SAS/STAT® 15.1 User's Guide

The SURVEYLOGISTIC Procedure

This document is an individual chapter from *SAS/STAT® 15.1 User's Guide*.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2018. *SAS/STAT® 15.1 User's Guide*. Cary, NC: SAS Institute Inc.

SAS/STAT® 15.1 User's Guide

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

November 2018

Chapter 117
The SURVEYLOGISTIC Procedure

Contents

Overview: SURVEYLOGISTIC Procedure

Categorical responses arise extensively in sample survey. Common examples of responses include the following:

- binary: for example, attended graduate school or not
- ordinal: for example, mild, moderate, and severe pain
- nominal: for example, ABC, NBC, CBS, FOX TV network viewed at a certain hour

Logistic regression analysis is often used to investigate the relationship between such discrete responses and a set of explanatory variables. For a description of logistic regression for sample survey data, see Binder (1981, 1983); Roberts, Rao, and Kumar (1987); Skinner, Holt, and Smith (1989); Morel (1989); Lehtonen and Pahkinen (1995).

For binary response models, the response of a sampling unit can take a specified value or not (for example, attended graduate school or not). Suppose $\mathbf{x}$ is a row vector of explanatory variables and π is the response probability to be modeled. The linear logistic model has the form

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

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

The logistic model shares a common feature with the more general class of generalized linear models—namely, that a function $g = g(\mu)$ of the expected value, μ, of the response variable is assumed to be linearly related to the explanatory variables. Since μ implicitly depends on the stochastic behavior of the response, and since t he 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. For this reason, Nelder and Wedderburn (1972) refer to $g(\cdot)$ 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 SURVEYLOGISTIC procedure enables you to choose one of these link functions, resulting in fitting a broad class of binary response models of the form

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

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

For example, pain severity can be classified into three response categories as 1=mild, 2=moderate, and 3=severe. The SURVEYLOGISTIC procedure fits a common slopes cumulative model, which is a parallel lines regression model based on the cumulative probabilities of the response categories rather than on their individual probabilities. The cumulative model has the form

$$g(\Pr(Y \leq d \mid \mathbf{x})) = \alpha_d + \mathbf{x}\boldsymbol{\beta}, \quad 1 \leq d \leq D$$

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.

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

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

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

The SURVEYLOGISTIC procedure fits linear logistic regression models for discrete response survey data by the method of maximum likelihood. For statistical inferences, PROC SURVEYLOGISTIC incorporates complex survey sample designs, including designs with stratification, clustering, and unequal weighting.

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 ordinal logistic regression models can be replaced by the probit function or the complementary log-log function.

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.

Variances of the regression parameters and odds ratios are computed by using either the Taylor series (linearization) method or replication (resampling) methods to estimate sampling errors of estimators based on complex sample designs (Binder 1983; Särndal, Swensson, and Wretman 1992; Wolter 2007; Rao, Wu, and Yue 1992).

The SURVEYLOGISTIC 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 LOGISTIC procedure.

Like many procedures in SAS/STAT software that allow the specification of CLASS variables, the SURVEYLOGISTIC 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.

Getting Started: SURVEYLOGISTIC Procedure

The SURVEYLOGISTIC procedure is similar to the LOGISTIC procedure and other regression procedures in the SAS System. See Chapter 76, "The LOGISTIC Procedure," for general information about how to perform logistic regression by using SAS. PROC SURVEYLOGISTIC is designed to handle sample survey data, and thus it incorporates the sample design information into the analysis.

The following example illustrates how to use PROC SURVEYLOGISTIC to perform logistic regression for sample survey data.

In the customer satisfaction survey example in the section "Getting Started: SURVEYSELECT Procedure" on page 10077 in Chapter 121, "The SURVEYSELECT Procedure," an Internet service provider conducts a customer satisfaction survey. The survey population consists of the company's current subscribers from four states: Alabama (AL), Florida (FL), Georgia (GA), and South Carolina (SC). The company plans to select a sample of customers from this population, interview the selected customers and ask their opinions on customer service, and then make inferences about the entire population of subscribers from the sample data. A stratified sample is selected by using the probability proportional to size (PPS) method. The sample design divides the customers into strata depending on their types ('Old' or 'New') and their states (AL, FL, GA, SC). There are eight strata in all. Within each stratum, customers are selected and interviewed by using the PPS with replacement method, where the size variable is Usage. The stratified PPS sample contains 192 customers. The data are stored in the SAS data set SampleStrata. Figure 117.1 displays the first 10 observations of this data set.

Figure 117.1 Stratified PPS Sample (First 10 Observations)

Customer Satisfaction Survey
Stratified PPS Sampling
(First 10 Observations)

Obs	State	Type	CustomerID	Rating	Usage	SamplingWeight
1	AL	New	24394278	Neutral	13.17	26.358
2	AL	New	64798692	Extremely Unsatisfied	15.53	22.352
3	AL	New	75375074	Unsatisfied	99.11	3.501
4	AL	New	262831809	Neutral	5.40	64.228
5	AL	New	294428658	Extremely Satisfied	1.17	297.488
6	AL	New	336222949	Unsatisfied	38.69	8.970
7	AL	New	351929023	Extremely Satisfied	2.72	127.475
8	AL	New	366142640	Satisfied	2.61	132.958
9	AL	New	371478614	Neutral	14.36	24.173
10	AL	New	477172230	Neutral	4.06	85.489

In the SAS data set SampleStrata, the variable CustomerID uniquely identifies each customer. The variable State contains the state of the customer's address. The variable Type equals 'Old' if the customer has subscribed to the service for more than one year; otherwise, the variable Type equals 'New'. The variable Usage contains the customer's average monthly service usage, in hours. The variable Rating contains the customer's responses to the survey. The sample design uses an unequal probability sampling method, with the sampling weights stored in the variable SamplingWeight.

The following SAS statements fit a cumulative logistic model between the satisfaction levels and the Internet usage by using the stratified PPS sample:

```
title 'Customer Satisfaction Survey';
proc surveylogistic data=SampleStrata;
   strata state type/list;
   model Rating (order=internal) = Usage;
   weight SamplingWeight;
run;
```

The PROC SURVEYLOGISTIC statement invokes the SURVEYLOGISTIC procedure. The STRATA statement specifies the stratification variables State and Type that are used in the sample design. The LIST option requests a summary of the stratification. In the MODEL statement, Rating is the response variable and Usage is the explanatory variable. The ORDER=internal is used for the response variable Rating to ask the procedure to order the response levels by using the internal numerical value (1–5) instead of the formatted character value. The WEIGHT statement specifies the variable SamplingWeight that contains the sampling weights.

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

Figure 117.2 Stratified PPS Sample, Model Information

Customer Satisfaction Survey

The SURVEYLOGISTIC Procedure

Model Information		
Data Set	WORK.SAMPLESTRATA	
Response Variable	Rating	
Number of Response Levels	5	
Stratum Variables	State	
	Type	
Number of Strata	8	
Weight Variable	SamplingWeight	Sampling Weight
Model	Cumulative Logit	
Optimization Technique	Fisher's Scoring	
Variance Adjustment	Degrees of Freedom (DF)	

PROC SURVEYLOGISTIC first lists the following model fitting information and sample design information in Figure 117.2:

- The link function is the logit of the cumulative of the lower response categories.
- The Fisher scoring optimization technique is used to obtain the maximum likelihood estimates for the regression coefficients.
- The response variable is Rating, which has five response levels.
- The stratification variables are State and Type.
- There are eight strata in the sample.
- The weight variable is SamplingWeight.

- The variance adjustment method used for the regression coefficients is the default degrees of freedom adjustment.

Figure 117.3 lists the number of observations in the data set and the number of observations used in the analysis. Since there is no missing value in this example, observations in the entire data set are used in the analysis. The sums of weights are also reported in this table.

Figure 117.3 Stratified PPS Sample, Number of Observations

Number of Observations Read	194
Number of Observations Used	194
Sum of Weights Read	14200.58
Sum of Weights Used	14200.58

The “Response Profile” table in Figure 117.4 lists the five response levels, their ordered values, and their total frequencies and total weights for each category. Due to the ORDER=INTERNAL option for the response variable Rating, the category “Extremely Unsatisfied” has the Ordered Value 1, the category “Unsatisfied” has the Ordered Value 2, and so on.

Figure 117.4 Stratified PPS Sample, Response Profile

Response Profile

Ordered Value	Rating	Total Frequency	Total Weight
1	Extremely Unsatisfied	61	2049.6278
2	Unsatisfied	42	3084.6615
3	Neutral	49	2295.9993
4	Satisfied	33	3754.7221
5	Extremely Satisfied	9	3015.5727

Probabilities modeled are cumulated over the lower Ordered Values.

Figure 117.5 displays the output of the stratification summary. There are a total of eight strata, and each stratum is defined by the customer types within each state. The table also shows the number of customers within each stratum.

Figure 117.5 Stratified PPS Sample, Stratification Summary

Stratum Information

Stratum Index	State	Type	N Obs
1	AL	New	24
2		Old	25
3	FL	New	25
4		Old	23
5	GA	New	24
6		Old	25
7	SC	New	24
8		Old	24

Figure 117.6 shows the iteration algorithm converged to obtain the MLE for this example. 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.

Figure 117.6 Stratified PPS Sample, Model Fitting Information

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

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	45064.144	44143.526
SC	45094.389	44181.332
-2 Log L	45056.144	44133.526

The table "Testing Global Null Hypothesis: BETA=0" in Figure 117.7 shows the likelihood ratio test, the efficient score test, and the Wald test for testing the significance of the explanatory variable (Usage). All tests are significant.

Figure 117.7 Stratified PPS Sample

Testing Global Null Hypothesis: BETA=0

Test	F Value	Num DF	Den DF	Pr > F
Likelihood Ratio	12.95	1	186	0.0004
Score	108.36	1	186	<.0001
Wald	5.93	1	186	0.0159

NOTE:
First-order Rao-Scott design correction 0.9734 applied to the likelihood ratio test.

Figure 117.8 shows the parameter estimates of the logistic regression and their standard errors.

Figure 117.8 Stratified PPS Sample, Parameter Estimates

Analysis of Maximum Likelihood Estimates

Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept	**Extremely Unsatisfied**	-2.1252	0.3727	-5.70	<.0001
Intercept	**Unsatisfied**	-0.8649	0.3616	-2.39	0.0178
Intercept	**Neutral**	-0.1768	0.3645	-0.49	0.6282
Intercept	**Satisfied**	1.0953	0.5255	2.08	0.0385
Usage		0.0441	0.0181	2.43	0.0159

NOTE: The degrees of freedom for the t tests is 186.

Figure 117.9 displays the odds ratio estimate and its confidence intervals.

Figure 117.9 Stratified PPS Sample, Odds Ratios

Odds Ratio Estimates			
Effect	Point Estimate	95% Confidence Limits	
Usage	1.045	1.008	1.083

NOTE: The degrees of freedom in computing the confidence limits is 186.

Syntax: SURVEYLOGISTIC Procedure

The following statements are available in the SURVEYLOGISTIC procedure:

```
PROC SURVEYLOGISTIC < options > ;
    BY variables ;
    CLASS variable < (v-options) > < variable < (v-options) > . . . > < / v-options > ;
    CLUSTER variables ;
    CONTRAST 'label' effect values < , . . . effect values > < / options > ;
    DOMAIN variables < variable∗variable variable∗variable∗variable . . . > ;
    EFFECT name = effect-type (variables < / options >) ;
    ESTIMATE < 'label' > estimate-specification < / options > ;
    FREQ variable ;
    LSMEANS < model-effects > < / options > ;
    LSMESTIMATE model-effect lsmestimate-specification < / options > ;
    MODEL events/trials = < effects < / options > > ;
    MODEL variable < (v-options) > = < effects > < / options > ;
    OUTPUT < OUT=SAS-data-set > < options > < / option > ;
    REPWEIGHTS variables < / options > ;
    SLICE model-effect < / options > ;
    STORE < OUT= >item-store-name < / LABEL='label' > ;
    STRATA variables < / option > ;
    < label: > TEST equation1 < , . . . , equationk > < / options > ;
    UNITS independent1 = list1 < . . . independentk = listk > < / option > ;
    WEIGHT variable ;
```

The PROC SURVEYLOGISTIC and MODEL statements are required.

The CLASS, CLUSTER, CONTRAST, EFFECT, ESTIMATE, LSMEANS, LSMESTIMATE, REPWEIGHTS, SLICE, STRATA, TEST statements can appear multiple times. You should use only one of each following statements: MODEL, WEIGHT, STORE, OUTPUT, and UNITS.

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, except the EFFECT, ESTIMATE, LSMEANS, LSMESTIMATE, SLICE, STORE statements. These statements are also available in many other procedures. Summary descriptions of functionality and syntax for these statements are shown in this chapter, and full documentation about them is available in Chapter 19, "Shared Concepts and Topics."

The syntax descriptions begin with the PROC SURVEYLOGISTIC statement; the remaining statements are covered in alphabetical order.

PROC SURVEYLOGISTIC Statement

PROC SURVEYLOGISTIC < *options* > **;**

The PROC SURVEYLOGISTIC statement invokes the SURVEYLOGISTIC procedure. Optionally, it identifies input data sets, controls the ordering of the response levels, and specifies the variance estimation method. The PROC SURVEYLOGISTIC statement is required.

Table 117.1 summarizes the *options* available in the PROC SURVEYLOGISTIC statement.

Table 117.1 PROC SURVEYLOGISTIC Statement Options

Option	Description
ALPHA=	Sets the confidence level for confidence intervals
DATA=	Names the SAS data set containing the data to be analyzed
INEST=	Names the SAS data set that contains initial estimates
MAXRESPONSELEVELS=	Specifies the maximum number of response levels allowed
MISSING	Treats missing values as a valid category
NAMELEN=	Specifies the length of effect names
NOMCAR	Treats missing values as *not missing completely at random*
NOSORT	Suppresses the internal sorting process
ORDER=	Specifies the sort order
RATE=	Specifies the sampling rate
TOTAL=	Specifies the total number of primary sampling units
VARMETHOD=	Specifies the variance estimation method

ALPHA=*value*

sets the confidence level for confidence intervals. The value of the ALPHA= option must be between 0 and 1, and the default value is 0.05. A confidence level of α produces $100(1-\alpha)\%$ confidence intervals. The default of ALPHA=0.05 produces 95% confidence intervals.

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.

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 9734 for more information.

MAXRESPONSELEVELS=*number*

specifies the maximum number of response levels that are allowed in your data set. By default, MAXRESPONSELEVELS=100. If you have more response levels than the maximum number allowed, then a message is displayed in the SAS log that provides the value of *number* required to continue the analysis, and the procedure stops.

MISSING

treats missing values as a valid (nonmissing) category for all categorical variables, which include CLASS, STRATA, CLUSTER, and DOMAIN variables.

By default, if you do not specify the MISSING option, an observation is excluded from the analysis if it has a missing value. For more information, see the section "Missing Values" on page 9724.

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.

NOMCAR

requests that the procedure treat missing values in the variance computation as *not missing completely at random* (NOMCAR) for Taylor series variance estimation. When you specify the NOMCAR option, PROC SURVEYLOGISTIC computes variance estimates by analyzing the nonmissing values as a domain or subpopulation, where the entire population includes both nonmissing and missing domains. See the section "Missing Values" on page 9724 for more details.

By default, PROC SURVEYLOGISTIC completely excludes an observation from analysis if that observation has a missing value, unless you specify the MISSING option. Note that the NOMCAR option has no effect on a classification variable when you specify the MISSING option, which treats missing values as a valid nonmissing level.

The NOMCAR option applies only to Taylor series variance estimation; it is ignored for replication methods.

NOSORT

suppresses the internal sorting process to shorten the computation time if the data set is presorted by the STRATA and CLUSTER variables. By default, the procedure sorts the data by the STRATA variables if you use the STRATA statement; then the procedure sorts the data by the CLUSTER variables within strata. If your data are already stored by the order of STRATA and CLUSTER variables, then you can specify this option to omit this sorting process to reduce the usage of computing resources, especially when your data set is very large. However, if you specify this NOSORT option while your data are not presorted by STRATA and CLUSTER variables, then any changes in these variables creates a new stratum or cluster.

ORDER=DATA | FORMATTED | FREQ | INTERNAL

specifies the sort order for the levels of the response variable. This option, except for ORDER=FREQ, also determines the sort order for the levels of ClUSTER and DOMAIN variables and controls STRATA variable levels in the "Stratum Information" table. By default, ORDER=INTERNAL. However, if an ORDER= option is specified after the response variable, in the MODEL statement, it overrides this option for the response variable. This option does not affect the ordering of the CLASS variable levels; see the ORDER= option in the CLASS statement for more information.

RATE=*value* | *SAS-data-set*

R=*value* | *SAS-data-set*

specifies the sampling rate, which PROC SURVEYLOGISTIC uses to compute a finite population correction for Taylor series or bootstrap variance estimation. This option is ignored for the jackknife or balanced repeated replication (BRR) variance estimation method.

If your sample design has multiple stages, you should specify the *first-stage sampling rate*, which is the ratio of the number of primary sampling units (PSUs) in the sample to the total number of PSUs in the population.

You can specify the sampling rate in either of the following ways:

value specifies a nonnegative number to use for a nonstratified design or for a stratified design that has the same sampling rate in each stratum.

SAS-data-set specifies a *SAS-data-set* that contains the stratification variables and the sampling rates for a stratified design that has different sampling rates in the strata. You must provide the sampling rates in the data set variable named _RATE_. The sampling rates must be nonnegative numbers.

You can specify sampling rates as numbers between 0 and 1. Or you can specify sampling rates in percentage form as numbers between 1 and 100, which PROC SURVEYLOGISTIC converts to proportions. The procedure treats the value 1 as 100% instead of 1%.

For more information, see the section "Specification of Population Totals and Sampling Rates" on page 9735.

If you do not specify either the RATE= or TOTAL= option, the Taylor series or bootstrap variance estimation does not include a finite population correction. You cannot specify both the RATE= and TOTAL= options.

TOTAL=*value* | *SAS-data-set*

N=*value* | *SAS-data-set*

specifies the total number of primary sampling units (PSUs) in the study population. PROC SURVEYLOGISTIC uses this information to compute a finite population correction for Taylor series or bootstrap variance estimation. This option is ignored for the jackknife or BRR variance estimation method.

You can specify the total number of PSUs in either of the following ways:

value specifies a positive number to use for a nonstratified design or for a stratified design that has the same population total in each stratum.

SAS-data-set specifies a *SAS-data-set* that contains the stratification variables and the population totals for a stratified design that has different population totals in the strata. You must provide the stratum totals in the data set variable named _TOTAL_. The stratum totals must be positive numbers.

For more information, see the section "Specification of Population Totals and Sampling Rates" on page 9735.

If you do not specify either the TOTAL= or RATE= option, the Taylor series or bootstrap variance estimation does not include a finite population correction. You cannot specify both the TOTAL= and RATE= options.

VARMETHOD=*method* **<** **(***method-options***)** **>**

specifies the variance estimation *method*. PROC SURVEYLOGISTIC provides the Taylor series method and the following replication (resampling) methods: balanced repeated replication (BRR), bootstrap, and jackknife.

Table 117.2 summarizes the available *methods* and *method-options*.

Table 117.2 Variance Estimation Methods

method	**Variance Estimation Method**	*method-options*
BOOTSTRAP	Bootstrap	CENTER= MH=*value* | *SAS-data-set* OUTWEIGHTS=*SAS-data-set* REPS=*number* SEED=*number*
BRR	Balanced repeated replication	CENTER= FAY <=*value*> HADAMARD=*SAS-data-set* OUTWEIGHTS=*SAS-data-set* PRINTH REPS=*number*
JACKKNIFE | JK	Jackknife	CENTER= OUTJKCOEFS=*SAS-data-set* OUTWEIGHTS=*SAS-data-set*
TAYLOR	Taylor series linearization	None

For VARMETHOD=BOOTSTRAP, VARMETHOD=BRR, and VARMETHOD=JACKKNIFE, you can specify *method-options* in parentheses after the variance estimation *method*. For example:

```
varmethod=BRR(reps=60 outweights=myReplicateWeights)
```

By default, VARMETHOD=JACKKNIFE if you also specify a REPWEIGHTS statement; otherwise, VARMETHOD=TAYLOR by default.

You can specify the following *methods*:

BOOTSTRAP <(*method-options*)>

requests variance estimation by the bootstrap method. For more information, see the section "Bootstrap Method" on page 9741.

The bootstrap method requires at least two primary sampling units (PSUs) in each stratum for stratified designs unless you use a REPWEIGHTS statement to provide replicate weights.

You can specify the following *method-options*:

CENTER=FULLSAMPLE | REPLICATES

defines how to compute the deviations for the bootstrap method. You can specify the following values:

FULLSAMPLE computes the deviations of the replicate estimates from the full sample estimate.

REPLICATES computes the deviations of the replicate estimates from the average of the replicate estimates.

By default, CENTER=FULLSAMPLE. For more information, see the section "Bootstrap Method" on page 9741.

MH=*value* | **(***values***)** | *SAS-data-set*

specifies the number of PSUs to select for the bootstrap replicate samples. You can provide bootstrap stratum sample sizes m_h by specifying a list of *values* or a *SAS-data-set*. Alternatively, you can provide a single bootstrap sample size *value* to use for all strata or for a nonstratified design. You can specify the number of replicate samples in the REPS= option. For more information, see the section "Bootstrap Method" on page 9741.

Each bootstrap sample size m_h must be a positive integer and must be less than n_h, which is the total number of PSUs in stratum h. By default, $m_h = n_h - 1$ for a stratified design. For a nonstratified design, the bootstrap sample size *value* must be less than n (the total number of PSUs in the sample). By default, $m = n - 1$ for a nonstratified design.

You can provide bootstrap sample sizes by specifying one of the following forms:

MH=*value*

specifies a single bootstrap sample size *value* to use for all strata or for a nonstratified design.

MH=(*values***)**

specifies a list of stratum bootstrap sample size *values*. You can separate the values with blanks or commas, and you must enclose the list of values in parentheses. The number of values must not be less than the number of strata in the DATA= input data set.

Each stratum sample size value must be a positive integer and must be less than the total number of PSUs in the corresponding stratum.

MH=*SAS-data-set*

names a *SAS-data-set* that contains the stratum bootstrap sample sizes. You must provide the sample sizes in a data set variable named _NSIZE_ or SampleSize.

The *SAS-data-set* must contain all stratification variables that you specify in the STRATA statement. It must also contain all stratum levels that appear in the DATA= input data set. If formats are associated with the STRATA variables, the formats must be consistent in the two data sets.

Each value of the _NSIZE_ or SampleSize variable must be a positive integer and must be less than the total number of PSUs in the corresponding stratum.

OUTWEIGHTS=*SAS-data-set*

names a *SAS-data-set* in which to store the bootstrap replicate weights that PROC SURVEYLOGISTIC creates. For information about replicate weights, see the section "Bootstrap Method" on page 9741. For information about the contents of the OUTWEIGHTS= data set, see the section "Replicate Weights Output Data Set" on page 9759.

This *method-option* is not available when you provide replicate weights in a REPWEIGHTS statement.

REPS=*number*

specifies the *number* of replicates for bootstrap variance estimation. The value of *number* must be an integer greater than 1. Increasing the number of replicates improves the estimation precision but also increases the computation time. By default, REPS=250.

SEED=*number*

specifies the initial seed for random number generation for bootstrap replicate sampling.

If you do not specify this option or if you specify a *number* that is negative or 0, PROC SURVEYLOGISTIC uses the time of day from the system clock to obtain an initial seed.

To reproduce the same bootstrap replicate weights and the same analysis in a subsequent execution of PROC SURVEYLOGISTIC, you can specify the same initial seed that was used in the original analysis.

PROC SURVEYLOGISTIC displays the value of the initial seed in the "Variance Estimation" table.

BRR < (*method-options***) >**

requests variance estimation by balanced repeated replication (BRR). This method requires a stratified sample design where each stratum contains two primary sampling units (PSUs). When you specify this method, you must also specify a STRATA statement unless you provide replicate weights by using the REPWEIGHTS statement. For more information, see the section "Balanced Repeated Replication (BRR) Method" on page 9743.

You can specify the following *method-options*:

CENTER=FULLSAMPLE | REPLICATES

defines how to compute the deviations for the bootstrap method. You can specify the following values:

FULLSAMPLE computes the deviations of the replicate estimates from the full sample estimate.

REPLICATES computes the deviations of the replicate estimates from the average of the replicate estimates.

By default, CENTER=FULLSAMPLE. For more information, see the section "Balanced Repeated Replication (BRR) Method" on page 9743.

FAY < =*value* **>**

requests Fay's method, which is a modification of the BRR method. For more information, see the section "Fay's BRR Method" on page 9744.

You can specify the *value* of the Fay coefficient, which is used in converting the original sampling weights to replicate weights. The Fay coefficient must be a nonnegative number less than 1. By default, the Fay coefficient is 0.5.

HADAMARD=*SAS-data-set*

H=*SAS-data-set*

names a *SAS-data-set* that contains the Hadamard matrix for BRR replicate construction. If you do not specify this *method-option*, PROC SURVEYLOGISTIC generates an appropriate Hadamard matrix for replicate construction. For more information, see the sections "Balanced Repeated Replication (BRR) Method" on page 9743 and "Hadamard Matrix" on page 9747.

If a Hadamard matrix of a particular dimension exists, it is not necessarily unique. Therefore, if you want to use a specific Hadamard matrix, you must provide the matrix as a *SAS-data-set* in this *method-option*.

In this *SAS-data-set*, each variable corresponds to a column and each observation corresponds to a row of the Hadamard matrix. You can use any variable names in this data set. All values in the data set must equal either 1 or –1. You must ensure that the matrix you provide is indeed a Hadamard matrix—that is, $\mathbf{A}'\mathbf{A} = R\,\mathbf{I}$, where $\mathbf{A}$ is the Hadamard matrix of dimension R and $\mathbf{I}$ is an identity matrix. PROC SURVEYLOGISTIC does not check the validity of the Hadamard matrix that you provide.

The *SAS-data-set* must contain at least H variables, where H denotes the number of first-stage strata in your design. If the data set contains more than H variables, PROC SURVEYLOGISTIC uses only the first H variables. Similarly, this data set must contain at least H observations.

If you do not specify the REPS= *method-option*, the number of replicates is assumed to be the number of observations in the *SAS-data-set*. If you specify the number of replicates—for example, REPS=*nreps*—the first *nreps* observations in the *SAS-data-set* are used to construct the replicates.

You can specify the PRINTH *method-option* to display the Hadamard matrix that PROC SURVEYLOGISTIC uses to construct replicates for BRR.

OUTWEIGHTS=*SAS-data-set*

names a *SAS-data-set* in which to store the replicate weights that PROC SURVEYLOGISTIC creates for BRR variance estimation. For information about replicate weights, see the section "Balanced Repeated Replication (BRR) Method" on page 9743. For information about the contents of the OUTWEIGHTS= data set, see the section "Replicate Weights Output Data Set" on page 9759.

This *method-option* is not available when you provide replicate weights in a REPWEIGHTS statement.

PRINTH

displays the Hadamard matrix that PROC SURVEYLOGISTIC uses to construct replicates for BRR variance estimation. When you provide the Hadamard matrix in the HADAMARD= *method-option*, PROC SURVEYLOGISTIC displays only the rows and columns that are actually used to construct replicates. For more information, see the sections "Balanced Repeated Replication (BRR) Method" on page 9743 and "Hadamard Matrix" on page 9747.

The PRINTH *method-option* is not available when you provide replicate weights in a REPWEIGHTS statement because the procedure does not use a Hadamard matrix in this case.

REPS=*number*

specifies the *number* of replicates for BRR variance estimation. The value of *number* must be an integer greater than 1.

If you do not use the HADAMARD= *method-option* to provide a Hadamard matrix, the number of replicates should be greater than the number of strata and should be a multiple of 4. For more information, see the section "Balanced Repeated Replication (BRR) Method" on page 9743. If PROC SURVEYLOGISTIC cannot construct a Hadamard matrix for the REPS= value that you specify, the value is increased until a Hadamard matrix of that dimension can be constructed. Therefore, the actual number of replicates that PROC SURVEYLOGISTIC uses might be larger than *number*.

If you use the HADAMARD= *method-option* to provide a Hadamard matrix, the value of *number* must not be greater than the number of rows in the Hadamard matrix. If you provide a Hadamard matrix and do not specify the REPS= *method-option*, the number of replicates is the number of rows in the Hadamard matrix.

If you do not specify the REPS= or the HADAMARD= *method-option* and do not use a REPWEIGHTS statement, the number of replicates is the smallest multiple of 4 that is greater than the number of strata.

If you use a REPWEIGHTS statement to provide replicate weights, PROC SURVEYLOGISTIC does not use the REPS= *method-option*; the number of replicates is the number of REPWEIGHTS variables.

JACKKNIFE < (*method-options*) >

JK < (*method-options*) >

requests variance estimation by the delete-1 jackknife method. For more information, see the section "Jackknife Method" on page 9746. If you use a REPWEIGHTS statement to provide replicate weights, VARMETHOD=JACKKNIFE is the default variance estimation method.

The delete-1 jackknife method requires at least two primary sampling units (PSUs) in each stratum for stratified designs unless you use a REPWEIGHTS statement to provide replicate weights.

You can specify the following *method-options*:

CENTER=FULLSAMPLE | REPLICATES

defines how to compute the deviations for the bootstrap method. You can specify the following values:

FULLSAMPLE computes the deviations of the replicate estimates from the full sample estimate.

REPLICATES computes the deviations of the replicate estimates from the average of the replicate estimates.

By default, CENTER=FULLSAMPLE. For more information, see the section "Jackknife Method" on page 9746.

OUTJKCOEFS=*SAS-data-set*

names a *SAS-data-set* in which to store the jackknife coefficients. For information about jackknife coefficients, see the section "Jackknife Method" on page 9746. For information about the contents of the OUTJKCOEFS= data set, see the section "Jackknife Coefficients Output Data Set" on page 9760.

OUTWEIGHTS=*SAS-data-set*

names a *SAS-data-set* in which to store the replicate weights that PROC SURVEYLOGISTIC creates for jackknife variance estimation. For information about replicate weights, see the section "Jackknife Method" on page 9746. For information about the contents of the OUTWEIGHTS= data set, see the section "Replicate Weights Output Data Set" on page 9759.

This *method-option* is not available when you use a REPWEIGHTS statement to provide replicate weights.

TAYLOR

requests Taylor series variance estimation. This is the default method if you do not specify the VARMETHOD= option or a REPWEIGHTS statement. For more information, see the section "Taylor Series (Linearization)" on page 9740.

BY Statement

BY *variables* **;**

You can specify a BY statement in PROC SURVEYLOGISTIC to obtain separate analyses of observations in groups that are 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 you specify more than one BY statement, only the last one specified is used.

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

- Sort the data by using the SORT procedure with a similar BY statement.
- Specify the NOTSORTED or DESCENDING option in the BY statement in the SURVEYLOGISTIC 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 by using the DATASETS procedure (in Base SAS software).

Note that using a BY statement provides completely separate analyses of the BY groups. It does not provide a statistically valid domain (subpopulation) analysis, where the total number of units in the subpopulation is not known with certainty. You should use the DOMAIN statement to obtain domain analysis. For more information about subpopulation analysis for sample survey data, see Cochran (1977).

For more information about BY-group processing, see the discussion in *SAS Language Reference: Concepts*. For more information about the DATASETS procedure, see the discussion in the *Base 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 sort 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 order in which to sort the levels of the classification variables. This option applies to the levels for all classification variables, except when you use the (default) ORDER=FORMATTED option with numeric classification variables that have no explicit format. In that case, the levels of such variables are ordered by their internal value.

The ORDER= option can take the following values:

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

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

For more information about sort order, see the chapter on the SORT procedure in the *Base 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.

EFFECT specifies effect coding

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

ORDINAL specifies the cumulative parameterization for an ordinal CLASS variable

POLYNOMIAL | POLY specifies polynomial coding

REFERENCE | REF specifies reference cell coding

ORTHEFFECT orthogonalizes PARAM=EFFECT

ORTHORDINAL | ORTHOTHERM orthogonalizes PARAM=ORDINAL

ORTHPOLY orthogonalizes PARAM=POLYNOMIAL

ORTHREF orthogonalizes PARAM=REFERENCE

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

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

REFERENCE=*'level'* | *keyword*

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

CLUSTER Statement

CLUSTER *variables* **;**

The CLUSTER statement names variables that identify the clusters in a clustered sample design. The combinations of categories of CLUSTER variables define the clusters in the sample. If there is a STRATA statement, clusters are nested within strata.

If you provide replicate weights for BRR or jackknife variance estimation with the REPWEIGHTS statement, you do not need to specify a CLUSTER statement.

If your sample design has clustering at multiple stages, you should identify only the first-stage clusters (primary sampling units (PSUs)), in the CLUSTER statement. See the section "Primary Sampling Units (PSUs)" on page 9735 for more information.

The CLUSTER *variables* are one or more variables in the DATA= input data set. These variables can be either character or numeric. The formatted values of the CLUSTER variables determine the CLUSTER variable levels. Thus, you can use formats to group values into levels. See the FORMAT procedure in the *Base SAS Procedures Guide* and the FORMAT statement and SAS formats in *SAS Formats and Informats: Reference* for more information.

When determining levels of a CLUSTER variable, an observation with missing values for this CLUSTER variable is excluded, unless you specify the MISSING option. For more information, see the section "Missing Values" on page 9724.

You can use multiple CLUSTER statements to specify cluster variables. The procedure uses variables from all CLUSTER statements to create clusters.

CONTRAST Statement

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

where a *row-description* is defined as follows:

effect values < , ... , *effect values* >

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

The CONTRAST statement enables you to specify a matrix, $\mathbf{L}$, for testing the hypothesis $\mathbf{L}\boldsymbol{\theta} = 0$, where $\boldsymbol{\theta}$ is the parameter vector. You must be familiar with the details of the model parameterization that PROC SURVEYLOGISTIC uses (for more information, see the PARAM= option in the section "CLASS Statement" on page 9696). Optionally, the CONTRAST statement enables you to estimate each row, $\mathbf{l}_i\boldsymbol{\theta}$, of $\mathbf{L}\boldsymbol{\theta}$ and test the hypothesis $\mathbf{l}_i\boldsymbol{\theta} = 0$. For more information, see the section "Testing Linear Hypotheses about the Regression Coefficients" on page 9753.

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 can be specified in the CONTRAST statement:

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

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

values are constants that are elements of the **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 that is coded CLASS variable A has four levels. Then there are three parameters ($\alpha_1, \alpha_2, \alpha_3$) that represent the first three levels, and the fourth parameter is represented by

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

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

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

or, equivalently,

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

which, in the form $\mathbf{L}\boldsymbol{\theta} = 0$, 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 SURVEYLOGISTIC finds a contrast to be nonestimable, it displays missing values in corresponding rows in the results. PROC SURVEYLOGISTIC handles missing level combinations of classification variables in the same manner as PROC LOGISTIC. Parameters corresponding to missing level combinations are not included in the model. This convention can affect the way in which you specify the **L** matrix in your CONTRAST statement. If the elements of **L** are not specified for an effect that contains a specified effect, then the elements of the specified effect are distributed over the levels of the higher-order effect just as the LOGISTIC procedure does for its CONTRAST and ESTIMATE statements. For example, suppose that the model contains effects A and B and their interaction A*B. If you specify a CONTRAST statement involving A alone, the **L** matrix contains nonzero terms for both A and A*B, since A*B contains A.

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

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

ALPHA=*value*

sets the confidence level for confidence intervals. The value of the ALPHA= option must be between 0 and 1, and the default value is 0.05. A confidence level of α produces $100(1-\alpha)\%$ confidence intervals. The default of ALPHA=0.05 produces 95% confidence intervals.

E

requests that the **L** matrix be displayed.

ESTIMATE=*keyword*

requests that each individual contrast (that is, each row, $\mathbf{l}_i\boldsymbol{\beta}$, of $\mathbf{L}\boldsymbol{\beta}$) or exponentiated contrast ($e^{\mathbf{l}_i\boldsymbol{\beta}}$) be estimated and tested. PROC SURVEYLOGISTIC displays the point estimate, its standard error, a *t* or Wald confidence interval, and a *t* or Wald chi-square test for each contrast. The significance level of the confidence interval is controlled by the ALPHA= option. You can estimate the contrast or the exponentiated contrast ($e^{\mathbf{l}_i\boldsymbol{\beta}}$), 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=*value*

tunes the estimability checking. If $\mathbf{v}$ is a vector, define ABS($\mathbf{v}$) to be the largest absolute value of the elements of $\mathbf{v}$. For a row vector $\mathbf{l}$ of the matrix $\mathbf{L}$, define

$$c = \begin{cases} \text{ABS}(\mathbf{l}) & \text{if ABS}(\mathbf{l}) > 0 \\ 1 & \text{otherwise} \end{cases}$$

If ABS($\mathbf{l} - \mathbf{lH}$) is greater than c**value*, then $\mathbf{l}\boldsymbol{\beta}$ is declared nonestimable. The $\mathbf{H}$ matrix is the Hermite form matrix $\mathbf{I}_0^{-}\mathbf{I}_0$, where $\mathbf{I}_0^{-}$ represents a generalized inverse of the information matrix $\mathbf{I}_0$ of the null model. The *value* must be between 0 and 1; the default is 10^{-4}.

DOMAIN Statement

DOMAIN *variables* < *variable*∗*variable variable*∗*variable*∗*variable . . .* > **;**

The DOMAIN statement requests analysis for domains (subpopulations) in addition to analysis for the entire study population. The DOMAIN statement names the *variables* that identify domains, which are called domain variables.

A domain variable can be either character or numeric. The procedure treats domain variables as categorical variables. If a *variable* appears by itself in a DOMAIN statement, each level of this variable determines a domain in the study population. If two or more *variables* are joined by asterisks (∗), then every possible combination of levels of these variables determines a domain. The procedure performs a descriptive analysis within each domain that is defined by the domain variables.

The formatted values of the domain variables determine the categorical variable levels. Thus, you can use formats to group values into levels. For more information, see the FORMAT procedure in *Base SAS Procedures Guide* and the FORMAT statement and SAS formats in *SAS Formats and Informats: Reference*.

When determining levels of a DOMAIN variable, an observation with missing values for this DOMAIN variable is excluded, unless you specify the MISSING option. For more information, see the section "Missing Values" on page 9724.

It is common practice to compute statistics for domains. Because formation of these domains might be unrelated to the sample design, the sample sizes for the domains are random variables. Use a DOMAIN statement to incorporate this variability into the variance estimation.

A DOMAIN statement is different from a BY statement. In a BY statement, you treat the sample sizes as fixed in each subpopulation, and you perform analysis within each BY group independently. For more information, see the section "Domain Analysis" on page 9747. Similarly, you should use a DOMAIN statement to perform a domain analysis over the entire data set. Creating a new data set from a single domain and analyzing that with PROC SURVEYLOGISTIC yields inappropriate estimates of variance.

By default, the SURVEYLOGISTIC procedure displays analyses for all levels of domains that are formed by the variables in a DOMAIN statement. Optionally, you can specify particular levels of each DOMAIN variable to be displayed by listing quoted *formatted-level-values* in parentheses after each variable name. You must enclose each *formatted-level-value* in single or double quotation marks. You can specify one or more levels of each variable; when you specify more than one level, separate the levels by a space or a comma. These examples illustrate the syntax:

```
domain Race*Gender(''Female'');
domain Race('White','Asian') Gender;
```

For example, **Race*Gender(''Female'')** requests that the procedure display analysis only for females within each race category, and **Race('White','Asian')** requests that the procedure display domain analysis only for people whose race is either white or Asian.

Specifying the same domain multiple times but with different levels for each corresponding domain variables is equivalent to specifying the union of different levels for the same variables. However, if you do not specify levels for a variable in a domain that is specified multiple times, only the specified levels are rendered. For example, the following two specifications together

```
domain Race('White')*Gender('Female');
domain Race('Asian')*Gender;
```

have the same effect as a single specification:

```
domain Race('White' 'Asian')*Gender('Female');
```

Also, the following specification

```
domain Race('White')*Gender Race('Asian')*Gender;
```

is equivalent to

```
domain Race('White' 'Asian')*Gender;
```

This syntax controls only the display of domain analysis results; it does not subset the data set, change the degrees of freedom, or otherwise affect the variance estimation.

EFFECT Statement

EFFECT *name*=*effect-type* **(***variables* < / *options* >**)** ;

The EFFECT statement enables you to construct special collections of columns for design matrices. These collections are referred to as *constructed effects* to distinguish them from the usual model effects that are formed from continuous or classification variables, as discussed in the section "GLM Parameterization of Classification Variables and Effects" on page 393 in Chapter 19, "Shared Concepts and Topics."

You can specify the following *effect-types*:

COLLECTION
specifies a collection effect that defines one or more variables as a single effect with multiple degrees of freedom. The variables in a collection are considered as a unit for estimation and inference.

LAG
specifies a classification effect in which the level that is used for a particular period corresponds to the level in the preceding period.

MULTIMEMBER | MM
specifies a multimember classification effect whose levels are determined by one or more variables that appear in a CLASS statement.

POLYNOMIAL | POLY
specifies a multivariate polynomial effect in the specified numeric variables.

SPLINE specifies a regression spline effect whose columns are univariate spline expansions of one or more variables. A spline expansion replaces the original variable with an expanded or larger set of new variables.

Table 117.3 summarizes the *options* available in the EFFECT statement.

Table 117.3 EFFECT Statement Options

Option	Description
Collection Effects Options	
DETAILS	Displays the constituents of the collection effect
Lag Effects Options	
DESIGNROLE=	Names a variable that controls to which lag design an observation is assigned
DETAILS	Displays the lag design of the lag effect
NLAG=	Specifies the number of periods in the lag
PERIOD=	Names the variable that defines the period. This option is required.
WITHIN=	Names the variable or variables that define the group within which each period is defined. This option is required.
Multimember Effects Options	
NOEFFECT	Specifies that observations with all missing levels for the multimember variables should have zero values in the corresponding design matrix columns
WEIGHT=	Specifies the weight variable for the contributions of each of the classification effects
Polynomial Effects Options	
DEGREE=	Specifies the degree of the polynomial
MDEGREE=	Specifies the maximum degree of any variable in a term of the polynomial
STANDARDIZE=	Specifies centering and scaling suboptions for the variables that define the polynomial
Spline Effects Options	
BASIS=	Specifies the type of basis (B-spline basis or truncated power function basis) for the spline effect
DEGREE=	Specifies the degree of the spline effect
KNOTMETHOD=	Specifies how to construct the knots for the spline effect

For more information about the syntax of these *effect-types* and how columns of constructed effects are computed, see the section “EFFECT Statement” on page 403 in Chapter 19, “Shared Concepts and Topics.”

ESTIMATE Statement

ESTIMATE < *'label'* > *estimate-specification* < **(***divisor***=***n***)** >
< **,** . . . < *'label'* > *estimate-specification* < **(***divisor***=***n***)** > >
< / *options* > **;**

The ESTIMATE statement provides a mechanism for obtaining custom hypothesis tests. Estimates are formed as linear estimable functions of the form $\mathbf{L}\boldsymbol{\beta}$. You can perform hypothesis tests for the estimable functions, construct confidence limits, and obtain specific nonlinear transformations.

Table 117.4 summarizes the *options* available in the ESTIMATE statement.

Table 117.4 ESTIMATE Statement Options

Option	Description
Construction and Computation of Estimable Functions	
DIVISOR=	Specifies a list of values to divide the coefficients
NOFILL	Suppresses the automatic fill-in of coefficients for higher-order effects
SINGULAR=	Tunes the estimability checking difference
Degrees of Freedom and *p*-Values	
ADJUST=	Determines the method of multiple comparison adjustment of estimates
ALPHA=α	Determines the confidence level $(1-\alpha)$
LOWER	Performs one-sided, lower-tailed inference
STEPDOWN	Adjusts multiplicity-corrected *p*-values further in a step-down fashion
TESTVALUE=	Specifies values under the null hypothesis for tests
UPPER	Performs one-sided, upper-tailed inference
Statistical Output	
CL	Constructs confidence limits
CORR	Displays the correlation matrix of estimates
COV	Displays the covariance matrix of estimates
E	Prints the **L** matrix
JOINT	Produces a joint *F* or chi-square test for the estimable functions
PLOTS=	Produces ODS statistical graphics if the analysis is sampling-based
SEED=	Specifies the seed for computations that depend on random numbers
Generalized Linear Modeling	
CATEGORY=	Specifies how to construct estimable functions for multinomial data
EXP	Exponentiates and displays estimates
ILINK	Computes and displays estimates and standard errors on the inverse linked scale

For more information about the syntax of the ESTIMATE statement, see the section "ESTIMATE Statement" on page 451 in Chapter 19, "Shared Concepts and Topics."

FREQ Statement

FREQ *variable* **;**

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

If you use the events/trials syntax in the MODEL statement, the FREQ statement is not allowed because the event and trial variables represent the frequencies in the data set.

If you use the FREQ statement and you specify the VARMETHOD=BRR, VARMETHOD=JACKKNIFE, or VARMETHOD=BOOTSTRAP option in the PROC SURVEYLOGISTIC statement to estimate the variance, then you must use the CLUSTER statement to identify the primary sampling units unless you also provide replicate weights with a REPWEIGHTS statement.

LSMEANS Statement

LSMEANS < *model-effects* > < / *options* > **;**

The LSMEANS statement computes and compares least squares means (LS-means) of fixed effects. LS-means are *predicted margins*—that is, they estimate the marginal means over a hypothetical balanced population on the linked scale. For example, in a binomial model with logit link, the least squares means are predicted population margins of the logits.

Table 117.5 the summarizes available *options* in the LSMEANS statement.

Table 117.5 LSMEANS Statement Options

Option	Description
Construction and Computation of LS-Means	
AT	Modifies the covariate value in computing LS-means
BYLEVEL	Computes separate margins
DIFF	Computes differences of LS-means
OM=	Specifies the weighting scheme for LS-means computation as determined by the input data set
SINGULAR=	Tunes estimability checking
Degrees of Freedom and p-Values	
ADJUST=	Determines the method of multiple-comparison adjustment of LS-means differences
ALPHA=α	Determines the confidence level $(1 - \alpha)$

Table 117.5 *continued*

Option	Description
STEPDOWN	Adjusts multiple-comparison p-values further in a step-down fashion
Statistical Output	
CL	Constructs confidence limits for means and mean differences
CORR	Displays the correlation matrix of LS-means
COV	Displays the covariance matrix of LS-means
E	Prints the **L** matrix
LINES	Uses connecting lines to indicate nonsignificantly different subsets of LS-means
LINESTABLE	Displays the results of the LINES option as a table
MEANS	Prints the LS-means
PLOTS=	Produces graphs of means and mean comparisons
SEED=	Specifies the seed for computations that depend on random numbers
Generalized Linear Modeling	
EXP	Exponentiates and displays estimates of LS-means or LS-means differences
ILINK	Computes and displays estimates and standard errors of LS-means (but not differences) on the inverse linked scale
ODDSRATIO	Reports (simple) differences of least squares means in terms of odds ratios if permitted by the link function

For details about the syntax of the LSMEANS statement, see the section "LSMEANS Statement" on page 467 in Chapter 19, "Shared Concepts and Topics."

LSMESTIMATE Statement

LSMESTIMATE *model-effect* < '*label*' > *values* < *divisor*=*n* >
< **,** . . . < '*label*' > *values* < *divisor*=*n* > >
< / *options* > **;**

The LSMESTIMATE statement provides a mechanism for obtaining custom hypothesis tests among least squares means.

Table 117.6 summarizes the *options* available in the LSMESTIMATE statement.

Table 117.6 LSMESTIMATE Statement Options

Option	Description
Construction and Computation of LS-Means	
AT	Modifies covariate values in computing LS-means

Table 117.6 *continued*

Option	Description
BYLEVEL	Computes separate margins
DIVISOR=	Specifies a list of values to divide the coefficients
OM=	Specifies the weighting scheme for LS-means computation as determined by a data set
SINGULAR=	Tunes estimability checking
Degrees of Freedom and *p*-Values	
ADJUST=	Determines the method of multiple-comparison adjustment of LS-means differences
ALPHA=α	Determines the confidence level $(1 - \alpha)$
LOWER	Performs one-sided, lower-tailed inference
STEPDOWN	Adjusts multiple-comparison *p*-values further in a step-down fashion
TESTVALUE=	Specifies values under the null hypothesis for tests
UPPER	Performs one-sided, upper-tailed inference
Statistical Output	
CL	Constructs confidence limits for means and mean differences
CORR	Displays the correlation matrix of LS-means
COV	Displays the covariance matrix of LS-means
E	Prints the $\mathbf{L}$ matrix
ELSM	Prints the $\mathbf{K}$ matrix
JOINT	Produces a joint F or chi-square test for the LS-means and LS-means differences
PLOTS=	Produces graphs of means and mean comparisons
SEED=	Specifies the seed for computations that depend on random numbers
Generalized Linear Modeling	
CATEGORY=	Specifies how to construct estimable functions for multinomial data
EXP	Exponentiates and displays LS-means estimates
ILINK	Computes and displays estimates and standard errors of LS-means (but not differences) on the inverse linked scale

For more information about the syntax of the LSMESTIMATE statement, see the section “LSMESTIMATE Statement” on page 487 in Chapter 19, “Shared Concepts and Topics.”

MODEL Statement

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

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

The MODEL statement names the response variable and the explanatory effects, including covariates, main effects, interactions, and nested effects; see the section "Specification of Effects" on page 4020 in Chapter 50, "The GLM Procedure," for more information. If you omit the explanatory variables, the procedure fits an intercept-only model. Model options can be specified after a slash (/).

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

In the events/trials syntax, you specify two variables that contain count data for a binomial experiment. These two variables are separated by a slash. The value of the first variable, *events*, is the number of positive responses (or events), and it must be nonnegative. The value of the second variable, *trials*, is the number of trials, and it must not be less than the value of *events*.

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

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

Response Variable Options

You specify the following *options* by enclosing them in parentheses after the response variable:

DESCENDING

DESC

reverses the order of response categories. If both the DESCENDING and the ORDER= options are specified, PROC SURVEYLOGISTIC orders the response categories according to the ORDER= option and then reverses that order. See the section "Response Level Ordering" on page 9724 for more detail.

EVENT='*category***' | ***keyword*

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

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

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

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

ORDER=DATA | FORMATTED | FREQ | INTERNAL

specifies the sort order for the levels of the response variable.

The ORDER= option can take the following values:

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

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

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

REFERENCE='*category*' | *keyword*

REF='*category*' | *keyword*

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

FIRST designates the first-ordered category as the reference

LAST designates the last-ordered category as the reference

Model Options

Model *options* can be specified after a slash (/). Table 117.7 summarizes the *options* available in the MODEL statement.

Table 117.7 MODEL Statement Options

Option	Description
Model Specification Options	
LINK=	Specifies link function
NOINT	Suppresses intercept(s)
OFFSET=	Specifies offset variable

Table 117.7 (continued)

Option	Description
Convergence Criterion Options	
ABSFCONV=	Specifies absolute function convergence criterion
FCONV=	Specifies relative function convergence criterion
GCONV=	Specifies relative gradient convergence criterion
XCONV=	Specifies relative parameter convergence criterion
MAXITER=	Specifies maximum number of iterations
NOCHECK	Suppresses checking for infinite parameters
RIDGING=	Specifies 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 Adjustment to Variance Estimation	
VADJUST=	Chooses variance estimation adjustment method
Options for Confidence Intervals	
DF=	Specifies the degrees of freedom
ALPHA=	Specifies α for the $100(1-\alpha)\%$ confidence intervals
CHISQ	Specifies the type of likelihood ratio chi-square test
CLPARM	Computes confidence intervals for parameters
CLODDS	Computes confidence intervals for odds ratios
Options for Display of Details	
CORRB	Displays correlation matrix
COVB	Displays covariance matrix
EXPB	Displays exponentiated values of estimates
GRADIENT	Displays gradients evaluated at null hypothesis
ITPRINT	Displays iteration history
NODUMMYPRINT	Suppresses "Class Level Information" table
PARMLABEL	Displays parameter labels
RSQUARE	Displays generalized R^2
STB	Displays 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 9732.

ALPHA=*value*

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

CHISQ (FIRSTORDER | NOADJUST | SECONDORDER)

specifies the type of likelihood ratio chi-square test. If you specify CHISQ(FIRSTORDER) or CHISQ(SECONDORDER), PROC SURVEYLOGISTIC provides a first-order or second-order (Satterthwaite) Rao-Scott likelihood ratio chi-square test, which is a design-adjusted test. If you specify CHISQ(NOADJUST), the procedure computes a chi-square test without the Rao-Scott design correction.

If you do not specify the CHISQ option, the default test that PROC SURVEYLOGISTIC uses depends on the design and model as follows:

- If you do not use a STRATA, CLUSTER, or REPWEIGHTS statement, then the default is CHISQ(NOADJUST).
- If you use a STRATA, CLUSTER, or REPWEIGHTS statement, and you need to estimate only one parameter excluding the intercepts in the model, then the default is CHISQ(FIRSTORDER).
- If you use a STRATA, CLUSTER, or REPWEIGHTS statement, and you need to estimate more than one parameter excluding the intercepts in the model, then the default is CHISQ(SECONDORDER).

For more information, see the section “Rao-Scott Likelihood Ratio Chi-Square Test” on page 9750.

Note that unless you specify the DF=INFINITY option, PROC SURVEYLOGISTIC displays an F test instead of a chi-square test.

CLODDS

requests confidence intervals for the odds ratios. Computation of these confidence intervals is based on individual t tests or Wald tests. The degrees of freedom for a t test is described in the section “Degrees of Freedom” on page 9748. The confidence coefficient can be specified with the ALPHA= option. See the section “Wald Confidence Intervals for Parameters” on page 9752 for more information.

CLPARM

requests confidence intervals for the parameters. Computation of these confidence intervals is based on the t tests or Wald tests. The degrees of freedom for a t test is described in the section “Degrees of Freedom” on page 9748. You can specify the confidence level by using the ALPHA= option.

CORRB

displays the correlation matrix of the parameter estimates.

COVB

displays the covariance matrix of the parameter estimates.

DF=*types* < (*value*) >

determines the denominator degrees of freedom (df) for F statistics in hypothesis testing, as well as the degrees of freedom in t tests for parameter estimates and odds ratio estimates, and for computing t distribution percentiles for confidence limits of these estimates.

You can specify *type* to be DESIGN, INFINITY, or PARMADJ. When you specify DF=DESIGN or DF=PARMADJ, you can optionally specify a positive *value* in parentheses to overwrite the default design degrees of freedom.

DF=PARMADJ is the default for the Taylor variance estimation method, and DF=DESIGN is the default for the replication variance estimation method.

For more information, see the section "Degrees of Freedom" on page 9748.

If you specify both DF=DESIGN(*value*) in the MODEL statement and the DF= option in a REPWEIGHTS statement, PROC SURVEYLOGISTIC uses the *value* in DF=DESIGN(*value*) in the MODEL statement to determine the *df* and ignores the one in the REPWEIGHTS statement.

You can specify one of the following *types*:

DESIGN

DESIGN < (*value*) >

specifies the *df* to be the design degrees of freedom. If you specify a positive *value* in DF=DESIGN(*value*), then *df*=*value*.

If you specify DF=DESIGN without the optional positive *value*, then *df* is determined as the design degrees of freedom.

For more information, see the section "Degrees of Freedom" on page 9748.

INFINITY

NONE

specifies that the *df* is infinite. As the denominator degrees of freedom grows, an *F* distribution approaches a chi-square distribution, and similarly a *t* distribution approaches a normal distribution. Therefore, when you specify DF=INFINITY, PROC SURVEYLOGISTIC uses chi-square tests and normal distribution percentiles to construct confidence intervals.

PARMADJ

PARMADJ < (*value*) >

requests that the *df* be modified as *f–r+1*, where *f* is the default design degrees of freedom or the *value* specified in this option, and *r* is the rank of the contrast of model parameters to be tested.

This option applies only when the Taylor variance estimation method is used (either by default or when you specify VARMETHOD=TAYLOR). This option can be useful when you have many parameters relative to the default design degrees of freedom.

EXPB

EXPEST

displays the exponentiated values ($e^{\hat{\theta}_i}$) of the parameter estimates $\hat{\theta}_i$ 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.

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 9732 for details.

GCONV=*value*

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

$$\frac{\mathbf{g}^{(i)\prime}\mathbf{I}^{(i)}\mathbf{g}^{(i)}}{|l^{(i)}| + 1\text{E–}6} < value$$

where $l^{(i)}$ is the value of the log-likelihood function, $\mathbf{g}^{(i)}$ is the gradient vector, and $\mathbf{I}^{(i)}$ the (expected) information matrix. All of these functions are evaluated at iteration i. This is the default convergence criterion, and the default value is 1E–8. For more information, see the section “Convergence Criteria” on page 9732.

GRADIENT

displays the gradient vector, which is evaluated at the global null hypothesis.

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 $-2 \log L$.

LINK=*keyword*

L=*keyword*

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

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

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

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

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

See the section “Link Functions and the Corresponding Distributions” on page 9728 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 created by the procedure contains results that are based on the last maximum likelihood iteration.

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 when the estimation takes more than eight iterations. For more information, see the section "Existence of Maximum Likelihood Estimates" on page 9732.

NODUMMYPRINT

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.

OFFSET=*name*

names the offset variable. The regression coefficient for this variable is fixed at 1.

PARMLABEL

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

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.

RSQUARE

requests a generalized R^2 measure for the fitted model.

For more information, see the section "Generalized Coefficient of Determination" on page 9734.

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 *value* times a norm of the matrix. Values of the SINGULAR= option must be numeric. By default, SINGULAR=10^{-12}.

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 θ_i is given by $\hat{\theta}_i/(s/s_i)$, 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.

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 where the LOGIT link is specified for binary response data. The default is TECHNIQUE=FISHER. If the LINK=GLOGIT option is specified, then Newton-Raphson is the default and only available method. See the section "Iterative Algorithms for Model Fitting" on page 9730 for details.

VADJUST=DF | MOREL < (*Morel-options*) > | NONE

specifies an adjustment to the variance estimation for the regression coefficients.

By default, PROC SURVEYLOGISTIC uses the degrees of freedom adjustment VADJUST=DF.

If you do not want to use any variance adjustment, you can specify the VADJUST=NONE option. You can specify the VADJUST=MOREL option for the variance adjustment proposed by Morel (1989).

You can specify the following *Morel-options* within parentheses after the VADJUST=MOREL option:

ADJBOUND=ϕ

sets the upper bound coefficient ϕ in the variance adjustment. This upper bound must be positive. By default, the procedure uses $\phi = 0.5$. See the section "Adjustments to the Variance Estimation" on page 9741 for more details on how this upper bound is used in the variance estimation.

DEFFBOUND=δ

sets the lower bound of the estimated design effect in the variance adjustment. This lower bound must be positive. By default, the procedure uses $\delta = 1$. See the section "Adjustments to the Variance Estimation" on page 9741 for more details about how this lower bound is used in the variance estimation.

XCONV=*value*

specifies the relative parameter convergence criterion. Convergence requires a small relative parameter change in subsequent iterations:

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

where

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

and $\theta_j^{(i)}$ is the estimate of the jth parameter at iteration i. See the section "Convergence Criteria" on page 9732 for details.

OUTPUT Statement

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

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. Formulas for the statistics are given in the section "Linear Predictor, Predicted Probability, and Confidence Limits" on page 9757.

If you use the single-trial syntax, the data set also contains a variable named _LEVEL_, which indicates the level of the response that the given row of output is referring to. For example, 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= Data Set in the OUTPUT Statement" on page 9759.

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.

Table 117.8 summarizes the *options* available in the OUTPUT statement.

Table 117.8 OUTPUT Statement Options

Option	Description
ALPHA=	Sets the level of significance
LOWER=	Names the variable that contains the lower confidence limits
OUT=	Names the output data set
PREDICTED=	Names the variable that contains the predicted probabilities
PREDPROBS=	Requests predicted probabilities
STDXBETA=	Names the variable that contains the standard error estimates
UPPER=	Names the variable that contains the upper confidence limits
XBETA=	Names the variable that contains the estimates of the linear predictor

You can specify the following *options* in the OUTPUT statement:

LOWER | L=*name*

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

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 by using the DATA*n* convention.

The statistic options in the OUTPUT statement specify the statistics to be included in the output data set and name the new variables that contain the statistics.

PREDICTED | P=*name*

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

PREDPROBS=(*keywords***)**

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

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

CUMULATIVE | C requests the cumulative predicted probability of each response level. For a response variable Y with three levels, 1, 2, and 3, the cumulative probabilities are Pr(Y≤1), Pr(Y≤2), and Pr(Y≤3). The cumulative probability for the last response level always has the constant value of 1. For generalized logit models, the cumulative predicted probabilities are not computed and are set to missing.

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

See the section "Details of the PREDPROBS= Option" on page 9718 at the end of this section for further details.

STDXBETA=*name*

names the variable that contains the standard error estimates of XBETA (the definition of which follows).

UPPER | U=*name*

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

XBETA=*name*

names the variable that contains the estimates of the linear predictor $\alpha_i + \mathbf{x}\boldsymbol{\beta}$, where i is the corresponding ordered value of _LEVEL_.

You can specify the following *option* in the OUTPUT statement after a slash (/):

ALPHA=*value*

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

Details of the PREDPROBS= Option

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

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

If you specify PREDPROBS=INDIVIDUAL, the OUTPUT data set contains *k* additional variables representing the individual probabilities, one for each response level, where *k* is the maximum number of response levels across all BY groups. The names of these variables have the form IP_*xxx*, where *xxx* represents the particular level. The representation depends on the following situations:

- If you specify the events/trials syntax, *xxx* is either Event or Nonevent. Thus, the variable that contains 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 that contains 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 can 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'), Pr(Y='Mild'), and Pr(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 that represent 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 they are arranged in variables in each output observation rather than in multiple output observations.

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

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

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

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

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

REPWEIGHTS Statement

REPWEIGHTS *variables* < / *options* > **;**

The REPWEIGHTS statement names *variables* that provide replicate weights for bootstrap, BRR, or jackknife variance estimation, which you request with the VARMETHOD=BOOTSTRAP, VARMETHOD=BRR, or VARMETHOD=JACKKNIFE option, respectively, in the PROC SURVEYLOGISTIC statement. For more information about the replication methods, see the section “Variance Estimation” on page 9739.

Each REPWEIGHTS *variable* contains the weights for a single replicate, and the number of replicates equals the number of REPWEIGHTS variables. The REPWEIGHTS variables must be numeric, and the variable values must be nonnegative numbers.

For more information about replicate weights that the SURVEYLOGISTIC procedure creates, see the sections “Balanced Repeated Replication (BRR) Method” on page 9743 and “Jackknife Method” on page 9746.

If you provide replicate weights with a REPWEIGHTS statement, you do not need to specify a CLUSTER or STRATA statement. If you use a REPWEIGHTS statement and do not specify the VARMETHOD= option in the PROC SURVEYLOGISTIC statement, the procedure uses VARMETHOD=JACKKNIFE by default.

If you specify a REPWEIGHTS statement but do not include a WEIGHT statement, the procedure uses the average of replicate weights of each observation as the observation’s weight.

You can specify the following *options* in the REPWEIGHTS statement after a slash (/):

DF=*df*

specifies the degrees of freedom for the analysis. The value of *df* must be a positive number. By default, the value of *df* is the number of REPWEIGHTS variables.

If you know the number of PSUs and the number of strata from which the replicate weights are generated, you should specify the number of PSUs minus the number of strata as the degrees of freedom in this option.

JKCOEFS=*value* | **<(>***values* **<)>** | *SAS-data-set*

specifies jackknife coefficients for the VARMETHOD=JACKKNIFE option in the PROC SURVEYLOGISTIC statement. The jackknife coefficient values must be nonnegative numbers. For more information about jackknife coefficients, see the section “Jackknife Method” on page 9746.

You can provide jackknife coefficients by specifying one of the following forms:

value

specifies a single jackknife coefficient *value* to use for all replicates, where *value* must be a nonnegative number.

values

specifies a list of jackknife coefficients, where each value in the *values* is a nonnegative number that corresponds to a single replicate that is identified by a REPWEIGHTS variable. You can separate the values with blanks or commas, and you can optionally enclose the *values* in parentheses. The number of values in the *values* must equal the number of replicate weight variables that you specify in the REPWEIGHTS statement.

You must list the jackknife coefficient values in the same order in which you list the corresponding replicate weight variables in the REPWEIGHTS statement.

SAS-data-set

names a *SAS-data-set* that contains the jackknife coefficients, where each coefficient value must be a nonnegative number. You must provide the jackknife coefficients in the data set variable named JKCoefficient. Each observation in this data set must correspond to a replicate that is identified by a REPWEIGHTS variable. The number of observations in the *SAS-data-set* must not be less than the number of REPWEIGHTS variables.

REPCOEFS=*value* **|** **<(>***values* **<)>** **|** *SAS-data-set*

specifies replicate coefficients for the VARMETHOD=BOOTSTRAP or VARMETHOD=JACKKNIFE option in the PROC SURVEYLOGISTIC statement, where each coefficient corresponds to an individual replicate weight that is identified by a REPWEIGHTS variable. The replicate coefficient values must be nonnegative numbers.

You can provide replicate coefficients by specifying one of the following forms:

value

specifies a single replicate coefficient *value* to use for all replicates, where *value* must be a nonnegative number.

values

specifies a list of replicate coefficients, where each value in the *values* is a nonnegative number that corresponds to a single replicate that is identified by a REPWEIGHTS variable. You can separate the values with blanks or commas, and you can optionally enclose the *values* in parentheses. The number of values in the *values* must equal the number of replicate weight variables that you specify in the REPWEIGHTS statement.

You must list the replicate coefficient values in the same order in which you list the corresponding replicate weight variables in the REPWEIGHTS statement.

SAS-data-set

names a *SAS-data-set* that contains the replicate coefficients, where each coefficient value must be a nonnegative number. You must provide the replicate coefficients in the data set variable named JKCoefficient. Each observation in this data set must correspond to a replicate that is identified by a REPWEIGHTS variable. The number of observations in the *SAS-data-set* must not be less than the number of REPWEIGHTS variables.

SLICE Statement

SLICE *model-effect* < / *options* > **;**

The SLICE statement provides a general mechanism for performing a partitioned analysis of the LS-means for an interaction. This analysis is also known as an analysis of simple effects.

This statement uses the same *options* as the LSMEANS statement, which are summarized in Table 19.23 in Chapter 19, "Shared Concepts and Topics." For more information about the syntax of the SLICE statement, see the section "SLICE Statement" on page 516 in Chapter 19, "Shared Concepts and Topics."

STORE Statement

STORE < **OUT=** >*item-store-name* < / **LABEL=**'*label*' > **;**

The STORE statement saves the context and results of the statistical analysis. The resulting item store has a binary file format that cannot be modified. The contents of the item store can be processed using the PLM procedure. For more information about the syntax of the STORE statement, see the section "STORE Statement" on page 520 in Chapter 19, "Shared Concepts and Topics."

STRATA Statement

STRATA *variables* < / *option* > **;**

The STRATA statement specifies variables that form the strata in a stratified sample design. The combinations of categories of STRATA variables define the strata in the sample.

If your sample design has stratification at multiple stages, you should identify only the first-stage strata in the STRATA statement. See the section "Specification of Population Totals and Sampling Rates" on page 9735 for more information.

If you provide replicate weights for BRR or jackknife variance estimation with the REPWEIGHTS statement, you do not need to specify a STRATA statement.

The STRATA *variables* are one or more variables in the DATA= input data set. These variables can be either character or numeric. The formatted values of the STRATA variables determine the levels. Thus, you can use formats to group values into levels. See the FORMAT procedure in the *Base SAS Procedures Guide* and the FORMAT statement and SAS formats in *SAS Formats and Informats: Reference* for more information.

When determining levels of a STRATA variable, an observation with missing values for this STRATA variable is excluded, unless you specify the MISSING option. For more information, see the section "Missing Values" on page 9724.

You can use multiple STRATA statements to specify stratum variables.

You can specify the following *option* in the STRATA statement after a slash (/):

LIST

displays a “Stratum Information” table, which includes values of the STRATA variables and the number of observations, number of clusters, population total, and sampling rate for each stratum. See the section “Stratum Information” on page 9763 for more details.

TEST Statement

< *label***: > TEST** *equation1* **< ,** *equation2***,** *. . .* **> < /** *option* **> ;**

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

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

The form of an *equation* is as follows:

term < $\pm$ *term* . . . > < = $\pm$ *term* < $\pm$ *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 illustrates possible uses of the TEST statement:

```
proc surveylogistic;
   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 $H_0\colon \mathbf{L}\boldsymbol{\theta} = \mathbf{c}$. This includes $\mathbf{L}\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})\mathbf{L}'$ bordered by $(\mathbf{L}\hat{\boldsymbol{\theta}} - \mathbf{c})$ and $[\mathbf{L}\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})\mathbf{L}']^{-1}$ bordered by $[\mathbf{L}\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})\mathbf{L}']^{-1}(\mathbf{L}\hat{\boldsymbol{\theta}} - \mathbf{c})$, where $\hat{\boldsymbol{\theta}}$ is the pseudo-estimator of $\boldsymbol{\theta}$ and $\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})$ is the estimated covariance matrix of $\hat{\boldsymbol{\theta}}$.

For more information, see the section “Testing Linear Hypotheses about the Regression Coefficients” on page 9753.

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 intervals 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 SURVEYLOGISTIC does not produce customized odds ratio estimates for any explanatory variable that is not listed in the UNITS statement.

For more information, see the section "Odds Ratio Estimation" on page 9754.

WEIGHT Statement

WEIGHT *variable* **;**

The WEIGHT statement names the variable that contains the sampling weights. This variable must be numeric, and the sampling weights must be positive numbers. If an observation has a weight that is nonpositive or missing, then the procedure omits that observation from the analysis. See the section "Missing Values" on page 9724 for more information. If you specify more than one WEIGHT statement, the procedure uses only the first WEIGHT statement and ignores the rest.

If you do not specify a WEIGHT statement but provide replicate weights with a REPWEIGHTS statement, PROC SURVEYLOGISTIC uses the average of replicate weights of each observation as the observation's weight.

If you do not specify a WEIGHT statement or a REPWEIGHTS statement, PROC SURVEYLOGISTIC assigns all observations a weight of one.

Details: SURVEYLOGISTIC Procedure

Missing Values

If you have missing values in your survey data for any reason, such as nonresponse, this can compromise the quality of your survey results. If the respondents are different from the nonrespondents with regard to a survey effect or outcome, then survey estimates might be biased and cannot accurately represent the survey population. There are a variety of techniques in sample design and survey operations that can reduce nonresponse. After data collection is complete, you can use imputation to replace missing values with acceptable values, and/or you can use sampling weight adjustments to compensate for nonresponse. You should complete this data preparation and adjustment before you analyze your data with PROC SURVEYLOGISTIC. For more information, see Cochran (1977); Kalton and Kasprzyk (1986); Brick and Kalton (1996).

If an observation has a missing value or a nonpositive value for the WEIGHT or FREQ variable, then that observation is excluded from the analysis.

An observation is also excluded if it has a missing value for any design (STRATA, CLUSTER, or DOMAIN) variable, unless you specify the MISSING option in the PROC SURVEYLOGISTIC statement. If you specify the MISSING option, the procedure treats missing values as a valid (nonmissing) category for all categorical variables.

By default, if an observation contains missing values for the response, offset, or any explanatory variables used in the independent effects, the observation is excluded from the analysis. This treatment is based on the assumption that the missing values are missing completely at random (MCAR). However, this assumption is not true sometimes. For example, evidence from other surveys might suggest that observations with missing values are systematically different from observations without missing values. If you believe that missing values are not missing completely at random, then you can specify the NOMCAR option to include these observations with missing values in the dependent variable and the independent variables in the variance estimation.

Whether or not the NOMCAR option is used, observations with missing or invalid values for WEIGHT, FREQ, STRATA, CLUSTER, or DOMAIN variables are always excluded, unless the MISSING option is also specified.

When you specify the NOMCAR option, the procedure treats observations with and without missing values for variables in the regression model as two different domains, and it performs a domain analysis in the domain of nonmissing observations.

If you use a REPWEIGHTS statement, all REPWEIGHTS variables must contain nonmissing values.

Model Specification

Response Level Ordering

Response level ordering is important because, by default, PROC SURVEYLOGISTIC models the probabilities of response levels with lower *Ordered Values.* Ordered Values, displayed in the “Response Profile” table, are assigned to response levels in ascending sorted order. That is, the lowest response level is assigned Ordered

Value 1, the next lowest is assigned Ordered Value 2, and so on. For example, if your response variable Y takes values in $\{1, \ldots, D+1\}$, then the functions of the response probabilities modeled with the cumulative model are

$$\text{logit}(\Pr(Y \le i|\mathbf{x})), i = 1, \ldots, D$$

and for the generalized logit model they are

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

where the highest Ordered Value $Y = D + 1$ is the reference level. You can change these default functions by specifying the EVENT=, REF=, DESCENDING, or ORDER= response variable options in the MODEL statement.

For binary response data with event and nonevent categories, the procedure models the function

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

where p is the probability of the response level assigned to Ordered Value 1 in the "Response Profiles" table. Since

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

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

If your event category has a higher Ordered Value than the nonevent category, the procedure models the nonevent probability. You can use response variable options to model the event probability. For example, suppose the binary response variable Y takes the values 1 and 0 for event and nonevent, respectively, and Exposure is the explanatory variable. By default, the procedure assigns Ordered Value 1 to response level Y=0, and Ordered Value 2 to response level Y=1. Therefore, the procedure models the probability of the nonevent (Ordered Value=1) category. To model the event probability, you can do the following:

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

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

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

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

- Specify the response variable option REF= in the MODEL statement as the nonevent category for the response variable. This option is most useful when you are fitting a generalized logit model.

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

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

```
proc format;
   value Disease 1='event' 0='nonevent';
run;

proc surveylogistic;
   format Y Disease.;
   model Y=Exposure;
run;
```

CLASS Variable Parameterization

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

EFFECT Three columns are created to indicate group membership of the nonreference levels. For the reference level, all three 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.

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

For CLASS main effects that use the EFFECT coding scheme, individual parameters correspond to the difference between the effect of each nonreference level and the average over all four levels.

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

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

For CLASS main effects that use the GLM coding scheme, individual parameters correspond to the difference between the effect of each level and the last level.

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

	Design Matrix		
A	**A2**	**A5**	**A7**
1	0	0	0
2	1	0	0
5	1	1	0
7	1	1	1

The first level of the effect is a control or baseline level.

For CLASS main effects that use the ORDINAL coding scheme, the first level of the effect is a control or baseline level; individual parameters correspond to the difference between effects of the current level and the preceding level. When the parameters for an ordinal main effect have the same sign, the response effect is monotonic across the levels.

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

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

REFERENCE | REF Three columns are created to indicate group membership of the nonreference levels. For the reference level, all three 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.

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

For CLASS main effects that use the REFERENCE coding scheme, individual parameters correspond to the difference between the effect of each nonreference level and the reference level.

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

	Design Matrix		
A	**AOEFF1**	**AOEFF2**	**AOEFF3**
1	1.41421	–0.81650	–0.57735
2	0.00000	1.63299	–0.57735
5	0.00000	0.00000	1.73205
7	–1.41421	–0.81649	–0.57735

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

	Design Matrix		
A	**AOORD1**	**AOORD2**	**AOORD3**
1	–1.73205	0.00000	0.00000
2	0.57735	–1.63299	0.00000
5	0.57735	0.81650	–1.41421
7	0.57735	0.81650	1.41421

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

	Design Matrix		
A	**AOPOLY1**	**AOPOLY2**	**AOPOLY5**
1	–1.153	0.907	–0.921
2	–0.734	–0.540	1.473
5	0.524	–1.370	–0.921
7	1.363	1.004	0.368

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

	Design Matrix		
A	**AOREF1**	**AOREF2**	**AOREF3**
1	1.73205	0.00000	0.00000
2	–0.57735	1.63299	0.00000
5	–0.57735	–0.81650	1.41421
7	–0.57735	–0.81650	–1.41421

Link Functions and the Corresponding Distributions

Four link functions are available in the SURVEYLOGISTIC 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

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

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

$$F(x) = \frac{1}{1+e^{-x}}$$

- The **probit** (or normit) function

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

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

  $$F(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}z^2} dz$$

  Traditionally, the probit function includes an additive constant 5, but throughout PROC SURVEYLOGISTIC, the terms probit and normit are used interchangeably, previously defined as $g(p)$.

- 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

  $$F(x) = 1 - e^{-e^x}$$

- The **generalized logit** function extends the binary logit link to a vector of levels $(\pi_1, \ldots, \pi_{k+1})$ by contrasting each level with a fixed level

  $$g(\pi_i) = \log\left(\frac{\pi_i}{\pi_{k+1}}\right) \quad i = 1, \ldots, k$$

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

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

where γ is the Euler constant. In comparing parameter estimates that use 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 from using the logit link function should be about $\pi/\sqrt{3} \approx 1.8$ larger than the estimates from the probit link function.

Model Fitting

Determining Observations for Likelihood Contributions

If you use the events/trials syntax, each observation is split into two observations. One has the response value 1 with a frequency equal to the value of the *events* variable. The other observation has the response value 2 and a frequency equal to the value of (*trials* – *events*). These two observations have the same explanatory variable values and the same WEIGHT values as the original observation.

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

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

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

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

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

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

Iterative Algorithms for Model Fitting

Two iterative maximum likelihood algorithms are available in PROC SURVEYLOGISTIC to obtain the pseudo-estimate $\hat{\boldsymbol{\theta}}$ of the model parameter $\boldsymbol{\theta}$. 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; the covariance matrix of $\hat{\boldsymbol{\theta}}$ is estimated in the section “Variance Estimation” on page 9739. For a generalized logit model, only the Newton-Raphson technique is available. You can use the TECHNIQUE= option in the MODEL statement to select a fitting algorithm.

Iteratively Reweighted Least Squares Algorithm (Fisher Scoring)

Let Y be the response variable that takes values $1, \ldots, k, k+1$ $(k \geq 1)$. Let j index all observations and Y_j be the value of response for the jth observation. Consider the multinomial variable $\mathbf{Z}_j = (Z_{1j}, \ldots, Z_{kj})'$ such that

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

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

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

where $\mathbf{W}_j = w_j f_j \mathbf{V}_j^{-1}$, and w_j and f_j are the WEIGHT and FREQ values of the jth observation.

With a starting value of $\boldsymbol{\theta}^{(0)}$, the pseudo-estimate of $\boldsymbol{\theta}$ is obtained iteratively as

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} + (\sum_j \mathbf{D}'_j \mathbf{W}_j \mathbf{D}_j)^{-1} \sum_j \mathbf{D}'_j \mathbf{W}_j (\mathbf{Z}_j - \boldsymbol{\pi}_j)$$

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

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

Newton-Raphson Algorithm

Let

$$\mathbf{g} = \sum_j w_j f_j \frac{\partial l_j}{\partial \boldsymbol{\theta}}$$

$$\mathbf{H} = \sum_j -w_j f_j \frac{\partial^2 l_j}{\partial \boldsymbol{\theta}^2}$$

be the gradient vector and the Hessian matrix, where $l_j = \log L_j$ is the log likelihood for the jth observation. With a starting value of $\boldsymbol{\theta}^{(0)}$, the pseudo-estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is obtained iteratively until convergence is obtained:

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} + \mathbf{H}^{-1}\mathbf{g}$$

where $\mathbf{H}$ and $\mathbf{g}$ are evaluated at the ith iteration $\boldsymbol{\theta}^{(i)}$. If the log likelihood evaluated at $\boldsymbol{\theta}^{(i+1)}$ is less than that evaluated at $\boldsymbol{\theta}^{(i)}$, then $\boldsymbol{\theta}^{(i+1)}$ is recomputed by step-halving or ridging. The iterative scheme continues until convergence is obtained—that is, until $\boldsymbol{\theta}^{(i+1)}$ is sufficiently close to $\boldsymbol{\theta}^{(i)}$. Then the maximum likelihood estimate of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(i+1)}$.

Convergence Criteria

Four convergence criteria are allowed: 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 pseudo-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 $\mathbf{x}_j$ be the row 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,

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

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

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

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 pseudo-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 SURVEYLOGISTIC 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 that gives 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, stops 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 5,000, 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.

Model Fitting Statistics

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

- –2 log likelihood:

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

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

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

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

- Akaike information criterion:

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

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

- Schwarz criterion:

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

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

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.

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:

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

where $L(\mathbf{0})$ is the likelihood of the intercept-only model, $L(\hat{\boldsymbol{\theta}})$ is the likelihood of the specified model, and N is the population size. The quantity R^2 achieves a maximum of less than 1 for discrete models, where the maximum is given by

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

Nagelkerke (1991) proposes the following adjusted coefficient, which can achieve a maximum value of 1:

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

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 $\tilde{R}^2$.

INEST= Data Set

You can specify starting values for the iterative algorithm in the INEST= data set.

The INEST= data set contains one observation for each BY group. The INEST= data set must contain the intercept variables (named Intercept for binary response models 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.

Survey Design Information

Specification of Population Totals and Sampling Rates

To include a finite population correction (*fpc*) in Taylor series or bootstrap variance estimation, you can input either the sampling rate or the population total by using the RATE= or TOTAL= option, respectively, in the PROC SURVEYLOGISTIC statement. (You cannot specify both of these options in the same PROC SURVEYLOGISTIC statement.) The RATE= and TOTAL= options apply only to Taylor series or bootstrap variance estimation. The procedure does not use a finite population correction for BRR or jackknife variance estimation.

If you do not specify the RATE= or TOTAL= option, the Taylor series or bootstrap variance estimation does not include a finite population correction. For fairly small sampling fractions, it is appropriate to ignore this correction. For more information, see Cochran (1977) and Kish (1965).

If your design has multiple stages of selection and you are specifying the RATE= option, you should input the first-stage sampling rate, which is the ratio of the number of primary sampling units (PSUs) in the sample to the total number of PSUs in the study population. If you are specifying the TOTAL= option for a multistage design, you should input the total number of PSUs in the study population. For more information, see the section "Primary Sampling Units (PSUs)" on page 9735.

For a nonstratified sample design, or for a stratified sample design that has the same sampling rate or the same population total in all strata, you can use the RATE=*value* or TOTAL=*value* option. If your sample design is stratified with different sampling rates or population totals in different strata, use the RATE=*SAS-data-set* or TOTAL=*SAS-data-set* option to name a SAS data set that contains the stratum sampling rates or totals. This data set is called a *secondary data set*, as opposed to the *primary data set* that you specify in the DATA= option.

The secondary data set must contain all the stratification variables that are listed in the STRATA statement and all the variables that are listed in the BY statement. If there are formats associated with the STRATA variables and the BY variables, then the formats must be consistent in the primary and the secondary data sets. If you specify the TOTAL=*SAS-data-set* option, the secondary data set must have a variable named _TOTAL_ that contains the stratum population totals. Or if you specify the RATE=*SAS-data-set* option, the secondary data set must have a variable named _RATE_ that contains the stratum sampling rates. If the secondary data set contains more than one observation for any one stratum, then the procedure uses the first value of _TOTAL_ or _RATE_ for that stratum and ignores the rest.

The *value* in the RATE= option or the values of _RATE_ in the secondary data set must be nonnegative numbers. You can specify *value* as a number between 0 and 1. Or you can specify *value* in percentage form as a number between 1 and 100, and PROC SURVEYLOGISTIC converts that number to a proportion. The procedure treats the value 1 as 100% instead of 1%.

If you specify the TOTAL=*value* option, *value* must not be less than the sample size. If you provide stratum population totals in a secondary data set, these values must not be less than the corresponding stratum sample sizes.

Primary Sampling Units (PSUs)

When you have clusters, or primary sampling units (PSUs), in your sample design, the procedure estimates variance from the variation among PSUs when the Taylor series variance method is used. For more information, see the section "Taylor Series (Linearization)" on page 9740.

BRR or jackknife variance estimation methods draw multiple replicates (also called subsamples) from the full sample by following a specific resampling scheme. These subsamples are constructed by deleting PSUs from the full sample.

The bootstrap variance estimation method repeatedly draws a simple random sample with replacement of primary sampling units (PSUs) from the full sample to form each replicate.

If you use a REPWEIGHTS statement to provide replicate weights for BRR, jackknife, bootstrap variance estimation, you do not need to specify a CLUSTER statement. Otherwise, you should specify a CLUSTER statement whenever your design includes clustering at the first stage of sampling. If you do not specify a CLUSTER statement, then PROC SURVEYLOGISTIC treats each observation as a PSU.

Logistic Regression Models and Parameters

The SURVEYLOGISTIC procedure fits a logistic regression model and estimates the corresponding regression parameters. Each model uses the link function you specified in the LINK= option in the MODEL statement. There are four types of model you can use with the procedure: cumulative logit model, complementary log-log model, probit model, and generalized logit model.

Notation

Let Y be the response variable with categories $1, 2, \ldots, D, D + 1$. The p covariates are denoted by a p-dimension row vector $\mathbf{x}$.

For a stratified clustered sample design, each observation is represented by a row vector, $(w_{hij}, \mathbf{y}'_{hij}, y_{hij(D+1)}, \mathbf{x}_{hij})$, where

- $h = 1, 2, \ldots, H$ is the stratum index
- $i = 1, 2, \ldots, n_h$ is the cluster index within stratum h
- $j = 1, 2, \ldots, m_{hi}$ is the unit index within cluster i of stratum h
- w_{hij} denotes the sampling weight
- $\mathbf{y}_{hij}$ is a D-dimensional column vector whose elements are indicator variables for the first D categories for variable Y. If the response of the jth unit of the ith cluster in stratum h falls in category d, the dth element of the vector is one, and the remaining elements of the vector are zero, where $d = 1, 2, \ldots, D$.
- $y_{hij(D+1)}$ is the indicator variable for the $(D + 1)$ category of variable Y
- $\mathbf{x}_{hij}$ denotes the k-dimensional row vector of explanatory variables for the jth unit of the ith cluster in stratum h. If there is an intercept, then $x_{hij1} \equiv 1$.
- $\tilde{n} = \sum_{h=1}^{H} n_h$ is the total number of clusters in the sample
- $n = \sum_{h=1}^{H} \sum_{i=1}^{n_h} m_{hi}$ is the total sample size

The following notations are also used:

- f_h denotes the sampling rate for stratum h

- $\boldsymbol{\pi}_{hij}$ is the expected vector of the response variable:

$$\begin{aligned}\boldsymbol{\pi}_{hij} &= E(\mathbf{y}_{hij}|\mathbf{x}_{hij}) \\ &= (\pi_{hij1}, \pi_{hij2}, \ldots, \pi_{hijD})' \\ \pi_{hij(D+1)} &= E(y_{hij(D+1)}|\mathbf{x}_{hij})\end{aligned}$$

Note that $\pi_{hij(D+1)} = 1 - \mathbf{1}'\boldsymbol{\pi}_{hij}$, where $\mathbf{1}$ is a D-dimensional column vector whose elements are 1.

Logistic Regression Models

If the response categories of the response variable Y can be restricted to a number of ordinal values, you can fit cumulative probabilities of the response categories with a cumulative logit model, a complementary log-log model, or a probit model. Details of cumulative logit models (or proportional odds models) can be found in McCullagh and Nelder (1989). If the response categories of Y are nominal responses without natural ordering, you can fit the response probabilities with a generalized logit model. Formulation of the generalized logit models for nominal response variables can be found in Agresti (2002). For each model, the procedure estimates the model parameter $\boldsymbol{\theta}$ by using a pseudo-log-likelihood function. The procedure obtains the pseudo-maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ by using iterations described in the section “Iterative Algorithms for Model Fitting” on page 9730 and estimates its variance described in the section “Variance Estimation” on page 9739.

Cumulative Logit Model

A cumulative logit model uses the **logit** function

$$g(t) = \log\left(\frac{t}{1-t}\right)$$

as the link function.

Denote the cumulative sum of the expected proportions for the first d categories of variable Y by

$$F_{hijd} = \sum_{r=1}^{d} \pi_{hijr}$$

for $d = 1, 2, \ldots, D$. Then the cumulative logit model can be written as

$$\log\left(\frac{F_{hijd}}{1 - F_{hijd}}\right) = \alpha_d + \mathbf{x}_{hij}\boldsymbol{\beta}$$

with the model parameters

$$\begin{aligned}\boldsymbol{\beta} &= (\beta_1, \beta_2, \ldots, \beta_k)' \\ \boldsymbol{\alpha} &= (\alpha_1, \alpha_2, \ldots, \alpha_D)', \ \ \alpha_1 < \alpha_2 < \cdots < \alpha_D \\ \boldsymbol{\theta} &= (\boldsymbol{\alpha}', \boldsymbol{\beta}')'\end{aligned}$$

Complementary Log-Log Model

A complementary log-log model uses the **complementary log-log** function

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

as the link function. Denote the cumulative sum of the expected proportions for the first d categories of variable Y by

$$F_{hijd} = \sum_{r=1}^{d} \pi_{hijr}$$

for $d = 1, 2, \ldots, D$. Then the complementary log-log model can be written as

$$\log(-\log(1 - F_{hijd})) = \alpha_d + \mathbf{x}_{hij}\boldsymbol{\beta}$$

with the model parameters

$$\begin{aligned}
\boldsymbol{\beta} &= (\beta_1, \beta_2, \ldots, \beta_k)' \\
\boldsymbol{\alpha} &= (\alpha_1, \alpha_2, \ldots, \alpha_D)', \ \ \alpha_1 < \alpha_2 < \cdots < \alpha_D \\
\boldsymbol{\theta} &= (\boldsymbol{\alpha}', \boldsymbol{\beta}')'
\end{aligned}$$

Probit Model

A probit model uses the **probit** (or normit) function, which is the inverse of the cumulative standard normal distribution function,

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

as the link function, where

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{1}{2}z^2} dz$$

Denote the cumulative sum of the expected proportions for the first d categories of variable Y by

$$F_{hijd} = \sum_{r=1}^{d} \pi_{hijr}$$

for $d = 1, 2, \ldots, D$. Then the probit model can be written as

$$F_{hijd} = \Phi(\alpha_d + \mathbf{x}_{hij}\boldsymbol{\beta})$$

with the model parameters

$$\begin{aligned}
\boldsymbol{\beta} &= (\beta_1, \beta_2, \ldots, \beta_k)' \\
\boldsymbol{\alpha} &= (\alpha_1, \alpha_2, \ldots, \alpha_D)', \ \ \alpha_1 < \alpha_2 < \cdots < \alpha_D \\
\boldsymbol{\theta} &= (\boldsymbol{\alpha}', \boldsymbol{\beta}')'
\end{aligned}$$

Generalized Logit Model

For nominal response, a generalized logit model is to fit the ratio of the expected proportion for each response category over the expected proportion of a reference category with a logit link function.

Without loss of generality, let category $D+1$ be the reference category for the response variable Y. Denote the expected proportion for the dth category by π_{hijd} as in the section "Notation" on page 9736. Then the generalized logit model can be written as

$$\log\left(\frac{\pi_{hijd}}{\pi_{hij(D+1)}}\right) = \mathbf{x}_{hij}\boldsymbol{\beta}_d$$

for $d = 1, 2, \ldots, D$, with the model parameters

$$\begin{aligned} \boldsymbol{\beta}_d &= (\beta_{d1}, \beta_{d2}, \ldots, \beta_{dk})' \\ \boldsymbol{\theta} &= (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2, \ldots, \boldsymbol{\beta}'_D)' \end{aligned}$$

Likelihood Function

Let $\mathbf{g}(\cdot)$ be a link function such that

$$\boldsymbol{\pi} = \mathbf{g}(\mathbf{x}, \boldsymbol{\theta})$$

where $\boldsymbol{\theta}$ is a column vector for regression coefficients. The pseudo-log likelihood is

$$l(\boldsymbol{\theta}) = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\left((\log(\boldsymbol{\pi}_{hij}))'\mathbf{y}_{hij} + \log(\pi_{hij(D+1)})y_{hij(D+1)}\right)$$

Denote the pseudo-estimator as $\hat{\boldsymbol{\theta}}$, which is a solution to the estimating equations:

$$\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\mathbf{D}_{hij}\left(\mathrm{diag}(\boldsymbol{\pi}_{hij}) - \boldsymbol{\pi}_{hij}\boldsymbol{\pi}'_{hij}\right)^{-1}(\mathbf{y}_{hij} - \boldsymbol{\pi}_{hij}) = \mathbf{0}$$

where $\mathbf{D}_{hij}$ is the matrix of partial derivatives of the link function $\mathbf{g}$ with respect to $\boldsymbol{\theta}$.

To obtain the pseudo-estimator $\hat{\boldsymbol{\theta}}$, the procedure uses iterations with a starting value $\boldsymbol{\theta}^{(0)}$ for $\boldsymbol{\theta}$. See the section "Iterative Algorithms for Model Fitting" on page 9730 for more details.

Variance Estimation

Due to the variability of characteristics among items in the population, researchers apply scientific sample designs in the sample selection process to reduce the risk of a distorted view of the population, and they make inferences about the population based on the information from the sample survey data. In order to make statistically valid inferences for the population, they must incorporate the sample design in the data analysis.

The SURVEYLOGISTIC procedure fits linear logistic regression models for discrete response survey data by using the maximum likelihood method. In the variance estimation, the procedure uses the Taylor series (linearization) method or replication (resampling) methods to estimate sampling errors of estimators based on complex sample designs, including designs with stratification, clustering, and unequal weighting (Binder 1981, 1983; Roberts, Rao, and Kumar 1987; Skinner, Holt, and Smith 1989; Binder and Roberts 2003; Morel 1989; Lehtonen and Pahkinen 1995; Woodruff 1971; Fuller 1975; Särndal, Swensson, and Wretman 1992; Fuller 2009; Wolter 2007; Rust 1985; Dippo, Fay, and Morganstein 1984; Rao and Shao 1999; Rao, Wu, and Yue 1992; Rao and Shao 1996).

You can use the VARMETHOD= option to specify a variance estimation method to use. By default, the Taylor series method is used. However, replication methods have recently gained popularity for estimating variances in complex survey data analysis. One reason for this popularity is the relative simplicity of replication-based estimates, especially for nonlinear estimators; another is that modern computational capacity has made replication methods feasible for practical survey analysis.

Replication methods draw multiple replicates (also called subsamples) from a full sample according to a specific resampling scheme. The most commonly used resampling schemes are the *balanced repeated replication* (BRR) method, the *jackknife* method, and the *bootstrap* method. For each replicate, the original weights are modified for the PSUs in the replicates in order to create replicate weights. The parameters of interest are estimated by using the replicate weights for each replicate. Then the variances of parameters of interest are estimated by the variability among the estimates derived from these replicates. You can use the REPWEIGHTS statement to provide your own replicate weights for variance estimation.

The following sections provide details about how the variance-covariance matrix of the estimated regression coefficients is estimated for each variance estimation method.

Taylor Series (Linearization)

The Taylor series (linearization) method is the most commonly used method to estimate the covariance matrix of the regression coefficients for complex survey data. It is the default variance estimation method used by PROC SURVEYLOGISTIC.

Using the notation described in the section “Notation” on page 9736, the estimated covariance matrix of model parameters $\hat{\boldsymbol{\theta}}$ by the Taylor series method is

$$\hat{V}(\hat{\boldsymbol{\theta}}) = \hat{\mathbf{Q}}^{-1}\hat{\mathbf{G}}\hat{\mathbf{Q}}^{-1}$$

where

$$\begin{aligned}
\hat{\mathbf{Q}} &= \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\hat{\mathbf{D}}_{hij}\left(\mathrm{diag}(\hat{\boldsymbol{\pi}}_{hij}) - \hat{\boldsymbol{\pi}}_{hij}\hat{\boldsymbol{\pi}}'_{hij}\right)^{-1}\hat{\mathbf{D}}'_{hij} \\
\hat{\mathbf{G}} &= \frac{n-1}{n-p}\sum_{h=1}^{H}\frac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_h}(\mathbf{e}_{hi\cdot} - \bar{\mathbf{e}}_{h\cdot\cdot})(\mathbf{e}_{hi\cdot} - \bar{\mathbf{e}}_{h\cdot\cdot})' \\
\mathbf{e}_{hi\cdot} &= \sum_{j=1}^{m_{hi}} w_{hij}\hat{\mathbf{D}}_{hij}\left(\mathrm{diag}(\hat{\boldsymbol{\pi}}_{hij}) - \hat{\boldsymbol{\pi}}_{hij}\hat{\boldsymbol{\pi}}'_{hij}\right)^{-1}(\mathbf{y}_{hij} - \hat{\boldsymbol{\pi}}_{hij}) \\
\bar{\mathbf{e}}_{h\cdot\cdot} &= \frac{1}{n_h}\sum_{i=1}^{n_h}\mathbf{e}_{hi\cdot}
\end{aligned}$$

and $\mathbf{D}_{hij}$ is the matrix of partial derivatives of the link function g with respect to $\boldsymbol{\theta}$ and $\hat{\mathbf{D}}_{hij}$ and the response probabilities $\hat{\boldsymbol{\pi}}_{hij}$ are evaluated at $\hat{\boldsymbol{\theta}}$.

If you specify the TECHNIQUE=NEWTON option in the MODEL statement to request the Newton-Raphson algorithm, the matrix $\hat{\mathbf{Q}}$ is replaced by the negative (expected) Hessian matrix when the estimated covariance matrix $\hat{V}(\hat{\boldsymbol{\theta}})$ is computed.

Adjustments to the Variance Estimation

The factor $(n-1)/(n-p)$ in the computation of the matrix $\hat{\mathbf{G}}$ reduces the small sample bias associated with using the estimated function to calculate deviations (Morel 1989; Hidiroglou, Fuller, and Hickman 1980). For simple random sampling, this factor contributes to the degrees-of-freedom correction applied to the residual mean square for ordinary least squares in which p parameters are estimated. By default, the procedure uses this adjustment in Taylor series variance estimation. It is equivalent to specifying the VADJUST=DF option in the MODEL statement. If you do not want to use this multiplier in the variance estimation, you can specify the VADJUST=NONE option in the MODEL statement to suppress this factor.

In addition, you can specify the VADJUST=MOREL option to request an adjustment to the variance estimator for the model parameters $\hat{\boldsymbol{\theta}}$, introduced by Morel (1989):

$$\hat{V}(\hat{\boldsymbol{\theta}}) = \hat{\mathbf{Q}}^{-1}\hat{\mathbf{G}}\hat{\mathbf{Q}}^{-1} + \kappa\lambda\hat{\mathbf{Q}}^{-1}$$

where for given nonnegative constants δ and ϕ,

$$\begin{aligned} \kappa &= \max\left(\delta,\ p^{-1}\mathrm{tr}\left(\hat{\mathbf{Q}}^{-1}\hat{\mathbf{G}}\right)\right) \\ \lambda &= \min\left(\phi, \frac{p}{\tilde{n}-p}\right) \end{aligned}$$

The adjustment $\kappa\lambda\hat{\mathbf{Q}}^{-1}$ does the following:

- reduces the small sample bias reflected in inflated Type I error rates
- guarantees a positive-definite estimated covariance matrix provided that $\hat{\mathbf{Q}}^{-1}$ exists
- is close to zero when the sample size becomes large

In this adjustment, κ is an estimate of the design effect, which has been bounded below by the positive constant δ. You can use DEFFBOUND=δ in the VADJUST=MOREL option in the MODEL statement to specify this lower bound; by default, the procedure uses $\delta = 1$. The factor λ converges to zero when the sample size becomes large, and λ has an upper bound ϕ. You can use ADJBOUND=ϕ in the VADJUST=MOREL option in the MODEL statement to specify this upper bound; by default, the procedure uses $\phi = 0.5$.

Bootstrap Method

The VARMETHOD=BOOTSTRAP option in the PROC SURVEYLOGISTIC statement requests the bootstrap method for variance estimation. This method can be used for stratified sample designs and for designs that have no stratification. If your design is stratified, the bootstrap method requires at least two PSUs in each stratum. You can provide your own bootstrap replicate weights for the analysis by using a REPWEIGHTS statement, or the procedure can construct bootstrap replicate weights for the analysis.

PROC SURVEYLOGISTIC estimates the parameter of interest (or requested statistics) from each replicate, and then uses the variability among replicate estimates to estimate the overall variance of these statistics.

This bootstrap method for complex survey data is similar to the method of Rao, Wu, and Yue (1992) and is also known as the bootstrap weights method (Mashreghi, Haziza, and Léger 2016). For more information, see Lohr (2010, Section 9.3.3), Wolter (2007, Chapter 5), Beaumont and Patak (2012), Fuller (2009, Section 4.5), and Shao and Tu (1995, Section 6.2.4). McCarthy and Snowden (1985), Rao and Wu (1988), Sitter (1992b), and Sitter (1992a) provide several adjusted bootstrap variance estimators that are consistent for complex survey data. The naive bootstrap variance estimator that is suitable for infinite populations is not consistent for complex survey data.

Replicate Weight Construction

If you do not provide replicate weights by using a REPWEIGHTS statement, PROC SURVEYLOGISTIC constructs bootstrap replicate weights for the analysis. The procedure selects replicate bootstrap samples by with-replacement random sampling of PSUs within strata. You can specify the number of bootstrap replicates in the REPS= *method-option*; by default, the number of replicates is 250. (Increasing the number of replicates can improve the estimation precision but also increases the computation time.) You can specify the bootstrap sample sizes m_h in the MH= *method-option*; by default, $m_h = n_h - 1$, where n_h is the number of PSUs in stratum h.

In each replicate sample, the original sampling weights of the selected units are adjusted to reflect the full sample. These adjusted weights are the *bootstrap replicate weights*. In replicate r, the bootstrap replicate weight for observation j in PSU i in stratum h is computed as

$$\tilde{w}_{hij}^{(r)} = \left\{ 1 + \sqrt{\frac{1-f_h}{m_h(n_h-1)}} n_h k_{hi}^{(r)} - \sqrt{\frac{m_h(1-f_h)}{n_h-1}} \right\} w_{hij}$$

where $k_{hi}^{(r)}$ is the number of times PSU i is selected for replicate r, and f_h is the sampling fraction in stratum h that you can input by using either the RATE= or TOTAL= option in the PROC SURVEYLOGISTIC statement.

You can use the OUTWEIGHTS=*SAS-data-set* method-option to store the bootstrap replicate weights in a SAS data set. For information about the contents of the OUTWEIGHTS= data set, see the section "Replicate Weights Output Data Set" on page 9759. You can provide these replicate weights to the procedure for subsequent analyses by using a REPWEIGHTS statement.

Variance Estimation

Let R be the total number of bootstrap replicates weights. Denote α_r as the replicate coefficient for the rth replicate ($r = 1, 2, \ldots, R$).

When the procedure generates the bootstrap replicate weights, $\alpha_r = 1/R$.

If you provide your own bootstrap replicate weights by including a REPWEIGHTS statement, you can specify α_r in the REPCOEFS= option. By default, $\alpha_r = 1/R$.

Let $\hat{\boldsymbol{\theta}}$ be the estimated regression coefficients from the full sample for $\boldsymbol{\theta}$, and let $\hat{\boldsymbol{\theta}}_r$ be the estimated regression coefficient when the rth set of bootstrap replicate weights are used. PROC SURVEYLOGISTIC estimates the covariance matrix of $\hat{\boldsymbol{\theta}}$ by the following equation:

$$\hat{\mathrm{V}}(\hat{\boldsymbol{\theta}}) = \sum_{r=1}^{R} \alpha_r \left(\hat{\boldsymbol{\theta}}_r - \hat{\boldsymbol{\theta}} \right) \left(\hat{\boldsymbol{\theta}}_r - \hat{\boldsymbol{\theta}} \right)'$$

Here, the degrees of freedom is the number of clusters minus the number of strata. If there are no clusters, then the degrees of freedom equals the number of observations minus the number of strata. If the design is not stratified, then the degrees of freedom equals the number of PSUs minus one.

If you provide your own replicate weights without specifying the DF= option, the degrees of freedom is set to be the number of replicates R.

If you specify the CENTER=REPLICATES *method-option*, then PROC SURVEYLOGISTIC computes the covariance matrix of $\hat{\boldsymbol{\theta}}$ by

$$\hat{\mathrm{V}}(\hat{\boldsymbol{\theta}}) = \sum_{r=1}^{R} \alpha_r \left(\hat{\boldsymbol{\theta}}_r - \overline{\hat{\boldsymbol{\theta}}}\right)\left(\hat{\boldsymbol{\theta}}_r - \overline{\hat{\boldsymbol{\theta}}}\right)'$$

where $\overline{\hat{\boldsymbol{\theta}}}$ is the average of the replicate estimates and is calculated as follows:

$$\overline{\hat{\boldsymbol{\theta}}} = \frac{1}{R}\sum_{r=1}^{R} \hat{\boldsymbol{\theta}}_r$$

If a parameter cannot be computed from one or more replicates, then the procedure computes the variance estimate by using those replicates from which the parameter can be estimated, and the number of those replicates, R', replaces the original number of replicates, R, in the variance estimation.

If you do not provide your own value for the degrees of freedom and if R' is less than the number of PSUs minus the number of strata, then the degrees of freedom is set to R'.

Balanced Repeated Replication (BRR) Method

The balanced repeated replication (BRR) method requires that the full sample be drawn by using a stratified sample design with two primary sampling units (PSUs) per stratum. Let H be the total number of strata. The total number of replicates R is the smallest multiple of 4 that is greater than H. However, if you prefer a larger number of replicates, you can specify the REPS=*number* option. If a *number* $\times$ *number* Hadamard matrix cannot be constructed, the number of replicates is increased until a Hadamard matrix becomes available.

Each replicate is obtained by deleting one PSU per stratum according to the corresponding Hadamard matrix and adjusting the original weights for the remaining PSUs. The new weights are called replicate weights.

Replicates are constructed by using the first H columns of the $R \times R$ Hadamard matrix. The rth ($r = 1, 2, \ldots, R$) replicate is drawn from the full sample according to the rth row of the Hadamard matrix as follows:

- If the (r, h) element of the Hadamard matrix is 1, then the first PSU of stratum h is included in the rth replicate and the second PSU of stratum h is excluded.
- If the (r, h) element of the Hadamard matrix is –1, then the second PSU of stratum h is included in the rth replicate and the first PSU of stratum h is excluded.

Note that the “first” and “second” PSUs are determined by data order in the input data set. Thus, if you reorder the data set and perform the same analysis by using BRR method, you might get slightly different results, because the contents in each replicate sample might change.

The replicate weights of the remaining PSUs in each half-sample are then doubled to their original weights. For more information about the BRR method, see Wolter (2007) and Lohr (2010).

By default, an appropriate Hadamard matrix is generated automatically to create the replicates. You can request that the Hadamard matrix be displayed by specifying the VARMETHOD=BRR(PRINTH) *method-option*. If you provide a Hadamard matrix by specifying the VARMETHOD=BRR(HADAMARD=) *method-option*, then the replicates are generated according to the provided Hadamard matrix.

You can use the VARMETHOD=BRR(OUTWEIGHTS=) *method-option* to save the replicate weights into a SAS data set.

Let $\hat{\boldsymbol{\theta}}$ be the estimated regression coefficients from the full sample for $\boldsymbol{\theta}$, and let $\hat{\boldsymbol{\theta}}_r$ be the estimated regression coefficient from the rth replicate by using replicate weights. PROC SURVEYLOGISTIC estimates the covariance matrix of $\hat{\boldsymbol{\theta}}$ by

$$\hat{\mathrm{V}}(\hat{\boldsymbol{\theta}}) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\boldsymbol{\theta}}_r - \hat{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}}_r - \hat{\boldsymbol{\theta}}\right)'$$

with H degrees of freedom, where H is the number of strata. If you provide your own replicate weights without specifying the DF= option, the degrees of freedom is set to be the number of replicates R.

If you specify the CENTER=REPLICATES *method-option*, then PROC SURVEYLOGISTIC computes the covariance matrix of $\hat{\boldsymbol{\theta}}$ by

$$\hat{\mathrm{V}}(\hat{\boldsymbol{\theta}}) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\boldsymbol{\theta}}_r - \overline{\hat{\boldsymbol{\theta}}}\right)\left(\hat{\boldsymbol{\theta}}_r - \overline{\hat{\boldsymbol{\theta}}}\right)'$$

where $\overline{\hat{\boldsymbol{\theta}}}$ is the average of the replicate estimates and is calculated as follows:

$$\overline{\hat{\boldsymbol{\theta}}} = \frac{1}{R}\sum_{r=1}^{R}\hat{\boldsymbol{\theta}}_r$$

If a parameter cannot be computed from one or more replicates, then the procedure computes the variance estimate by using those replicates from which the parameter can be estimated, and the number of those replicates, R', replaces the original number of replicates, R, in the variance estimation.

If you do not provide your own value for the degrees of freedom, then the degrees of freedom equals the minimum between R' and the number of strata, H.

Fay's BRR Method

Fay's method is a modification of the BRR method, and it requires a stratified sample design with two primary sampling units (PSUs) per stratum. The total number of replicates R is the smallest multiple of 4 that is greater than the total number of strata H. However, if you prefer a larger number of replicates, you can specify the REPS= *method-option*.

For each replicate, Fay's method uses a Fay coefficient $0 \le \epsilon < 1$ to impose a perturbation of the original weights in the full sample that is gentler than using only half-samples, as in the traditional BRR method. The Fay coefficient $0 \le \epsilon < 1$ can be set by specifying the FAY $= \epsilon$ *method-option*. By default, $\epsilon = 0.5$ if the FAY *method-option* is specified without providing a value for ϵ (Judkins 1990; Rao and Shao 1999).

When $\epsilon = 0$, Fay's method becomes the traditional BRR method. For more information, see Dippo, Fay, and Morganstein (1984); Fay (1984, 1989); Judkins (1990).

Let H be the number of strata. Replicates are constructed by using the first H columns of the $R \times R$ Hadamard matrix, where R is the number of replicates, $R > H$. The rth ($r = 1, 2, \ldots, R$) replicate is created from the full sample according to the rth row of the Hadamard matrix as follows:

- If the (r, h) element of the Hadamard matrix is 1, then the full sample weight of the first PSU in stratum h is multiplied by ϵ and the full sample weight of the second PSU is multiplied by $2 - \epsilon$ to obtain the rth replicate weights.

- If the (r, h) element of the Hadamard matrix is –1, then the full sample weight of the first PSU in stratum h is multiplied by $2 - \epsilon$ and the full sample weight of the second PSU is multiplied by ϵ to obtain the rth replicate weights.

You can use the VARMETHOD=BRR(OUTWEIGHTS=) *method-option* to save the replicate weights into a SAS data set.

By default, an appropriate Hadamard matrix is generated automatically to create the replicates. You can request that the Hadamard matrix be displayed by specifying the VARMETHOD=BRR(PRINTH) *method-option*. If you provide a Hadamard matrix by specifying the VARMETHOD=BRR(HADAMARD=) *method-option*, then the replicates are generated according to the provided Hadamard matrix.

Let $\hat{\boldsymbol{\theta}}$ be the estimated regression coefficients from the full sample for $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\theta}}_r$ be the estimated regression coefficient obtained from the rth replicate by using replicate weights. PROC SURVEYLOGISTIC estimates the covariance matrix of $\hat{\boldsymbol{\theta}}$ by

$$\hat{\mathrm{V}}(\hat{\boldsymbol{\theta}}) = \frac{1}{R(1-\epsilon)^2} \sum_{r=1}^{R} \left(\hat{\boldsymbol{\theta}}_r - \hat{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}}_r - \hat{\boldsymbol{\theta}}\right)'$$

with H degrees of freedom, where H is the number of strata. If you provide your own replicate weights without specifying the DF= option, the degrees of freedom is set to be the number of replicates R.

If you specify the CENTER=REPLICATES *method-option*, then PROC SURVEYLOGISTIC computes the covariance matrix of $\hat{\boldsymbol{\theta}}$ by

$$\hat{\mathrm{V}}(\hat{\boldsymbol{\theta}}) = \frac{1}{R(1-\epsilon)^2} \sum_{r=1}^{R} \left(\hat{\boldsymbol{\theta}}_r - \overline{\hat{\boldsymbol{\theta}}}\right)\left(\hat{\boldsymbol{\theta}}_r - \overline{\hat{\boldsymbol{\theta}}}\right)'$$

where $\overline{\hat{\boldsymbol{\theta}}}$ is the average of the replicate estimates and is calculated as follows:

$$\overline{\hat{\boldsymbol{\theta}}} = \frac{1}{R} \sum_{r=1}^{R} \hat{\boldsymbol{\theta}}_r$$

If a parameter cannot be computed from one or more replicates, then the procedure computes the variance estimate by using those replicates from which the parameter can be estimated, and the number of those replicates, R', replaces the original number of replicates, R, in the variance estimation.

If you do not provide your own value for the degrees of freedom, then the degrees of freedom equals the minimum between R' and the number of strata, H.

Jackknife Method

The jackknife method of variance estimation deletes one PSU at a time from the full sample to create replicates. The total number of replicates R is the same as the total number of PSUs. In each replicate, the sample weights of the remaining PSUs are modified by the jackknife coefficient α_r. The modified weights are called replicate weights.

The jackknife coefficient and replicate weights are described as follows.

Without Stratification If there is no stratification in the sample design (no STRATA statement), the jackknife coefficients α_r are the same for all replicates:

$$\alpha_r = \frac{R-1}{R} \quad \text{where } r = 1, 2, \ldots, R$$

Denote the original weight in the full sample for the jth member of the ith PSU as w_{ij}. If the ith PSU is included in the rth replicate ($r = 1, 2, \ldots, R$), then the corresponding replicate weight for the jth member of the ith PSU is defined as

$$w_{ij}^{(r)} = w_{ij}/\alpha_r$$

With Stratification If the sample design involves stratification, each stratum must have at least two PSUs to use the jackknife method.

Let stratum $\tilde{h}_r$ be the stratum from which a PSU is deleted for the rth replicate. Stratum $\tilde{h}_r$ is called the *donor stratum.* Let $n_{\tilde{h}_r}$ be the total number of PSUs in the donor stratum $\tilde{h}_r$. The jackknife coefficients are defined as

$$\alpha_r = \frac{n_{\tilde{h}_r} - 1}{n_{\tilde{h}_r}} \quad \text{where } r = 1, 2, \ldots, R$$

Denote the original weight in the full sample for the jth member of the ith PSU as w_{ij}. If the ith PSU is included in the rth replicate ($r = 1, 2, \ldots, R$), then the corresponding replicate weight for the jth member of the ith PSU is defined as

$$w_{ij}^{(r)} = \begin{cases} w_{ij} & \text{if } i\text{th PSU is not in the donor stratum } \tilde{h}_r \\ w_{ij}/\alpha_r & \text{if } i\text{th PSU is in the donor stratum } \tilde{h}_r \end{cases}$$

You can use the VARMETHOD=JACKKNIFE(OUTJKCOEFS=) *method-option* to save the jackknife coefficients into a SAS data set and use the VARMETHOD=JACKKNIFE(OUTWEIGHTS=) *method-option* to save the replicate weights into a SAS data set.

If you provide your own replicate weights in a REPWEIGHTS statement, then you can also provide corresponding jackknife coefficients in the JKCOEFS= or REPCOEFS= option. If you provide replicate weights but do not provide jackknife coefficients, PROC SURVEYLOGISTIC uses $\alpha_r = (R-1)/R$ as the jackknife coefficient for all replicates by default.

Let $\hat{\boldsymbol{\theta}}$ be the estimated regression coefficients from the full sample for $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\theta}}_r$ be the estimated regression coefficient obtained from the rth replicate by using replicate weights. PROC SURVEYLOGISTIC estimates the covariance matrix of $\hat{\boldsymbol{\theta}}$ by

$$\hat{\mathrm{V}}(\hat{\boldsymbol{\theta}}) = \sum_{r=1}^{R} \alpha_r \left(\hat{\boldsymbol{\theta}}_r - \hat{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}}_r - \hat{\boldsymbol{\theta}}\right)'$$

with $R - H$ degrees of freedom, where R is the number of replicates and H is the number of strata, or $R - 1$ when there is no stratification. If you provide your own replicate weights without specifying the DF= option, the degrees of freedom is set to be the number of replicates R.

If you specify the CENTER=REPLICATES *method-option*, then PROC SURVEYLOGISTIC computes the covariance matrix of $\hat{\boldsymbol{\theta}}$ by

$$\hat{\mathrm{V}}(\hat{\boldsymbol{\theta}}) = \sum_{r=1}^{R} \alpha_r \left(\hat{\boldsymbol{\theta}}_r - \overline{\hat{\boldsymbol{\theta}}}\right)\left(\hat{\boldsymbol{\theta}}_r - \overline{\hat{\boldsymbol{\theta}}}\right)'$$

where $\overline{\hat{\boldsymbol{\theta}}}$ is the average of the replicate estimates and is calculated as follows:

$$\overline{\hat{\boldsymbol{\theta}}} = \frac{1}{R}\sum_{r=1}^{R} \hat{\boldsymbol{\theta}}_r$$

If a parameter cannot be computed from one or more replicates, then the procedure computes the variance estimate by using those replicates from which the parameter can be estimated, and the number of those replicates, R', replaces the original number of replicates, R, in the variance estimation.

If you do not provide your own value for the degrees of freedom, then the degrees of freedom equals $R' - H$.

Hadamard Matrix

A Hadamard matrix **H** is a square matrix whose elements are either 1 or –1 such that

$$\mathbf{H}\mathbf{H}' = k\mathbf{I}$$

where k is the dimension of **H** and **I** is the identity matrix of order k. The order k is necessarily 1, 2, or a positive integer that is a multiple of 4.

For example, the following matrix is a Hadamard matrix of dimension k = 8:

$$\begin{array}{rrrrrrrr}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\
1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\
1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\
1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\
1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\
1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\
1 & -1 & -1 & 1 & -1 & 1 & 1 & -1
\end{array}$$

Domain Analysis

A DOMAIN statement requests that the procedure perform logistic regression analysis for each domain.

For a domain Ω, let I_Ω be the corresponding indicator variable:

$$I_\Omega(h,i,j) = \begin{cases} 1 & \text{if observation } (h,i,j) \text{ belongs to } \Omega \\ 0 & \text{otherwise} \end{cases}$$

Let

$$v_{hij} = w_{hij} I_{\Omega}(h,i,j) = \begin{cases} w_{hij} & \text{if observation } (h,i,j) \text{ belongs to } \Omega \\ 0 & \text{otherwise} \end{cases}$$

The regression in domain Ω uses v as the weight variable.

Hypothesis Testing and Estimation

Degrees of Freedom

In this section, degrees of freedom (*df*) refers to the denominator degrees of freedom for F statistics in hypothesis testing. It also refers to the degrees of freedom in t tests for parameter estimates and odds ratio estimates, and for computing t distribution percentiles for confidence limits of these estimates. The value of *df* is determined by the design degrees of freedom f and by what you specify in the DF= option in the MODEL statement.

The default *df* is determined as

$$df = \begin{cases} f - r + 1 & \text{for Taylor variance estimation method} \\ f & \text{for replication variance estimation methods} \end{cases}$$

where f is the design degrees of freedom and r is the rank of the contrast of model parameters to be tested.

Design Degrees of Freedom f

The design degrees of freedom f is determined by the survey design and the variance estimation method.

Design Degrees of Freedom *f* for the Taylor Series Method For Taylor series variance estimation, the design degrees of freedom f can depend on the number of clusters, the number of strata, and the number of observations. These numbers are based on the observations that are included in the analysis; they do not count observations that are excluded from the analysis because of missing values. If all values in a stratum are excluded from the analysis as missing values, then that stratum is called an *empty stratum*. Empty strata are not counted in the total number of strata for the analysis. Similarly, empty clusters and missing observations are not included in the totals counts of clusters and observations that are used to compute the f for the analysis.

If you specify the MISSING option in the CLASS statement, missing values are treated as valid nonmissing levels and are included in determining the f. If you specify the NOMCAR option for Taylor series variance estimation, observations that have missing values for variables in the regression model are included. For more information about missing values, see the section "Missing Values" on page 9724.

Using the notation that is defined in the section "Notation" on page 9736, let $\tilde{n}$ be the total number of clusters if the design has a CLUSTER statement; let n be the total sample size; and let H be the number of strata if there is a STRATA statement, or 1 otherwise. Then for Taylor series variance estimation, the design degrees of freedom is

$$f = \begin{cases} \tilde{n} - H & \text{if the design contains clusters} \\ n - H & \text{if the design does not contain clusters} \end{cases}$$

Design Degrees of Freedom f for Replication Methods For replication variance estimation methods that include a REPWEIGHTS statement, the design degrees of freedom f equals the number of REPWEIGHTS variables, unless you specify an alternative in the DF= option in a REPWEIGHTS statement.

When the BRR method (including Fay's method) is requested but no REPWEIGHTS statement is specified, the design degrees of freedom f equals the number of strata.

When the jackknife method or the bootstrap method is requested but no REPWEIGHTS statement is specified, the design degrees of freedom f is the same as in the Taylor series method described in the previous section.

Setting Design Degrees of Freedom f to a Specific Value

If you do not want to use the default design degrees of freedom, then you can specify the DF=DESIGN(*value*) or DF=PARMADJ(*value*) (for Taylor method only) option in the MODEL statement, where *value* is a positive number. Then f=*value*.

However, if you specify the *value* in the DF= option in the MODEL statement as well as with the DF= option in a REPWEIGHTS statement, then the *df* is determined by the *value* in the MODEL statement, and the DF= option in the REPWEIGHTS statement is ignored.

Setting Design Degrees of Freedom to Infinity

If you specify DF=INFINITY in the MODEL statement, then the *df* is set to be infinite.

As the denominator degrees of freedom grows, an F distribution approaches a chi-square distribution, and similarly a t distribution approaches a normal distribution. Therefore, when you specify DF=INFINITY in the MODEL statement), PROC SURVEYLOGISTIC uses chi-square tests and normal distribution percentiles to construct confidence intervals.

Modifying Degrees of Freedom with the Number of Parameters

When you use Taylor series variance estimation (by default or when you specify VARMETHOD=TAYLOR in the MODEL statement), and you are fitting a model that has many parameters relative to the design degrees of freedom, it is appropriate to modify the design degrees of freedom by using the number of nonsingular parameters p in the model (Korn and Graubard (1999, section 5.2), Rao, Scott, and Skinner (1998)). You can specify DF=PARMADJ or DF=PARMADJ(*value*) in the MODEL statement to request this modification only for Taylor series variance estimation method; and this option does not apply to the replication variance estimation method.

Let f be the design degrees of freedom that is described in the section "Design Degrees of Freedom f for the Taylor Series Method" on page 9748. By default, or if you specify the DF=PARMADJ option, the *df* is modified as $df = f - p + 1$.

Testing Global Null Hypothesis: BETA=0

The *global null hypothesis* refers to the null hypothesis that all the explanatory effects can be eliminated and the model can contain only intercepts. By using the notations in the section "Logistic Regression Models" on page 9737, the global null hypothesis is defined as follows:

- If you have a cumulative model whose model parameters are $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$, where $\boldsymbol{\alpha}$ are the parameters for the intercepts and $\boldsymbol{\beta}$ are the parameters for the explanatory effects, then $H_0\colon \boldsymbol{\beta} = \mathbf{0}$. The number of restrictions r that are imposed on $\boldsymbol{\theta}$ is the number of parameters in slope parameter $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)'$: $r = k$.

- If you have a generalized logit model whose model parameters are $\boldsymbol{\theta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2, \ldots, \boldsymbol{\beta}'_D)'$ and $\boldsymbol{\beta}_d = (\beta_{d1}, \beta_{d2}, \ldots, \beta_{dk})'$ $(d = 1, 2, \ldots, D)$, then $H_0\colon (\beta_{d2}, \ldots, \beta_{dk})' = \mathbf{0}$ $(d = 1, 2, \ldots, D)$. The number of restrictions r that are imposed on $\boldsymbol{\theta}$ is the total number of slope parameters in $\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2, \ldots, \boldsymbol{\beta}'_D$: $r = (k-1) * D$.

PROC SURVEYLOGISTIC displays these tests in the "Testing Global Null Hypothesis: BETA=0" table.

Rao-Scott Likelihood Ratio Chi-Square Test

For complex survey design, you can use a design-adjusted Rao-Scott likelihood ratio chi-square test to test the global null hypothesis. For information about design-adjusted chi-square tests, see Lohr (2010, Section 10.3.2), Rao and Scott (1981), Rao and Scott (1984), Rao and Scott (1987), Thomas and Rao (1987), Rao and Thomas (1989), and Thomas, Singh, and Roberts (1996).

If you specify the CHISQ(NOADJUST) option, PROC SURVEYLOGISTIC computes the likelihood ratio chi-square test without the Rao-Scott design correction. If you specify the CHISQ(FIRSTORDER) option, PROC SURVEYLOGISTIC performs a first-order Rao-Scott likelihood ratio chi-square test. If you specify the CHISQ(SECONDORDER) option, PROC SURVEYLOGISTIC performs a second-order Rao-Scott likelihood ratio chi-square test.

If you do not specify the CHISQ option, the default test depends on the design and the model. By default, PROC SURVEYLOGISTIC performs a first-order or second-order Rao-Scott likelihood ratio chi-square (Satterthwaite) test if your design is not simple random sampling or when you provide replicate weights. Otherwise, if your design is simple random sampling and you do not provide replicate weights, PROC SURVEYLOGISTIC does not make any adjustment for the likelihood ratio test. In other words:

- If your design does not contain stratification nor clustering, and you do not provide replicate weights, then by default PROC SURVEYLOGISTIC performs a likelihood ratio chi-square test without any adjustment.
- If your design contains either stratification or clustering, or if you provide replicate weights, then by default PROC SURVEYLOGISTIC performs a likelihood ratio chi-square test with Rao-Scott adjustment. However, the default order of the adjustment depends on the number of model parameters excluding the intercepts.
  - If there is more than one nonintercept parameter in the model, the default is the second-order Rao-Scott likelihood ratio test.
  - If there is only one nonintercept parameter in the model, there is no need to compute the second-order adjustment. Therefore, the default is the first-order Rao-Scott likelihood ratio test.

Let $\hat{\boldsymbol{\theta}}$ be the estimated parameters, let $\hat{\boldsymbol{\theta}}_{H_0}$ be the estimated parameters under the global null hypothesis, and let r be the restrictions imposed on $\boldsymbol{\theta}$ under the global null hypothesis H_0. Let $L(\boldsymbol{\theta})$ be the log-likelihood function that is computed by using normalized weights.

Denote the estimated covariance matrix of $\hat{\boldsymbol{\theta}}$ under simple random sampling as $\hat{V}^{\text{srs}}(\hat{\boldsymbol{\theta}})$, and its partition corresponding to the r slope parameters as $\hat{V}^{\text{srs}}_{rr}(\hat{\boldsymbol{\theta}})$. Similarly, denote the estimated covariance matrix of $\hat{\boldsymbol{\theta}}$ under the sample design as $\hat{V}(\hat{\boldsymbol{\theta}})$, and its partition corresponding to the r slope parameters as $\hat{V}_{rr}(\hat{\boldsymbol{\theta}})$.

Define the *design effect matrix* E as

$$E = \hat{V}_{rr}(\hat{\boldsymbol{\theta}}) \left(\hat{V}^{\text{srs}}_{rr}(\hat{\boldsymbol{\theta}}) \right)^{-1}$$

Denote r^* as the rank of E and the positive eigenvalues of the design matrix E as $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_{r^*} > 0$.

Likelihood Ratio Chi-Square Test

Without the Rao-Scott design correction, the global null hypothesis is tested using either the chi-square statistics,

$$Q_{\chi^2} = 2\left[L(\hat{\boldsymbol{\theta}}) - L(\hat{\boldsymbol{\theta}}_{H_0})\right]$$

with r degrees of freedom, or an equivalent F statistics,

$$F = 2\left[L(\hat{\boldsymbol{\theta}}) - L(\hat{\boldsymbol{\theta}}_{H_0})\right]/r$$

with (r, ∞) degrees of freedom.

Rao-Scott First-Order Chi-Square Test

To address the impact of a complex survey design on the significance level of the likelihood ratio test, Rao and Scott (1984) proposed a first-order correction to the chi-square statistics as

$$Q_{RS1} = Q_{\chi^2}/\bar{\delta}.$$

where the first-order design correction,

$$\bar{\delta}. = \sum_{i=1}^{r^*} \delta_i / r^*$$

is the average of positive eigenvalues of the design effect matrix E.

Under the null hypothesis, the first-order Rao-Scott chi-square Q_{RS1} approximately follows a chi-square distribution with r^* degrees of freedom.

The corresponding F statistic is

$$F_{RS1} = Q_{RS1}/r^*$$

which has an F distribution with r^* and $df \cdot r^*$ degrees of freedom under the null hypothesis (Thomas and Rao 1984, 1987), and *df* is the design degrees of freedom as described in the section "Design Degrees of Freedom f for the Taylor Series Method" on page 9748.

Rao-Scott Second-Order Chi-Square Test

Rao and Scott (1987) further proposed the second-order (Satterthwaite) Rao-Scott chi-square statistic as

$$Q_{RS2} = Q_{RS1}/(1 + \hat{a}^2)$$

where Q_{RS1} is the first-order Rao-Scott chi-square statistic and the *second-order design correction* is computed from the coefficient of variation of the eigenvalues of the design effect matrix E as

$$\hat{a}^2 = \frac{1}{r^* - 1} \sum_{i=1}^{r^*} (\delta_i - \bar{\delta}.)^2 / \bar{\delta}.^2$$

Under the null hypothesis, the second-order Rao-Scott chi-square Q_{RS2} approximately follows a chi-square distribution with $r^*/(1 + \hat{a}^2)$ degrees of freedom.

The corresponding F statistic is

$$F_{RS2} = Q_{RS2}(1 + \hat{a}^2)/r^*$$

which has an F distribution with $r^*/(1 + \hat{a}^2)$ and $df \cdot r^*/(1 + \hat{a}^2)$ degrees of freedom under the null hypothesis (Thomas and Rao 1984, 1987), and df is the design degrees of freedom as described in the section "Design Degrees of Freedom f for the Taylor Series Method" on page 9748.

Score Statistics and Tests

To express the general form of the score statistic, let $\boldsymbol{\theta}$ be the parameter vector you want to estimate and let $\mathbf{g}(\boldsymbol{\theta})$ be the vector of first partial derivatives (gradient vector) of the log likelihood with respect to the parameter vector $\boldsymbol{\theta}$.

Consider a null hypothesis H_0 that has r restrictions imposed on $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta}$ under H_0, let $\mathbf{g}(\hat{\boldsymbol{\theta}})$ be the gradient vector evaluated at $\hat{\boldsymbol{\theta}}$, and let $\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})$ be the estimated covariance matrix for $\hat{\boldsymbol{\theta}}$, which is described in the section "Variance Estimation" on page 9739.

For the Taylor series variance estimation method, by default (or when DF=PARMADJ), PROC SURVEYLOGISTIC computes the score test statistic for the null hypothesis H_0 as

$$W_F = \left(\frac{f - r + 1}{f \ r}\right) \mathbf{g}(\hat{\boldsymbol{\theta}})' \left[\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})\right]^{-1} \mathbf{g}(\hat{\boldsymbol{\theta}})$$

where f is the design degrees of freedom as described in the section "Design Degrees of Freedom f for the Taylor Series Method" on page 9748.

If you specify DF=DESIGN option or if you use the replication variance estimation method, PROC SURVEYLOGISTIC computes the score test statistic for the null hypothesis H_0 as

$$W_F = \frac{1}{r} \mathbf{g}(\hat{\boldsymbol{\theta}})' \left[\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})\right]^{-1} \mathbf{g}(\hat{\boldsymbol{\theta}})$$

Under H_0, W_F has an F distribution with (r, df) degrees of freedom, where the denominator degrees of freedom df is described in the section "Degrees of Freedom" on page 9748.

As the denominator degrees of freedom grows, an F distribution approaches a chi-square distribution, and similarly a t distribution approaches a normal distribution. If you specify DF=INFINITY in the MODEL statement, the score test statistic for both Taylor series and replication methods for testing the null hypothesis H_0 can be expressed as

$$W_{\chi^2} = \mathbf{g}(\hat{\boldsymbol{\theta}})' \left[\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})\right]^{-1} \mathbf{g}(\hat{\boldsymbol{\theta}})$$

W_{χ^2} has a chi-square distribution with r degrees of freedom under the null hypothesis H_0.

Wald Confidence Intervals for Parameters

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

$$\hat{\theta}_j \pm z_{1-\alpha/2}\hat{\sigma}_j$$

where $z_{1-\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the standard normal distribution, $\hat{\theta}_j$ is the pseudo-estimate of θ_j, and $\hat{\sigma}_j$ is the standard error estimate of $\hat{\theta}_j$ in the section "Variance Estimation" on page 9739.

Testing Linear Hypotheses about the Regression Coefficients

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

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

where $\mathbf{L}$ is a matrix of coefficients for the linear hypotheses and $\mathbf{c}$ is a vector of constants whose rank is r. The vector of regression coefficients $\boldsymbol{\theta}$ includes both slope parameters and intercept parameters.

Let $\hat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta}$, and let $\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})$ be the estimated covariance matrix that is described in the section "Variance Estimation" on page 9739.

For the Taylor series variance estimation method, PROC SURVEYLOGISTIC computes the test statistic for the null hypothesis H_0 as

$$W_F = \left(\frac{f-r+1}{f\ r}\right)(\mathbf{L}\hat{\boldsymbol{\theta}} - \mathbf{c})'[\mathbf{L}\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})\mathbf{L}']^{-1}(\mathbf{L}\hat{\boldsymbol{\theta}} - \mathbf{c})$$

where f is the design degrees of freedom as described in the section "Design Degrees of Freedom f for the Taylor Series Method" on page 9748.

For the replication variance estimation method, PROC SURVEYLOGISTIC computes the test statistic for the null hypothesis H_0 as

$$W_F = \frac{1}{r}(\mathbf{L}\hat{\boldsymbol{\theta}} - \mathbf{c})'[\mathbf{L}\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})\mathbf{L}']^{-1}(\mathbf{L}\hat{\boldsymbol{\theta}} - \mathbf{c})$$

Under the H_0, W_F has an F distribution with (r, df) degrees of freedom, and the denominator degrees of freedom df is described in the section "Degrees of Freedom" on page 9748.

As the denominator degrees of freedom grows, an F distribution approaches a chi-square distribution, and similarly a t distribution approaches a normal distribution. If you specify DF=INFINITY in the MODEL statement, PROC SURVEYLOGISTIC computes the test statistic for both Taylor series and replication methods for testing the null hypothesis H_0 as

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

Under H_0, χ^2_W has an asymptotic chi-square distribution with r degrees of freedom.

Type 3 Tests

For models that use less-than-full-rank parameterization (as specified by the PARAM=GLM option in the CLASS statement), a Type 3 test of an effect of interest (main effect or interaction) is a test of the Type III estimable functions that are defined for that effect. When the model contains no missing cells, performing the Type 3 test of a main effect corresponds to testing the hypothesis of equal marginal means. For more information about Type III estimable functions, see Chapter 50, "The GLM Procedure," and Chapter 15, "The Four Types of Estimable Functions." Also see Littell, Freund, and Spector (1991).

For models that use full-rank parameterization, all parameters are estimable when there are no missing cells, so it is unnecessary to define estimable functions. The standard test of an effect of interest in this case is the joint test that the values of the parameters associated with that effect are zero. For a model that uses effects parameterization (as specified by the PARAM=EFFECT option in the CLASS statement), performing the joint test for a main effect is equivalent to testing the equality of marginal means. For a model that uses reference parameterization (as specified by the PARAM=REF option in the CLASS statement), performing the joint test is equivalent to testing the equality of cell means at the reference level of the other model effects. For more information about the coding scheme and the associated interpretation of results, see Muller and Fetterman (2002, Chapter 14).

If there is no interaction term, the Type 3 test of an effect for a model that uses GLM parameterization is the same as the joint test of the effect for the model that uses full-rank parameterization. In this situation, the joint test is also called the Type 3 test. For a model that contains an interaction term and no missing cells, the Type 3 test of a component main effect under GLM parameterization is the same as the joint test of the component main effect under effect parameterization. Both test the equality of cell means. But this Type 3 test differs from the joint test under reference parameterization, which tests the equality of cell means at the reference level of the other component main effect. If some cells are missing, you can obtain meaningful tests only by testing a Type III estimable function, so in this case you should use GLM parameterization.

The results of a Type 3 test or a joint test do not depend on the order in which you specify the terms in the MODEL statement.

Odds Ratio Estimation

Consider a dichotomous response variable with outcomes *event* and *nonevent*. Let a dichotomous risk factor variable X take 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 might be too small to be considered important, while a change of 10 pounds might be more meaningful. The odds ratio for a change in X from a to b is estimated by raising the odds ratio estimate for a unit change in X to the power of $c = b - a$, as shown previously.

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

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

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

The log odds for Black is

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

The log odds for White is

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

Therefore, the log odds ratio of Black versus White becomes

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

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

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

The log odds ratio of Black versus White is given by

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

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

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

The log odds ratio of Black versus White is

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

Consider the hypothetical example of heart disease among race in Hosmer and Lemeshow (2000, p. 51). 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 shown in Table 117.9.

Table 117.9 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 Estimates**
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 SURVEYLOGISTIC, the "Odds Ratio Estimates" table contains the odds ratio estimates and the corresponding t or Wald confidence intervals computed by using the covariance matrix in the section "Variance Estimation" on page 9739. 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 in the MODEL statement to request confidence intervals for the odds ratios.

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

Rank Correlation of Observed Responses and Predicted Probabilities

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

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

Let N be the sum of observation frequencies in the data. Suppose there are 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 SURVEYLOGISTIC computes the following four indices of rank correlation for assessing the predictive ability of a model:

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

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

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

Linear Predictor, Predicted Probability, and Confidence Limits

This section describes how predicted probabilities and confidence limits are calculated by using the pseudo-estimates (MLEs) obtained from PROC SURVEYLOGISTIC. For a specific example, see the section "Getting Started: SURVEYLOGISTIC Procedure" on page 9682. Predicted probabilities and confidence limits can be output to a data set with the OUTPUT statement.

Let $\Delta_{\alpha/2}$ is the $100(1-\alpha/2)$th percentile point of a standard normal distribution or a t distribution according to the DF= specification:

$$\Delta_{\alpha/2} = \begin{cases} 100(1-\alpha/2)\text{th percentile point of a standard normal distribution } z_{\alpha/2} & \text{if DF=INFINITY} \\ 100(1-\alpha/2)\text{th percentile point of a } t \text{ distribution } t_{\alpha/2} & \text{otherwise} \end{cases}$$

Cumulative Response Models

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

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

is estimated by

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

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

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

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**	$100(1-\alpha)$ **Confidence Limits**
LOGIT	$1/(1+e^{-\hat{\eta}_i})$	$1/(1+e^{-\hat{\eta}_i \pm \Delta_{\alpha/2}\hat{\sigma}(\hat{\eta}_i)})$
PROBIT	$\Phi(\hat{\eta}_i)$	$\Phi(\hat{\eta}_i \pm \Delta_{\alpha/2}\hat{\sigma}(\hat{\eta}_i))$
CLOGLOG	$1-e^{-e^{\hat{\eta}_i}}$	$1-e^{-e^{\hat{\eta}_i \pm \Delta_{\alpha/2}\hat{\sigma}(\hat{\eta}_i)}}$

Generalized Logit Model

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

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

By the *delta method*,

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

A 100(1−α)% confidence level for π_i is given by

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

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

Output Data Sets

You can use the Output Delivery System (ODS) to create a SAS data set from any piece of PROC SURVEYLOGISTIC output. See the section "ODS Table Names" on page 9765 for more information. For a more detailed description of using ODS, see Chapter 20, "Using the Output Delivery System."

PROC SURVEYLOGISTIC also provides an OUTPUT statement to create a data set that contains estimated linear predictors, the estimates of the cumulative or individual response probabilities, and their confidence limits.

If you use replication variance estimation, PROC SURVEYLOGISTIC provides an output data set that stores the replicate weights and an output data set that stores the jackknife coefficients for jackknife variance estimation.

OUT= Data Set in the OUTPUT Statement

The OUT= data set in the OUTPUT statement contains all the variables in the input data set along with statistics you request by using *keyword=name* options or the PREDPROBS= option in the OUTPUT statement. In addition, if you use the single-trial syntax and you request any of the XBETA=, STDXBETA=, PREDICTED=, LCL=, and UCL= options, the OUT= data set contains the automatic variable _LEVEL_. The value of _LEVEL_ identifies the response category upon which the computed values of XBETA=, STDXBETA=, PREDICTED=, LCL=, and UCL= are based.

When there are more than two response levels, only variables named by the XBETA=, STDXBETA=, PREDICTED=, LOWER=, and UPPER= options and the variables given by PREDPROBS=(INDIVIDUAL CUMULATIVE) have their values computed; the other variables have missing values. If you fit a generalized logit model, the cumulative predicted probabilities are not computed.

When there are only two response categories, each input observation produces one observation in the OUT= data set.

If there are more than two response categories and you specify only the PREDPROBS= option, then each input observation produces one observation in the OUT= data set. However, if you fit an ordinal (cumulative) model and specify options other than the PREDPROBS= options, each input observation generates as many output observations as one fewer than the number of response levels, and the predicted probabilities and their confidence limits correspond to the cumulative predicted probabilities. If you fit a generalized logit model and specify options other than the PREDPROBS= options, each input observation generates as many output observations as the number of response categories; the predicted probabilities and their confidence limits correspond to the probabilities of individual response categories.

For observations in which only the response variable is missing, values of the XBETA=, STDXBETA=, PREDICTED=, UPPER=, LOWER=, and PREDPROBS= options are computed even though these observations do not affect the model fit. This enables, for instance, predicted probabilities to be computed for new observations.

Replicate Weights Output Data Set

If you specify the OUTWEIGHTS= *method-option* for VARMETHOD=BRR, VARMETHOD=JACKKNIFE, or VARMETHOD=BOOTSTRAP, PROC SURVEYLOGISTIC stores the replicate weights in an output data set. The OUTWEIGHTS= output data set contains all observations from the DATA= input data set that are valid (used in the analysis). (A valid observation is an observation that has a positive value of the WEIGHT variable. Valid observations must also have nonmissing values of the STRATA and CLUSTER variables, unless you specify the MISSING option.)

The OUTWEIGHTS= data set contains all variables in the DATA= input data set and the replicate weight variables, RepWt_1, RepWt_2, . . ., RepWt_n, where *n* is the total number of replicates in the analysis. Each replicate weight variable contains the replicate weights for the corresponding replicate. Replicate weights equal 0 for observations that are not included in the replicate.

After the procedure creates replicate weights for a particular input data set and survey design, you can use the OUTWEIGHTS= *method-option* to store these replicate weights and then use them again in subsequent analyses, either in PROC SURVEYLOGISTIC or in other survey procedures. You can use the REPWEIGHTS statement to provide replicate weights for the procedure.

Jackknife Coefficients Output Data Set

If you specify the OUTJKCOEFS= *method-option* for VARMETHOD=JACKKNIFE, PROC SURVEYLOGISTIC stores the jackknife coefficients in an output data set. The OUTJKCOEFS= output data set contains one observation for each replicate. The OUTJKCOEFS= data set contains the following variables:

- Replicate, which is the replicate number for the jackknife coefficient
- JKCoefficient, which is the jackknife coefficient
- DonorStratum, which is the stratum of the PSU that was deleted to construct the replicate, if you specify a STRATA statement

After the procedure creates jackknife coefficients for a particular input data set and survey design, you can use the OUTJKCOEFS= *method-option* to store these coefficients and then use them again in subsequent analyses, either in PROC SURVEYLOGISTIC or in the other survey procedures. You can use the JKCOEFS= option in the REPWEIGHTS statement to provide jackknife coefficients for the procedure.

Displayed Output

The SURVEYLOGISTIC procedure produces output that is described in the following sections.

Output that is generated by the EFFECT, ESTIMATE, LSMEANS, LSMESTIMATE, and SLICE statements are not listed below. For information about the output that is generated by these statements, see the corresponding sections of Chapter 19, “Shared Concepts and Topics.”

Model Information

By default, PROC SURVEYLOGISTIC displays the following information in the “Model Information” table:

- name of the input Data Set
- name and label of the Response Variable if the single-trial syntax is used
- number of Response Levels

- name of the Events Variable if the events/trials syntax is used
- name of the Trials Variable if the events/trials syntax is used
- name of the Offset Variable if the OFFSET= option is specified
- name of the Frequency Variable if the FREQ statement is specified
- name(s) of the Stratum Variable(s) if the STRATA statement is specified
- total Number of Strata if the STRATA statement is specified
- name(s) of the Cluster Variable(s) if the CLUSTER statement is specified
- total Number of Clusters if the CLUSTER statement is specified
- name of the Weight Variable if the WEIGHT statement is specified
- Variance Adjustment method
- Upper Bound ADJBOUND parameter used in the VADJUST=MOREL(ADJBOUND=) option
- Lower Bound DEFFBOUND parameter used in the VADJUST=MOREL(DEFFBOUND=) option
- whether *fpc* (finite population correction) is used

Variance Estimation

By default, PROC SURVEYLOGISTIC displays the following variance estimation information in the "Variance Estimation" table:

- Method, which is the variance estimation method
- Variance Adjustment method
- Initial seed for the random number that the bootstrap method uses
- Upper Bound ADJBOUND parameter specified in the VADJUST=MOREL(ADJBOUND=) option
- Lower Bound DEFFBOUND parameter specified in the VADJUST=MOREL(DEFFBOUND=) option
- whether *fpc* (finite population correction) is used
- Number of Replicates
- Number of Replicates Used, which is the number of replicates that are used in variance estimation (the rest are excluded because of nonconvergence)
- Hadamard Data Set name, if you specify the VARMETHOD=BRR(HADAMARD=) *method-option*
- Fay Coefficient, if you specify the VARMETHOD=BRR(FAY) *method-option*
- Replicate Weights input data set name, if you use a REPWEIGHTS statement
- whether Missing Levels are created for categorical variables by the MISSING option
- whether observations with Missing Values are included in the analysis by the NOMCAR option

Data Summary

By default, PROC SURVEYLOGISTIC displays the following information for the entire data set:

- Number of Observations read from the input data set
- Number of Observations used in the analysis

If there is a DOMAIN statement, PROC SURVEYLOGISTIC also displays the following:

- Number of Observations in the current domain
- Number of Observations not in the current domain

If there is a FREQ statement, PROC SURVEYLOGISTIC also displays the following:

- Sum of Frequencies of all the observations read from the input data set
- Sum of Frequencies of all the observations used in the analysis

If there is a WEIGHT statement, PROC SURVEYLOGISTIC also displays the following:

- Sum of Weights of all the observations read from the input data set
- Sum of Weights of all the observations used in the analysis
- Sum of Weights of all the observations in the current domain, if DOMAIN statement is also specified.

Response Profile

By default, PROC SURVEYLOGISTIC displays a "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.

Class Level Information

If you use a CLASS statement to name classification variables, PROC SURVEYLOGISTIC displays a "Class Level Information" table. This table contains the following information for each classification variable:

- Class, which lists each CLASS variable name
- Value, which lists the values of the classification variable. The values are separated by a white space character; therefore, to avoid confusion, you should not include a white space character within a classification variable value.
- Design Variables, which lists the parameterization used for the classification variables

Stratum Information

When you specify the LIST option in the STRATA statement, PROC SURVEYLOGISTIC displays a "Stratum Information" table, which provides the following information for each stratum:

- Stratum Index, which is a sequential stratum identification number
- STRATA variable(s), which lists the levels of STRATA variables for the stratum
- Population Total, if you specify the TOTAL= option
- Sampling Rate, if you specify the TOTAL= or RATE= option. If you specify the TOTAL= option, the sampling rate is based on the number of nonmissing observations in the stratum.
- N Obs, which is the number of observations
- number of Clusters, if you specify a CLUSTER statement

Maximum Likelihood Iteration History

The “Maximum Likelihood Iterative Phase” table gives the iteration number, the step size (in the scale of 1.0, 0.5, 0.25, 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. You need to use the ITPRINT option in the MODEL statement to obtain this table.

Model Fit Statistics

By default, PROC SURVEYLOGISTIC displays the following information in the “Model Fit Statistics” table:

- “Model Fit Statistics” and “Testing Global Null Hypothesis: BETA=0” tables, which give the various criteria (–2 Log L, AIC, SC) based on the likelihood for fitting a model with intercepts only and for fitting a model with intercepts and explanatory variables. If you specify the NOINT option, these statistics are calculated without considering the intercept parameters. The third column of the table gives the chi-square statistics and p-values for the –2 Log L statistic and for the Score statistic. These test the joint effect of the explanatory variables included in the model. The Score criterion is always missing for the models identified by the first two columns of the table. Note also that the first two rows of the Chi-Square column are always missing, since tests cannot be performed for AIC and SC.
- generalized R^2 measures for the fitted model if you specify the RSQUARE option in the MODEL statement

Type III Analysis of Effects

PROC SURVEYLOGISTIC displays the “Type III Analysis of Effects” table if the model contains an effect that involves a CLASS variable. This table gives the degrees of freedom for F statistics or Wald chi-square statistic, and the p-value for each effect in the model.

Analysis of Maximum Likelihood Estimates

By default, PROC SURVEYLOGISTIC displays the following information in the “Analysis of Maximum Likelihood Estimates” table:

- maximum likelihood estimate of the parameter
- estimated standard error of the parameter estimate, computed as the square root of the corresponding diagonal element of the estimated covariance matrix
- t value, which is the t statistic for testing H_0: Parameter $= 0$
- Pr $>$ | t |, which is the two-sided p-value for the t test
- $100(1-\alpha)\%$ confidence intervals for estimated parameters. You need to specify the CLPARM option in the MODEL statement to display these estimates.
- 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

  $$s = \begin{cases} \pi/\sqrt{3} & \text{logistic} \\ 1 & \text{normal} \\ \pi/\sqrt{6} & \text{extreme-value} \end{cases}$$

  You need to specify the STB option in the MODEL statement to obtain these estimates. Standardized estimates of the intercept parameters are set to missing.
- value of ($e^{\hat{\beta}_i}$) for each slope parameter β_i if you specify the EXPB option in the MODEL statement. For continuous variables, this is equivalent to the estimated odds ratio for a one-unit change.
- label of the variable (if space permits) if you specify the PARMLABEL option in the MODEL statement. Because of constraints on the line size, the variable label might be suppressed in order to display the table in one panel. Use the SAS system option LINESIZE= to specify a larger line size to accommodate variable labels. A shorter line size can break the table into two panels, allowing labels to be displayed.

Odds Ratio Estimates

The “Odds Ratio Estimates” table displays the odds ratio estimates and the corresponding 95% confidence intervals. For continuous explanatory variables, these odds ratios correspond to a unit increase in the risk factors.

Association of Predicted Probabilities and Observed Responses

The “Association of Predicted Probabilities and Observed Responses” table displays 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.

Estimated Covariance Matrix

PROC SURVEYLOGISTIC displays the following information in the “Estimated Covariance Matrix” table:

- estimated covariance matrix of the parameter estimates if you use the COVB option in the MODEL statement
- estimated correlation matrix of the parameter estimates if you use the CORRB option in the MODEL statement

Linear Hypotheses Testing Results

The “Linear Hypothesis Testing” table gives the result of the *F* test or Wald test for each TEST statement (if specified).

Hadamard Matrix

If you specify the VARMETHOD=BRR(PRINTH) *method-option* in the PROC SURVEYLOGISTIC statement, the procedure displays the Hadamard matrix.

When you provide a Hadamard matrix with the VARMETHOD=BRR(HADAMARD=) *method-option*, the procedure displays only used rows and columns of the Hadamard matrix.

ODS Table Names

PROC SURVEYLOGISTIC assigns a name to each table it creates; these names are listed in Table 117.10. You can use these names to refer the table when using the Output Delivery System (ODS) to select tables and create output data sets. The EFFECT, ESTIMATE, LSMEANS, LSMESTIMATE, and SLICE statements also create tables, which are not listed in Table 117.10. For information about these tables, see the corresponding sections of Chapter 19, “Shared Concepts and Topics.”

For more information about ODS, see Chapter 20, “Using the Output Delivery System.”

Table 117.10 ODS Tables Produced by PROC SURVEYLOGISTIC

ODS Table Name	Description	Statement	Option
Association	Association of predicted probabilities and observed responses	MODEL	Default
ClassLevelInfo	CLASS variable levels and design variables	MODEL	Default (with CLASS variables)
CLOdds	Confidence intervals for odds ratios	MODEL	CLODDS
CLparm	Confidence intervals for parameters	MODEL	CLPARM
ContrastCoeff	**L** matrix from CONTRAST	CONTRAST	E
ContrastEstimate	Estimates from CONTRAST	CONTRAST	ESTIMATE=
ContrastTest	Wald test for CONTRAST	CONTRAST	Default

Table 117.10 *continued*

ODS Table Name	Description	Statement	Option
ConvergenceStatus	Convergence status	MODEL	Default
CorrB	Estimated correlation matrix of parameter estimators	MODEL	CORRB
CovB	Estimated covariance matrix of parameter estimators (Ordinal response)	MODEL	COVB
DomainSummary	Domain summary	DOMAIN	Default
FitStatistics	Model fit statistics	MODEL	Default
GlobalTests	Test for global null hypothesis	MODEL	Default
Gradient	Gradient evaluated at global null hypothesis	MODEL	GRADIENT
HadamardMatrix	Hadamard matrix	PROC	PRINTH
IterHistory	Iteration history	MODEL	ITPRINT
LastGradient	Last evaluation of gradient	MODEL	ITPRINT
Linear	Linear combination	PROC	Default
LogLikeChange	Final change in the log likelihood	MODEL	ITPRINT
ModelANOVA	Joint or Type 3 tests of effects	MODEL	Default (with CLASS variables)
ModelInfo	Model information	PROC	Default
NObs	Number of observations	PROC	Default
OddsEst	Adjusted odds ratios	UNITS	Default
OddsRatios	Odds ratios	MODEL	Default
ParameterEstimates	Maximum likelihood estimates of model parameters	MODEL	Default
RSquare	R-square	MODEL	RSQUARE
ResponseProfile	Response profile	PROC	Default
StrataInfo	Stratum information	STRATA	LIST
TestPrint1	$\mathbf{L}[\mathrm{cov}(\mathbf{b})]\mathbf{L}'$ and $\mathbf{Lb}-\mathbf{c}$	TEST	PRINT
TestPrint2	$\mathrm{Ginv}(\mathbf{L}[\mathrm{cov}(\mathbf{b})]\mathbf{L}')$ and $\mathrm{Ginv}(\mathbf{L}[\mathrm{cov}(\mathbf{b})]\mathbf{L}')(\mathbf{Lb}-\mathbf{c})$	TEST	PRINT
TestStmts	Linear hypotheses testing results	TEST	Default
VarianceEstimation	Variance estimation	PROC	Default

ODS Graphics

Statistical procedures use ODS Graphics to create graphs as part of their output. ODS Graphics is described in detail in Chapter 21, "Statistical Graphics Using ODS."

Before you create graphs, ODS Graphics must be enabled (for example, by specifying the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" on page 623 in Chapter 21, "Statistical Graphics Using ODS."

The overall appearance of graphs is controlled by ODS styles. Styles and other aspects of using ODS Graphics are discussed in the section "A Primer on ODS Statistical Graphics" on page 622 in Chapter 21, "Statistical Graphics Using ODS."

When ODS Graphics is enabled, then the ESTIMATE, LSMEANS, LSMESTIMATE, and SLICE statements can produce plots that are associated with their analyses. For information about these plots, see the corresponding sections of Chapter 19, "Shared Concepts and Topics."

Examples: SURVEYLOGISTIC Procedure

Example 117.1: Stratified Cluster Sampling

A market research firm conducts a survey among undergraduate students at a certain university to evaluate three new Web designs for a commercial Web site targeting undergraduate students at the university.

The sample design is a stratified sample where the strata are students' classes. Within each class, 300 students are randomly selected by using simple random sampling without replacement. The total number of students in each class in the fall semester of 2001 is shown in the following table:

Class	Enrollment
1 - Freshman	3,734
2 - Sophomore	3,565
3 - Junior	3,903
4 - Senior	4,196

This total enrollment information is saved in the SAS data set Enrollment by using the following SAS statements:

```
proc format;
   value Class
      1='Freshman' 2='Sophomore'
      3='Junior'   4='Senior';
run;

data Enrollment;
   format Class Class.;
   input Class _TOTAL_;
   datalines;
1 3734
2 3565
3 3903
4 4196
;
```

In the data set Enrollment, the variable _TOTAL_ contains the enrollment figures for all classes. They are also the population size for each stratum in this example.

Each student selected in the sample evaluates one randomly selected Web design by using the following scale:

1	Dislike very much
2	Dislike
3	Neutral
4	Like
5	Like very much

The survey results are collected and shown in the following table, with the three different Web designs coded as A, B, and C.

Evaluation of New Web Designs

		Rating Counts				
Strata	**Design**	**1**	**2**	**3**	**4**	**5**
Freshman	A	10	34	35	16	15
	B	5	6	24	30	25
	C	11	14	20	34	21
Sophomore	A	19	12	26	18	25
	B	10	18	32	23	26
	C	15	22	34	9	20
Junior	A	8	21	23	26	22
	B	1	4	15	33	47
	C	16	19	30	23	12
Senior	A	11	14	24	33	18
	B	8	15	25	30	22
	C	2	34	30	18	16

The survey results are stored in a SAS data set WebSurvey by using the following SAS statements:

```
proc format;
   value Design 1='A' 2='B' 3='C';
   value Rating
      1='dislike very much'
      2='dislike'
      3='neutral'
      4='like'
      5='like very much';
run;

data WebSurvey;
   format Class Class. Design Design. Rating Rating.;
   do Class=1 to 4;
      do Design=1 to 3;
         do Rating=1 to 5;
            input Count @@;
            output;
         end;
      end;
   end;
   datalines;
10 34 35 16 15   8 21 23 26 22   5 10 24 30 21
```

```
 1 14 25 23 37  11 14 20 34 21  16 19 30 23 12
19 12 26 18 25  11 14 24 33 18  10 18 32 23 17
 8 15 35 30 12  15 22 34  9 20   2 34 30 18 16
;

data WebSurvey;
   set WebSurvey;
   if Class=1 then Weight=3734/300;
   if Class=2 then Weight=3565/300;
   if Class=3 then Weight=3903/300;
   if Class=4 then Weight=4196/300;
run;
```

The data set WebSurvey contains the variables Class, Design, Rating, Count, and Weight. The variable class is the stratum variable, with four strata: freshman, sophomore, junior, and senior. The variable Design specifies the three new Web designs: A, B, and C. The variable Rating contains students' evaluations of the new Web designs. The variable counts gives the frequency with which each Web design received each rating within each stratum. The variable weight contains the sampling weights, which are the reciprocals of selection probabilities in this example.

Output 117.1.1 shows the first 20 observations of the data set.

Output 117.1.1 Web Design Survey Sample (First 20 Observations)

Obs	Class	Design	Rating	Count	Weight
1	Freshman	A	dislike very much	10	12.4467
2	Freshman	A	dislike	34	12.4467
3	Freshman	A	neutral	35	12.4467
4	Freshman	A	like	16	12.4467
5	Freshman	A	like very much	15	12.4467
6	Freshman	B	dislike very much	8	12.4467
7	Freshman	B	dislike	21	12.4467
8	Freshman	B	neutral	23	12.4467
9	Freshman	B	like	26	12.4467
10	Freshman	B	like very much	22	12.4467
11	Freshman	C	dislike very much	5	12.4467
12	Freshman	C	dislike	10	12.4467
13	Freshman	C	neutral	24	12.4467
14	Freshman	C	like	30	12.4467
15	Freshman	C	like very much	21	12.4467
16	Sophomore	A	dislike very much	1	11.8833
17	Sophomore	A	dislike	14	11.8833
18	Sophomore	A	neutral	25	11.8833
19	Sophomore	A	like	23	11.8833
20	Sophomore	A	like very much	37	11.8833

The following SAS statements perform the logistic regression:

```
proc surveylogistic data=WebSurvey total=Enrollment;
   stratum Class;
   freq Count;
   class Design;
```

```
      model Rating (order=internal) = design;
      weight Weight;
   run;
```

The PROC SURVEYLOGISTIC statement invokes the procedure. The TOTAL= option specifies the data set Enrollment, which contains the population totals in the strata. The population totals are used to calculate the finite population correction factor in the variance estimates. The response variable Rating is in the ordinal scale. A cumulative logit model is used to investigate the responses to the Web designs. In the MODEL statement, rating is the response variable, and Design is the effect in the regression model. The ORDER=INTERNAL option is used for the response variable Rating to sort the ordinal response levels of Rating by its internal (numerical) values rather than by the formatted values (for example, 'like very much'). Because the sample design involves stratified simple random sampling, the STRATA statement is used to specify the stratification variable Class. The WEIGHT statement specifies the variable Weight for sampling weights.

The sample and analysis summary is shown in Output 117.1.2. There are five response levels for the Rating, with 'dislike very much' as the lowest ordered value. The regression model is modeling lower cumulative probabilities by using logit as the link function. Because the TOTAL= option is used, the finite population correction is included in the variance estimation. The sampling weight is also used in the analysis.

Output 117.1.2 Web Design Survey, Model Information

The SURVEYLOGISTIC Procedure

Model Information	
Data Set	WORK.WEBSURVEY
Response Variable	Rating
Number of Response Levels	5
Frequency Variable	Count
Stratum Variable	Class
Number of Strata	4
Weight Variable	Weight
Model	Cumulative Logit
Optimization Technique	Fisher's Scoring
Variance Adjustment	Degrees of Freedom (DF)
Finite Population Correction	Used

Response Profile

Ordered Value	Rating	Total Frequency	Total Weight
1	dislike very much	116	1489.0733
2	dislike	227	2933.0433
3	neutral	338	4363.3767
4	like	283	3606.8067
5	like very much	236	3005.7000

Probabilities modeled are cumulated over the lower Ordered Values.

An alternative model is to use the generalized logit model with the LINK=GLOGIT option, as shown in the following SAS statements:

```
proc surveylogistic data=WebSurvey total=Enrollment;
   stratum Class;
   freq Count;
   class Design;
   model Rating (ref='neutral') = Design /link=glogit;
   weight Weight;
run;
```

The REF='neutral' option is used for the response variable Rating to indicate that all other response levels are referenced to the level 'neutral.' The option LINK=GLOGIT option requests that the procedure fit a generalized logit model.

The summary of the analysis is shown in Output 117.1.3, which indicates that the generalized logit model is used in the analysis.

Output 117.1.3 Web Design Survey, Model Information

The SURVEYLOGISTIC Procedure

Model Information	
Data Set	WORK.WEBSURVEY
Response Variable	Rating
Number of Response Levels	5
Frequency Variable	Count
Stratum Variable	Class
Number of Strata	4
Weight Variable	Weight
Model	Generalized Logit
Optimization Technique	Newton-Raphson
Variance Adjustment	Degrees of Freedom (DF)
Finite Population Correction	Used

Response Profile

Ordered Value	Rating	Total Frequency	Total Weight
1	dislike	227	2933.0433
2	dislike very much	116	1489.0733
3	like	283	3606.8067
4	like very much	236	3005.7000
5	neutral	338	4363.3767

Logits modeled use Rating='neutral' as the reference category.

Output 117.1.4 shows the parameterization for the main effect Design.

Output 117.1.4 Web Design Survey, Class Level Information

Class Level Information

Class	Value	Design Variables	
Design	**A**	1	0
	B	0	1
	C	-1	-1

The parameter and odds ratio estimates are shown in Output 117.1.5. For each odds ratio estimate, the 95% confidence intervals shown in the table contain the value 1.0. Therefore, no conclusion about which Web design is preferred can be made based on this survey.

Output 117.1.5 Web Design Survey, Parameter and Odds Ratio Estimates

Analysis of Maximum Likelihood Estimates

Parameter		Rating	Estimate	Standard Error	t Value	Pr > |t|
Intercept		dislike	-0.3964	0.0832	-4.77	<.0001
Intercept		dislike very much	-1.0826	0.1045	-10.36	<.0001
Intercept		like	-0.1892	0.0780	-2.43	0.0154
Intercept		like very much	-0.3767	0.0824	-4.57	<.0001
Design	A	dislike	-0.0942	0.1166	-0.81	0.4196
Design	A	dislike very much	-0.0647	0.1469	-0.44	0.6597
Design	A	like	-0.1370	0.1104	-1.24	0.2149
Design	A	like very much	0.0446	0.1130	0.39	0.6934
Design	B	dislike	0.0391	0.1201	0.33	0.7451
Design	B	dislike very much	0.2721	0.1448	1.88	0.0605
Design	B	like	0.1669	0.1102	1.52	0.1300
Design	B	like very much	0.1420	0.1174	1.21	0.2265

NOTE: The degrees of freedom for the t tests is 1196.

Odds Ratio Estimates

Effect	Rating	Point Estimate	95% Confidence Limits	
Design A vs C	dislike	0.861	0.583	1.272
Design A vs C	dislike very much	1.153	0.691	1.924
Design A vs C	like	0.899	0.618	1.306
Design A vs C	like very much	1.260	0.851	1.866
Design B vs C	dislike	0.984	0.658	1.471
Design B vs C	dislike very much	1.615	0.975	2.677
Design B vs C	like	1.218	0.838	1.769
Design B vs C	like very much	1.389	0.924	2.087

NOTE:
The degrees of freedom in computing the confidence limits is 1196.

Example 117.2: The Medical Expenditure Panel Survey (MEPS)

The U.S. Department of Health and Human Services conducts the Medical Expenditure Panel Survey (MEPS) to produce national and regional estimates of various aspects of health care. The MEPS has a complex sample design that includes both stratification and clustering. The sampling weights are adjusted for nonresponse and raked with respect to population control totals from the Current Population Survey. See the MEPS (2006) and Machlin, Yu, and Zodet (2005) for details.

In this example, the 1999 full-year consolidated data file HC-038 (MEPS HC-038, 2002) from the MEPS is used to investigate the relationship between medical insurance coverage and the demographic variables. The data can be downloaded directly from the Agency for Healthcare Research and Quality (AHRQ)

Web site at https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-038 in either ASCII format or SAS transport format. The Web site includes a detailed description of the data as well as the SAS program used to access and format it.

For this example, the SAS transport format data file for HC-038 is downloaded to ‘C:H38.ssp’ on a Windows-based PC. The instructions on the Web site lead to the following SAS statements for creating a SAS data set MEPS, which contains only the sample design variables and other variables necessary for this analysis.

```
proc format;
   value racex
      -9 = 'NOT ASCERTAINED'
      -8 = 'DK'
      -7 = 'REFUSED'
      -1 = 'INAPPLICABLE'
      1 = 'AMERICAN INDIAN'
      2 = 'ALEUT, ESKIMO'
      3 = 'ASIAN OR PACIFIC ISLANDER'
      4 = 'BLACK'
      5 = 'WHITE'
      91 = 'OTHER'
      ;
   value sex
      -9 = 'NOT ASCERTAINED'
      -8 = 'DK'
      -7 = 'REFUSED'
      -1 = 'INAPPLICABLE'
      1 = 'MALE'
      2 = 'FEMALE'
      ;
   value povcat9h
      1 = 'NEGATIVE OR POOR'
      2 = 'NEAR POOR'
      3 = 'LOW INCOME'
      4 = 'MIDDLE INCOME'
      5 = 'HIGH INCOME'
      ;
   value inscov9f
      1 = 'ANY PRIVATE'
      2 = 'PUBLIC ONLY'
      3 = 'UNINSURED'
      ;
run;

libname mylib '';
filename in1 'H38.SSP';
proc xcopy in=in1 out=mylib import;
run;

data meps;
   set mylib.H38;
   label racex= sex= inscov99= povcat99=
      varstr99= varpsu99= perwt99f= totexp99=;
   format racex racex. sex sex.
      povcat99 povcat9h. inscov99 inscov9f.;
   keep inscov99 sex racex povcat99 varstr99
      varpsu99 perwt99f totexp99;
run;
```

There are a total of 24,618 observations in this SAS data set. Each observation corresponds to a person in the survey. The stratification variable is VARSTR99, which identifies the 143 strata in the sample. The variable VARPSU99 identifies the 460 PSUs in the sample. The sampling weights are stored in the variable PERWT99F. The response variable is the health insurance coverage indicator variable, INSCOV99, which has three values:

1	The person had any private insurance coverage any time during 1999
2	The person had only public insurance coverage during 1999
3	The person was uninsured during all of 1999

The demographic variables include gender (SEX), race (RACEX), and family income level as a percent of the poverty line (POVCAT99). The variable RACEX has five categories:

1	American Indian
2	Aleut, Eskimo
3	Asian or Pacific Islander
4	Black
5	White

The variable POVCAT99 is constructed by dividing family income by the applicable poverty line (based on family size and composition), with the resulting percentages grouped into five categories:

1	Negative or poor (less than 100%)
2	Near poor (100% to less than 125%)
3	Low income (125% to less than 200%)
4	Middle income (200% to less than 400%)
5	High income (greater than or equal to 400%)

The data set also contains the total health care expenditure in 1999, TOTEXP99, which is used as a covariate in the analysis.

Output 117.2.1 displays the first 30 observations of this data set.

Output 117.2.1 1999 Full-Year MEPS (First 30 Observations)

Obs	SEX	RACEX	POVCAT99	INSCOV99	TOTEXP99	PERWT99F	VARSTR99	VARPSU99
1	MALE	WHITE	MIDDLE INCOME	PUBLIC ONLY	2735	14137.86	131	2
2	FEMALE	WHITE	MIDDLE INCOME	ANY PRIVATE	6687	17050.99	131	2
3	MALE	WHITE	MIDDLE INCOME	ANY PRIVATE	60	35737.55	131	2
4	MALE	WHITE	MIDDLE INCOME	ANY PRIVATE	60	35862.67	131	2
5	FEMALE	WHITE	MIDDLE INCOME	ANY PRIVATE	786	19407.11	131	2
6	MALE	WHITE	MIDDLE INCOME	ANY PRIVATE	345	18499.83	131	2
7	MALE	WHITE	MIDDLE INCOME	ANY PRIVATE	680	18499.83	131	2
8	MALE	WHITE	MIDDLE INCOME	ANY PRIVATE	3226	22394.53	136	1
9	FEMALE	WHITE	MIDDLE INCOME	ANY PRIVATE	2852	27008.96	136	1
10	MALE	WHITE	MIDDLE INCOME	ANY PRIVATE	112	25108.71	136	1
11	MALE	WHITE	MIDDLE INCOME	ANY PRIVATE	3179	17569.81	136	1
12	MALE	WHITE	MIDDLE INCOME	ANY PRIVATE	168	21478.06	136	1
13	FEMALE	WHITE	MIDDLE INCOME	ANY PRIVATE	1066	21415.68	136	1
14	MALE	WHITE	NEGATIVE OR POOR	PUBLIC ONLY	0	12254.66	125	1
15	MALE	WHITE	NEGATIVE OR POOR	ANY PRIVATE	0	17699.75	125	1
16	FEMALE	WHITE	NEGATIVE OR POOR	UNINSURED	0	18083.15	125	1
17	MALE	BLACK	NEGATIVE OR POOR	PUBLIC ONLY	230	6537.97	78	10
18	MALE	WHITE	LOW INCOME	UNINSURED	408	8951.36	95	2
19	FEMALE	WHITE	LOW INCOME	UNINSURED	0	11833.00	95	2
20	MALE	WHITE	LOW INCOME	UNINSURED	40	12754.07	95	2
21	FEMALE	WHITE	LOW INCOME	UNINSURED	51	14698.57	95	2
22	MALE	WHITE	LOW INCOME	UNINSURED	0	3890.20	92	19
23	FEMALE	WHITE	LOW INCOME	UNINSURED	610	5882.29	92	19
24	MALE	WHITE	LOW INCOME	PUBLIC ONLY	24	8610.47	92	19
25	FEMALE	BLACK	MIDDLE INCOME	UNINSURED	1758	0.00	64	1
26	MALE	BLACK	MIDDLE INCOME	PUBLIC ONLY	551	7049.70	64	1
27	MALE	BLACK	MIDDLE INCOME	ANY PRIVATE	65	34067.03	64	1
28	FEMALE	BLACK	NEGATIVE OR POOR	PUBLIC ONLY	0	9313.84	73	12
29	FEMALE	BLACK	NEGATIVE OR POOR	PUBLIC ONLY	10	14697.03	73	12
30	MALE	BLACK	NEGATIVE OR POOR	PUBLIC ONLY	0	4574.73	73	12

The following SAS statements fit a generalized logit model for the 1999 full-year consolidated MEPS data:

```
proc surveylogistic data=meps;
   stratum VARSTR99;
   cluster VARPSU99;
   weight PERWT99F;
   class SEX RACEX POVCAT99;
   model INSCOV99 = TOTEXP99 SEX RACEX POVCAT99 / link=glogit;
run;
```

The STRATUM statement specifies the stratification variable VARSTR99. The CLUSTER statement specifies the PSU variable VARPSU99. The WEIGHT statement specifies the sample weight variable PERWT99F. The demographic variables SEX, RACEX, and POVCAT99 are listed in the CLASS statement to indicate that they are categorical independent variables in the MODEL statement. In the MODEL statement, the response variable is INSCOV99, and the independent variables are TOTEXP99 along with the selected demographic variables. The LINK= option requests that the procedure fit the generalized logit model because the response variable INSCOV99 has nominal responses.

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

PROC SURVEYLOGISTIC lists the model fitting information and sample design information in Output 117.2.2.

Output 117.2.2 MEPS, Model Information

The SURVEYLOGISTIC Procedure

Model Information	
Data Set	WORK.MEPS
Response Variable	INSCOV99
Number of Response Levels	3
Stratum Variable	VARSTR99
Number of Strata	143
Cluster Variable	VARPSU99
Number of Clusters	460
Weight Variable	PERWT99F
Model	Generalized Logit
Optimization Technique	Newton-Raphson
Variance Adjustment	Degrees of Freedom (DF)

Output 117.2.3 displays the number of observations and the total of sampling weights both in the data set and used in the analysis. Only the observations with positive person-level weight are used in the analysis. Therefore, 1,053 observations with zero person-level weights were deleted.

Output 117.2.3 MEPS, Number of Observations

Number of Observations Read	24618
Number of Observations Used	23565
Sum of Weights Read	2.7641E8
Sum of Weights Used	2.7641E8

Output 117.2.4 lists the three insurance coverage levels for the response variable INSCOV99. The "UNINSURED" category is used as the reference category in the model.

Output 117.2.4 MEPS, Response Profile

Ordered Value	INSCOV99	Total Frequency	Total Weight
1	ANY PRIVATE	16130	204403997
2	PUBLIC ONLY	4241	41809572
3	UNINSURED	3194	30197198

Logits modeled use INSCOV99='UNINSURED' as the reference category.

Output 117.2.5 shows the parameterization in the regression model for each categorical independent variable.

Output 117.2.5 MEPS, Classification Levels

Class Level Information

Class	Value	Design Variables			
SEX	FEMALE	1			
	MALE	-1			
RACEX	ALEUT, ESKIMO	1	0	0	0
	AMERICAN INDIAN	0	1	0	0
	ASIAN OR PACIFIC ISLANDER	0	0	1	0
	BLACK	0	0	0	1
	WHITE	-1	-1	-1	-1
POVCAT99	HIGH INCOME	1	0	0	0
	LOW INCOME	0	1	0	0
	MIDDLE INCOME	0	0	1	0
	NEAR POOR	0	0	0	1
	NEGATIVE OR POOR	-1	-1	-1	-1

Output 117.2.6 displays the parameter estimates and their standard errors.

Output 117.2.6 MEPS, Parameter Estimates

Analysis of Maximum Likelihood Estimates

Parameter		INSCOV99	Estimate	Standard Error	t Value	Pr > |t|
Intercept		ANY PRIVATE	2.7703	0.1907	14.53	<.0001
Intercept		PUBLIC ONLY	1.9216	0.1563	12.30	<.0001
TOTEXP99		ANY PRIVATE	0.000215	0.000071	3.03	0.0026
TOTEXP99		PUBLIC ONLY	0.000241	0.000072	3.34	0.0009
SEX	FEMALE	ANY PRIVATE	0.1208	0.0248	4.87	<.0001
SEX	FEMALE	PUBLIC ONLY	0.1741	0.0308	5.65	<.0001
RACEX	ALEUT, ESKIMO	ANY PRIVATE	7.1457	0.6978	10.24	<.0001
RACEX	ALEUT, ESKIMO	PUBLIC ONLY	7.6303	0.5028	15.18	<.0001
RACEX	AMERICAN INDIAN	ANY PRIVATE	-2.0904	0.2616	-7.99	<.0001
RACEX	AMERICAN INDIAN	PUBLIC ONLY	-1.8992	0.2910	-6.53	<.0001
RACEX	ASIAN OR PACIFIC ISLANDER	ANY PRIVATE	-1.8055	0.2299	-7.85	<.0001
RACEX	ASIAN OR PACIFIC ISLANDER	PUBLIC ONLY	-1.9914	0.2286	-8.71	<.0001
RACEX	BLACK	ANY PRIVATE	-1.7517	0.1984	-8.83	<.0001
RACEX	BLACK	PUBLIC ONLY	-1.7038	0.1692	-10.07	<.0001
POVCAT99	HIGH INCOME	ANY PRIVATE	1.4560	0.0685	21.26	<.0001
POVCAT99	HIGH INCOME	PUBLIC ONLY	-0.6092	0.0903	-6.75	<.0001
POVCAT99	LOW INCOME	ANY PRIVATE	-0.3066	0.0666	-4.60	<.0001
POVCAT99	LOW INCOME	PUBLIC ONLY	-0.0239	0.0754	-0.32	0.7512
POVCAT99	MIDDLE INCOME	ANY PRIVATE	0.6467	0.0587	11.01	<.0001
POVCAT99	MIDDLE INCOME	PUBLIC ONLY	-0.3496	0.0807	-4.33	<.0001
POVCAT99	NEAR POOR	ANY PRIVATE	-0.8015	0.1076	-7.45	<.0001
POVCAT99	NEAR POOR	PUBLIC ONLY	0.2985	0.0952	3.14	0.0019

NOTE: The degrees of freedom for the t tests is 317.

Output 117.2.7 displays the odds ratio estimates and their standard errors.

Output 117.2.7 MEPS, Odds Ratios

Odds Ratio Estimates

Effect	INSCOV99	Point Estimate	95% Confidence Limits	
TOTEXP99	ANY PRIVATE	1.000	1.000	1.000
TOTEXP99	PUBLIC ONLY	1.000	1.000	1.000
SEX FEMALE vs MALE	ANY PRIVATE	1.273	1.155	1.404
SEX FEMALE vs MALE	PUBLIC ONLY	1.417	1.255	1.599
RACEX ALEUT, ESKIMO vs WHITE	ANY PRIVATE	>999.999	>999.999	>999.999
RACEX ALEUT, ESKIMO vs WHITE	PUBLIC ONLY	>999.999	>999.999	>999.999
RACEX AMERICAN INDIAN vs WHITE	ANY PRIVATE	0.553	0.339	0.903
RACEX AMERICAN INDIAN vs WHITE	PUBLIC ONLY	1.146	0.602	2.185
RACEX ASIAN OR PACIFIC ISLANDER vs WHITE	ANY PRIVATE	0.735	0.499	1.084
RACEX ASIAN OR PACIFIC ISLANDER vs WHITE	PUBLIC ONLY	1.045	0.655	1.670
RACEX BLACK vs WHITE	ANY PRIVATE	0.776	0.638	0.944
RACEX BLACK vs WHITE	PUBLIC ONLY	1.394	1.129	1.721
POVCAT99 HIGH INCOME vs NEGATIVE OR POOR	ANY PRIVATE	11.595	9.293	14.467
POVCAT99 HIGH INCOME vs NEGATIVE OR POOR	PUBLIC ONLY	0.274	0.213	0.353
POVCAT99 LOW INCOME vs NEGATIVE OR POOR	ANY PRIVATE	1.990	1.606	2.466
POVCAT99 LOW INCOME vs NEGATIVE OR POOR	PUBLIC ONLY	0.492	0.395	0.615
POVCAT99 MIDDLE INCOME vs NEGATIVE OR POOR	ANY PRIVATE	5.162	4.197	6.348
POVCAT99 MIDDLE INCOME vs NEGATIVE OR POOR	PUBLIC ONLY	0.356	0.280	0.452
POVCAT99 NEAR POOR vs NEGATIVE OR POOR	ANY PRIVATE	1.213	0.901	1.632
POVCAT99 NEAR POOR vs NEGATIVE OR POOR	PUBLIC ONLY	0.680	0.526	0.878

NOTE: The degrees of freedom in computing the confidence limits is 317.

For example, after adjusting for the effects of sex, race, and total health care expenditures, a person with high income is estimated to be 11.595 times more likely than a poor person to choose private health care insurance over no insurance, but only 0.274 times as likely to choose public health insurance over no insurance.

References

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

Agresti, A. (2002). *Categorical Data Analysis*. 2nd ed. New York: John Wiley & Sons.

Aitchison, J., and Silvey, S. (1957). "The Generalization of Probit Analysis to the Case of Multiple Responses." *Biometrika* 44:131–140.

Albert, A., and Anderson, J. A. (1984). "On the Existence of Maximum Likelihood Estimates in Logistic Regression Models." *Biometrika* 71:1–10.

Ashford, J. R. (1959). "An Approach to the Analysis of Data for Semi-quantal Responses in Biology Response." *Biometrics* 15:573–581.

Beaumont, J.-F., and Patak, Z. (2012). "On the Generalized Bootstrap for Sample Surveys with Special Attention to Poisson Sampling." *International Statistical Review* 80:127–148.

Binder, D. A. (1981). "On the Variances of Asymptotically Normal Estimators from Complex Surveys." *Survey Methodology* 7:157–170.

Binder, D. A. (1983). "On the Variances of Asymptotically Normal Estimators from Complex Surveys." *International Statistical Review* 51:279–292.

Binder, D. A., and Roberts, G. R. (2003). "Design-Based and Model-Based Methods for Estimating Model Parameters." In *Analysis of Survey Data*, edited by C. Skinner and R. Chambers, 29–48. New York: John Wiley & Sons.

Brick, J. M., and Kalton, G. (1996). "Handling Missing Data in Survey Research." *Statistical Methods in Medical Research* 5:215–238.

Cochran, W. G. (1977). *Sampling Techniques*. 3rd ed. New York: John Wiley & Sons.

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

Cox, D. R., and Snell, E. J. (1989). *The Analysis of Binary Data*. 2nd ed. London: Chapman & Hall.

Dippo, C. S., Fay, R. E., and Morganstein, D. H. (1984). "Computing Variances from Complex Samples with Replicate Weights." In *Proceedings of the Survey Research Methods Section*, 489–494. Alexandria, VA: American Statistical Association.

Fay, R. E. (1984). "Some Properties of Estimates of Variance Based on Replication Methods." In *Proceedings of the Survey Research Methods Section*, 495–500. Alexandria, VA: American Statistical Association.

Fay, R. E. (1989). "Theory and Application of Replicate Weighting for Variance Calculations." In *Proceedings of the Survey Research Methods Section*, 212–217. Alexandria, VA: American Statistical Association.

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

Fuller, W. A. (1975). "Regression Analysis for Sample Survey." *Sankhyā, Series C* 37:117–132.

Fuller, W. A. (2009). *Sampling Statistics*. Hoboken, NJ: John Wiley & Sons.

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.

Hidiroglou, M. A., Fuller, W. A., and Hickman, R. D. (1980). *SUPER CARP*. Ames: Iowa State University Statistical Laboratory.

Hosmer, D. W., Jr., and Lemeshow, S. (2000). *Applied Logistic Regression*. 2nd ed. New York: John Wiley & Sons.

Judkins, D. R. (1990). "Fay's Method for Variance Estimation." *Journal of Official Statistics* 6:223–239.

Kalton, G., and Kasprzyk, D. (1986). "The Treatment of Missing Survey Data." *Survey Methodology* 12:1–16.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.

Korn, E. L., and Graubard, B. I. (1999). *Analysis of Health Surveys*. New York: John Wiley & Sons.

Lancaster, H. O. (1961). "Significance Tests in Discrete Distributions." *Journal of the American Statistical Association* 56:223–234.

Lehtonen, R., and Pahkinen, E. (1995). *Practical Methods for Design and Analysis of Complex Surveys*. Chichester, UK: John Wiley & Sons.

Littell, R. C., Freund, R. J., and Spector, P. C. (1991). *SAS System for Linear Models*. 3rd ed. Cary, NC: SAS Institute Inc.

Lohr, S. L. (2010). *Sampling: Design and Analysis*. 2nd ed. Boston: Brooks/Cole.

Machlin, S., Yu, W., and Zodet, M. (2005). "Computing Standard Errors for MEPS Estimates." Rockville, MD: Agency for Healthcare Research and Quality. http://www.meps.ahrq.gov/mepsweb/survey_comp/standard_errors.jsp.

Mashreghi, Z., Haziza, D., and Léger, C. (2016). "A Survey of Bootstrap Methods in Finite Population Sampling." *Statistics Surveys* 10:1–52.

McCarthy, P. J., and Snowden, C. B. (1985). *The Bootstrap and Finite Population Sampling.* Vital and Health Statistics, Series 2, no. 95. DHHS Publication Number 85–1369. Washington, DC: US Government Printing Office. https://www.cdc.gov/nchs/data/series/sr_02/sr02_095.pdf.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd ed. London: Chapman & Hall.

McFadden, D. (1974). "Conditional Logit Analysis of Qualitative Choice Behavior." In *Frontiers in Econometrics*, edited by P. Zarembka, 105–142. New York: Academic Press.

MEPS (2002). "MEPS HC-038: 1999 Full Year Consolidated Data File." Rockville, MD: Agency for Healthcare Research and Quality. Accessed July 22, 2011. http://www.meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-038.

MEPS (2006). "MEPS Survey Background." Rockville, MD: Agency for Healthcare Research and Quality. Accessed July 22, 2011. http://www.meps.ahrq.gov/mepsweb/about_meps/survey_back.jsp.

Morel, J. G. (1989). "Logistic Regression under Complex Survey Designs." *Survey Methodology* 15:203–223.

Muller, K. E., and Fetterman, B. A. (2002). *Regression and ANOVA: An Integrated Approach Using SAS Software*. Cary, NC: SAS Institute Inc.

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* 135:370–384.

Rao, J. N. K., and Scott, A. J. (1981). "The Analysis of Categorical Data from Complex Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables." *Journal of the American Statistical Association* 76:221–230.

Rao, J. N. K., and Scott, A. J. (1984). "On Chi-Squared Tests for Multiway Contingency Tables with Cell Properties Estimated from Survey Data." *Annals of Statistics* 12:46–60.

Rao, J. N. K., and Scott, A. J. (1987). "On Simple Adjustments to Chi-Square Tests with Survey Data." *Annals of Statistics* 15:385–397.

Rao, J. N. K., Scott, A. J., and Skinner, C. J. (1998). "Quasi-score Tests with Survey Data." *Statistica Sinica* 8:1059–1070.

Rao, J. N. K., and Shao, J. (1996). "On Balanced Half-Sample Variance Estimation in Stratified Random Sampling." *Journal of the American Statistical Association* 91:343–348.

Rao, J. N. K., and Shao, J. (1999). "Modified Balanced Repeated Replication for Complex Survey Data." *Biometrika* 86:403–415.

Rao, J. N. K., and Thomas, D. R. (1989). "Chi-Squared Tests for Contingency Tables." In *Analysis of Complex Surveys*, edited by C. J. Skinner, D. Holt, and T. M. F. Smith, 89–114. New York: John Wiley & Sons.

Rao, J. N. K., and Wu, C. F. J. (1988). "Resampling Inference with Complex Survey Data." *Journal of the American Statistical Association* 83:231–241.

Rao, J. N. K., Wu, C. F. J., and Yue, K. (1992). "Some Recent Work on Resampling Methods for Complex Surveys." *Survey Methodology* 18:209–217.

Roberts, G., Rao, J. N. K., and Kumar, S. (1987). "Logistic Regression Analysis of Sample Survey Data." *Biometrika* 74:1–12.

Rust, K. (1985). "Variance Estimation for Complex Estimators in Sample Surveys." *Journal of Official Statistics* 1:381–397.

Rust, K., and Kalton, G. (1987). "Strategies for Collapsing Strata for Variance Estimation." *Journal of Official Statistics* 3:69–81.

Santner, T. J., and Duffy, D. E. (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.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Shao, J., and Tu, D. (1995). *The Jackknife and Bootstrap*. New York: Springer-Verlag.

Sitter, R. R. (1992a). "Comparing Three Bootstrap Methods for Survey Data." *Canadian Journal of Statistics* 20:135–154.

Sitter, R. R. (1992b). "A Resampling Procedure for Complex Survey Data." *Journal of the American Statistical Association* 87:755–765.

Skinner, C. J., Holt, D., and Smith, T. M. F. (1989). *Analysis of Complex Surveys*. New York: John Wiley & Sons.

Stokes, M. E., Davis, C. S., and Koch, G. G. (2000). *Categorical Data Analysis Using the SAS System*. 2nd ed. Cary, NC: SAS Institute Inc.

Thomas, D. R., and Rao, J. N. K. (1984). "A Monte Carlo Study of Exact Levels of Goodness-of-Fit Statistics under Cluster Sampling." In *Proceedings of the Survey Research Methods Section*, 207–211. Alexandria, VA: American Statistical Association.

Thomas, D. R., and Rao, J. N. K. (1987). "Small-Sample Comparisons of Level and Power for Simple Goodness-of-Fit Statistics under Cluster Sampling." *Journal of the American Statistical Association* 82:630–636.

Thomas, D. R., Singh, A. C., and Roberts, G. R. (1996). "Tests of Independence on Two-Way Tables under Cluster Sampling: An Evaluation." *International Statistical Review* 64:295–311.

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.

Wolter, K. M. (2007). *Introduction to Variance Estimation*. 2nd ed. New York: Springer.

Woodruff, R. S. (1971). "A Simple Method for Approximating the Variance of a Complicated Estimate." *Journal of the American Statistical Association* 66:411–414.

Subject Index

Syntax Index