

SAS/STAT[®] 14.3

User's Guide

The SURVEYREG

Procedure

This document is an individual chapter from *SAS/STAT® 14.3 User's Guide*.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2017. *SAS/STAT® 14.3 User's Guide*. Cary, NC: SAS Institute Inc.

SAS/STAT® 14.3 User's Guide

Copyright © 2017, SAS Institute Inc., Cary, NC, USA

All Rights Reserved. Produced in the United States of America.

For a hard-copy book: No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

For a web download or e-book: Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

U.S. Government License Rights; Restricted Rights: The Software and its documentation is commercial computer software developed at private expense and is provided with RESTRICTED RIGHTS to the United States Government. Use, duplication, or disclosure of the Software by the United States Government is subject to the license terms of this Agreement pursuant to, as applicable, FAR 12.212, DFAR 227.7202-1(a), DFAR 227.7202-3(a), and DFAR 227.7202-4, and, to the extent required under U.S. federal law, the minimum restricted rights as set out in FAR 52.227-19 (DEC 2007). If FAR 52.227-19 is applicable, this provision serves as notice under clause (c) thereof and no other notice is required to be affixed to the Software or documentation. The Government's rights in Software and documentation shall be only those set forth in this Agreement.

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

September 2017

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

SAS software may be provided with certain third-party software, including but not limited to open-source software, which is licensed under its applicable third-party software license agreement. For license information about third-party software distributed with SAS software, refer to <http://support.sas.com/thirdpartylicenses>.

Chapter 117

The SURVEYREG Procedure

Contents

Overview: SURVEYREG Procedure	9624
Getting Started: SURVEYREG Procedure	9625
Simple Random Sampling	9625
Stratified Sampling	9627
Output Data Sets	9630
Syntax: SURVEYREG Procedure	9631
PROC SURVEYREG Statement	9632
BY Statement	9643
CLASS Statement	9643
CLUSTER Statement	9644
CONTRAST Statement	9645
DOMAIN Statement	9646
EFFECT Statement	9648
ESTIMATE Statement	9649
LSMEANS Statement	9650
LSMESTIMATE Statement	9651
MODEL Statement	9653
OUTPUT Statement	9655
REPWEIGHTS Statement	9656
SLICE Statement	9658
STORE Statement	9658
STRATA Statement	9658
TEST Statement	9659
WEIGHT Statement	9659
Details: SURVEYREG Procedure	9660
Missing Values	9660
Survey Design Information	9661
Computational Details	9662
Variance Estimation	9665
Testing	9672
Domain Analysis	9673
Computational Resources	9673
Output Data Sets	9674
Displayed Output	9675
ODS Table Names	9680
ODS Graphics	9681

Examples: SURVEYREG Procedure	9682
Example 117.1: Simple Random Sampling	9682
Example 117.2: Cluster Sampling	9684
Example 117.3: Regression Estimator for Simple Random Sample	9687
Example 117.4: Stratified Sampling	9688
Example 117.5: Regression Estimator for Stratified Sample	9694
Example 117.6: Stratum Collapse	9697
Example 117.7: Domain Analysis	9701
Example 117.8: Comparing Domain Statistics	9704
Example 117.9: Variance Estimate Using the Jackknife Method	9710
References	9713

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 9680.

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 9681.

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 9631. The section “[Examples: SURVEYREG Procedure](#)” on page 9682 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 117.1 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 117.1 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 117.2 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 117.2 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 117.3.

Figure 117.3 Regression Coefficients

Estimated Regression Coefficients				
Parameter	Estimate	Standard 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 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 117.1 provides the number of students in each grade.

Table 117.1 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 117.1. 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;
```

Comparing these statements to those in the section “Simple Random Sampling” on page 9625, 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 117.4 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 117.1.

Figure 117.4 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
Design Summary	
Number of Strata	3

Figure 117.4 *continued*

Fit Statistics	
R-Square	0.8037
Root MSE	2.4371
Denominator DF	37

Figure 117.5 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 117.5 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 117.6 displays the tests for significance of the model effects. The Income effect is strongly significant at the 5% level.

Figure 117.6 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 117.7 displays the regression coefficient estimates, their standard errors, and the associated *t* tests for the stratified sample.

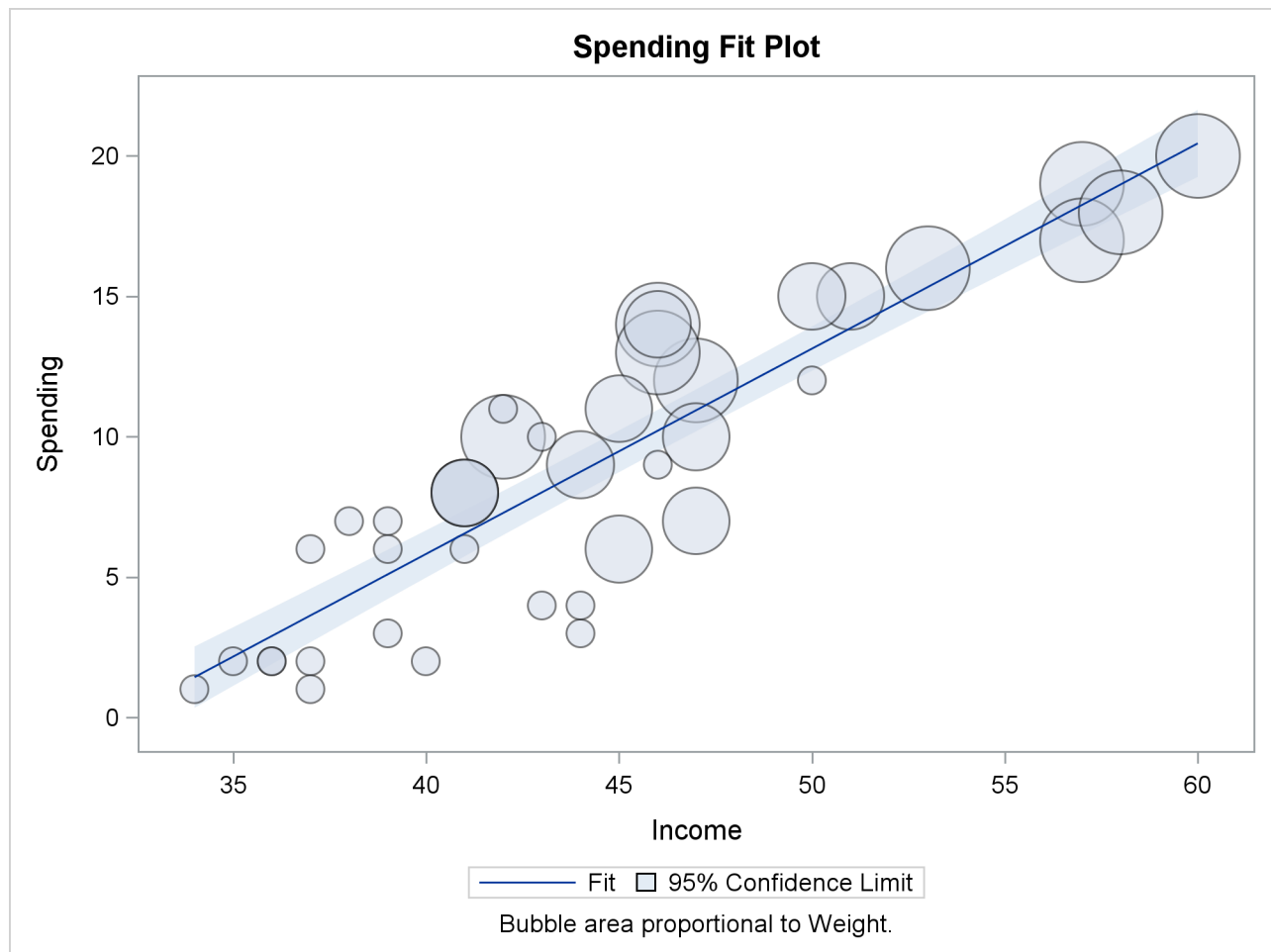
Figure 117.7 Regression Coefficients

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 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 117.8 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 117.8 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 9655 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 117.7](#)) 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 117.7](#) be placed into a SAS data set MyParmEst.

The PRINT procedure displays observations of the data set MyParmEst:

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

[Figure 117.9](#) displays the observations in the data set MyParmEst. The section “ODS Table Names” on page 9680 gives the complete list of the tables produced by PROC SURVEYREG.

Figure 117.9 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 117.2 summarizes the *options* available in the PROC SURVEYREG statement.

Table 117.2 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 valid category
NAMELEN=	Specifies the length of effect names
NOMCAR	Treats missing values as <i>not missing completely at random</i>
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 9660.

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

treats that missing values in the variance computation as *not missing completely at random* (NOMCAR) for Taylor series variance estimation. When you specify this 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. For more information, see the section “[Missing Values](#)” on page 9660.

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; it is ignored for replication methods.

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 *SAS Visual Data Management and Utility 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;
```

For more information about enabling and disabling ODS Graphics, see the section “[Enabling and Disabling ODS Graphics](#)” on page 615 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. The number of bins also depends on the value of the **SHAPE=** option:

- If **SHAPE=RECTANGULAR**, the number of bins is calculated as follows:
 - If you specify only one number, *nbin1*, then it is used for both the horizontal and vertical axes. For example, if you specify **NBINS=10**, then PROC SURVEYREG creates 100 bins (10×10).
 - If you specify two numbers, then *nbin1* is used for the horizontal axis and *nbin2* is used for the vertical axis. For example, if you specify **NBINS=10 20**, then PROC SURVEYREG creates 200 bins (10×20).
- If you specify **SHAPE=HEXAGONAL**, then PROC SURVEYREG calculates the number of rectangular bins by the method previously described. Then the procedure creates a heat map that has hexagonal bins of approximately the same size as those rectangular bins.

If you specify this option, then by default **WEIGHT=HEATMAP**.

If you do not specify this option, then the number of bins is determined by first using the algorithm that is discussed in the section “[ODS Graphics](#)” on page 5085 in Chapter 68, “[The KDE Procedure](#),” and then multiplying the resulting numbers of bins by three.

WEIGHT=BUBBLE | HEATMAP

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*. You can specify the following options:

BUBBLE	overlays a bubble plot in which the bubble area is proportional to the weight of an observation.
HEATMAP	overlays a heat map in which the heat color represents the sum of the weights at the corresponding location.

If you specify the **NBINS=** option, then by default **WEIGHT=HEATMAP**.

If you do not specify this option, then the default plot depends on the number of observations in your data. For a data set that contains 100 observations or less, the default is **WEIGHT=BUBBLE**. For a data set that contains more than 100 observations, the default is **WEIGHT=HEATMAP**.

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. The number of bins also depends on the value of the **SHAPE=** option:

- If **SHAPE=RECTANGULAR**, the number of bins is calculated as follows:
 - If you specify only one number, *nbin1*, then it is used for both the horizontal and vertical axes. For example, if you specify **NBINS=10**, then PROC SURVEYREG creates 100 bins (10×10).
 - If you specify two numbers, then *nbin1* is used for the horizontal axis and *nbin2* is used for the vertical axis. For example, if you specify **NBINS=10 20**, then PROC SURVEYREG creates 200 bins (10×20).
- If you specify **SHAPE=HEXAGONAL**, then PROC SURVEYREG calculates the number of rectangular bins by the method previously described. Then the procedure creates a heat map that has hexagonal bins of approximately the same size as those rectangular bins.

If you specify this option, then by default **WEIGHT=HEATMAP**.

If you do not specify this option, then the number of bins is determined by first using the algorithm that is discussed in the section “[ODS Graphics](#)” on page 5085 in Chapter 68, “[The KDE Procedure](#),” and then multiplying the resulting numbers of bins by three.

WEIGHT=BUBBLE | HEATMAP

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*. You can specify the following options:

- | | |
|----------------|--|
| BUBBLE | overlays a bubble plot in which the bubble area is proportional to the weight of an observation. |
| HEATMAP | overlays a heat map in which the heat color represents the sum of the weights at the corresponding location. |

If you specify the **NBINS=** option, then by default **WEIGHT=HEATMAP**.

If you do not specify this option, then the default plot depends on the number of observations in your data. For a data set that contains 100 observations or less, the default is **WEIGHT=BUBBLE**. For a data set that contains more than 100 observations, the default is **WEIGHT=HEATMAP**.

SHAPE=RECTANGULAR | HEXAGONAL

SHAPE=REC | HEX

requests either rectangular or hexagonal bins for a heat map of the data. Thus, this option implies **WEIGHT=HEATMAP** by default. By default, **SHAPE=RECTANGULAR**.

NONE

suppresses all plots.

RATE=value | SAS-data-set

R=value | SAS-data-set

specifies the sampling rate, which PROC SURVEYREG uses to compute a finite population correction for Taylor series variance estimation. This option is ignored for replication variance estimation methods.

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

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

<i>value</i>	specifies a nonnegative number to use for a nonstratified design or for a stratified design that has the same sampling rate in each stratum.
<i>SAS-data-set</i>	specifies a <i>SAS-data-set</i> that contains the stratification variables and the sampling rates for a stratified design that has different sampling rates in the strata. You must provide the sampling rates in the data set variable named <code>_RATE_</code> . The sampling rates must be nonnegative numbers.

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

For more information, see the section “[Specification of Population Totals and Sampling Rates](#)” on page 9661.

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

TOTAL=value | SAS-data-set

N=value | SAS-data-set

specifies the total number of primary sampling units (PSUs) in the study population. PROC SURVEYREG uses this information to compute a finite population correction for Taylor series variance estimation. This option is ignored for replication variance estimation methods.

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

<i>value</i>	specifies a positive number to use for a nonstratified design or for a stratified design that has the same population total in each stratum.
<i>SAS-data-set</i>	specifies a <i>SAS-data-set</i> that contains the stratification variables and the population totals for a stratified design that has different population totals in the strata. You must provide the stratum totals in the data set variable named <code>_TOTAL_</code> . The stratum totals must be positive numbers.

For more information, see the section “[Specification of Population Totals and Sampling Rates](#)” on page 9661.

If you do not specify either the `TOTAL=` or the `RATE=` option, 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=*method* <(method-options)>

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

[Table 117.3](#) summarizes the available *methods* and *method-options*.

Table 117.3 Variance Estimation Methods

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

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

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

By default, VARMETHOD=JACKKNIFE if you also specify a [REPWEIGHTS](#) statement; otherwise, VARMETHOD=TAYLOR by default.

You can specify the following *methods*:

BOOTSTRAP <(method-options)>

requests variance estimation by the bootstrap method. For more information, see the section “[Bootstrap Method](#)” on page 9667.

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

You can specify the following *method-options*:

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

specifies the number of PSUs to select for the bootstrap replicate samples. You can provide bootstrap stratum sample sizes m_h by specifying a list of *values* or a *SAS-data-set*. Alternatively, you can provide a single bootstrap sample size *value* to use for all strata or for a nonstratified design. You can specify the number of replicate samples in the [REPS=](#) option. For more information, see the section “[Bootstrap Method](#)” on page 9667.

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

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

MH=*value*

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

MH=(*values*)

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

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

MH=*SAS-data-set*

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

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

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

OUTWEIGHTS=SAS-data-set

names a *SAS-data-set* in which to store the bootstrap replicate weights that PROC SURVEYREG creates. For information about replicate weights, see the section “[Bootstrap Method](#)” on page 9667. For information about the contents of the OUTWEIGHTS= data set, see the section “[Replicate Weights Output Data Set](#)” on page 9675.

This *method-option* is not available when you provide replicate weights in a [REPWEIGHTS](#) statement.

REPS=number

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

SEED=number

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

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

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

PROC SURVEYREG displays the value of the initial seed in the “Variance Estimation” table.

BRR <(method-options)>

requests variance estimation by balanced repeated replication (BRR). This method requires a stratified sample design where each stratum contains two primary sampling units (PSUs). When you specify this method, you must also specify a [STRATA](#) statement unless you provide replicate weights by using the [REPWEIGHTS](#) statement. For more information, see the section “[Balanced Repeated Replication \(BRR\) Method](#)” on page 9668.

You can specify the following *method-options*:

FAY <=value>

requests Fay’s method, which is a modification of the BRR method. For more information, see the section “[Fay’s BRR Method](#)” on page 9669.

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

HADAMARD=SAS-data-set**H=SAS-data-set**

names a *SAS-data-set* that contains the Hadamard matrix for BRR replicate construction. If you do not specify this *method-option*, PROC SURVEYREG generates an appropriate Hadamard matrix for replicate construction. For more information, see the sections “[Balanced Repeated Replication \(BRR\) Method](#)” on page 9668 and “[Hadamard Matrix](#)” on page 9671.

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

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

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

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

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

OUTWEIGHTS=SAS-data-set

names a *SAS-data-set* in which to store the replicate weights that PROC SURVEYREG creates for BRR variance estimation. For information about replicate weights, see the section “[Balanced Repeated Replication \(BRR\) Method](#)” on page 9668. For information about the contents of the OUTWEIGHTS= data set, see the section “[Replicate Weights Output Data Set](#)” on page 9675.

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

PRINTH

displays the Hadamard matrix that PROC SURVEYREG uses to construct replicates for BRR variance estimation. When you provide the Hadamard matrix in the *HADAMARD= method-option*, PROC SURVEYREG displays only the rows and columns that are actually used to construct replicates. For more information, see the sections “[Balanced Repeated Replication \(BRR\) Method](#)” on page 9668 and “[Hadamard Matrix](#)” on page 9671.

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

REPS=number

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

If you do not use the *HADAMARD= method-option* to provide a Hadamard matrix, the number of replicates should be greater than the number of strata and should be a multiple of 4. For more information, see the section “[Balanced Repeated Replication \(BRR\) Method](#)”

on page 9668. If PROC SURVEYREG cannot construct a Hadamard matrix for the REPS= value that you specify, the value is increased until a Hadamard matrix of that dimension can be constructed. Therefore, the actual number of replicates that PROC SURVEYREG uses might be larger than *number*.

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

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

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

JACKKNIFE <(method-options)>

JK <(method-options)>

requests variance estimation by the delete-1 jackknife method. For more information, see the section “[Jackknife Method](#)” on page 9670. If you use a **REPWEIGHTS** statement to provide replicate weights, VARMETHOD=JACKKNIFE is the default variance estimation method.

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

You can specify the following *method-options*:

OUTJKCOEFS=SAS-data-set

names a *SAS-data-set* in which to store the jackknife coefficients. For information about jackknife coefficients, see the section “[Jackknife Method](#)” on page 9670. For information about the contents of the OUTJKCOEFS= data set, see the section “[Jackknife Coefficients Output Data Set](#)” on page 9675.

OUTWEIGHTS=SAS-data-set

names a *SAS-data-set* in which to store the replicate weights that PROC SURVEYREG creates for jackknife variance estimation. For information about replicate weights, see the section “[Jackknife Method](#)” on page 9670. For information about the contents of the OUTWEIGHTS= data set, see the section “[Replicate Weights Output Data Set](#)” on page 9675.

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

TAYLOR

requests Taylor series variance estimation. This is the default method if you do not specify the VARMETHOD= option or a **REPWEIGHTS** statement. For more information, see the section “[Taylor Series \(Linearization\)](#)” on page 9666.

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 *SAS Visual Data Management and Utility Procedures Guide*.

CLASS Statement

CLASS *variable* < (REF= *option*) > ... < *variable* < (REF= *option*) > > < / *global-options* > ;

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 *SAS Visual Data Management and Utility 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 specify the following REF= option to indicate how the levels of an individual classification variable are to be ordered by enclosing it in parentheses after the variable name:

REF= 'level' | FIRST | LAST

specifies a level of the classification variable to be put at the end of the list of levels. This level thus corresponds to the reference level in the usual interpretation of the estimates with PROC SURVEYREG's singular parameterization. You can specify the *level* of the variable to use as the reference level; specify a value that corresponds to the formatted value of the variable if a format is assigned. Alternatively, you can specify REF=FIRST to designate that the first ordered level serve as the reference, or REF=LAST to designate that the last ordered level serve as the reference. To specify that REF=FIRST or REF=LAST be used for all classification variables, use the **REF=** *global-option* after the slash (/) in the CLASS statement.

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

REF=FIRST | LAST

specifies a level of all classification variables to be put at the end of the list of levels. This level thus corresponds to the reference level in the usual interpretation of the estimates with PROC SURVEYREG's singular parameterization. Specify REF=FIRST to designate that the first ordered level for each classification variable serve as the reference. Specify REF=LAST to designate that the last ordered level serve as the reference. This option applies to all the variables specified in the CLASS statement. To specify different reference levels for different classification variables, use **REF=** options for individual 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 9661 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 *SAS Visual Data Management and Utility 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 9660.

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 $H_0: \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$, 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 391 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 3773 in Chapter 48, “[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 $\text{rank}(\mathbf{L})$. 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,

<i>label</i>	identifies the contrast in the output. A label is required for every contrast specified. Labels must be enclosed in single quotes.
<i>effect</i>	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.
<i>values</i>	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 3792 in Chapter 48, “[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 $\text{ABS}(\mathbf{v})$ to be the largest absolute value of the elements of $\mathbf{v}$. For a row vector $\mathbf{l}$ of the matrix $\mathbf{L}$, define

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

If $\text{ABS}(\mathbf{l} - \mathbf{H}\mathbf{l})$ is greater than $c \cdot \text{value}$, then $\mathbf{l}\boldsymbol{\beta}$ is declared nonestimable. Here, $\mathbf{H}$ is the matrix $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}$. 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 $H_0: \mathbf{L}\boldsymbol{\beta} = 0$.

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.

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

The formatted values of the domain variables determine the categorical variable levels. Thus, you can use formats to group values into levels. For more information, see the FORMAT procedure in *SAS Visual Data*.

Management and Utility Procedures Guide and the FORMAT statement and SAS formats in *SAS Formats and Informats: Reference*.

When determining levels of a DOMAIN variable, an observation with missing values for this DOMAIN variable is excluded, unless you specify the **MISSING** option. For more information, see the section “[Missing Values](#)” on page 9660.

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

A DOMAIN statement is different from a BY statement. In a BY statement, you treat the sample sizes as fixed in each subpopulation, and you perform analysis within each BY group independently. For more information, see the section “[Domain Analysis](#)” on page 9673. Similarly, you should use a DOMAIN statement to perform a domain analysis over the entire data set. Creating a new data set from a single domain and analyzing that with PROC SURVEYREG yields inappropriate estimates of variance.

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

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

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

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

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

have the same effect as a single specification:

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

Also, the following specification

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

is equivalent to

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

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

EFFECT Statement

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

The EFFECT statement enables you to construct special collections of columns for design matrices. These collections are referred to as *constructed effects* to distinguish them from the usual model effects that are formed from continuous or classification variables, as discussed in the section “GLM Parameterization of Classification Variables and Effects” on page 391 in Chapter 19, “Shared Concepts and Topics.”

You can specify the following *effect-types*:

COLLECTION	specifies a collection effect that defines one or more variables as a single effect with multiple degrees of freedom. The variables in a collection are considered as a unit for estimation and inference.
LAG	specifies a classification effect in which the level that is used for a particular period corresponds to the level in the preceding period.
MULTIMEMBER MM	specifies a multimember classification effect whose levels are determined by one or more variables that appear in a CLASS statement.
POLYNOMIAL POLY	specifies a multivariate polynomial effect in the specified numeric variables.
SPLINE	specifies a regression spline effect whose columns are univariate spline expansions of one or more variables. A spline expansion replaces the original variable with an expanded or larger set of new variables.

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

Table 117.4 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. This option is required.
WITHIN=	Names the variable or variables that define the group within which each period is defined. This option is required.
Multimember Effects Options	
NOEFFECT	Specifies that observations with all missing levels for the multimember variables should have zero values in the corresponding design matrix columns

Table 117.4 *continued*

Option	Description
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 401 in Chapter 19, “Shared Concepts and Topics.”

ESTIMATE Statement

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

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

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

Table 117.5 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 <i>p</i>-values	
ADJUST=	Determines the method for multiple comparison adjustment of estimates

Table 117.5 *continued*

Option	Description
ALPHA= α	Determines the confidence level ($1 - \alpha$)
LOWER	Performs one-sided, lower-tailed inference
STEPDOWN	Adjusts multiplicity-corrected p -values further in a step-down fashion
TESTVALUE=	Specifies values under the null hypothesis for tests
UPPER	Performs one-sided, upper-tailed inference
Statistical Output	
CL	Constructs confidence limits
CORR	Displays the correlation matrix of estimates
COV	Displays the covariance matrix of estimates
E	Prints the L matrix
JOINT	Produces a joint F or chi-square test for the estimable functions
PLOTS=	Requests ODS statistical graphics if the analysis is sampling-based
SEED=	Specifies the seed for computations that depend on random numbers
Generalized Linear Modeling	
CATEGORY=	Specifies how to construct estimable functions with multinomial data
EXP	Exponentiates and displays estimates
ILINK	Computes and displays estimates and standard errors on the inverse linked scale

For details about the syntax of the ESTIMATE statement, see the section “[ESTIMATE Statement](#)” on page 448 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 117.6 summarizes available *options* in the LSMEANS statement.

Table 117.6 LSMEANS Statement Options

Option	Description
Construction and Computation of LS-Means	
AT	Modifies the covariate value in computing LS-means
BYLEVEL	Computes separate margins

Table 117.6 *continued*

Option	Description
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	Uses connecting lines to indicate nonsignificantly different subsets of LS-means
LINESTABLE	Displays the results of the LINES option as a table
MEANS	Prints the LS-means
PLOTS=	Requests graphs of means and mean comparisons
SEED=	Specifies the seed for computations that depend on random numbers
Generalized Linear Modeling	
EXP	Exponentiates and displays estimates of LS-means or LS-means differences
ILINK	Computes and displays estimates and standard errors of LS-means (but not differences) on the inverse linked scale
ODDSRATIO	Reports (simple) differences of least squares means in terms of odds ratios if permitted by the link function

For details about the syntax of the LSMEANS statement, see the section “[LSMEANS Statement](#)” on page 464 in Chapter 19, “[Shared Concepts and Topics](#).”

LSMESTIMATE Statement

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

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

Table 117.7 summarizes the *options* available in the LSMESTIMATE statement.

Table 117.7 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
PLOTS=	Requests graphs of means and mean comparisons
SEED=	Specifies the seed for computations that depend on random numbers
Generalized Linear Modeling	
CATEGORY=	Specifies how to construct estimable functions with multinomial data
EXP	Exponentiates and displays LS-means estimates
ILINK	Computes and displays estimates and standard errors of LS-means (but not differences) on the inverse linked scale

For details about the syntax of the LSMESTIMATE statement, see the section “[LSMESTIMATE Statement](#)” on page 484 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 3773 in Chapter 48, “[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 117.8 summarizes the *options* available in the MODEL statement.

Table 117.8 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 $X'X$ 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 $X'X$ matrix, or the $X'WX$ 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 9671 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 $\text{ABS}(\mathbf{v})$ to be the largest absolute value of the elements of $\mathbf{v}$. For a row vector $\mathbf{l}$ of the matrix $\mathbf{L}$, define

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

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

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 $X'X$ matrix, or the $X'WX$ matrix when there is a **WEIGHT** variable, where **W** is the diagonal matrix constructed from **WEIGHT** variable values. The **X** option also displays the crossproducts vector $X'y$ or $X'Wy$.

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 bootstrap, BRR, or jackknife variance estimation, which you request with the **VARMETHOD=BOOTSTRAP**, **VARMETHOD=BRR**, or **VARMETHOD=JACKKNIFE** option, respectively, in the PROC SURVEYREG statement. For more information about the replication methods, see the section “Variance Estimation” on page 9665.

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

For more information about replicate weights that the SURVEYREG procedure creates, see the sections “Balanced Repeated Replication (BRR) Method” on page 9668 and “Jackknife Method” on page 9670.

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 value of *df* is the number of REPWEIGHTS variables.

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

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

specifies jackknife coefficients for the **VARMETHOD=JACKKNIFE** option in the PROC SURVEYREG statement. The jackknife coefficient values must be nonnegative numbers. For more information about jackknife coefficients, see the section “Jackknife Method” on page 9670.

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

value

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

values

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

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

SAS-data-set

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

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

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

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

value

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

values

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

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

SAS-data-set

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

SLICE Statement

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

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

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 512 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 515 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 9661 for more information.

If you provide replicate weights for BRR, jackknife, or bootstrap variance estimation by including 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 *SAS Visual Data Management and Utility 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 9660.

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 9677 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 9664 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 117.9](#) summarizes the *options* available in the TEST statement.

Table 117.9 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 517 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 9660 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 or bootstrap variance estimation, you can input either the sampling rate or the population total by using the **RATE=** or **TOTAL=** option, respectively, in the PROC 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 or bootstrap variance estimation. The procedure does not use a finite population correction for BRR or jackknife variance estimation.

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

If your design has multiple stages of selection and you are specifying the **RATE=** option, you should input the first-stage sampling rate, which is the ratio of the number of primary sampling units (PSUs) in the sample to the total number of PSUs in the study population. If you are specifying the **TOTAL=** option for a multistage design, you should input the total number of PSUs in the study population. For more information, see the section “[Primary Sampling Units \(PSUs\)](#)” on page 9661.

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

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

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

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

Primary Sampling Units (PSUs)

When you have clusters, or primary sampling units (PSUs), in your sample design, the procedure estimates variance from the variation among PSUs when the Taylor series variance method is used. For more information, see the section “[Taylor Series \(Linearization\)](#)” on page 9666.

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

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

If you use a **REPWEIGHTS** statement to provide replicate weights for BRR, jackknife, bootstrap variance estimation, you do not need to specify a **CLUSTER** statement. Otherwise, you should specify a **CLUSTER** statement whenever your design includes clustering at the first stage of sampling. If you do not specify a **CLUSTER** statement, then PROC 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.)
- $h = 1, 2, \dots, H$ is the stratum index
- $i = 1, 2, \dots, n_h$ is the cluster index within stratum h
- $j = 1, 2, \dots, m_{hi}$ 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 9661 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 9665. 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

$$f_{\text{SRS}} = n / w_{\dots}$$

where n is the sample size and $w_{\dots}$ (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:

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

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:

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

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 replication methods.

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 9662 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 $h_1, h_2, \dots, h_c$ into the pooled stratum, the procedure calculates the sampling rate for the pooled stratum as

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

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

$$R^2 = 1 - \frac{SS_{error}}{SS_{total}}$$

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

$$SS_{error} = \mathbf{r}'\mathbf{W}\mathbf{r}$$

and SS_{total} is the total sum of squares

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

where $w_{...}$ 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

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

where R^2 is the multiple R-square.

Root Mean Square Errors

PROC SURVEYREG computes the square root of mean square errors as

$$\sqrt{MSE} = \sqrt{n SS_{error} / (n - p) w_{...}}$$

where $w_{...}$ 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, the *jackknife* method, and the *bootstrap* 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 9662 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

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

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

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \mathbf{G} (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$$

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.

Bootstrap Method

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

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

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

Replicate Weight Construction

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

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

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

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

You can use the **OUTWEIGHTS=SAS-data-set** method-option to store the bootstrap replicate weights in a SAS data set. For information about the contents of the **OUTWEIGHTS=** data set, see the section “[Replicate Weights Output Data Set](#)” on page 9675. You can provide these replicate weights to the procedure for subsequent analyses by using a **REPWEIGHTS** statement.

Variance Estimation

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

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

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

Let $\hat{\beta}$ be the estimated regression coefficients from the full sample for β , and let $\hat{\beta}_r$ be the estimated regression coefficient when the r th set of bootstrap replicate weights is used. PROC SURVEYREG estimates the covariance matrix of $\hat{\beta}$ by the following equation

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

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

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

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 $\text{number} \times \text{number}$ 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 r th ($r = 1, 2, \dots, R$) replicate is drawn from the full sample according to the r th 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 r th 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 r th 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 $\hat{\beta}$ be the estimated regression coefficients from the full sample for β , and let $\hat{\beta}_r$ be the estimated regression coefficient from the r th replicate by using replicate weights. PROC SURVEYREG estimates the covariance matrix of $\hat{\beta}$ by

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

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 =  $\epsilon$` *method-option*. By default, $\epsilon = 0.5$ if the `FAY` *method-option* is specified without providing a value for ϵ (Judkins 1990; Rao and Shao 1999). When $\epsilon = 0$, Fay's method becomes the traditional BRR method. For more details, see Dipbo, 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 r th ($r = 1, 2, \dots, R$) replicate is created from the full sample according to the r th 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 r th 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 r th 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 $\hat{\beta}$ be the estimated regression coefficients from the full sample for β . Let $\hat{\beta}_r$ be the estimated regression coefficient obtained from the r th replicate by using replicate weights. PROC SURVEYREG estimates the covariance matrix of $\hat{\beta}$ by

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

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:

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

Denote the original weight in the full sample for the j th member of the i th PSU as w_{ij} . If the i th PSU is included in the r th replicate ($r = 1, 2, \dots, R$), then the corresponding replicate weight for the j th member of the i th 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 r th 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

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

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

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

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

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

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

$$\hat{V}(\hat{\beta}) = \sum_{r=1}^R \alpha_r (\hat{\beta}_r - \hat{\beta}) (\hat{\beta}_r - \hat{\beta})'$$

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 $\mathbf{H}$ 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 $\mathbf{H}$ and $\mathbf{I}$ 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 $k = 8$:

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

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

For replication variance estimation methods that include a **REPWEIGHTS** statement, the denominator df equals the number of **REPWEIGHTS** variables, unless you specify an alternative in the **DF= option** in the **REPWEIGHTS** statement.

When the **BRR** method (including **Fay's method**) is specified but no **REPWEIGHTS** statement is included, the denominator df equals the number of strata.

When the **jackknife** method or **bootstrap** method is specified but no **REPWEIGHTS** statement is included, the denominator df is the same as the Taylor series method described earlier.

Alternatively, you can specify your own denominator df by using the **DF=** option in the **MODEL** 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 9672 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 $H_0: \mathbf{L}\boldsymbol{\beta} = 0$. The Wald F statistic equals

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

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

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 9671. 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 517 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

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

where $\mathbf{L}$ is the contrast vector or matrix you specify, $\boldsymbol{\beta}$ is the vector of regression parameters, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}$, $\hat{\mathbf{V}}$ is the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$, $\text{rank}(\mathbf{L})$ 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 3825 in Chapter 48, “**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 9680 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 replication 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`, `VARMETHOD=JACKKNIFE`, or `VARMETHOD=BOOTSTRAP`, 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 all variables in the `DATA=` input data set and the replicate weight variables, `RepWt_1`, `RepWt_2`, . . . , `RepWt_n`, where n is the total number of replicates in the analysis. Each replicate weight variable contains the replicate weights for the corresponding replicate. Replicate weights equal 0 for observations that are not included in the replicate.

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

Jackknife Coefficients Output Data Set

If you specify the `OUTJKCOEFS= method-option` for `VARMETHOD=JACKKNIFE`, PROC 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 9665) 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
- Initial seed for the random number that the bootstrap method uses
- 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 $X'X$ matrix. When there is a WEIGHT variable, the procedure displays the $X'WX$ matrix. This option also displays the crossproducts vector $X'y$ or $X'Wy$, where 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 $X'X$ matrix. When there is a WEIGHT variable, the procedure displays the inverse or the generalized inverse of the $X'WX$ 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 $H_0: \text{Parameter} = 0$
- $\text{Pr} > |t|$, 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 $H_0: \text{Contrast} = 0$
- $\text{Pr} > F$, 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 117.10. 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 117.10. For information about these tables, see the corresponding sections of Chapter 19, “Shared Concepts and Topics.”

Table 117.10 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	Default
FitStatistics	Fit statistics	MODEL	Default
HadamardMatrix	Hadamard matrix	PROC	PRINTH
InvXPX	Inverse matrix of $X'X$	MODEL	I
ParameterEstimates	Estimated regression coefficients	MODEL	SOLUTION
StrataInfo	Stratum information	STRATA	LIST
VarianceEstimation	Variance estimation	PROC	Default
XPX	$X'X$ 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 9627:

```

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 615 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 614 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 117.11](#) lists the name of the graph that PROC SURVEYREG generates, together with its description and the [PLOTS=](#) option *plot-request* that produces it.

Table 117.11 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 117.1: 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:

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

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;
```

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 117.1.1 summarizes the data information and the fit information.

Output 117.1.1 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 117.1.2 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 117.1.2 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				
Parameter	Estimate	Standard		
		Error	t Value	Pr > t
Intercept	0.20331056	0.09444296	2.15	0.0452
LFPR1968	0.65604048	0.17635810	3.72	0.0016

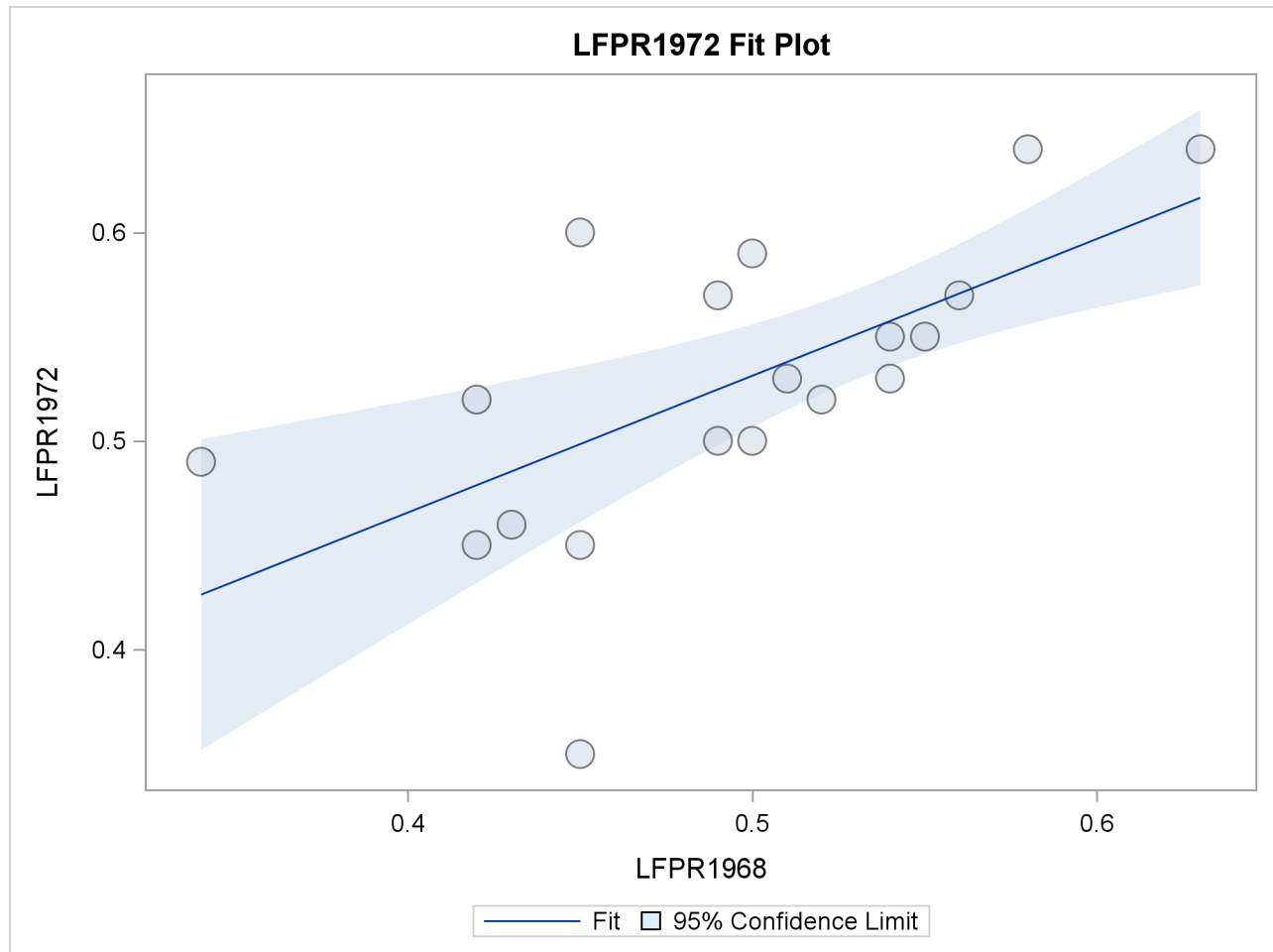
Note: The degrees of freedom for the t tests is 18.

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 [Output 117.1.3](#).

Output 117.1.3 Regression Fitting



Example 117.2: 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 117.2.1](#) 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 117.2.1 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 117.2.2 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 117.2.2 Regression Analysis for Cluster Sampling

Fit Statistics	
R-Square	0.9860
Root MSE	3.0488
Denominator DF	4

Estimated Regression Coefficients				
Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	-0.0191292	0.89204053	-0.02	0.9839
Population75	1.0546253	0.05167565	20.41	<.0001

Note: The 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 117.2.3 displays the regression results ignoring the clusters. Compared to the results in Output 117.2.2, the regression coefficient estimates are the same. However, without using clusters, the regression coefficients have a smaller variance estimate, as in Output 117.2.3. 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 117.2.2](#); without using the clusters, the estimate is 1.05, but with the estimated standard error 0.04, as displayed in [Output 117.2.3](#). To estimate the variance of the regression coefficients correctly, you should include the clustering information in the regression analysis.

Output 117.2.3 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				
Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	-0.0191292	0.67417606	-0.03	0.9775
Population75	1.0546253	0.03668414	28.75	<.0001

Note: The degrees of freedom for the t tests is 31.

Example 117.3: 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 117.2](#).

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 117.3.1 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 117.3.1 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

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

Example 117.4: 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 117.12.

Table 117.12 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;

```

Output 117.4.1 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 117.4.1 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

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 117.4.2 displays tests of model effects and the estimated regression coefficients.

Output 117.4.2 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.

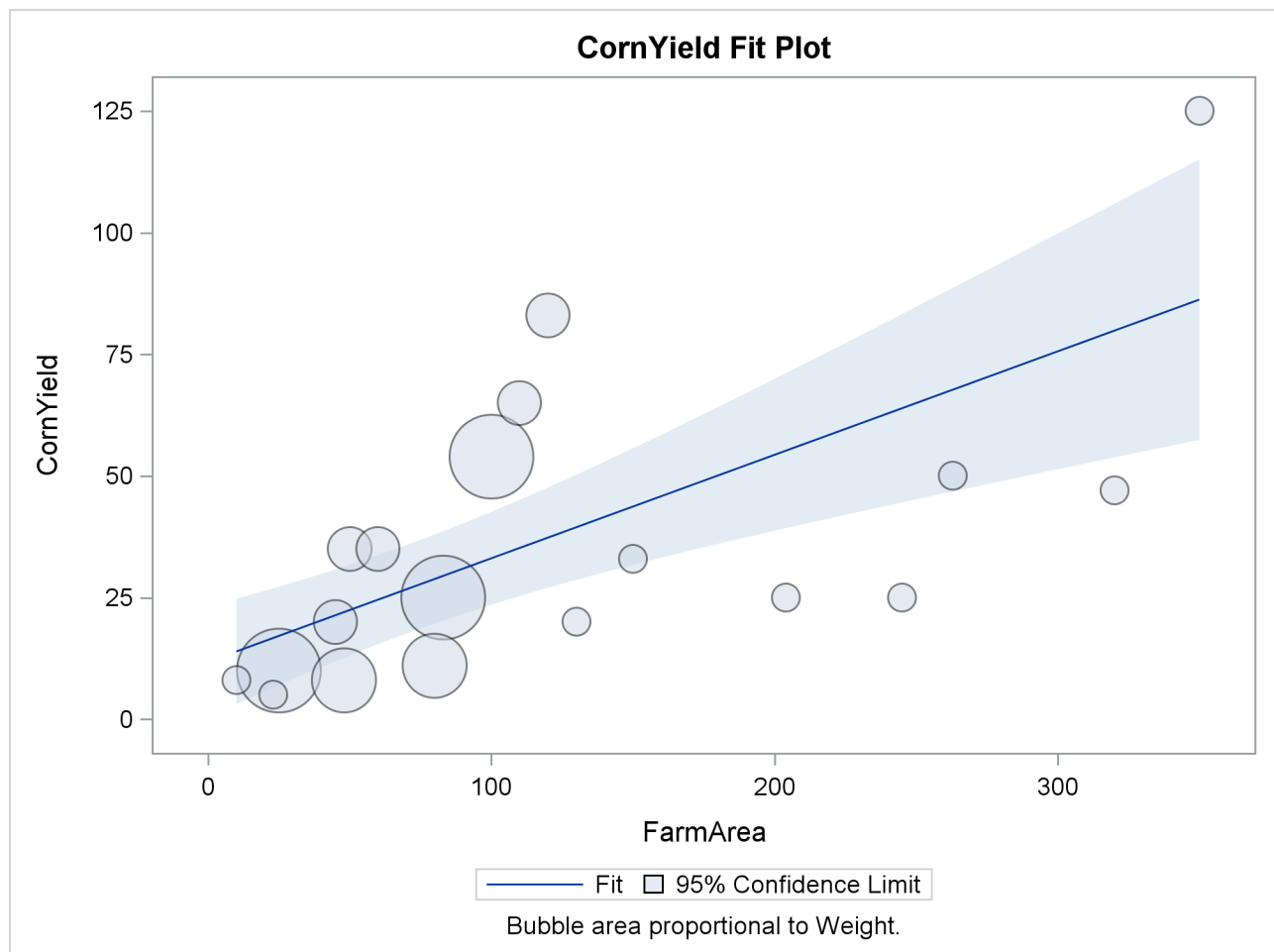
Estimated Regression Coefficients				
Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	11.8162978	5.31981027	2.22	0.0433
FarmArea	0.2126576	0.04560949	4.66	0.0004

Note: The 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 117.4.3 displays the fit of the regression.

Output 117.4.3 Regression Fitting



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:

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

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

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 117.4.4 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 117.4.4 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

Output 117.4.4 *continued*

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 degrees of freedom for the t tests is 14.				
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 117.4.5 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 117.4.5 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 117.4.5 *continued*

Estimated Regression Coefficients				
Parameter	Estimate	Standard 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 degrees of freedom for the t tests is 14.

Covariance of Estimated Regression Coefficients				
	State Iowa	State 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 117.5: Regression Estimator for Stratified Sample

This example uses the corn yield data set FARMS from [Example 117.4](#) to illustrate how to construct a regression estimator for a stratified sample design.

As in [Example 117.3](#), 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 117.13](#). You want to estimate the total corn yield by using this information under the three linear models given in [Example 117.4](#).

Table 117.13 Information for Each Stratum

Stratum	State	Region	Number of Farms		Total Farm Area
			Population	Sample	
1	Iowa	1	100	3	13,200
2		2	50	5	
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 117.13](#) (21950); and the coefficients for effect State*Region are the total number of farms in each strata (as displayed in [Table 117.13](#)).

[Output 117.5.1](#) 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 117.5.1 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

Label	Estimate		Standard		
	Estimate	Error	DF	t Value	Pr > t
Estimate of CornYield under Model I	7463.52	926.84	14	8.05	<.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 117.4](#)). 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 117.13. 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 117.5.2 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 117.5.2 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

Label	Estimate		Standard		
	Estimate	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 117.13](#).

[Output 117.5.3](#) 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 117.5.3 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

Label	Estimate		Standard			Pr > t
	Estimate	Error	DF	t Value		
Total CornYield in Iowa under Model III	6246.11	851.27	14	7.34		<.0001

Example 117.6: 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 9664.

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 117.6.1 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 9671).

Output 117.6.1 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 117.6.2 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 9664).

Output 117.6.3 displays the parameter estimates and the tests of the significance of the model effects.

Output 117.6.2 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 117.6.3 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 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 117.6.4 does not contain the stratum collapse information displayed in Output 117.6.1, and the denominator degrees of freedom are 3 instead of 4.

Output 117.6.4 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 117.6.5](#), although the fourth stratum and the fifth stratum contain only one observation, no stratum collapse occurs.

Output 117.6.5 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 117.6.6](#), are different from those in [Output 117.6.3](#), as are the tests of the significance of model effects.

Output 117.6.6 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				
Parameter	Estimate	Standard 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 degrees of freedom for the t tests is 3.

Example 117.7: 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 116.2](#) in Chapter 116, “[The SURVEYREG 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. For more information about the survey and the data sets, see the Center of Disease Control and Prevention's website (<http://www.cdc.gov/>).

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

... 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. The PLOT= option in the PROC statement requests that weights be represented as a heat map with hexagonal bins.

```

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

```

Output 117.7.1 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 117.7.1 Domain Analysis Among Cancer Patients

Study of Body Weight and Age among Cancer Patients

The SURVEYREG Procedure

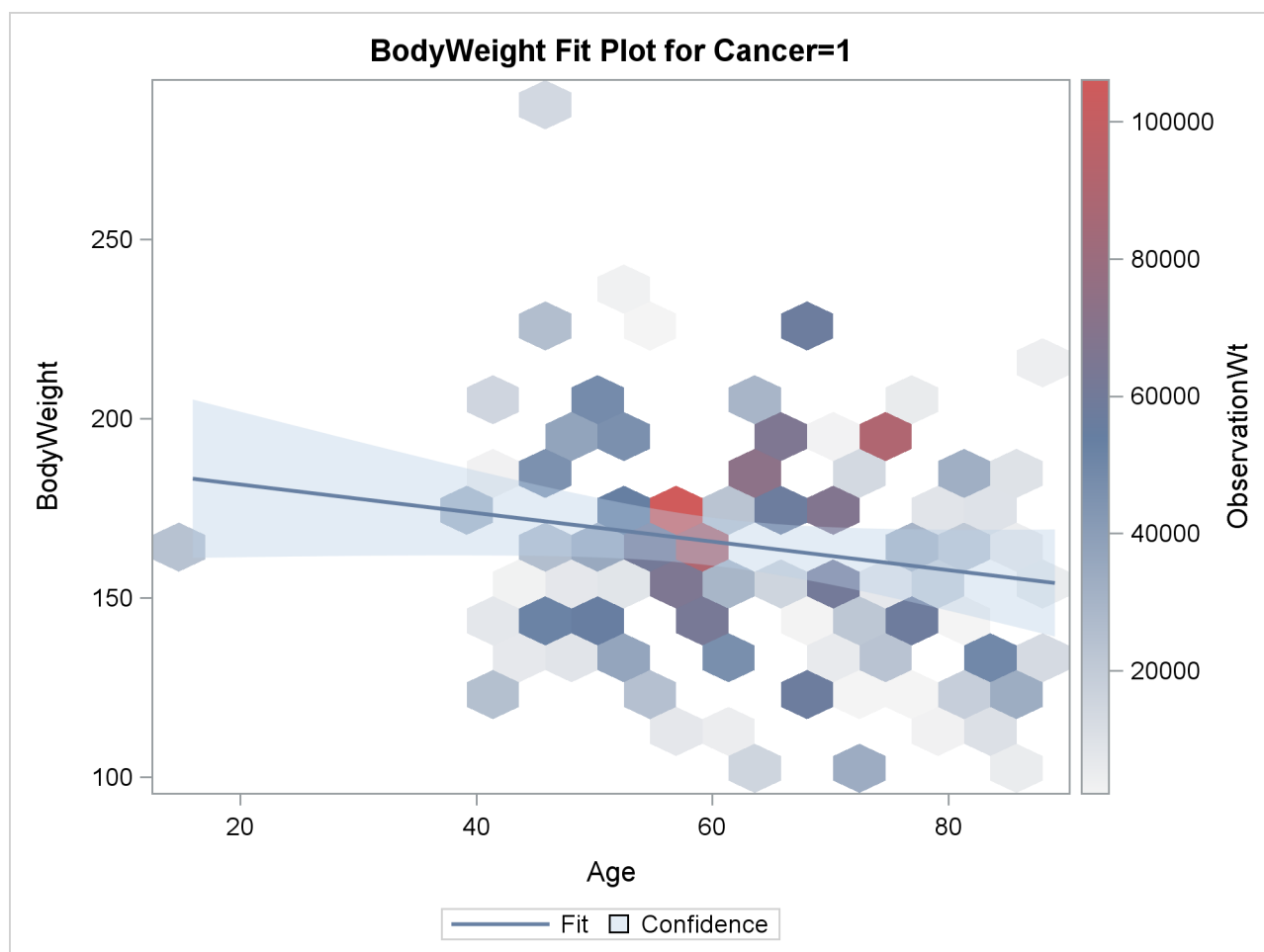
Cancer=1

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				
Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	189.614789	14.9467889	12.69	<.0001
Age	-0.398556	0.2398447	-1.66	0.0971

Note: The 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 117.7.2.

Output 117.7.2 Regression Fitting**Example 117.8: Comparing Domain Statistics**

Recall the example in the section “[Getting Started: SURVEYREG Procedure](#)” on page 9625, 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 117.8.1 shows the estimated spending among male and female students.

Output 117.8.1 Estimated Domain Means
The SURVEYMEANS Procedure

Statistics for Gender Domains					
Gender	Variable	N	Mean	Std Error	
				of Mean	95% CL for Mean
F	Spending	19	9.376111	1.077927	7.19202418 11.5601988
M	Spending	21	8.923052	1.003423	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 117.8.2 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 117.8.1 which are produced by PROC SURVEYMEANS.

Output 117.8.2 Domain Means between Gender**Ice Cream Spending Analysis
Compare Domain Statistics****The SURVEYREG Procedure****Regression Analysis for Dependent Variable Spending**

Gender Least Squares Means					
		Standard			
Gender	Estimate	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 117.8.3 shows the estimated difference for weekly ice cream 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 117.8.3 Domain Means Comparison

Differences of Gender Least Squares Means					
		Standard			
Gender	_Gender	Estimate	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;
```

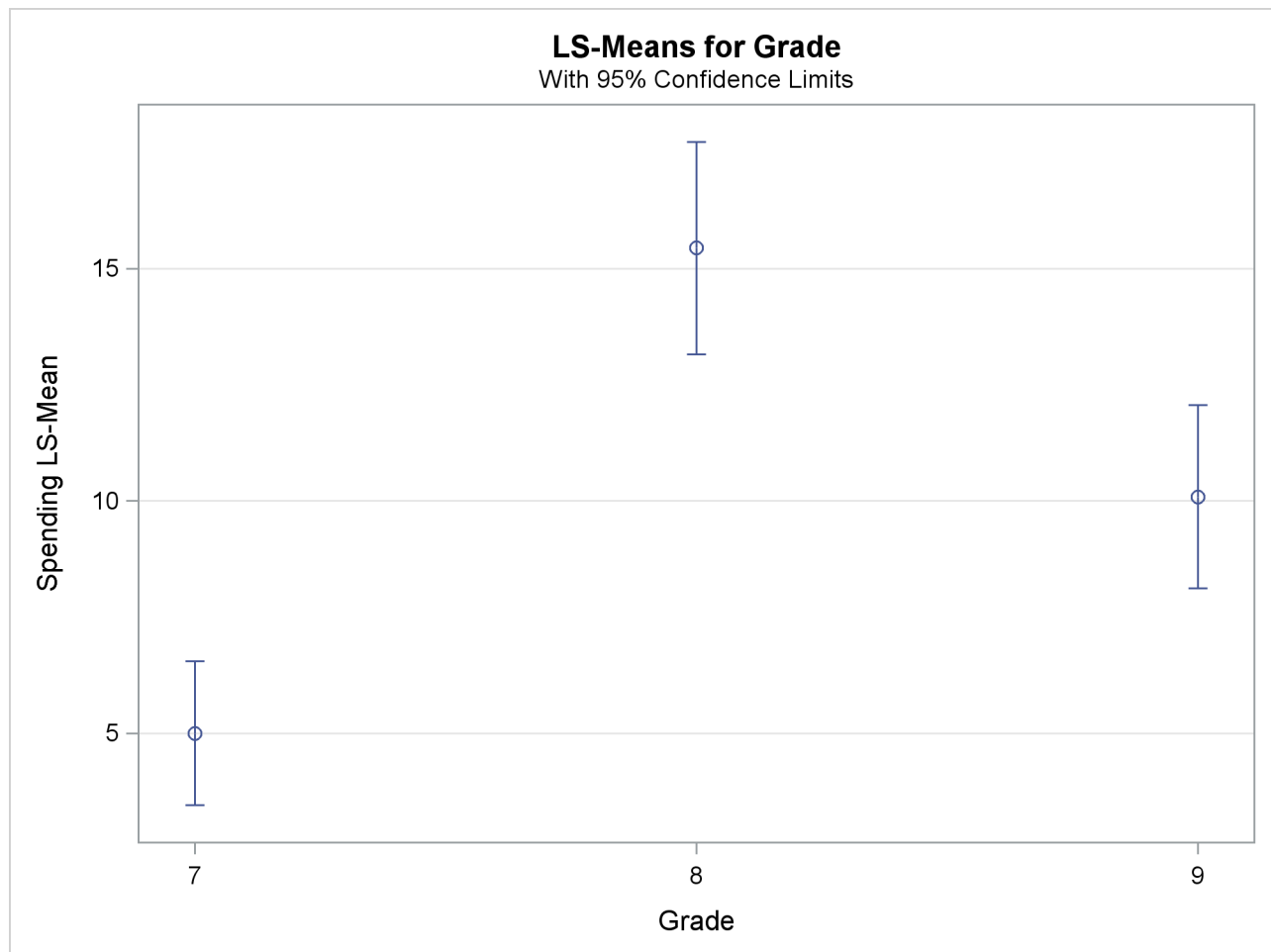
The Grade is specified in the CLASS statement to be used as an effect in the MODEL statement. The DIFF option in the LSMEANS statement requests that the procedure compute the difference among the domain means for the effect Grade. The ODS GRAPHICS statement enables ODS to create graphics. The PLOTS=(DIFF 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 117.8.4 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 117.8.4 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 117.8.5 plots the weekly spending results that are shown in [Output 117.8.4](#).

Output 117.8.5 Plot of Means of Ice Cream Spending within Grades

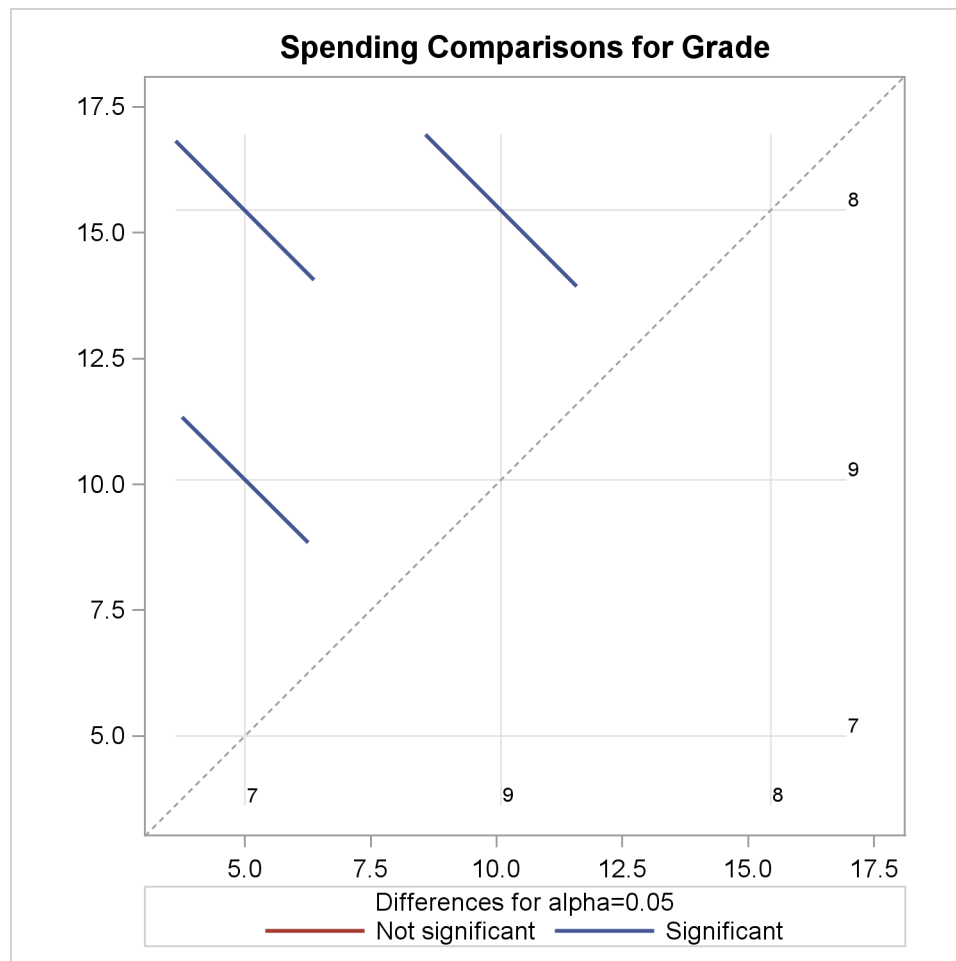
Output 117.8.6 displays pairwise comparisons for weekly ice scream spending among grades. All the differences are significant based on t tests.

Output 117.8.6 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 117.8.7 plots the comparisons that are shown in Output 117.8.6.

Output 117.8.7 Plot of Pairwise Comparisons of Spending among Grades



In Output 117.8.7, 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 117.8.6.

Example 117.9: Variance Estimate Using the Jackknife Method

This example uses the stratified sample from the section “Getting Started: SURVEYREG Procedure” on page 9625 to illustrate how to estimate the variances with replication methods.

As shown in the section “Stratified Sampling” on page 9627, 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 117.9.1 shows the summary of the data and the variance estimation method. There are a total of 40 replicates generated by the procedure.

Output 117.9.1 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

Design Summary	
Number of Strata	3

Variance Estimation	
Method	Jackknife
Number of Replicates	40

Output 117.9.2 displays the parameter estimates and their standard errors, as well as the tests of model effects that use the jackknife method.

Output 117.9.2 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 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 117.9.3 prints the first 6 observation in the output data set JKWeights, which contains the replicate weights.

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.

Output 117.9.3 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	RepWt_5
1	7	7	39	2	0.010965	91.200	0.000	96.000	91.200	91.200	96.000
2	7	7	38	1	0.010965	91.200	96.000	0.000	91.200	91.200	96.000
3	8	12	47	1	0.008780	113.889	113.889	113.889	0.000	113.889	113.889
4	9	10	47	4	0.009557	104.636	104.636	104.636	104.636	0.000	104.636
5	7	1	34	4	0.010965	91.200	96.000	96.000	91.200	91.200	0.000
6	7	10	43	2	0.010965	91.200	96.000	96.000	91.200	91.200	96.000

Obs	RepWt_6	RepWt_7	RepWt_8	RepWt_9	RepWt_10	RepWt_11	RepWt_12	RepWt_13	RepWt_14
1	96.000	96.000	91.200	91.200	96.000	96.000	91.200	91.200	96.000
2	96.000	96.000	91.200	91.200	96.000	96.000	91.200	91.200	96.000
3	113.889	113.889	128.125	128.125	113.889	113.889	113.889	128.125	113.889
4	104.636	104.636	104.636	104.636	104.636	104.636	115.100	104.636	104.636
5	96.000	96.000	91.200	91.200	96.000	96.000	91.200	91.200	96.000
6	0.000	96.000	91.200	91.200	96.000	96.000	91.200	91.200	96.000

Obs	RepWt_15	RepWt_16	RepWt_17	RepWt_18	RepWt_19	RepWt_20	RepWt_21	RepWt_22	RepWt_23
1	96.000	96.000	91.200	91.200	91.200	91.200	91.200	91.200	96.000
2	96.000	96.000	91.200	91.200	91.200	91.200	91.200	91.200	96.000
3	113.889	113.889	113.889	128.125	128.125	113.889	113.889	113.889	113.889
4	104.636	104.636	115.100	104.636	104.636	115.100	115.100	115.100	104.636
5	96.000	96.000	91.200	91.200	91.200	91.200	91.200	91.200	96.000
6	96.000	96.000	91.200	91.200	91.200	91.200	91.200	91.200	96.000

Obs	RepWt_24	RepWt_25	RepWt_26	RepWt_27	RepWt_28	RepWt_29	RepWt_30	RepWt_31	RepWt_32
1	96.000	96.000	91.200	91.200	91.200	96.000	96.000	96.000	96.000
2	96.000	96.000	91.200	91.200	91.200	96.000	96.000	96.000	96.000
3	113.889	113.889	113.889	128.125	113.889	113.889	113.889	113.889	113.889
4	104.636	104.636	115.100	104.636	115.100	104.636	104.636	104.636	104.636
5	96.000	96.000	91.200	91.200	91.200	96.000	96.000	96.000	96.000
6	96.000	96.000	91.200	91.200	91.200	96.000	96.000	96.000	96.000

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

References

- Beaumont, J.-F., and Patak, Z. (2012). “On the Generalized Bootstrap for Sample Surveys with Special Attention to Poisson Sampling.” *International Statistical Review* 80:127–148.
- Brick, J. M., and Kalton, G. (1996). “Handling Missing Data in Survey Research.” *Statistical Methods in Medical Research* 5:215–238.
- Cochran, W. G. (1977). *Sampling Techniques*. 3rd ed. New York: John Wiley & Sons.
- 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 ed. Boston: Brooks/Cole.
- Mashreghi, Z., Haziza, D., and Léger, C. (2016). “A Survey of Bootstrap Methods in Finite Population Sampling.” *Statistics Surveys* 10:1–52.
- McCarthy, P. J., and Snowden, C. B. (1985). *The Bootstrap and Finite Population Sampling*. Vital and Health Statistics, Series 2, no. 95. DHHS Publication Number 85–1369. Washington, DC: US Government Printing Office. https://www.cdc.gov/nchs/data/series/sr_02/sr02_095.pdf.
- 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., and Wu, C. F. J. (1988). "Resampling Inference with Complex Survey Data." *Journal of the American Statistical Association* 83:231–241.
- Rao, J. N. K., Wu, C. F. J., and Yue, K. (1992). "Some Recent Work on Resampling Methods for Complex Surveys." *Survey Methodology* 18:209–217.
- 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.
- Shao, J., and Tu, D. (1995). *The Jackknife and Bootstrap*. New York: Springer-Verlag.
- Sitter, R. R. (1992a). "Comparing Three Bootstrap Methods for Survey Data." *Canadian Journal of Statistics* 20:135–154.
- Sitter, R. R. (1992b). "A Resampling Procedure for Complex Survey Data." *Journal of the American Statistical Association* 87:755–765.
- Wolter, K. M. (2007). *Introduction to Variance Estimation*. 2nd ed. New York: Springer.
- Woodruff, R. S. (1971). "A Simple Method for Approximating the Variance of a Complicated Estimate." *Journal of the American Statistical Association* 66:411–414.

Subject Index

- ADJRSQ
 - SURVEYREG procedure, [9653](#)
- adjusted R-square
 - SURVEYREG procedure, [9665](#)
- alpha level
 - SURVEYREG procedure, [9633](#), [9656](#)
- analysis of variance
 - SURVEYREG procedure, [9664](#)
- ANOVA
 - SURVEYREG procedure, [9653](#), [9664](#)
- balanced repeated replication
 - SURVEYREG procedure, [9668](#)
 - variance estimation (SURVEYREG), [9668](#)
- Bootstrap
 - SURVEYREG procedure, [9667](#)
- Bootstrap Method
 - variance estimation (SURVEYREG), [9667](#)
- bootstrap replicate weights
 - SURVEYREG procedure, [9667](#)
- Bootstrap variance estimation
 - SURVEYREG procedure, [9667](#)
- BRR
 - SURVEYREG procedure, [9668](#)
- BRR variance estimation
 - SURVEYREG procedure, [9668](#)
- bubble plots
 - SURVEYREG procedure, [9634](#)
- cluster sampling
 - SURVEYREG procedure, [9684](#)
- clustering
 - SURVEYREG procedure, [9644](#)
- computational details
 - SURVEYREG procedure, [9662](#)
- computational resources
 - SURVEYREG procedure, [9673](#)
- confidence level
 - SURVEYREG procedure, [9633](#)
- confidence limits
 - SURVEYREG procedure, [9653](#)
- contrasts
 - SURVEYREG procedure, [9645](#), [9673](#)
- creating bootstrap replicates
 - SURVEYREG procedure, [9639](#)
- degrees of freedom
 - SURVEYREG procedure, [9671](#)
- design effects
 - SURVEYREG procedure, [9663](#)
- design information, [9661](#)
- domain analysis
 - SURVEYREG procedure, [9673](#), [9701](#), [9704](#)
- domain means comparison
 - SURVEYREG procedure, [9704](#)
- donor stratum
 - SURVEYREG procedure, [9670](#)
- effect testing
 - SURVEYREG procedure, [9672](#)
- Fay coefficient
 - SURVEYREG procedure, [9669](#)
- Fay's BRR method
 - variance estimation (SURVEYREG), [9669](#)
- finite population correction
 - SURVEYREG procedure, [9637](#), [9661](#)
- fit plots
 - SURVEYREG procedure, [9634](#)
- Hadamard matrix
 - SURVEYREG procedure, [9671](#)
- heat map plots
 - SURVEYREG procedure, [9634](#)
- initial seed
 - SURVEYREG procedure, [9640](#)
- jackknife
 - SURVEYREG procedure, [9670](#)
- jackknife coefficients
 - SURVEYREG procedure, [9670](#)
 - SURVEYREG procedure, [9675](#)
- jackknife variance estimation
 - SURVEYREG procedure, [9670](#)
- linearization method
 - SURVEYREG procedure, [9666](#)
- missing values
 - SURVEYREG procedure, [9633](#), [9660](#)
- MSE
 - SURVEYREG procedure, [9665](#)
- multiple R-square
 - SURVEYREG procedure, [9665](#)
- number of replicates
 - SURVEYREG procedure, [9667–9670](#)

- ODS graph names
 - SURVEYREG procedure, 9681
- ODS Graphics
 - SURVEYREG procedure, 9634, 9681
- ODS table names
 - SURVEYREG procedure, 9680
- options summary
 - EFFECT statement, 9648
 - ESTIMATE statement, 9649
- output data sets
 - SURVEYREG procedure, 9674
- output jackknife coefficient
 - SURVEYREG procedure, 9675
- output replicate weights
 - SURVEYREG procedure, 9675
- output table names
 - SURVEYREG procedure, 9680
- pooled stratum
 - SURVEYREG procedure, 9664
- primary sampling units (PSUs)
 - SURVEYREG procedure, 9661
- regression analysis
 - survey sampling, 9624
- regression coefficients
 - SURVEYREG procedure, 9663
- regression estimators
 - SURVEYREG procedure, 9687, 9694
- replicate weights
 - SURVEYREG procedure, 9665
- replication methods
 - SURVEYREG procedure, 9638
 - SURVEYREG procedure, 9665, 9710
- root MSE
 - SURVEYREG procedure, 9665
- sampling rates
 - SURVEYREG procedure, 9637, 9661
- sampling weights
 - SURVEYREG procedure, 9656, 9659
- seed
 - initial (SURVEYREG), 9640
- simple random sampling
 - SURVEYREG procedure, 9625, 9682
- singularity level
 - SURVEYREG procedure, 9646, 9654
- stratification
 - SURVEYREG procedure, 9658
- stratified sampling
 - SURVEYREG procedure, 9627, 9688
- stratum collapse
 - SURVEYREG procedure, 9664, 9697
- subdomain analysis, *see also* domain analysis
- subgroup analysis, *see also* domain analysis

- subpopulation analysis, *see also* domain analysis
- survey sampling, *see also* SURVEYREG procedure
 - regression analysis, 9624
- SURVEYREG procedure
 - balanced repeated replication, 9668
 - Bootstrap, 9667
 - bootstrap replicate weights, 9667
 - Bootstrap variance estimation, 9667
 - BRR, 9668
 - BRR variance estimation, 9668
 - clustering, 9644
 - creating bootstrap replicates, 9639
 - domain variable, 9646
 - donor stratum, 9670
 - Fay coefficient, 9669
 - Fay's BRR variance estimation, 9669
 - finite population correction, 9637, 9661
 - Hadamard matrix, 9671
 - initial seed, 9640
 - jackknife, 9670
 - jackknife coefficients, 9670
 - jackknife variance estimation, 9670
 - number of replicates, 9667–9670
 - ordering of effects, 9633
 - output replicate weights, 9675
 - population totals, 9637, 9661
 - primary sampling units (PSUs), 9661
 - replication methods, 9638
 - sampling rates, 9637, 9661
 - sampling weights, 9656, 9659
 - specify domain level, 9647
 - subset of domain levels, 9647
 - VARMETHOD=Bootstrap option, 9667
 - VARMETHOD=BRR option, 9668
 - VARMETHOD=JACKKNIFE option, 9670
 - VARMETHOD=JK option, 9670
 - weighting, 9656, 9659
- SURVEYREG procedure, 9624
 - ADJRSQ, 9653
 - adjusted R-square, 9665
 - alpha level, 9633, 9656
 - analysis of contrasts table, 9679
 - analysis of variance, 9664
 - ANOVA, 9653, 9664
 - ANOVA table, 9678
 - bubble plots, 9634
 - classification level table, 9678
 - cluster sampling, 9684
 - coefficients of contrast table, 9679
 - computational details, 9662
 - computational resources, 9673
 - confidence level, 9633
 - confidence limits, 9653
 - contrasts, 9645, 9673

covariance of estimated regression coefficients
table, 9679

data summary table, 9675

degrees of freedom, 9671

design effects, 9663

design summary table, 9676

domain analysis, 9673, 9701, 9704

domain means comparison, 9704

domain summary table, 9676

effect testing, 9672

fit plots, 9634

fit statistics table, 9676

Hadamard matrix, 9679

heat map plots, 9634

inverse matrix of $X'X$, 9678

jackknife coefficients, 9675

linearization method, 9666

list of strata, 9659

missing values, 9633, 9660

MSE, 9665

multiple R-square, 9665

ODS graph names, 9681

ODS Graphics, 9634, 9681

output data sets, 9630, 9674

output jackknife coefficient, 9675

output table names, 9680

pooled stratum, 9664

regression coefficients, 9663

regression coefficients table, 9679

regression estimators, 9687, 9694

replicate weights, 9665

replication methods, 9665, 9710

root MSE, 9665

simple random sampling, 9625, 9682

singularity level, 9646, 9654

stratification, 9658

stratified sampling, 9627, 9688

stratum collapse, 9664, 9697

stratum information table, 9677

subpopulation analysis, 9701, 9704

Taylor series variance estimation, 9666

testing effect, 9672

tests of model effects table, 9678

variance estimation, 9665

variance estimation table, 9677

Wald test, 9672, 9673

$X'X$ matrix, 9678

Taylor series variance estimation

SURVEYREG procedure, 9666

testing effect

SURVEYREG procedure, 9672

variance estimation

Bootstrap (SURVEYREG), 9667

BRR (SURVEYREG), 9668

jackknife (SURVEYREG), 9670

SURVEYREG procedure, 9665

Taylor series (SURVEYREG), 9666

VARMETHOD=Bootstrap option

SURVEYREG procedure, 9667

VARMETHOD=BRR option

SURVEYREG procedure, 9668

VARMETHOD=JACKKNIFE option

SURVEYREG procedure, 9670

VARMETHOD=JK option

SURVEYREG procedure, 9670

Wald test

SURVEYREG procedure, 9672, 9673

weighting

SURVEYREG procedure, 9656, 9659

Syntax Index

- ADJRSQ option
 - MODEL statement (SURVEYREG), [9653](#)
- ALPHA= option
 - OUTPUT statement (SURVEYREG), [9656](#)
 - PROC SURVEYREG statement, [9633](#)
- ANOVA option
 - MODEL statement (SURVEYREG), [9653](#)
- BY statement
 - SURVEYREG procedure, [9643](#)
- CLASS statement
 - SURVEYREG procedure, [9643](#)
- CLPARM option
 - MODEL statement (SURVEYREG), [9653](#)
- CLUSTER statement
 - SURVEYREG procedure, [9644](#)
- CONTRAST statement
 - SURVEYREG procedure, [9645](#)
- COVB option
 - MODEL statement (SURVEYREG), [9654](#)
- DATA= option
 - PROC SURVEYREG statement, [9633](#)
- DEFF option
 - MODEL statement (SURVEYREG), [9654](#)
- DF= option
 - MODEL statement (SURVEYREG), [9654](#)
 - REPWEIGHTS statement (SURVEYREG), [9656](#)
- DOMAIN statement
 - SURVEYREG procedure, [9646](#)
- E option
 - CONTRAST statement (SURVEYREG), [9645](#)
- EFFECT statement
 - SURVEYREG procedure, [9648](#)
- ESTIMATE statement
 - SURVEYREG procedure, [9649](#)
- FAY= option
 - VARMETHOD=BRR (PROC SURVEYREG statement), [9640](#)
- HADAMARD= option
 - VARMETHOD=BRR (PROC SURVEYREG statement), [9640](#)
- INVERSE option
 - MODEL statement (SURVEYREG), [9654](#)
- JKCOEFS= option
 - REPWEIGHTS statement (SURVEYREG), [9656](#)
- keyword= option
 - OUTPUT statement (SURVEYREG), [9655](#)
- LCLM keyword
 - OUTPUT statement (SURVEYREG), [9655](#)
- LIST option
 - STRATA statement (SURVEYREG), [9659](#)
- LSMESTIMATE statement
 - SURVEYREG procedure, [9651](#)
- MH= method-option
 - PROC SURVEYREG statement, [9639](#)
- MISSING option
 - PROC SURVEYREG statement, [9633](#)
- MODEL statement
 - SURVEYREG procedure, [9653](#)
- N= option
 - PROC SURVEYREG statement, [9637](#)
- NAMELEN= option
 - PROC SURVEYREG statement, [9633](#)
- NBINS= global plot option
 - PROC SURVEYREG statement, [9635](#)
- NBINS= option
 - PROC SURVEYREG statement, [9636](#)
- NOCOLLAPSE option
 - STRATA statement (SURVEYREG), [9659](#)
- NOFILL option
 - CONTRAST statement (SURVEYREG), [9645](#)
- NOINT option
 - MODEL statement (SURVEYREG), [9654](#)
- NOMCAR option
 - PROC SURVEYREG statement, [9633](#)
- ORDER= option
 - PROC SURVEYREG statement, [9633](#)
- OUT= option
 - OUTPUT statement (SURVEYREG), [9655](#)
- OUTJKCOEFS= option
 - VARMETHOD=JACKKNIFE (PROC SURVEYREG statement), [9642](#)
- OUTPUT statement
 - SURVEYREG procedure, [9655](#)
- OUTWEIGHTS= option
 - VARMETHOD=BOOTSTRAP (PROC SURVEYREG statement), [9640](#)

- VARMETHOD=BRR (PROC SURVEYREG statement), [9641](#)
- VARMETHOD=JACKKNIFE (PROC SURVEYREG statement), [9642](#)
- PARMLABEL option
 - MODEL statement (SURVEYREG), [9654](#)
- PLOTS= option
 - PROC SURVEYREG statement, [9634](#)
- PLOTS=FIT option
 - PROC SURVEYREG statement, [9636](#)
- PLOTS=FIT(NBINS=) option
 - PROC SURVEYREG statement, [9636](#)
- PREDICTED keyword
 - OUTPUT statement (SURVEYREG), [9655](#)
- PRINTH option
 - VARMETHOD=BRR (PROC SURVEYREG statement), [9641](#)
- PROC SURVEYREG statement, *see* SURVEYREG procedure
- R= option
 - PROC SURVEYREG statement, [9637](#)
- RATE= option
 - PROC SURVEYREG statement, [9637](#)
- REF= option
 - CLASS statement (SURVEYREG), [9644](#)
- REPCOEF= option
 - REPWEIGHTS statement (SURVEYREG), [9657](#)
- REPS= option
 - VARMETHOD=BOOTSTRAP (PROC SURVEYREG statement), [9640](#)
 - VARMETHOD=BRR (PROC SURVEYREG statement), [9641](#)
- REPWEIGHTS statement
 - SURVEYREG procedure, [9656](#)
- RESIDUAL keyword
 - OUTPUT statement (SURVEYREG), [9655](#)
- SEED= method-option
 - PROC SURVEYREG statement, [9640](#)
- SHAPE= plot option
 - PROC SURVEYREG statement, [9637](#)
- SHAPE=HEXAGONAL option
 - PROC SURVEYREG statement, [9637](#)
- SHAPE=RECTANGULAR option
 - PROC SURVEYREG statement, [9637](#)
- SINGULAR= option
 - CONTRAST statement (SURVEYREG), [9646](#)
 - MODEL statement (SURVEYREG), [9654](#)
- SLICE statement
 - SURVEYREG procedure, [9658](#)
- SOLUTION option
 - MODEL statement (SURVEYREG), [9654](#)
- STB option

- MODEL statement (SURVEYREG), [9654](#)
- STD keyword
 - OUTPUT statement (SURVEYREG), [9655](#)
- STDP keyword
 - OUTPUT statement (SURVEYREG), [9655](#)
- STORE statement
 - SURVEYREG procedure, [9658](#)
- STRATA statement
 - SURVEYREG procedure, [9658](#)
- SUBGROUP statement
 - SURVEYREG procedure, [9646](#)
- SURVEYREG procedure, BY statement, [9643](#)
- SURVEYREG procedure, CLASS statement, [9643](#)
 - REF= option, [9644](#)
 - REF= variable option, [9644](#)
- SURVEYREG procedure, CLUSTER statement, [9644](#)
- SURVEYREG procedure, DOMAIN statement, [9646](#)
- SURVEYREG procedure, EFFECT statement, [9648](#)
- SURVEYREG procedure, ESTIMATE statement, [9649](#)
- SURVEYREG procedure, LSMESTIMATE statement, [9651](#)
- SURVEYREG procedure, PROC SURVEYREG statement
 - FAY= option (VARMETHOD=BRR), [9640](#)
 - HADAMARD= option (VARMETHOD=BRR), [9640](#)
 - M= method-option, [9639](#)
 - N= option, [9637](#)
 - ORDER= option, [9633](#)
 - OUTJKCOEF= option
 - (VARMETHOD=JACKKNIFE), [9642](#)
 - OUTWEIGHTS= option
 - (VARMETHOD=BOOTSTRAP), [9640](#)
 - OUTWEIGHTS= option (VARMETHOD=BRR), [9641](#)
 - OUTWEIGHTS= option
 - (VARMETHOD=JACKKNIFE), [9642](#)
 - PRINTH option (VARMETHOD=BRR), [9641](#)
 - RATE= option, [9637](#)
 - REPS= option (VARMETHOD=BOOTSTRAP), [9640](#)
 - REPS= option (VARMETHOD=BRR), [9641](#)
 - SEED= method-option, [9640](#)
 - TOTAL= option, [9637](#)
 - VARMETHOD= option, [9638](#)
 - VARMETHOD=BOOTSTRAP option, [9639](#)
 - VARMETHOD=BRR option, [9640](#)
 - VARMETHOD=JACKKNIFE option, [9642](#)
 - VARMETHOD=TAYLOR option, [9642](#)
- SURVEYREG procedure, REPWEIGHTS statement, [9656](#)
 - DF= option, [9656](#)
 - JKCOEF= option, [9656](#)
 - REPCOEF= option, [9657](#)

SURVEYREG procedure, SLICE statement, [9658](#)
SURVEYREG procedure, STORE statement, [9658](#)
SURVEYREG procedure, TEST statement, [9659](#)
SURVEYREG procedure, WEIGHT statement, [9659](#)
SURVEYREG procedure
syntax, [9631](#)

SURVEYREG procedure, CONTRAST statement,
[9645](#)

E option, [9645](#)

NOFILL option, [9645](#)

SINGULAR= option, [9646](#)

SURVEYREG procedure, MODEL statement, [9653](#)

ADJRSQ option, [9653](#)

ANOVA option, [9653](#)

CLPARM option, [9653](#)

COVB option, [9654](#)

DEFF option, [9654](#)

INVERSE option, [9654](#)

NOINT option, [9654](#)

PARMLABEL option, [9654](#)

SINGULAR= option, [9654](#)

SOLUTION option, [9654](#)

STB option, [9654](#)

VADJUST= option, [9654](#)

XPX option, [9655](#)

SURVEYREG procedure, MODEL statement
(SURVEYREG)

DF= option, [9654](#)

SURVEYREG procedure, OUTPUT statement, [9655](#)

ALPHA= option, [9656](#)

keyword= option, [9655](#)

LCLM keyword, [9655](#)

OUT= option, [9655](#)

PREDICTED keyword, [9655](#)

RESIDUAL keyword, [9655](#)

STD keyword, [9655](#)

STDP keyword, [9655](#)

UCLM keyword, [9655](#)

SURVEYREG procedure, PROC SURVEYREG
statement, [9632](#)

ALPHA= option, [9633](#)

DATA= option, [9633](#)

MISSING option, [9633](#)

NAMELEN= option, [9633](#)

NOMCAR option, [9633](#)

PLOTS= option, [9634](#)

PLOTS=FIT option, [9636](#)

TRUNCATE option, [9638](#)

SURVEYREG procedure, STRATA statement, [9658](#)

LIST option, [9659](#)

NOCOLLAPSE option, [9659](#)

TRUNCATE option

PROC SURVEYREG statement, [9638](#)

UCLM keyword

OUTPUT statement (SURVEYREG), [9655](#)

VADJUST= option

MODEL statement (SURVEYREG), [9654](#)

VARMETHOD= option

PROC SURVEYREG statement, [9638](#)

VARMETHOD=BOOTSTRAP option

PROC SURVEYREG statement, [9639](#)

VARMETHOD=BRR option

PROC SURVEYREG statement, [9640](#)

VARMETHOD=JACKKNIFE option

PROC SURVEYREG statement, [9642](#)

VARMETHOD=TAYLOR option

PROC SURVEYREG statement, [9642](#)

WEIGHT statement

SURVEYREG procedure, [9659](#)

WEIGHT= global plot option

PROC SURVEYREG statement, [9635](#)

WEIGHT= plot option

PROC SURVEYREG statement, [9636](#)

WEIGHT=BUBBLE option

PROC SURVEYREG statement, [9635](#), [9636](#)

WEIGHT=HEATMAP option

PROC SURVEYREG statement, [9635](#), [9636](#)

XPX option

MODEL statement (SURVEYREG), [9655](#)