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. The procedure 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).

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 7514. The section “Examples: SURVEYREG Procedure” on page 7557 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:

```sas
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:

```sas
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 90.1 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 90.1 Summary of Data

<table>
<thead>
<tr>
<th>Ice Cream Spending Analysis</th>
<th>Simple Random Sample Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>The SURVEYREG Procedure</td>
<td></td>
</tr>
<tr>
<td>Regression Analysis for Dependent Variable Spending</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Class</th>
<th>Variable</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kids</td>
<td></td>
<td>4</td>
<td>1 2 3 4</td>
</tr>
</tbody>
</table>

Figure 90.2 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 90.2 Testing Effects in the Regression

Tests of Model Effects

<table>
<thead>
<tr>
<th>Effect</th>
<th>Num DF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>119.15</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Intercept</td>
<td>1</td>
<td>153.32</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Income</td>
<td>1</td>
<td>324.45</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Kids</td>
<td>3</td>
<td>0.92</td>
<td>0.4385</td>
</tr>
</tbody>
</table>

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 90.3.
Figure 90.3 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 denominator degrees of freedom for the t tests is 39.
Matrix X'X is singular and a generalized inverse was used to solve the normal equations. Estimates are not unique.

Stratified Sampling

Suppose that the previous student sample is actually drawn using a stratified sample design. The strata are grades in the junior high school: 7, 8, and 9. Within strata, simple random samples are selected. Table 90.1 provides the number of students in each grade.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Number of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>1,824</td>
</tr>
<tr>
<td>8</td>
<td>1,025</td>
</tr>
<tr>
<td>9</td>
<td>1,151</td>
</tr>
<tr>
<td>Total</td>
<td>4,000</td>
</tr>
</tbody>
</table>

In order to analyze this sample by using PROC SURVEYREG, you need to input the stratification information by creating a SAS data set with the information in Table 90.1. The following SAS statements create such a data set called 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 total information.
In a stratified sample design, when the sampling rates in the strata are unequal, you need to use sampling weights to reflect this information. 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):

```
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;
  weight Weight;
run;
```

Comparing these statements to those in the section “Simple Random Sampling” on page 7507, 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 included in the output. The WEIGHT statement specifies the weight variable.

Figure 90.4 summarizes the data information, the sample design information, and the fit information. Note that, due to the stratification, the denominator degrees of freedom for $F$ tests and $t$ tests are 37, which is different from the analysis in Figure 90.1.

**Figure 90.4** Summary of the Regression

<table>
<thead>
<tr>
<th>Ice Cream Spending Analysis</th>
<th>Stratified Sample Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>The SURVEYREG Procedure</td>
<td></td>
</tr>
<tr>
<td>Regression Analysis for Dependent Variable Spending</td>
<td></td>
</tr>
<tr>
<td>Data Summary</td>
<td></td>
</tr>
<tr>
<td>Number of Observations</td>
<td>40</td>
</tr>
<tr>
<td>Sum of Weights</td>
<td>4000.0</td>
</tr>
<tr>
<td>Weighted Mean of Spending</td>
<td>9.14130</td>
</tr>
<tr>
<td>Weighted Sum of Spending</td>
<td>36565.2</td>
</tr>
</tbody>
</table>
Chapter 90: The SURVEYREG Procedure

Figure 90.4 continued

<table>
<thead>
<tr>
<th>Design Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Strata</td>
</tr>
<tr>
<td>Fit Statistics</td>
</tr>
<tr>
<td>R-square</td>
</tr>
<tr>
<td>Root MSE</td>
</tr>
<tr>
<td>Denominator DF</td>
</tr>
</tbody>
</table>

For each stratum, Figure 90.5 displays the value of identifying variables, the number of observations (sample size), the total population size, and the calculated sampling rate or fraction.

Figure 90.5 Stratification and Classification Information

<table>
<thead>
<tr>
<th>Stratum Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stratum Population Sampling</td>
</tr>
<tr>
<td>Index</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
</tbody>
</table>

Class Level Information

<table>
<thead>
<tr>
<th>Class Variable</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kids</td>
<td>4</td>
<td>1 2 3 4</td>
</tr>
</tbody>
</table>

Figure 90.6 displays the tests for the significance of model effects under the stratified sample design. The Income effect is strongly significant, while the Kids effect is not significant at the 5% level.

Figure 90.6 Testing Effects

<table>
<thead>
<tr>
<th>Tests of Model Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect</td>
</tr>
<tr>
<td>Model</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>Income</td>
</tr>
<tr>
<td>Kids</td>
</tr>
</tbody>
</table>

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

The regression coefficient estimates for the stratified sample, along with their standard errors and associated t tests, are displayed in Figure 90.7.
Figure 90.7 Regression Coefficients

| Parameter | Estimate | Standard Error | t Value | Pr > |t| |
|-----------|----------|----------------|---------|------|--------|
| Intercept | -26.086882 | 2.44108058 | -10.69 | <.0001 |
| Income    | 0.776699  | 0.04295904 | 18.08  | <.0001 |
| Kids 1    | 0.888631  | 1.07000634 | 0.83   | 0.4116 |
| Kids 2    | 1.545726  | 1.20815863 | 1.28   | 0.2087 |
| Kids 3    | -0.526817 | 1.32748011 | -0.40  | 0.6938 |
| Kids 4    | 0.000000  | 0.00000000 | .      | .     |

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

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.

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 7531 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 90.7) in the previous section in an output data set, you use the ODS OUTPUT statement as follows:

```sas
OPTIONS nodate nocenter nodate date2 name=32; 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;
  weight Weight;
  ods output ParameterEstimates = MyParmEst;
run;
```

The statement

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

requests that the “ParameterEstimates” table that appears in Figure 90.7 be placed into a SAS data set MyParmEst.
The PRINT procedure displays observations of the data set MyParmEst:

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

Figure 90.8 displays the observations in the data set MyParmEst. The section “ODS Table Names” on page 7555 gives the complete list of the tables produced by PROC SURVEYREG.

**Figure 90.8 The Data Set MyParmEst**

<table>
<thead>
<tr>
<th>Obs</th>
<th>Parameter</th>
<th>Estimate</th>
<th>StdErr</th>
<th>DenDF</th>
<th>tValue</th>
<th>Probt</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intercept</td>
<td>-26.086882</td>
<td>2.44108058</td>
<td>37</td>
<td>-10.69</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>2</td>
<td>Income</td>
<td>0.776699</td>
<td>0.04295904</td>
<td>37</td>
<td>18.08</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>3</td>
<td>Kids 1</td>
<td>0.888631</td>
<td>1.07000634</td>
<td>37</td>
<td>0.83</td>
<td>0.4116</td>
</tr>
<tr>
<td>4</td>
<td>Kids 2</td>
<td>1.545726</td>
<td>1.20815863</td>
<td>37</td>
<td>1.28</td>
<td>0.2087</td>
</tr>
<tr>
<td>5</td>
<td>Kids 3</td>
<td>-0.526817</td>
<td>1.32748011</td>
<td>37</td>
<td>-0.40</td>
<td>0.6938</td>
</tr>
<tr>
<td>6</td>
<td>Kids 4</td>
<td>0.000000</td>
<td>0.00000000</td>
<td>37</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

**Syntax: SURVEYREG Procedure**

The following statements are available in PROC SURVEYREG:

```plaintext
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 lsmeanste-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 pre-
cede 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**

```plaintext
PROC SURVEYREG < options > ;
```

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

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

- **ALPHA=\(\alpha\)**
  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 \(\alpha\) 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 7536.

- **NOMCAR**
  requests that the procedure treat missing values in the variance computation as not missing completely at random (NOMCAR) for Taylor series variance estimation. When you specify the NOMCAR option, PROC SURVEYREG computes variance estimates by analyzing the nonmissing values as a domain or subpopulation, where the entire population includes both nonmissing and missing domains. See the section “Missing Values” on page 7536 for more details.
By default, PROC SURVEYREG completely excludes an observation from analysis if that observation has a missing value, unless you specify the MISSING option. Note that the NOMCAR option has no effect on a classification variable when you specify the MISSING option, which treats missing values as a valid nonmissing level.

The NOMCAR option applies only to Taylor series variance estimation. The replication methods, which you request with the VARMETHOD=BRR and VARMETHOD=JACKKNIFE options, do not use the NOMCAR option.

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

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

This option also determines the sorting 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. With this option, the levels of such variables are ordered by their internal value.

The ORDER= option can take the following values:

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

By default, ORDER=FORMATTED. For ORDER=FORMATTED and ORDER=INTERNAL, the sort order is machine-dependent. For more information about sorting order, see the chapter on the SORT procedure in the *Base SAS Procedures Guide* and the discussion of BY-group processing in *SAS Language Reference: Concepts*.

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

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

If your sample design has multiple stages, you should specify the first-stage sampling rate, which is the ratio of the number of PSUs selected to the total number of PSUs in the population.
For a nonstratified sample design, or for a stratified sample design with the same sampling rate in all strata, you should specify a nonnegative value for the RATE= option. If your design is stratified with different sampling rates in the strata, then you should name a SAS data set that contains the stratification variables and the sampling rates. See the section “Specification of Population Totals and Sampling Rates” on page 7537 for more details.

The value in the RATE= option or the values of _RATE_ in the secondary data set must be nonnegative numbers. You can specify value as a number between 0 and 1. Or you can specify value in percentage form as a number between 1 and 100, and PROC SURVEYREG converts that number to a proportion. The procedure treats the value 1 as 100%, and not the percentage form 1%.

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

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

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

For a nonstratified sample design, or for a stratified sample design with the same population total in all strata, you should specify a positive value for the TOTAL= option. If your sample design is stratified with different population totals in the strata, then you should name a SAS data set that contains the stratification variables and the population totals. See the section “Specification of Population Totals and Sampling Rates” on page 7537 for more details.

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

TRUNCATE

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

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

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

For VARMETHOD=BRR and VARMETHOD=JACKKNIFE you can specify method-options in parentheses. Table 90.2 summarizes the available method-options.
### Table 90.2 Variance Estimation Options

<table>
<thead>
<tr>
<th>METHOD</th>
<th>Variance Estimation Method</th>
<th>Method-Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRR</td>
<td>Balanced repeated replication</td>
<td>FAY &lt;=value &gt; HADAMARD=SAS-data-set OUTWEIGHTS=SAS-data-set PRINTH REPS=number</td>
</tr>
<tr>
<td>jackknife</td>
<td>Jackknife</td>
<td>OUTJSCOEF=SAS-data-set OUTWEIGHTS=SAS-data-set</td>
</tr>
<tr>
<td>TAYLOR</td>
<td>Taylor series linearization</td>
<td>None</td>
</tr>
</tbody>
</table>

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

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

The following values are available for the VARMETHOD= option:

- **BRR** requests balanced repeated replication (BRR) variance estimation. The BRR method requires a stratified sample design with two primary sampling units (PSUs) per stratum. See the section “Balanced Repeated Replication (BRR) Method” on page 7543 for more information.

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

  - **FAY <=value >** requests Fay’s method, a modification of the BRR method, for variance estimation. See the section “Fay’s BRR Method” on page 7544 for more information.

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

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

  names a SAS data set that contains the Hadamard matrix for BRR replicate construction. If you do not provide a Hadamard matrix with the HADAMARD= method-option, PROC SURVEYREG generates an appropriate Hadamard matrix for replicate construction. See the sections “Balanced Repeated Replication (BRR) Method” on page 7543 and “Hadamard Matrix” on page 7546 for details.

  If a Hadamard matrix of a given dimension exists, it is not necessarily unique. Therefore, if you want to use a specific Hadamard matrix, you must provide the matrix as a SAS data set in the HADAMARD= method-option.
In the HADAMARD= input data set, each variable corresponds to a column of the Hadamard matrix, and each observation corresponds to a row of the matrix. You can use any variable names in the HADAMARD= data set. All values in the data set must equal either 1 or -1. You must ensure that the matrix you provide is indeed a Hadamard matrix—that is, \( A' A = RI \), where \( A \) is the Hadamard matrix of dimension \( R \) and \( I \) is an identity matrix. PROC SURVEYREG does not check the validity of the Hadamard matrix that you provide.

The HADAMARD= input data set must contain at least \( H \) variables, where \( H \) denotes the number of first-stage strata in your design. If the data set contains more than \( H \) variables, the procedure uses only the first \( H \) variables. Similarly, the HADAMARD= input data set must contain at least \( H \) observations.

If you do not specify the REPS= method-option, then the number of replicates is taken to be the number of observations in the HADAMARD= input data set. If you specify the number of replicates—for example, \( \text{REPS=}n\text{reps} \)—then the first \( n\text{reps} \) observations in the HADAMARD= data set are used to construct the replicates.

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

\textbf{OUTWEIGHTS=}SAS-data-set

names a SAS data set that contains replicate weights. See the section “Balanced Repeated Replication (BRR) Method” on page 7543 for information about replicate weights. See the section “Replicate Weights Output Data Set” on page 7550 for more details about the contents of the OUTWEIGHTS= data set.

The \textbf{OUTWEIGHTS=} method-option is not available when you provide replicate weights with the REPWEIGHTS statement.

\textbf{PRINTH}

displays the Hadamard matrix.

When you provide your own Hadamard matrix with the HADAMARD= method-option, only the rows and columns of the Hadamard matrix that are used by the procedure are displayed. See the sections “Balanced Repeated Replication (BRR) Method” on page 7543 and “Hadamard Matrix” on page 7546 for details.

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

\textbf{REPS=}number

specifies the number of replicates for BRR variance estimation. The value of \( \text{number} \) must be an integer greater than 1.
If you do not provide a Hadamard matrix with the HADAMARD= method-option, the number of replicates should be greater than the number of strata and should be a multiple of 4. See the section “Balanced Repeated Replication (BRR) Method” on page 7543 for more information. If a Hadamard matrix cannot be constructed for the REPS= value that you specify, the value is increased until a Hadamard matrix of that dimension can be constructed. Therefore, it is possible for the actual number of replicates used to be larger than the REPS= value that you specify.

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

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

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

JACKKNIFE | JK < (method-options) > requests variance estimation by the delete-1 jackknife method. See the section “Jackknife Method” on page 7545 for details. If you provide replicate weights with a REPWEIGHTS statement, VARMETHOD=JACKKNIFE is the default variance estimation method.

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

OUTJKCOEFS=SAS-data-set
names a SAS data set that contains jackknife coefficients. See the section “Jackknife Method” on page 7545 for information about jackknife coefficients. See the section “Jackknife Coefficients Output Data Set” on page 7550 for more details about the contents of the OUTJKCOEFS= data set.

OUTWEIGHTS=SAS-data-set
names a SAS data set that contains replicate weights. See the section “Jackknife Method” on page 7545 for information about replicate weights. See the section “Replicate Weights Output Data Set” on page 7550 for more details about the contents of the OUTWEIGHTS= data set.

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

TAYLOR
requests Taylor series variance estimation. This is the default method if you do not specify the VARMETHOD= option or a REPWEIGHTS statement. See the section “Taylor Series (Linearization)” on page 7542 for more information.
BY Statement

BY variables;

You can specify a BY statement with PROC SURVEYREG to obtain separate analyses on observations in groups that are defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. If you specify more than one BY statement, only the last one specified is used.

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

- Sort the data by using the SORT procedure with a similar BY statement.
- Specify the NOTSORTED or DESCENDING option in the BY statement for the SURVEYREG procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables by using the DATASETS procedure (in Base SAS software).

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

For more information about BY-group processing, see the discussion in SAS Language Reference: Concepts. For more information about the DATASETS procedure, see the discussion in the Base SAS Procedures Guide.

CLASS Statement

CLASS variables;

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

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

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

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

You can use multiple CLASS statements to specify classification variables.

---

**CLUSTER Statement**

```
CLUSTER variables;
```

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

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

If your sample design has clustering at multiple stages, you should identify only the first-stage clusters (primary sampling units (PSUs)), in the CLUSTER statement. See the section “Primary Sampling Units (PSUs)” on page 7538 for more information.

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

When determining levels of a CLUSTER variable, an observation with missing values for this CLUSTER variable is excluded, unless you specify the MISSING option. For more information, see the section “Missing Values” on page 7536.

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 $H_0: L\beta = 0$, where $L$ is the vector or matrix you specify and $\beta$ 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 394 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 3186 in Chapter 41, “The GLM Procedure.”

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

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

In the CONTRAST statement,

- label identifies the contrast in the output. A label is required for every contrast specified. Labels must be enclosed in single quotes.
- effect identifies an effect that appears in the MODEL statement. You can use the INTERCEPT keyword as an effect when an intercept is fitted in the model. You do not need to include all effects that are in the MODEL statement.
- values are constants that are elements of L associated with the effect.

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

- **E** displays the entire coefficient L vector or matrix.
- **NOFILL** requests no filling in higher-order effects. When you specify only certain portions of L, by default PROC SURVEYREG constructs the remaining elements from the context. (For more information, see the section “Specification of ESTIMATE Expressions” on page 3207 in Chapter 41, “The GLM Procedure.”)

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

- **SINGULAR=value** tunes the estimability checking. If v is a vector, define ABS(v) to be the largest absolute value of the elements of v. For a row vector I of the matrix L, define

  \[ c = \begin{cases} 
  \text{ABS}(I) & \text{if } \text{ABS}(I) > 0 \\
  1 & \text{otherwise} \end{cases} \]

  If ABS(I − LH) is greater than c*value, then Iβ is declared nonestimable. Here, H is the matrix \((X'X)^{-1}X'X\). The value must be between 0 and 1; the default is 10^{-4}.

As stated previously, the CONTRAST statement enables you to perform hypothesis tests \(H_0: L\beta = 0\).
If the $L$ matrix contains more than one contrast, then you can separate the rows of the $L$ matrix with commas. For example, for the model

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

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

$$ \begin{pmatrix} \mu & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & \beta_1 & \beta_2 \end{pmatrix} $$

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

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

The corresponding CONTRAST statement is

```plaintext
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*variable ... > ;

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

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

Note that a DOMAIN statement is different from a BY statement. In a BY statement, you treat the sample sizes as fixed in each subpopulation, and you perform analysis within each BY group independently. See the section “Domain Analysis” on page 7548 for more details.

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

A domain variable can be either character or numeric. The procedure treats domain variables as categorical variables. If a variable appears by itself in a DOMAIN statement, each level of this variable determines a domain in the study population. If two or more variables are joined by asterisks (*), then every possible combination of levels of these variables determines a domain. The procedure performs a descriptive analysis within each domain that is defined by the domain variables.
When determining levels of a DOMAIN variable, an observation with missing values for this DOMAIN variable is excluded, unless you specify the MISSING option. For more information, see the section “Missing Values” on page 7536.

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

**EFFECT Statement**


definition of the EFFECT statement

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 formed from continuous or classification variables, as discussed in the section “GLM Parameterization of Classification Variables and Effects” on page 394 of Chapter 19, “Shared Concepts and Topics.”

The following effect-types are available:

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

Table 90.3 summarizes important options for each type of EFFECT statement.

<table>
<thead>
<tr>
<th>Table 90.3 Important EFFECT Statement Options</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Option</strong></td>
</tr>
<tr>
<td><strong>Options for Collection Effects</strong></td>
</tr>
<tr>
<td>DETAILS</td>
</tr>
<tr>
<td><strong>Options for Lag Effects</strong></td>
</tr>
<tr>
<td>DESIGNROLE=</td>
</tr>
<tr>
<td>DETAILS</td>
</tr>
</tbody>
</table>
Table 90.3 continued

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLAG=</td>
<td>Specifies the number of periods in the lag</td>
</tr>
<tr>
<td>PERIOD=</td>
<td>Names the variable that defines the period</td>
</tr>
<tr>
<td>WITHIN=</td>
<td>Names the variable or variables that define the group within which each period is defined</td>
</tr>
</tbody>
</table>

Options for Multimember Effects

NOEFFECT Specifies that observations with all missing levels for the multimember variables should have zero values in the corresponding design matrix columns

WEIGHT= Specifies the weight variable for the contributions of each of the classification effects

Options for Polynomial Effects

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

Options for Spline Effects

BASIS= Specifies the type of basis (B-spline basis or truncated power function basis) for the spline expansion

DEGREE= Specifies the degree of the spline transformation

KNOTMETHOD= Specifies how to construct the knots for spline effects

For further details about the syntax of these effect-types and how columns of constructed effects are computed, see the section “EFFECT Statement” on page 403 of 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 \( L\beta \). You can perform hypothesis tests for the estimable functions, construct confidence limits, and obtain specific nonlinear transformations.

Table 90.4 summarizes important options in the ESTIMATE statement.
### Table 90.4 Important ESTIMATE Statement Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Construction and Computation of Estimable Functions</strong></td>
<td></td>
</tr>
<tr>
<td>DIVISOR=</td>
<td>Specifies a list of values to divide the coefficients</td>
</tr>
<tr>
<td>NOFILL</td>
<td>Suppresses the automatic fill-in of coefficients for higher-order effects</td>
</tr>
<tr>
<td>SINGULAR=</td>
<td>Tunes the estimability checking difference</td>
</tr>
<tr>
<td><strong>Degrees of Freedom and p-values</strong></td>
<td></td>
</tr>
<tr>
<td>ADJUST=</td>
<td>Determines the method for multiple comparison adjustment of estimates</td>
</tr>
<tr>
<td>ALPHA=α</td>
<td>Determines the confidence level ((1 - \alpha))</td>
</tr>
<tr>
<td>LOWER</td>
<td>Performs one-sided, lower-tailed inference</td>
</tr>
<tr>
<td>STEPDOWN</td>
<td>Adjusts multiplicity-corrected (p)-values further in a step-down fashion</td>
</tr>
<tr>
<td>TESTVALUE=</td>
<td>Specifies values under the null hypothesis for tests</td>
</tr>
<tr>
<td>UPPER</td>
<td>Performs one-sided, upper-tailed inference</td>
</tr>
<tr>
<td><strong>Statistical Output</strong></td>
<td></td>
</tr>
<tr>
<td>CL</td>
<td>Constructs confidence limits</td>
</tr>
<tr>
<td>CORR</td>
<td>Displays the correlation matrix of estimates</td>
</tr>
<tr>
<td>COV</td>
<td>Displays the covariance matrix of estimates</td>
</tr>
<tr>
<td>E</td>
<td>Prints the (L) matrix</td>
</tr>
<tr>
<td>JOINT</td>
<td>Produces a joint (F) or chi-square test for the estimable functions</td>
</tr>
<tr>
<td>SEED=</td>
<td>Specifies the seed for computations that depend on random numbers</td>
</tr>
</tbody>
</table>

For details about the syntax of the ESTIMATE statement, see the section “ESTIMATE Statement” on page 448 of Chapter 19, “Shared Concepts and Topics.”

### LSMEANS Statement

```plaintext
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 90.5** summarizes important options in the LSMEANS statement.
Table 90.5  Important LSMEANS Statement Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Construction and Computation of LS-Means</strong></td>
<td></td>
</tr>
<tr>
<td>AT</td>
<td>Modifies the covariate value in computing LS-means</td>
</tr>
<tr>
<td>BYLEVEL</td>
<td>Computes separate margins</td>
</tr>
<tr>
<td>DIFF</td>
<td>Requests differences of LS-means</td>
</tr>
<tr>
<td>OM=</td>
<td>Specifies the weighting scheme for LS-means computation as determined by the input data set</td>
</tr>
<tr>
<td>SINGULAR=</td>
<td>Tunes estimability checking</td>
</tr>
<tr>
<td><strong>Degrees of Freedom and p-values</strong></td>
<td></td>
</tr>
<tr>
<td>ADJUST=</td>
<td>Determines the method for multiple comparison adjustment of LS-means differences</td>
</tr>
<tr>
<td>ALPHA=α</td>
<td>Determines the confidence level ((1 - \alpha))</td>
</tr>
<tr>
<td>STEPDOWN</td>
<td>Adjusts multiple comparison (p)-values further in a step-down fashion</td>
</tr>
<tr>
<td><strong>Statistical Output</strong></td>
<td></td>
</tr>
<tr>
<td>CL</td>
<td>Constructs confidence limits for means and mean differences</td>
</tr>
<tr>
<td>CORR</td>
<td>Displays the correlation matrix of LS-means</td>
</tr>
<tr>
<td>COV</td>
<td>Displays the covariance matrix of LS-means</td>
</tr>
<tr>
<td>E</td>
<td>Prints the (L) matrix</td>
</tr>
<tr>
<td>LINES</td>
<td>Produces a “Lines” display for pairwise LS-means differences</td>
</tr>
<tr>
<td>MEANS</td>
<td>Prints the LS-means</td>
</tr>
<tr>
<td>PLOTS=</td>
<td>Requests ODS statistical graphics of means and mean comparisons</td>
</tr>
<tr>
<td>SEED=</td>
<td>Specifies the seed for computations that depend on random numbers</td>
</tr>
</tbody>
</table>

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

**LSMESTRIMATE Statement**

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

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

Table 90.6 summarizes important options in the LSMESTRIMATE statement.
Model Statement

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 3186 in Chapter 41, “The GLM Procedure.”

### Table 90.6  Important LSMESTIMATE Statement Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Construction and Computation of LS-Means</strong></td>
<td></td>
</tr>
<tr>
<td>AT</td>
<td>Modifies covariate values in computing LS-means</td>
</tr>
<tr>
<td>BYLEVEL</td>
<td>Computes separate margins</td>
</tr>
<tr>
<td>DIVISOR=</td>
<td>Specifies a list of values to divide the coefficients</td>
</tr>
<tr>
<td>OM=</td>
<td>Specifies the weighting scheme for LS-means computation as determined by a data set</td>
</tr>
<tr>
<td>SINGULAR=</td>
<td>Tunes estimability checking</td>
</tr>
<tr>
<td><strong>Degrees of Freedom and p-values</strong></td>
<td></td>
</tr>
<tr>
<td>ADJUST=</td>
<td>Determines the method for multiple comparison adjustment of LS-means differences</td>
</tr>
<tr>
<td>ALPHA=α</td>
<td>Determines the confidence level $(1 - \alpha)$</td>
</tr>
<tr>
<td>LOWER</td>
<td>Performs one-sided, lower-tailed inference</td>
</tr>
<tr>
<td>STEPDOWN</td>
<td>Adjusts multiple comparison $p$-values further in a step-down fashion</td>
</tr>
<tr>
<td>TESTVALUE=</td>
<td>Specifies values under the null hypothesis for tests</td>
</tr>
<tr>
<td>UPPER</td>
<td>Performs one-sided, upper-tailed inference</td>
</tr>
<tr>
<td><strong>Statistical Output</strong></td>
<td></td>
</tr>
<tr>
<td>CL</td>
<td>Constructs confidence limits for means and mean differences</td>
</tr>
<tr>
<td>CORR</td>
<td>Displays the correlation matrix of LS-means</td>
</tr>
<tr>
<td>COV</td>
<td>Displays the covariance matrix of LS-means</td>
</tr>
<tr>
<td>E</td>
<td>Prints the $L$ matrix</td>
</tr>
<tr>
<td>ELSM</td>
<td>Prints the $K$ matrix</td>
</tr>
<tr>
<td>JOINT</td>
<td>Produces a joint $F$ or chi-square test for the LS-means and LS-means differences</td>
</tr>
<tr>
<td>SEED=</td>
<td>Specifies the seed for computations that depend on random numbers</td>
</tr>
</tbody>
</table>

For details about the syntax of the LSMESTIMATE statement, see the section “LSMESTIMATE Statement” on page 480 of Chapter 19, “Shared Concepts and Topics.”
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.

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 7546 for details.

**INVERSE**
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, where W is the diagonal matrix constructed from WEIGHT variable values.

**NOINT**
oms 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 v is a vector, define $\text{ABS}(v)$ to be the largest absolute value of the elements of v. For a row vector l of the matrix L, define

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

If $\text{ABS}(l - lH) > c*value$, then $l\beta$ is declared nonestimable. Here, H is the matrix $(X'X)^{-1}X'X$. The value must be between 0 and 1; the default is $10^{-4}$. 
SOLUTION displays a solution to the normal equations, which are the parameter estimates. The SOLUTION option is useful only when you use a CLASS statement. If you do not specify a CLASS statement, PROC SURVEYREG displays parameter estimates by default. But if you specify a CLASS statement, PROC SURVEYREG does not display parameter estimates unless you also specify the SOLUTION option.

VADJUST=DF | NONE specifies whether to use degrees of freedom adjustment \((n - 1)/(n - p)\) in the computation of the matrix \(G\) 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

\[
\text{OUTPUT} < \text{OUT}=\text{SAS-data-set} > < \text{keyword} < =\text{variable-name} > ... \text{keyword} < =\text{variable-name} > > < /\text{option}> ;
\]

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:

\[ \text{OUT}=\text{SAS-data-set} \]
gives the name of the new output data set. By default, the procedure uses the DATA\text{n} convention to name the new data set.

\[ \text{keyword} < =\text{variable-name} > \]
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,

\[
\text{output out=myOutDataSet p=myPredictor;}
\]

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:

\[ \text{LCLM} | \text{L} \]
lower bound of a 100\((1 - \alpha)\)% confidence interval for the expected value (mean) of the predicted value. The \(\alpha\) level is equal to the value of the ALPHA= option in
the OUTPUT statement or, if this option is not specified, to the \texttt{ALPHA=} option in the \texttt{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 \texttt{_LCLM\_}.

\textbf{PREDICTED | P} \quad \text{predicted values. If no variable name is given for this keyword, the default variable name is \texttt{_PREDICTED\_}.}

\textbf{RESIDUAL | R} \quad \text{residuals, calculated as \texttt{ACTUAL} — \texttt{PREDICTED}. If no variable name is given for this keyword, the default variable name is \texttt{_RESIDUAL\_}.}

\textbf{STDP | STD} \quad \text{standard error of the mean predicted value. If no variable name is given for this keyword, the default variable name is \texttt{_STD\_}.}

\textbf{UCLM | U} \quad \text{upper bound of a 100(1 — \( \alpha \))\% confidence interval for the expected value (mean) of the predicted value. The \( \alpha \) level is equal to the value of the \texttt{ALPHA=} option in the OUTPUT statement or, if this option is not specified, to the \texttt{ALPHA=} option in the \texttt{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 \texttt{_UCLM\_}.}

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

\textbf{ALPHA=} \( \alpha \)

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

\section*{REPWEIGHTS Statement}

\texttt{REPWEIGHTS variables < / \text{options}> ;}

The \texttt{REPWEIGHTS} statement names variables that provide replicate weights for BRR or jackknife variance estimation, which you request with the \texttt{VARMETHOD=BRR} or \texttt{VARMETHOD=JACKKNIFE} option in the \texttt{PROC \SURVEYREG} statement. If you do not provide replicate weights for these methods by using a \texttt{REPWEIGHTS} statement, then the procedure constructs replicate weights for the analysis. See the sections “\textit{Balanced Repeated Replication (BRR) Method}” on page 7543 and “\textit{Jackknife Method}” on page 7545 for information about replicate weights.

Each \texttt{REPWEIGHTS} variable should contain the weights for a single replicate, and the number of replicates equals the number of \texttt{REPWEIGHTS} variables. The \texttt{REPWEIGHTS} variables must be numeric, and the variable values must be nonnegative numbers.

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

If you specify a \texttt{REPWEIGHTS} statement but do not include a \texttt{WEIGHT} statement, the procedure uses the average of replicate weights of each observation as the observation’s weight.
You can specify the following options in the REPWEIGHTS statement after a slash (/):

**DF=df**

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

**JKCOEFS=value**

specifies a jackknife coefficient for VARMETHOD=JACKKNIFE. The coefficient value must be a nonnegative number. See the section “Jackknife Method” on page 7545 for details about jackknife coefficients.

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

**JKCOEFS=values**

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

See the section “Jackknife Method” on page 7545 for details about jackknife coefficients.

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

**JKCOEFS=SAS-data-set**

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

See the section “Jackknife Method” on page 7545 for details about jackknife coefficients.

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

**SLICE Statement**

```sas
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 19.19. For details about the syntax of the SLICE statement, see the section “SLICE Statement” on page 510 of 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 is 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 513 of 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 7537 for more information.

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

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

When determining levels of a STRATA variable, an observation with missing values for this STRATA variable is excluded, unless you specify the MISSING option. For more information, see the section “Missing Values” on page 7536.

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 7553 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 5140 for details.

**TEST Statement**

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

The TEST statement enables you to perform $F$ tests for model effects that test Type I, 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 90.7 summarizes options in the TEST statement.

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

For details about the syntax of the TEST statement, see the section “TEST Statement” on page 514 of 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 5136 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. See Cochran (1977), Kalton and Kaspyzyk (1986), and Brick and Kalton (1996) for more information.

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

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

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

Whether or not you specify the NOMCAR option, the procedure always excludes observations with missing or invalid values for the WEIGHT, STRATA, CLUSTER, and DOMAIN variables, unless you specify the MISSING option.

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

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

Specification of Population Totals and Sampling Rates

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

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

If your design has multiple stages of selection and you are specifying the \texttt{RATE=} option, you should input the first-stage sampling rate, which is the ratio of the number of PSUs in the sample to the total number of PSUs in the study population. If you are specifying the \texttt{TOTAL=} option for a multistage design, you should input the total number of PSUs in the study population. See the section “Primary Sampling Units (PSUs)” on page 7538 for more details.

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

The secondary data set must contain all the stratification variables listed in the \texttt{STRATA} statement and all the variables in the \texttt{BY} statement. If there are formats associated with the \texttt{STRATA} variables and the \texttt{BY} variables, then the formats must be consistent in the primary and the secondary data sets. If you specify the \texttt{TOTAL=SAS-data-set} option, the secondary data set must have a variable named \texttt{_TOTAL_} that contains the stratum population totals. Or if you specify the \texttt{RATE=SAS-data-set} option, the secondary data set must have a variable named \texttt{_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 \texttt{_TOTAL_} or \texttt{_RATE_} for that stratum and ignores the rest.

The \texttt{value} in the \texttt{RATE=} option or the values of \texttt{_RATE_} in the secondary data set must be nonnegative numbers. You can specify \texttt{value} as a number between 0 and 1. Or you can specify \texttt{value} in percentage form as a number between 1 and 100, and \texttt{PROC SURVEYREG} converts that number to a proportion. The procedure treats the value 1 as 100\%, and not the percentage form 1\%.

If you specify the \texttt{TOTAL=}\texttt{value} option, \texttt{value} must not be less than the sample size. If you provide stratum population totals in a secondary data set, these values must not be less than the corresponding stratum sample sizes.
Primary Sampling Units (PSUs)

When you have clusters, or primary sampling units (PSUs), in your sample design, the procedure estimates variance from the variation among PSUs when the Taylor series variance method is used. See the section “Variance Estimation” on page 7542 for more information.

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

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

Computational Details

Notation

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

\[
(w, y, X) = (w_{hij}, y_{hij}, x_{hij})
\]

where

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

Also, \( f_h \) denotes the sampling rate for stratum \( h \). You can use the TOTAL= or RATE= option to input population totals or sampling rates. See the section “Specification of Population Totals and Sampling Rates” on page 7537 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 $X'WX\mathbf{\beta} = X'Wy$ by using a modified sweep routine that produces a generalized (g2) inverse $(X'WX)^{-}$ and a solution (Pringle and Rayner 1971)

$$\hat{\mathbf{\beta}} = (X'WX)^{-}X'Wy$$

where $W$ 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 7542. 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

$$f_{SRS} = n / w_...$$

where $n$ is the sample size and $w_...$ (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:

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

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:

$$f_{SRS} = 0$$
Stratum Collapse

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

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

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

If you specify stratum sampling rates by using the RATE=SAS-data-set option, PROC SURVEYREG computes the sampling rate for the new pooled stratum as the weighted average of the sampling rates for the collapsed strata. See the section “Computational Details” on page 7538 for details. If the specified sampling rate equals 0 for any of the collapsed strata, then the pooled stratum is assigned a sampling rate of 0. If you specify stratum totals by using the TOTAL=SAS-data-set option, PROC SURVEYREG combines the totals for the collapsed strata to compute the sampling rate for the new pooled stratum.

Sampling Rate of the Pooled Stratum from Collapse

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

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

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

$$SS_{error} = r^t W r$$

and $$SS_{total}$$ is the total sum of squares

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

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

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

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 (Woodruff (1971); Fuller (1975); Fuller, Kennedy, Schnell, Sullivan, and Park (1989); Särndal, Swensson, and Wretman (1992); Wolter (2007); Rust (1985); Dippo, Fay, and Morganstein (1984); Rao, Wu, and Yue (1992); and Rao and Shao (1996)). You can use the VARMETHOD= option to specify a variance estimation method to use. By default, the Taylor series method is used. However, replication methods have recently gained popularity for estimating variances in complex survey data analysis. One reason for this popularity is the relative simplicity of replication-based estimates, especially for nonlinear estimators; another is that modern computational capacity has made replication methods feasible for practical survey analysis.

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

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

Taylor Series (Linearization)

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

Use the notation described in the section “Notation” on page 7538 to denote the residuals from the linear regression as

\[ r = y - X\hat{\beta} \]

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

\[ G = \frac{n - 1}{n - p} \sum_{h=1}^{H} \frac{n_h(1 - f_h)}{n_h - 1} \sum_{i=1}^{n_h} (e_{hi} - \bar{e}_h .)'(e_{hi} - \bar{e}_h .) \]

where

\[ e_{hij} = w_{hij} r_{hij} x_{hij} \]

\[ e_{hi} = \sum_{j=1}^{m_{hi}} e_{hij} \]

\[ \bar{e}_h . = \frac{1}{n_h} \sum_{i=1}^{n_h} e_{hi} . \]
The Taylor series estimate of the covariance matrix of \( \hat{\beta} \) is

\[
\hat{V}(\hat{\beta}) = (X'WX)^{-1}G(X'WX)^{-1}
\]

The factor \((n - 1)/(n - p)\) in the computation of the matrix \( G \) 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 \texttt{VADJUST=NONE} option in the \texttt{MODEL} statement to suppress this factor.

**Balanced Repeated Replication (BRR) Method**

The balanced repeated replication (BRR) method requires that the full sample be drawn by using a stratified sample design with two primary sampling units (PSUs) per stratum. Let \( H \) be the total number of strata. The total number of replicates \( R \) is the smallest multiple of 4 that is greater than \( H \). However, if you prefer a larger number of replicates, you can specify the \texttt{REPS=number} option. If a \( n \times n \) 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 \( (r = 1, 2, \ldots, R) \) replicate is drawn from the full sample according to the \( r \)th row of the Hadamard matrix as follows:

- If the \((r, h)\)th 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 \((r, h)\)th 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 (2009).

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 \texttt{VARMETHOD=BRR(PRINTH)} method-option. If you provide a Hadamard matrix by specifying the \texttt{VARMETHOD=BRR(HADAMARD=)} method-option, then the replicates are generated according to the provided Hadamard matrix.

You can use the \texttt{VARMETHOD=BRR(OUTWEIGHTS=)} method-option to save the replicate weights into a SAS data set.
Chapter 90: The SURVEYREG Procedure

Let $\hat{\beta}$ be the estimated regression coefficients from the full sample for $\beta$, 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

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

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 $\epsilon$ (Judkins 1990; Rao and Shao 1999). When $\epsilon = 0$, Fay’s method becomes the traditional BRR method. For more details, see Dippo, Fay, and Morganstein (1984), Fay (1984), Fay (1989), and 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 replicate is created from the full sample according to the $r$th row of the Hadamard matrix as follows:

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

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

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 $\alpha_r$. 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 $\alpha_r$ are the same for all replicates:

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

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

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

**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

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

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 ($r = 1, 2, \ldots, R$), then the corresponding replicate weight for the $j$th member of the $i$th PSU is defined as

$$w_{ij}^{(r)} = \left\{ \begin{array}{ll}
w_{ij} & \text{if } i \text{th PSU is not in the donor stratum } \tilde{h}_r \\
\frac{w_{ij}}{\alpha_r} & \text{if } i \text{th PSU is in the donor stratum } \tilde{h}_r
\end{array} \right.$$  

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

If you provide your own replicate weights with a `REPWEIGHTS` statement, then you can also provide corresponding jackknife coefficients with the `JKCOEFS=` option.

Let $\hat{\beta}$ be the estimated regression coefficients from the full sample for $\beta$. 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

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

with $R - H$ degrees of freedom, where $R$ is the number of replicates and $H$ is the number of strata, or $R - 1$ when there is no stratification.
Hadamard Matrix

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

$$HH' = kI$$

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

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

\[
\begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\
1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\
1 & -1 & -1 & 1 & -1 & -1 & -1 & 1 \\
1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 \\
1 & -1 & 1 & -1 & -1 & -1 & -1 & -1 \\
1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 \\
1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\
\end{bmatrix}
\]

Degrees of Freedom

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

Denominator Degrees of Freedom

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

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

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

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

For the BRR method (including Fay’s method) without a REPWEIGHTS statement, the denominator $df$ equals the number of strata.
For the jackknife method without a REPWEIGHTS statement, the denominator \( df \) is equal to the number of replicates minus the actual number of strata.

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

**Numerator Degrees of Freedom**

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

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

See the section “Testing Effects” on page 7547 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 \( L\hat{\beta} \) 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 \( H_0: L\hat{\beta} = 0 \). The Wald \( F \) statistic equals

\[
F_{\text{Wald}} = \frac{(L\hat{\beta})'(L'W^2L)^{-1}(L\hat{\beta})}{\text{rank}(L'W^2L)}
\]

with numerator degrees of freedom equal to \( \text{rank}(L'W^2L) \).

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 7546. 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 514 of 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 $H_0 : L\beta = 0$ is computed as

$$
F_{Wald} = \frac{(L_{\text{Full}}\hat{\beta})(L_{\text{Full}}'\hat{\Sigma}L_{\text{Full}})^{-1}(L_{\text{Full}}\hat{\beta})}{\text{rank}(L)}
$$

where $L$ is the contrast vector or matrix you specify, $\beta$ is the vector of regression parameters, $\hat{\beta} = (X'WX)^{-1}X'WY$, $\hat{\Sigma}$ is the estimated covariance matrix of $\hat{\beta}$, rank($L$) is the rank of $L$, and $L_{\text{Full}}$ is a matrix such that

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

If $L$ is a full-rank matrix and all rows of $L$ are estimable functions, then $L_{\text{Full}}$ is the same as $L$. It is possible that $L_{\text{Full}}$ 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 3243 in Chapter 41, “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 7555 for more information. For a more detailed description of using ODS, see Chapter 20, “Using the Output Delivery System.”

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

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

**OUT= Data Set Created by the OUTPUT Statement**

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

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

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

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

Replicate Weights Output Data Set

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

The OUTWEIGHTS= data set contains the following variables:

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

where \( n \) is the total number of replicates in the analysis. Each replicate weight variable contains the replicate
weights for the corresponding replicate. Replicate weights equal zero for those observations not included in
the replicate.

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

Jackknife Coefficients Output Data Set

If you specify the OUTJKCOEFS= method-option for VARMETHOD=JACKKNIFE, PROC SURVEYREG
stores the jackknife coefficients in an output data set. The OUTJKCOEFS= output data set contains one
observation for each replicate. The OUTJKCOEFS= output 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 7542) or if the NOMCAR option is used, by default, PROC SURVEYREG displays the following variance estimation information in the “Variance Estimation” table:

- Method, which is the variance estimation method
- Number of Replicates, if you specify the VARMETHOD=BRR or VARMETHOD=JACKKNIFE option
- Hadamard Data Set name, if you specify the VARMETHOD=BRR(HADAMARD=) method-option
- Fay Coefficient, if you specify the VARMETHOD=BRR(FAY) method-option
- Replicate Weights input data set name, if you provide replicate weights with a REPWEIGHTS statement
- Missing Levels, which indicates whether missing levels of categorical variables are included by the MISSING option
- Missing Values, which indicates whether observations with missing values are included in the analysis by the NOMCAR option
Stratum Information

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

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

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

Class Level Information

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

- Class Variable, which lists each CLASS variable name
- Levels, which is the number of values or levels of the classification variable
- Values, which lists the values of the classification variable. The values are separated by a white space character; therefore, to avoid confusion, you should not include a white space character within a classification variable value.

X’X Matrix

If you specify the XPX option in the MODEL statement, PROC SURVEYREG displays the 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
- Standard Error, which is the standard error of the estimate
- t Value, which is the $t$ statistic for testing $H_0: \text{Parameter} = 0$
- Pr $>|t|$, which is the two-sided significance probability corresponding to the $t$ Value

**Covariance of Estimated Regression Coefficients**

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

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

**Analysis of Contrasts**

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

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

**Hadamard Matrix**

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

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

**ODS Table Names**

PROC SURVEYREG assigns a name to each table it creates; these names are listed in Table 90.8. 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 90.8. For information about these tables, see the corresponding sections of Chapter 19, “Shared Concepts and Topics.”
Table 90.8  ODS Tables Produced by PROC SURVEYREG

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

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 7510:

```
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, with the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section “Enabling and Disabling ODS Graphics” on page 609 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 608 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.”

Examples: SURVEYREG Procedure

Example 90.1: 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:

```sas
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
```

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

\[ \text{LFPR1972} = \beta_0 + \beta_1 \times \text{LFPR1968} + \text{error} \]

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:

```plaintext
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 90.1.1 summarizes the data information and the fit information.

### Output 90.1.1  Summary of Regression Using Simple Random Sampling

<table>
<thead>
<tr>
<th>Study of Labor Force Participation Rates of Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>The SURVEYREG Procedure</td>
</tr>
<tr>
<td>Regression Analysis for Dependent Variable LFPR1972</td>
</tr>
<tr>
<td>Data Summary</td>
</tr>
<tr>
<td>Number of Observations</td>
</tr>
<tr>
<td>Mean of LFPR1972</td>
</tr>
<tr>
<td>Sum of LFPR1972</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fit Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-square</td>
</tr>
<tr>
<td>Root MSE</td>
</tr>
<tr>
<td>Denominator DF</td>
</tr>
</tbody>
</table>

Output 90.1.2 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 90.1.2  Regression Coefficient Estimates

<table>
<thead>
<tr>
<th>Effect</th>
<th>Num DF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>13.84</td>
<td>0.0016</td>
</tr>
<tr>
<td>Intercept</td>
<td>1</td>
<td>4.63</td>
<td>0.0452</td>
</tr>
<tr>
<td>LFPR1968</td>
<td>1</td>
<td>13.84</td>
<td>0.0016</td>
</tr>
</tbody>
</table>

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

| Parameter | Estimate | Error    | t Value | Pr > |t| |
|-----------|----------|----------|---------|-------|
| Intercept | 0.20331056 | 0.09444296 | 2.15   | 0.0452 |
| LFPR1968  | 0.65604048 | 0.17635810 | 3.72   | 0.0016 |

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

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

Example 90.2: 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:

```plaintext
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
```
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:

```sas
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 90.2.1 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 90.2.1  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 90.2.2 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 90.2.2  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 denominator degrees of freedom for the t tests is 4.
```

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

```
title1 'Regression Analysis for Swedish Municipalities';
title2 'Simple Random Sampling';
proc surveyreg data=Municipalities total=284;
   model Population85=Population75;
run;
```
The analysis ignores the clusters in the sample, assuming that the sample design is a simple random sampling. Therefore, the TOTAL= option specifies the total number of municipalities, which is 284.

Output 90.2.3 displays the regression results ignoring the clusters. Compared to the results in Output 90.2.2, the regression coefficient estimates are the same. However, without using clusters, the regression coefficients have a smaller variance estimate, as in Output 90.2.3. 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 90.2.2; without using the clusters, the estimate is 1.05, but with the estimated standard error 0.04, as displayed in Output 90.2.3. To estimate the variance of the regression coefficients correctly, you should include the clustering information in the regression analysis.

Output 90.2.3  Regression Analysis for Simple Random Sampling

| 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 denominator degrees of freedom for the t tests is 31.

Example 90.3: 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 90.2.
In this sample, a linear model between the Swedish populations in 1975 and in 1985 is established:

\[ \text{Population85} = \alpha + \beta \times \text{Population75} + \text{error} \]

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:

```plaintext
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 (\(\alpha\)) in the ESTIMATE statement is 284 and the coefficient for Population75 (\(\beta\)) is the total population in 1975 (8.2 million).

Output 90.3.1 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 90.3.1** Use the Regression Estimator to Estimate the Population Total

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

**Example 90.4: 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 90.9.
Table 90.9  Number of Farms in Each Stratum

<table>
<thead>
<tr>
<th>Stratum</th>
<th>State</th>
<th>Region</th>
<th>Number of Farms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Iowa</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>2</td>
<td>50</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>Nebraska</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>2</td>
<td>40</td>
</tr>
</tbody>
</table>

Total 235 19

Three models for the data are considered:

- **Model I** — Common intercept and slope:
  
  \[ \text{Corn Yield} = \alpha + \beta \times \text{Farm Area} \]

- **Model II** — Common intercept, different slope:
  
  \[ \text{Corn Yield} = \begin{cases} 
  \alpha + \beta_{\text{Iowa}} \times \text{Farm Area} & \text{if the farm is in Iowa} \\
  \alpha + \beta_{\text{Nebraska}} \times \text{Farm Area} & \text{if the farm is in Nebraska} 
\end{cases} \]

- **Model III** — Different intercept and different slope:
  
  \[ \text{Corn Yield} = \begin{cases} 
  \alpha_{\text{Iowa}} + \beta_{\text{Iowa}} \times \text{Farm Area} & \text{if the farm is in Iowa} \\
  \alpha_{\text{Nebraska}} + \beta_{\text{Nebraska}} \times \text{Farm Area} & \text{if the farm is in Nebraska} 
\end{cases} \]

Data from the stratified sample are saved in the SAS data set Farms. In the data set Farms, the variable Weight represents the sampling weight. In the following DATA step, the sampling weights are the reciprocals of selection probabilities:

```sas
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
```
Example 90.4: Stratified Sampling

The information about population size in each stratum is saved in the SAS data set StratumTotals:

```
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 PROC SURVEYREG with the following statements:

```
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 90.4.1 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 90.4.1 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
```
**Chapter 90: The SURVEYREG Procedure**

**Output 90.4.1 continued**

<table>
<thead>
<tr>
<th>Fit Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-square</td>
</tr>
<tr>
<td>Root MSE</td>
</tr>
<tr>
<td>Denominator DF</td>
</tr>
</tbody>
</table>

**Stratum Information**

<table>
<thead>
<tr>
<th>Stratum Index</th>
<th>State</th>
<th>Region</th>
<th>N Obs</th>
<th>Total</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Iowa</td>
<td>1</td>
<td>3</td>
<td>100</td>
<td>3.00%</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>2</td>
<td>5</td>
<td>50</td>
<td>10.0%</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>3</td>
<td>15</td>
<td>15</td>
<td>20.0%</td>
</tr>
<tr>
<td>4</td>
<td>Nebraska</td>
<td>1</td>
<td>6</td>
<td>30</td>
<td>20.0%</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>2</td>
<td>2</td>
<td>40</td>
<td>5.00%</td>
</tr>
</tbody>
</table>

Output 90.4.2 displays tests of model effects and the estimated regression coefficients.

**Output 90.4.2 Estimated Regression Coefficients and the Estimated Covariance Matrix**

<table>
<thead>
<tr>
<th>Tests of Model Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect</td>
</tr>
<tr>
<td>Model</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>FarmArea</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Estimated Regression Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>FarmArea</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Covariance of Estimated Regression Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>FarmArea</td>
</tr>
</tbody>
</table>

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:

```r
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:

```r
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 90.4.3 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 90.4.3 Regression Results from Fitting Model II**

<table>
<thead>
<tr>
<th>Analysis of Farm Area and Corn Yield</th>
<th>Model II: Same Intercept, Different Slopes</th>
</tr>
</thead>
<tbody>
<tr>
<td>The SURVEYREG Procedure</td>
<td></td>
</tr>
<tr>
<td>Regression Analysis for Dependent Variable CornYield</td>
<td></td>
</tr>
<tr>
<td>Fit Statistics</td>
<td></td>
</tr>
<tr>
<td>R-square 0.8158</td>
<td></td>
</tr>
<tr>
<td>Root MSE 11.6759</td>
<td></td>
</tr>
<tr>
<td>Denominator DF 14</td>
<td></td>
</tr>
</tbody>
</table>
Output 90.4.3 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 denominator degrees of freedom for the t tests is 14.

### Covariance of Estimated Regression Coefficients

<table>
<thead>
<tr>
<th></th>
<th>Intercept</th>
<th>FarmAreaIA</th>
<th>FarmAreaNE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>14.51135861</td>
<td>-0.118001232</td>
<td>-0.079908772</td>
</tr>
<tr>
<td>FarmAreaIA</td>
<td>-0.118001232</td>
<td>0.0035654381</td>
<td>0.0006501109</td>
</tr>
<tr>
<td>FarmAreaNE</td>
<td>-0.079908772</td>
<td>0.0006501109</td>
<td>0.0006227496</td>
</tr>
</tbody>
</table>

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:

```plaintext
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 90.4.4 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 90.4.4 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

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 denominator degrees of freedom for the t tests is 14.

Covariance of Estimated Regression Coefficients

<table>
<thead>
<tr>
<th>State</th>
<th>State Iowa</th>
<th>Nebraska</th>
<th>FarmAreaIA</th>
<th>FarmAreaNE</th>
</tr>
</thead>
<tbody>
<tr>
<td>State Iowa</td>
<td>27.790863033</td>
<td>0</td>
<td>-0.205517205</td>
<td>0</td>
</tr>
<tr>
<td>State Nebraska</td>
<td>0</td>
<td>2.8910750385</td>
<td>0</td>
<td>-0.027354011</td>
</tr>
<tr>
<td>FarmAreaIA</td>
<td>-0.205517205</td>
<td>0</td>
<td>0.0041711265</td>
<td>0</td>
</tr>
<tr>
<td>FarmAreaNE</td>
<td>0</td>
<td>-0.027354011</td>
<td>0</td>
<td>0.0003988349</td>
</tr>
</tbody>
</table>

Example 90.5: Regression Estimator for Stratified Sample

This example uses the corn yield data set FARMS from Example 90.4 to illustrate how to construct a regression estimator for a stratified sample design.

As in Example 90.3, 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 90.10. You want to estimate the total corn yield by using this information under the three linear models given in Example 90.4.
Table 90.10  Information for Each Stratum

<table>
<thead>
<tr>
<th>Stratum</th>
<th>State</th>
<th>Region</th>
<th>Population</th>
<th>Sample</th>
<th>Total Farm Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Iowa</td>
<td>1</td>
<td>100</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>2</td>
<td>50</td>
<td>5</td>
<td>13,200</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>3</td>
<td>15</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Nebraska</td>
<td>1</td>
<td>30</td>
<td>6</td>
<td>8,750</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>2</td>
<td>40</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>235</td>
<td>19</td>
<td>21,950</td>
</tr>
</tbody>
</table>

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

```plaintext
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:

```plaintext
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 90.10 (21950); and the coefficients for effect State*Region are the total number of farms in each strata (as displayed in Table 90.10).

Output 90.5.1 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 90.5.1** Regression Estimator for the Total of CornYield under Model I

<table>
<thead>
<tr>
<th>Estimate Corn Yield from Farm Size</th>
<th>Model I: Same Intercept and Slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>The SURVEYREG Procedure</td>
<td></td>
</tr>
<tr>
<td>Regression Analysis for Dependent Variable CornYield</td>
<td></td>
</tr>
<tr>
<td>Estimate</td>
<td></td>
</tr>
<tr>
<td>Label Estimate</td>
<td>Standard Error DF t Value</td>
</tr>
<tr>
<td>Estimate of CornYield under Model I</td>
<td>7463.52 926.84 14 8.05</td>
</tr>
</tbody>
</table>

Estimate

| Label                     | Pr > |t|   |
|---------------------------|------|----|
| Estimate of CornYield under Model I | <.0001 |

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

```plaintext
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 90.4). 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 90.10. The total number of farms in each strata are the coefficients for the strata effect:

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

Output 90.5.2 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.
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:

```sql
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

```sql
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 90.10.
Output 90.5.3 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 90.5.3** Regression Estimator for the Total of CornYield under Model III

<table>
<thead>
<tr>
<th>Estimate Corn Yield from Farm Size</th>
<th>Model III: Different Intercepts and Slopes</th>
</tr>
</thead>
<tbody>
<tr>
<td>The SURVEYREG Procedure</td>
<td></td>
</tr>
<tr>
<td>Regression Analysis for Dependent Variable CornYield</td>
<td></td>
</tr>
<tr>
<td>Estimate</td>
<td></td>
</tr>
<tr>
<td>Label</td>
<td>Estimate Error DF t Value</td>
</tr>
<tr>
<td>Total CornYield in Iowa under Model III</td>
<td>6246.11 851.27 14 7.34</td>
</tr>
</tbody>
</table>

| Estimate | Pr > |t| |
|-----------|------|-----|
| Total CornYield in Iowa under Model III | <.0001 |

**Example 90.6: 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 7540.

Suppose that you have the following data:

```r
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:

```sas
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:

```sas
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 90.6.1 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 7546).
**Output 90.6.1** Summary of Data and Regression

<table>
<thead>
<tr>
<th>Stratified Sample with Single Sampling Unit in Strata With Stratum Collapse</th>
</tr>
</thead>
<tbody>
<tr>
<td>The SURVEYREG Procedure</td>
</tr>
<tr>
<td>Regression Analysis for Dependent Variable Y</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Data Summary</strong></td>
</tr>
<tr>
<td>Number of Observations: 8</td>
</tr>
<tr>
<td>Sum of Weights: 157.00000</td>
</tr>
<tr>
<td>Weighted Mean of Y: 4.31210</td>
</tr>
<tr>
<td>Weighted Sum of Y: 677.00000</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Design Summary</strong></td>
</tr>
<tr>
<td>Number of Strata: 5</td>
</tr>
<tr>
<td>Number of Strata Collapsed: 2</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Fit Statistics</strong></td>
</tr>
<tr>
<td>R-square: 0.9564</td>
</tr>
<tr>
<td>Root MSE: 0.5111</td>
</tr>
<tr>
<td>Denominator DF: 4</td>
</tr>
</tbody>
</table>

Output 90.6.2 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 7540).

Output 90.6.3 displays the parameter estimates and the tests of the significance of the model effects.

**Output 90.6.2** Stratification Information

<table>
<thead>
<tr>
<th>Stratum Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stratum Index</td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

NOTE: Strata with only one observation are collapsed into the stratum with Stratum Index "0".
### Output 90.3 Parameter Estimates and Effect Tests

#### Tests of Model Effects

<table>
<thead>
<tr>
<th>Effect</th>
<th>Num DF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>173.01</td>
<td>0.0002</td>
</tr>
<tr>
<td>Intercept</td>
<td>1</td>
<td>0.00</td>
<td>0.9961</td>
</tr>
<tr>
<td>X</td>
<td>1</td>
<td>173.01</td>
<td>0.0002</td>
</tr>
</tbody>
</table>

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

#### Estimated Regression Coefficients

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

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

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

```plaintext
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 90.6.4 does not contain the stratum collapse information displayed in Output 90.6.1, and the denominator degrees of freedom are 3 instead of 4.

### Output 90.4 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

<table>
<thead>
<tr>
<th>Number of Observations</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum of Weights</td>
<td>157.00000</td>
</tr>
<tr>
<td>Weighted Mean of Y</td>
<td>4.31210</td>
</tr>
<tr>
<td>Weighted Sum of Y</td>
<td>677.00000</td>
</tr>
</tbody>
</table>
**Output 90.6.4 continued**

```
<table>
<thead>
<tr>
<th>Design Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Strata</td>
</tr>
</tbody>
</table>

**Fit Statistics**

| R-square | 0.9564 |
| Root MSE  | 0.5111 |
| Denominator DF | 3 |
```

In **Output 90.6.5**, although the fourth stratum and the fifth stratum contain only one observation, no stratum collapse occurs.

**Output 90.6.5 Stratification Information**

```
<table>
<thead>
<tr>
<th>Stratum Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stratum Population Sampling</td>
</tr>
<tr>
<td>Index</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
</tbody>
</table>
```

As a result of not collapsing strata, the standard error estimates of the parameters, shown in **Output 90.6.6**, are different from those in **Output 90.6.3**, as are the tests of the significance of model effects.

**Output 90.6.6 Parameter Estimates and Effect Tests**

```
<table>
<thead>
<tr>
<th>Tests of Model Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>Model</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>X</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Estimated Regression Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>X</td>
</tr>
</tbody>
</table>

NOTE: The denominator degrees of freedom for the t tests is 3. |
### Example 90.7: Domain Analysis

Recall the example in the section “Getting Started: SURVEYREG Procedure” on page 7507, 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).

This example shows how you can use PROC SURVEYREG to perform domain analysis. The data set follows:

```plaintext
data IceCreamDataDomain;
  input Grade Spending Income Gender$ @@;
datelines;
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;
```

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, as follows:

```plaintext
data StudentTotals;
  input Grade _TOTAL_;
datelines;
7 1824
8 1025
9 1151
;
```
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.

The following statements demonstrate how you can analyze the relationship between spending and income among male and female students:

```sas
title1 'Ice Cream Spending Analysis';
title2 'Domain Analysis by Gender';
proc surveyreg data=IceCreamDataDomain total=StudentTotals;
   strata Grade;
   model Spending = Income;
   domain Gender;
   weight Weight;
run;
```

Output 90.7.1 gives a summary of the domains.

**Output 90.7.1  Domain Analysis Summary**

<table>
<thead>
<tr>
<th>Ice Cream Spending Analysis</th>
<th>Domain Analysis by Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>The SURVEYREG Procedure</td>
<td></td>
</tr>
<tr>
<td>Gender=F</td>
<td></td>
</tr>
<tr>
<td>Domain Regression Analysis</td>
<td></td>
</tr>
<tr>
<td>for Variable Spending</td>
<td></td>
</tr>
<tr>
<td><strong>Domain Summary</strong></td>
<td></td>
</tr>
<tr>
<td>Number of Observations</td>
<td>40</td>
</tr>
<tr>
<td>Number of Observations in Domain</td>
<td>19</td>
</tr>
<tr>
<td>Number of Observations Not in Domain</td>
<td>21</td>
</tr>
<tr>
<td>Sum of Weights in Domain</td>
<td>1926.9</td>
</tr>
<tr>
<td>Weighted Mean of Spending</td>
<td>9.37611</td>
</tr>
<tr>
<td>Weighted Sum of Spending</td>
<td>18066.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ice Cream Spending Analysis</th>
<th>Domain Analysis by Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>The SURVEYREG Procedure</td>
<td></td>
</tr>
<tr>
<td>Gender=M</td>
<td></td>
</tr>
<tr>
<td>Domain Regression Analysis</td>
<td></td>
</tr>
<tr>
<td>for Variable Spending</td>
<td></td>
</tr>
<tr>
<td><strong>Domain Summary</strong></td>
<td></td>
</tr>
<tr>
<td>Number of Observations</td>
<td>40</td>
</tr>
<tr>
<td>Number of Observations in Domain</td>
<td>21</td>
</tr>
<tr>
<td>Number of Observations Not in Domain</td>
<td>19</td>
</tr>
<tr>
<td>Sum of Weights in Domain</td>
<td>2073.1</td>
</tr>
<tr>
<td>Weighted Mean of Spending</td>
<td>8.92305</td>
</tr>
<tr>
<td>Weighted Sum of Spending</td>
<td>18498.7</td>
</tr>
</tbody>
</table>
Output 90.7.2 shows the parameter estimates for the model within each domain.

**Output 90.7.2** Parameter Estimates within Domain

| Parameter | Estimate | Error   | t Value | Pr > |t| |
|-----------|---------|---------|---------|-------|---|
| Intercept | -23.751681 | 2.30795437 | -10.29 | <.0001 |
| Income    | 0.735366  | 0.04757001  | 15.46 | <.0001 |

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

**Ice Cream Spending Analysis**
**Domain Analysis by Gender**

The SURVEYREG Procedure

Gender=F

Domain Regression Analysis for Variable Spending

Estimated Regression Coefficients

| Parameter | Estimate | Error   | t Value | Pr > |t| |
|-----------|---------|---------|---------|-------|---|
| Intercept | -23.213291 | 2.13361241 | -10.88 | <.0001 |
| Income    | 0.729419  | 0.04589801  | 15.89 | <.0001 |

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

For this particular example, the effect Income is significant for both models built within subgroups of male and female students, and the models are quite similar. In many other cases, regression models vary from subgroup to subgroup.
Example 90.8: Compare Domain Statistics

This example is a continuation of Example 90.7 in which domain analyses for male and female students were performed. 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:

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

Output 90.8.1 shows the estimated spending among male and female students.

Output 90.8.1 Estimated Domain Means

```
The SURVEYMEANS Procedure

Domain Analysis: Gender

<table>
<thead>
<tr>
<th>Gender</th>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Error of Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>Spending</td>
<td>19</td>
<td>9.376111</td>
<td>1.077927</td>
</tr>
<tr>
<td>M</td>
<td>Spending</td>
<td>21</td>
<td>8.923052</td>
<td>1.003423</td>
</tr>
</tbody>
</table>

Domain Analysis: Gender

<table>
<thead>
<tr>
<th>Gender</th>
<th>Variable</th>
<th>95% CL for Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>Spending</td>
<td>7.19202418   11.5601988</td>
</tr>
<tr>
<td>M</td>
<td>Spending</td>
<td>6.88992385 10.9561807</td>
</tr>
</tbody>
</table>
```

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;

Output 90.8.2 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 90.8.1 which are produced by PROC SURVEYMEANS.

Output 90.8.3 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.
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:

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

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

Output 90.8.4 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 90.8.4** Domain Means among Grades

| 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 90.8.5 plots the weekly spending results that are shown in Output 90.8.4.
Output 90.8.5  Plot of Means of Ice Cream Spending within Grades

Output 90.8.6 displays pairwise comparisons for weekly ice scream spending among grades. All the differences are significant based on \( t \) tests.

Output 90.8.6  Domain Means Comparison

| 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 90.8.7 plots the comparisons that are shown in Output 90.8.6.
In Output 90.8.7, 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 90.8.6.

Example 90.9: Variance Estimate Using the Jackknife Method

This example uses the stratified sample from the section “Getting Started: SURVEYREG Procedure” on page 7507 to illustrate how to estimate the variances with replication methods.

As shown in the section “Stratified Sampling” on page 7510, 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, we use the jackknife method to estimate the variance, saving the replicate weights generated by the procedure into a SAS data set:

```sas
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 90.9.1 shows the summary of the data and the variance estimation method. There are a total of 40 replicates generated by the procedure.

**Output 90.9.1  Variance Estimation Using the Jackknife Method**

<table>
<thead>
<tr>
<th>Ice Cream Spending Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use the Jackknife Method to Estimate the Variance</td>
</tr>
<tr>
<td>The SURVEYREG Procedure</td>
</tr>
<tr>
<td>Regression Analysis for Dependent Variable Spending</td>
</tr>
<tr>
<td>Data Summary</td>
</tr>
<tr>
<td>Number of Observations</td>
</tr>
<tr>
<td>Sum of Weights</td>
</tr>
<tr>
<td>Weighted Mean of Spending</td>
</tr>
<tr>
<td>Weighted Sum of Spending</td>
</tr>
<tr>
<td>Design Summary</td>
</tr>
<tr>
<td>Number of Strata</td>
</tr>
<tr>
<td>Variance Estimation</td>
</tr>
<tr>
<td>Method</td>
</tr>
<tr>
<td>Number of Replicates</td>
</tr>
</tbody>
</table>

Output 90.9.2 displays the parameter estimates and their standard errors, as well as the tests of model effects that use the jackknife method.
Output 90.9.2 Variance Estimation Using the Jackknife Method

<table>
<thead>
<tr>
<th>Tests of Model Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect</td>
</tr>
<tr>
<td>Model</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>Income</td>
</tr>
<tr>
<td>Kids</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Estimated Regression Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>Income</td>
</tr>
<tr>
<td>Kids 1</td>
</tr>
<tr>
<td>Kids 2</td>
</tr>
<tr>
<td>Kids 3</td>
</tr>
<tr>
<td>Kids 4</td>
</tr>
</tbody>
</table>

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

Matrix X'WX is singular and a generalized inverse was used to solve the normal equations. Estimates are not unique.

Output 90.9.3 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 90.9.3  The Jackknife Replicate Weights for the First 6 Observations

<table>
<thead>
<tr>
<th>Obs</th>
<th>Grade</th>
<th>Spending</th>
<th>Income</th>
<th>Kids</th>
<th>Prob</th>
<th>Weight</th>
<th>RepWt_1</th>
<th>RepWt_2</th>
<th>RepWt_3</th>
<th>RepWt_4</th>
<th>RepWt_5</th>
<th>RepWt_6</th>
<th>RepWt_7</th>
<th>RepWt_8</th>
<th>RepWt_9</th>
<th>RepWt_10</th>
<th>RepWt_11</th>
<th>RepWt_12</th>
<th>RepWt_13</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>7</td>
<td>39</td>
<td>2</td>
<td>0.010965</td>
<td>91.200</td>
<td>0.000</td>
<td>96.000</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>96.000</td>
<td>91.200</td>
<td>96.000</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>7</td>
<td>38</td>
<td>1</td>
<td>0.010965</td>
<td>91.200</td>
<td>0.000</td>
<td>96.000</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>96.000</td>
<td>91.200</td>
<td>96.000</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>12</td>
<td>47</td>
<td>1</td>
<td>0.008780</td>
<td>113.889</td>
<td>113.889</td>
<td>113.889</td>
<td>91.200</td>
<td>96.000</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>10</td>
<td>47</td>
<td>4</td>
<td>0.009557</td>
<td>104.636</td>
<td>104.636</td>
<td>104.636</td>
<td>104.636</td>
<td>104.636</td>
<td>104.636</td>
<td>104.636</td>
<td>104.636</td>
<td>104.636</td>
<td>104.636</td>
<td>104.636</td>
<td>104.636</td>
<td>104.636</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>1</td>
<td>34</td>
<td>4</td>
<td>0.010965</td>
<td>91.200</td>
<td>0.000</td>
<td>96.000</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>96.000</td>
<td>91.200</td>
<td>96.000</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>10</td>
<td>43</td>
<td>3</td>
<td>0.010965</td>
<td>91.200</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obs</th>
<th>RepWt_14</th>
<th>RepWt_15</th>
<th>RepWt_16</th>
<th>RepWt_17</th>
<th>RepWt_18</th>
<th>RepWt_19</th>
<th>RepWt_20</th>
<th>RepWt_21</th>
<th>RepWt_22</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
</tr>
<tr>
<td>2</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
</tr>
<tr>
<td>3</td>
<td>113.889</td>
<td>113.889</td>
<td>113.889</td>
<td>128.125</td>
<td>128.125</td>
<td>128.125</td>
<td>128.125</td>
<td>113.889</td>
<td>113.889</td>
</tr>
<tr>
<td>4</td>
<td>104.636</td>
<td>104.636</td>
<td>104.636</td>
<td>104.636</td>
<td>104.636</td>
<td>104.636</td>
<td>104.636</td>
<td>104.636</td>
<td>115.100</td>
</tr>
<tr>
<td>5</td>
<td>96.000</td>
<td>0.000</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
</tr>
<tr>
<td>6</td>
<td>96.000</td>
<td>0.000</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obs</th>
<th>RepWt_23</th>
<th>RepWt_24</th>
<th>RepWt_25</th>
<th>RepWt_26</th>
<th>RepWt_27</th>
<th>RepWt_28</th>
<th>RepWt_29</th>
<th>RepWt_30</th>
<th>RepWt_31</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>96.000</td>
<td>96.000</td>
</tr>
<tr>
<td>2</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>96.000</td>
<td>96.000</td>
</tr>
<tr>
<td>3</td>
<td>113.889</td>
<td>113.889</td>
<td>113.889</td>
<td>128.125</td>
<td>128.125</td>
<td>128.125</td>
<td>128.125</td>
<td>113.889</td>
<td>113.889</td>
</tr>
<tr>
<td>4</td>
<td>104.636</td>
<td>104.636</td>
<td>104.636</td>
<td>104.636</td>
<td>104.636</td>
<td>104.636</td>
<td>104.636</td>
<td>104.636</td>
<td>115.100</td>
</tr>
<tr>
<td>5</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>96.000</td>
<td>96.000</td>
</tr>
<tr>
<td>6</td>
<td>96.000</td>
<td>96.000</td>
<td>96.000</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>96.000</td>
<td>96.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obs</th>
<th>RepWt_32</th>
<th>RepWt_33</th>
<th>RepWt_34</th>
<th>RepWt_35</th>
<th>RepWt_36</th>
<th>RepWt_37</th>
<th>RepWt_38</th>
<th>RepWt_39</th>
<th>RepWt_40</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>96.000</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>96.000</td>
<td>91.200</td>
<td>91.200</td>
<td>96.000</td>
<td>96.000</td>
</tr>
<tr>
<td>2</td>
<td>96.000</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>96.000</td>
<td>91.200</td>
<td>91.200</td>
<td>96.000</td>
<td>96.000</td>
</tr>
<tr>
<td>3</td>
<td>113.889</td>
<td>113.889</td>
<td>128.125</td>
<td>128.125</td>
<td>113.889</td>
<td>113.889</td>
<td>113.889</td>
<td>113.889</td>
<td>113.889</td>
</tr>
<tr>
<td>4</td>
<td>104.636</td>
<td>115.100</td>
<td>104.636</td>
<td>104.636</td>
<td>115.100</td>
<td>104.636</td>
<td>115.100</td>
<td>104.636</td>
<td>104.636</td>
</tr>
<tr>
<td>5</td>
<td>96.000</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>96.000</td>
<td>91.200</td>
<td>91.200</td>
<td>96.000</td>
<td>96.000</td>
</tr>
<tr>
<td>6</td>
<td>96.000</td>
<td>91.200</td>
<td>91.200</td>
<td>91.200</td>
<td>96.000</td>
<td>91.200</td>
<td>91.200</td>
<td>96.000</td>
<td>96.000</td>
</tr>
</tbody>
</table>
References


Subject Index

A

ADJRSQ
SURVEYREG procedure, 7530
adjusted R-square
SURVEYREG procedure, 7541
alpha level
SURVEYREG procedure, 7515, 7532
analysis of variance
SURVEYREG procedure, 7540
ANOVA
SURVEYREG procedure, 7530, 7540

B

balanced repeated replication
SURVEYREG procedure, 7543
variance estimation (SURVEYREG), 7543
BRR
SURVEYREG procedure, 7543
BRR variance estimation
SURVEYREG procedure, 7543

C

classification variables
SURVEYREG procedure, 7521
cluster sampling
SURVEYREG procedure, 7559
clustering
SURVEYREG procedure, 7522
computational details
SURVEYREG procedure, 7538
computational resources
SURVEYREG procedure, 7548
confidence level
SURVEYREG procedure, 7515
confidence limits
SURVEYREG procedure, 7530
contrasts
SURVEYREG procedure, 7522, 7548

degrees of freedom
SURVEYREG procedure, 7546
design effects
SURVEYREG procedure, 7539
domain analysis
SURVEYREG procedure, 7548, 7578, 7581
domain means comparison
SURVEYREG procedure, 7581
donor stratum
SURVEYREG procedure, 7545

effect testing
SURVEYREG procedure, 7547

F

Fay coefficient
SURVEYREG procedure, 7518, 7544
Fay’s BRR method
variance estimation (SURVEYREG), 7544
finite population correction
SURVEYREG procedure, 7516, 7517, 7537

H

Hadamard matrix
SURVEYREG procedure, 7518, 7546

J

jackknife
SURVEYREG procedure, 7545
jackknife coefficients
SURVEYREG procedure, 7545, 7550
jackknife variance estimation
SURVEYREG procedure, 7545

L

linearization method
SURVEYREG procedure, 7542

M

missing values
SURVEYREG procedure, 7515, 7536
MSE
SURVEYREG procedure, 7541
multiple R-square
SURVEYREG procedure, 7541
N
number of replicates
SURVEYREG procedure, 7519, 7543–7545

O
ODS graph names
SURVEYREG procedure, 7557
ODS Graphics
SURVEYREG procedure, 7557
ODS table names
SURVEYREG procedure, 7555
options summary
EFFECT statement, 7525
ESTIMATE statement, 7526
output data sets
SURVEYREG procedure, 7549
output jackknife coefficient
SURVEYREG procedure, 7550
output replicate weights
SURVEYREG procedure, 7550
output table names
SURVEYREG procedure, 7555

P
pooled stratum
SURVEYREG procedure, 7540
primary sampling units (PSUs)
SURVEYREG procedure, 7538

R
regression coefficients
SURVEYREG procedure, 7539
regression estimators
SURVEYREG procedure, 7562, 7569
replicate weights
SURVEYREG procedure, 7542
replication methods
SURVEYREG procedure, 7517, 7542, 7585
root MSE
SURVEYREG procedure, 7541

S
sampling rates
SURVEYREG procedure, 7516, 7537
sampling weights
SURVEYREG procedure, 7532, 7535
simple random sampling
SURVEYREG procedure, 7507, 7557
singularity level
SURVEYREG procedure, 7523, 7530
stratification
SURVEYREG procedure, 7534
stratified sampling
SURVEYREG procedure, 7510, 7563
stratum collapse
SURVEYREG procedure, 7540, 7573
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, 7507
SURVEYREG procedure, 7507
ADJRSQ, 7530
adjusted R-square, 7541
alpha level, 7515, 7532
analysis of contrasts table, 7555
analysis of variance, 7540
ANOVA, 7530, 7540
ANOVA table, 7554
balanced repeated replication, 7543
BRR, 7543
BRR variance estimation, 7543
classification level table, 7553
classification variables, 7521
cluster sampling, 7559
clustering, 7522
coefficients of contrast table, 7555
computational details, 7538
computational resources, 7548
confidence level, 7515
confidence limits, 7530
contrasts, 7522, 7548
covariance of estimated regression coefficients
table, 7554
data summary table, 7551
degrees of freedom, 7546
design effects, 7539
design summary table, 7551
domain analysis, 7548, 7578, 7581
domain means comparison, 7581
domain summary table, 7552
domain variable, 7524
donor stratum, 7545
effect testing, 7547
Fay coefficient, 7518, 7544
Fay’s BRR variance estimation, 7544
finite population correction, 7516, 7517, 7537
first-stage sampling rate, 7516
fit statistics table, 7552
Hadamard matrix, 7518, 7546, 7555
inverse matrix of X’X, 7553
jackknife, 7545
jackknife coefficients, 7545, 7550
jackknife variance estimation, 7545
linearization method, 7542
list of strata, 7534
missing values, 7515, 7536
MSE, 7541
multiple R-square, 7541
number of replicates, 7519, 7543–7545
ODS graph names, 7557
ODS Graphics, 7557
ordering of effects, 7516
output data sets, 7513, 7549
output jackknife coefficient, 7550
output replicate weights, 7550
output table names, 7555
pooled stratum, 7540
population totals, 7517, 7537
primary sampling units (PSUs), 7538
regression coefficients, 7539
regression coefficients table, 7554
regression estimators, 7562, 7569
replicate weights, 7542
replication methods, 7517, 7542, 7585
root MSE, 7541
sampling rates, 7516, 7537
sampling weights, 7532, 7535
simple random sampling, 7507, 7557
singularity level, 7523, 7530
stratification, 7534
stratified sampling, 7510, 7563
stratum collapse, 7540, 7573
stratum information table, 7553
subpopulation analysis, 7578, 7581
Taylor series variance estimation, 7520, 7542
testing effect, 7547
tests of model effects table, 7554
variance estimation, 7542
variance estimation table, 7552
VARMETHOD=BRR option, 7543
VARMETHOD=JACKKNIFE option, 7545
VARMETHOD=JK option, 7545
Wald test, 7547, 7548
weighting
SURVEYREG procedure, 7532, 7535

T
Taylor series variance estimation
SURVEYREG procedure, 7520, 7542
testing effect
SURVEYREG procedure, 7547

V
variance estimation
Syntax Index

A
ADJRSQ option
  MODEL statement (SURVEYREG), 7530
ALPHA= option
  OUTPUT statement (SURVEYREG), 7532
  PROC SURVEYREG statement, 7515
ANOVA option
  MODEL statement (SURVEYREG), 7530
B
BY statement
  SURVEYREG procedure, 7521
C
CLASS statement
  SURVEYREG procedure, 7521
CLPARM option
  MODEL statement (SURVEYREG), 7530
CLUSTER statement
  SURVEYREG procedure, 7522
CONTRAST statement
  SURVEYREG procedure, 7522
COVB option
  MODEL statement (SURVEYREG), 7530
D
DATA= option
  PROC SURVEYREG statement, 7515
DEFF option
  MODEL statement (SURVEYREG), 7530
DF= option
  MODEL statement (SURVEYREG), 7530
  REPWEIGHTS statement (SURVEYREG), 7533
DOMAIN statement
  SURVEYREG procedure, 7524
E
E option
  CONTRAST statement (SURVEYREG), 7523
  EFFECT statement
    SURVEYREG procedure, 7525
ESTIMATE statement
  SURVEYREG procedure, 7526
F
FAY= option
  VARMETHOD=BRR (PROC SURVEYREG statement), 7518
H
H= option
  VARMETHOD=BRR (PROC SURVEYREG statement), 7518
HADAMARD= option
  VARMETHOD=BRR (PROC SURVEYREG statement), 7518
I
INVERSE option
  MODEL statement (SURVEYREG), 7530
J
JKCOEFS= option
  REPWEIGHTS statement (SURVEYREG), 7533
K
keyword= option
  OUTPUT statement (SURVEYREG), 7531
L
LCLM keyword
  OUTPUT statement (SURVEYREG), 7531
LIST option
  STRATA statement (SURVEYREG), 7534
LSMEAN statement
  SURVEYREG procedure, 7528
M
MISSING option
  PROC SURVEYREG statement, 7515
  MODEL statement
    SURVEYREG procedure, 7529
N
N= option
PROC SURVEYREG statement, 7517
NOCOLLABSE option
STRATA statement (SURVEYREG), 7535
NOFILL option
CONTRAST statement (SURVEYREG), 7523
NOINT option
MODEL statement (SURVEYREG), 7530
NOMCAR option
PROC SURVEYREG statement, 7515
O
ORDER= option
PROC SURVEYREG statement, 7516
OUT= option
OUTPUT statement (SURVEYREG), 7531
OUTJKCOEFS= option
V ARMETHOD=JACKKNIFE (PROC SURVEYREG statement), 7520
V ARMETHOD=JK (PROC SURVEYREG statement), 7520
OUTPUT statement
SURVEYREG procedure, 7531
OUTWEIGHTS= option
V ARMETHOD=BRR (PROC SURVEYREG statement), 7519
V ARMETHOD=JACKKNIFE (PROC SURVEYREG statement), 7520
V ARMETHOD=JK (PROC SURVEYREG statement), 7520
PROC SURVEYREG statement, see SURVEYREG procedure
S
PARMLABEL option
MODEL statement (SURVEYREG), 7530
PREDICTED keyword
OUTPUT statement (SURVEYREG), 7532
PRINTH option
V ARMETHOD=BRR (PROC SURVEYREG statement), 7519
V ARMETHOD=JACKKNIFE (PROC SURVEYREG statement), 7520
V ARMETHOD=JK (PROC SURVEYREG statement), 7520
PROC SURVEYREG statement, see SURVEYREG procedure
R
R= option
PROC SURVEYREG statement, 7516
RATE= option
PROC SURVEYREG statement, 7516
REPS= option
V ARMETHOD=BRR (PROC SURVEYREG statement), 7519
REPWEEKTS statement
SURVEYREG procedure, 7532
RESIDUAL keyword
OUTPUT statement (SURVEYREG), 7532
S
SINGULAR= option
CONTRAST statement (SURVEYREG), 7523
MODEL statement (SURVEYREG), 7530
SLICE statement
SURVEYREG procedure, 7533
SOLUTION option
MODEL statement (SURVEYREG), 7531
STD keyword
OUTPUT statement (SURVEYREG), 7532
STDP keyword
OUTPUT statement (SURVEYREG), 7532
STORE statement
SURVEYREG procedure, 7534
STRATA statement
SURVEYREG procedure, 7534
SUBGROUP statement
SURVEYREG procedure, 7524
SURVEYREG procedure
syntax, 7514
SURVEYREG procedure, BY statement, 7521
SURVEYREG procedure, CLASS statement, 7521
SURVEYREG procedure, CLUSTER statement, 7522
SURVEYREG procedure, CONTRAST statement, 7522
E option, 7523
NOFILL option, 7523
SINGULAR= option, 7523
SURVEYREG procedure, DOMAIN statement, 7524
SURVEYREG procedure, EFFECT statement, 7525
SURVEYREG procedure, ESTIMATE statement, 7526
SURVEYREG procedure, LSMESTIMATE statement, 7528
SURVEYREG procedure, MODEL statement, 7529
ADJRSQ option, 7530
ANOVA option, 7530
CLPARM option, 7530
COVB option, 7530
DEFF option, 7530
INVERSE option, 7530
NOINT option, 7530
PARMLABEL option, 7530
SINGULAR= option, 7530
SOLUTION option, 7531
VADJUST= option, 7531
XPP option, 7531
SURVEYREG procedure, MODEL statement (SURVEYREG)
DF= option, 7530

SURVEYREG procedure, OUTPUT statement, 7531
    ALPHA= option, 7532
    keyword= option, 7531
    LCLM keyword, 7531
    OUT= option, 7531
    PREDICTED keyword, 7532
    RESIDUAL keyword, 7532
    STD keyword, 7532
    STDP keyword, 7532
    UCLM keyword, 7532

SURVEYREG procedure, PROC SURVEYREG statement, 7515
    ALPHA= option, 7515
    DATA= option, 7515
    FAY= option (V ARMETHOD=BRR), 7518
    H= option (V ARMETHOD=BRR), 7518
    HADAMARD= option (V ARMETHOD=BRR), 7518
    MISSING option, 7515
    N= option, 7517
    NOMCAR option, 7515
    ORDER= option, 7516
    OUTJKCOEFS= option
        (V ARMETHOD=JACKKNIFE), 7520
    OUTJKCOEFS= option (V ARMETHOD=JK), 7520
    OUTWEIGHTS= option
        (V ARMETHOD=BRR), 7519
    OUTWEIGHTS= option
        (V ARMETHOD=JACKKNIFE), 7520
    OUTWEIGHTS= option (V ARMETHOD=JK), 7520
    PRINTH option (V ARMETHOD=BRR), 7519
    R= option, 7516
    RATE= option, 7516
    REPS= option (V ARMETHOD=BRR), 7519
    TOTAL= option, 7517
    TRUNCATE option, 7517
    V ARMETHOD= option, 7517

SURVEYREG procedure, REPWEIGHTS statement, 7532
    DF= option, 7533
    JKCOEFS= option, 7533

SURVEYREG procedure, SLICE statement, 7533

SURVEYREG procedure, STORE statement, 7534

SURVEYREG procedure, STRATA statement, 7534
    LIST option, 7534
    NOCOLLAPSE option, 7535

SURVEYREG procedure, TEST statement, 7535

SURVEYREG procedure, WEIGHT statement, 7535

T

TEST statement
Your Turn

We welcome your feedback.

- If you have comments about this book, please send them to yourturn@sas.com. Include the full title and page numbers (if applicable).
- If you have comments about the software, please send them to suggest@sas.com.
Whether you are new to the work force or an experienced professional, you need to distinguish yourself in this rapidly changing and competitive job market. SAS® Publishing provides you with a wide range of resources to help you set yourself apart. Visit us online at support.sas.com/bookstore.

**SAS® Press**

Need to learn the basics? Struggling with a programming problem? You’ll find the expert answers that you need in example-rich books from SAS Press. Written by experienced SAS professionals from around the world, SAS Press books deliver real-world insights on a broad range of topics for all skill levels.

**SAS® Documentation**

To successfully implement applications using SAS software, companies in every industry and on every continent all turn to the one source for accurate, timely, and reliable information: SAS documentation. We currently produce the following types of reference documentation to improve your work experience:

- Online help that is built into the software.
- Tutorials that are integrated into the product.
- Reference documentation delivered in HTML and PDF – **free** on the Web.
- Hard-copy books.

**SAS® Publishing News**

Subscribe to SAS Publishing News to receive up-to-date information about all new SAS titles, author podcasts, and new Web site features via e-mail. Complete instructions on how to subscribe, as well as access to past issues, are available at our Web site.

Support the development of SAS® software by purchasing SAS® products and services from SAS Institute Inc. at support.sas.com/bookstore.