SAS/STAT® 13.1 User's Guide
The SURVEYREG Procedure

This document is an individual chapter from *SAS/STAT® 13.1 User's Guide*.

The correct bibliographic citation for the complete manual is as follows: SAS Institute Inc. 2013. *SAS/STAT® 13.1 User's Guide*. Cary, NC: SAS Institute Inc.

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513-2414.

December 2013

Chapter 98

The SURVEYREG Procedure

Contents

Overview: SURVEYREG Procedure

The SURVEYREG procedure performs regression analysis for sample survey data. This procedure can handle complex survey sample designs, including designs with stratification, clustering, and unequal weighting. The procedure fits linear models for survey data and computes regression coefficients and their variance-covariance matrix. PROC SURVEYREG also provides significance tests for the model effects and for any specified estimable linear functions of the model parameters. Using the regression model, the procedure can compute predicted values for the sample survey data.

PROC SURVEYREG uses elementwise regression to compute the regression coefficient estimators by generalized least squares estimation. The procedure assumes that the regression coefficients are the same across strata and primary sampling units (PSUs). To estimate the variance-covariance matrix for the regression coefficients, PROC SURVEYREG uses either the Taylor series (linearization) method or replication (resampling) methods to estimate sampling errors of estimators, based on complex sample designs. For details see Woodruff (1971); Fuller (1975); Särndal, Swensson, and Wretman (1992); Wolter (2007); Rust (1985); Dippo, Fay, and Morganstein (1984); Rao and Shao (1999); Rao, Wu, and Yue (1992); and Rao and Shao (1996).

PROC SURVEYREG uses the Output Delivery System (ODS), a SAS subsystem that provides capabilities for displaying and controlling the output from SAS procedures. ODS enables you to convert any of the output from PROC SURVEYREG into a SAS data set. For more information, see the section "ODS Table Names" on page 8406.

PROC SURVEYREG uses ODS Graphics to create graphs as part of its output. For general information about ODS Graphics, see Chapter 21, "Statistical Graphics Using ODS." For specific information about the statistical graphics available with the SURVEYREG procedure, see the PLOTS= option in the PROC SURVEYREG statement and the section "ODS Graphics" on page 8408.

Getting Started: SURVEYREG Procedure

This section demonstrates how you can use PROC SURVEYREG to perform a regression analysis for sample survey data. For a complete description of the usage of PROC SURVEYREG, see the section "Syntax: SURVEYREG Procedure" on page 8363. The section "Examples: SURVEYREG Procedure" on page 8409 provides more detailed examples that illustrate the applications of PROC SURVEYREG.

Simple Random Sampling

Suppose that, in a junior high school, there are a total of 4,000 students in grades 7, 8, and 9. You want to know how household income and the number of children in a household affect students' average weekly spending for ice cream.

In order to answer this question, you draw a sample by using simple random sampling from the student population in the junior high school. You randomly select 40 students and ask them their average weekly expenditure for ice cream, their household income, and the number of children in their household. The answers from the 40 students are saved as the following SAS data set IceCream:

```
data IceCream;
   input Grade Spending Income Kids @@;
   datalines;
7   7  39  2   7   7  38  1   8  12  47  1
9  10  47  4   7   1  34  4   7  10  43  2
7   3  44  4   8  20  60  3   8  19  57  4
7   2  35  2   7   2  36  1   9  15  51  1
8  16  53  1   7   6  37  4   7   6  41  2
7   6  39  2   9  15  50  4   8  17  57  3
8  14  46  2   9   8  41  2   9   8  41  1
9   7  47  3   7   3  39  3   7  12  50  2
7   4  43  4   9  14  46  3   8  18  58  4
9   9  44  3   7   2  37  1   7   1  37  2
7   4  44  2   7  11  42  2   9   8  41  2
8  10  42  2   8  13  46  1   7   2  40  3
9   6  45  1   9  11  45  4   7   2  36  1
7   9  46  1
;
```

In the data set IceCream, the variable Grade indicates a student's grade. The variable Spending contains the dollar amount of each student's average weekly spending for ice cream. The variable Income specifies the household income, in thousands of dollars. The variable Kids indicates how many children are in a student's family.

The following PROC SURVEYREG statements request a regression analysis:

```
title1 'Ice Cream Spending Analysis';
title2 'Simple Random Sample Design';
proc surveyreg data=IceCream total=4000;
   class Kids;
   model Spending = Income Kids / solution;
run;
```

The PROC SURVEYREG statement invokes the procedure. The TOTAL=4000 option specifies the total in the population from which the sample is drawn. The CLASS statement requests that the procedure use the variable Kids as a classification variable in the analysis. The MODEL statement describes the linear model that you want to fit, with Spending as the dependent variable and Income and Kids as the independent variables. The SOLUTION option in the MODEL statement requests that the procedure output the regression coefficient estimates.

Figure 98.1 displays the summary of the data, the summary of the fit, and the levels of the classification variable Kids. The "Fit Statistics" table displays the denominator degrees of freedom, which are used in *F* tests and *t* tests in the regression analysis.

Figure 98.1 Summary of Data

```
                  Ice Cream Spending Analysis
                  Simple Random Sample Design

                    The SURVEYREG Procedure

      Regression Analysis for Dependent Variable Spending

                         Data Summary

          Number of Observations              40
          Mean of Spending               8.75000
          Sum of Spending              350.00000


                        Fit Statistics

               R-Square            0.8132
               Root MSE            2.4506
               Denominator DF          39


                    Class Level Information

            CLASS
            Variable      Levels    Values

            Kids               4    1 2 3 4
```

Figure 98.2 displays the tests for model effects. The effect Income is significant in the linear regression model, while the effect Kids is not significant at the 5% level.

Figure 98.2 Testing Effects in the Regression

```
                     Tests of Model Effects

          Effect        Num DF     F Value    Pr > F

          Model              4      119.15    <.0001
          Intercept          1      153.32    <.0001
          Income             1      324.45    <.0001
          Kids               3        0.92    0.4385

    NOTE: The denominator degrees of freedom for the F tests is 39.
```

The regression coefficient estimates and their standard errors and associated *t* tests are displayed in Figure 98.3.

Figure 98.3 Regression Coefficients

```
                    Estimated Regression Coefficients

                                      Standard
         Parameter       Estimate        Error    t Value    Pr > |t|

         Intercept     -26.084677   2.46720403     -10.57      <.0001
         Income          0.775330   0.04304415      18.01      <.0001
         Kids 1          0.897655   1.12352876       0.80      0.4292
         Kids 2          1.494032   1.24705263       1.20      0.2381
         Kids 3         -0.513181   1.33454891      -0.38      0.7027
         Kids 4          0.000000   0.00000000        .          .

NOTE: The denominator degrees of freedom for the t tests is 39.
      Matrix X'X is singular and a generalized inverse was used to solve the
      normal equations.  Estimates are not unique.
```

Stratified Sampling

Suppose that the previous student sample is actually selected by using a stratified sample design. The strata are the grades in the junior high school: 7, 8, and 9. Within the strata, simple random samples are selected. Table 98.1 provides the number of students in each grade.

Table 98.1 Students in Grades

Grade	Number of Students
7	1,824
8	1,025
9	1,151
Total	**4,000**

In order to analyze this sample by using PROC SURVEYREG, you need to input the stratification information by creating a SAS data set that contains the information in Table 98.1. The following SAS statements create such a data set, named StudentTotals:

```
data StudentTotals;
   input Grade _TOTAL_;
   datalines;
7 1824
8 1025
9 1151
;
```

The variable Grade is the stratification variable, and the variable _TOTAL_ contains the total numbers of students in each stratum in the survey population. PROC SURVEYREG requires you to use the keyword _TOTAL_ as the name of the variable that contains the population totals.

When the sample design is stratified and the stratum sampling rates are unequal, you should use sampling weights to reflect this information in the analysis. For this example, the appropriate sampling weights are the reciprocals of the probabilities of selection. You can use the following DATA step to create the sampling weights:

```
data IceCream;
   set IceCream;
   if Grade=7 then Prob=20/1824;
   if Grade=8 then Prob=9/1025;
   if Grade=9 then Prob=11/1151;
   Weight=1/Prob;
run;
```

If you use PROC SURVEYSELECT to select your sample, PROC SURVEYSELECT creates these sampling weights for you.

The following statements demonstrate how you can fit a linear model while incorporating the sample design information (stratification and unequal weighting):

```
ods graphics on;
title1 'Ice Cream Spending Analysis';
title2 'Stratified Sample Design';
proc surveyreg data=IceCream total=StudentTotals;
   strata Grade /list;
   model Spending = Income;
   weight Weight;
run;
ods graphics off;
```

Comparing these statements to those in the section "Simple Random Sampling" on page 8355, you can see how the TOTAL=StudentTotals option replaces the previous TOTAL=4000 option.

The STRATA statement specifies the stratification variable Grade. The LIST option in the STRATA statement requests that the stratification information be displayed. The WEIGHT statement specifies the weight variable.

Figure 98.4 summarizes the data information, the sample design information, and the fit information. Because of the stratification, the denominator degrees of freedom for F tests and t tests are 37, which are different from those in the analysis in Figure 98.1.

Figure 98.4 Summary of the Regression

```
                    Ice Cream Spending Analysis
                      Stratified Sample Design

                      The SURVEYREG Procedure

          Regression Analysis for Dependent Variable Spending

                           Data Summary

                Number of Observations              40
                Sum of Weights                  4000.0
                Weighted Mean of Spending      9.14130
                Weighted Sum of Spending       36565.2
```

Figure 98.4 *continued*

Design Summary

Number of Strata	3

Fit Statistics

R-Square	0.8037
Root MSE	2.4371
Denominator DF	37

Figure 98.5 displays the following information for each stratum: the value of the stratification variable, the number of observations (sample size), the total population size, and the sampling rate (fraction).

Figure 98.5 Stratification Information

Stratum Information

Stratum Index	Grade	N Obs	Population Total	Sampling Rate
1	7	20	1824	1.10%
2	8	9	1025	0.88%
3	9	11	1151	0.96%

Figure 98.6 displays the tests for significance of the model effects. The Income effect is strongly significant at the 5% level.

Figure 98.6 Testing Effects

Tests of Model Effects

Effect	Num DF	F Value	Pr > F
Model	1	492.39	<.0001
Intercept	1	225.81	<.0001
Income	1	492.39	<.0001

NOTE: The denominator degrees of freedom for the F tests is 37.

Figure 98.7 displays the regression coefficient estimates, their standard errors, and the associated *t* tests for the stratified sample.

Figure 98.7 Regression Coefficients

Estimated Regression Coefficients

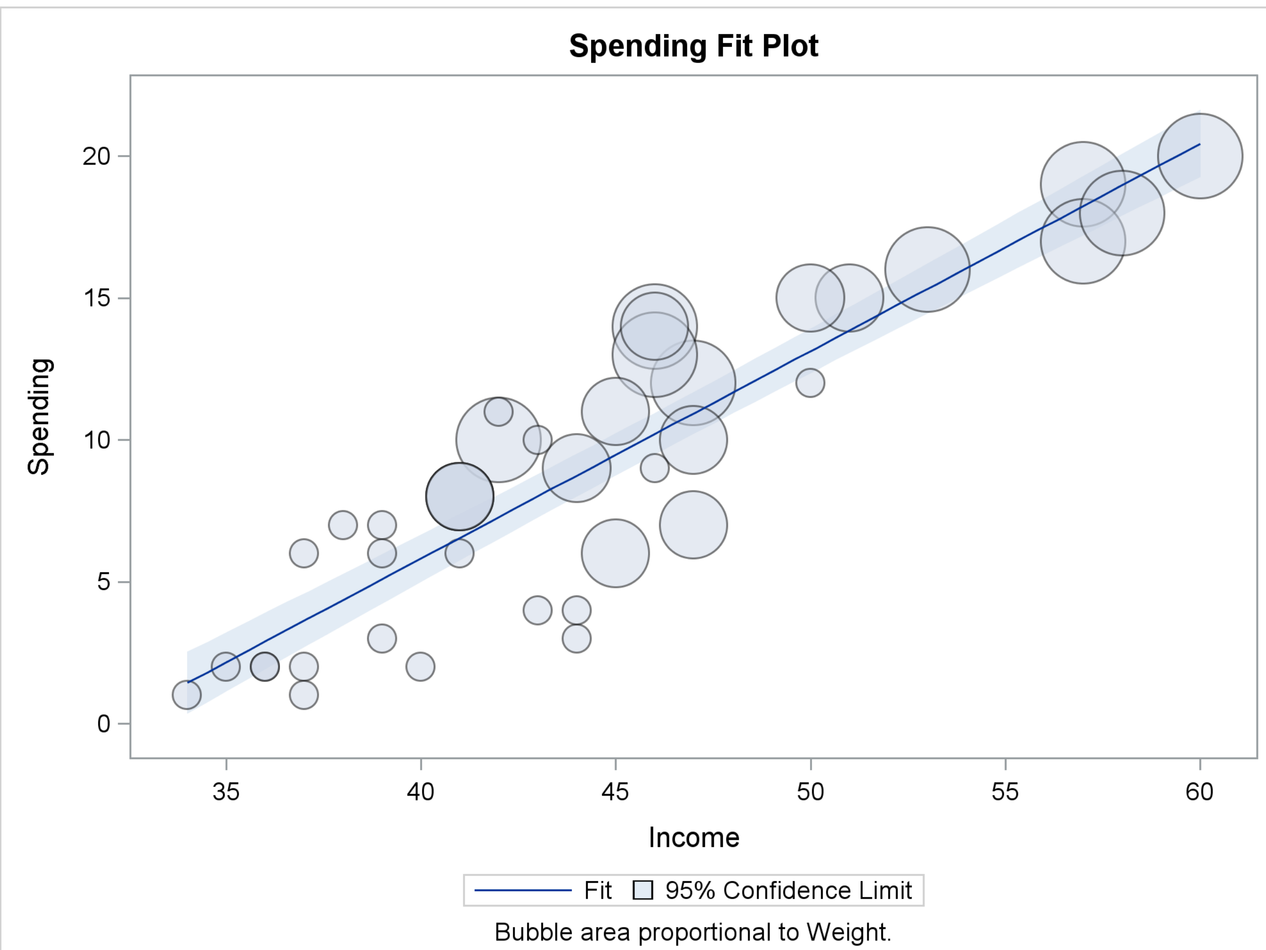
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-23.416322	1.55827214	-15.03	<.0001
Income	0.731052	0.03294520	22.19	<.0001

NOTE: The denominator degrees of freedom for the t tests is 37.

You can request other statistics and tests by using PROC SURVEYREG. You can also analyze data from a more complex sample design. The remainder of this chapter provides more detailed information.

When ODS Graphics is enabled and the model contains a single continuous regressor, PROC SURVEYREG provides a fit plot that displays the regression line and the confidence limits of the mean predictions. Figure 98.8 displays the fit plot for the regression model of Spending as a function of Income. The regression line and confidence limits of mean prediction are overlaid by a bubble plot of the data, in which the bubble area is proportional to the sampling weight of an observation.

Figure 98.8 Regression Fitting



Output Data Sets

You can use the OUTPUT statement to create a new SAS data set that contains the estimated linear predictors and their standard error estimates, the residuals from the linear regression, and the confidence limits for the predictors. See the section "OUTPUT Statement" on page 8383 for more details.

You can use the Output Delivery System (ODS) to create SAS data sets that capture the outputs from PROC SURVEYREG. For more information about ODS, see Chapter 20, "Using the Output Delivery System."

For example, to save the ParameterEstimates table (Figure 98.7) in the previous section in an output data set, you use the ODS OUTPUT statement as follows:

```
title1 'Ice Cream Spending Analysis';
title2 'Stratified Sample Design';
proc surveyreg data=IceCream total=StudentTotals;
   strata Grade /list;
   model Spending = Income;
   weight Weight;
   ods output ParameterEstimates = MyParmEst;
run;
```

The statement

```
ods output ParameterEstimates = MyParmEst;
```

requests that the ParameterEstimates table that appears in Figure 98.7 be placed into a SAS data set MyParmEst.

The PRINT procedure displays observations of the data set MyParmEst:

```
proc print data=MyParmEst;
run;
```

Figure 98.9 displays the observations in the data set MyParmEst. The section "ODS Table Names" on page 8406 gives the complete list of the tables produced by PROC SURVEYREG.

Figure 98.9 The Data Set MyParmEst

Ice Cream Spending Analysis
Stratified Sample Design

Obs	Parameter	Estimate	StdErr	DenDF	tValue	Probt
1	Intercept	-23.416322	1.55827214	37	-15.03	<.0001
2	Income	0.731052	0.03294520	37	22.19	<.0001

Syntax: SURVEYREG Procedure

The following statements are available in the SURVEYREG procedure:

PROC SURVEYREG < *options* > **;**
BY *variables* **;**
CLASS *variables* **;**
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* > **;**
LSMEANS < *model-effects* > < */ options* > **;**
LSMESTIMATE *model-effect lsmestimate-specification* < */ options* > **;**
MODEL *dependent* **=** < *effects* > < */ options* > **;**
OUTPUT < *keyword*< **=***variable-name* > *. . . keyword*< **=***variable-name* > > < */ option* > **;**
REPWEIGHTS *variables* < */ options* > **;**
SLICE *model-effect* < */ options* > **;**
STORE < **OUT=** >*item-store-name* < /**LABEL=***'label'* > **;**
STRATA *variables* < */ options* > **;**
TEST < *model-effects* > < */ options* > **;**
WEIGHT *variable* **;**

The PROC SURVEYREG and MODEL statements are required. If your model contains classification effects, you must list the classification variables in a CLASS statement, and the CLASS statement must precede the MODEL statement. If you use a CONTRAST statement or an ESTIMATE statement, the MODEL statement must precede the CONTRAST or ESTIMATE statement.

The rest of this section provides detailed syntax information for each of the preceding statements, except the EFFECT, ESTIMATE, LSMEANS, LSMESTIMATE, SLICE, STORE, and TEST 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 CLASS, CLUSTER, CONTRAST, EFFECT, ESTIMATE, LSMEANS, LSMESTIMATE, REPWEIGHTS, SLICE, STRATA, TEST statements can appear multiple times. You should use only one of each of the following statements: MODEL, WEIGHT, STORE, and OUTPUT.

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

PROC SURVEYREG Statement

PROC SURVEYREG < *options* > **;**

The PROC SURVEYREG statement invokes the SURVEYREG procedure. It optionally names the input data sets and specifies the variance estimation method.

Table 98.2 summarizes the *options* available in the PROC SURVEYREG statement.

Table 98.2 PROC SURVEYREG Statement Options

Option	Description
ALPHA=	Sets the confidence level
DATA=	Specifies the SAS data set to be analyzed
MISSING	Treats missing values as a nonmissing
NAMELEN=	Specifies the length of effect names
NOMCAR	Treats missing values as *not missing completely at random*
ORDER=	Specifies the sort order
PLOTS=	Requests plots from ODS Graphics
RATE=	Specifies the sampling rate
TOTAL=	Specifies the total number of primary sampling units
TRUNCATE	Specifies class levels using no more than the first 16 characters of the formatted values
VARMETHOD=	Specifies the variance estimation method

You can specify the following *options* in the PROC SURVEYREG statement:

ALPHA=α

sets the confidence level for confidence limits. 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 limits. The default of ALPHA=0.05 produces 95% confidence limits.

DATA=*SAS-data-set*

specifies the SAS data set to be analyzed by PROC SURVEYREG. If you omit the DATA= option, the procedure uses the most recently created SAS data set.

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 8387.

NAMELEN=*n*

specifies the length of effect names in tables and output data sets to be *n* characters, where *n* is a value between 40 and 200. The default length is 40 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 SURVEYREG 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 8387 for more details.

By default, PROC SURVEYREG 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. The replication methods, which you request with the VARMETHOD=BRR and VARMETHOD=JACKKNIFE options, do not use the NOMCAR option.

ORDER=DATA | FORMATTED | FREQ | INTERNAL

specifies the sort order for the levels of the classification variables (which are specified in the CLASS statement).

This option also determines the sort order for the levels of DOMAIN 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*.

PLOTS < (*global-plot-options*) > < = *plot-request* < (*plot-option*) > >

PLOTS < (*global-plot-options*) > < = (*plot-request* < (*plot-option*) > < . . . *plot-request* < (*plot-option*) > >) >

controls the plots that are produced through ODS Graphics.

When ODS Graphics is enabled and when the regression model depends on at most one continuous variable as a regressor, excluding the intercept, the PLOTS= option in the PROC SURVEYREG statement controls fit plots for the regression.

A *plot-request* identifies the plot, and a *plot-option* controls the appearance and content of the plot. You can specify *plot-options* in parentheses after a *plot-request*. A *global-plot-option* applies to all plots for which it is available unless it is altered by a specific *plot-option*. You can specify *global-plot-options* in parentheses after the PLOTS option.

When you specify only one *plot-request*, you can omit the parentheses around it. Here are a few examples of requesting plots:

```
plots=all
plots(weight=heatmap)=fit
```

When the regression model depends on at most one continuous variable as a regressor, excluding the intercept, PROC SURVEYREG provides a bubble plot or a heat map for model fitting. In a bubble plot, the bubble area is proportional to the weight of an observation. In a heat map, the heat color represents the sum of the weights at the corresponding location. The default plot depends on the number of observations in your data. That is, for a data set that contains 100 observations or less, a bubble plot is the default. For a data set that contains more than 100 observations, a heat map is the default.

ODS Graphics must be enabled before you can request a plot. For example:

```
ods graphics on;
proc surveyreg plots=fit;
   model height=weight;
run;
ods graphics off;
```

For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" on page 606 in Chapter 21, "Statistical Graphics Using ODS."

When ODS Graphics is enabled, 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."

For general information about ODS Graphics, see Chapter 21, "Statistical Graphics Using ODS."

Global Plot Option

A *global-plot-option* applies to all plots for which the option is available unless it is altered by a specific *plot-option*. You can specify the following *global-plot-options*:

ONLY

suppresses the default plots and requests only the plots that are specified as *plot-requests*.

NBINS=*nbin1* < *nbin2* >

specifies the number of bins for the heat map of the observation weights in the fit plot. If you specify only one number, *nbin1*, then it is used for both the horizontal and vertical axes; if you specify two numbers, *nbin1* and *nbin2*, then the first, *nbin1*, is used for the horizontal axis and the second, *nbin2*, is used for the vertical axis. If you do not specify this option for numbers of bins, then by default they are determined according to the algorithm that is discussed in the section "ODS Graphics" on page 4089 in Chapter 52, "The KDE Procedure."

WEIGHT=BUBBLE

WEIGHT=HEATMAP | HEAT

requests either a bubble plot or a heat map of the data as an overlay on the regression line and confidence limits band of the prediction in a fit plot. In a bubble plot, the bubble area is proportional to the weight of an observation. In a heat map, the heat color represents the sum of the weights at the corresponding location.

If you do not specify this option, the default plot depends on the number of observations in your data. That is, for a data set that contains 100 observations or less, the default is a bubble plot. For a data set that contains more than 100 observations, the default is a heat map.

Plot Requests

You can specify the following *plot-requests*:

ALL

requests all appropriate plots.

FIT < (*plot-options*) >

requests a plot that displays the model fitting for a model that depends on at most one regressor, excluding the intercept. The plot is either a bubble plot or a heat map that is overlaid with the regression line and confidence band of the prediction.

The FIT plot request has the following *plot-options*:

NBINS=*nbin1* < *nbin2* >

specifies the number of bins for the heat map of the observation weights in the fit plot. If you specify only one number, *nbin1*, then it is used for both the horizontal and vertical axes; if you specify two numbers, *nbin1* and *nbin2*, then the first, *nbin1*, is used for the horizontal axis and the second, *nbin2*, is used for the vertical axis. If you do not specify this option for numbers of bins, then by default they are determined according to the algorithm that is discussed in the section "ODS Graphics" on page 4089 in Chapter 52, "The KDE Procedure."

WEIGHT=BUBBLE

WEIGHT=HEATMAP | HEAT

requests either a bubble plot or a heat map of the data as an overlay on the regression line and confidence limits band of the prediction in a fit plot. In a bubble plot, the bubble area is proportional to the weight of an observation. In a heat map, the heat color represents the sum of the weights at the corresponding location.

If you do not specify this option, the default plot depends on the number of observations in your data. That is, for a data set that contains 100 observations or less, the default is a bubble plot. For a data set that contains more than 100 observations, the default is a heat map.

NONE

suppresses all plots.

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

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

specifies the sampling rate as a nonnegative *value*, or specifies an input data set that contains the stratum sampling rates. The procedure uses this information to compute a finite population correction for Taylor series variance estimation. The procedure does not use the RATE= option for BRR or jackknife variance estimation, which you request with the VARMETHOD=BRR or VARMETHOD=JACKKNIFE option.

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

For a nonstratified sample design, or for a stratified sample design with the same sampling rate in all strata, you should specify a nonnegative *value* for the RATE= option. If your design is stratified with different sampling rates in the strata, then you should name a SAS data set that contains the stratification variables and the sampling rates. See the section "Specification of Population Totals and Sampling Rates" on page 8388 for more details.

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 SURVEYREG converts that number to a proportion. The procedure treats the value 1 as 100%, and not the percentage form 1%.

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

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

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

specifies the total number of primary sampling units in the study population as a positive *value*, or specifies an input data set that contains the stratum population totals. The procedure uses this information to compute a finite population correction for Taylor series variance estimation. The procedure does not use the TOTAL= option for BRR or jackknife variance estimation, which you request with the VARMETHOD=BRR or VARMETHOD=JACKKNIFE option.

For a nonstratified sample design, or for a stratified sample design with the same population total in all strata, you should specify a positive *value* for the TOTAL= option. If your sample design is stratified with different population totals in the strata, then you should name a SAS data set that contains the stratification variables and the population totals. See the section "Specification of Population Totals and Sampling Rates" on page 8388 for more details.

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

TRUNCATE

specifies that class levels should be determined using no more than the first 16 characters of the formatted values of the CLASS, STRATA, and CLUSTER variables. When formatted values are longer than 16 characters, you can use this option in order to revert to the levels as determined in releases before SAS 9.

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

VARMETHOD=JACKKNIFE | JK < **(***method-options***)** >

VARMETHOD=TAYLOR

specifies the variance estimation method. VARMETHOD=TAYLOR requests the Taylor series method, which is the default if you do not specify the VARMETHOD= option or the REPWEIGHTS statement. VARMETHOD=BRR requests variance estimation by balanced repeated replication (BRR), and VARMETHOD=JACKKNIFE requests variance estimation by the delete-1 jackknife method.

For VARMETHOD=BRR and VARMETHOD=JACKKNIFE you can specify *method-options* in parentheses. Table 98.3 summarizes the available *method-options*.

Table 98.3 Variance Estimation Options

VARMETHOD=	Variance Estimation Method	*Method-Options*
BRR	Balanced repeated replication	FAY < =*value* > HADAMARD=*SAS-data-set* OUTWEIGHTS=*SAS-data-set* PRINTH REPS=*number*
JACKKNIFE	Jackknife	OUTJKCOEFS=*SAS-data-set* OUTWEIGHTS=*SAS-data-set*
TAYLOR	Taylor series linearization	None

Method-options must be enclosed in parentheses following the method keyword. For example:

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

The following values are available for the VARMETHOD= option:

BRR < (*method-options*) >

requests balanced repeated replication (BRR) variance estimation. The BRR method requires a stratified sample design with two primary sampling units (PSUs) per stratum. See the section “Balanced Repeated Replication (BRR) Method” on page 8394 for more information.

You can specify the following *method-options* in parentheses following VARMETHOD=BRR:

FAY < =*value* >

requests Fay’s method, a modification of the BRR method, for variance estimation. See the section “Fay’s BRR Method” on page 8395 for more information.

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 value of the Fay coefficient equals 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 provide a Hadamard matrix with the HADAMARD= *method-option*, PROC SURVEYREG generates an appropriate Hadamard matrix for replicate construction. See the sections "Balanced Repeated Replication (BRR) Method" on page 8394 and "Hadamard Matrix" on page 8397 for details.

If a Hadamard matrix of a given 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 the HADAMARD= *method-option*.

In the HADAMARD= input data set, each variable corresponds to a column of the Hadamard matrix, and each observation corresponds to a row of the matrix. You can use any variable names in the HADAMARD= 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 SURVEYREG does not check the validity of the Hadamard matrix that you provide.

The HADAMARD= input 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, the procedure uses only the first H variables. Similarly, the HADAMARD= input data set must contain at least H observations.

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

You can specify the PRINTH option to display the Hadamard matrix that the procedure uses to construct replicates for BRR.

OUTWEIGHTS=*SAS-data-set*

names a SAS data set that contains replicate weights. See the section "Balanced Repeated Replication (BRR) Method" on page 8394 for information about replicate weights. See the section "Replicate Weights Output Data Set" on page 8401 for more details about the contents of the OUTWEIGHTS= data set.

The OUTWEIGHTS= *method-option* is not available when you provide replicate weights with the REPWEIGHTS statement.

PRINTH

displays the Hadamard matrix.

When you provide your own Hadamard matrix with the HADAMARD= *method-option*, only the rows and columns of the Hadamard matrix that are used by the procedure are displayed. See the sections "Balanced Repeated Replication (BRR) Method" on page 8394 and "Hadamard Matrix" on page 8397 for details.

The PRINTH *method-option* is not available when you provide replicate weights with the 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 provide a Hadamard matrix with the HADAMARD= *method-option*, the number of replicates should be greater than the number of strata and should be a multiple of 4. See the section "Balanced Repeated Replication (BRR) Method" on page 8394 for more information. If a Hadamard matrix cannot be constructed for the REPS= value that you specify, the value is increased until a Hadamard matrix of that dimension can be constructed. Therefore, it is possible for the actual number of replicates used to be larger than the REPS= value that you specify.

If you provide a Hadamard matrix with the HADAMARD= *method-option*, the value of REPS= must not be less 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 equals the number of rows in the Hadamard matrix.

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

If you provide replicate weights with the REPWEIGHTS statement, the procedure does not use the REPS= *method-option*. With a REPWEIGHTS statement, the number of replicates equals the number of REPWEIGHTS variables.

JACKKNIFE | JK < (*method-options*) >

requests variance estimation by the delete-1 jackknife method. See the section "Jackknife Method" on page 8396 for details. If you provide replicate weights with a REPWEIGHTS statement, VARMETHOD=JACKKNIFE is the default variance estimation method.

You can specify the following *method-options* in parentheses following VARMETHOD=JACKKNIFE:

OUTJKCOEFS=*SAS-data-set*

names a SAS data set that contains jackknife coefficients. See the section "Jackknife Method" on page 8396 for information about jackknife coefficients. See the section "Jackknife Coefficients Output Data Set" on page 8401 for more details about the contents of the OUTJKCOEFS= data set.

OUTWEIGHTS=*SAS-data-set*

names a SAS data set that contains replicate weights. See the section "Jackknife Method" on page 8396 for information about replicate weights. See the section "Replicate Weights Output Data Set" on page 8401 for more details about the contents of the OUTWEIGHTS= data set.

The OUTWEIGHTS= *method-option* is not available when you provide replicate weights with the REPWEIGHTS statement.

TAYLOR

requests Taylor series variance estimation. This is the default method if you do not specify the VARMETHOD= option or a REPWEIGHTS statement. See the section "Taylor Series (Linearization)" on page 8393 for more information.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC SURVEYREG 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 for the SURVEYREG 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 *variables* ;

The CLASS statement names the classification variables to be used in the model. Typical classification variables are Treatment, Sex, Race, Group, and Replication. If you use the CLASS statement, it must appear before the MODEL statement.

Classification variables can be either character or numeric. By default, class levels are determined from the entire set of formatted values of the CLASS variables.

NOTE: Prior to SAS 9, class levels were determined by using no more than the first 16 characters of the formatted values. To revert to this previous behavior, you can use the TRUNCATE option in the PROC SURVEYREG statement.

In any case, you can use formats to group values into levels. See the discussion of the FORMAT procedure in the *Base SAS Procedures Guide* and the discussions of the FORMAT statement and SAS formats in *SAS Formats and Informats: Reference*. You can adjust the order of CLASS variable levels with the ORDER= option in the PROC SURVEYREG statement.

You can use multiple CLASS statements to specify classification variables.

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 8389 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 8387.

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

Prior to SAS 9, clusters were determined by using no more than the first 16 characters of the formatted values. If you want to revert to this previous behavior, you can use the TRUNCATE option in the PROC SURVEYREG statement.

CONTRAST Statement

CONTRAST *'label' effect values* < */ options* > **;**

CONTRAST *'label' effect values* < *. . . effect values* > < */ options* > **;**

The CONTRAST statement provides custom hypothesis tests for linear combinations of the regression parameters $H_0\colon \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$, where $\mathbf{L}$ is the vector or matrix you specify and $\boldsymbol{\beta}$ is the vector of regression parameters. Thus, to use this feature, you must be familiar with the details of the model parameterization used by PROC SURVEYREG. For information about the parameterization, see the section "GLM Parameterization of Classification Variables and Effects" on page 387 in Chapter 19, "Shared Concepts and Topics."

Each term in the MODEL statement, called an *effect*, is a variable or a combination of variables. You can specify an effect with a variable name or a special notation by using variable names and operators. For more details about how to specify an effect, see the section "Specification of Effects" on page 3495 in Chapter 44, "The GLM Procedure."

For each CONTRAST statement, PROC SURVEYREG computes Wald's F test. The procedure displays this value with the degrees of freedom, and identifies it with the contrast label. The numerator degrees of freedom for Wald's F test equal rank($\mathbf{L}$). The denominator degrees of freedom equal the number of clusters (or the number of observations if there is no CLUSTER statement) minus the number of strata. Alternatively, you can use the DF= option in the MODEL statement to specify the denominator degrees of freedom.

You can specify any number of CONTRAST statements, but they must appear after the MODEL statement.

In the CONTRAST statement,

label identifies the contrast in the output. A label is required for every contrast specified. Labels must be enclosed in single quotes.

effect identifies an effect that appears in the MODEL statement. You can use the INTERCEPT keyword as an effect when an intercept is fitted in the model. You do not need to include all effects that are in the MODEL statement.

values are constants that are elements of $\mathbf{L}$ associated with the effect.

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

E

displays the entire coefficient $\mathbf{L}$ vector or matrix.

NOFILL

requests no filling in higher-order effects. When you specify only certain portions of $\mathbf{L}$, by default PROC SURVEYREG constructs the remaining elements from the context. (For more information, see the section "Specification of ESTIMATE Expressions" on page 3515 in Chapter 44, "The GLM Procedure.")

When you specify the NOFILL option, PROC SURVEYREG does not construct the remaining portions and treats the vector or matrix $\mathbf{L}$ as it is defined in the CONTRAST statement.

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

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

If ABS($\mathbf{l} - \mathbf{lH}$) is greater than c**value*, then $\mathbf{l}\boldsymbol{\beta}$ is declared nonestimable. Here, $\mathbf{H}$ is the matrix $(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}$. The *value* must be between 0 and 1; the default is 10^{-7}.

As stated previously, the CONTRAST statement enables you to perform hypothesis tests $H_0\colon \mathbf{L}\boldsymbol{\beta} = 0$.

If the $\mathbf{L}$ matrix contains more than one contrast, then you can separate the rows of the $\mathbf{L}$ matrix with commas.

For example, for the model

```
proc surveyreg;
   class A B;
   model Y=A B;
run;
```

with A at 5 levels and B at 2 levels, the parameter vector is

$$(\mu \;\; \alpha_1 \;\; \alpha_2 \;\; \alpha_3 \;\; \alpha_4 \;\; \alpha_5 \;\; \beta_1 \;\; \beta_2)$$

To test the hypothesis that the pooled A linear and A quadratic effect is zero, you can use the following $\mathbf{L}$ matrix:

$$\mathbf{L} = \begin{bmatrix} 0 & -2 & -1 & 0 & 1 & 2 & 0 & 0 \\ 0 & 2 & -1 & -2 & -1 & 2 & 0 & 0 \end{bmatrix}$$

The corresponding CONTRAST statement is

```
contrast 'A Linear & Quadratic'
         a -2 -1  0  1  2,
         a  2 -1 -2 -1  2;
```

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.

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

Note that 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. See the section “Domain Analysis” on page 8399 for more details.

Use the DOMAIN statement on the entire data set to perform a domain analysis. Creating a new data set from a single domain and analyzing that with PROC SURVEYREG yields inappropriate estimates of variance.

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.

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 8387.

The formatted values of the domain variables determine the categorical 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.

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 387 in Chapter 19, "Shared Concepts and Topics."

You can specify the following *effect-types*:

COLLECTION is 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 is a classification effect in which the level that is used for a given period corresponds to the level in the preceding period.

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

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

SPLINE is 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 98.4 summarizes the *options* available in the EFFECT statement.

Table 98.4 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
WITHIN=	Names the variable or variables that define the group within which each period is defined
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 397 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 98.5 summarizes the *options* available in the ESTIMATE statement.

Table 98.5 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 for 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 $\mathbf{L}$ matrix
JOINT	Produces a joint F or chi-square test for the estimable functions
SEED=	Specifies the seed for computations that depend on random numbers

For details about the syntax of the ESTIMATE statement, see the section "ESTIMATE Statement" on page 444 in Chapter 19, "Shared Concepts and Topics."

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.

Table 98.6 the summarizes available *options* in the LSMEANS statement.

Table 98.6 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	Requests 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 for multiple-comparison adjustment of LS-means differences
ALPHA=α	Determines the confidence level $(1-\alpha)$
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	Produces a "Lines" display for pairwise LS-means differences
MEANS	Prints the LS-means
PLOTS=	Requests graphs of means and mean comparisons
SEED=	Specifies the seed for computations that depend on random numbers

For details about the syntax of the LSMEANS statement, see the section "LSMEANS Statement" on page 460 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 98.7 summarizes the *options* available in the LSMESTIMATE statement.

Table 98.7 LSMESTIMATE Statement Options

Option	Description
Construction and Computation of LS-Means	
AT	Modifies covariate values in computing LS-means
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 for 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 **L** matrix
ELSM	Prints the **K** matrix
JOINT	Produces a joint *F* or chi-square test for the LS-means and LS-means differences
SEED=	Specifies the seed for computations that depend on random numbers

For details about the syntax of the LSMESTIMATE statement, see the section "LSMESTIMATE Statement" on page 476 in Chapter 19, "Shared Concepts and Topics."

MODEL Statement

MODEL *dependent* **=** < *effects* > < / *options* > **;**

The MODEL statement specifies the dependent (response) variable and the independent (regressor) variables or effects. The dependent variable must be numeric. Each term in a MODEL statement, called an *effect*, is a variable or a combination of variables. You can specify an effect with a variable name or with special notation by using variable names and operators. For more information about how to specify an effect, see the section "Specification of Effects" on page 3495 in Chapter 44, "The GLM Procedure."

Only one MODEL statement is allowed for each PROC SURVEYREG statement. If you specify more than one MODEL statement, the procedure uses the first model and ignores the rest.

Table 98.8 summarizes the *options* available in the MODEL statement.

Table 98.8 MODEL Statement Options

Option	Description
ADJRSQ	Compute the adjusted multiple R-square
ANOVA	Produces the ANOVA table
CLPARM	Requests confidence limits
COVB	Displays the estimated covariance matrix
DEFF	Displays design effects
DF=	Specifies the denominator degrees of freedom
I	Displays the inverse or the generalized inverse of the $\mathbf{X}'\mathbf{X}$ matrix
NOINT	Omits the intercept
PARMLABEL	Displays the labels of the parameters
SINGULAR=	Tunes the estimability checking
SOLUTION	Displays parameter estimates
STB	Displays standardized parameter estimates
VADJUST=	Specifies whether to use degrees of freedom adjustment
X	Displays the $\mathbf{X}'\mathbf{X}$ matrix, or the $\mathbf{X}'\mathbf{W}\mathbf{X}$ matrix

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

ADJRSQ

requests the procedure compute the adjusted multiple R-square.

ANOVA

requests the ANOVA table be produced in the output. By default, the ANOVA table is not printed in the output.

CLPARM

requests confidence limits for the parameter estimates. The SURVEYREG procedure determines the confidence coefficient by using the ALPHA= option, which by default equals 0.05 and produces 95% confidence bounds. The CLPARM option also requests confidence limits for all the estimable linear functions of regression parameters in the ESTIMATE statements.

Note that when there is a CLASS statement, you need to use the SOLUTION option with the CLPARM option to obtain the parameter estimates and their confidence limits.

COVB
displays the estimated covariance matrix of the estimated regression estimates.

DEFF
displays design effects for the regression coefficient estimates.

DF=*value*
specifies the denominator degrees of freedom for the F tests and the degrees of freedom for the t tests. For details about the default denominator degrees of freedom, see the section "Denominator Degrees of Freedom" on page 8397 for details.

I | INVERSE
displays the inverse or the generalized inverse of the $\mathbf{X}'\mathbf{X}$ matrix. When there is a WEIGHT variable, the procedure displays the inverse or the generalized inverse of the $\mathbf{X}'\mathbf{W}\mathbf{X}$ matrix, where $\mathbf{W}$ is the diagonal matrix constructed from WEIGHT variable values.

NOINT
omits the intercept from the model.

PARMLABEL
displays the labels of the parameters in the "Estimated Regression Coefficients" table, if the effect contains a single continuous variable that has a label.

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($\mathbf{l} - \mathbf{l}\mathbf{H}$) is greater than c**value*, then $\mathbf{l}\boldsymbol{\beta}$ is declared nonestimable. Here, $\mathbf{H}$ is the matrix $(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}$. The *value* must be between 0 and 1; the default is 10^{-4}.

SOLUTION
displays a solution to the normal equations, which are the parameter estimates. The SOLUTION option is useful only when you use a CLASS statement. If you do not specify a CLASS statement, PROC SURVEYREG displays parameter estimates by default. But if you specify a CLASS statement, PROC SURVEYREG does not display parameter estimates unless you also specify the SOLUTION option.

STB
produces standardized regression coefficients. A standardized regression coefficient is computed by dividing a parameter estimate by the ratio of the sample standard deviation of the dependent variable to the sample standard deviation of the regressor.

VADJUST=DF | NONE
specifies whether to use degrees of freedom adjustment $(n-1)/(n-p)$ in the computation of the matrix $\mathbf{G}$ for the variance estimation. If you do not specify the VADJUST= option, by default, PROC SURVEYREG uses the degrees-of-freedom adjustment that is equivalent to the VARADJ=DF option. If you do not want to use this variance adjustment, you can specify the VADJUST=NONE option.

X | XPX
displays the $\mathbf{X}'\mathbf{X}$ matrix, or the $\mathbf{X}'\mathbf{W}\mathbf{X}$ matrix when there is a WEIGHT variable, where $\mathbf{W}$ is the diagonal matrix constructed from WEIGHT variable values. The X option also displays the crossproducts vector $\mathbf{X}'\mathbf{y}$ or $\mathbf{X}'\mathbf{W}\mathbf{y}$.

OUTPUT Statement

OUTPUT < **OUT=***SAS-data-set*> < *keyword*< **=***variable-name*> . . . *keyword*< **=***variable-name*> > < / *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 residuals from the linear regression, and the confidence limits for the predictors.

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

OUT=*SAS-data-set*
gives the name of the new output data set. By default, the procedure uses the DATA*n* convention to name the new data set.

keyword < =*variable-name*>
specifies the statistics to include in the output data set and names the new variables that contain the statistics. You can specify a *keyword* for each desired statistic (see the following list of *keywords*). Optionally, you can name a statistic by providing a variable name followed an equal sign to contain the statistic. For example,

```
output out=myOutDataSet p=myPredictor;
```

creates a SAS data set myOutDataSet that contains the predicted values in the variable myPredictor.

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

LCLM | L
lower bound of a $100(1-\alpha)\%$ confidence interval for the expected value (mean) of the predicted value. The α level is equal to the value of the ALPHA= option in the OUTPUT statement or, if this option is not specified, to the ALPHA= option in the PROC SURVEYREG statement. If neither of these options is set, then $\alpha = 0.05$ by default, resulting in the lower bound for a 95% confidence interval. If no variable name is given for this keyword, the default variable name is _LCLM_.

PREDICTED | PRED | P
predicted values. If no variable name is given for this keyword, the default variable name is _PREDICTED_.

RESIDUAL | R
residuals, calculated as ACTUAL – PREDICTED. If no variable name is given for this keyword, the default variable name is _RESIDUAL_.

STDP | STD
standard error of the mean predicted value. If no variable name is given for this keyword, the default variable name is _STD_.

UCLM | U
upper bound of a $100(1-\alpha)\%$ confidence interval for the expected value (mean) of the predicted value. The α level is equal to the value of the ALPHA= option in the OUTPUT statement or, if this option is not specified, to the ALPHA= option in the

PROC SURVEYREG statement. If neither of these options is set, then $\alpha = 0.05$ by default, resulting in the upper bound for a 95% confidence interval. If no variable name is given for this keyword, the default variable name is _UCLM_.

The following *option* is available in the OUTPUT statement and is specified after a slash (/):

ALPHA=α

specifies the level of significance α for $100(1-\alpha)\%$ confidence intervals. By default, α is equal to the value of the ALPHA= option in the PROC SURVEYREG statement or 0.05 if that option is not specified. You can use values between 0 and 1.

REPWEIGHTS Statement

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

The REPWEIGHTS statement names variables that provide replicate weights for BRR or jackknife variance estimation, which you request with the VARMETHOD=BRR or VARMETHOD=JACKKNIFE option in the PROC SURVEYREG statement. If you do not provide replicate weights for these methods by using a REPWEIGHTS statement, then the procedure constructs replicate weights for the analysis. See the sections "Balanced Repeated Replication (BRR) Method" on page 8394 and "Jackknife Method" on page 8396 for information about replicate weights.

Each REPWEIGHTS variable should contain 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.

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 SURVEYREG 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 degrees of freedom equals the number of REPWEIGHTS variables.

JKCOEFS=*value*

specifies a jackknife coefficient for VARMETHOD=JACKKNIFE. The coefficient *value* must be a nonnegative number. See the section "Jackknife Method" on page 8396 for details about jackknife coefficients.

You can use this option to specify a single value of the jackknife coefficient, which the procedure uses for all replicates. To specify different coefficients for different replicates, use the JKCOEFS=*values* or JKCOEFS=*SAS-data-set* option.

JKCOEFS=*values*

specifies jackknife coefficients for VARMETHOD=JACKKNIFE, where each coefficient corresponds to an individual replicate that is identified by a REPWEIGHTS variable. You can separate *values* with blanks or commas. The coefficient *values* must be nonnegative numbers. The number of *values* must equal the number of replicate weight variables named in the REPWEIGHTS statement. List these values in the same order in which you list the corresponding replicate weight variables in the REPWEIGHTS statement.

See the section "Jackknife Method" on page 8396 for details about jackknife coefficients.

To specify different coefficients for different replicates, you can also use the JKCOEFS=*SAS-data-set* option. To specify a single jackknife coefficient for all replicates, use the JKCOEFS=*value* option.

JKCOEFS=*SAS-data-set*

names a SAS data set that contains the jackknife coefficients for VARMETHOD=JACKKNIFE. You provide the jackknife coefficients in the JKCOEFS= data set variable JKCoefficient. Each coefficient value must be a nonnegative number. The observations in the JKCOEFS= data set should correspond to the replicates that are identified by the REPWEIGHTS variables. Arrange the coefficients or observations in the JKCOEFS= data set in the same order in which you list the corresponding replicate weight variables in the REPWEIGHTS statement. The number of observations in the JKCOEFS= data set must not be less than the number of REPWEIGHTS variables.

See the section "Jackknife Method" on page 8396 for details about jackknife coefficients.

To specify different coefficients for different replicates, you can also use the JKCOEFS=*values* option. To specify a single jackknife coefficient for all replicates, use the JKCOEFS=*value* option.

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.

The SLICE statement uses the same *options* as the LSMEANS statement, which are summarized in Table 19.21. For details about the syntax of the SLICE statement, see the section "SLICE Statement" on page 505 in Chapter 19, "Shared Concepts and Topics."

STORE Statement

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

The STORE statement requests that the procedure save 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 with the PLM procedure.

For details about the syntax of the STORE statement, see the section "STORE Statement" on page 508 in Chapter 19, "Shared Concepts and Topics."

STRATA Statement

STRATA *variables* < / *options* > **;**

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 8388 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 8387.

You can use multiple STRATA statements to specify stratum variables.

You can specify the following *options* 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 8404 for more details.

NOCOLLAPSE
: prevents the procedure from collapsing (combining) strata that have only one sampling unit for the Taylor series variance estimation. By default, the procedure collapses strata that contain only one sampling unit for the Taylor series method. See the section "Stratum Collapse" on page 8391 for details.

TEST Statement

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

The TEST statement enables you to perform F tests for model effects that test Type I, Type II, or Type III hypotheses. See Chapter 15, "The Four Types of Estimable Functions," for details about the construction of Type I, II, and III estimable functions.

Table 98.9 summarizes the *options* available in the TEST statement.

Table 98.9 TEST Statement Options

Option	Description
CHISQ	Requests chi-square tests
DDF=	Specifies denominator degrees of freedom for fixed effects
E	Requests Type I, Type II, and Type III coefficients
E1	Requests Type I coefficients
E2	Requests Type II coefficients
E3	Requests Type III coefficients
HTYPE=	Indicates the type of hypothesis test to perform
INTERCEPT	Adds a row that corresponds to the overall intercept

For details about the syntax of the TEST statement, see the section "TEST Statement" on page 509 in Chapter 19, "Shared Concepts and Topics."

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 8387 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 SURVEYREG 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 SURVEYREG assigns all observations a weight of one.

Details: SURVEYREG 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 SURVEYREG. 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 variable, then that observation is excluded from the analysis.

An observation is also excluded from the analysis if it has a missing value for any design (STRATA, CLUSTER, or DOMAIN) variable, unless you specify the MISSING option in the PROC SURVEYREG 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 dependent variable or for any variable 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 sometimes is not true. 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 you specify the NOMCAR option, the procedure always excludes observations with missing or invalid values for the WEIGHT, STRATA, CLUSTER, and DOMAIN variables, unless you specify the MISSING option.

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.

Survey Design Information

Specification of Population Totals and Sampling Rates

To include a finite population correction (*fpc*) in Taylor series variance estimation, you can input either the sampling rate or the population total by using the RATE= or TOTAL= option in the PROC SURVEYREG statement. (You cannot specify both of these options in the same PROC SURVEYREG statement.) The RATE= and TOTAL= options apply only to Taylor series 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 variance estimation does not include a finite population correction. For fairly small sampling fractions, it is appropriate to ignore this correction. See Cochran (1977) and Kish (1965) for more information.

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 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. See the section "Primary Sampling Units (PSUs)" on page 8389 for more details.

For a nonstratified sample design, or for a stratified sample design with 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 with the DATA= option.

The secondary data set must contain all the stratification variables listed in the STRATA statement and all the variables 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 SURVEYREG converts that number to a proportion. The procedure treats the value 1 as 100%, and not the percentage form 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. See the section "Variance Estimation" on page 8393 for more information.

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

If you use a REPWEIGHTS statement to provide replicate weights for BRR or jackknife 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 SURVEYREG treats each observation as a PSU.

Computational Details

Notation

For a stratified clustered sample design, observations are represented by an $n \times (p + 2)$ matrix

$$(\mathbf{w}, \mathbf{y}, \mathbf{X}) = (w_{hij}, y_{hij}, \mathbf{x}_{hij})$$

where

- $\mathbf{w}$ denotes the sampling weight vector
- $\mathbf{y}$ denotes the dependent variable

- $\mathbf{X}$ denotes the $n \times p$ design matrix. (When an effect contains only classification variables, the columns of $\mathbf{X}$ that correspond this effect contain only 0s and 1s; no reparameterization is made.)
- $h = 1, 2, \ldots, H$ 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
- p is the total number of parameters (including an intercept if the INTERCEPT effect is included in the MODEL statement)
- $n = \sum_{h=1}^{H} \sum_{i=1}^{n_h} m_{hi}$ is the total number of observations in the sample

Also, f_h denotes the sampling rate for stratum h. You can use the TOTAL= or RATE= option to input population totals or sampling rates. See the section "Specification of Population Totals and Sampling Rates" on page 8388 for details. If you input stratum totals, PROC SURVEYREG computes f_h as the ratio of the stratum sample size to the stratum total. If you input stratum sampling rates, PROC SURVEYREG uses these values directly for f_h. If you do not specify the TOTAL= or RATE= option, then the procedure assumes that the stratum sampling rates f_h are negligible, and a finite population correction is not used when computing variances.

Regression Coefficients

PROC SURVEYREG solves the normal equations $\mathbf{X}'\mathbf{W}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{W}\mathbf{y}$ by using a modified sweep routine that produces a generalized (g2) inverse $(\mathbf{X}'\mathbf{W}\mathbf{X})^-$ and a solution (Pringle and Rayner 1971)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^-\mathbf{X}'\mathbf{W}\mathbf{y}$$

where $\mathbf{W}$ is the diagonal matrix constructed from WEIGHT variable values.

For models with CLASS variables, there are more design matrix columns than there are degrees of freedom (*df*) for the effect. Thus, there are linear dependencies among the columns. In this case, the parameters are not estimable; there is an infinite number of least squares solutions. PROC SURVEYREG uses a generalized (g2) inverse to obtain values for the estimates. The solution values are not displayed unless you specify the SOLUTION option in the MODEL statement. The solution has the characteristic that estimates are zero whenever the design column for that parameter is a linear combination of previous columns. (In strict terms, the solution values should not be called estimates.) With this full parameterization, hypothesis tests are constructed to test linear functions of the parameters that are estimable.

Design Effect

If you specify the DEFF option in the MODEL statement, PROC SURVEYREG calculates the design effects for the regression coefficients. The design effect of an estimate is the ratio of the actual variance to the variance computed under the assumption of simple random sampling:

$$\text{DEFF} = \frac{\text{variance under the sample design}}{\text{variance under simple random sampling}}$$

See Kish (1965, p. 258) for more details. PROC SURVEYREG computes the numerator as described in the section "Variance Estimation" on page 8393. And the denominator is computed under the assumption that the sample design is simple random sampling, with no stratification and no clustering.

To compute the variance under the assumption of simple random sampling, PROC SURVEYREG calculates the sampling rate as follows. If you specify both sampling weights and sampling rates (or population totals) for the analysis, then the sampling rate under simple random sampling is calculated as

$$f_{\mathrm{SRS}} = n \;/\; w_{\cdots}$$

where n is the sample size and $w_{\cdots}$ (the sum of the weights over all observations) estimates the population size. If the sum of the weights is less than the sample size, f_{SRS} is set to zero. If you specify sampling rates for the analysis but not sampling weights, then PROC SURVEYREG computes the sampling rate under simple random sampling as the average of the stratum sampling rates:

$$f_{\mathrm{SRS}} = \frac{1}{H} \sum_{h=1}^{H} f_h$$

If you do not specify sampling rates (or population totals) for the analysis, then the sampling rate under simple random sampling is assumed to be zero:

$$f_{\mathrm{SRS}} = 0$$

Stratum Collapse

If there is only one sampling unit in a stratum, then PROC SURVEYREG cannot estimate the variance for this stratum for the Taylor series method. To estimate stratum variances, by default the procedure collapses, or combines, those strata that contain only one sampling unit. If you specify the NOCOLLAPSE option in the STRATA statement, PROC SURVEYREG does not collapse strata and uses a variance estimate of zero for any stratum that contains only one sampling unit.

Note that stratum collapse only applies to Taylor series variance estimation (the default method, also specified by VARMETHOD=TAYLOR). The procedure does not collapse strata for BRR or jackknife variance estimation, which you request with the VARMETHOD=BRR or VARMETHOD=JACKKNIFE option.

If you do not specify the NOCOLLAPSE option for the Taylor series method, PROC SURVEYREG collapses strata according to the following rules. If there are multiple strata that contain only one sampling unit each, then the procedure collapses, or combines, all these strata into a new pooled stratum. If there is only one stratum with a single sampling unit, then PROC SURVEYREG collapses that stratum with the preceding stratum, where strata are ordered by the STRATA variable values. If the stratum with one sampling unit is the first stratum, then the procedure combines it with the following stratum.

If you specify stratum sampling rates by using the RATE=*SAS-data-set* option, PROC SURVEYREG computes the sampling rate for the new pooled stratum as the weighted average of the sampling rates for the collapsed strata. See the section "Computational Details" on page 8389 for details. If the specified sampling rate equals 0 for any of the collapsed strata, then the pooled stratum is assigned a sampling rate of 0. If you specify stratum totals by using the TOTAL=*SAS-data-set* option, PROC SURVEYREG combines the totals for the collapsed strata to compute the sampling rate for the new pooled stratum.

Sampling Rate of the Pooled Stratum from Collapse

Assuming that PROC SURVEYREG collapses single-unit strata $h_1, h_2, \ldots, h_c$ into the pooled stratum, the procedure calculates the sampling rate for the pooled stratum as

$$f_{\text{Pooled Stratum}} = \begin{cases} 0 & \text{if any of } f_{h_l} = 0 \text{ where } l = 1, 2, \ldots, c \\ \left(\sum_{l=1}^{c} n_{h_l} f_{h_l}^{-1} \right)^{-1} \sum_{l=1}^{c} n_{h_l} & \text{otherwise} \end{cases}$$

Analysis of Variance (ANOVA)

PROC SURVEYREG produces an analysis of variance table for the model specified in the MODEL statement. This table is identical to the one produced by the GLM procedure for the model. PROC SURVEYREG computes ANOVA table entries by using the sampling weights, but not the sample design information about stratification and clustering.

The degrees of freedom (*df*) displayed in the ANOVA table are the same as those in the ANOVA table produced by PROC GLM. The Total DF is the total degrees of freedom used to obtain the regression coefficient estimates. The Total DF equals the total number of observations minus 1 if the model includes an intercept. If the model does not include an intercept, the Total DF equals the total number of observations. The Model DF equals the degrees of freedom for the effects in the MODEL statement, not including the intercept. The Error DF equals the Total DF minus the Model DF.

Multiple R-Square

PROC SURVEYREG computes a multiple R-square for the weighted regression as

$$R^2 = 1 - \frac{\mathrm{SS}_{error}}{\mathrm{SS}_{total}}$$

where SS_{error} is the error sum of squares in the ANOVA table

$$\mathrm{SS}_{error} = \mathbf{r}'\mathbf{W}\mathbf{r}$$

and SS_{total} is the total sum of squares

$$\mathrm{SS}_{total} = \begin{cases} \mathbf{y}'\mathbf{W}\mathbf{y} & \text{if no intercept} \\ \mathbf{y}'\mathbf{W}\mathbf{y} - \left(\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} y_{hij} \right)^2 / \, w_{\ldots} & \text{otherwise} \end{cases}$$

where $w_{\ldots}$ is the sum of the sampling weights over all observations.

Adjusted R-Square

If you specify the ADJRSQ option in the MODEL statement, PROC SURVEYREG computes an multiple R-square adjusted as the weighted regression as

$$\mathrm{ADJRSQ} = \begin{cases} 1 - \dfrac{n(1 - R^2)}{n - p} & \text{if no intercept} \\ 1 - \dfrac{(n-1)(1 - R^2)}{n - p} & \text{otherwise} \end{cases}$$

where R^2 is the multiple R-square.

Root Mean Square Errors

PROC SURVEYREG computes the square root of mean square errors as

$$\sqrt{\text{MSE}} = \sqrt{n\ \text{SS}_{error}\ /\ (n-p)\ w_{\dots}}$$

where $w_{\dots}$ is the sum of the sampling weights over all observations.

Variance Estimation

PROC SURVEYREG uses the Taylor series method or replication (resampling) methods to estimate sampling errors of estimators based on complex sample designs (Fuller 2009; Woodruff 1971; Fuller 1975; Fuller et al. 1989; Särndal, Swensson, and Wretman 1992; 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 and the *jackknife* method. For each replicate, the original weights are modified for the PSUs in the replicates 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 SURVEYREG.

Use the notation described in the section "Notation" on page 8389 to denote the residuals from the linear regression as

$$\mathbf{r} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

with r_{hij} as its elements. Let the $p \times p$ matrix $\mathbf{G}$ be defined as

$$\mathbf{G} = \frac{n-1}{n-p} \sum_{h=1}^{H} \frac{n_h(1-f_h)}{n_h - 1} \sum_{i=1}^{n_h} (\mathrm{e}_{hi\cdot} - \bar{\mathrm{e}}_{h\cdot\cdot})'(\mathrm{e}_{hi\cdot} - \bar{\mathrm{e}}_{h\cdot\cdot})$$

where

$$\mathbf{e}_{hij} = w_{hij} r_{hij} \mathbf{x}_{hij}$$
$$\mathbf{e}_{hi\cdot} = \sum_{j=1}^{m_{hi}} \mathbf{e}_{hij}$$
$$\bar{\mathbf{e}}_{h\cdot\cdot} = \frac{1}{n_h} \sum_{i=1}^{n_h} \mathbf{e}_{hi\cdot}$$

The Taylor series estimate of the covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\widehat{\mathbf{V}}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-}\mathbf{G}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}$$

The factor $(n-1)/(n-p)$ in the computation of the matrix $\mathbf{G}$ reduces the small sample bias associated with using the estimated function to calculate deviations (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 use this adjustment in the variance estimation. If you do not want to use this multiplier in variance estimation, you can specify the VADJUST=NONE option in the MODEL statement to suppress this factor.

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* × *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 ($r = 1, 2, ..., R$) 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 details 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 $\hat{\boldsymbol{\beta}}$ be the estimated regression coefficients from the full sample for $\boldsymbol{\beta}$, and let $\hat{\boldsymbol{\beta}}_r$ be the estimated regression coefficient from the rth replicate by using replicate weights. PROC SURVEYREG estimates the covariance matrix of $\hat{\boldsymbol{\beta}}$ by

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

with H degrees of freedom, where H is the number of strata.

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 $0 \leq \epsilon < 1$ 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 $0 \leq \epsilon < 1$ can be set by specifying the FAY = ϵ *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 details, 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 ($r = 1, 2, ..., R$) 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 $\hat{\boldsymbol{\beta}}$ be the estimated regression coefficients from the full sample for $\boldsymbol{\beta}$. Let $\hat{\boldsymbol{\beta}}_r$ be the estimated regression coefficient obtained from the rth replicate by using replicate weights. PROC SURVEYREG estimates the covariance matrix of $\hat{\boldsymbol{\beta}}$ by

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

with H degrees of freedom, where H is the number of strata.

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:

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

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 ($r = 1, 2, ..., R$), 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

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

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 ($r = 1, 2, ..., R$), 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 with a REPWEIGHTS statement, then you can also provide corresponding jackknife coefficients with the JKCOEFS= option.

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

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

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.

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:

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

Degrees of Freedom

PROC SURVEYREG produces tests for the significance of model effects, regression parameters, estimable functions specified in the ESTIMATE statement, and contrasts specified in the CONTRAST statement. It computes all these tests taking into account the sample design. The degrees of freedom for these tests differ from the degrees of freedom for the ANOVA table, which does not consider the sample design.

Denominator Degrees of Freedom

The denominator *df* refers to the denominator degrees of freedom for *F* tests and to the degrees of freedom for *t* tests in the analysis.

For the Taylor series method, the denominator *df* equals the number of clusters minus the actual number of strata. If there are no clusters, the denominator *df* equals the number of observations minus the actual number of strata. The *actual number of strata* equals the following:

- one, if there is no STRATA statement
- the number of strata in the sample, if there is a STRATA statement but the procedure does not collapse any strata
- the number of strata in the sample after collapsing, if there is a STRATA statement and the procedure collapses strata that have only one sampling unit

Alternatively, you can specify your own denominator *df* by using the DF= option in the MODEL statement.

For the BRR method (including Fay's method) without a REPWEIGHTS statement, the denominator *df* equals the number of strata.

For the jackknife method without a REPWEIGHTS statement, the denominator *df* is equal to the number of replicates minus the *actual number of strata*.

When there is a REPWEIGHTS statement, the denominator *df* equals the number of REPWEIGHTS variables, unless you specify an alternative in the DF= option in a REPWEIGHTS statement.

Numerator Degrees of Freedom

The numerator *df* refers to the numerator degrees of freedom for the Wald *F* statistic associated with an effect or with a contrast. The procedure computes the Wald *F* statistic for an effect as a Type III test; that is, the test has the following properties:

- The hypothesis for an effect does not involve parameters of other effects except for containing effects (which it must involve to be estimable).
- The hypotheses to be tested are invariant to the ordering of effects in the model.

See the section "Testing Effects" on page 8398 for more information. The numerator *df* for the Wald *F* statistic for a contrast is the rank of the **L** matrix that defines the contrast.

Testing

Testing Effects

For each effect in the model, PROC SURVEYREG computes an **L** matrix such that every element of $\mathbf{L}\boldsymbol{\beta}$ is estimable; the **L** matrix has the maximum possible rank that is associated with the effect. To test the effect, the procedure uses the Wald *F* statistic for the hypothesis $H_0\colon \mathbf{L}\boldsymbol{\beta} = 0$. The Wald *F* statistic equals

$$F_{\text{Wald}} = \frac{(\mathbf{L}\hat{\boldsymbol{\beta}})'(\mathbf{L}'\widehat{\mathbf{V}}\mathbf{L})^{-1}(\mathbf{L}\hat{\boldsymbol{\beta}})}{\text{rank}(\mathbf{L}'\widehat{\mathbf{V}}\mathbf{L})}$$

with numerator degrees of freedom equal to $\text{rank}(\mathbf{L}'\widehat{\mathbf{V}}\mathbf{L})$.

In the Taylor series method, the denominator degrees of freedom is equal to the number of clusters minus the number of strata (unless you specify the denominator degrees of freedom with the DF= option in the MODEL statement). For details about denominator degrees of freedom in replication methods, see the section "Denominator Degrees of Freedom" on page 8397. It is possible that the **L** matrix cannot be constructed for an effect, in which case that effect is not testable. For more information about how the matrix **L** is constructed, see the discussion in Chapter 15, "The Four Types of Estimable Functions."

You can use the TEST statement to perform F tests that test Type I, Type II, or Type III hypotheses. For details about the syntax of the TEST statement, see the section "TEST Statement" on page 509 in Chapter 19, "Shared Concepts and Topics."

Contrasts

You can use the CONTRAST statement to perform custom hypothesis tests. If the hypothesis is testable in the univariate case, the Wald F statistic for $H_0 : \mathbf{L}\boldsymbol{\beta} = 0$ is computed as

$$F_{\text{Wald}} = \frac{(\mathbf{L}_{\text{Full}}\hat{\boldsymbol{\beta}})'(\mathbf{L}'_{\text{Full}}\widehat{\mathbf{V}}\mathbf{L}_{\text{Full}})^{-1}(\mathbf{L}_{\text{Full}}\hat{\boldsymbol{\beta}})}{\text{rank}(\mathbf{L})}$$

where $\mathbf{L}$ is the contrast vector or matrix you specify, $\boldsymbol{\beta}$ is the vector of regression parameters, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}\mathbf{Y}$, $\widehat{\mathbf{V}}$ is the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$, rank($\mathbf{L}$) is the rank of $\mathbf{L}$, and $\mathbf{L}_{\text{Full}}$ is a matrix such that

- $\mathbf{L}_{\text{Full}}$ has the same number of columns as $\mathbf{L}$
- $\mathbf{L}_{\text{Full}}$ has full row rank
- the rank of $\mathbf{L}_{\text{Full}}$ equals the rank of the $\mathbf{L}$ matrix
- all rows of $\mathbf{L}_{\text{Full}}$ are estimable functions
- the Wald F statistic computed using the $\mathbf{L}_{\text{Full}}$ matrix is equivalent to the Wald F statistic computed by using the $\mathbf{L}$ matrix with any row deleted that is a linear combination of previous rows

If $\mathbf{L}$ is a full-rank matrix and all rows of $\mathbf{L}$ are estimable functions, then $\mathbf{L}_{\text{Full}}$ is the same as $\mathbf{L}$. It is possible that $\mathbf{L}_{\text{Full}}$ matrix cannot be constructed for contrasts in a CONTRAST statement, in which case the contrasts are not testable.

Domain Analysis

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

For a domain D, let I_D be the corresponding indicator variable:

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

Let

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

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

Computational Resources

Due to the complex nature of survey data analysis, the SURVEYREG procedure requires more memory than an analysis of the same regression model by the GLM procedure. For details about the amount of memory related to the modeling, see the section “Computational Resources” on page 3549 in Chapter 44, “The GLM Procedure.”

The memory needed by the SURVEYREG procedure to handle the survey design is described as follows.

Let

- H be the total number of strata
- n_c be the total number of clusters in your sample across all H strata, if you specify a CLUSTER statement
- p be the total number of parameters in the model

The memory needed (in bytes) is

$$48H + 8pH + 4p(p+1)H$$

For a cluster sample, the additional memory needed (in bytes) is

$$48H + 8pH + 4p(p+1)H + 4p(p+1)n_c + 16n_c$$

The SURVEYREG procedure also uses other small amounts of additional memory. However, when you have a large number of clusters or strata, or a large number of parameters in your model, the memory described previously dominates the total memory required by the procedure.

Output Data Sets

You can use the Output Delivery System (ODS) to create a SAS data set from any piece of PROC SURVEYREG output. See the section “ODS Table Names” on page 8406 for more information. For a more detailed description of using ODS, see Chapter 20, “Using the Output Delivery System.”

PROC SURVEYREG also provides an OUTPUT statement to create a data set that contains estimated linear predictors and their standard error estimates, the residuals from the linear regression, and the confidence limits for the predictors.

If you use BRR or jackknife variance estimation, PROC SURVEYREG 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 Created by the OUTPUT Statement

The OUTPUT statement produces an output data set that contains the following:

- all original data from the SAS data set input to PROC SURVEYREG
- the new variables corresponding to the diagnostic measures specified with statistics *keywords* in the OUTPUT statement (PREDICTED=, RESIDUAL=, and so on)

When any independent variable in the analysis (including all classification variables) is missing for an observation, then all new variables that correspond to diagnostic measures are missing for the observation in the output data set.

When a dependent variable in the analysis is missing for an observation, then the residual variable that corresponds to R is also missing in the output data set. However, the variables corresponding to LCLM, P, STDP, and UCLM are not missing.

Replicate Weights Output Data Set

If you specify the OUTWEIGHTS= *method-option* for VARMETHOD=BRR or VARMETHOD=JACKKNIFE, PROC SURVEYREG 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 the following variables:

- all variables in the DATA= input data set
- RepWt_1, RepWt_2, . . ., RepWt_n, which are the replicate weight variables

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 zero for those observations 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 SURVEYREG or in the 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 SURVEYREG 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 SURVEYREG 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 SURVEYREG procedure produces output that is described in the following sections.

Output that is generated by the EFFECT, ESTIMATE, LSMEANS, LSMESTIMATE, and SLICE statements is 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."

Data Summary

By default, PROC SURVEYREG displays the following information in the "Data Summary" table:

- Number of Observations, which is the total number of observations used in the analysis, excluding observations with missing values
- Sum of Weights, if you specify a WEIGHT statement
- Mean of the dependent variable in the MODEL statement, or Weighted Mean if you specify a WEIGHT statement
- Sum of the dependent variable in the MODEL statement, or Weighted Sum if you specify a WEIGHT statement

Design Summary

When you specify a CLUSTER statement or a STRATA statement, the procedure displays a "Design Summary" table, which provides the following sample design information:

- Number of Strata, if you specify a STRATA statement
- Number of Strata Collapsed, if the procedure collapses strata
- Number of Clusters, if you specify a CLUSTER statement
- Overall Sampling Rate used to calculate the design effect, if you specify the DEFF option in the MODEL statement

Domain Summary

By default, PROC SURVEYREG displays the following information in the "Domain Summary" table:

- Number of Observations, which is the total number of observations used in the analysis
- total number of observations in the current domain
- total number of observations not in the current domain
- Sum of Weights for the observations in the current domain, if you specify a WEIGHT statement

Fit Statistics

By default, PROC SURVEYREG displays the following regression statistics in the "Fit Statistics" table:

- R-square for the regression
- Root MSE, which is the square root of the mean square error
- Denominator DF, which is the denominator degrees of freedom for the *F* tests and also the degrees of freedom for the *t* tests produced by the procedure

Variance Estimation

If the variance method is not Taylor series (see the section "Variance Estimation" on page 8393) or if the NOMCAR option is used, by default, PROC SURVEYREG displays the following variance estimation information in the "Variance Estimation" table:

- Method, which is the variance estimation method
- Number of Replicates, if you specify the VARMETHOD=BRR or VARMETHOD=JACKKNIFE option
- 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 provide replicate weights with a REPWEIGHTS statement
- Missing Levels, which indicates whether missing levels of categorical variables are included by the MISSING option
- Missing Values, which indicates whether observations with missing values are included in the analysis by the NOMCAR option

Stratum Information

When you specify the LIST option in the STRATA statement, PROC SURVEYREG 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= option or the 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
- Collapsed, which has the value 'Yes' if the stratum is collapsed with another stratum before analysis

If PROC SURVEYREG collapses strata, the "Stratum Information" table also displays stratum information for the new, collapsed stratum. The new stratum has a Stratum Index of 0 and is labeled 'Pooled.'

Class Level Information

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

- CLASS Variable, which lists each CLASS variable name
- Levels, which is the number of values or levels of the classification variable
- Values, 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.

X'X Matrix

If you specify the XPX option in the MODEL statement, PROC SURVEYREG displays the $\mathbf{X}'\mathbf{X}$ matrix. When there is a WEIGHT variable, the procedure displays the $\mathbf{X}'\mathbf{W}\mathbf{X}$ matrix. This option also displays the crossproducts vector $\mathbf{X}'\mathbf{y}$ or $\mathbf{X}'\mathbf{W}\mathbf{y}$, where $\mathbf{y}$ is the response vector (dependent variable).

Inverse Matrix of X'X

If you specify the INVERSE option in the MODEL statement, PROC SURVEYREG displays the inverse or the generalized inverse of the $\mathbf{X}'\mathbf{X}$ matrix. When there is a WEIGHT variable, the procedure displays the inverse or the generalized inverse of the $\mathbf{X}'\mathbf{W}\mathbf{X}$ matrix.

ANOVA for Dependent Variable

If you specify the ANOVA option in the model statement, PROC SURVEYREG displays an analysis of variance table for the dependent variable. This table is identical to the ANOVA table displayed by the GLM procedure.

Tests of Model Effects

By default, PROC SURVEYREG displays a “Tests of Model Effects” table, which provides Wald’s F test for each effect in the model. The table contains the following information for each effect:

- Effect, which is the effect name
- Num DF, which is the numerator degrees of freedom for Wald’s F test
- F Value, which is Wald’s F statistic
- Pr > F, which is the significance probability corresponding to the F Value

A footnote displays the denominator degrees of freedom, which is the same for all effects.

Estimated Regression Coefficients

PROC SURVEYREG displays the “Estimated Regression Coefficients” table by default when there is no CLASS statement. Also, the procedure displays this table when you specify a CLASS statement and also specify the SOLUTION option in the MODEL statement. This table contains the following information for each regression parameter:

- Parameter, which identifies the effect or regressor variable
- Estimate, which is the estimate of the regression coefficient
- Standardized Estimate, which is the standardized regression coefficient
- Standard Error, which is the standard error of the estimate
- t Value, which is the t statistic for testing $H_0\colon \text{Parameter} = 0$
- Pr > | t |, which is the two-sided significance probability corresponding to the t Value
- Pr > | t |, which is the two-sided significance probability corresponding to the t Value

Covariance of Estimated Regression Coefficients

When you specify the COVB option in the MODEL statement, PROC SURVEYREG displays the “Covariance of Estimated Regression Coefficients” matrix.

Coefficients of Contrast

When you specify the E option in a CONTRAST statement, PROC SURVEYREG displays a “Coefficients of Contrast” table for the contrast. You can use this table to check the coefficients you specified in the CONTRAST statement. Also, this table gives a note for a nonestimable contrast.

Analysis of Contrasts

If you specify a CONTRAST statement, PROC SURVEYREG produces an “Analysis of Contrasts” table, which displays Wald’s F test for the contrast. If you use more than one CONTRAST statement, the procedure displays all results in the same table. The “Analysis of Contrasts” table contains the following information for each contrast:

- Contrast, which is the label of the contrast
- Num DF, which is the numerator degrees of freedom for Wald’s F test
- F Value, which is Wald’s F statistic for testing H_0: Contrast $= 0$
- Pr > F, which is the significance probability corresponding to the F Value

Hadamard Matrix

If you specify the VARMETHOD=BRR(PRINTH) *method-option* in the PROC SURVEYREG statement, the procedure displays the Hadamard matrix.

When you provide a Hadamard matrix with the VARMETHOD=BRR(HADAMARD=) *method-option* but the procedure does not use the entire matrix, the procedure displays only the rows and columns that are actually used to construct replicates.

ODS Table Names

PROC SURVEYREG assigns a name to each table it creates; these names are listed in Table 98.10. You can use these names to refer to tables when you use the Output Delivery System (ODS) to select tables and create output data sets. For more information about ODS, see Chapter 20, “Using the Output Delivery System.”

To improve the consistency among procedures, tables that are generated by the ESTIMATE statements are changed slightly in appearance and formatting compared to releases prior to SAS/STAT 9.22. However, the statistics in the “Estimates” table remain unchanged. The Coef table replaces the previous EstimateCoef table that displays the **L** matrix coefficients of an estimable function of the parameters.

The EFFECT, ESTIMATE, LSMEANS, LSMESTIMATE, and SLICE statements also create tables, which are not listed in Table 98.10. For information about these tables, see the corresponding sections of Chapter 19, “Shared Concepts and Topics.”

Table 98.10 ODS Tables Produced by PROC SURVEYREG

ODS Table Name	Description	Statement	Option
ANOVA	ANOVA for dependent variable	MODEL	ANOVA
ClassVarInfo	Class level information	CLASS	Default
ContrastCoef	Coefficients of contrast	CONTRAST	E
Contrasts	Analysis of contrasts	CONTRAST	Default
CovB	Covariance of estimated regression coefficients	MODEL	COVB
DataSummary	Data summary	PROC	Default
DesignSummary	Design summary	STRATA | CLUSTER	Default
DomainSummary	Domain summary	DOMAIN	Default
Effects	Tests of model effects	MODEL	Defect
FitStatistics	Fit statistics	MODEL	Default
HadamardMatrix	Hadamard matrix	PROC	PRINTH
InvXPX	Inverse matrix of $\mathbf{X}'\mathbf{X}$	MODEL	I
ParameterEstimates	Estimated regression coefficients	MODEL	SOLUTION
StrataInfo	Stratum information	STRATA	LIST
VarianceEstimation	Variance estimation	PROC	Default
XPX	$\mathbf{X}'\mathbf{X}$ matrix	MODEL	XPX

By referring to the names of such tables, you can use the ODS OUTPUT statement to place one or more of these tables in output data sets.

For example, the following statements create an output data set MyStrata, which contains the StrataInfo table, an output data set MyParmEst, which contains the ParameterEstimates table, and an output data set Cov, which contains the CovB table for the ice cream study discussed in the section "Stratified Sampling" on page 8358:

```
title1 'Ice Cream Spending Analysis';
title2 'Stratified Sample Design';
proc surveyreg data=IceCream total=StudentTotals;
   strata Grade /list;
   class Kids;
   model Spending = Income Kids / solution covb;
   weight Weight;
   ods output StrataInfo = MyStrata
              ParameterEstimates = MyParmEst
              CovB = Cov;
run;
```

Note that the option CovB is specified in the MODEL statement in order to produce the covariance matrix table.

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 606 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 605 in Chapter 21, “Statistical Graphics Using ODS.”

When ODS Graphics is enabled, 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.”

When ODS Graphics is enabled and when the regression model depends on at most one continuous variable as a regressor, excluding the intercept, the PLOTS= option in the PROC SURVEYREG statement controls fit plots for the regression.

PROC SURVEYREG provides a bubble plot or a heat map for model fitting. You can request a specific type of presentation of the weights by specifying the PLOTS(WEIGHT=) global plot option to request either a bubble plot or a heat map plot of the data that overlays the regression line and confidence limits band of the prediction in a fit plot. If you do not specify this option, the default plot depends on the number of observations in your data. That is, for a data set that contains 100 observations or less, the default is a bubble plot, in which the bubble area is proportional to the sampling weight of an observation. For a data set that contains more than 100 observations, the default is a heat map, in which the color of heat represents the sum of weights at the corresponding location.

PROC SURVEYREG assigns a name to each graph that it creates using ODS Graphics. You can use the name to refer to the graph. Table 98.11 lists the name of the graph that PROC SURVEYREG generates, together with its description and the PLOTS= option *plot-request* that produces it.

Table 98.11 ODS Graphs Produced by PROC SURVEYREG

ODS Graph Name	Description	PLOTS= Option
FitPlot	Regression line and confidence limits band of the prediction overlaid on a bubble plot or a heat map of the data	FIT

Examples: SURVEYREG Procedure

Example 98.1: Simple Random Sampling

This example investigates the relationship between the labor force participation rate (LFPR) of women in 1968 and 1972 in large cities in the United States. A simple random sample of 19 cities is drawn from a total of 200 cities. For each selected city, the LFPRs are recorded and saved in a SAS data set Labor. In the following DATA step, LFPR in 1972 is contained in the variable LFPR1972, and the LFPR in 1968 is identified by the variable LFPR1968:

```
data Labor;
   input City $ 1-16 LFPR1972 LFPR1968;
   datalines;
New York         .45      .42
Los Angeles      .50      .50
Chicago          .52      .52
Philadelphia     .45      .45
Detroit          .46      .43
San Francisco    .55      .55
Boston           .60      .45
Pittsburgh       .49      .34
St. Louis        .35      .45
Connecticut      .55      .54
Washington D.C. .52      .42
Cincinnati       .53      .51
Baltimore        .57      .49
Newark           .53      .54
Minn/St. Paul    .59      .50
Buffalo          .64      .58
Houston          .50      .49
Patterson        .57      .56
Dallas           .64      .63
;
```

Assume that the LFPRs in 1968 and 1972 have a linear relationship, as shown in the following model:

$$\mathsf{LFPR1972} = \beta_0 + \beta_1 * \mathsf{LFPR1968} + \text{error}$$

You can use PROC SURVEYREG to obtain the estimated regression coefficients and estimated standard errors of the regression coefficients. The following statements perform the regression analysis:

```
ods graphics on;
title 'Study of Labor Force Participation Rates of Women';
proc surveyreg data=Labor total=200;
   model LFPR1972 = LFPR1968;
run;
ods graphics off;
```

Here, the TOTAL=200 option specifies the finite population total from which the simple random sample of 19 cities is drawn. You can specify the same information by using the sampling rate option RATE=0.095 (19/200=.095).

Output 98.1.1 summarizes the data information and the fit information.

Output 98.1.1 Summary of Regression Using Simple Random Sampling

```
Study of Labor Force Participation Rates of Women

The SURVEYREG Procedure

Regression Analysis for Dependent Variable LFPR1972

Data Summary

Number of Observations           19
Mean of LFPR1972            0.52684
Sum of LFPR1972            10.01000


Fit Statistics

R-Square              0.3970
Root MSE             0.05657
Denominator DF            18
```

Output 98.1.2 presents the significance tests for the model effects and estimated regression coefficients. The *F* tests and *t* tests for the effects in the model are also presented in these tables.

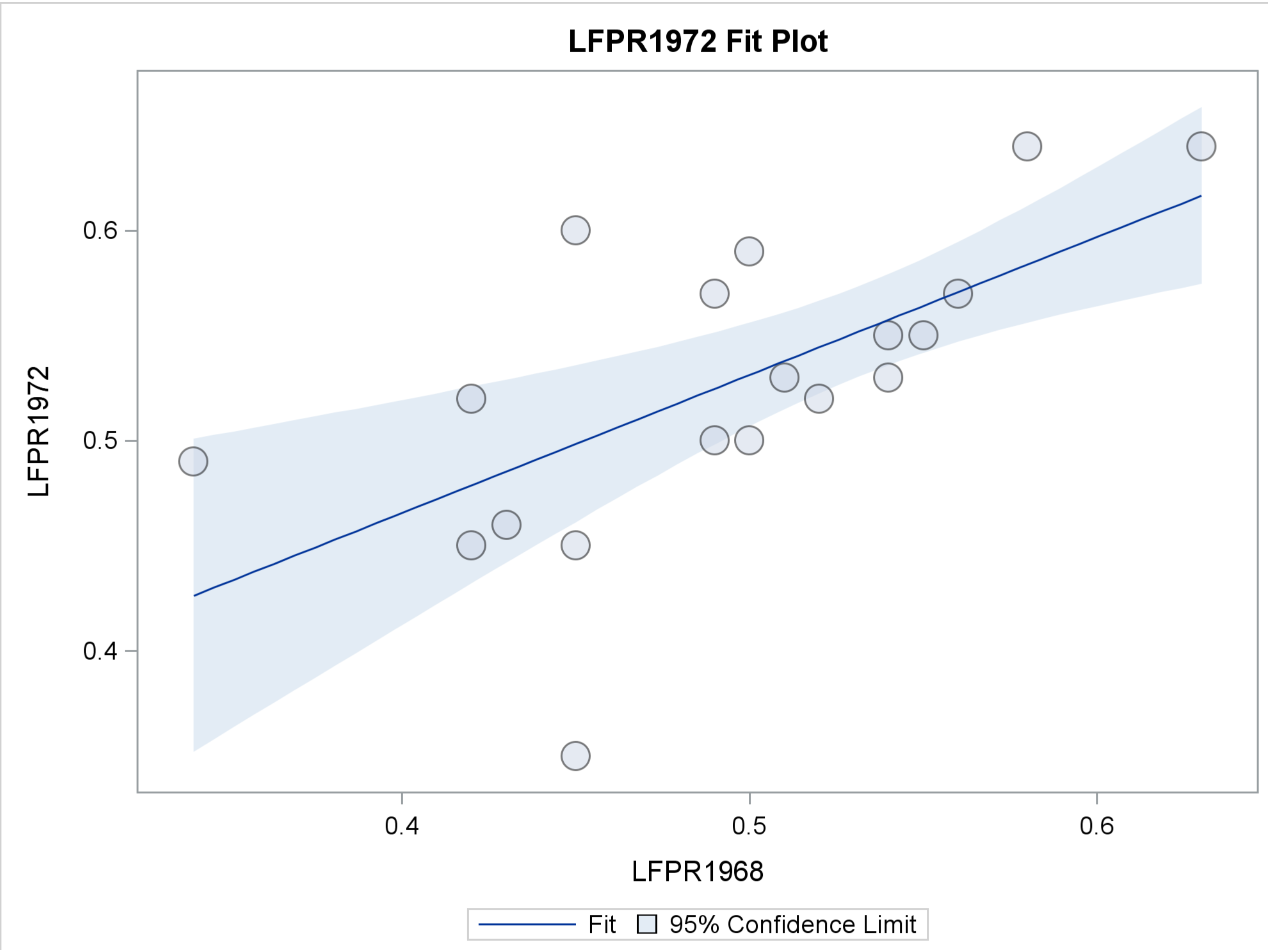
Output 98.1.2 Regression Coefficient Estimates

```
Tests of Model Effects

Effect        Num DF    F Value    Pr > F

Model              1      13.84    0.0016
Intercept          1       4.63    0.0452
LFPR1968           1      13.84    0.0016

NOTE: The denominator degrees of freedom for the F tests is 18.


Estimated Regression Coefficients

                                Standard
Parameter      Estimate            Error    t Value    Pr > |t|

Intercept    0.20331056       0.09444296       2.15      0.0452
LFPR1968     0.65604048       0.17635810       3.72      0.0016

NOTE: The denominator degrees of freedom for the t tests is 18.
```

From the regression performed by PROC SURVEYREG, you obtain a positive estimated slope for the linear relationship between the LFPR in 1968 and the LFPR in 1972. The regression coefficients are all significant at the 5% level. The effects Intercept and LFPR1968 are significant in the model at the 5% level. In this example, the *F* test for the overall model without intercept is the same as the effect LFPR1968.

When ODS graphics is enabled and you have only one regressor in the model, PROC SURVEYREG displays a plot of the model fitting, which is shown in Figure 98.1.3.

Output 98.1.3 Regression Fitting



Example 98.2: Cluster Sampling

This example illustrates the use of regression analysis in a simple random cluster sample design. The data are from Särndal, Swensson, and Wretman (1992, p. 652). A total of 284 Swedish municipalities are grouped into 50 clusters of neighboring municipalities. Five clusters with a total of 32 municipalities are randomly selected. The results from the regression analysis in which clusters are used in the sample design are compared to the results of a regression analysis that ignores the clusters. The linear relationship between the population in 1975 and in 1985 is investigated.

The 32 selected municipalities in the sample are saved in the data set Municipalities:

```
data Municipalities;
   input Municipality Cluster Population85 Population75;
   datalines;
205   37    5    5
206   37   11   11
207   37   13   13
208   37    8    8
209   37   17   19
  6    2   16   15
  7    2   70   62
  8    2   66   54
  9    2   12   12
 10    2   60   50
 94   17    7    7
 95   17   16   16
 96   17   13   11
 97   17   12   11
 98   17   70   67
 99   17   20   20
100   17   31   28
101   17   49   48
276   50    6    7
277   50    9   10
278   50   24   26
279   50   10    9
280   50   67   64
281   50   39   35
282   50   29   27
283   50   10    9
284   50   27   31
 52   10    7    6
 53   10    9    8
 54   10   28   27
 55   10   12   11
 56   10  107  108
;
```

The variable Municipality identifies the municipalities in the sample; the variable Cluster indicates the cluster to which a municipality belongs; and the variables Population85 and Population75 contain the municipality populations in 1985 and in 1975 (in thousands), respectively. A regression analysis is performed by PROC SURVEYREG with a CLUSTER statement:

```
title1 'Regression Analysis for Swedish Municipalities';
title2 'Cluster Sampling';
proc surveyreg data=Municipalities total=50;
   cluster Cluster;
   model Population85=Population75;
run;
```

The TOTAL=50 option specifies the total number of clusters in the sampling frame.

Output 98.2.1 displays the data and design summary. Since the sample design includes clusters, the procedure displays the total number of clusters in the sample in the “Design Summary” table.

Output 98.2.1 Regression Analysis for Cluster Sampling

```
              Regression Analysis for Swedish Municipalities
                             Cluster Sampling

                         The SURVEYREG Procedure

          Regression Analysis for Dependent Variable Population85

                               Data Summary

                    Number of Observations              32
                    Mean of Population85          27.50000
                    Sum of Population85          880.00000


                              Design Summary

                      Number of Clusters              5
```

Output 98.2.2 displays the fit statistics and regression coefficient estimates. In the "Estimated Regression Coefficients" table, the estimated slope for the linear relationship is 1.05, which is significant at the 5% level; but the intercept is not significant. This suggests that a regression line crossing the original can be established between populations in 1975 and in 1985.

Output 98.2.2 Regression Analysis for Cluster Sampling

```
                              Fit Statistics

                         R-Square            0.9860
                         Root MSE            3.0488
                         Denominator DF           4


                    Estimated Regression Coefficients

                                            Standard
         Parameter         Estimate            Error    t Value    Pr > |t|

         Intercept       -0.0191292       0.89204053      -0.02      0.9839
         Population75     1.0546253       0.05167565      20.41      <.0001

          NOTE: The denominator degrees of freedom for the t tests is 4.
```

The CLUSTER statement is necessary in PROC SURVEYREG in order to incorporate the sample design. If you do not specify a CLUSTER statement in the regression analysis, as in the following statements, the standard deviation of the regression coefficients are incorrectly estimated.

```
title1 'Regression Analysis for Swedish Municipalities';
title2 'Simple Random Sampling';
proc surveyreg data=Municipalities total=284;
   model Population85=Population75;
run;
```

The analysis ignores the clusters in the sample, assuming that the sample design is a simple random sampling. Therefore, the TOTAL= option specifies the total number of municipalities, which is 284.

Output 98.2.3 displays the regression results ignoring the clusters. Compared to the results in Output 98.2.2, the regression coefficient estimates are the same. However, without using clusters, the regression coefficients have a smaller variance estimate, as in Output 98.2.3. By using clusters in the analysis, the estimated regression coefficient for effect Population75 is 1.05, with the estimated standard error 0.05, as displayed in Output 98.2.2; without using the clusters, the estimate is 1.05, but with the estimated standard error 0.04, as displayed in Output 98.2.3. To estimate the variance of the regression coefficients correctly, you should include the clustering information in the regression analysis.

Output 98.2.3 Regression Analysis for Simple Random Sampling

```
                 Regression Analysis for Swedish Municipalities
                            Simple Random Sampling

                           The SURVEYREG Procedure

             Regression Analysis for Dependent Variable Population85

                                Data Summary

                    Number of Observations            32
                    Mean of Population85        27.50000
                    Sum of Population85        880.00000


                               Fit Statistics

                         R-Square              0.9860
                         Root MSE              3.0488
                         Denominator DF            31


                      Estimated Regression Coefficients

                                          Standard
        Parameter          Estimate          Error    t Value    Pr > |t|

        Intercept        -0.0191292     0.67417606      -0.03      0.9775
        Population75      1.0546253     0.03668414      28.75      <.0001

        NOTE: The denominator degrees of freedom for the t tests is 31.
```

Example 98.3: Regression Estimator for Simple Random Sample

By using auxiliary information, you can construct regression estimators to provide more accurate estimates of population characteristics. With ESTIMATE statements in PROC SURVEYREG, you can specify a regression estimator as a linear function of the regression parameters to estimate the population total. This example illustrates this application by using the data set Municipalities from Example 98.2.

In this sample, a linear model between the Swedish populations in 1975 and in 1985 is established:

$$\text{Population85} = \alpha + \beta * \text{Population75} + \text{error}$$

Assuming that the total population in 1975 is known to be 8200 (in thousands), you can use the ESTIMATE statement to predict the 1985 total population by using the following statements:

```
title1 'Regression Analysis for Swedish Municipalities';
title2 'Estimate Total Population';
proc surveyreg data=Municipalities total=50;
   cluster Cluster;
   model Population85=Population75;
   estimate '1985 population' Intercept 284 Population75 8200;
run;
```

Since each observation in the sample is a municipality and there is a total of 284 municipalities in Sweden, the coefficient for Intercept (α) in the ESTIMATE statement is 284 and the coefficient for Population75 (β) is the total population in 1975 (8.2 million).

Output 98.3.1 displays the regression results and the estimation of the total population. By using the linear model, you can predict the total population in 1985 to be 8.64 million, with a standard error of 0.26 million.

Output 98.3.1 Use the Regression Estimator to Estimate the Population Total

Regression Analysis for Swedish Municipalities
Estimate Total Population

The SURVEYREG Procedure

Regression Analysis for Dependent Variable Population85

Estimate

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
1985 population	8642.49	258.56	4	33.43	<.0001

Example 98.4: Stratified Sampling

This example illustrates the use of the SURVEYREG procedure to perform a regression in a stratified sample design. Consider a population of 235 farms producing corn in Nebraska and Iowa. You are interested in the relationship between corn yield (CornYield) and total farm size (FarmArea).

Each state is divided into several regions, and each region is used as a stratum. Within each stratum, a simple random sample with replacement is drawn. A total of 19 farms is selected by using a stratified simple random sample. The sample size and population size within each stratum are displayed in Table 98.12.

Table 98.12 Number of Farms in Each Stratum

Stratum	State	Region	Number of Farms Population	Sample
1	Iowa	1	100	3
2		2	50	5
3		3	15	3
4	Nebraska	1	30	6
5		2	40	2
Total			**235**	**19**

The following three models are considered:

- Model I — Common intercept and slope:

$$\text{Corn Yield} = \alpha + \beta * \text{Farm Area}$$

- Model II — Common intercept, different slope:

$$\text{Corn Yield} = \begin{cases} \alpha + \beta_{\text{Iowa}} * \text{Farm Area} & \text{if the farm is in Iowa} \\ \alpha + \beta_{\text{Nebraska}} * \text{Farm Area} & \text{if the farm is in Nebraska} \end{cases}$$

- Model III — Different intercept and different slope:

$$\text{Corn Yield} = \begin{cases} \alpha_{\text{Iowa}} + \beta_{\text{Iowa}} * \text{Farm Area} & \text{if the farm is in Iowa} \\ \alpha_{\text{Nebraska}} + \beta_{\text{Nebraska}} * \text{Farm Area} & \text{if the farm is in Nebraska} \end{cases}$$

Data from the stratified sample are saved in the SAS data set Farms. The variable Weight contains the sampling weights, which are reciprocals of the selection probabilities.

```
data Farms;
   input State $ Region FarmArea CornYield Weight;
   datalines;
Iowa      1 100  54 33.333
Iowa      1  83  25 33.333
Iowa      1  25  10 33.333
Iowa      2 120  83 10.000
Iowa      2  50  35 10.000
Iowa      2 110  65 10.000
Iowa      2  60  35 10.000
Iowa      2  45  20 10.000
Iowa      3  23   5  5.000
Iowa      3  10   8  5.000
Iowa      3 350 125  5.000
Nebraska 1 130  20  5.000
Nebraska 1 245  25  5.000
Nebraska 1 150  33  5.000
Nebraska 1 263  50  5.000
Nebraska 1 320  47  5.000
```

```
Nebraska 1 204  25  5.000
Nebraska 2  80  11 20.000
Nebraska 2  48   8 20.000
;
```

The SAS data set StratumTotals contains the stratum population sizes.

```
data StratumTotals;
   input State $ Region _TOTAL_;
   datalines;
Iowa     1 100
Iowa     2  50
Iowa     3  15
Nebraska 1  30
Nebraska 2  40
;
```

Using the sample data from the data set Farms and the control information data from the data set StratumTotals, you can fit Model I by using the following statements in PROC SURVEYREG:

```
ods graphics on;
title1 'Analysis of Farm Area and Corn Yield';
title2 'Model I: Same Intercept and Slope';
proc surveyreg data=Farms total=StratumTotals;
   strata State Region / list;
   model CornYield = FarmArea / covB;
   weight Weight;
run;
ods graphics off;
```

Output 98.4.1 displays the data summary and stratification information fitting Model I. The sampling rates are automatically computed by the procedure based on the sample sizes and the population totals in strata.

Output 98.4.1 Data Summary and Stratum Information Fitting Model I

```
                 Analysis of Farm Area and Corn Yield
                  Model I: Same Intercept and Slope

                       The SURVEYREG Procedure

          Regression Analysis for Dependent Variable CornYield

                            Data Summary

                Number of Observations                19
                Sum of Weights                 234.99900
                Weighted Mean of CornYield      31.56029
                Weighted Sum of CornYield         7416.6


                           Design Summary

                     Number of Strata             5
```

Output 98.4.1 *continued*

Fit Statistics

R-Square	0.3882
Root MSE	20.6422
Denominator DF	14

Stratum Information

Stratum Index	State	Region	N Obs	Population Total	Sampling Rate
1	Iowa	1	3	100	3.00%
2		2	5	50	10.0%
3		3	3	15	20.0%
4	Nebraska	1	6	30	20.0%
5		2	2	40	5.00%

Output 98.4.2 displays tests of model effects and the estimated regression coefficients.

Output 98.4.2 Estimated Regression Coefficients and the Estimated Covariance Matrix

Tests of Model Effects

Effect	Num DF	F Value	Pr > F
Model	1	21.74	0.0004
Intercept	1	4.93	0.0433
FarmArea	1	21.74	0.0004

NOTE: The denominator degrees of freedom for the F tests is 14.

Estimated Regression Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	11.8162978	5.31981027	2.22	0.0433
FarmArea	0.2126576	0.04560949	4.66	0.0004

NOTE: The denominator degrees of freedom for the t tests is 14.

Covariance of Estimated Regression Coefficients

	Intercept	FarmArea
Intercept	28.300381277	-0.146471538
FarmArea	-0.146471538	0.0020802259

Output 98.4.3 Regression Fitting

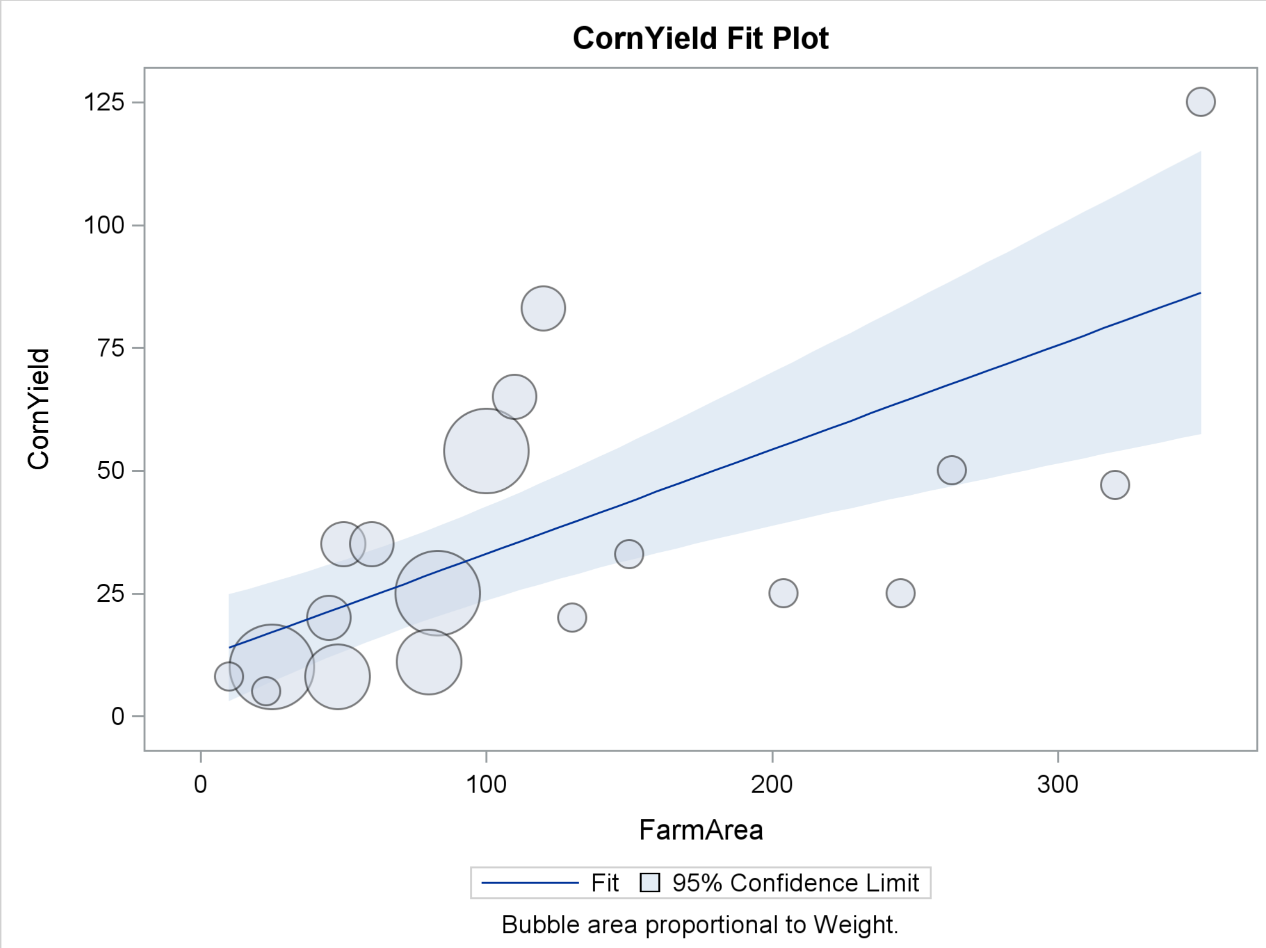


Figure 98.4.3 displays the fit of the regression.

Alternatively, you can assume that the linear relationship between corn yield (CornYield) and farm area (FarmArea) is different among the states (Model II). In order to analyze the data by using this model, you create auxiliary variables FarmAreaNE and FarmAreaIA to represent farm area in different states:

$$\mathsf{FarmAreaNE} = \begin{cases} 0 & \text{if the farm is in Iowa} \\ \mathsf{FarmArea} & \text{if the farm is in Nebraska} \end{cases}$$

$$\mathsf{FarmAreaIA} = \begin{cases} \mathsf{FarmArea} & \text{if the farm is in Iowa} \\ 0 & \text{if the farm is in Nebraska} \end{cases}$$

The following statements create these variables in a new data set called FarmsByState and use PROC SURVEYREG to fit Model II:

```
data FarmsByState;
   set Farms;
   if State='Iowa' then do;
      FarmAreaIA=FarmArea;
      FarmAreaNE=0;
   end;

   else do;
      FarmAreaIA=0;
      FarmAreaNE=FarmArea;
   end;
run;
```

The following statements perform the regression by using the new data set FarmsByState. The analysis uses the auxiliary variables FarmAreaIA and FarmAreaNE as the regressors:

```
title1 'Analysis of Farm Area and Corn Yield';
title2 'Model II: Same Intercept, Different Slopes';
proc surveyreg data=FarmsByState total=StratumTotals;
   strata State Region;
   model CornYield = FarmAreaIA FarmAreaNE / covB;
   weight Weight;
run;
```

Output 98.4.4 displays the fit statistics and parameter estimates. The estimated slope parameters for each state are quite different from the estimated slope in Model I. The results from the regression show that Model II fits these data better than Model I.

Output 98.4.4 Regression Results from Fitting Model II

Analysis of Farm Area and Corn Yield
Model II: Same Intercept, Different Slopes

The SURVEYREG Procedure

Regression Analysis for Dependent Variable CornYield

Fit Statistics

R-Square	0.8158
Root MSE	11.6759
Denominator DF	14

Estimated Regression Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	4.04234816	3.80934848	1.06	0.3066
FarmAreaIA	0.41696069	0.05971129	6.98	<.0001
FarmAreaNE	0.12851012	0.02495495	5.15	0.0001

NOTE: The denominator degrees of freedom for the t tests is 14.

Output 98.4.4 *continued*

Covariance of Estimated Regression Coefficients

	Intercept	FarmAreaIA	FarmAreaNE
Intercept	14.511135861	-0.118001232	-0.079908772
FarmAreaIA	-0.118001232	0.0035654381	0.0006501109
FarmAreaNE	-0.079908772	0.0006501109	0.0006227496

For Model III, different intercepts are used for the linear relationship in two states. The following statements illustrate the use of the NOINT option in the MODEL statement associated with the CLASS statement to fit Model III:

```
title1 'Analysis of Farm Area and Corn Yield';
title2 'Model III: Different Intercepts and Slopes';
proc surveyreg data=FarmsByState total=StratumTotals;
   strata State Region;
   class State;
   model CornYield = State FarmAreaIA FarmAreaNE / noint covB solution;
   weight Weight;
run;
```

The model statement includes the classification effect State as a regressor. Therefore, the parameter estimates for effect State present the intercepts in two states.

Output 98.4.5 displays the regression results for fitting Model III, including parameter estimates, and covariance matrix of the regression coefficients. The estimated covariance matrix shows a lack of correlation between the regression coefficients from different states. This suggests that Model III might be the best choice for building a model for farm area and corn yield in these two states.

However, some statistics remain the same under different regression models—for example, Weighted Mean of CornYield. These estimators do not rely on the particular model you use.

Output 98.4.5 Regression Results for Fitting Model III

Analysis of Farm Area and Corn Yield
Model III: Different Intercepts and Slopes

The SURVEYREG Procedure

Regression Analysis for Dependent Variable CornYield

Fit Statistics

R-Square	0.9300
Root MSE	11.9810
Denominator DF	14

Output 98.4.5 *continued*

```
                Estimated Regression Coefficients

                                       Standard
Parameter              Estimate           Error    t Value    Pr > |t|

State Iowa           5.27797099      5.27170400       1.00      0.3337
State Nebraska       0.65275201      1.70031616       0.38      0.7068
FarmAreaIA           0.40680971      0.06458426       6.30      <.0001
FarmAreaNE           0.14630563      0.01997085       7.33      <.0001

 NOTE: The denominator degrees of freedom for the t tests is 14.


         Covariance of Estimated Regression Coefficients

                                        State
                  State Iowa         Nebraska       FarmAreaIA       FarmAreaNE

State Iowa      27.790863033                0     -0.205517205                0
State Nebraska             0     2.8910750385                0     -0.027354011
FarmAreaIA      -0.205517205                0     0.0041711265                0
FarmAreaNE                 0     -0.027354011                0     0.0003988349
```

Example 98.5: Regression Estimator for Stratified Sample

This example uses the corn yield data set FARMS from Example 98.4 to illustrate how to construct a regression estimator for a stratified sample design.

As in Example 98.3, by incorporating auxiliary information into a regression estimator, the procedure can produce more accurate estimates of the population characteristics that are of interest. In this example, the sample design is a stratified sample design. The auxiliary information is the total farm areas in regions of each state, as displayed in Table 98.13. You want to estimate the total corn yield by using this information under the three linear models given in Example 98.4.

Table 98.13 Information for Each Stratum

			Number of Farms		
Stratum	**State**	**Region**	**Population**	**Sample**	**Total Farm Area**
1	Iowa	1	100	3	
2		2	50	5	13,200
3		3	15	3	
4	Nebraska	1	30	6	8,750
5		2	40	2	
Total			**235**	**19**	**21,950**

The regression estimator to estimate the total corn yield under Model I can be obtained by using PROC SURVEYREG with an ESTIMATE statement:

```
title1 'Estimate Corn Yield from Farm Size';
title2 'Model I: Same Intercept and Slope';
proc surveyreg data=Farms total=StratumTotals;
   strata State Region / list;
   class  State Region;
   model  CornYield = FarmArea State*Region /solution;
   weight Weight;
   estimate 'Estimate of CornYield under Model I'
           INTERCEPT 235 FarmArea 21950
           State*Region 100 50 15 30 40 /e;
run;
```

To apply the constraint in each stratum that the weighted total number of farms equals to the total number of farms in the stratum, you can include the strata as an effect in the MODEL statement, effect State*Region. Thus, the CLASS statement must list the STRATA variables, State and Region, as classification variables. The following ESTIMATE statement specifies the regression estimator, which is a linear function of the regression parameters:

```
   estimate 'Estimate of CornYield under Model I'
           INTERCEPT 235 FarmArea 21950
           State*Region 100 50 15 30 40 /e;
```

This linear function contains the total for each explanatory variable in the model. Because the sampling units are farms in this example, the coefficient for Intercept in the ESTIMATE statement is the total number of farms (235); the coefficient for FarmArea is the total farm area listed in Table 98.13 (21950); and the coefficients for effect State*Region are the total number of farms in each strata (as displayed in Table 98.13).

Output 98.5.1 displays the results of the ESTIMATE statement. The regression estimator for the total of CornYield in Iowa and Nebraska is 7464 under Model I, with a standard error of 927.

Output 98.5.1 Regression Estimator for the Total of CornYield under Model I

```
                   Estimate Corn Yield from Farm Size
                   Model I: Same Intercept and Slope

                        The SURVEYREG Procedure

            Regression Analysis for Dependent Variable CornYield

                               Estimate

                                                 Standard
Label                                 Estimate      Error       DF    t Value

Estimate of CornYield under Model I    7463.52     926.84       14       8.05

                               Estimate

                Label                                 Pr > |t|

                Estimate of CornYield under Model I     <.0001
```

Under Model II, a regression estimator for totals can be obtained by using the following statements:

```
title1 'Estimate Corn Yield from Farm Size';
title2 'Model II: Same Intercept, Different Slopes';
proc surveyreg data=FarmsByState total=StratumTotals;
   strata State Region;
   class  State Region;
   model  CornYield = FarmAreaIA FarmAreaNE
                      state*region /solution;
   weight Weight;
   estimate 'Total of CornYield under Model II'
           INTERCEPT 235 FarmAreaIA 13200 FarmAreaNE 8750
           State*Region 100 50 15 30 40 /e;
run;
```

In this model, you also need to include strata as a fixed effect in the MODEL statement. Other regressors are the auxiliary variables FarmAreaIA and FarmAreaNE (defined in Example 98.4). In the following ESTIMATE statement, the coefficient for Intercept is still the total number of farms; and the coefficients for FarmAreaIA and FarmAreaNE are the total farm area in Iowa and Nebraska, respectively, as displayed in Table 98.13. The total number of farms in each strata are the coefficients for the strata effect:

```
   estimate 'Total of CornYield under Model II'
           INTERCEPT 235 FarmAreaIA 13200 FarmAreaNE 8750
           State*Region 100 50 15 30 40 /e;
```

Output 98.5.2 displays that the results of the regression estimator for the total of corn yield in two states under Model II is 7580 with a standard error of 859. The regression estimator under Model II has a slightly smaller standard error than under Model I.

Output 98.5.2 Regression Estimator for the Total of CornYield under Model II

Estimate Corn Yield from Farm Size
Model II: Same Intercept, Different Slopes

The SURVEYREG Procedure

Regression Analysis for Dependent Variable CornYield

Estimate

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
Total of CornYield under Model II	7580.49	859.18	14	8.82	<.0001

Finally, you can apply Model III to the data and estimate the total corn yield. Under Model III, you can also obtain the regression estimators for the total corn yield for each state. Three ESTIMATE statements are used in the following statements to create the three regression estimators:

```
title1 'Estimate Corn Yield from Farm Size';
title2 'Model III: Different Intercepts and Slopes';
proc surveyreg data=FarmsByState total=StratumTotals;
   strata State Region;
```

```
      class  State Region;
      model  CornYield = state FarmAreaIA FarmAreaNE
         State*Region /noint solution;
      weight Weight;
      estimate 'Total CornYield in Iowa under Model III'
                State 165 0 FarmAreaIA 13200 FarmAreaNE  0
                State*region 100 50 15  0  0 /e;
      estimate 'Total CornYield in Nebraska under Model III'
                State 0 70 FarmAreaIA 0 FarmAreaNE 8750
                State*Region 0 0 0 30 40 /e;
      estimate 'Total CornYield in both states under Model III'
                State 165 70 FarmAreaIA 13200 FarmAreaNE 8750
                State*Region 100 50 15 30 40 /e;
   run;
```

The fixed effect State is added to the MODEL statement to obtain different intercepts in different states, by using the NOINT option. Among the ESTIMATE statements, the coefficients for explanatory variables are different depending on which regression estimator is estimated. For example, in the ESTIMATE statement

```
      estimate 'Total CornYield in Iowa under Model III'
                State 165 0 FarmAreaIA 13200 FarmAreaNE  0
                State*region 100 50 15  0  0 /e;
```

the coefficients for the effect State are 165 and 0, respectively. This indicates that the total number of farms in Iowa is 165 and the total number of farms in Nebraska is 0, because the estimation is the total corn yield in Iowa only. Similarly, the total numbers of farms in three regions in Iowa are used for the coefficients of the strata effect State*Region, as displayed in Table 98.13.

Output 98.5.3 displays the results from the three regression estimators by using Model III. Since the estimations are independent in each state, the total corn yield from both states is equal to the sum of the estimated total of corn yield in Iowa and Nebraska, $6246 + 1334 = 7580$. This regression estimator is the same as the one under Model II. The variance of regression estimator of the total corn yield in both states is the sum of variances of regression estimators for total corn yield in each state. Therefore, it is not necessary to use Model III to obtain the regression estimator for the total corn yield unless you need to estimate the total corn yield for each individual state.

Output 98.5.3 Regression Estimator for the Total of CornYield under Model III

```
                    Estimate Corn Yield from Farm Size
                 Model III: Different Intercepts and Slopes

                          The SURVEYREG Procedure

             Regression Analysis for Dependent Variable CornYield

                                 Estimate

                                                   Standard
Label                                  Estimate       Error      DF   t Value

Total CornYield in Iowa under Model III   6246.11      851.27      14      7.34

                                  Estimate

               Label                                  Pr > |t|

               Total CornYield in Iowa under Model III     <.0001
```

Example 98.6: Stratum Collapse

In a stratified sample, it is possible that some strata might have only one sampling unit. When this happens, PROC SURVEYREG collapses the strata that contain a single sampling unit into a pooled stratum. For more detailed information about stratum collapse, see the section "Stratum Collapse" on page 8391.

Suppose that you have the following data:

```
data Sample;
   input Stratum X Y W;
   datalines;
10 0 0 5
10 1 1 5
11 1 1 10
11 1 2 10
12 3 3 16
33 4 4 45
14 6 7 50
12 3 4 16
;
```

The variable Stratum is again the stratification variable, the variable X is the independent variable, and the variable Y is the dependent variable. You want to regress Y on X. In the data set Sample, both Stratum=33 and Stratum=14 contain one observation. By default, PROC SURVEYREG collapses these strata into one pooled stratum in the regression analysis.

To input the finite population correction information, you create the SAS data set StratumTotals:

```
data StratumTotals;
   input Stratum _TOTAL_;
   datalines;
10 10
11 20
12 32
33 40
33 45
14 50
15  .
66 70
;
```

The variable Stratum is the stratification variable, and the variable _TOTAL_ contains the stratum totals. The data set StratumTotals contains more strata than the data set Sample. Also in the data set StratumTotals, more than one observation contains the stratum totals for Stratum=33:

```
33 40
33 45
```

PROC SURVEYREG allows this type of input. The procedure simply ignores strata that are not present in the data set Sample; for the multiple entries of a stratum, the procedure uses the first observation. In this example, Stratum=33 has the stratum total _TOTAL_=40.

The following SAS statements perform the regression analysis:

```
title1 'Stratified Sample with Single Sampling Unit in Strata';
title2 'With Stratum Collapse';
proc surveyreg data=Sample total=StratumTotals;
   strata Stratum/list;
   model Y=X;
   weight W;
run;
```

Output 98.6.1 shows that there are a total of five strata in the input data set and two strata are collapsed into a pooled stratum. The denominator degrees of freedom is 4, due to the collapse (see the section "Denominator Degrees of Freedom" on page 8397).

Output 98.6.1 Summary of Data and Regression

```
            Stratified Sample with Single Sampling Unit in Strata
                           With Stratum Collapse

                          The SURVEYREG Procedure

               Regression Analysis for Dependent Variable Y

                               Data Summary

                   Number of Observations               8
                   Sum of Weights               157.00000
                   Weighted Mean of Y             4.31210
                   Weighted Sum of Y            677.00000


                              Design Summary

                 Number of Strata                          5
                 Number of Strata Collapsed                2


                              Fit Statistics

                        R-Square            0.9564
                        Root MSE            0.5111
                        Denominator DF           4
```

Output 98.6.2 displays the stratification information, including stratum collapse. Under the column Collapsed, the fourth stratum (Stratum=14) and the fifth (Stratum=33) are marked as 'Yes,' which indicates that these two strata are collapsed into the pooled stratum (Stratum Index=0). The sampling rate for the pooled stratum is 2% (see the section "Sampling Rate of the Pooled Stratum from Collapse" on page 8392).

Output 98.6.3 displays the parameter estimates and the tests of the significance of the model effects.

Output 98.6.2 Stratification Information

Stratum Information

Stratum Index	Collapsed	Stratum	N Obs	Population Total	Sampling Rate
1		10	2	10	20.0%
2		11	2	20	10.0%
3		12	2	32	6.25%
4	Yes	14	1	50	2.00%
5	Yes	33	1	40	2.50%
0	Pooled		2	90	2.22%

NOTE: Strata with only one observation are collapsed into the stratum with Stratum Index "0".

Output 98.6.3 Parameter Estimates and Effect Tests

Tests of Model Effects

Effect	Num DF	F Value	Pr > F
Model	1	173.01	0.0002
Intercept	1	0.00	0.9961
X	1	173.01	0.0002

NOTE: The denominator degrees of freedom for the F tests is 4.

Estimated Regression Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.00179469	0.34306373	0.01	0.9961
X	1.12598708	0.08560466	13.15	0.0002

NOTE: The denominator degrees of freedom for the t tests is 4.

Alternatively, if you prefer not to collapse strata with a single sampling unit, you can specify the NOCOLLAPSE option in the STRATA statement:

```
title1 'Stratified Sample with Single Sampling Unit in Strata';
title2 'Without Stratum Collapse';
proc surveyreg data=Sample total=StratumTotals;
   strata Stratum/list nocollapse;
   model Y = X;
   weight W;
run;
```

Output 98.6.4 does not contain the stratum collapse information displayed in Output 98.6.1, and the denominator degrees of freedom are 3 instead of 4.

Output 98.6.4 Summary of Data and Regression

```
     Stratified Sample with Single Sampling Unit in Strata
                  Without Stratum Collapse

                  The SURVEYREG Procedure

          Regression Analysis for Dependent Variable Y

                        Data Summary

              Number of Observations              8
              Sum of Weights              157.00000
              Weighted Mean of Y            4.31210
              Weighted Sum of Y           677.00000


                       Design Summary

                 Number of Strata             5


                       Fit Statistics

                   R-Square            0.9564
                   Root MSE            0.5111
                   Denominator DF           3
```

In Output 98.6.5, although the fourth stratum and the fifth stratum contain only one observation, no stratum collapse occurs.

Output 98.6.5 Stratification Information

Stratum Information

Stratum Index	Stratum	N Obs	Population Total	Sampling Rate
1	10	2	10	20.0%
2	11	2	20	10.0%
3	12	2	32	6.25%
4	14	1	50	2.00%
5	33	1	40	2.50%

As a result of not collapsing strata, the standard error estimates of the parameters, shown in Output 98.6.6, are different from those in Output 98.6.3, as are the tests of the significance of model effects.

Output 98.6.6 Parameter Estimates and Effect Tests

```
                     Tests of Model Effects

           Effect        Num DF     F Value     Pr > F

           Model              1      347.27     0.0003
           Intercept          1        0.00     0.9962
           X                  1      347.27     0.0003

NOTE: The denominator degrees of freedom for the F tests is 3.


              Estimated Regression Coefficients

                                 Standard
 Parameter       Estimate           Error     t Value     Pr > |t|

 Intercept     0.00179469     0.34302581        0.01       0.9962
 X             1.12598708     0.06042241       18.64       0.0003

NOTE: The denominator degrees of freedom for the t tests is 3.
```

Example 98.7: Domain Analysis

You can use PROC SURVEYREG to perform domain analysis in a subgroup of your interest. To illustrate, this example uses a data set from the National Health and Nutrition Examination Survey I (NHANES I) Epidemiologic Followup Study (NHEFS), described in Example 97.2 in Chapter 97, "The SURVEYPHREG Procedure."

The NHEFS is a national longitudinal survey that is conducted by the National Center for Health Statistics, the National Institute on Aging, and some other agencies of the Public Health Service in the United States. Some important objectives of this survey are to determine the relationships between clinical, nutritional, and behavioral factors; to determine the relationship between mortality and hospital utilization; and to monitor changes in risk factors for the initial cohort that represents the NHANES I population. A cohort of size 14,407, which includes all persons 25 to 74 years old who completed a medical examination at NHANES I in 1971–1975, was selected for the NHEFS. Personal interviews were conducted for every selected unit during the first wave of data collection from the year 1982 to 1984. Follow-up studies were conducted in 1986, 1987, and 1992. In the year 1986, only nondeceased persons 55 to 74 years old (as reported in the base year survey) were interviewed. The 1987 and 1992 NHEFS contain the entire nondeceased NHEFS cohort. Vital and tracing status data, interview data, health care facility stay data, and mortality data for all four waves are available for public use. See `http://www.cdc.gov/nchs/nhanes/nhefs/nhefs.htm` for more information about the survey and the data sets.

For illustration purposes, 1,018 observations from the 1987 NHEFS public use interview data are used to create the data set cancer. The observations are obtained from 10 strata that contain 596 PSUs. The sum of observation weights for these selected units is over 19 million. Observation weights range from 359 to 129,359 with a mean of 18,747.69 and a median of 11,414.

The following variables are used in this example:

- ObsNo, unit identification
- Strata, stratum identification
- PSU, identification for primary sampling units
- ObservationWt, sampling weight associated with each unit
- Age, the event-time variable, defined as follows:
  - age of the subject when the first cancer was reported for subjects with reported cancer
  - age of the subject at death for deceased subjects without reported cancer
  - age of the subject as reported in 1987 follow-up (this value is used for nondeceased subjects who never reported cancer)
  - age of the subject for the entry year 1971–1975 survey if the subject has cancer (or is deceased) but the date of incident is not reported
- Cancer, cancer indicator (1 = cancer reported, 0 = cancer not reported)
- BodyWeight, body weight of the subject as reported in the 1987 follow-up, or an imputed body weight based on the subject's age in the entry year 1971–1975 survey

The following SAS statements create the data set cancer. Note that BodyWeight for a few observations (8%) is imputed based on Age by using a deterministic regression imputation model (Särndal and Lundström (2005, chapter 12)). The imputed values are treated as observed values in this example. In other words, this example treats the data set Cancer as the observed data set.

```
data cancer;
   input ObsNo Strata PSU ObservationWt Age Cancer BodyWeight;
   datalines;
   1  3  002  3805   53  1  175
   2  3  002  6107   77  0  175
   3  3  039  2968   50  0  160
   4  3  084  30438  52  0  145
   5  3  007  5081   80  0  127
   6  3  009  3891   62  0  180
   7  3  009  3580   50  0  157
   8  3  022  2968   56  0  142
   9  3  050  23748  60  0  140
  10  3  060  48264  69  0  168

   ... more lines ...

1016  4  002   2689  40  0  120
1017  4  092  45888  52  0  166
1018  4  035   4347  58  0  156
;
```

Suppose you want to study how aging affects body weight in the subgroup of cancer patients for the base year survey population. Because whether an individual has cancer or not is unrelated to the design of the sample, this kind of analysis is called domain analysis (subgroup analysis).

The following statements request a linear regression of BodyWeight on Age among cancer patients. The STRATA, CLUSTER, and WEIGHT statements identify the variance strata, PSUs, and analysis weights, respectively. The DOMAIN statement defines the subgroups of people who have been diagnosed with cancer and people who do not have cancer. The ODS SELECT statement requests that PROC SURVEYREG display only the analysis in the subgroup Cancer = 1 in the output.

```
title1 'Study of Body Weight and Age among Cancer Patients';
ods graphics on;
proc surveyreg data=cancer plot=fit;
   strata strata;
   cluster psu;
   weight ObservationWt;
   model bodyweight = age;
   domain cancer;
   ods select where=(_labelpath_ ? 'Cancer=1');
run;
ods graphics off;
```

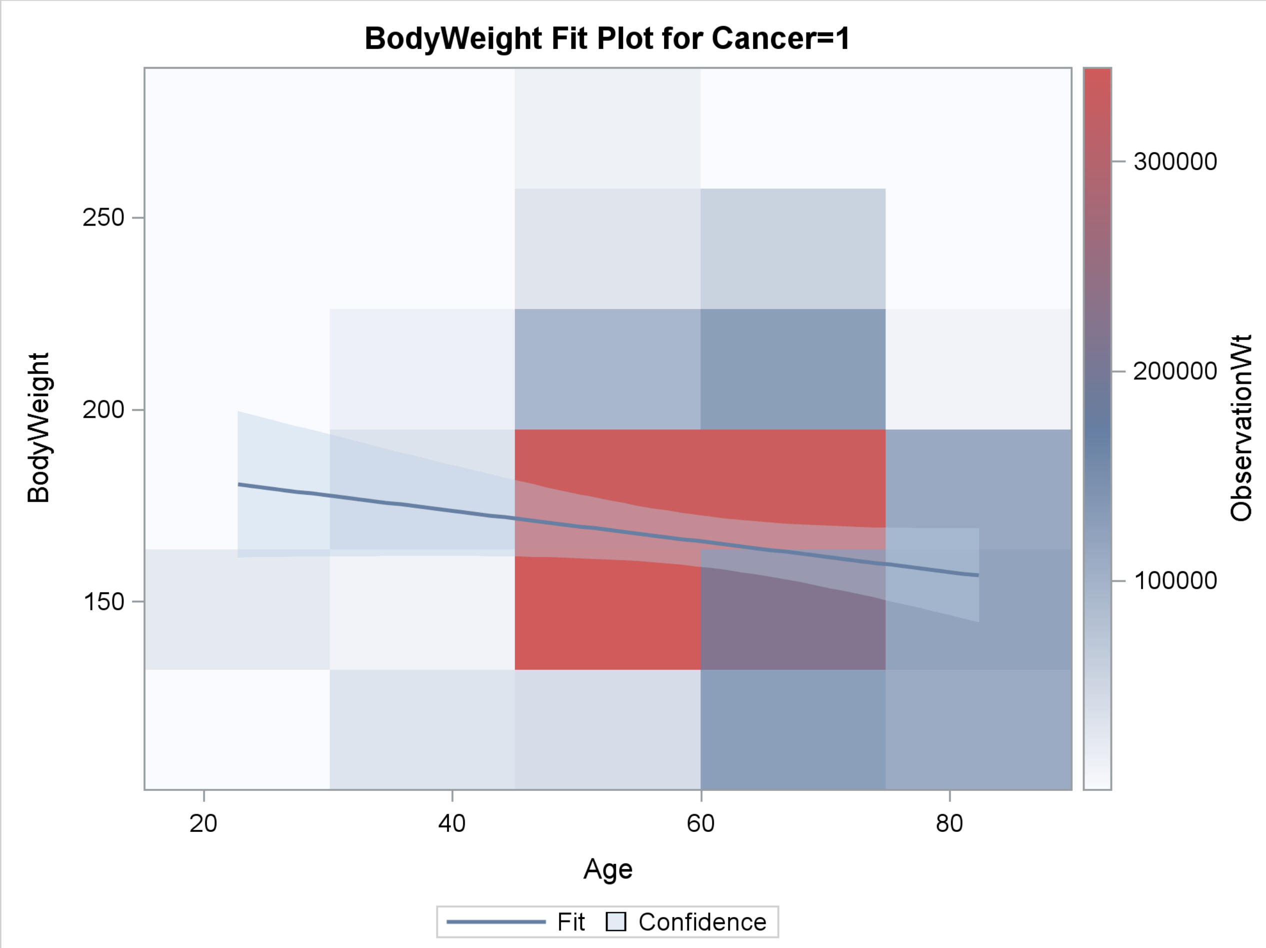
Output 98.7.1 gives a summary of the data and the parameter estimates of the linear regression in domain Cancer = 1. The analysis indicates that aging does not significantly affect body weight among cancer patients.

Output 98.7.1 Domain Analysis Among Cancer Patients

```
           Study of Body Weight and Age among Cancer Patients

                        The SURVEYREG Procedure

                               Cancer=1

           Domain Regression Analysis for Variable BodyWeight

                            Domain Summary

           Number of Observations                          1017
           Number of Observations in Domain                 119
           Number of Observations Not in Domain             898
           Sum of Weights in Domain                   2211545.0
           Weighted Mean of BodyWeight                164.87655
           Weighted Sum of BodyWeight                 364631909


                   Estimated Regression Coefficients

                                  Standard
      Parameter      Estimate        Error    t Value    Pr > |t|

      Intercept    189.614789   14.9467889      12.69      <.0001
      Age           -0.398556    0.2398447      -1.66      0.0971

    NOTE: The denominator degrees of freedom for the t tests is 586.
```

When ODS Graphics is enabled and the model contains a single continuous regressor, PROC SURVEYREG displays a plot of the model fitting, which is shown in Figure 98.7.2.

Output 98.7.2 Regression Fitting



Example 98.8: Compare Domain Statistics

Recall the example in the section "Getting Started: SURVEYREG Procedure" on page 8355, which analyzed a stratified simple random sample from a junior high school to examine how household income and the number of children in a household affect students' average weekly spending for ice cream. You can use the same sample to analyze the average weekly spending among male and female students. Because student gender is unrelated to the design of the sample, this kind of analysis is called domain analysis (subgroup analysis).

The data set follows:

```
data IceCreamDataDomain;
   input Grade Spending Income Gender$ @@;
   datalines;
7   7  39  M    7   7  38  F    8  12  47  F
9  10  47  M    7   1  34  M    7  10  43  M
7   3  44  M    8  20  60  F    8  19  57  M
7   2  35  M    7   2  36  F    9  15  51  F
8  16  53  F    7   6  37  F    7   6  41  M
```

```
7   6  39  M   9  15  50  M   8  17  57  F
8  14  46  M   9   8  41  M   9   8  41  F
9   7  47  F   7   3  39  F   7  12  50  M
7   4  43  M   9  14  46  F   8  18  58  M
9   9  44  F   7   2  37  F   7   1  37  M
7   4  44  M   7  11  42  M   9   8  41  M
8  10  42  M   8  13  46  F   7   2  40  F
9   6  45  F   9  11  45  M   7   2  36  F
7   9  46  F
;
data IceCreamDataDomain;
   set IceCreamDataDomain;
   if Grade=7 then Prob=20/1824;
   if Grade=8 then Prob=9/1025;
   if Grade=9 then Prob=11/1151;
   Weight=1/Prob;
run;
data StudentTotals;
   input Grade _TOTAL_;
   datalines;
7 1824
8 1025
9 1151
;
```

In the data set IceCreamDataDomain, the variable Grade indicates a student's grade, which is the stratification variable. The variable Spending contains the dollar amount of each student's average weekly spending for ice cream. The variable Income specifies the household income, in thousands of dollars. The variable Gender indicates a student's gender. The sampling weights are created by using the reciprocals of the probabilities of selection.

In the data set StudentTotals, the variable Grade is the stratification variable, and the variable _TOTAL_ contains the total numbers of students in the strata in the survey population.

Suppose that you are now interested in estimating the gender domain means of weekly ice cream spending (that is, the average spending for males and females, respectively). You can use the SURVEYMEANS procedure to produce these domain statistics by using the following statements:

```
proc surveymeans data=IceCreamDataDomain total=StudentTotals;
   strata Grade;
   var spending;
   domain Gender;
   weight Weight;
run;
```

Output 98.8.1 shows the estimated spending among male and female students.

Output 98.8.1 Estimated Domain Means

The SURVEYMEANS Procedure

Domain Statistics in Gender

Gender	Variable	N	Mean	Std Error of Mean
F	Spending	19	9.376111	1.077927
M	Spending	21	8.923052	1.003423

Domain Statistics in Gender

Gender	Variable	95% CL for Mean	
F	Spending	7.19202418	11.5601988
M	Spending	6.88992385	10.9561807

You can also use PROC SURVEYREG to estimate these domain means. The benefit of this alternative approach is that PROC SURVEYREG provides more tools for additional analysis, such as domain means comparisons in a LSMEANS statement.

Suppose that you want to test whether there is a significant difference for the ice cream spending between male and female students. You can use the following statements to perform the test:

```
title1 'Ice Cream Spending Analysis';
title2 'Compare Domain Statistics';
proc surveyreg data=IceCreamDataDomain total=StudentTotals;
   strata Grade;
   class Gender;
   model Spending = Gender / vadjust=none;
   lsmeans Gender / diff;
   weight Weight;
run;
```

The variable Gender is used as a model effect. The VADJUST=NONE option is used to produce variance estimates for domain means that are identical to those produced by PROC SURVEYMEANS. The LSMEANS statement requests that PROC SURVEYREG estimate the average spending in each gender group. The DIFF option requests that the procedure compute the difference among domain means.

Output 98.8.2 displays the estimated weekly spending on ice cream among male and female students, respectively, and their standard errors. Female students spend $9.38 per week on average, and male students spend $8.92 per week on average. These domain means, including their standard errors, are identical to those in Output 98.8.1 which are produced by PROC SURVEYMEANS.

Output 98.8.2 Domain Means between Gender

Ice Cream Spending Analysis
Compare Domain Statistics

The SURVEYREG Procedure

Regression Analysis for Dependent Variable Spending

Gender Least Squares Means

Gender	Estimate	Standard Error	DF	t Value	Pr > |t|
F	9.3761	1.0779	37	8.70	<.0001
M	8.9231	1.0034	37	8.89	<.0001

Output 98.8.3 shows the estimated difference for weekly ice scream spending between the two gender groups. The female students spend $0.45 more than male students on average, and the difference is not statistically significant based on the *t* test.

Output 98.8.3 Domain Means Comparison

Differences of Gender Least Squares Means

Gender	_Gender	Estimate	Standard Error	DF	t Value	Pr > |t|
F	M	0.4531	1.7828	37	0.25	0.8008

If you want to investigate whether there is any significant difference in ice cream spending among grades, you can use the following similar statements to compare:

```
ods graphics on;
title1 'Ice Cream Spending Analysis';
title2 'Compare Domain Statistics';
proc surveyreg data=IceCreamDataDomain total=StudentTotals;
   strata Grade;
   class Grade;
   model Spending = Grade / vadjust=none;
   lsmeans Grade / diff plots=(diff meanplot(cl));
   weight Weight;
run;
ods graphics off;
```

The Grade is specified in the CLASS statement to be used as an effect in the MODEL statement. The DIFF option in the LSMEANS statement requests that the procedure compute the difference among the domain means for the effect Grade. The ODS GRAPHICS statement enables ODS to create graphics. The PLOTS=(DIFF MEANPLOT(CL)) option requests two graphics: the domain means plot MeanPlot and their pairwise difference plot DiffPlot. The CL suboption requests the MeanPlot to display confidence. For information about ODS Graphics, see Chapter 21, “Statistical Graphics Using ODS.”

Output 98.8.4 shows the estimated weekly spending on ice cream for students within each grade. Students in Grade 7 spend the least, only $5.00 per week. Students in Grade 8 spend the most, $15.44 per week. Students in Grade 9 spend a little less at $10.09 per week.

Output 98.8.4 Domain Means among Grades

Ice Cream Spending Analysis
Compare Domain Statistics

The SURVEYREG Procedure

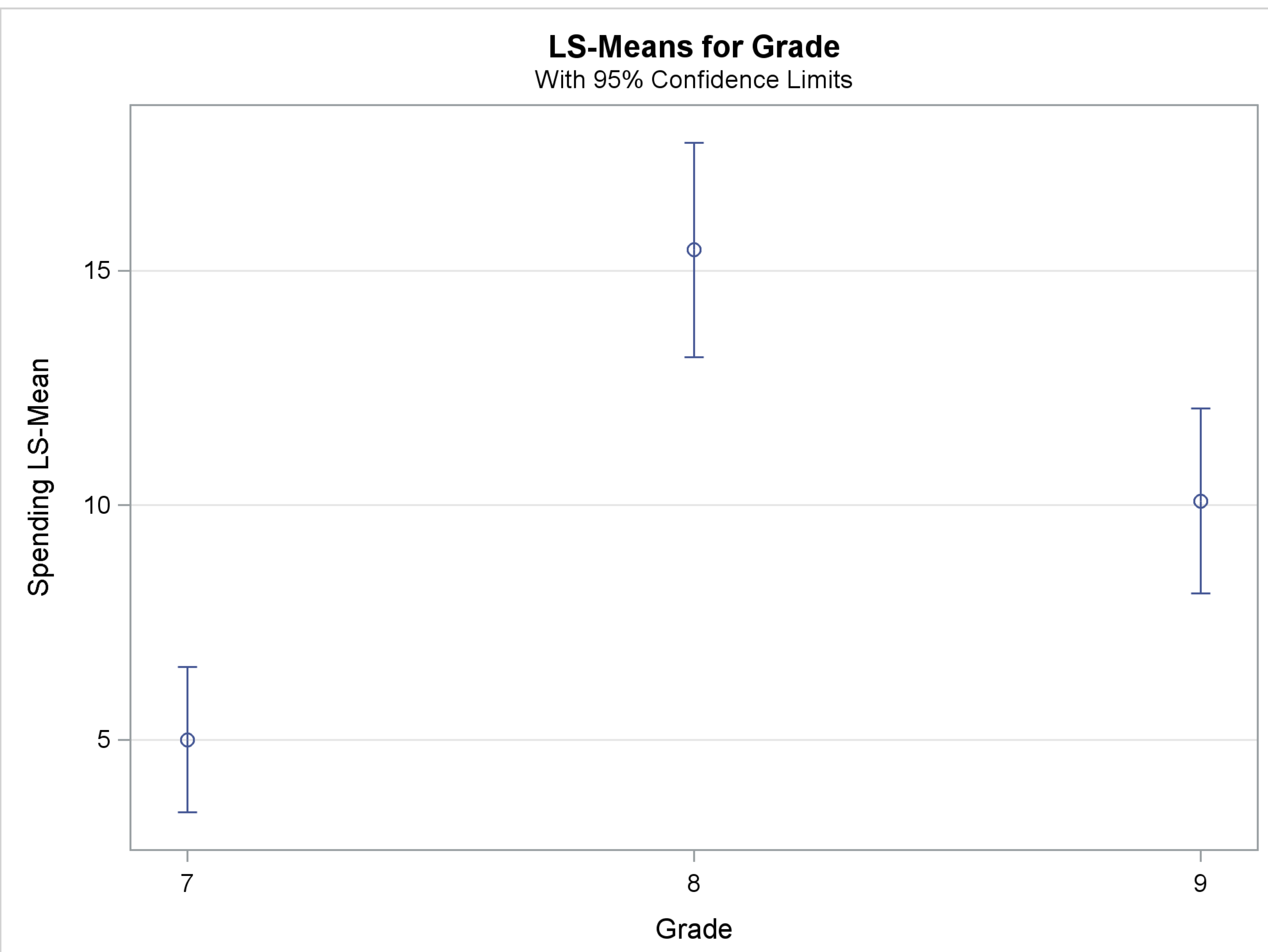
Regression Analysis for Dependent Variable Spending

Grade Least Squares Means

Grade	Estimate	Standard Error	DF	t Value	Pr > |t|
7	5.0000	0.7636	37	6.55	<.0001
8	15.4444	1.1268	37	13.71	<.0001
9	10.0909	0.9719	37	10.38	<.0001

Output 98.8.5 plots the weekly spending results that are shown in Output 98.8.4.

Output 98.8.5 Plot of Means of Ice Cream Spending within Grades



Output 98.8.6 displays pairwise comparisons for weekly ice scream spending among grades. All the differences are significant based on *t* tests.

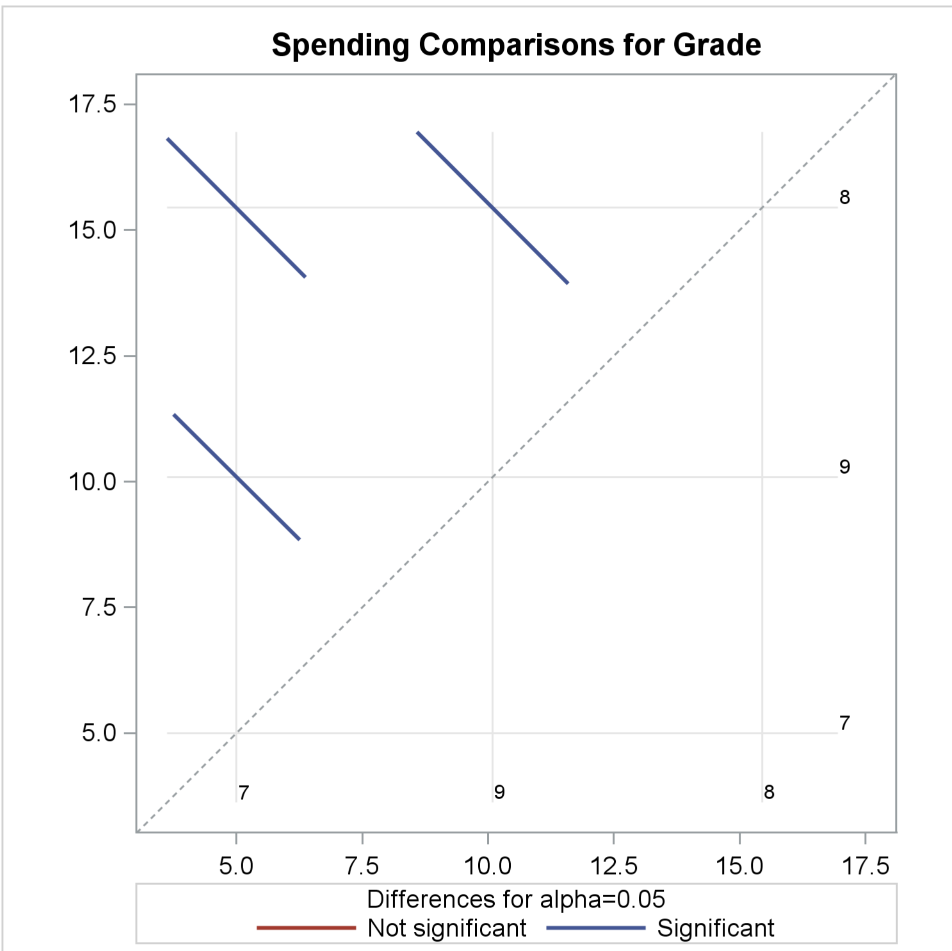
Output 98.8.6 Domain Means Comparison

Differences of Grade Least Squares Means

Grade	_Grade	Estimate	Standard Error	DF	t Value	Pr > |t|
7	8	-10.4444	1.3611	37	-7.67	<.0001
7	9	-5.0909	1.2360	37	-4.12	0.0002
8	9	5.3535	1.4880	37	3.60	0.0009

Output 98.8.7 plots the comparisons that are shown in Output 98.8.6.

Output 98.8.7 Plot of Pairwise Comparisons of Spending among Grades



In Output 98.8.7, the spending for each grade is shown in the background grid on both axes. Comparisons for each pair of domain means are shown by colored bars at intersections of these grids. The length of each bar represents the width of the confidence intervals for the corresponding difference between domain means.

The significance of these pairwise comparisons are indicated in the plot by whether these bars cross the 45-degree background dash-line across the plot. Since none of the three bars cross the dash-line, all pairwise comparisons are significant, as shown in Output 98.8.6.

Example 98.9: Variance Estimate Using the Jackknife Method

This example uses the stratified sample from the section "Getting Started: SURVEYREG Procedure" on page 8355 to illustrate how to estimate the variances with replication methods.

As shown in the section "Stratified Sampling" on page 8358, the sample is saved in the SAS data set IceCream. The variable Grade that indicates a student's grade is the stratification variable. The variable Spending contains the dollar amount of each student's average weekly spending for ice cream. The variable Income specifies the household income, in thousands of dollars. The variable Kids indicates how many children are in a student's family. The variable Weight contains sampling weights.

In this example, the procedure uses the jackknife method to estimate the variance, saving the replicate weights that PROC SURVEYREG generates in a SAS data set:

```
title1 'Ice Cream Spending Analysis';
title2 'Use the Jackknife Method to Estimate the Variance';
proc surveyreg data=IceCream
   varmethod=JACKKNIFE(outweights=JKWeights);
   strata Grade;
   class Kids;
   model Spending = Income Kids / solution;
   weight Weight;
run;
```

The VARMETHOD=JACKKNIFE option requests the procedure to estimate the variance by using the jackknife method. The OUTWEIGHTS=*JKWeights* option provides a SAS data set named JKWeights that contains the replicate weights used in the computation.

Output 98.9.1 shows the summary of the data and the variance estimation method. There are a total of 40 replicates generated by the procedure.

Output 98.9.1 Variance Estimation Using the Jackknife Method

```
                 Ice Cream Spending Analysis
       Use the Jackknife Method to Estimate the Variance

                   The SURVEYREG Procedure

      Regression Analysis for Dependent Variable Spending

                        Data Summary

            Number of Observations              40
            Sum of Weights                  4000.0
            Weighted Mean of Spending      9.14130
            Weighted Sum of Spending       36565.2
```

Output 98.9.1 *continued*

Design Summary

Number of Strata	3

Variance Estimation

Method	Jackknife
Number of Replicates	40

Output 98.9.2 displays the parameter estimates and their standard errors, as well as the tests of model effects that use the jackknife method.

Output 98.9.2 Variance Estimation Using the Jackknife Method

Tests of Model Effects

Effect	Num DF	F Value	Pr > F
Model	4	110.48	<.0001
Intercept	1	133.30	<.0001
Income	1	289.16	<.0001
Kids	3	0.90	0.4525

NOTE: The denominator degrees of freedom for the F tests is 37.

Estimated Regression Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-26.086882	2.58771182	-10.08	<.0001
Income	0.776699	0.04567521	17.00	<.0001
Kids 1	0.888631	1.12799263	0.79	0.4358
Kids 2	1.545726	1.25598146	1.23	0.2262
Kids 3	-0.526817	1.42555453	-0.37	0.7138
Kids 4	0.000000	0.00000000	.	.

NOTE: The denominator degrees of freedom for the t tests is 37.
Matrix X'WX is singular and a generalized inverse was used to solve the normal equations. Estimates are not unique.

Output 98.9.3 prints the first 6 observation in the output data set JKWeights, which contains the replicate weights.

Output 98.9.3 The Jackknife Replicate Weights for the First 6 Observations

The Jackknife Weights for the First 6 Obs

Obs	Grade	Spending	Income	Kids	Prob	Weight	RepWt_1	RepWt_2	RepWt_3	RepWt_4
1	7	7	39	2	0.010965	91.200	0.000	96.000	91.200	91.200
2	7	7	38	1	0.010965	91.200	96.000	0.000	91.200	91.200
3	8	12	47	1	0.008780	113.889	113.889	113.889	0.000	113.889
4	9	10	47	4	0.009557	104.636	104.636	104.636	104.636	0.000
5	7	1	34	4	0.010965	91.200	96.000	96.000	91.200	91.200
6	7	10	43	2	0.010965	91.200	96.000	96.000	91.200	91.200

Obs	RepWt_5	RepWt_6	RepWt_7	RepWt_8	RepWt_9	RepWt_10	RepWt_11	RepWt_12	RepWt_13
1	96.000	96.000	96.000	91.200	91.200	96.000	96.000	91.200	91.200
2	96.000	96.000	96.000	91.200	91.200	96.000	96.000	91.200	91.200
3	113.889	113.889	113.889	128.125	128.125	113.889	113.889	113.889	128.125
4	104.636	104.636	104.636	104.636	104.636	104.636	104.636	115.100	104.636
5	0.000	96.000	96.000	91.200	91.200	96.000	96.000	91.200	91.200
6	96.000	0.000	96.000	91.200	91.200	96.000	96.000	91.200	91.200

Obs	RepWt_14	RepWt_15	RepWt_16	RepWt_17	RepWt_18	RepWt_19	RepWt_20	RepWt_21	RepWt_22
1	96.000	96.000	96.000	91.200	91.200	91.200	91.200	91.200	91.200
2	96.000	96.000	96.000	91.200	91.200	91.200	91.200	91.200	91.200
3	113.889	113.889	113.889	113.889	128.125	128.125	113.889	113.889	113.889
4	104.636	104.636	104.636	115.100	104.636	104.636	115.100	115.100	115.100
5	96.000	96.000	96.000	91.200	91.200	91.200	91.200	91.200	91.200
6	96.000	96.000	96.000	91.200	91.200	91.200	91.200	91.200	91.200

Obs	RepWt_23	RepWt_24	RepWt_25	RepWt_26	RepWt_27	RepWt_28	RepWt_29	RepWt_30	RepWt_31
1	96.000	96.000	96.000	91.200	91.200	91.200	96.000	96.000	96.000
2	96.000	96.000	96.000	91.200	91.200	91.200	96.000	96.000	96.000
3	113.889	113.889	113.889	113.889	128.125	113.889	113.889	113.889	113.889
4	104.636	104.636	104.636	115.100	104.636	115.100	104.636	104.636	104.636
5	96.000	96.000	96.000	91.200	91.200	91.200	96.000	96.000	96.000
6	96.000	96.000	96.000	91.200	91.200	91.200	96.000	96.000	96.000

Obs	RepWt_32	RepWt_33	RepWt_34	RepWt_35	RepWt_36	RepWt_37	RepWt_38	RepWt_39	RepWt_40
1	96.000	91.200	91.200	91.200	96.000	91.200	91.200	96.000	96.000
2	96.000	91.200	91.200	91.200	96.000	91.200	91.200	96.000	96.000
3	113.889	113.889	128.125	128.125	113.889	113.889	113.889	113.889	113.889
4	104.636	115.100	104.636	104.636	104.636	115.100	115.100	104.636	104.636
5	96.000	91.200	91.200	91.200	96.000	91.200	91.200	96.000	96.000
6	96.000	91.200	91.200	91.200	96.000	91.200	91.200	96.000	96.000

The data set JKWeights contains all the variable in the data set IceCream, in addition to the replicate weights variables named RepWt_1, RepWt_2, ..., RepWt_40.

For example, the first observation (student) from stratum Grade=7 is deleted to create the first replicate. Therefore, stratum Grade=7 is the donor stratum for the first replicate, and the corresponding replicate weights are saved in the variable RepWt_1.

Because the first observation is deleted in the first replicate, RepWt_1=0 for the first observation. For observations from strata other than the donor stratum Grade=7, their replicate weights remain the same as in the variable Weight, while the rest of the observations in stratum Grade=7 are multiplied by the reciprocal of the corresponding jackknife coefficient, 0.95 for the first replicate.

References

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 Edition, New York: John Wiley & Sons.

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.

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.

Fuller, W. A., Kennedy, W. J., Schnell, D., Sullivan, G., and Park, H. J. (1989), *PC CARP*, Ames: Iowa State University Statistical Laboratory.

Hidiroglou, M. A., Fuller, W. A., and Hickman, R. D. (1980), *SUPER CARP*, Ames: Iowa State University Statistical Laboratory.

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.

Lohr, S. L. (2010), *Sampling: Design and Analysis*, 2nd Edition, Boston: Brooks/Cole.

Pringle, R. M. and Rayner, A. A. (1971), *Generalized Inverse Matrices with Applications to Statistics*, New York: Hafner Publishing.

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., Wu, C. F. J., and Yue, K. (1992), "Some Recent Work on Resampling Methods for Complex Surveys," *Survey Methodology*, 18, 209–217.

Rust, K. (1985), "Variance Estimation for Complex Estimators in Sample Surveys," *Journal of Official Statistics*, 1, 381–397.

Särndal, C. E. and Lundström, S. (2005), *Estimation in Surveys with Nonresponse*, Chichester, UK: John Wiley & Sons.

Särndal, C. E., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.

Wolter, K. M. (2007), *Introduction to Variance Estimation*, 2nd Edition, 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