

# Chapter 49

## The PROBIT Procedure

### Chapter Table of Contents

---

<b>OVERVIEW</b> . . . . .	2607
<b>GETTING STARTED</b> . . . . .	2608
Estimating the Natural Response Threshold Parameter . . . . .	2608
<b>SYNTAX</b> . . . . .	2612
PROC PROBIT Statement . . . . .	2613
BY Statement . . . . .	2616
CLASS Statement . . . . .	2616
MODEL Statement . . . . .	2617
OUTPUT Statement . . . . .	2620
WEIGHT Statement . . . . .	2621
<b>DETAILS</b> . . . . .	2621
Missing Values . . . . .	2621
Response Level Ordering . . . . .	2621
Computational Method . . . . .	2622
Distributions . . . . .	2624
Model Specification . . . . .	2624
Lack of Fit Tests . . . . .	2625
Tolerance Distribution . . . . .	2626
Inverse Confidence Limits . . . . .	2626
OUTEST= Data Set . . . . .	2628
Displayed Output . . . . .	2628
ODS Table Names . . . . .	2630
<b>EXAMPLES</b> . . . . .	2630
Example 49.1 Dosage Levels . . . . .	2630
Example 49.2 Multilevel Response . . . . .	2639
Example 49.3 Logistic Regression . . . . .	2643
<b>REFERENCES</b> . . . . .	2646



# Chapter 49

## The PROBIT Procedure

---

### Overview

The PROBIT procedure calculates maximum likelihood estimates of regression parameters and the natural (or threshold) response rate for quantal response data from biological assays or other discrete event data. This includes probit, logit, ordinal logistic, and extreme value (or gompit) regression models.

Probit analysis developed from the need to analyze qualitative (dichotomous or polytomous) dependent variables within the regression framework. Many response variables are binary by nature (yes/no), while others are measured ordinally rather than continuously (degree of severity). Ordinary least squares (OLS) regression has been shown to be inadequate when the dependent variable is discrete (Collett, 1991 and Agresti, 1990). Probit or logit analyses are more appropriate in this case.

The PROBIT procedure computes maximum likelihood estimates of the parameters  $\beta$  and  $C$  of the probit equation using a modified Newton-Raphson algorithm. When the response  $Y$  is binary, with values 0 and 1, the probit equation is

$$p = \Pr(Y = 0) = C + (1 - C)F(\mathbf{x}'\beta)$$

where

- $\beta$  is a vector of parameter estimates
- $F$  is a cumulative distribution function (the normal, logistic, or extreme value)
- $\mathbf{x}$  is a vector of explanatory variables
- $p$  is the probability of a response
- $C$  is the natural (threshold) response rate

Notice that PROC PROBIT, by default, models the probability of the *lower* response levels. The choice of the distribution function  $F$  (normal for the probit model, logistic for the logit model, and extreme value or Gompertz for the gompit model) determines the type of analysis. For most problems, there is relatively little difference between the normal and logistic specifications of the model. Both distributions are symmetric about the value zero. The extreme value (or Gompertz) distribution, however, is not symmetric, approaching 0 on the left more slowly than it approaches 1 on the right. You can use the extreme value distribution where such asymmetry is appropriate.

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  $1, \dots, k, k + 1$ . 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 PROBIT 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

$$\Pr(Y \leq 1 \mid \mathbf{x}) = F(\mathbf{x}'\boldsymbol{\beta})$$

$$\Pr(Y \leq i \mid \mathbf{x}) = F(\alpha_i + \mathbf{x}'\boldsymbol{\beta}), \quad 2 \leq i \leq k$$

where  $\alpha_2, \dots, \alpha_k$  are  $k - 1$  intercept parameters. By default, the covariate vector  $\mathbf{x}$  contains an overall intercept term.

You can set or estimate the natural (threshold) response rate  $C$ . Estimation of  $C$  can begin either from an initial value that you specify or from the rate observed in a control group. By default, the natural response rate is fixed at zero.

An observation in the data set analyzed by the PROBIT procedure may contain the response and explanatory values for one subject. Alternatively, it may provide the number of observed events from a number of subjects at a particular setting of the explanatory variables. In this case, PROC PROBIT models the probability of an event.

---

## Getting Started

The following example illustrates how you can use the PROBIT procedure to compute the threshold response rate and regression parameter estimates for quantal response data.

---

### Estimating the Natural Response Threshold Parameter

Suppose you want to test the effect of a drug at 12 dosage levels. You randomly divide 180 subjects into 12 groups of 15—one group for each dosage level. You then conduct the experiment and, for each subject, record the presence or absence of a positive response to the drug. You summarize the data by counting the number of subjects responding positively in each dose group. Your data set is as follows:

```
data study;
  input Dose Respond;
  Number = 15;
  Observed=Respond/Number;
  datalines;
0      3
1.1    4
1.3    4
2.0    3
2.2    5
2.8    4
```

```

3.7  5
3.9  9
4.4  8
4.8  11
5.9  12
6.8  13
;
run;

```

The variable `dose` represents the amount of drug administered. The first group, receiving a dose level of 0, is the control group. The variable `number` represents the number of subjects in each group. All groups are equal in size; hence, `number` has the value 15 for all observations. The variable `respond` represents the number of subjects responding to the associated drug dosage. The variable `observed` is used in subsequent statements for comparison with the predicted probabilities output from the PROBIT procedure.

You can model the probability of positive response as a function of dosage using the following statements:

```

proc probit data=study log10 optc;
  model respond/number=dose;
  output out=new p=p_hat;
run;

```

The `DATA=` option specifies that PROC PROBIT analyze the SAS data set `study`. The `LOG10` option replaces the first continuous independent variable (`dose`) by its common logarithm. The `OPTC` option estimates the natural response rate. When you use the `LOG10` option with the `OPTC` option, any observations with a dose value less than or equal to zero are used in the estimation as a control group.

The `OUTPUT` statement creates a new data set, `new`, that contains all the variables in the original data set, and a new variable, `p_hat`, that represents the predicted probabilities.

The `MODEL` statement specifies a proportional response using the variables `respond` and `number` in *events/trials* syntax. The variable `dose` is the stimulus or explanatory variable. The results from this analysis are displayed in the following figures.

```

                                The SAS System

                                Probit Procedure

                                Model Information

Data Set                               WORK.STUDY
Events Variable                         Respond
Trials Variable                         Number
Number of Observations                   12
Number of Events                         81
Number of Trials                         180
Number of Events In Control Group        3
Number of Trials In Control Group        15
Name of Distribution                     NORMAL
Log Likelihood                           -104.3945783

Algorithm converged.

```

**Figure 49.1.** Model Fitting Information for the PROBIT Procedure

Figure 49.1 displays background information about the model fit. Included are the name of the input data set, the response variables used, and the number of observations, events, and trials. The last line in Figure 49.1 shows the final value of the log-likelihood function.

Figure 49.2 displays the table of parameter estimates for the model. The parameter  $C$ , which is the natural response threshold or the proportion of individuals responding at zero dose, is estimated to be 0.2409. Since both the intercept and the slope coefficient have significant  $p$ -values (0.0020, 0.0010), you can write the model for

$$\text{Pr}(\text{response}) = C + (1 - C)F(\mathbf{x}'\beta)$$

as

$$\text{Pr}(\text{response}) = 0.2409 + 0.7591(\Phi(-4.1439 + 6.2308 \times \log_{10}(\text{dose})))$$

where  $\Phi$  is the normal cumulative distribution function.

```

                                Probit Procedure

                                Analysis of Parameter Estimates

Variable      DF      Estimate      Standard
                                Error Chi-Square Pr > ChiSq Label
Intercept      1      -4.14385      1.34149      9.5419      0.0020 Intercept
Log10(Dose)    1       6.23076      1.89958     10.7588      0.0010
_C_            1       0.24088      0.05226
                                Lower threshold

```

**Figure 49.2.** Model Parameter Estimates for the PROBIT Procedure

Finally, PROC PROBIT specifies the resulting tolerance distribution by providing the mean MU and scale parameter SIGMA as well as the covariance matrix of the distribution parameters.

Probit Procedure			
Probit Model in Terms of Tolerance Distribution			
	MU	SIGMA	
	0.66506312	0.16049411	
Estimated Covariance Matrix for Tolerance Parameters			
	MU	SIGMA	_C_
MU	0.001158	-0.000493	0.000954
SIGMA	-0.000493	0.002394	-0.000999
_C_	0.000954	-0.000999	0.002731

**Figure 49.3.** Tolerance Distribution Estimates for the PROBIT Procedure

The following PROC GPLOT statements request a plot of the fitted probabilities and observed proportions versus the variable dose.

```

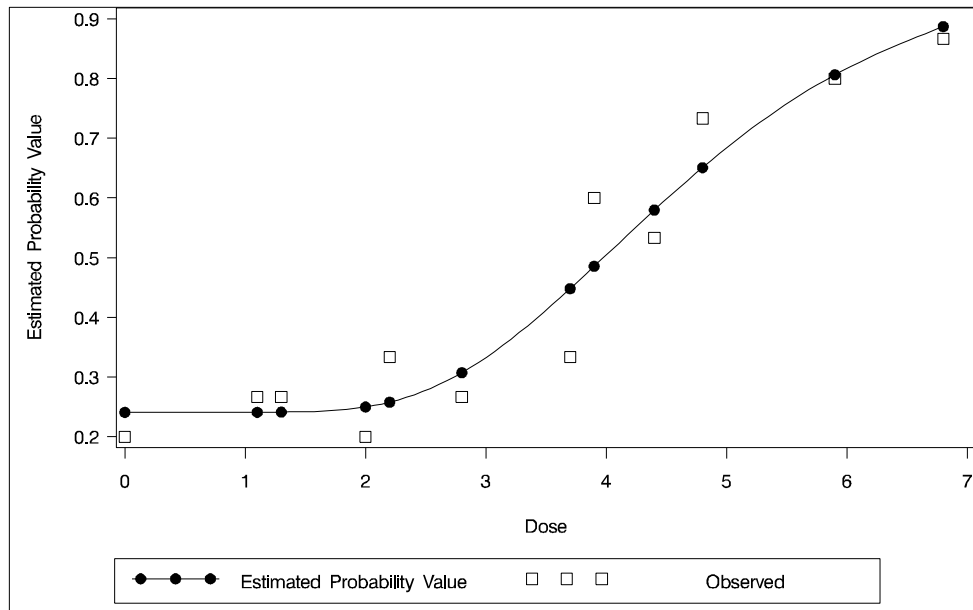
symbol1 i=spline v=dot c=white;
symbol2 i=none v=dot c=yellow;
legend1 frame cframe=ligr cborder=black position=center
      label=none value=(justify=center);
axis1 minor=none color=black label=(angle=90 rotate=0);
axis2 minor=none color=black;

proc gplot data=new;
  plot (p_hat observed)*dose/overlay
      frame cframe=ligr vaxis=axis1 haxis=axis2 legend=legend1;
run;

```

The SYMBOL statements determine line type, plotting symbol, and color. The AXIS and LEGEND statements determine settings for the plot axes and legend.

The OVERLAY option in the PLOT statement specifies that both plots (p\_hat versus dose and observed versus dose) are displayed on the same axes. The VAXIS, HAXIS, and LEGEND options direct the GPLOT procedure to use the settings defined in the previous AXIS and LEGEND statements.



**Figure 49.4.** Plot of Observed and Fitted Probabilities versus Dose Level

The plot in Figure 49.4 shows the relationship between dosage level, observed response proportions, and estimated probability values.

---

## Syntax

The following statements are available in PROC PROBIT:

```

PROC PROBIT < options > ;
  CLASS variables ;
  MODEL response=independents < / options > ;
  BY variables ;
  OUTPUT < OUT=SAS-data-set > < options > ;
  WEIGHT variable ;

```

The MODEL statement is required. If a CLASS statement is used, it must precede the MODEL statement.

---

## PROC PROBIT Statement

**PROC PROBIT** < *options* > ;

The PROC PROBIT statement starts the procedure. You can specify the following options in the PROC PROBIT statement.

### COVOUT

writes the parameter estimate covariance matrix to the OUTEST= data set.

### C=rate

### OPTC

controls how the natural response is handled. Specify the OPTC option to request that the natural response rate  $C$  be estimated. Specify the C=rate option to set the natural response rate or to provide the initial estimate of the natural response rate. The natural response rate value must be a number between 0 and 1.

- If you specify neither the OPTC nor the C= option, a natural response rate of zero is assumed.
- If you specify both the OPTC and the C= option, the C= option should be a reasonable initial estimate of the natural response rate. For example, you could use the ratio of the number of responses to the number of subjects in a control group.
- If you specify the C= option but not the OPTC option, the natural response rate is set to the specified value and not estimated.
- If you specify the OPTC option but not the C= option, PROC PROBIT's action depends on the response variable, as follows:
  - If you specify either the LN or LOG10 option and some subjects have the first independent variable (dose) values less than or equal to zero, these subjects are treated as a control group. The initial estimate of  $C$  is then the ratio of the number of responses to the number of subjects in this group.
  - If you do not specify the LN or LOG10 option or if there is no control group, then one of the following occurs:
    - \* If all responses are greater than zero, the initial estimate of the natural response rate is the minimal response rate ( the ratio of the number of responses to the number of subjects in a dose group) across all dose levels.
    - \* If one or more of the responses is zero (making the response rate zero in that dose group), the initial estimate of the natural rate is the reciprocal of twice the largest number of subjects in any dose group in the experiment.

**DATA=SAS-data-set**

names the SAS data set to be used by PROC PROBIT. By default, the procedure uses the most recently created SAS data set.

**HPROB= $\rho$** 

specifies a minimum probability level for the Pearson chi-square to indicate a good fit. The default value is 0.10. The LACKFIT option must also be specified for this option to have any effect. For Pearson goodness of fit chi-square values with probability greater than the HPROB= value, the fiducial limits, if requested with the INVERSECL option, are computed using a critical value of 1.96. For chi-square values with probability less than the value of the HPROB= option, the critical value is a 0.95 two-sided quantile value taken from the  $t$  distribution with degrees of freedom equal to  $(k - 1) \times m - q$ , where  $k$  is the number of levels for the response variable,  $m$  is the number of different sets of independent variable values, and  $q$  is the number of parameters fit in the model. Note that the HPROB= option can also appear in the MODEL statement.

**INVERSECL**

computes confidence limits for the values of the first continuous independent variable (such as dose) that yield selected response rates. If the algorithm fails to converge (this can happen when  $C$  is nonzero), missing values are reported for the confidence limits. See the section “Inverse Confidence Limits” on page 2626 for details. Note that the INVERSECL option can also appear in the MODEL statement.

**LACKFIT**

performs two goodness-of-fit tests (a Pearson chi-square test and a log-likelihood ratio chi-square test) for the fitted model.

**Note:** The data set must be sorted by the independent variables before the PROBIT procedure is run if you want to perform a test of fit. This test is not appropriate if the data are very sparse, with only a few values at each set of the independent variable values.

If the Pearson chi-square test statistic is significant, then the covariance estimates and standard error estimates are adjusted. See the “Lack of Fit Tests” section on page 2625 for a description of the tests. Note that the LACKFIT option can also appear in the MODEL statement.

**LOG****LN**

analyzes the data by replacing the first continuous independent variable by its natural logarithm. This variable is usually the level of some treatment such as dosage. In addition to the usual output given by the INVERSECL option, the estimated dose values and 95% fiducial limits for dose are also displayed. If you specify the OPTC option, any observations with a dose value less than or equal to zero are used in the estimation as a control group. If you do not specify the OPTC option with the LOG or LN option, then any observations with the first continuous independent variable values less than or equal to zero are ignored.

**LOG10**

specifies an analysis like that of the LN or LOG option except that the common logarithm (log to the base 10) of the dose value is used rather than the natural logarithm.

**NOPRINT**

suppresses the display of all output. Note that this option temporarily disables the Output Delivery System (ODS). For more information, see Chapter 14, “Using the Output Delivery System.”

**OPTC**

controls how the natural response is handled. See the description of the C= option on page 2613 for details.

**ORDER=DATA | FORMATTED | FREQ | INTERNAL**

specifies the sorting order for the levels of the classification variables specified in the CLASS statement, including the levels of the response variable. Response level ordering is important since PROC PROBIT always models the probability of response levels at the beginning of the ordering. See the section “Response Level Ordering” on page 2621 for further details. This ordering also determines which parameters in the model correspond to each level in the data. The following table shows how PROC PROBIT interprets values of the ORDER= option.

Value of ORDER=	Levels Sorted By
DATA	order of appearance in the input data set
FORMATTED	formatted value
FREQ	descending frequency count; levels with the most observations come first in the order
INTERNAL	unformatted value

By default, ORDER=FORMATTED. For the values 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*.

**OUTEST= SAS-data-set**

specifies a SAS data set to contain the parameter estimates and, if the COVOUT option is specified, their estimated covariances. If you omit this option, the output data set is not created. The contents of the data set are described in the section “OUTEST= Data Set” on page 2628. This data set is not created if class variables are used.

---

## BY Statement

**BY** *variables* ;

You can specify a BY statement with PROC PROBIT 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.

If your input data set is not sorted in ascending order on each of the BY variables, 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 PROBIT 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.

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** *variables* ;

The CLASS statement names the classification variables to be used in the analysis. Classification variables can be either character or numeric. If a single response variable is specified in the MODEL statement, it must also be specified in a CLASS statement.

Class levels are determined from the formatted values of the CLASS variables. Thus, you can use formats to group values into levels. See the discussion of the FORMAT procedure in *SAS Language Reference: Dictionary*.

If you use a CLASS statement, you cannot output parameter estimates to the OUTEST= data set (you can output them to a data set via ODS). If the CLASS statement is used, it must appear before any of the MODEL statements.

---

## MODEL Statement

```
<label:> MODEL response=independents < / options > ;
```

```
<label:> MODEL events/trials=independents < / options > ;
```

The MODEL statement names the variables used as the response and the independent variables. Additionally, you can specify the distribution used to model the response, as well as other options. More than one MODEL statement can be specified with the PROBIT procedure. The optional *label* is used to label output from the matching MODEL statement.

The *response* can be a single variable with a value that is used to indicate the level of the observed response. Such a response variable must be listed in the CLASS statement. For example, the response might be a variable called Symptoms that takes on the values 'None,' 'Mild,' or 'Severe.' Note that, for dichotomous response variables, the probability of the lower sorted value is modeled by default (see the "Details" section beginning on page 2621). Because the model fit by the PROBIT procedure requires ordered response levels, you may need to use either the ORDER=DATA option in the PROC statement or a numeric coding of the response to get the desired ordering of levels.

Alternatively, the response can be specified as a pair of variable names 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. Both variables must be numeric and nonnegative, and the ratio of the first variable value to the second variable value must be between 0 and 1, inclusive. For example, the variables might be hits, a variable containing the number of hits for a baseball player, and AtBats, a variable containing the number of times at bat. A model for hitting proportion (batting average) as a function of age could be specified as

```
model hits/AtBats=age;
```

If no independent variables are specified, PROC PROBIT fits an intercept-only model.

The following options are available in the MODEL statement.

### **CONVERGE=***value*

specifies the convergence criterion. Convergence is declared when the maximum change in the parameter estimates between Newton-Raphson steps is less than the value specified. The change is a relative change if the parameter is greater than 0.01 in absolute value; otherwise, it is an absolute change.

By default, CONVERGE=0.001.

### **CORRB**

displays the estimated correlation matrix of the parameter estimates.

**COVB**

displays the estimated covariance matrix of the parameter estimates.

**DISTRIBUTION**=*distribution-type*

**DIST**=*distribution-type*

**D**=*distribution-type*

specifies the cumulative distribution function used to model the response probabilities. The distributions are described in the “Details” section beginning on page 2621. Valid values for *distribution-type* are

NORMAL the normal distribution for the probit model

LOGISTIC the logistic distribution for the logit model

EXTREMEVALUE | EXTREME | GOMPERTZ the extreme value, or Gompertz distribution for the gompit model

By default, DISTRIBUTION=NORMAL.

**HPROB**=*value*

specifies a minimum probability level for the Pearson chi-square to indicate a good fit. The default value is 0.10. The LACKFIT option must also be specified for this option to have any effect. For Pearson goodness of fit chi-square values with probability greater than the HPROB= value, the fiducial limits, if requested with the INVERSECL option, are computed using a critical value of 1.96. For chi-square values with probability less than the value of the HPROB= option, the critical value is a 0.95 two-sided quantile value taken from the *t* distribution with degrees of freedom equal to  $(k - 1) \times m - q$ , where *k* is the number of levels for the response variable, *m* is the number of different sets of independent variable values, and *q* is the number of parameters fit in the model. If you specify the HPROB= option in both the PROC and MODEL statements, the MODEL statement option takes precedence.

**INITIAL**=*values*

sets initial values for the parameters in the model other than the intercept. The values must be given in the order in which the variables are listed in the MODEL statement. If some of the independent variables listed in the MODEL statement are classification variables, then there must be as many values given for that variable as there are classification levels minus 1. The INITIAL option can be specified as follows.

Type of List	Specification
list separated by blanks	<b>initial=3 4 5</b>
list separated by commas	<b>initial=3,4,5</b>

By default, all parameters have initial estimates of zero.

**INTERCEPT**=*value*

initializes the intercept parameter to *value*. By default, INTERCEPT=0.

**INVERSECL**

computes confidence limits for the values of the first continuous independent variable (such as dose) that yield selected response rates. If the algorithm fails to converge (this can happen when  $C$  is nonzero), missing values are reported for the confidence limits. See the section “Inverse Confidence Limits” on page 2626 for details.

**ITPRINT**

displays the iteration history, the final evaluation of the gradient, and the second derivative matrix (Hessian).

**LACKFIT**

performs two goodness-of-fit tests (a Pearson chi-square test and a log-likelihood ratio chi-square test) for the fitted model.

**Note:** The data set must be sorted by the independent variables before the PROBIT procedure is run if you want to perform a test of fit. This test is not appropriate if the data are very sparse, with only a few values at each set of the independent variable values.

If the Pearson chi-square test statistic is significant, then the covariance estimates and standard error estimates are adjusted. See the “Lack of Fit Tests” section on page 2625 for a description of the tests. If you specify the LACKFIT option in both the PROC and MODEL statements, the MODEL statement option takes precedence.

**MAXITER=*value***

specifies the maximum number of iterations to be performed in estimating the parameters. By default, MAXITER=50.

**NOINT**

fits a model with no intercept parameter. If the INTERCEPT= option is also specified, the intercept is fixed at the specified value; otherwise, it is set to zero. This is most useful when the response is binary. When the response has  $k$  levels, then  $k - 1$  intercept parameters are fit. The NOINT option sets the intercept parameter corresponding to the lowest response level equal to zero. A Lagrange multiplier, or score, test for the restricted model is computed when the NOINT option is specified.

**SINGULAR=*value***

specifies the singularity criterion for determining linear dependencies in the set of independent variables. The sum of squares and crossproducts matrix of the independent variables is formed and swept. If the relative size of a pivot becomes less than the value specified, then the variable corresponding to the pivot is considered to be linearly dependent on the previous set of variables considered. By default, SINGULAR=1E-12.

---

## OUTPUT Statement

**OUTPUT** <OUT=SAS-data-set> <keyword=name...keyword=name>;

The OUTPUT statement creates a new SAS data set containing all variables in the input data set and, optionally, the fitted probabilities, the estimate of  $\mathbf{x}'\beta$ , and the estimate of its standard error. Estimates of the probabilities,  $\mathbf{x}'\beta$ , and the standard errors are computed for observations with missing response values as long as the values of all the explanatory variables are nonmissing. This enables you to compute these statistics for additional settings of the explanatory variables that are of interest but for which responses are not observed.

You can specify multiple OUTPUT statements. Each OUTPUT statement creates a new data set and applies only to the preceding MODEL statement. If you want to create a permanent SAS data set, you must specify a two-level name (refer to *SAS Language Reference: Concepts* for more information on permanent SAS data sets).

Details on the specifications in the OUTPUT statement are as follows:

*keyword=name* specifies the statistics to include in the output data set and assigns names to the new variables that contain the statistics. Specify a keyword for each desired statistic (see the following list of keywords), an equal sign, and the variable to contain the statistic.

The keywords allowed and the statistics they represent are as follows:

PROB | P cumulative probability estimates

$$p = C + (1 - C)F(a_j + \mathbf{x}'\beta)$$

STD standard error estimates of  $a_j + \mathbf{x}'\mathbf{b}$

XBETA estimates of  $a_j + \mathbf{x}'\beta$

OUT=SAS-data-set names the output data set. By default, the new data set is named using the DATA $n$  convention.

When the *single variable response* syntax is used, the `_LEVEL_` variable is added to the output data set, and there are  $k - 1$  output observations for each input observation, where  $k$  is the number of response levels. There is no observation output corresponding to the highest response level. For each of the  $k - 1$  observations, the PROB variable contains the fitted probability of obtaining a response level up to the level indicated by the `_LEVEL_` variable, the XBETA variable contains  $a_j + \mathbf{x}'\mathbf{b}$ , where  $j$  references the levels ( $a_1 = 0$ ), and the STD variable contains the standard error estimate of the XBETA variable. See the “Details” section, which follows, for the formulas for the parameterizations.

---

## WEIGHT Statement

**WEIGHT** *variable* ;

A WEIGHT statement can be used with PROC PROBIT to weight each observation by the value of the variable specified. The contribution of each observation to the likelihood function is multiplied by the value of the weight variable. Observations with zero, negative, or missing weights are not used in model estimation.

---

## Details

---

### Missing Values

PROC PROBIT does not use any observations having missing values for any of the independent variables, the response variables, or the weight variable. If only the response variables are missing, statistics requested in the OUTPUT statement are computed.

---

### Response Level Ordering

For binary response data, PROC PROBIT fits the following model by default.

$$\Phi^{-1} \left( \frac{p - C}{1 - C} \right) = \mathbf{x}'\boldsymbol{\beta}$$

where  $p$  is the probability of the response level identified as the first level in the “Weighted Frequency Counts for the Ordered Response Categories” table in the output and  $\Phi$  is the normal cumulative distribution function. By default, the covariate vector  $\mathbf{x}$  contains an intercept term. This is sometimes called Abbot’s formula.

Because of the symmetry of the normal (and logistic) distribution, the effect of reversing the order of the two response values is to change the signs of  $\boldsymbol{\beta}$  in the preceding equation.

By default, response levels appear in ascending, sorted order (that is, the lowest level appears first and then the next lowest, 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 the first ordered level. One of the most common sets of response levels is  $\{0,1\}$ , with 1 representing the event with the probability that 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 PROBIT assigns the first ordered level 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

- 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,  $Y=0$  could be assigned formatted value 'nonevent' and  $Y=1$  could be assigned formatted value 'event.' Since  $ORDER=FORMATTED$  by default,  $Y=1$  becomes the first ordered level. See Example 49.3 for an illustration of this method.

```
proc format;
  value disease 1='event' 0='nonevent';
run;
proc probit;
  model y=exposure;
  format y disease.;
run;
```

- arrange the input data set so that  $Y=1$  appears first and use the  $ORDER=DATA$  option in the PROC PROBIT statement. Since  $ORDER=DATA$  sorts levels in order of their appearance in the data set,  $Y=1$  becomes the first ordered level. Note that this option causes class variables to be sorted by their order of appearance in the data set, also.

---

## Computational Method

The log-likelihood function is maximized by means of a ridge-stabilized Newton-Raphson algorithm. Initial parameter estimates are set to zero. The  $INITIAL=$  and  $INTERCEPT=$  options in the MODEL statement can be used to give nonzero initial estimates.

The log-likelihood function,  $L$ , is computed as

$$L = \sum_i w_i \ln(p_i)$$

where the sum is over the observations in the data set,  $w_i$  is the weight for the  $i$ th observation, and  $p_i$  is the modeled probability of the observed response. In the case of the events/trials syntax in the MODEL statement, each observation contributes two terms corresponding to the probability of the event and the probability of its complement:

$$L = \sum_i w_i [r_i \ln(p_i) + (n_i - r_i) \ln(1 - p_i)]$$

where  $r_i$  is the number of events and  $n_i$  is the number of trials for observation  $i$ . This log-likelihood function differs from the log-likelihood function for a binomial

or multinomial distribution by additive terms consisting of the log of binomial or multinomial coefficients. These terms are parameter-independent and do not affect the model estimation or the standard errors and tests.

The estimated covariance matrix,  $\mathbf{V}$ , of the parameter estimates is computed as the negative inverse of the information matrix of second derivatives of  $L$  with respect to the parameters evaluated at the final parameter estimates. Thus, the estimated covariance matrix is derived from the observed information matrix rather than the expected information matrix (these are generally not the same). The standard error estimates for the parameter estimates are taken as the square roots of the corresponding diagonal elements of  $\mathbf{V}$ .

For a classification effect, an overall chi-square statistic is computed as

$$\chi^2 = \mathbf{b}'_1 \mathbf{V}^{-1}_{11} \mathbf{b}_1$$

where  $\mathbf{V}_{11}$  is the submatrix of  $\mathbf{V}$  corresponding to the indicator variables for the classification effect and  $\mathbf{b}_1$  is the vector of parameter estimates corresponding to the classification effect. This chi-square statistic has degrees of freedom equal to the rank of  $\mathbf{V}_{11}$ .

If some of the independent variables are perfectly correlated with the response pattern, then the theoretical parameter estimates may be infinite. Although fitted probabilities of 0 and 1 are not especially pathological, infinite parameter estimates are required to yield these probabilities. Due to the finite precision of computer arithmetic, the actual parameter estimates are not infinite. Indeed, since the tails of the distributions allowed in the PROBIT procedure become small rapidly, an argument to the cumulative distribution function of around 20 becomes effectively infinite. In the case of such parameter estimates, the standard error estimates and the corresponding chi-square tests are not trustworthy.

The chi-square tests for the individual parameter values are Wald tests based on the observed information matrix and the parameter estimates. The theory behind these tests assumes large samples. If the samples are not large, it may be better to base the tests on log-likelihood ratios. These changes in log likelihood can be obtained by fitting the model twice, once with all the parameters of interest and once leaving out the parameters to be tested. Refer to Cox and Oakes (1984) for a discussion of the merits of some possible test methods.

---

## Distributions

The distributions,  $F(x)$ , allowed in the PROBIT procedure are specified with the DISTRIBUTION= option in the model statement. The cumulative distribution functions for the available distributions are

$$\int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz \quad (\text{normal})$$

$$\frac{1}{1 + e^{-x}} \quad (\text{logistic})$$

$$1 - e^{-e^x} \quad (\text{extreme value or Gompertz})$$

The variances of these three distributions are not all equal to 1, and their means are not all equal to zero. Their means and variances are shown in the following table, where  $\gamma$  is the Euler constant.

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

When comparing parameter estimates using different distributions, you need to take into account the different scalings and, for the extreme value (or Gompertz) distribution, 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 from the logistic model should be about  $\pi/\sqrt{3}$  larger than the estimates from the probit model.

---

## Model Specification

For a two-level response, the probability that the lesser response occurs is modeled by the probit equation as

$$p = C + (1 - C)F(\mathbf{x}'\mathbf{b})$$

The probability of the other (complementary) event is  $1 - p$ .

For a multilevel response with outcomes labeled  $l_i$  for  $i = 1, 2, \dots, k$ , the probability,  $p_j$ , of observing level  $l_j$  is

$$\begin{aligned}
 p_1 &= C + (1 - C)F(\mathbf{x}'\mathbf{b}) \\
 p_2 &= (1 - C) (F(a_2 + \mathbf{x}'\mathbf{b}) - F(\mathbf{x}'\mathbf{b})) \\
 &\vdots \\
 p_j &= (1 - C) (F(a_j + \mathbf{x}'\mathbf{b}) - F(a_{j-1} + \mathbf{x}'\mathbf{b})) \\
 &\vdots \\
 p_k &= (1 - C)(1 - F(a_{k-1} + \mathbf{x}'\mathbf{b}))
 \end{aligned}$$

Thus, for a  $k$ -level response, there are  $k - 2$  additional parameters,  $a_2, a_3, \dots, a_{k-1}$ , estimated. These parameters are denoted by `INTER.j`,  $j = 2, 3, \dots, k - 1$  in the output.

An intercept parameter is always added to the set of independent variables as the first term in the model unless the `NOINT` option is specified in the `MODEL` statement. If a classification variable taking on  $k$  levels is used as one of the independent variables, a set of  $k$  indicator variables is generated to model the effect of this variable. Because of the presence of the intercept term, there are at most  $k - 1$  degrees of freedom for this effect in the model.

---

## Lack of Fit Tests

Two goodness-of-fit tests can be requested from the `PROBIT` procedure—a Pearson chi-square test and a log-likelihood ratio chi-square test.

If there is only a single continuous independent variable, the data are internally sorted to group response values by the independent variable. Otherwise, the data are aggregated into groupings that are delimited whenever a change is observed in one of the independent variables.

**Note:** Because of this grouping, the data set should be sorted by the independent variables before the `PROBIT` procedure is run if the `LACKFIT` option is specified.

If the Pearson goodness-of-fit chi-square test is requested and the  $p$ -value for the test is too small, variances and covariances are adjusted by a heterogeneity factor (the goodness-of-fit chi-square divided by its degrees of freedom) and a critical value from the  $t$  distribution is used to compute the fiducial limits. The Pearson chi-square test statistic is computed as

$$\sum_i \sum_j \frac{(r_{ij} - n_i p_{ij})^2}{n_i p_{ij}}$$

where the sum on  $i$  is over grouping, the sum on  $j$  is over levels of response, the  $r_{ij}$  is the frequency of response level  $j$  for the  $i$ th grouping,  $n_i$  is the total frequency for the  $i$ th grouping, and  $p_{ij}$  is the fitted probability for the  $j$ th level at the  $i$ th grouping.

The log-likelihood ratio chi-square test statistic is computed as

$$2 \sum_i \sum_j r_{ij} \ln \left( \frac{r_{ij}}{n_i p_{ij}} \right)$$

This quantity is sometimes called the deviance. If the modeled probabilities fit the data, these statistics should be approximately distributed as chi-square with degrees of freedom equal to  $(k-1) \times m - q$ , where  $k$  is the number of levels of the multinomial or binomial response,  $m$  is the number of sets of independent variable values (covariate patterns), and  $q$  is the number of parameters fit in the model.

In order for the Pearson statistic and the deviance to be distributed as chi-square, there must be sufficient replication within the groupings. 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.

## Tolerance Distribution

For a single independent variable, such as a dosage level, the models for the probabilities can be justified on the basis of a population with mean  $\mu$  and scale parameter  $\sigma$  of tolerances for the subjects. Then, given a dose  $x$ , the probability,  $P$ , of observing a response in a particular subject is the probability that the subject's tolerance is less than the dose or

$$P = F \left( \frac{x - \mu}{\sigma} \right)$$

Thus, in this case, the intercept parameter,  $\mathbf{b}_0$ , and the regression parameter,  $\mathbf{b}_1$ , are related to  $\mu$  and  $\sigma$  by

$$\begin{aligned} \mathbf{b}_1 &= \frac{1}{\sigma} \\ \mathbf{b}_0 &= -\frac{\mu}{\sigma} \end{aligned}$$

**Note:** The parameter  $\sigma$  is not equal to the standard deviation of the population of tolerances for the logistic and extreme value distributions.

## Inverse Confidence Limits

In bioassay problems, estimates of the values of the independent variables that yield a desired response are often needed. For instance, the value yielding a 50% response rate (called the ED50 or LD50) is often used. The INVERSECL option requests that confidence limits be computed for the value of the independent variable that yields a specified response. These limits are computed only for the first continuous variable

effect in the model. The other variables are set either at their mean values if they are continuous or at the reference (last) level if they are discrete variables. For a discussion of inverse confidence limits, refer to Hubert, Bohidar, and Peace (1988).

For the PROBIT procedure, the response variable is a probability. An estimate of the first continuous variable value needed to achieve a response of  $p$  is given by

$$\hat{x}_1 = \frac{1}{b_1} (F^{-1}(p) - \mathbf{x}^* \mathbf{b}^*)$$

where  $F$  is the cumulative distribution function used to model the probability,  $\mathbf{x}^*$  is the vector of independent variables excluding the first one,  $\mathbf{b}^*$  is the vector of parameter estimates excluding the first one, and  $b_1$  is the estimated regression coefficient for the independent variable of interest. Note that, for both binary and ordinal models, the INVERSECL option provides estimates of the value of  $x_1$  yielding  $\Pr(\text{first response level}) = p$ , for various values of  $p$ .

This estimator is given as a ratio of random variables, for example,  $r = a/b$ . Confidence limits for this ratio can be computed using Fieller's theorem. A brief description of this theorem follows. Refer to Finney (1971) for a more complete description of Fieller's theorem.

If the random variables  $a$  and  $b$  are thought to be distributed as jointly normal, then for any fixed value  $r$  the following probability statement holds if  $z$  is an  $\alpha/2$  quantile from the standard normal distribution and  $\mathbf{V}$  is the variance-covariance matrix of  $a$  and  $b$ .

$$\Pr((a - rb)^2 > z^2(V_{aa} - 2rV_{ab} + r^2V_{bb})) = \alpha$$

Usually the inequality can be solved for  $r$  to yield a confidence interval. The PROBIT procedure uses a value of 1.96 for  $z$ , corresponding to an  $\alpha$  value of 0.05, unless the goodness-of-fit  $p$ -value is less than the specified value of the HPROB= option. When this happens, the covariance matrix is scaled by the heterogeneity factor, and a  $t$  distribution quantile is used for  $z$ .

It is possible for the roots of the equation for  $r$  to be imaginary or for the confidence interval to be all points outside of an interval. In these cases, the limits are set to missing by the PROBIT procedure.

Although the normal and logistic distribution give comparable fitted values of  $p$  if the empirically observed proportions are not too extreme, they can give appreciably different values when extrapolated into the tails. Correspondingly, the estimates of the confidence limits and dose values can be different for the two distributions even when they agree quite well in the body of the data. Extrapolation outside of the range of the actual data is often sensitive to model assumptions, and caution is advised if extrapolation is necessary.

---

## OUTEST= Data Set

The OUTEST= data set contains parameter estimates and the log likelihood for the specified models. A set of observations is created for each MODEL statement specified. You can use a label in the MODEL statement to distinguish between the estimates for different MODEL statements. If you specify the COVOUT option, the OUTEST= data set also contains the estimated covariance matrix of the parameter estimates.

The OUTEST= data set is not created if there are any CLASS variables in any model. If created, this data set contains each variable used as a dependent or independent variable in any MODEL statement. One observation consists of parameter values for the model with the dependent variable having the value  $-1$ . If you specify the COVOUT option, there are additional observations containing the rows of the estimated covariance matrix. For these observations, the dependent variable contains the parameter estimate for the corresponding row variable. The following variables are also added to the data set:

<code>_MODEL_</code>	a character variable of length 8 containing the label of the MODEL statement, if present, or blank otherwise
<code>_NAME_</code>	a character variable containing the name of the dependent variable for the parameter estimates observations or the name of the row for the covariance matrix estimates
<code>_TYPE_</code>	a character variable containing the type of the observation, either PARMS for parameter estimates or COV for covariance estimates
<code>_DIST_</code>	a character variable containing the name of the distribution modeled
<code>_LNLIKE_</code>	a numeric variable containing the last computed value of the log likelihood
<code>_C_</code>	a numeric variable containing the estimated threshold parameter
<code>INTERCEPT</code>	a numeric variable containing the intercept parameter estimates and covariances

Any BY variables specified are also added to the OUTEST= data set.

---

## Displayed Output

If you request the iteration history (ITPRINT), PROC PROBIT displays

- the current value of the log likelihood
- the ridging parameter for the modified Newton-Raphson optimization process
- the current estimate of the parameters
- the current estimate of the parameter  $C$  for a natural (threshold) model
- the values of the gradient and the Hessian on the last iteration

If you include CLASS variables, PROC PROBIT displays

- the numbers of levels for each CLASS variable
- the (ordered) values of the levels
- the number of observations used

After the model is fit, PROC PROBIT displays

- the name of the input data set
- the name of the dependent variables
- the number of observations used
- the number of events and the number of trials
- the final value of the log-likelihood function
- the parameter estimates
- the standard error estimates of the parameter estimates
- approximate chi-square test statistics for the test

If you specify the COVB or CORRB options, PROC PROBIT displays

- the estimated covariance matrix for the parameter estimates
- the estimated correlation matrix for the parameter estimates

If you specify the LACKFIT option, PROC PROBIT displays

- a count of the number of levels of the response and the number of distinct sets of independent variables
- a goodness-of-fit test based on the Pearson chi-square
- a goodness-of-fit test based on the likelihood-ratio chi-square

If you specify only one independent variable, the normal distribution is used to model the probabilities, and the response is binary, PROC PROBIT displays

- the mean MU of the stimulus tolerance
- the scale parameter SIGMA of the stimulus tolerance
- the covariance matrix for MU, SIGMA, and the natural response parameter  $C$

If you specify the INVERSECL options, PROC PROBIT also displays

- the estimated dose along with the 95% fiducial limits for probability levels 0.01 to 0.10, 0.15 to 0.85 by 0.05, and 0.90 to 0.99

## ODS Table Names

PROC PROBIT assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in the following table. For more information on ODS, see Chapter 14, “Using the Output Delivery System.”

**Table 49.1.** ODS Tables Produced in PROC PROBIT

ODS Table Name	Description	Statement	Option
ClassLevels	Class variable levels	CLASS	default
ConvergenceStatus	Convergence status	MODEL	default
CorrB	Parameter estimate correlation matrix	MODEL	CORRB
CovB	Parameter estimate covariance matrix	MODEL	COVB
CovTolerance	Covariance matrix for location and scale	MODEL *	default
GoodnessOfFit	Goodness of fit tests	MODEL	LACKFIT
IterHistory	Iteration history	MODEL	ITPRINT
LagrangeStatistics	Lagrange statistics	MODEL	NOINT
LastGrad	Last evaluation of the gradient	MODEL	ITPRINT
LastHess	Last evaluation of the Hessian	MODEL	ITPRINT
LogProbitAnalysis	Probit analysis for log dose	MODEL	INVERSECL
ModelInfo	Model information	MODEL	default
MuSigma	Location and scale	MODEL *	default
OrdinalResp	Counts for ordinal data	MODEL *	default
ParameterEstimates	Parameter estimates	MODEL	default
ProbitAnalysis	Probit analysis for linear dose	MODEL	INVERSECL
ResponseLevels	Response-covariate profile	MODEL	LACKFIT

\*Depends on data.

## Examples

### Example 49.1. Dosage Levels

In this example, Dose is the variable representing the level of the stimulus, N represents the number of subjects tested at each level of the stimulus, and Response is the number of subjects responding to that level of the stimulus. Both probit and logit response models are fit to the data. The LOG10 option in the PROC statement requests that the log base 10 of Dose is used as the independent variable. Specifically, for a given level of Dose, the probability  $p$  of a positive response is modeled as

$$p = \Pr(\text{Response}) = F(b_0 + b_1 \times \log_{10}(\text{Dose}))$$

The probabilities are estimated first using the normal distribution function and then using the logistic distribution function. Note that, in this model specification, the natural rate is assumed to be zero.

Lack-of-fit tests and inverse confidence limits are also requested.

In the DATA step that reads the data, a number of observations are generated that have a missing value for the response. Although the PROBIT procedure does not use the observations with the missing values to fit the model, it does give predicted values for all nonmissing sets of independent variables. These data points fill in the plot of fitted and observed values in the logistic model. The plot displays the estimated logistic cumulative distribution function and the observed response rates.

The following statements produce Output 49.1.1:

```

data a;
  infile cards eof=eof;
  input Dose N Response;
  Observed= Response/N;
  output;
  return;
eof: do Dose=0.5 to 7.5 by 0.25;
      output;
      end;
      datalines;
1 10 1
2 12 2
3 10 4
4 10 5
5 12 8
6 10 8
7 10 10
;

proc probit log10;
  model Response/N=Dose / lackfit inversecl itprint;
  model Response/N=Dose / d=logistic inversecl;
  output out=B p=Prob std=std xbeta=xbeta;
  title 'Output from Probit Procedure';
run;

legend1 label=none frame cframe=ligr cborder=black
  position=center value=(justify=center);
axis1 minor=none color=black label=(angle=90 rotate=0) ;
axis2 minor=none color=black;
proc gplot;
  plot Observed*Dose Prob*Dose / overlay frame cframe=ligr
  vaxis=axis1 haxis=axis2 legend=legend1;
  title 'Plot of Observed and Fitted Probabilities';
run;

```

## Output 49.1.1. Dosage Levels: PROC PROBIT

Output from Probit Procedure				
Probit Procedure				
Iteration History for Parameter Estimates				
Iter	Ridge	Loglikelihood	Intercept	Log10(Dose)
0	0	-51.292891	0	0
1	0	-37.881166	-1.355817008	2.635206083
2	0	-37.286169	-1.764939171	3.3408954936
3	0	-37.280389	-1.812147863	3.4172391614
4	0	-37.280388	-1.812704962	3.418117919

## Output 49.1.1. (continued)

Output from Probit Procedure				
Probit Procedure				
Model Information				
Data Set		WORK.B		
Events Variable		Response		
Trials Variable		N		
Number of Observations		7		
Number of Events		38		
Number of Trials		74		
Missing Values		29		
Name of Distribution		NORMAL		
Log Likelihood		-37.28038802		
Last Evaluation of the Negative of the Gradient				
	Intercept	Log10(Dose)		
	3.4349069E-7	-2.09809E-8		
Last Evaluation of the Negative of the Hessian				
		Intercept	Log10(Dose)	
Intercept	36.005280383	20.152675982		
Log10(Dose)	20.152675982	13.078826305		
Goodness-of-Fit Tests				
Statistic	Value	DF	Pr >	ChiSq
Pearson Chi-Square	3.6497	5	0.6009	
L.R. Chi-Square	4.6381	5	0.4616	
Response-Covariate Profile				
Response Levels		2		
Number of Covariate Values		7		

The  $p$ -values in the Goodness-of-Fit table of 0.6009 for the Pearson chi-square and 0.4616 for the likelihood ratio chi-square indicate an adequate fit for the model fit with the normal distribution.

**Output 49.1.1.** (continued)

```

Output from Probit Procedure

      Probit Procedure

Analysis of Parameter Estimates

Variable          DF      Estimate      Standard
                  DF      Estimate      Error Chi-Square Pr > ChiSq Label
Intercept          1     -1.81270      0.44934      16.2743      <.0001 Intercept
Log10(Dose)        1       3.41812      0.74555      21.0196      <.0001

Probit Model in Terms of Tolerance Distribution

                          MU          SIGMA
                          0.53032254  0.29255866

Estimated Covariance Matrix
for Tolerance Parameters

                          MU          SIGMA
MU          0.002418          -0.000409
SIGMA      -0.000409          0.004072

```

Tolerance distribution parameter estimates for the normal distribution indicate a mean tolerance for the population of 0.5303.

## Output 49.1.1. (continued)

Output from Probit Procedure			
Probit Procedure			
Probit Analysis on Log10(Dose)			
Probability	Log10(Dose)	95% Fiducial Limits	
		Lower	Upper
0.01	-0.15027	-0.69520	0.07710
0.02	-0.07052	-0.55768	0.13475
0.03	-0.01992	-0.47066	0.17157
0.04	0.01814	-0.40535	0.19941
0.05	0.04911	-0.35235	0.22218
0.06	0.07546	-0.30733	0.24165
0.07	0.09857	-0.26794	0.25882
0.08	0.11926	-0.23275	0.27426
0.09	0.13807	-0.20081	0.28837
0.10	0.15539	-0.17148	0.30142
0.15	0.22710	-0.05087	0.35631
0.20	0.28410	0.04368	0.40124
0.25	0.33299	0.12342	0.44116
0.30	0.37690	0.19348	0.47857
0.35	0.41759	0.25658	0.51505
0.40	0.45620	0.31428	0.55183
0.45	0.49356	0.36754	0.58999
0.50	0.53032	0.41693	0.63057
0.55	0.56709	0.46296	0.67451
0.60	0.60444	0.50618	0.72271
0.65	0.64305	0.54734	0.77603
0.70	0.68374	0.58745	0.83551
0.75	0.72765	0.62776	0.90265
0.80	0.77655	0.66999	0.98009
0.85	0.83354	0.71675	1.07280
0.90	0.90525	0.77313	1.19192
0.91	0.92257	0.78645	1.22098
0.92	0.94139	0.80083	1.25266
0.93	0.96208	0.81653	1.28760
0.94	0.98519	0.83394	1.32673
0.95	1.01154	0.85367	1.37150
0.96	1.04250	0.87669	1.42425
0.97	1.08056	0.90479	1.48930
0.98	1.13116	0.94189	1.57603
0.99	1.21092	0.99987	1.71322

The LD50 (ED50 for log dose) is 0.5303, the dose corresponding to a probability of 0.5. This is the same as the mean tolerance for the normal distribution.

## Output 49.1.1. (continued)

Output from Probit Procedure			
Probit Procedure			
Probit Analysis on Dose			
Probability	Dose	95% Fiducial Limits	
		Lower	Upper
0.01	0.70750	0.20174	1.19428
0.02	0.85012	0.27690	1.36381
0.03	0.95517	0.33833	1.48445
0.04	1.04266	0.39323	1.58275
0.05	1.11971	0.44428	1.66794
0.06	1.18976	0.49280	1.74444
0.07	1.25478	0.53959	1.81474
0.08	1.31600	0.58513	1.88043
0.09	1.37427	0.62978	1.94253
0.10	1.43019	0.67379	2.00182
0.15	1.68696	0.88948	2.27148
0.20	1.92353	1.10582	2.51907
0.25	2.15276	1.32868	2.76162
0.30	2.38180	1.56126	3.01001
0.35	2.61573	1.80541	3.27375
0.40	2.85893	2.06198	3.56308
0.45	3.11573	2.33096	3.89040
0.50	3.39096	2.61173	4.27141
0.55	3.69051	2.90372	4.72622
0.60	4.02199	3.20757	5.28094
0.65	4.39594	3.52649	5.97082
0.70	4.82770	3.86764	6.84712
0.75	5.34134	4.24384	7.99198
0.80	5.97787	4.67723	9.55182
0.85	6.81617	5.20898	11.82500
0.90	8.03992	5.93102	15.55685
0.91	8.36704	6.11581	16.63355
0.92	8.73752	6.32162	17.89203
0.93	9.16385	6.55428	19.39079
0.94	9.66463	6.82242	21.21933
0.95	10.26925	7.13946	23.52336
0.96	11.02811	7.52812	26.56140
0.97	12.03830	8.03145	30.85292
0.98	13.52585	8.74757	37.67327
0.99	16.25233	9.99702	51.66816

The ED50 for dose is 3.39 with a 95% confidence interval of (2.61, 4.27).

## Output 49.1.1. (continued)

Output from Probit Procedure	
Probit Procedure	
Model Information	
Data Set	WORK.B
Events Variable	Response
Trials Variable	N
Number of Observations	7
Number of Events	38
Number of Trials	74
Missing Values	29
Name of Distribution	LOGISTIC
Log Likelihood	-37.11065336

Algorithm converged.

## Output 49.1.1. (continued)

Output from Probit Procedure						
Probit Procedure						
Analysis of Parameter Estimates						
Variable	DF	Estimate	Standard Error	Chi-Square	Pr >	ChiSq Label
Intercept	1	-3.22464	0.88606	13.2447	0.0003	Intercept
Log10(Dose)	1	5.97018	1.44917	16.9721	<.0001	

The regression parameter estimates for the logistic model of -3.22 and 5.97 are approximately  $\pi/\sqrt{3}$  times as large as those for the normal model.

## Output 49.1.1. (continued)

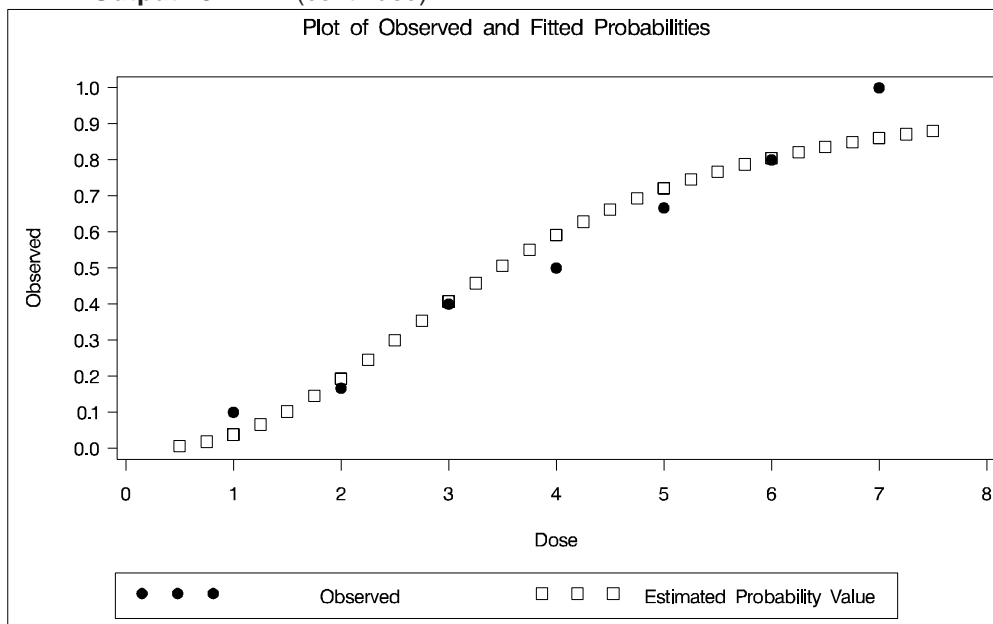
Output from Probit Procedure			
Probit Procedure			
Probit Analysis on Log10(Dose)			
Probability	Log10(Dose)	95% Fiducial Limits	
		Lower	Upper
0.01	-0.22955	-0.97443	0.04234
0.02	-0.11175	-0.75160	0.12404
0.03	-0.04212	-0.62020	0.17266
0.04	0.00780	-0.52620	0.20771
0.05	0.04693	-0.45266	0.23533
0.06	0.07925	-0.39207	0.25827
0.07	0.10686	-0.34039	0.27796
0.08	0.13103	-0.29522	0.29530
0.09	0.15259	-0.25503	0.31085
0.10	0.17209	-0.21876	0.32498
0.15	0.24958	-0.07553	0.38207
0.20	0.30792	0.03091	0.42645
0.25	0.35611	0.11742	0.46451
0.30	0.39820	0.19143	0.49933
0.35	0.43644	0.25684	0.53275
0.40	0.47221	0.31587	0.56619
0.45	0.50651	0.36985	0.60090
0.50	0.54013	0.41957	0.63807
0.55	0.57374	0.46559	0.67895
0.60	0.60804	0.50846	0.72475
0.65	0.64381	0.54895	0.77673
0.70	0.68205	0.58815	0.83638
0.75	0.72414	0.62752	0.90583
0.80	0.77233	0.66915	0.98877
0.85	0.83067	0.71631	1.09243
0.90	0.90816	0.77561	1.23344
0.91	0.92766	0.79014	1.26932
0.92	0.94922	0.80607	1.30913
0.93	0.97339	0.82378	1.35392
0.94	1.00100	0.84384	1.40524
0.95	1.03332	0.86713	1.46548
0.96	1.07245	0.89511	1.53866
0.97	1.12237	0.93053	1.63230
0.98	1.19200	0.97952	1.76331
0.99	1.30980	1.06166	1.98571

## Output 49.1.1. (continued)

Output from Probit Procedure			
Probit Procedure			
Probit Analysis on Dose			
Probability	Dose	95% Fiducial Limits	
		Lower	Upper
0.01	0.58945	0.10606	1.10241
0.02	0.77312	0.17717	1.33059
0.03	0.90757	0.23977	1.48818
0.04	1.01813	0.29772	1.61328
0.05	1.11413	0.35264	1.71923
0.06	1.20018	0.40545	1.81245
0.07	1.27896	0.45668	1.89655
0.08	1.35218	0.50673	1.97380
0.09	1.42100	0.55586	2.04573
0.10	1.48625	0.60428	2.11340
0.15	1.77656	0.84036	2.41031
0.20	2.03199	1.07377	2.66962
0.25	2.27043	1.31043	2.91417
0.30	2.50152	1.55391	3.15737
0.35	2.73172	1.80650	3.40997
0.40	2.96627	2.06954	3.68293
0.45	3.21006	2.34343	3.98929
0.50	3.46837	2.62766	4.34580
0.55	3.74746	2.92137	4.77469
0.60	4.05546	3.22449	5.30576
0.65	4.40366	3.53960	5.98046
0.70	4.80891	3.87389	6.86087
0.75	5.29836	4.24153	8.05054
0.80	5.92009	4.66819	9.74470
0.85	6.77126	5.20363	12.37174
0.90	8.09391	5.96506	17.11758
0.91	8.46559	6.16797	18.59179
0.92	8.89644	6.39834	20.37650
0.93	9.40575	6.66466	22.59024
0.94	10.02317	6.97974	25.42373
0.95	10.79732	7.36425	29.20649
0.96	11.81534	7.85434	34.56649
0.97	13.25466	8.52168	42.88406
0.98	15.55972	9.53935	57.98471
0.99	20.40815	11.52540	96.76344

Both the ED50 and the LD50 are similar to those for the normal model.

Output 49.1.1. (continued)



## Example 49.2. Multilevel Response

In this example, two preparations, a standard preparation and a test preparation, are each given at several dose levels to groups of insects. The symptoms are recorded for each insect within each group, and two multilevel probit models are fit. Because the natural sort order of the three levels is not the same as the response order, the ORDER=DATA option is specified in the PROC statement to get the desired order.

The following statements produce Output 49.2.1:

```

data multi;
  input Prep $ Dose Symptoms $ N;
  LDose=log10(Dose);
  if Prep='test' then PrepDose=LDose;
  else PrepDose=0;
  datalines;
stand    10    None    33
stand    10    Mild    7
stand    10    Severe  10
stand    20    None    17
stand    20    Mild    13
stand    20    Severe  17
stand    30    None    14
stand    30    Mild    3
stand    30    Severe  28
stand    40    None    9
stand    40    Mild    8
stand    40    Severe  32
test     10    None    44
test     10    Mild    6
test     10    Severe  0
test     20    None    32

```

```

test      20      Mild      10
test      20      Severe     12
test      30      None       23
test      30      Mild       7
test      30      Severe    21
test      40      None      16
test      40      Mild       6
test      40      Severe    19
;

proc probit order=data;
  class Prep Symptoms;
  nonpara: model Symptoms=Prep LDose PrepDose / lackfit;
  weight N;
  parallel: model Symptoms=Prep LDose / lackfit;
  weight N;
  title 'Probit Models for Symptom Severity';
run;

```

The first model uses the PrepDose variable to allow for nonparallelism between the dose response curves for the two preparations. The results of this first model indicate that the parameter for the PrepDose variable is not significant, having a Wald chi-square of 0.73. Also, since the first model is a generalization of the second, a likelihood ratio test statistic for this same parameter can be obtained by multiplying the difference in log likelihoods between the two models by 2. The value obtained,  $2 \times (-345.94 - (-346.31))$ , is 0.73. This is in close agreement with the Wald chi-square from the first model. The lack-of-fit test statistics for the two models do not indicate a problem with either fit.

#### Output 49.2.1. Multilevel Response: PROC PROBIT

Probit Models for Symptom Severity		
Probit Procedure		
Class Level Information		
Name	Levels	Values
Symptoms	3	None Mild Severe
Prep	2	stand test

Output 49.2.1. (continued)

```

Probit Models for Symptom Severity

Probit Procedure

Model Information

Data Set                WORK.MULTI
Dependent Variable      Symptoms
Weight Variable         N
Number of Observations  23
Name of Distribution     NORMAL
Log Likelihood          -345.9401767

Weighted Frequency Counts for the Ordered Response Categories

Level      Count
None       188
Mild       60
Severe     139

Goodness-of-Fit Tests

Statistic                Value      DF      Pr > ChiSq
Pearson Chi-Square      12.7930   11      0.3071
L.R.    Chi-Square      15.7869   11      0.1492

Response-Covariate Profile

Response Levels          3
Number of Covariate Values  8
    
```

Output 49.2.1. (continued)

```

Probit Models for Symptom Severity

Probit Procedure

Analysis of Parameter Estimates

Variable      DF      Estimate      Standard
              Error Chi-Square Pr > ChiSq Label
Intercept    1      3.80803      0.62517      37.1030      <.0001 Intercept
Prep         1      -1.25728     0.81897      2.3568      0.1247
              0      0            0            .            . stand
              test
LDose       1      -2.15120     0.39088      30.2874      <.0001
PrepDose    1      -0.50722     0.59449      0.7279      0.3935
Inter.2     1      0.46844     0.05591      .            . Mild
    
```

## Output 49.2.1. (continued)

Probit Models for Symptom Severity		
Probit Procedure		
Class Level Information		
Name	Levels	Values
Symptoms	3	None Mild Severe
Prep	2	stand test

## Output 49.2.1. (continued)

Probit Models for Symptom Severity			
Probit Procedure			
Model Information			
Data Set	WORK.MULTI		
Dependent Variable	Symptoms		
Weight Variable	N		
Number of Observations	23		
Name of Distribution	NORMAL		
Log Likelihood	-346.306141		
Weighted Frequency Counts for the Ordered Response Categories			
	Level	Count	
	None	188	
	Mild	60	
	Severe	139	
Goodness-of-Fit Tests			
Statistic	Value	DF	Pr > ChiSq
Pearson Chi-Square	12.7864	12	0.3848
L.R. Chi-Square	16.5189	12	0.1686
Response-Covariate Profile			
Response Levels	3		
Number of Covariate Values	8		

Output 49.2.1. (continued)

Probit Models for Symptom Severity						
Probit Procedure						
Analysis of Parameter Estimates						
Variable	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq	Label
Intercept	1	3.41482	0.41260	68.4962	<.0001	Intercept
Prep	1			20.3304	<.0001	
	1	-0.56752	0.12586	20.3304	<.0001	stand
	0	0	0	.	.	test
LDose	1	-2.37213	0.29495	64.6824	<.0001	
Inter.2	1	0.46780	0.05584			Mild

The negative coefficient associated with LDose indicates that the probability of having no symptoms (Symptoms='None') or no or mild symptoms (Symptoms='None' or Symptoms='Mild') decreases as LDose increases; that is, the probability of a severe symptom increases with LDose. This association is apparent for both treatment groups.

The negative coefficient associated with the standard treatment group (Prep = stand) indicates that the standard treatment is associated with more severe symptoms across all ldose values.

### Example 49.3. Logistic Regression

In this example, a series of people are questioned as to whether or not they would subscribe to a new newspaper. For each person, the variables sex (Female, Male), age, and subs (1=yes,0=no) are recorded. The PROBIT procedure is used to fit a logistic regression model to the probability of a positive response (subscribing) as a function of the variables sex and age. Specifically, the probability of subscribing is modeled as

$$p = \Pr(\text{subs} = 1) = F(b_0 + b_1 \times \text{sex} + b_2 \times \text{age})$$

where  $F$  is the cumulative logistic distribution function.

By default, the PROBIT procedure models the probability of the lower response level for binary data. One way to model  $\Pr(\text{subs} = 1)$  is to format the response variable so that the formatted value corresponding to subs=1 is the lower level. The following statements format the values of subs as 1 = 'accept' and 0 = 'reject', so that PROBIT models  $\Pr(\text{accept}) = \Pr(\text{subs} = 1)$ .

The following statements produce Output 49.3.1:

```
data news;
  input sex $ age subs;
  datalines;
Female    35    0
Male     44    0
Male     45    1
```

```
Female      47      1
Female      51      0
Female      47      0
Male        54      1
Male        47      1
Female      35      0
Female      34      0
Female      48      0
Female      56      1
Male        46      1
Female      59      1
Female      46      1
Male        59      1
Male        38      1
Female      39      0
Male        49      1
Male        42      1
Male        50      1
Female      45      0
Female      47      0
Female      30      1
Female      39      0
Female      51      0
Female      45      0
Female      43      1
Male        39      1
Male        31      0
Female      39      0
Male        34      0
Female      52      1
Female      46      0
Male        58      1
Female      50      1
Female      32      0
Female      52      1
Female      35      0
Female      51      0
;

proc format;
  value subscrib 1 = 'accept' 0 = 'reject';
run;

proc probit;
  class subs sex;
  model subs=sex age / d=logistic itprint;
  format subs subscrib.;
  title 'Logistic Regression of Subscription Status';
run;
```

**Output 49.3.1.** Logistic Regression: PROC PROBIT

```

Logistic Regression of Subscription Status

      Probit Procedure

      Class Level Information

      Name          Levels  Values
      ----          -
      subs           2      accept reject
      sex            2      Female Male
    
```

**Output 49.3.1.** (continued)

```

Logistic Regression of Subscription Status

      Probit Procedure

      Iteration History for Parameter Estimates

      Iter   Ridge   Loglikelihood      Intercept          age          sex.1
      ----   ----   -
      0       0      -27.725887         0                0                0
      1       0      -20.142659      -3.634567629     0.1051634384    -1.648455751
      2       0      -19.52245        -5.254865196     0.1506493473    -2.234724956
      3       0      -19.490439      -5.728485385     0.1639621828    -2.409827238
      4       0      -19.490303      -5.76187293      0.1649007124    -2.422349862
      5       0      -19.490303      -5.7620267       0.1649050312    -2.422407743

      Model Information

      Data Set          WORK.NEWS
      Dependent Variable      subs
      Number of Observations      40
      Name of Distribution      LOGISTIC
      Log Likelihood          -19.49030281

      Weighted Frequency Counts for the Ordered Response Categories

      Level          Count
      ----          -
      accept          20
      reject          20
    
```

## Output 49.3.1. (continued)

```

Logistic Regression of Subscription Status

Probit Procedure

Last Evaluation of the Negative of the Gradient

Intercept      sex.1      age
-5.95379E-12   8.76834E-10 -1.636692E-8

Last Evaluation of the Negative of the Hessian

Intercept      sex.1      age
Intercept      6.4597397447 4.6042218284 292.04051848
sex.1          4.6042218284 4.6042218284 216.20829515
age           292.04051848 216.20829515 13487.329973

Algorithm converged.

```

## Output 49.3.1. (continued)

```

Logistic Regression of Subscription Status

Probit Procedure

Analysis of Parameter Estimates

Variable  DF  Estimate  Standard
          DF  Error Chi-Square Pr > ChiSq Label
Intercept 1  -5.76203  2.76345  4.3476  0.0371 Intercept
sex        1  -2.42241  0.95590  6.4220  0.0113
           1  -2.42241  0.95590  6.4220  0.0113 Female
           0  0  0  .  . Male
age        1  0.16491  0.06519  6.3992  0.0114

```

From Output 49.3.1, there appears to be an effect due to both the variables sex and age. The positive coefficient for age indicates that older people are more likely to subscribe than younger people. The negative coefficient for sex indicates that females are less likely to subscribe than males.

---

## References

- Agresti, A. (1990), *Categorical Data Analysis*, New York: John Wiley & Sons.
- Collett, D (1991), *Modelling Binary Data*, London: Chapman and Hall.
- Cox, D.R. (1970), *Analysis of Binary Data*, London: Chapman and Hall.
- Cox, D.R. and Oakes, D. (1984), *Analysis of Survival Data*, London: Chapman and Hall.
- Finney, D.J. (1971), *Probit Analysis*, Third Edition, London: Cambridge University Press.

Hubert, J. J., Bohidar, N. R., and Peace, K. E. (1988), "Assessment of Pharmacological Activity," *Biopharmaceutical Statistics for Drug Development*, ed. K. E. Peace, New York: Marcel Dekker.