, response levels are assigned to Ordered Values in ascending, sorted order (that is, the lowest level is assigned Ordered Value 1, the next lowest is assigned 2, and so on). There are a number of ways that you can control the sort order of the response categories and, therefore, which level is assigned Ordered Value 1. 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. By default, PROC LOGISTIC assigns Ordered Value 1 to response level 0, causing the probability of the nonevent to be modeled. There are several ways to change this. Besides recoding the variable Y, you can do the following.

- specify the DESCENDING option in the PROC LOGISTIC statement, which reverses the default ordering of Y from (0,1) to (1,0), making 1 (the event) the level with Ordered Value 1:

```
proc logistic descending;
   model Y=Exposure;
run;
```

- 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, Y=1 becomes Ordered Value 1.

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

Link Functions and the Corresponding Distributions

Three 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

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

- 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 variances of these three corresponding 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

Suppose the response variable can take on the ordered values $1, \ldots, k, k+1$ where k is an integer ≥ 1. If you use *events/trials* 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.

The likelihood for the jth observation with ordered response value y_j and explanatory variables vector $\boldsymbol{x}_j$ is given by

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

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

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. You can use the TECHNIQUE= option to select a fitting algorithm.

Iteratively Reweighted Least-Squares Algorithm

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

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

With p_{ij} denoting the probability that the *j*th observation has response value *i*, the expected value of $\mathbf{Z}_j$ is $\mathbf{p}_j = (p_{1j}, \ldots, p_{(k+1)j})'$. 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 $\mathbf{p}_j$. Let $\boldsymbol{\gamma}$ be the vector of regression parameters; in other words, $\boldsymbol{\gamma}' = (\alpha_1, \ldots, \alpha_k, \boldsymbol{\beta}')$. And let $\mathbf{D}_j$ be the matrix of partial derivatives of $\mathbf{p}_j$ with respect to $\boldsymbol{\gamma}$. The estimating equation for the regression parameters is

$$\sum_j \mathbf{D}_j' \mathbf{W}_j (\mathbf{Z}_j - \mathbf{p}_j) = \mathbf{0}$$

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 $\mathbf{p}_j$ as the diagonal.

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

$$\boldsymbol{\gamma}_{m+1} = \boldsymbol{\gamma}_m + (\sum_j \mathbf{D}_j' \mathbf{W_j} \mathbf{D}_j)^{-1} \sum_j \mathbf{D}_j' \mathbf{W}_j (\mathbf{Z}_j - \mathbf{p}_j)$$

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

The covariance matrix of $\hat{\boldsymbol{\gamma}}$ is estimated by

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

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

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

With parameter vector $\boldsymbol{\gamma}' = (\alpha_1, \ldots, \alpha_k, \boldsymbol{\beta}')$, the gradient vector and the Hessian matrix are given, respectively, by

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

With a starting value of $\boldsymbol{\gamma}_0$, the maximum likelihood estimate $\hat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$ is obtained iteratively until convergence is obtained:

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

If the likelihood evaluated at $\boldsymbol{\gamma}_{m+1}$ is less than that evaluated at $\boldsymbol{\gamma}_m$, then $\boldsymbol{\gamma}_{m+1}$ is recomputed by step-halving or ridging.

The covariance matrix of $\hat{\boldsymbol{\gamma}}$ is estimated by

$$\widehat{\text{cov}}(\hat{\boldsymbol{\gamma}}) = \mathbf{H}_{\hat{\boldsymbol{\gamma}}}^{-1}$$

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.

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).

Consider a binary response model. Let Y_j be the response of the ith subject and let $\boldsymbol{x}_j$ 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,

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

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

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

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 $\hat{p}_j$ be the estimated probability of the observed response. The three criteria displayed by the LOGISTIC procedure are calculated as follows:

- -2 Log Likelihood:

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

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

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

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

- Akaike Information Criterion:

  $$\text{AIC} = -2 \text{ Log L} + 2(k + s)$$

  where k is the total number of response levels minus one, and s is the number of explanatory effects.

- Schwarz Criterion:

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

  where k and s are 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:

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

where $L(\mathbf{0})$ is the likelihood of the intercept-only model, $L(\widehat{\boldsymbol{\beta}})$ 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 $\mathbf{U}(\boldsymbol{\gamma})$be the vector of first partial derivatives of the log likelihood with respect to the parameter vector $\boldsymbol{\gamma}$, and let $\mathbf{H}(\boldsymbol{\gamma})$ be the matrix of second partial derivatives of the log likelihood with respect to $\boldsymbol{\gamma}$. That is, $\mathbf{U}(\boldsymbol{\gamma})$ is the gradient vector, and $\mathbf{H}(\boldsymbol{\gamma})$ is the Hessian matrix. Let $\mathbf{I}(\boldsymbol{\gamma})$ be either $-\mathbf{H}(\boldsymbol{\gamma})$ or the expected value of $-\mathbf{H}(\boldsymbol{\gamma})$. Consider a null hypothesis H_0. Let $\hat{\boldsymbol{\gamma}}_0$ be the MLE of $\boldsymbol{\gamma}$ under H_0. The chi-square score statistic for testing H_0 is defined by

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

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 $\boldsymbol{\gamma}$ 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 model has a parameter vector

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

where $\alpha_1, \ldots, \alpha_k$ are intercept parameters, and $\beta_1, \ldots, \beta_s$ are slope parameters for the explanatory effects. Consider the null hypothesis H_0: $\beta_{t+1} = \ldots = \beta_s = 0$ where t < s. For the reduced model with t explanatory effects, let $\hat{\alpha}_1, \ldots, \hat{\alpha}_k$ be the MLEs of the unknown intercept parameters, and let $\hat{\beta}_1, \ldots, \hat{\beta}_t$ be the MLEs of the unknown slope parameters. The residual chi-square is the chi-square score statistic testing the null hypothesis H_0; that is, the residual chi-square is

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

where $\hat{\boldsymbol{\gamma}}_0 = (\hat{\alpha}_1, \ldots, \hat{\alpha}_k, \hat{\beta}_1, \ldots, \hat{\beta}_t, 0, \ldots, 0)'$.

The residual chi-square has an asymptotic chi-square distribution with $s - t$ degrees of freedom. 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 "Model-Fitting Information and 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. 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) Procedure" 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.

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. 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

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

where g(.) is the link function, and $\boldsymbol{\gamma}_i = (\alpha_i, \beta_{i1}, \ldots, \beta_{is})'$ 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

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

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 $\boldsymbol{\gamma}$ is

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

and the chi-squared score statistic $\mathbf{U}'(\hat{\boldsymbol{\gamma}}_0)\mathbf{I}^{-1}(\hat{\boldsymbol{\gamma}}_0)\mathbf{U}(\hat{\boldsymbol{\gamma}}_0)$ 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 CLPARMS= 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 $\beta_j = \theta$ is defined as

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

where $\mathcal{B}_j(\theta)$ is the set of all $\boldsymbol{\beta}$ with the jth element fixed at θ, 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 $l_0 = l_{\max} - .5\chi^2_{1-\alpha,1}$, where $\chi^2_{1-\alpha,1}$ 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

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

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

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

where λ is the Lagrange multiplier, $\mathbf{e}_j$ is the jth unit vector, and θ 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

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

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

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

The log odds for $White$ is

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

Therefore, the log odds ratio of $Black$ versus $White$ becomes

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

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

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

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

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

Consider the hypothetical example of heart disease among race in Hosmer and Lemeshow (1989, p 44). 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.

Rank Correlation of Observed Responses and Predicted Probabilities

Define an event response as the response having Ordered Value of 1. A pair of observations with different responses is said to be concordant (discordant) if the observation with the response that has the larger Ordered Value has the lower (higher) predicted event probability. If a pair of observations with different responses 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 probabilities into intervals of length 0.002 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 the area under the receiver operating characteristic (ROC) curve when the response is binary (Hanley and McNeil 1982).

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 1906. Predicted probabilities and confidence limits can be output to a data set with the OUTPUT statement.

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

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

is estimated by

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

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 $\Pr(Y \leq i \mid \mathbf{x})$ are obtained by back-transforming the corresponding measures for the linear predictor.

Link	Predicted Probability	100(1-0.5α)% 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	$1-\mathrm{e}^{-\mathrm{e}^{\hat{\eta}_i}}$	$1-\mathrm{e}^{-\mathrm{e}^{\hat{\eta}_i \pm z_{\alpha/2}\hat{\sigma}(\hat{\eta}_i)}}$

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, 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

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

where

y_j is 1 for an event response and 0 otherwise

w_j is the WEIGHT value

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

h_{jj} is the hat diagonal element (defined on page 1964) 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 ($\hat{p}_j^*$) is large. The probability $\hat{p}_j^*$ is the reduced-bias estimate based on a one-step approximation given in the previous section. For a given cutpoint z, the jth individual is predicted to give a positive response if $\hat{p}_j^* \geq z$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

At the start, let $w_i^* = 1$ and let p_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

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

where m 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 39.8 on page 2021 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 (1989) 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 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 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 $\hat{p}_j$ of the event of interest is calculated. Note that $\hat{p}_j$ is computed directly without resorting to the one-step approximation, as used in the calculation of the classification table.

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:

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

where $I(.)$ 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_). The area under the ROC curve, as determined by the trapezoidal rule, is given by the statistic *c* in the "Association of Predicted Probabilities and Observed Responses" table.

Testing Linear Hypotheses about the Regression Coefficients

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

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

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 $\boldsymbol{\beta}$ includes slope parameters as well as intercept parameters. The Wald chi-square statistic for testing H_0 is computed as

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

where $\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}})$ 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.

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 total weight (the product of the WEIGHT and FREQ values) of the jth observation.
p_j	is the probability of an event response for the jth observation given by $p_j = F(\alpha + \boldsymbol{\beta}'\boldsymbol{x}_j)$, where $F(.)$ is the inverse link function defined on page 1940.
$\mathbf{b}$	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 p_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

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

where

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

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

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

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

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 1989).

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

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

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 DFBETA diagnostic for each observation. The DFBETA 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 1957. For the jth observation, the DFBETAS are given by

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

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

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

$\mathbf{\Delta b}_j^1$ 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

$$\Delta_j \mathrm{D} = d_j^2 + \overline{C}_j$$

and

$$\Delta_j \chi^2 = \overline{C}_j / h_{jj}$$

OUTEST= Output Data Set

The OUTEST= data set contains estimates of the regression coefficients. If you use the COVOUT option in the PROC LOGISTIC statement, the OUTEST= data set also contains the estimated covariance matrix of the parameter estimates.

Number of Variables and Number of Observations

The data set contains one variable for each intercept parameter and one variable for each explanatory variable in the MODEL statement.

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 of the parameter estimators. 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 Data Set

The OUTEST= data set contains the following variables:

- any BY variables specified
- _LINK_, a character variable of length 8 with three possible values: CLOGLOG for the complementary log-log function, LOGIT for the logit function, or NORMIT for the normit function
- _TYPE_, a character variable of length 8 with two possible values: PARMS for parameter estimates or COV for covariance estimates
- _NAME_, a character variable containing the name of the response variable when _TYPE_=PARMS or the name of a model parameter when _TYPE_=COV

- _STATUS_, a character variable which indicates whether the estimates have converged
- one variable for each intercept parameter. In the case that one BY group fits a binary response model and another BY group fits an ordinal response model with more than two response levels, the data set contains the intercept variables Intercept (for the only intercept of the binary response model) and Intercept1, ..., Intercept$r$, where r is the largest number (greater than 1) of intercept parameters among the BY groups. Any of these variables not pertinent to a specific BY group have their values set to missing.
- one variable for each model parameter and the OFFSET= variable if specified. If an explanatory variable is not included in the final model in a model building process, the corresponding estimates of parameters and covariances are set to missing values.
- _LNLIKE_, the log likelihood

INEST= 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 model and Intercept, Intercept2, Intercept3, and so forth, for ordinal 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

The OUT= data set contains all the variables in the input data set and variables for statistics you specify using *keyword=name* or the PREDPROBS= option in the OUTPUT statement.

In addition, if you use the *single-trial* syntax and the statistics you requested include XBETA=, STDXBETA=, PREDICTED=, LCL=, and UCL= options, the OUT= data set contains the variable _LEVEL_ ; and for a model with $k+1$ response categories $(k \geq 1)$, each input observation generates k observations with k different values of _LEVEL_ . When there are more than two response levels (i.e., k>1), 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.

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 allows, for instance, predicted probabilities to be computed for new observations.

OUTROC= Data Set

The OUTROC= data set contains data necessary for producing the ROC curve. 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 1956. An ROC curve is obtained by plotting _SENSIT_ against _1MSPEC_. For more information, see the section "Receiver Operating Characteristic Curves" on page 1962.

Computational Resources

The memory needed to fit a model is approximately $24(n+2)^2$ bytes, where n is the number of parameters estimated. For 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.

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:

- the name of the input Data Set
- the name and label of the Response Variable if the *single-trial* syntax is used
- the number of Response Levels
- the name of the Events Variable if the *events/trials* syntax is used
- the name of the Trials Variable if the *events/trials* syntax is used
- the Number of Observations used in the analysis
- the name of the Offset Variable if the OFFSET= option is specified
- the name of the Frequency Variable if the FREQ statement is specified
- the name of the Weight Variable if the WEIGHT statement is specified
- the Sum of Weights of all the observations used in the analysis
- the Link Function
- the "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 "NO EVENT" if the *events/trials* syntax is used; the count or frequency; and the sum of weights if the WEIGHT statement is specified
- the "Class Level Information" table, which gives the level and the design variables for each CLASS explanatory variable
- if you specify the SIMPLE option in the PROC LOGISTIC statement, the "Descriptive Statistics for Continuous Explanatory Variables" table for continuous explanatory variables, and the "Frequency Distribution of Class Variables" and the "Weight Distribution of Class Variables" tables (if the WEIGHT statement is specified). 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.
- if you use the ITPRINT option in the MODEL statement, the "Maximum Likelihood Iterative Phase" table, which 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.
- if you use the SCALE= option in the MODEL statement, the Pearson and deviance goodness-of-fit statistics
- if an ordinal response model is fitted, the score test result for testing the parallel lines assumption. 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."

- the "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.
- if you specify the RSQUARE option in the MODEL statement, generalized R^2 measures for the fitted model
- if the model contains an effect involving a CLASS variable, the "Type III Analysis of Effects" table, which gives the Wald Chi-square statistic, the degrees of freedom, and the p-value for each effect in the model
- the "Analysis of Maximum Likelihood Estimates" table, which includes
  - the maximum likelihood estimate of the parameter
  - the estimated standard error of the parameter estimate, computed as the square root of the corresponding diagonal element of the estimated covariance matrix
  - the Wald chi-square statistic, computed by squaring the ratio of the parameter estimate divided by its standard error estimate
  - the p-value of the Wald chi-square statistic with respect to a chi-square distribution with one degree of freedom
  - if you specify the STB option in the MODEL statement, the standardized estimate for the slope parameter, 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.
  - if you specify the EXPB option in the MODEL statement, the value of $(\mathrm{e}^{\hat{\beta}_i})$ for each slope parameter β_i. For continuous variables, this is equivalent to the estimated odds ratio for a 1 unit change.
  - if you specify the PARMLABEL option in the MODEL statement, the label of the variable (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.
- the "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.

- measures of association between predicted probabilities and observed responses, which include 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*
- if you use the CLPARM= option in the MODEL statement, confidence intervals for all the parameters
- if you use the CLODDS= option in the MODEL statement, confidence intervals for all the odds ratios
- if you use a FORWARD, BACKWARD, or STEPWISE selection method, a summary of the model-building process, which 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)
- if you specify the FAST option in the MODEL statement, the "Analysis of Variables Removed by Fast Backward Elimination" table, which 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
- if you specify the DETAILS option in the MODEL statement, the "Analysis of Effects not in the Model" table, which 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
- the 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.
- if you use the COVB option in the MODEL statement, the estimated covariance matrix of the parameter estimates
- if you use the CORRB option in the MODEL statement, the estimated correlation matrix of the parameter estimates
- the "Linear Hypothesis Testing" table, which gives the result of the Wald test for each TEST statement (if specified)
- if you use the LACKFIT option in the MODEL statement, the results of the Hosmer and Lemeshow test for the goodness of fit of the fitted model

- if you use the INFLUENCE option in the MODEL statement, the "Regression Diagnostics" table, which 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 (DIFCHSQ), and the change in the deviance (DIFDEV)
- if you specified the IPLOTS option in the MODEL statement,
  - the index plot of Pearson residuals
  - the index plot of deviance residuals
  - the index plot of the diagonal elements of the hat matrix
  - index plots of the 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
  - the index plot of confidence interval displacement diagnostics C
  - the index plot of confidence interval displacement diagnostics CBAR
  - the index plot of the changes in the Pearson chi-square statistic
  - the index plot of the changes in the deviance

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 15, "Using the Output Delivery System."

Table 39.2. 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 39.2. (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
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
OddsRatios	Odds ratios	MODEL	default
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
TestPrint1	**L**[cov(**b**)]**L**' and **Lb-c**	TEST	PRINT

Table 39.2. (continued)

ODS Table Name	Description	Statement	Option
TestPrint2	Ginv(**L**[cov(**b**)]**L**') and Ginv(**L**[cov(**b**)]**L**')(**Lb-c**)	TEST	PRINT
TestStmts	Linear hypotheses testing results	TEST	default
TypeIII	Type III tests of effects	MODEL	default (with CLASS variables)

Examples

Example 39.1. 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 DESCENDING option causes 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 descending outest=betas covout;
   model remiss=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
          predprobs=(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 39.1.1–Output 39.1.5.

Output 39.1.1. 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
     Number of Observations        27
     Link Function                 Logit
     Optimization Technique        Fisher's scoring


                            Response Profile

                      Ordered                      Total
                        Value       remiss     Frequency

                            1            1             9
                            2            0            18


                       Stepwise Selection Procedure


Step  0. Intercept entered:

                          Model Convergence Status

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


                  Analysis of Maximum Likelihood Estimates

                                     Standard
      Parameter    DF    Estimate       Error    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 Not in the Model

                                       Score
                  Effect      DF    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 39.1.2. 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	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

Analysis of Effects Not in the Model

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 39.1.3. 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	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

Analysis of Effects Not in the Model

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 39.1.4. 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	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

Analysis of Effects Not in the Model

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 39.1.5. 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 39.1.1). In Step 1 (Output 39.1.2), 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 39.1.3), 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 39.1.4), 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 39.1.5.

Output 39.1.6. 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	4	0	0.00	4	4.00
2	3	0	0.03	3	2.97
3	3	0	0.34	3	2.66
4	3	1	0.65	2	2.35
5	3	0	0.84	3	2.16
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
7.1966	7	0.4087

Results of the Hosmer and Lemeshow test are shown in Output 39.1.6. There is no evidence of a lack of fit in the selected model ($p = 0.4087$).

Output 39.1.7. 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	ESTIMATE	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 39.1.7. 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 39.1.8. 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_1	IP_0	XP_1	XP_0	_LEVEL_	phat	lcl	ucl
1	1	0.80	0.83	0.66	1.9	1.100	0.996	1	1	0.72265	0.27735	0.56127	0.43873	1	0.72265	0.16892	0.97093
2	1	0.90	0.36	0.32	1.4	0.740	0.992	1	1	0.57874	0.42126	0.52539	0.47461	1	0.57874	0.26788	0.83762
3	0	0.80	0.88	0.70	0.8	0.176	0.982	0	0	0.10460	0.89540	0.12940	0.87060	1	0.10460	0.00781	0.63419
4	0	1.00	0.87	0.87	0.7	1.053	0.986	0	0	0.28258	0.71742	0.32741	0.67259	1	0.28258	0.07498	0.65683
5	1	0.90	0.75	0.68	1.3	0.519	0.980	1	1	0.71418	0.28582	0.63099	0.36901	1	0.71418	0.25218	0.94876
6	0	1.00	0.65	0.65	0.6	0.519	0.982	0	0	0.27089	0.72911	0.32731	0.67269	1	0.27089	0.05852	0.68951
7	1	0.95	0.97	0.92	1.0	1.230	0.992	1	0	0.32156	0.67844	0.27077	0.72923	1	0.32156	0.13255	0.59516
8	0	0.95	0.87	0.83	1.9	1.354	1.020	0	1	0.60723	0.39277	0.90094	0.09906	1	0.60723	0.10572	0.95287
9	0	1.00	0.45	0.45	0.8	0.322	0.999	0	0	0.16632	0.83368	0.19136	0.80864	1	0.16632	0.03018	0.56123
10	0	0.95	0.36	0.34	0.5	0.000	1.038	0	0	0.00157	0.99843	0.00160	0.99840	1	0.00157	0.00000	0.68962
11	0	0.85	0.39	0.33	0.7	0.279	0.988	0	0	0.07285	0.92715	0.08277	0.91723	1	0.07285	0.00614	0.49982
12	0	0.70	0.76	0.53	1.2	0.146	0.982	0	0	0.17286	0.82714	0.36162	0.63838	1	0.17286	0.00637	0.87206
13	0	0.80	0.46	0.37	0.4	0.380	1.006	0	0	0.00346	0.99654	0.00356	0.99644	1	0.00346	0.00001	0.46530
14	0	0.20	0.39	0.08	0.8	0.114	0.990	0	0	0.00018	0.99982	0.00019	0.99981	1	0.00018	0.00000	0.96482
15	0	1.00	0.90	0.90	1.1	1.037	0.990	0	1	0.57122	0.42878	0.64646	0.35354	1	0.57122	0.25303	0.83973
16	1	1.00	0.84	0.84	1.9	2.064	1.020	1	1	0.71470	0.28530	0.52787	0.47213	1	0.71470	0.15362	0.97189
17	0	0.65	0.42	0.27	0.5	0.114	1.014	0	0	0.00062	0.99938	0.00063	0.99937	1	0.00062	0.00000	0.62665
18	0	1.00	0.75	0.75	1.0	1.322	1.004	0	0	0.22289	0.77711	0.26388	0.73612	1	0.22289	0.04483	0.63670
19	0	0.50	0.44	0.22	0.6	0.114	0.990	0	0	0.00154	0.99846	0.00158	0.99842	1	0.00154	0.00000	0.79644
20	1	1.00	0.63	0.63	1.1	1.072	0.986	1	1	0.64911	0.35089	0.57947	0.42053	1	0.64911	0.26305	0.90555
21	0	1.00	0.33	0.33	0.4	0.176	1.010	0	0	0.01693	0.98307	0.01830	0.98170	1	0.01693	0.00029	0.50475
22	0	0.90	0.93	0.84	0.6	1.591	1.020	0	0	0.00622	0.99378	0.00652	0.99348	1	0.00622	0.00003	0.56062
23	1	1.00	0.58	0.58	1.0	0.531	1.002	1	0	0.25261	0.74739	0.15577	0.84423	1	0.25261	0.06137	0.63597
24	0	0.95	0.32	0.30	1.6	0.886	0.988	0	1	0.87011	0.12989	0.96363	0.03637	1	0.87011	0.40910	0.98481
25	1	1.00	0.60	0.60	1.7	0.964	0.990	1	1	0.93132	0.06868	0.91983	0.08017	1	0.93132	0.44114	0.99573
26	1	1.00	0.69	0.69	0.9	0.398	0.986	1	0	0.46051	0.53949	0.37688	0.62312	1	0.46051	0.16612	0.78529
27	0	1.00	0.73	0.73	0.7	0.398	0.986	0	0	0.28258	0.71742	0.32741	0.67259	1	0.28258	0.07498	0.65683

The data set pred created by the OUTPUT statement is displayed in Output 39.1.8. 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 descending;
   model remiss=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 39.1.9. 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
     Number of Observations        27
     Link Function                 Logit
     Optimization Technique        Fisher's scoring


                               Response Profile

                      Ordered                       Total
                        Value       remiss     Frequency

                            1            1             9
                            2            0            18


                          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 39.1.10. Fast Elimination Step

Step 1. Fast Backward Elimination:

Analysis of Variables Removed by Fast Backward Elimination

Effect Removed	Chi-Square	Pr > ChiSq	Residual Chi-Square	DF	Pr > Residual ChiSq
blast	0.0008	0.9768	0.0008	1	0.9768
smear	0.0951	0.7578	0.0959	2	0.9532
cell	1.5134	0.2186	1.6094	3	0.6573
temp	0.6535	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

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	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 39.1.9 and Output 39.1.10. Initially, a full model containing all six risk factors is fit to the data (Output 39.1.9). In the next step (Output 39.1.10), 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 39.1.11. 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 39.1.11. 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 classified 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 39.2. 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 39.2.1, and the estimated covariance matrix is displayed in Output 39.2.2.

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 39.2.1. 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	
Number of Observations	28	
Frequency Variable	freq	
Sum of Frequencies	208	
Link Function	Logit	
Optimization Technique	Fisher's scoring	

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

NOTE: 8 observations having zero frequencies or weights were excluded since they do not contribute to the analysis.

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

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	Chi-Square	Pr > ChiSq
Intercept	1	-7.0801	0.5624	158.4851	<.0001
Intercept2	1	-6.0249	0.4755	160.5500	<.0001
Intercept3	1	-4.9254	0.4272	132.9484	<.0001
Intercept4	1	-3.8568	0.3902	97.7087	<.0001
Intercept5	1	-2.5205	0.3431	53.9704	<.0001
Intercept6	1	-1.5685	0.3086	25.8374	<.0001
Intercept7	1	-0.0669	0.2658	0.0633	0.8013
Intercept8	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 39.2.2. Estimated Covariance Matrix

Multiple Response Cheese Tasting Experiment

The LOGISTIC Procedure

Estimated Covariance Matrix

Variable	Intercept	Intercept2	Intercept3	Intercept4	Intercept5
Intercept	0.316291	0.219581	0.176278	0.147694	0.114024
Intercept2	0.219581	0.226095	0.177806	0.147933	0.11403
Intercept3	0.176278	0.177806	0.182473	0.148844	0.114092
Intercept4	0.147694	0.147933	0.148844	0.152235	0.114512
Intercept5	0.114024	0.11403	0.114092	0.114512	0.117713
Intercept6	0.091085	0.091081	0.091074	0.091109	0.091821
Intercept7	0.057814	0.057813	0.057807	0.05778	0.057721
Intercept8	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

Variable	Intercept6	Intercept7	Intercept8	Additive1	Additive2	Additive3
Intercept	0.091085	0.057814	0.041304	-0.09419	-0.18686	-0.13565
Intercept2	0.091081	0.057813	0.041304	-0.09421	-0.18161	-0.13569
Intercept3	0.091074	0.057807	0.0413	-0.09427	-0.1687	-0.1352
Intercept4	0.091109	0.05778	0.041277	-0.09428	-0.14717	-0.13118
Intercept5	0.091821	0.057721	0.041162	-0.09246	-0.11415	-0.11207
Intercept6	0.09522	0.058312	0.041324	-0.08521	-0.09113	-0.09122
Intercept7	0.058312	0.07064	0.04878	-0.06041	-0.05781	-0.05802
Intercept8	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 39.3. 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 dummy variables are created for Treatment and one for Sex, as shown in Output 39.3.1.

Output 39.3.1. 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 III analysis of effects based on the Wald test (Output 39.3.2). 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 39.3.2. Wald Tests of Individual Effects

The LOGISTIC Procedure

Type III 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 39.3.3. The Exp(Est) column contains the exponentiated parameter estimates. 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 39.3.3. Parameter Estimates with Effect Coding

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	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. Although it is the default, the option RULE=SINGLE is explicitly specified to select one effect in each step where the selection must maintain model hierarchy. 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
                                           rule=single
                                           expb;
   run;
```

Results of the forward selection process are summarized in Output 39.3.4. The variable Treatment is selected first, followed by Age and then Sex. The results are consistent with the previous analysis (Output 39.3.2) in which the Treatment*Sex interaction and Duration are not statistically significant.

Output 39.3.4. Effects Selected into the Model

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 39.3.5 shows the Type III 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 39.3.5, 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 39.3.5. Type III Effects and Parameter Estimates with Effect Coding

Type III 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	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 statments 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 -1,
                       Treatment 0 1 -1,
                       Treatment 1 -1 0 / estimate=exp;
   contrast 'Female vs Male' Sex 1 -1 / estimate=exp;
   run;
```

Output 39.3.6. Reference Coding of CLASS Variables

The LOGISTIC Procedure

Class Level Information

		Design Variables	
Class	Value	1	2
Treatment	A	1	0
	B	0	1
	P	0	0
Sex	F	1	
	M	0	

The reference coding is shown in Output 39.3.6. The Type III analysis of effects, the parameter estimates for the reference coding, and the odds ratio estimates are displayed in Output 39.3.7. Although the parameter estimates are different (because of the different parameterizations), the "Type III Analysis of Effects" table and the "Odds Ratio" table remain the same as in Output 39.3.5. 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 39.3.7. Type III Effects and Parameter Estimates with Reference Coding

The LOGISTIC Procedure

Type III 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	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 39.3.8 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 specifed in the 'Pairwise' CONTRAST statement, there are only two degrees of freedom, and the Wald test result is identical to the Type III analysis of Treatment in Output 39.3.7. 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 39.3.7). 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 39.3.8. Results of CONTRAST Statements

The LOGISTIC Procedure

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	Lower Limit	Upper Limit
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 39.4. 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.

```
title 'Occurrence of Vaso-Constriction';
proc logistic data=vaso;
   model Response=LogRate LogVolume/influence iplots;
run;
```

Results of the model fit are shown in Output 39.4.1. 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 39.4.1. 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
Number of Observations	39
Link Function	Logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	Response	Total Frequency
1	constrict	20
2	no_constrict	19

Model Convergence Status

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

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	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 regression diagnostics produced by the INFLUENCE option are shown in Output 39.4.2–Output 39.4.7.

The values of the explanatory variables (LogRate and LogVolume) are listed for each observation (Output 39.4.2). For each diagnostic, the case number, representing the sequence number of the observation, is displayed along with the diagnostic value. Also displayed is a plot where the vertical axis represents the case number and the horizontal axis represents the value of the diagnostic statistic.

Output 39.4.2. Covariates and Pearson Residuals

Occurrence of Vaso-Constriction

The LOGISTIC Procedure

Regression Diagnostics

Case Number	Covariates LogRate	Covariates Log Volume	Pearson Residual Value	Pearson Residual (1 unit = 0.44) -8 -4 0 2 4 6 8
1	-0.1924	1.3083	0.2205	| |* |
2	0.0862	1.2528	0.1349	| * |
3	0.9163	0.2231	0.2923	| |* |
4	0.4055	-0.2877	3.5181	| | *|
5	1.1632	-0.2231	0.5287	| |* |
6	1.2528	-0.3567	0.6090	| |* |
7	-0.2877	-0.5108	-0.0328	| * |
8	0.5306	0.0953	-1.0196	| * | |
9	-0.2877	-0.1054	-0.0938	| * |
10	-0.7985	-0.1054	-0.0293	| * |
11	-0.5621	-0.2231	-0.0370	| * |
12	1.0116	-0.5978	-0.5073	| *| |
13	1.0986	-0.5108	-0.7751	| * | |
14	0.8459	0.3365	0.2559	| |* |
15	1.3218	-0.2877	0.4352	| |* |
16	0.4947	0.8329	0.1576	| * |
17	0.4700	1.1632	0.0709	| * |
18	0.3471	-0.1625	2.9062	| | * |
19	0.0583	0.5306	-1.0718	| * | |
20	0.5878	0.5878	0.2405	| |* |
21	0.6931	-0.9163	-0.1076	| * |
22	0.3075	-0.0513	-0.4193	| *| |
23	0.3001	0.3001	-1.0242	| * | |
24	0.3075	0.4055	-1.3684	| * | |
25	0.5766	0.4700	0.3347	| |* |
26	0.4055	-0.5108	-0.1595	| * |
27	0.4055	0.5878	0.3645	| |* |
28	0.6419	-0.0513	-0.8989	| * | |
29	-0.0513	0.6419	0.8981	| | * |
30	-0.9163	0.4700	-0.0992	| * |
31	-0.2877	0.9933	0.6198	| |* |
32	-3.5066	0.8544	-0.00073	| * |
33	0.6043	0.0953	-1.2062	| * | |
34	0.7885	0.0953	0.5447	| |* |
35	0.6931	0.1823	0.5404	| |* |
36	1.2030	-0.2231	0.4828	| |* |
37	0.6419	-0.0513	-0.8989	| * | |
38	0.6419	-0.2877	-0.4874	| |* |
39	0.4855	0.2624	0.7053	| | | * |

Output 39.4.3. Deviance Residuals and Hat Matrix Diagonal Elements

Regression Diagnostics

Case Number	Deviance Residual Value	Deviance Residual (1 unit = 0.28) -8 -4 0 2 4 6 8	Hat Matrix Diagonal Value	Hat Matrix Diagonal (1 unit = 0.02) 0 2 4 6 8 12 16
1	0.3082	|*|	0.0927	|*|
2	0.1899	|*|	0.0429	|*|
3	0.4049	|*|	0.0612	|*|
4	2.2775	|*|	0.0867	|*|
5	0.7021	|*|	0.1158	|*|
6	0.7943	|*|	0.1524	|*|
7	-0.0464	|*|	0.00761	|*|
8	-1.1939	|*|	0.0559	|*|
9	-0.1323	|*|	0.0342	|*|
10	-0.0414	|*|	0.00721	|*|
11	-0.0523	|*|	0.00969	|*|
12	-0.6768	|*|	0.1481	|*|
13	-0.9700	|*|	0.1628	|*|
14	0.3562	|*|	0.0551	|*|
15	0.5890	|*|	0.1336	|*|
16	0.2215	|*|	0.0402	|*|
17	0.1001	|*|	0.0172	|*|
18	2.1192	|*|	0.0954	|*|
19	-1.2368	|*|	0.1315	|*|
20	0.3353	|*|	0.0525	|*|
21	-0.1517	|*|	0.0373	|*|
22	-0.5691	|*|	0.1015	|*|
23	-1.1978	|*|	0.0761	|*|
24	-1.4527	|*|	0.0717	|*|
25	0.4608	|*|	0.0587	|*|
26	-0.2241	|*|	0.0548	|*|
27	0.4995	|*|	0.0661	|*|
28	-1.0883	|*|	0.0647	|*|
29	1.0876	|*|	0.1682	|*|
30	-0.1400	|*|	0.0507	|*|
31	0.8064	|*|	0.2459	|*|
32	-0.00103	|*|	0.000022	|*|
33	-1.3402	|*|	0.0510	|*|
34	0.7209	|*|	0.0601	|*|
35	0.7159	|*|	0.0552	|*|
36	0.6473	|*|	0.1177	|*|
37	-1.0883	|*|	0.0647	|*|
38	-0.6529	|*|	0.1000	|*|
39	0.8987	|*|	0.0531	|*|

Output 39.4.4. DFBETA for Intercept and DFBETA for LogRate

Regression Diagnostics

Case Number	Intercept DfBeta Value	(1 unit = 0.13) -8 -4 0 2 4 6 8	LogRate DfBeta Value	(1 unit = 0.12) -8 -4 0 2 4 6 8
1	-0.0165	| * |	0.0193	| * |
2	-0.0134	| * |	0.0151	| * |
3	-0.0492	| * |	0.0660	| |* |
4	1.0734	| | * |	-0.9302	|* | |
5	-0.0832	| *| |	0.1411	| |* |
6	-0.0922	| *| |	0.1710	| |* |
7	-0.00280	| * |	0.00274	| * |
8	-0.1444	| *| |	0.0613	| |* |
9	-0.0178	| * |	0.0173	| * |
10	-0.00245	| * |	0.00246	| * |
11	-0.00361	| * |	0.00358	| * |
12	-0.1173	| *| |	0.0647	| |* |
13	-0.0931	| *| |	-0.00946	| * |
14	-0.0414	| * |	0.0538	| * |
15	-0.0940	| *| |	0.1408	| |* |
16	-0.0198	| * |	0.0234	| * |
17	-0.00630	| * |	0.00701	| * |
18	0.9595	| | * |	-0.8279	| * | |
19	-0.2591	| * | |	0.2024	| | * |
20	-0.0331	| * |	0.0421	| * |
21	-0.0180	| * |	0.0158	| * |
22	-0.1449	| *| |	0.1237	| |* |
23	-0.1961	| *| |	0.1275	| |* |
24	-0.1281	| *| |	0.0410	| * |
25	-0.0403	| * |	0.0570	| * |
26	-0.0366	| * |	0.0329	| * |
27	-0.0327	| * |	0.0496	| * |
28	-0.1423	| *| |	0.0617	| |* |
29	0.2367	| | * |	-0.1950	| * | |
30	-0.0224	| * |	0.0227	| * |
31	0.1165	| |* |	-0.0996	| *| |
32	-3.22E-6	| * |	3.405E-6	| * |
33	-0.0882	| *| |	-0.0137	| * |
34	-0.0425	| * |	0.0877	| |* |
35	-0.0340	| * |	0.0755	| |* |
36	-0.0867	| *| |	0.1381	| |* |
37	-0.1423	| *| |	0.0617	| |* |
38	-0.1395	| *| |	0.1032	| |* |
39	0.0326	| * |	0.0190	| * |

Output 39.4.5. DFBETA for LogVolume and Confidence Interval Displacement C

Regression Diagnostics

Case Number	Log Volume DfBeta Value	(1 unit = 0.13) -8 -4 0 2 4 6 8	Confidence Interval Displacement C Value	(1 unit = 0.08) 0 2 4 6 8 12 16
1	0.0556	| * |	0.00548	|* |
2	0.0261	| * |	0.000853	|* |
3	0.0589	| * |	0.00593	|* |
4	-1.0180	|* |	1.2873	| *|
5	0.0583	| * |	0.0414	| * |
6	0.0381	| * |	0.0787	| * |
7	0.00265	| * |	8.321E-6	|* |
8	0.0570	| * |	0.0652	| * |
9	0.0153	| * |	0.000322	|* |
10	0.00211	| * |	6.256E-6	|* |
11	0.00319	| * |	0.000014	|* |
12	0.1651	| |* |	0.0525	| * |
13	0.1775	| |* |	0.1395	| * |
14	0.0527	| * |	0.00404	|* |
15	0.0643	| |* |	0.0337	|* |
16	0.0307	| * |	0.00108	|* |
17	0.00914	| * |	0.000089	|* |
18	-0.8477	| * |	0.9845	| * |
19	-0.00488	| * |	0.2003	| * |
20	0.0518	| * |	0.00338	|* |
21	0.0208	| * |	0.000465	|* |
22	0.1179	| |* |	0.0221	|* |
23	0.0357	| * |	0.0935	| * |
24	-0.1004	| *| |	0.1558	| * |
25	0.0708	| |* |	0.00741	|* |
26	0.0373	| * |	0.00156	|* |
27	0.0788	| |* |	0.0101	|* |
28	0.1025	| |* |	0.0597	| * |
29	0.0286	| * |	0.1961	| * |
30	0.0159	| * |	0.000554	|* |
31	0.1322	| |* |	0.1661	| * |
32	2.48E-6	| * |	1.18E-11	|* |
33	-0.00216	| * |	0.0824	| * |
34	0.0671	| |* |	0.0202	|* |
35	0.0711	| |* |	0.0180	|* |
36	0.0631	| * |	0.0352	|* |
37	0.1025	| |* |	0.0597	| * |
38	0.1397	| |* |	0.0293	|* |
39	0.0489	| * |	0.0295	|* |

Output 39.4.6. Confidence Interval Displacement CBAR

Regression Diagnostics

Case Number	Confidence Interval Displacement CBar Value	(1 unit = 0.07) 0 2 4 6 8 12 16	Delta Deviance Value	(1 unit = 0.4) 0 2 4 6 8 12 16
1	0.00497	|* |	0.1000	|* |
2	0.000816	|* |	0.0369	|* |
3	0.00557	|* |	0.1695	|* |
4	1.1756	| *|	6.3626	| *|
5	0.0366	|* |	0.5296	| * |
6	0.0667	| * |	0.6976	| * |
7	8.258E-6	|* |	0.00216	|* |
8	0.0616	| * |	1.4870	| * |
9	0.000311	|* |	0.0178	|* |
10	6.211E-6	|* |	0.00172	|* |
11	0.000013	|* |	0.00274	|* |
12	0.0447	| * |	0.5028	| * |
13	0.1168	| * |	1.0577	| * |
14	0.00382	|* |	0.1307	|* |
15	0.0292	|* |	0.3761	| * |
16	0.00104	|* |	0.0501	|* |
17	0.000088	|* |	0.0101	|* |
18	0.8906	| * |	5.3817	| * |
19	0.1740	| * |	1.7037	| * |
20	0.00320	|* |	0.1156	|* |
21	0.000448	|* |	0.0235	|* |
22	0.0199	|* |	0.3437	| * |
23	0.0864	| * |	1.5212	| * |
24	0.1447	| * |	2.2550	| * |
25	0.00698	|* |	0.2193	| * |
26	0.00147	|* |	0.0517	|* |
27	0.00941	|* |	0.2589	| * |
28	0.0559	| * |	1.2404	| * |
29	0.1631	| * |	1.3460	| * |
30	0.000526	|* |	0.0201	|* |
31	0.1253	| * |	0.7755	| * |
32	1.18E-11	|* |	1.065E-6	|* |
33	0.0782	| * |	1.8744	| * |
34	0.0190	|* |	0.5387	| * |
35	0.0170	|* |	0.5295	| * |
36	0.0311	|* |	0.4501	| * |
37	0.0559	| * |	1.2404	| * |
38	0.0264	|* |	0.4526	| * |
39	0.0279	|* |	0.8355	| * |

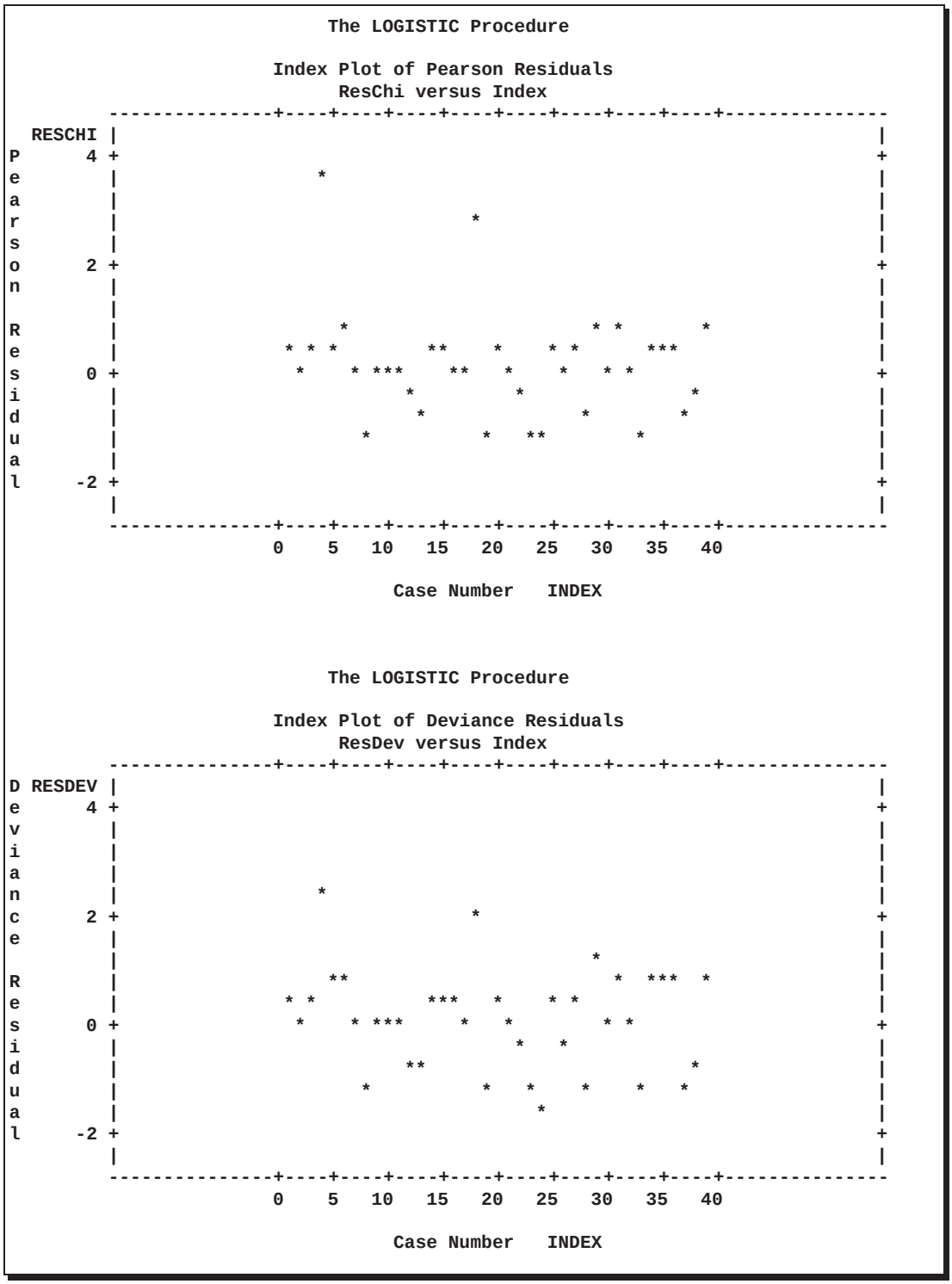
Output 39.4.7. Changes in Deviance and Pearson χ^2

Regression Diagnostics

Delta Chi-Square

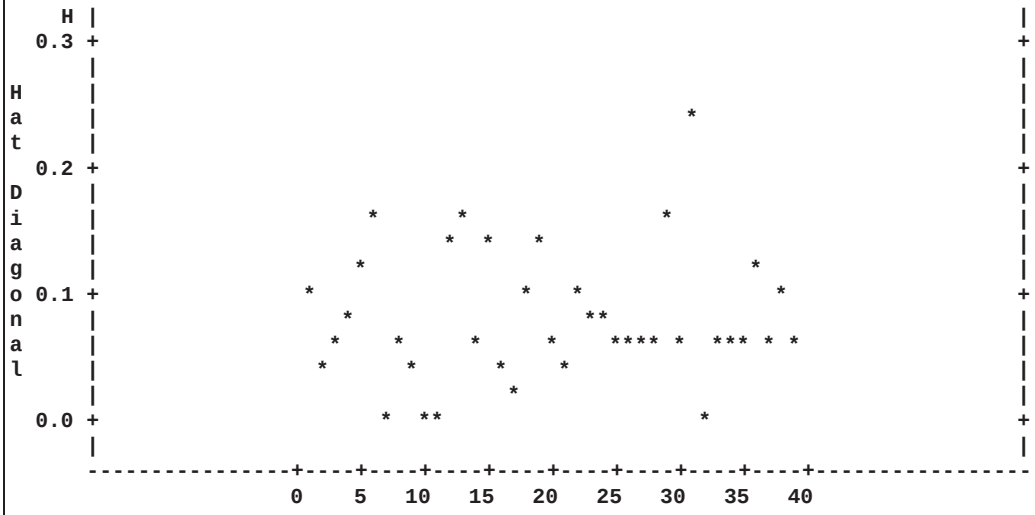
Case Number	Value	(1 unit = 0.85) 0 2 4 6 8 12 16
1	0.0536	|* |
2	0.0190	|* |
3	0.0910	|* |
4	13.5523	| *|
5	0.3161	|* |
6	0.4376	| * |
7	0.00109	|* |
8	1.1011	| * |
9	0.00911	|* |
10	0.000862	|* |
11	0.00138	|* |
12	0.3021	|* |
13	0.7175	| * |
14	0.0693	|* |
15	0.2186	|* |
16	0.0259	|* |
17	0.00511	|* |
18	9.3363	| * |
19	1.3227	| * |
20	0.0610	|* |
21	0.0120	|* |
22	0.1956	|* |
23	1.1355	| * |
24	2.0171	| * |
25	0.1190	|* |
26	0.0269	|* |
27	0.1423	|* |
28	0.8639	| * |
29	0.9697	| * |
30	0.0104	|* |
31	0.5095	| * |
32	5.324E-7	|* |
33	1.5331	| * |
34	0.3157	|* |
35	0.3091	|* |
36	0.2641	|* |
37	0.8639	| * |
38	0.2639	|* |
39	0.5254	| * |

The index plots produced by the IPLOTS option are essentially the same plots as those produced by the INFLUENCE option with a 90-degree rotation and perhaps on a more refined scale. 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.


The LOGISTIC Procedure
Index Plot of Pearson Residuals
ResChi versus Index
RESCHI
Pearson Residual
4
2
0
-2
0 5 10 15 20 25 30 35 40
Case Number INDEX
The LOGISTIC Procedure
Index Plot of Deviance Residuals
ResDev versus Index
RESDEV
Deviance Residual
4
2
0
-2
0 5 10 15 20 25 30 35 40
Case Number INDEX

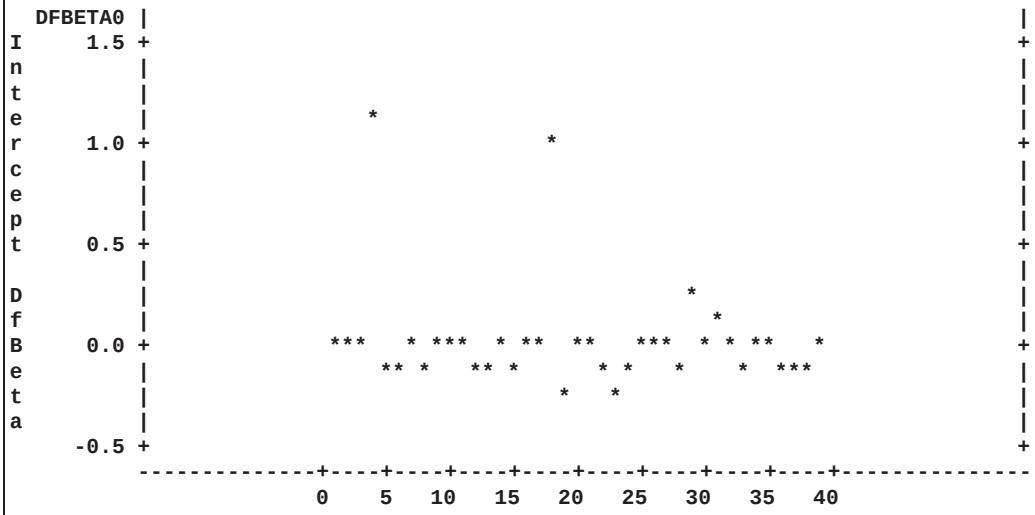
The LOGISTIC Procedure

Index Plot of Diagonal Elements of the Hat Matrix
H versus Index



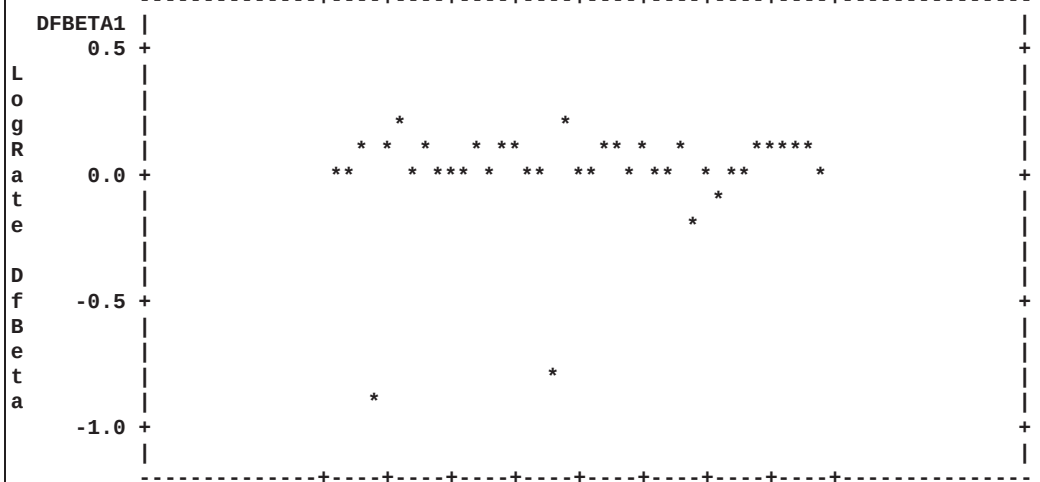
The LOGISTIC Procedure

Index Plot of Standardized Changes in the Estimate for Intercept
DfBeta0 versus Index



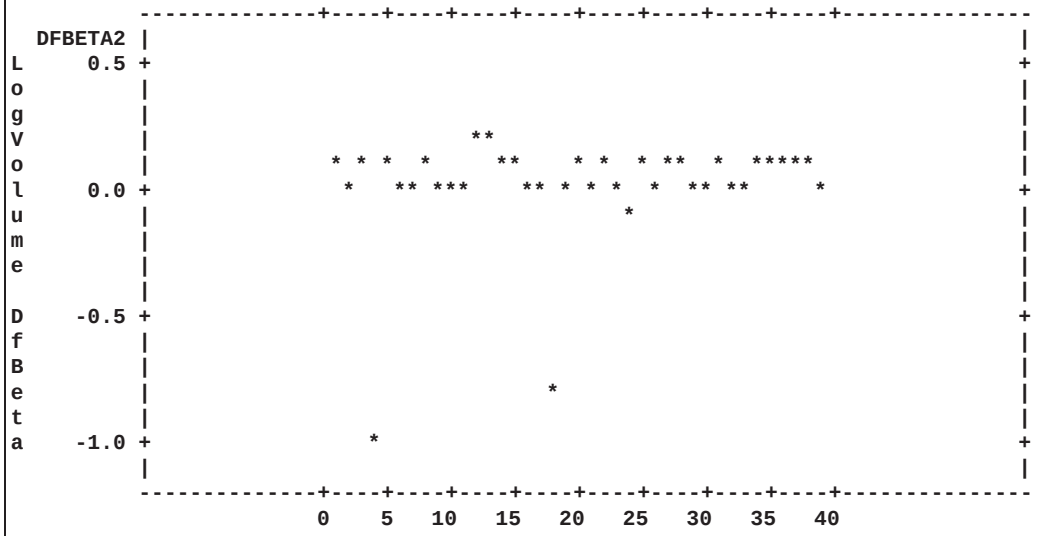
The LOGISTIC Procedure

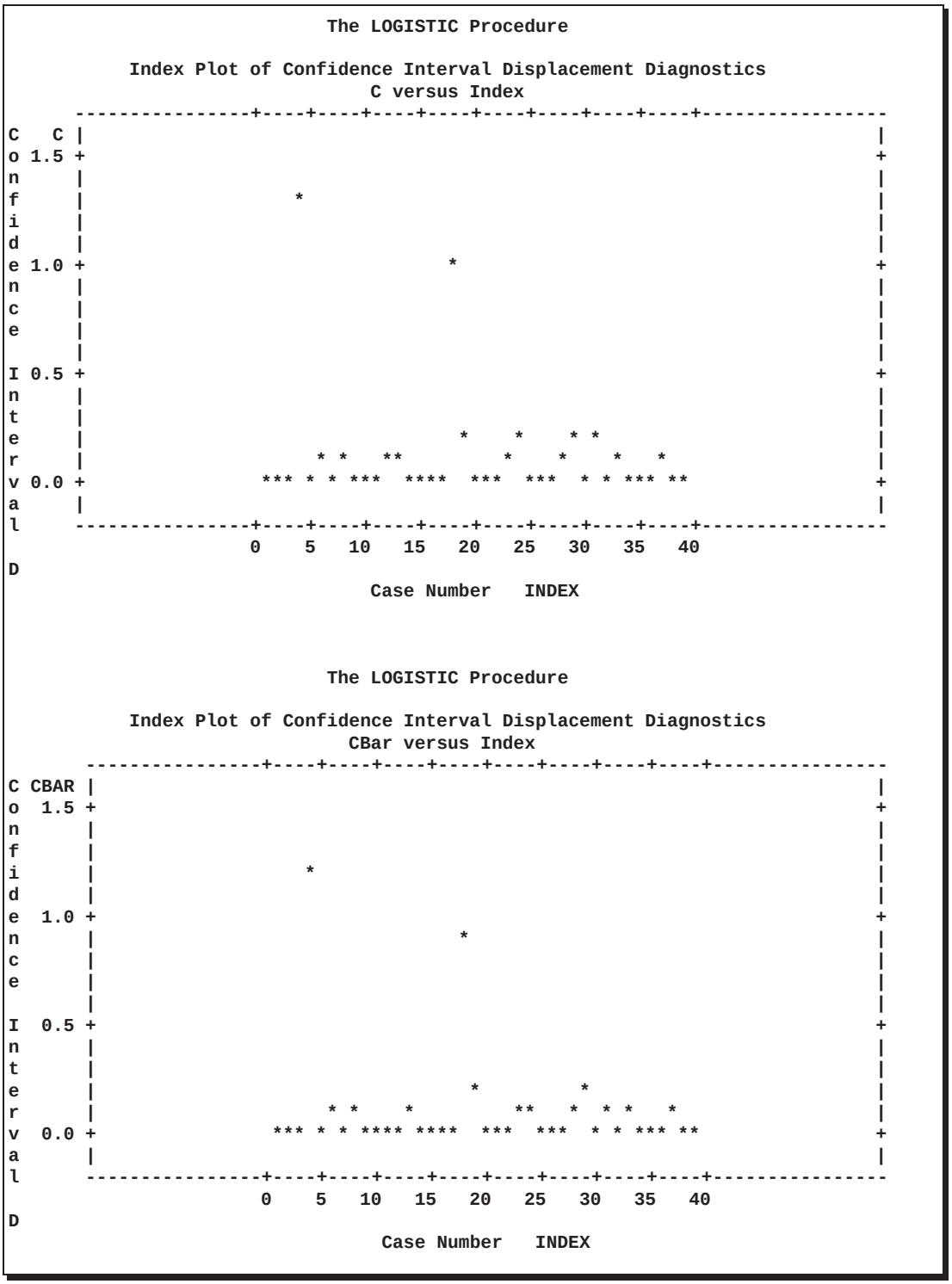
Index Plot of Standardized Changes in the Estimate for LogRate
DfBeta1 versus Index

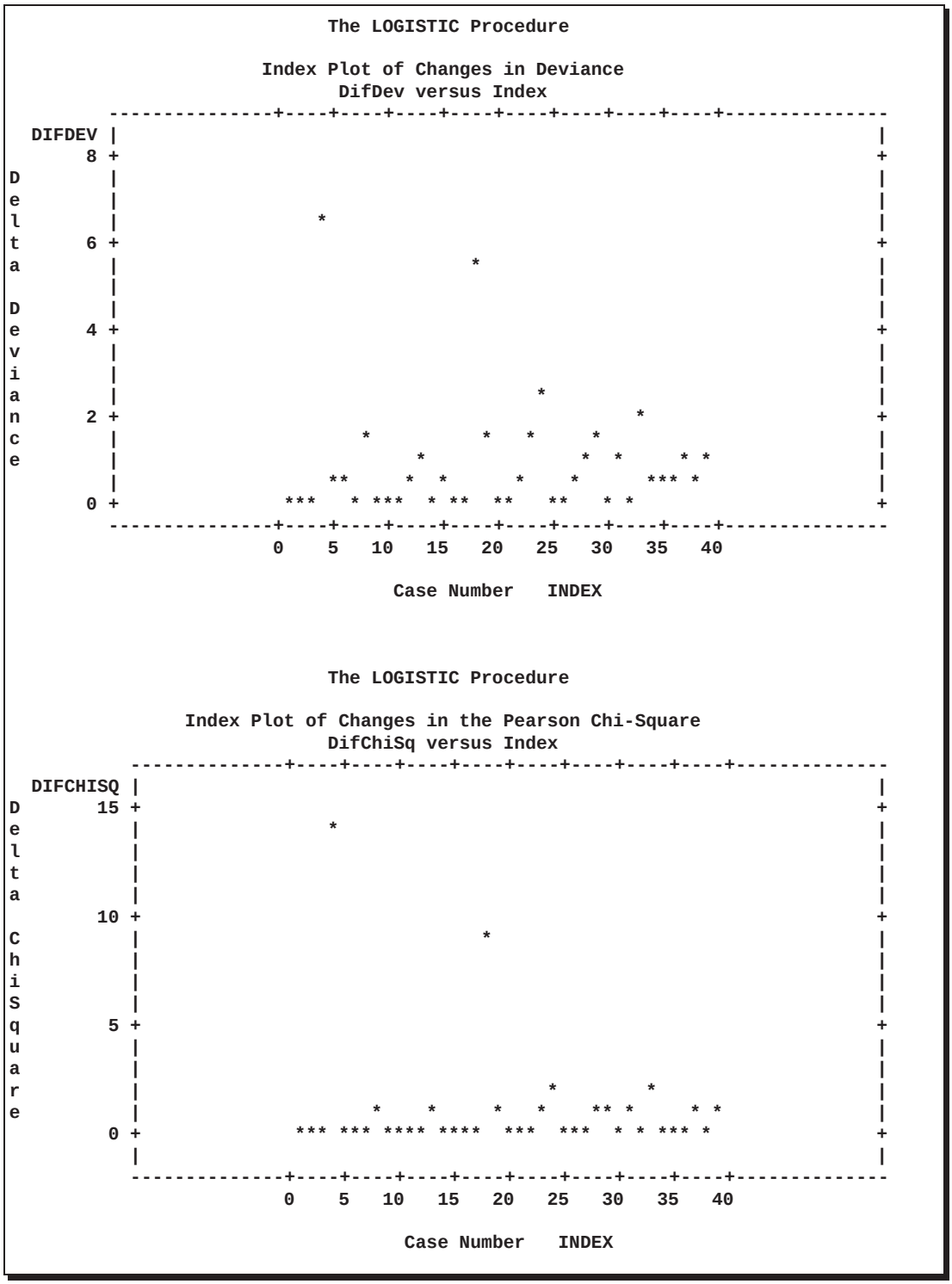


The LOGISTIC Procedure

Index Plot of Standardized Changes in the Estimate for LogVolume
DfBeta2 versus Index




The LOGISTIC Procedure
Index Plot of Confidence Interval Displacement Diagnostics
C versus Index
C
Confidence Interval Displacement
C
1.5
1.0
0.5
0.0
0
5
10
15
20
25
30
35
40
Case Number
INDEX
The LOGISTIC Procedure
Index Plot of Confidence Interval Displacement Diagnostics
CBar versus Index
CBAR
Confidence Interval Displacement
CBAR
1.5
1.0
0.5
0.0
0
5
10
15
20
25
30
35
40
Case Number
INDEX


The LOGISTIC Procedure
Index Plot of Changes in Deviance
DifDev versus Index
DIFDEV
Delta Deviance
Case Number INDEX
The LOGISTIC Procedure
Index Plot of Changes in the Pearson Chi-Square
DifChiSq versus Index
DIFCHISQ
Delta ChiSquare
Case Number INDEX

The index plots of the Pearson residuals and the deviance residuals 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 suggests that case 31 is an extreme point in the design space. The index plots of DFBETAS indicate that case 4 and case 18 are causing instability in all three parameter estimates. The other four index plots also point to these two cases as having a large impact on the coefficients and goodness of fit.

Example 39.5. Stratified Sampling

Consider the hypothetical example in Fleiss (1981, pp. 6–7) in which a test is applied to a sample of 1000 people known to have a disease and to another sample of 1000 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 order=data data=Screen;
   freq Count;
   model Disease=Test / pevent=.5 .01 ctable pprob=.5;
run;
```

The ORDER=DATA option causes the Disease level of the first observation in the input data set to be the event. So, 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 39.5.1.

Output 39.5.1. False Positive and False Negative Rates

The LOGISTIC Procedure

Classification Table

Prob Event	Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages 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 39.6. 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, and calculates the normal confidence intervals for the regression parameters. 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 code is 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
;
```

```
proc logistic data=Data1;
   model disease/n=age / scale=none
                         clparm=wald
                         clodds=pl
                         rsquare
                         outroc=roc1;
   units age=10;
run;
```

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 39.6.1 and Output 39.6.2.

Output 39.6.1. Deviance and Pearson Goodness-of-Fit Analysis

The LOGISTIC Procedure

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	DF	Value	Value/DF	Pr > ChiSq
Deviance	4	7.7756	1.9439	0.1002
Pearson	4	6.6020	1.6505	0.1585

Number of events/trials observations: 6

Output 39.6.2. 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	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

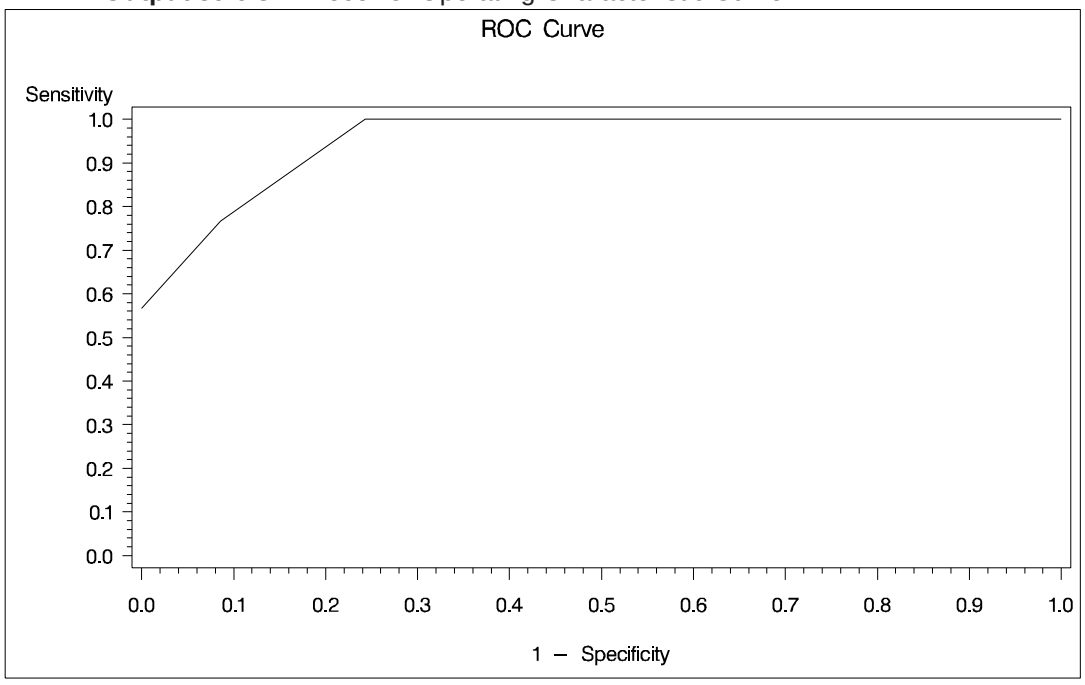
Profile Likelihood Confidence Interval for Adjusted Odds Ratios

Effect	Unit	Estimate	95% Confidence Limits	
age	10.0000	7.892	3.881	21.406

The ROC curve is plotted by the GPLOT procedure, and the plot is shown in Output 39.6.3.

```
symbol1 i=join v=none c=blue;
proc gplot data=roc1;
   title 'ROC Curve';
   plot _sensit_*_1mspec_=1 / vaxis=0 to 1 by .1 cframe=ligr;
run;
```

Output 39.6.3. Receiver Operating Characteristic Curve



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.

Example 39.7. 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 39.7.1. Full Model Fit

The LOGISTIC Procedure

Model Information

Data Set	WORK.A
Response Variable	Y
Number of Response Levels	2
Number of Observations	8
Frequency Variable	F
Sum of Frequencies	528
Link Function	Logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	Y	Total Frequency
1	1	191
2	2	337

Model Convergence Status

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

Model Fit Statistics

Criterion	Intercept Only	Intercept and 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

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	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 39.7.1. 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 39.7.2. Reduced Model Fit

The LOGISTIC Procedure

Model Information

Data Set	WORK.A
Response Variable	Y
Number of Response Levels	2
Number of Observations	8
Frequency Variable	F
Sum of Frequencies	528
Link Function	Logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	Y	Total Frequency
1	1	191
2	2	337

Model Convergence Status

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

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	DF	Value	Value/DF	Pr > ChiSq
Deviance	2	0.3541	0.1770	0.8377
Pearson	2	0.3531	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

```
              Analysis of Maximum Likelihood Estimates

                              Standard
Parameter    DF    Estimate       Error    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 39.7.2) 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 39.8. 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 39.8.1. 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        DF          Value     Value/DF     Pr > ChiSq

   Deviance         16        68.3465       4.2717         <.0001
   Pearson          16        66.7617       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
   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 39.8.1. 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 39.8.2. 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
Number of Observations	20
Weight Variable	1 / (1 + 0.075941 * (n - 1))
Sum of Weights	198.32164573
Link Function	Logit
Optimization Technique	Fisher's scoring

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	DF	Value	Value/DF	Pr > ChiSq
Deviance	16	16.4402	1.0275	0.4227
Pearson	16	16.0000	1.0000	0.4530

Number of events/trials observations: 20

NOTE: Since the Williams method was used to accomodate 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

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

                              Standard
Parameter    DF    Estimate       Error    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 39.8.2. 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 39.8.3. Reduced Model with Overdispersion Controlled

```
            Analysis of Maximum Likelihood Estimates

                              Standard
Parameter    DF    Estimate       Error    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 39.8.3. Soil condition remains a significant factor ($p = 0.0064$) for the seed germination.

Example 39.9. Conditional Logistic Regression for Matched Pairs Data

In matched case-control studies, conditional logistic regression is used to investigate the relationship between an outcome of being a case or a control and a set of prognostic factors. When each matched set consists of a single case and a single control, 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 case and control, 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).

The data in this example are a subset of the data from the Los Angeles Study of the Endometrial Cancer Data in Breslow and Days (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.

Before PROC LOGISTIC is used for the logistic regression analysis, 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 Data1;
   drop id1 gall1 hyper1;
   retain id1 gall1 hyper1 0;
   input ID Outcome Gall Hyper @@ ;
   if (ID = id1) then do;
      Gall=gall1-Gall; Hyper=hyper1-Hyper;
      output;
   end;
   else do;
      id1=ID; gall1=Gall; hyper1=Hyper;
   end;
   datalines;
 1  1 0  0  1  0  0  0   2  1 0  0  2  0  0  0
 3  1 0  1  3  0  0  1   4  1 0  0  4  0  1  0
 5  1 1  0  5  0  0  1   6  1 0  1  6  0  0  0
 7  1 1  0  7  0  0  0   8  1 1  1  8  0  0  1
 9  1 0  0  9  0  0  0  10  1 0  0 10  0  0  0
11  1 1  0 11  0  0  0  12  1 0  0 12  0  0  1
```

```
13  1 1  0 13  0  0  1  14  1 1  0 14  0  1  0
15  1 1  0 15  0  0  1  16  1 0  1 16  0  0  0
17  1 0  0 17  0  1  1  18  1 0  0 18  0  1  1
19  1 0  0 19  0  0  1  20  1 0  1 20  0  0  0
21  1 0  0 21  0  1  1  22  1 0  1 22  0  0  1
23  1 0  1 23  0  0  0  24  1 0  0 24  0  0  0
25  1 0  0 25  0  0  0  26  1 0  0 26  0  0  1
27  1 1  0 27  0  0  1  28  1 0  0 28  0  0  1
29  1 1  0 29  0  0  0  30  1 0  1 30  0  0  0
31  1 0  1 31  0  0  0  32  1 0  1 32  0  0  0
33  1 0  1 33  0  0  0  34  1 0  0 34  0  0  0
35  1 1  1 35  0  1  1  36  1 0  0 36  0  0  1
37  1 0  1 37  0  0  0  38  1 0  1 38  0  0  1
39  1 0  1 39  0  0  1  40  1 0  1 40  0  0  0
41  1 0  0 41  0  0  0  42  1 0  1 42  0  1  0
43  1 0  0 43  0  0  1  44  1 0  0 44  0  0  0
45  1 1  0 45  0  0  0  46  1 0  0 46  0  0  0
47  1 1  1 47  0  0  0  48  1 0  1 48  0  0  0
49  1 0  0 49  0  0  0  50  1 0  1 50  0  0  1
51  1 0  0 51  0  0  0  52  1 0  1 52  0  0  1
53  1 0  1 53  0  0  0  54  1 0  1 54  0  0  0
55  1 1  0 55  0  0  0  56  1 0  0 56  0  0  0
57  1 1  1 57  0  1  0  58  1 0  0 58  0  0  0
59  1 0  0 59  0  0  0  60  1 1  1 60  0  0  0
61  1 1  0 61  0  1  0  62  1 0  1 62  0  0  0
63  1 1  0 63  0  0  0
;
```

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. 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=PL is specified, PROC LOGISTIC computes a 95% profile likelihood confidence interval for the odds ratio for each predictor variable.

```
proc logistic data=Data1;
   model outcome=Gall / noint CLODDS=PL;
run;

proc logistic data=Data1;
   model outcome=Gall Hyper / noint CLODDS=PL;
run;
```

Results from the two conditional logistic analyses are shown in Output 39.9.1 and Output 39.9.2. 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 39.9.1. Conditional Logistic Regression (Gall as risk factor)

```
                          The LOGISTIC Procedure

                            Model Information

                Data Set                      WORK.DATA1
                Response Variable             Outcome
                Number of Response Levels     1
                Number of Observations        63
                Link Function                 Logit
                Optimization Technique        Fisher's scoring


                         Response Profile

                Ordered                          Total
                  Value      Outcome          Frequency

                      1            0                 63


                      Model Convergence Status

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


                      Model Fit Statistics

                                    Without         With
                   Criterion      Covariates     Covariates

                   AIC                87.337         85.654
                   SC                 87.337         87.797
                   -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

                                       Standard
      Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq

      Gall          1      0.9555      0.5262        3.2970        0.0694

NOTE: Since there is only one response level, measures of association between
      the observed and predicted values were not calculated.


         Profile Likelihood Confidence Interval for Adjusted Odds Ratios

           Effect         Unit     Estimate     95% Confidence Limits

           Gall         1.0000        2.600        0.981        8.103
```

Output 39.9.2. Conditional Logistic Regression (Gall and Hyper as risk factors)

The LOGISTIC Procedure

Model Information

Data Set	WORK.DATA1
Response Variable	Outcome
Number of Response Levels	1
Number of Observations	63
Link Function	Logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	Outcome	Total Frequency
1	0	63

Model Convergence Status

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

Model Fit Statistics

Criterion	Without Covariates	With Covariates
AIC	87.337	86.788
SC	87.337	91.074
-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	Chi-Square	Pr > ChiSq
Gall	1	0.9704	0.5307	3.3432	0.0675
Hyper	1	0.3481	0.3770	0.8526	0.3558

NOTE: Since there is only one response level, measures of association between the observed and predicted values were not calculated.

Profile Likelihood Confidence Interval for Adjusted Odds Ratios

Effect	Unit	Estimate	95% Confidence Limits	
Gall	1.0000	2.639	0.987	8.299
Hyper	1.0000	1.416	0.682	3.039

In the first model, where Gall is the only predictor variable (Output 39.9.1), the odds ratio estimate for Gall is 2.60, which is an estimate of the relative risk for gall bladder disease. A 95% confidence interval for this relative risk is (0.981, 8.103).

In the second model, where both Gall and Hyper are present (Output 39.9.2), 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 (0.987, 8.299). 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 not statistically significant. The 95% profile likelihood confidence interval for the odds ratio for Hyper is (0.682, 3.039), which contains unity.

Example 39.10. 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 et al. 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

$$\gamma = 1 - \mathrm{e}^{-\mu}$$

The following SAS statements create the data set sero, which contains the results of a serological survey of malarial infection. Individuals of nine age groups 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

$$\log(-\log(1-p_i)) = \log(u) + \log(A_i)$$

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 39.10.1. 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
Number of Observations       9
Offset Variable              X                      Log of Midpoint of Age Range
Link Function                Complementary log-log
Optimization Technique       Fisher's scoring


                               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


             Deviance and Pearson Goodness-of-Fit Statistics

        Criterion        DF          Value     Value/DF     Pr > ChiSq

        Deviance          8        41.5032       5.1879         <.0001
        Pearson           8        50.6883       6.3360         <.0001

                Number of events/trials observations: 9


                Analysis of Maximum Likelihood Estimates

                                    Standard
     Parameter    DF    Estimate       Error    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 39.10.1. 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

$$\widehat{\gamma} = 1 - \mathrm{e}^{-\widehat{\mu}} = 1 - \mathrm{e}^{-\mathrm{e}^{\widehat{\beta}_0}} = 1 - \mathrm{e}^{-\mathrm{e}^{-4.6605}} = 0.00942$$

The 95% confidence interval for γ, obtained by back-transforming the 95% confidence interval for β_0, is (0.0082, 0.0011); 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 SAS statements, nine dummy variables (agegrp1–agegrp9) are created as the design variables for the age groups. PROC LOGISTIC is invoked to fit a complementary log-log model that contains agegrp1–agegrp9 as the only explanatory variables with no intercept term and with X=log(A) as an offset term. Note that $\log(\mu_i)$ is the regression parameter associated with agegrp$i$.

```
data two;
   array agegrp(9) agegrp1-agegrp9 (0 0 0 0 0 0 0 0 0);
   set sero;
   agegrp[group]=1;
   output;
   agegrp[group]=0;
run;
proc logistic data=two;
   model R/N=agegrp1-agegrp9 / offset=X
                               noint
                               link=cloglog
                               clparm=pl;
   title 'Infectious Rates and 95% Confidence Intervals';
run;
```

Output 39.10.2. Modeling Separate Risk of Infection

Infectious Rates and 95% Confidence Intervals

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
agegrp1	1	-3.1048	0.3536	77.0877	<.0001
agegrp2	1	-4.4542	0.4083	119.0164	<.0001
agegrp3	1	-4.2769	0.2358	328.9593	<.0001
agegrp4	1	-4.7761	0.2674	319.0600	<.0001
agegrp5	1	-4.7165	0.2238	443.9920	<.0001
agegrp6	1	-4.5012	0.1606	785.1350	<.0001
agegrp7	1	-5.4252	0.2296	558.1114	<.0001
agegrp8	1	-4.9987	0.2008	619.4666	<.0001
agegrp9	1	-4.1965	0.1559	724.3157	<.0001
X	1	1.0000	0	.	.

Profile Likelihood Confidence Interval for Parameters

Parameter	Estimate	95% Confidence Limits	
agegrp1	-3.1048	-3.8880	-2.4833
agegrp2	-4.4542	-5.3769	-3.7478
agegrp3	-4.2769	-4.7775	-3.8477
agegrp4	-4.7761	-5.3501	-4.2940
agegrp5	-4.7165	-5.1896	-4.3075
agegrp6	-4.5012	-4.8333	-4.2019
agegrp7	-5.4252	-5.9116	-5.0063
agegrp8	-4.9987	-5.4195	-4.6289
agegrp9	-4.1965	-4.5164	-3.9037

Table 39.3. Infection Rate in One Year

	Number Infected per 1000 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

Results of fitting the model for separate risk of infection are shown in Output 39.10.2. For the first age group, the point estimate of $\log(\mu_1)$ is -3.1048. This translates into an infection rate of $1 - \exp(-\exp(-3.1048)) = 0.0438$. A 95% confidence interval for the 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. Table 39.3 on page 2034 shows the estimated infection rate in one year's time for each age group. Note that the infection rate for the first age group is high compared to the other age groups.

Example 39.11. 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 $\boldsymbol{x}_i$. The discrete-time hazard rate λ_{it} is defined as

$$\lambda_{it} = \Pr(T_i = t \mid T_i \geq t, \boldsymbol{x}_i), \quad t = 1, 2, \ldots$$

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

$$\lambda_{ij} = 1 - \exp(-\exp(\alpha_j + \boldsymbol{\beta}' \boldsymbol{x}_i))$$

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 $\boldsymbol{x}_i = \mathbf{0}$. 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 39.11.1 shows an observation in the beetles data set with time=3, and Output 39.11.2 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 39.11.1. An Observation with Time=3 in Data Set Beetles

Obs	time	conc	freq	sex
17	3	0.2	10	1

Output 39.11.2. Corresponding Beetle-day Observations in Days

Obs	time	conc	freq	sex	day	y	day1	day2	day3	day4	day5	day6	day7	day8	day9	day10	day11	day12	day13
25	3	0.2	10	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
26	3	0.2	10	1	2	0	0	1	0	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	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. Since the values of y are coded 0 and 1, specifying the DESCENDING 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 descending outest=est1;
   model y= day1-day13 sex conc / noint link=cloglog
                                  technique=newton;
   freq freq;
run;
```

Output 39.11.3. Parameter Estimates for the Grouped Proportional Hazards Model

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	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 39.11.3. 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

$$\hat{S}(t) = [\hat{S}_0(t)]^{\exp(-0.5651x_1 + 3.0918x_2)}$$

The following DATA step computes the survivor curves for male and female flour-beetles exposed to the insecticide of concentrations .20 mg/cm^2 and .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 39.11.4.

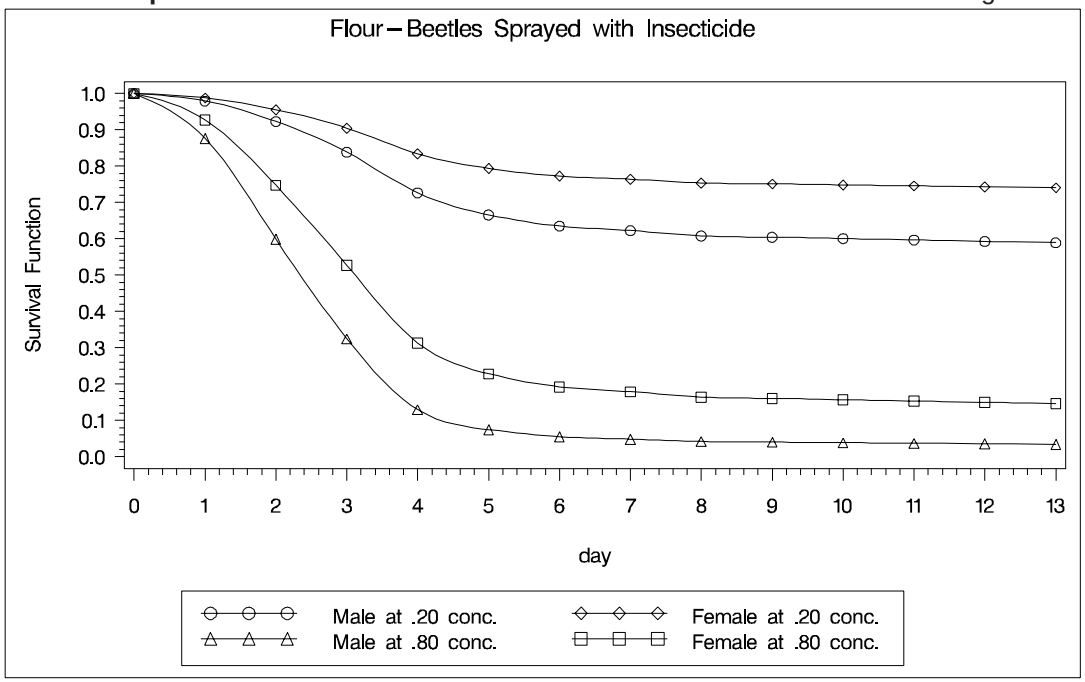
```
data one (keep=day survival element s_m20 s_f20 s_m80 s_f80);
   array dd day1-day13;
   array sc[4] m20 f20 m80 f80;
   array s_sc[4] s_m20 s_f20 s_m80 s_f80 (1 1 1 1);
   set est1;
   m20= exp(sex + .20 * conc);
   f20= exp(2 * sex + .20 * conc);
   m80= exp(sex + .80 * conc);
   f80= exp(2 * sex + .80 * conc);
   survival=1;
   day=0;
   output;
   do over dd;
      element= exp(-exp(dd));
      survival= survival * element;
      do i=1 to 4;
        s_sc[i] = survival ** sc[i];
      end;
      day + 1;
      output;
   end;
   label s_m20= 'Male at .20 conc.'
         s_m80= 'Male at .80 conc.'
         s_f20= 'Female at .20 conc.'
         s_f80= 'Female at .80 conc.';
run;

title1 'Flour-Beetles Sprayed with Insecticide';
legend1 label=none frame cframe=ligr cborder=black
        position=center value=(justify=center);
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 cframe=ligr;
   symbol1  v=dot i=spline c=black  height=.8;
   symbol2  v=dot i=spline c=red    height=.8;
   symbol3  v=dot i=spline c=blue   height=.8;
   symbol4  v=dot i=spline c=yellow 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 39.11.4. Predicted Survival at Concentrations of 0.20 and 0.80 mg/cm^2



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.

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.

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. and Days W. (1980), *Statistical Methods in Cancer Research, Volume 1–The Analysis of Case-Control Studies*, Lyon: IARC Scientific Publication No. 32.

Breslow, N.E. (1982), "Covariance Adjustment of Relative-Risk Estimates in Matched Studies," *Biometrics*, 38, 661–672.

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. 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.

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.

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

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.

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.

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. (1995), *Categorical Data Analysis Using the SAS System*, Cary, NC: SAS Institute Inc.

Venzon, D.J. and Moolgavkar, S.H. (1988), "A Method for Computing Profile-Likelihood Based Confidence Intervals," *Applied Statistics*, 37, 87–94.

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.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/STAT® User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999.

SAS/STAT® User's Guide, Version 8

ISBN 1–58025–494–2

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, October 1999