Chapter 62
The SURVEYREG Procedure

Chapter Table of Contents

Chapter 62
The SURVEYREG Procedure

Overview

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 computes the regression coefficient estimators by generalized least squares estimation using element-wise regression. 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 the Taylor expansion theory for estimating sampling errors of estimators based on complex sample designs (Woodruff 1971; Fuller 1975; Särndal, Swenson, and Wretman 1992, Chapter 5 and Chapter 13). This method obtains a linear approximation for the estimator and then uses the variance estimator for this approximation to estimate the variance of the estimator itself.

PROC SURVEYREG uses ODS (Output Delivery System) to place results in output data sets. This is a departure from older SAS procedures that provide OUTPUT statements for similar functionality.

Getting Started

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" on page 3225. The "Examples" section on page 3246 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 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 a SAS data set.

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

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

The following PROC SURVEYREG statements request a regression analysis.

```
title1 'Ice Cream Spending Analysis';
title2 'Simple Random Sampling 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.

```
                 Ice Cream Spending Analysis
                 Simple Random Sampling Design

                   The SURVEYREG Procedure

        Regression Analysis for Dependent Variable Spending

                        Data Summary

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


                       Fit Statistics

                 R-square            0.8132
                 Root MSE            2.4506
                 Denominator DF          39


                   Class Level Information

               Class
               Variable      Levels    Values

               Kids               4    1 2 3 4
```

Figure 62.1. Summary of Data

Figure 62.1 displays the summary of the data, the summary of the fit, and the levels of the classification variable Kids. The "Fit Summary" table displays the denominator degrees of freedom, which are used in *F* tests and *t* tests in the regression analysis.

```
                 Ice Cream Spending Analysis
                 Simple Random Sampling Design

                   The SURVEYREG Procedure

        Regression Analysis for Dependent Variable Spending

             ANOVA for Dependent Variable Spending

                                 Sum of        Mean
Source                   DF     Squares      Square    F Value    Pr > F

Model                     4     915.310    228.8274      38.10    <.0001
Error                    35     210.190      6.0054
Corrected Total          39    1125.500


                    Tests of Model Effects

             Effect       Num DF    F Value    Pr > F

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

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

Figure 62.2. Testing Effects in the Regression

Figure 62.2 displays the ANOVA table for the regression and 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.

```
                        Ice Cream Spending Analysis
                       Simple Random Sampling Design

                          The SURVEYREG Procedure

              Regression Analysis for Dependent Variable Spending

                       Estimated Regression Coefficients

                                      Standard
          Parameter      Estimate         Error    t Value    Pr > |t|

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

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

Figure 62.3. Regression Coefficients

The regression coefficient estimates and their standard errors and associated *t* tests are displayed in Figure 62.3.

Stratified Sampling

Suppose that the previous student sample is actually drawn from a stratified sampling. The strata are grades in the junior high school: the 7th grade, the 8th grade, and the 9th grade. Within strata, simple random samples are selected. Table 62.1 provides the number of students in each grade.

Table 62.1. Students in Grades

Grade	Number of Students
7	1,824
8	1,025
3	1,151
Total	4,000

In order to analyze this sample using PROC SURVEYREG, you need to input the stratification information by creating a SAS data set for Table 62.1. The following SAS statements create a data set called StudentTotal.

```
data StudentTotal;
   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 the strata 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.

The following statements demonstrate how you can fit the linear model while incorporating the sample design information (stratification).

```
title1 'Ice Cream Spending Analysis';
title2 'Stratified Simple Random Sampling Design';
proc surveyreg data=IceCream total=StudentTotal;
   strata Grade /list;
   class Kids;
   model Spending = Income Kids / solution;
   run;
```

By comparing these statements to those in the section "Simple Random Sampling" on page 3217, the TOTAL=StudentTotal option replaces the previous TOTAL=4000 option. When the population totals and sample sizes differ among strata, the population totals must be provided by a data set.

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.

Ice Cream Spending Analysis
Stratified Simple Random Sampling Design

The SURVEYREG Procedure

Regression Analysis for Dependent Variable Spending

Data Summary

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

Design Summary

Number of Strata	3

Fit Statistics

R-square	0.8132
Root MSE	2.4506
Denominator DF	37

Figure 62.4. Summary of the Regression

Figure 62.4 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 is 37, which is different from the analysis in Figure 62.1.

Ice Cream Spending Analysis
Stratified Simple Random Sampling Design

The SURVEYREG Procedure

Regression Analysis for Dependent Variable Spending

Stratum Information

Stratum Index	Grade	N Obs	Population Total	Sampling Rate
1	7	20	1824	0.01
2	8	9	1025	0.01
3	9	11	1151	0.01

Class Level Information

Class Variable	Levels	Values
Kids	4	1 2 3 4

Figure 62.5. Stratification and Classification Information

Figure 62.5 displays the identifications of strata, numbers of observations or sample sizes in strata, total numbers of students in strata, and calculated sampling rates or sampling fractions in strata.

```
                    Ice Cream Spending Analysis
                Stratified Simple Random Sampling Design

                       The SURVEYREG Procedure

            Regression Analysis for Dependent Variable Spending

                  ANOVA for Dependent Variable Spending

                                 Sum of        Mean
Source                     DF    Squares      Square    F Value    Pr > F

Model                       4    915.310    228.8274      38.10    <.0001
Error                      35    210.190      6.0054
Corrected Total            39   1125.500


                         Tests of Model Effects

                Effect       Num DF    F Value    Pr > F

                Model             4     114.60    <.0001
                Intercept         1     150.05    <.0001
                Income            1     317.63    <.0001
                Kids              3       0.93    0.4355

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

Figure 62.6. Testing Effects

Figure 62.6 displays the ANOVA table for the regression and the tests for the significance of model effects under the stratified sample design. The income effect is significant, while the kids effect is not significant at the 5% level.

```
                    Ice Cream Spending Analysis
                Stratified Simple Random Sampling Design

                       The SURVEYREG Procedure

            Regression Analysis for Dependent Variable Spending

                    Estimated Regression Coefficients

                                  Standard
        Parameter      Estimate       Error    t Value    Pr > |t|

        Intercept    -26.084677    2.48241893     -10.51      <.0001
        Income         0.775330    0.04350401      17.82      <.0001
        Kids 1         0.897655    1.11778377       0.80      0.4271
        Kids 2         1.494032    1.25209199       1.19      0.2404
        Kids 3        -0.513181    1.36853454      -0.37      0.7098
        Kids 4         0.000000    0.00000000          .         .

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

Figure 62.7. Regression Coefficients

The regression coefficient estimates for the stratified sample are displayed in Figure 62.7. The standard errors of the estimates and associated *t* tests are also shown in this table.

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

Create Output Data Set

PROC SURVEYREG uses the Output Delivery System (ODS) to create output data sets. This is a departure from older SAS procedures that provide OUTPUT statements for similar functionality. For more information on ODS, see Chapter 15, "Using the Output Delivery System."

For example, to place the "ParameterEstimates" table (Figure 62.7) in the previous section in an output data set, you use the ODS OUTPUT statement as follows.

```
title1 'Ice Cream Spending Analysis';
title2 'Stratified Simple Random Sampling Design';
proc surveyreg data=IceCream total=StudentTotal;
   strata Grade /list;
   class Kids;
   model Spending = Income Kids / solution;
   ods output ParameterEstimates = MyParmEst;
   run;
```

The statement

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

requests that the "ParameterEstimates" table that appears in Figure 62.7 to be placed in a SAS data set named MyParmEst.

The PRINT procedure displays observations of the data set MyParmEst.

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

Figure 62.8 displays the observations in the data set MyParmEst.

Ice Cream Spending Analysis
Stratified Simple Random Sampling Design

Obs	Parameter	Estimate	StdErr	DenDF	tValue	Probt
1	Intercept	-26.084677	2.48241893	37	-10.51	<.0001
2	Income	0.775330	0.04350401	37	17.82	<.0001
3	Kids 1	0.897655	1.11778377	37	0.80	0.4271
4	Kids 2	1.494032	1.25209199	37	1.19	0.2404
5	Kids 3	-0.513181	1.36853454	37	-0.37	0.7098
6	Kids 4	0.000000	0.00000000	37	.	.

Figure 62.8. The Data Set MyParmEst

Syntax

The following statements are available in PROC SURVEYREG.

PROC SURVEYREG < *options* > **;**
BY *variables* **;**
CLASS *variables* **;**
CLUSTER *variables* **;**
CONTRAST *'label' effect values*
< . . . *effect values* > < */ options* > **;**
ESTIMATE *'label' effect values*
< . . . *effect values* > < */ options* > **;**
MODEL *dependent* = < *effects* > < */ options* > **;**
STRATA *variables* < */ options* > **;**
WEIGHT *variable* **;**

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

The CONTRAST and ESTIMATE statements can appear multiple times; all other statements can appear only once.

PROC SURVEYREG Statement

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

The PROC SURVEYREG statement invokes the procedure. You can specify the following options in the PROC SURVEYREG statement.

ALPHA=α
sets the confidence level for confidence limits. The value of the ALPHA= option must be between 0.0001 and 0.9999, and the default value is 0.05. A confidence level of α produces $100(1-\alpha)$% confidence limits. The default of ALPHA=0.05 produces 95% confidence limits. If α is between 0 and 1 but outside the range of 0.0001 to 0.9999, the procedure uses the closest range endpoint. For example, if you specify ALPHA=0.000001, the procedure uses 0.0001 to determine 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.

NOPRINT

suppresses the display of all output. Note that this option temporarily disables the Output Delivery System (ODS). For more information, see Chapter 15, "Using the Output Delivery System."

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

specifies the sampling rate as a positive *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 variance estimation. 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 positive *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 3234 for details.

The *value* in the RATE= option or the values of _RATE_ in the secondary data set must be positive 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 will convert 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= option or the RATE= option, then the variance estimation does not include a finite population correction. You cannot specify both the TOTAL= option and the RATE= option.

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 variance estimation.

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 3234 for details.

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

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC SURVEYREG to obtain separate analyses on observations in groups defined by the BY variables.

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

When a BY statement appears, the procedure expects the input data sets to be sorted in order of the BY variables. If you specify more than one BY statement, the procedure uses only the latest BY statement and ignores any previous ones.

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

- Sort the data using the SORT procedure with a similar BY statement.
- Specify the BY statement option NOTSORTED or DESCENDING in the BY statement for the 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 using the DATASETS procedure.

For more information on the BY statement, refer to the discussion in *SAS Language Reference: Concepts*. For more information on the DATASETS procedure, refer to the discussion in the *SAS Procedures Guide.*

CLASS Statement

CLASS | **CLASSES** *variables* **;**

The CLASS statement specifies the classification variables to be used in the model. Typical class variables are TREATMENT, GENDER, RACE, GROUP, and REPLICATION. If you specify the CLASS statement, it must appear before the MODEL statement.

Classification variables can be either character or numeric. The procedure uses only the first 16 characters of a character variable. Class levels are determined from the formatted values of the CLASS variables. Thus, you can use formats to group values into levels. Refer to the discussion of the FORMAT procedure in the *SAS Procedures*

Guide and to the discussions of the FORMAT statement and SAS formats in *SAS Language Reference: Concepts*.

You can use multiple CLASS statements to specify classification variables.

CLUSTER Statement

CLUSTER | **CLUSTERS** *variables* **;**

The CLUSTER statement specifies variables that identify 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 your sample design has clustering at multiple stages, you should identify only the first-stage clusters, or primary sampling units (PSUs), in the CLUSTER statement.

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. Refer to the discussion of the FORMAT procedure in the *SAS Procedures Guide* and to the discussions of the FORMAT statement and SAS formats in *SAS Language Reference: Dictionary*.

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

CONTRAST Statement

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

The CONTRAST statement provides custom hypothesis tests for linear combinations of the regression parameters H_0: $\mathbf{L}\boldsymbol{\beta} = 0$, where $\mathbf{L}$ is the vector or matrix you specify and $\boldsymbol{\beta}$ is the vector of regression parameters. Thus, to use this feature, you must be familiar with the details of the model parameterization used by PROC SURVEYREG. For information on the parameterization, see the section "Parameterization of PROC GLM Models" on page 1521 in Chapter 30, "The GLM Procedure.".

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 using variable names and operators. For more details on how to specify an effect, see the section "Specification of Effects" on page 1517 in Chapter 30, "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 equals rank($\mathbf{L}$).

The denominator degrees of freedom equals the number of clusters (or the number of observations if there is no CLUSTER statement) minus the number of strata. Alternatively, you can use the DF= option in the MODEL statement to specify the denominator degrees of freedom.

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

In the CONTRAST statement,

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

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

E

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

NOFILL

requests no filling in higher-order effects. When you specify only certain portions of $\mathbf{L}$, by default PROC SURVEYREG constructs the remaining elements from the context (for more information, the section "Specification of ESTIMATE Expressions" on page 1536 in Chapter 30, "The GLM Procedure.").

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

SINGULAR=*value*

specifies the sensitivity for checking estimability. If $\mathbf{v}$ is a vector, define ABS($\mathbf{v}$) to be the largest absolute value of the elements of $\mathbf{v}$. Say $\mathbf{H}$ is the $(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}$ matrix, and C is ABS($\mathbf{L}$) except for elements of $\mathbf{L}$ that equal 0, and then C is 1. If ABS$(\mathbf{L} - \mathbf{LH}) > C \cdot$ *value*, then $\mathbf{L}$ is declared nonestimable. The SINGULAR=*value* must be between 0 and 1, and the default is 10^{-4}.

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

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

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

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

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

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

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

The corresponding CONTRAST statement is

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

ESTIMATE Statement

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

You can use an ESTIMATE statement to estimate a linear function of the regression parameters by multiplying a row vector **L** by the parameter estimate vector $\widehat{\boldsymbol{\beta}}$.

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 with a special notation using variable names and operators. For more details on how to specify an effect, see the section "Specification of Effects" on page 1517 in Chapter 30, "The GLM Procedure.".

PROC SURVEYREG checks the linear function for estimability. (See the SINGULAR= option described on page 3231).

The procedure displays the estimate $\mathbf{L}\widehat{\boldsymbol{\beta}}$ along with its standard error and *t* test. If you specify the CLPARM option in the MODEL statement, PROC SURVEYREG also displays confidence limits for the linear function. By default, the degrees of freedom for the *t* test equals the number of clusters (or the number of observations if there is no CLUSTER statement) minus the number of strata. Alternatively, you can specify the degrees of freedom with the DF= option in the MODEL statement.

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

In the ESTIMATE statement,

label identifies the linear function $\mathbf{L}$ in the output. A label is required for every function 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 values are constants that are elements of the vector $\mathbf{L}$ associated with the effect. For example, the following code forms an estimate that is the difference between the parameters estimated for the first and second levels of the CLASS variable A.

```
estimate 'A1 vs A2' A  1 -1;
```

You can specify the following options in the ESTIMATE statement after a slash (/).

DIVISOR=*value*
specifies a value by which to divide all coefficients so that fractional coefficients can be entered as integers. For example, note the difference between the following two ESTIMATE statements.

```
estimate '1/3(A1+A2) - 2/3A3' a 1 1 -2 / divisor=3;
estimate '1/3(A1+A2) - 2/3A3' a .33333 .33333  -.66667;
```

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

NOFILL
requests no filling in higher-order effects. When you specify only certain portions of the vector $\mathbf{L}$, by default PROC SURVEYREG constructs the remaining elements from the context. (See the section "Specification of ESTIMATE Expressions" on page 1536 in Chapter 30, "The GLM Procedure.") When you specify the NOFILL option, PROC SURVEYREG does not construct the remaining portions and treats the vector $\mathbf{L}$ as it is defined in the ESTIMATE statement.

SINGULAR=*value*
specifies the sensitivity for checking estimability. If $\mathbf{v}$ is a vector, define ABS($\mathbf{v}$) to be the largest absolute value of the elements of $\mathbf{v}$. Say $\mathbf{H}$ is the $(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}$ matrix, and C is ABS($\mathbf{L}$) except for elements of $\mathbf{L}$ that equal 0, and then C is 1. If ABS($\mathbf{L} - \mathbf{L}\mathbf{H}) > C \times$*value*, then $\mathbf{L}$ is declared nonestimable. The SINGULAR= *value* must be between 0 and 1, and the default is 10^{-4}.

MODEL Statement

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

The MODEL statement specifies the dependent (response) variable and the independent (regressor) variables or effects. 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 a special notation using variable names and operators. For more information on how to specify an effect, see the section "Specification of Effects" on page 1517 in Chapter 30, "The GLM Procedure." The dependent variable must be numeric. 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 (/).

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

DF=*value*
specifies the denominator degrees of freedom for the *F* tests and the degrees of freedom for the *t* tests. The default is the number of clusters (or the number of observations if there is no CLUSTER statement) minus the number of actual strata. The number of actual strata equals the number of strata in the data before collapsing minus the number of strata collapsed plus 1.

CLPARM
requests confidence limits for the parameter estimates. The SURVEYREG procedure determines the confidence coefficient using the ALPHA= option described on page 3225, 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. See the SOLUTION option on page 3233.

DEFF
displays design effects for the regression coefficient estimates.

NOINT
omits the intercept from the model.

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

X
XPX

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

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.

STRATA Statement

STRATA | **STRATUM** *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 3234 for more information.

The STRATA *variables* are one or more variables in the DATA= input data set. These variables can be either character or numeric. The procedure uses only the first 16 characters of the value of a character variable. The formatted values of the STRATA variables determine the levels. Thus, you can use formats to group values into levels. Refer to the discussion of the FORMAT procedure in the *SAS Procedures Guide*.

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. This table also displays stratum collapse information.

NOCOLLAPSE

prevents the procedure from collapsing, or combining, strata that have only one sampling unit. By default, the procedure collapses strata that contain only one sampling unit. See the section "Stratum Collapse" on page 3236 for details.

WEIGHT Statement

WEIGHT | **WGT** *variable* **;**

The WEIGHT statement specifies the variable that contains the sampling weights. This variable must be numeric. If you do not specify a WEIGHT statement, PROC SURVEYREG assigns all observations a weight of 1. 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. If you specify more than one WEIGHT statement, the procedure uses only the first WEIGHT statement and ignores the rest.

Details

Specification of Population Totals and Sampling Rates

If your analysis should include a finite population correction (*fpc*), you can input either the sampling rate or the population total using the RATE= option or the TOTAL= option. You cannot specify both of these options in the same PROC SURVEYREG statement. If you do not specify one of these options, the procedure does not use the *fpc* when computing variance estimates. For fairly small sampling fractions, it is appropriate to ignore this correction. Refer to Cochran (1977) and Kish (1965).

If your design has multiple stages of selection and you are specifying the RATE= option, you should input the first-stage sampling rate, which is the ratio of the number of PSUs in the sample to the total number of PSUs in the study population. If you are specifying the TOTAL= option for a multistage design, you should input the total number of PSUs in the study population.

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

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

The secondary data set must contain all BY and STRATA groups that occur in the primary data set. If the secondary data set contains more than one observation for any one stratum, then the procedure uses the first value of _TOTAL_ or _RATE_ for that stratum and ignores the rest.

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

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

Primary Sampling Units (PSUs)

When you have clusters, or primary sampling units (PSUs), in your sample design, the procedure estimates variance from the variation among PSUs. For more information, see the section "Variance Estimation" on page 3239. You can use the CLUSTER statement to identify the first stage clusters in your design. PROC SURVEYREG assumes that each cluster represents a PSU in the sample and that each observation is an element of a PSU. If you do not specify a CLUSTER statement, the procedure treats each observation as a PSU.

Missing Values

If an observation has a missing value or a nonpositive value for the WEIGHT variable, then PROC SURVEYREG excludes that observation from the analysis. An observation is also excluded if it has a missing value for any STRATA variable, CLUSTER variable, dependent variable, or any variable used in the independent effects. The analysis includes all observations in the data set that have nonmissing values for all these design and analysis variables.

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 will be biased and will not accurately represent the survey population. There are a variety of techniques in sample design and survey operations that can reduce nonresponse. Once data collection is complete, you can use imputation to replace missing values with acceptable values, and 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. Refer to Cochran (1977) for more details.

Stratum Collapse

If there is only one sampling unit in a stratum, then PROC SURVEYREG cannot estimate the variance for this stratum. 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 0 for any stratum that contains only one sampling unit.

If you do not specify the NOCOLLAPSE option, PROC SURVEYREG collapses strata according to the following rules. If there are multiple strata that each contain only one sampling unit, 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 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 Method" on page 3237 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 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.

Analysis of Variance

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 using the sampling weights, but not the sample design information on 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.

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. By default, 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

- 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 the denominator DF using the DF= option in the MODEL statement.

Numerator Degrees of Freedom

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

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

See the section "Testing Effects" on page 3239 for more information. The numerator DF for the Wald *F* statistic for a contrast is the rank of the $\mathbf{L}$ matrix that defines the contrast.

Computational Method

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

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

where

- w denotes the sampling weight vector
- y denotes the dependent variable

- $\mathbf{X}$ denotes the design matrix. (When an effect contains only classification variables, the columns of $\mathbf{X}$ corresponding to this effect contain only 0s and 1s; no reparameterization is made.)
- $h = 1, 2, \ldots, H$ is the stratum number with a total of H strata
- $i = 1, 2, \ldots, n_h$ is the cluster number within stratum h, with a total of n_h clusters
- $j = 1, 2, \ldots, m_{hi}$ is the unit number within cluster i of stratum h, with a total of m_{hi} units
- 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= option or the RATE= option to input population totals or sampling rates. See the section "Specification of Population Totals and Sampling Rates" on page 3234 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= option or the RATE= option, then the procedure assumes that the stratum sampling rates f_h are negligible, and a finite population correction is not used when computing variances.

Regression Coefficients

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

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

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

For models with class variables, there are more design matrix columns than there are degrees of freedom (DF) for the effect. Thus, there are linear dependencies among the columns. In this case, the parameters are not estimable; there is an infinite number of least-squares solutions. PROC SURVEYREG uses a generalized (g2) inverse to obtain values for the estimates. The solution values are not displayed unless you specify the SOLUTION option in the MODEL statement. The solution has the characteristic that estimates are 0 whenever the design column for that parameter is a linear combination of previous columns. (Strictly termed, 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.

Variance Estimation

PROC SURVEYREG uses the Taylor series expansion theory to estimate the covariance-variance matrix of the estimated regression coefficients (Fuller 1975). Let

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

where the (h, i, j)th element is r_{hij}. Compute $1 \times p$ row vectors

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

and calculate the $p \times p$ matrix

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

PROC SURVEYREG computes the covariance matrix of $\boldsymbol{\beta}$ as

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

Testing Effects

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

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

with numerator degrees of freedom equal to $\text{rank}(\mathbf{L})$ and denominator degrees of freedom equal to the number of clusters minus the number of strata (unless you have specified the denominator degrees of freedom with the DF= option in the MODEL statement; see the section "Denominator Degrees of Freedom" on page 3237). It is possible that the $\mathbf{L}$ matrix cannot be constructed for an effect, in which case that effect is not testable. For more information on how the matrix $\mathbf{L}$ is constructed, see the discussion in Chapter 12, "The Four Types of Estimable Functions."

Multiple R-squared

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

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

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

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

and $\mathrm{SS}_{\mathrm{total}}$ is the total sum of squares

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

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

Root Mean Square Errors

PROC SURVEYREG computes the square root of mean square errors as

$$\sqrt{\mathrm{MSE}} = \sqrt{n \, \mathrm{SS}_{\mathrm{error}} \, / \, (n-p) \, w_{...}}$$

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

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.

$$\mathrm{DEFF} = \frac{\text{Variance under the Sample Design}}{\text{Variance under Simple Random Sampling}}$$

Refer to Kish (1965, p.258). PROC SURVEYREG computes the numerator as described in the section "Variance Estimation" on page 3239. And the denominator is computed under the assumption that the sample design is simple random sampling, with no stratification and no clustering.

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

$$f_{\,\mathrm{SRS}} = n \, / \, w_{...}$$

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.

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

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

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

Sampling Rate of the Pooled Stratum from Collapse

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

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

Contrasts

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

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

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

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

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

Output Data Sets

Output data sets from PROC SURVEYREG are produced using the ODS (Output Delivery System). ODS encompasses more than just the production of output data sets. For more information on using ODS, see Chapter 15, "Using the Output Delivery System."

Displayed Output

The SURVEYREG procedure produces the following output.

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

Fit Summary

By default, PROC SURVEYREG displays the following regression statistics in the "Fit Summary" 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

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.

$\mathbf{X}'\mathbf{X}$ *Matrix*

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

Inverse Matrix of $\mathbf{X}'\mathbf{X}$

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

ANOVA for Dependent Variable

By default, 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 SOLUTIONS 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: 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

Coefficients of Estimate

When you specify the E option in an ESTIMATE statement, PROC SURVEYREG displays a "Coefficients of Estimate" table for the linear function of the regression parameters in the ESTIMATE statement. You can use this table to check the coefficients you specified in the ESTIMATE statement. Also, this table gives a note for a nonestimable function.

Analysis of Estimable Functions

If you specify an ESTIMATE statement, PROC SURVEYREG checks the function for estimability. If the function is estimable, PROC SURVEYREG produces an "Analysis of Estimable Functions" table, which displays the estimate and the corresponding *t* test. If you use more than one ESTIMATE statement, the procedure displays all results in the same table. The table contains the following information for each estimable function.

- Parameter, which is the label of the function
- Estimate, which is the estimate of the estimable liner function
- Standard Error, which is the standard error of the estimate
- t Value, which is the *t* statistic for testing H_0: Estimable Function $= 0$
- Pr > | t |, which is the two-sided significance probability corresponding to the t Value

ODS Table Names

PROC SURVEYREG assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in the following table. For more information on ODS, see Chapter 15, "Using the Output Delivery System."

Table 62.2. ODS Tables Produced in PROC SURVEYREG

ODS Table Name	Description	Statement	Option
ANOVA	ANOVA for dependent variable	MODEL	default
ClassVarInfo	Class level information	CLASS	default
ContrastCoef	Coefficients of contrast	CONTRAST	E
Contrasts	Analysis of contrasts	CONTRAST	default
CovB	Covariance of estimated regression coefficients	MODEL	COVB
DataSummary	Data summary	MODEL	default
DesignSummary	Design summary	STRATA | CLUSTER	default
Effects	Tests of model effects	MODEL	
EstimateCoef	Coefficients of estimate	ESTIMATE	E
Estimates	Analysis of estimable functions	ESTIMATE	default
FitStatistics	Fit summary	MODEL	default
InvXPX	Inverse matrix of $\mathbf{X}'\mathbf{X}$	MODEL	INV
ParameterEstimates	Estimated regression coefficients	MODEL	default
StrataInfo	Stratum information	STRATA	LIST
XPX	$\mathbf{X}'\mathbf{X}$ matrix	MODEL	XPX

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

For example, the following statements create an output data set named MyStrata, which contains the "StrataInfo" table, an output data set named MyParmEst, which contains the "ParameterEstimates" table, and an output data set named Cov, which contains the "CovB" table for the ice cream study discussed in the section "Stratified Sampling" on page 3220.

```
title1 'Ice Cream Spending Analysis';
title2 'Stratified Simple Random Sampling Design';
proc surveyreg data=IceCream total=StudentTotal;
   strata Grade /list;
   class Kids;
   model Spending = Income Kids / solution covb;
   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.

Examples

Example 62.1. 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 named Labor. The LFPR in 1972 is contained in the variable LFPR1972, and the LFPR in 1968 is identified by the variable LFPR1968.

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

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

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

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

```
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 62.1.1. Summary of Regression Using Simple Random Sampling

```
           Study of Labor Force Participation Rates of Women

                        The SURVEYREG Procedure

          Regression Analysis for Dependent Variable LFPR1972

                             Data Summary

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


                            Fit Statistics

                      R-square           0.3970
                      Root MSE          0.05657
                      Denominator DF         18


                ANOVA for Dependent Variable LFPR1972

                                 Sum of        Mean
Source                   DF     Squares      Square    F Value    Pr > F

Model                     1    0.035813    0.035813      11.19    0.0038
Error                    17    0.054398    0.003200
Corrected Total          18    0.090211
```

Output 62.1.1 summarizes the data information, the fit information, and the ANOVA table.

Output 62.1.2. Regression Coefficient Estimates

```
          Study of Labor Force Participation Rates of Women

                      The SURVEYREG Procedure

       Regression Analysis for Dependent Variable LFPR1972

                      Tests of Model Effects

            Effect       Num DF    F Value    Pr > F

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

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


                  Estimated Regression Coefficients

                                 Standard
        Parameter      Estimate         Error    t Value    Pr > |t|

        Intercept    0.20331056    0.09444296       2.15      0.0452
        LFPR1968     0.65604048    0.17635810       3.72      0.0016

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

Output 62.1.2 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.

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. 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 62.2. Simple Random Cluster Sampling

This example illustrates the use of regression analysis in a simple random cluster sampling design. The data are from Särndal, Swenson, and Wretman (1992, p. 652).

A total of 284 Swedish municipalities are grouped into 50 clusters of neighboring municipalities. Five clusters with a total of 32 municipalities are randomly selected. The results from the regression analysis in which clusters are used in the sample design are compared to the results of a regression analysis that ignores the clusters. The linear relationship between the population in 1975 and in 1985 is investigated.

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

```
data Municipalities;
   input Municipality Cluster Population85 Population75;
   datalines;
   205   37    5    5
   206   37   11   11
   207   37   13   13
```

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

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

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

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

Output 62.2.1. Regression Analysis for Simple Random Cluster Sampling

```
          Regression Analysis for Swedish Municipalities
                  Cluster Simple Random Sampling

                     The SURVEYREG Procedure

     Regression Analysis for Dependent Variable Population85

                          Data Summary

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


                          Design Summary

                  Number of Clusters              5


                          Fit Statistics

                     R-square             0.9860
                     Root MSE             3.0488
                     Denominator DF            4


                Estimated Regression Coefficients

                                    Standard
Parameter          Estimate            Error    t Value    Pr > |t|

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

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

Output 62.2.1 displays the data summary, design summary, fit summary, and regression coefficient estimates. Since the sample design includes clusters, the procedure displays the total number of clusters in the sample in the "Design Summary" table. 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.

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, the standard deviation of the regression coefficients will be 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 62.2.2. Regression Analysis for Simple Random Sampling

```
          Regression Analysis for Swedish Municipalities
                     Simple Random Sampling

                    The SURVEYREG Procedure

      Regression Analysis for Dependent Variable Population85

                          Data Summary

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


                          Fit Statistics

                     R-square            0.9860
                     Root MSE            3.0488
                     Denominator DF          31


                 Estimated Regression Coefficients

                                  Standard
Parameter          Estimate          Error    t Value    Pr > |t|

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

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

Output 62.2.2 displays the regression results ignoring the clusters. Compared to the results in Output 62.2.1 on page 3250, the regression coefficient estimates are the same. However, without using clusters, the regression coefficients have a smaller variance estimate in Output 62.2.2. Using clusters in the analysis, the estimated regression coeffiecient for effect Population75 is 1.05, with the estimated standard error 0.05, as displayed in Output 62.2.1; without using the clusters, the estimate is 1.05, but with the estimated standard error 0.04, as displayed in Output 62.2.2. To estimated the variance of the regression coefficients correctly, you should include the clustering information in the regression analysis.

Example 62.3. Regresson Estimator for Simple Random Sample

Using auxilary information, you can construct the regression estimators to provide more accurate estimates of the population characteristics that are of interest. With ESTIMATE statements in PROC SURVEYREG, you can speciofy a regression estimator as a linear function of the regression parameters to estimate the population total. This example illustrates this application, using the data in the previous example.

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

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

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

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

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

Output 62.3.1. Use the Regression Estimator to Estimate the Population Total

```
          Regression Analysis for Swedish Municipalities
                    Estimate Total Population

                       The SURVEYREG Procedure

      Regression Analysis for Dependent Variable Population85

                  Analysis of Estimable Functions

                                      Standard
 Parameter            Estimate         Error    t Value    Pr > |t|

 1985 population    8642.49485    258.558613      33.43      <.0001

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

Output 62.3.1 displays the regression results and the estimation of the total population. 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.

Example 62.4. Stratified Sampling

This example illustrates the SURVEYREG procedure to perform a regression in a stratified sample design. Consider a population of 235 farms producing corn in the states of Nebraska and Iowa. You are interested in the relationship between corn yield (CornYield) and the 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 to the stratified simple random sample. The sample size and population size within each stratum are displayed in Table 62.3.

Table 62.3. Number of Farms in Each Stratum

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

Three models are considered to represent the data:

- Model I — Common intercept and slope:

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

- Model II — Common intercept, different slope:

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

- Model III — Different intercept and slope:

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

Data from the stratified sample are saved in the SAS data set Farms.

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

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

In the data set Farms, the variable Weight represents the sampling weight. In this example, the sampling weight is proportional to the reciprocal of the sampling rate within each stratum from which a farm is selected. The information on population size in each stratum is saved in the SAS data set TotalInStrata.

```
data TotalInStrata;
   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 TotalInStrata, you can fit Model I using PROC SURVEYREG.

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

Output 62.4.1. Data Summary and Stratum Information Fitting Model I

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

The SURVEYREG Procedure

Regression Analysis for Dependent Variable CornYield

Data Summary

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

Design Summary

Number of Strata	5

Fit Statistics

R-square	0.3882
Root MSE	20.6422
Denominator DF	14

Stratum Information

Stratum Index	State	Region	N Obs	Population Total	Sampling Rate
1	Iowa	1	3	100	0.03
2		2	5	50	0.10
3		3	3	15	0.20
4	Nebraska	1	6	30	0.20
5		2	2	40	0.05

Output 62.4.1 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 62.4.2. Estimated Regression Coefficients and the Estimated Covariance Matrix

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

                    The SURVEYREG Procedure

      Regression Analysis for Dependent Variable CornYield

                     Tests of Model Effects

           Effect       Num DF    F Value    Pr > F

           Model             1      21.74    0.0004
           Intercept         1       4.93    0.0433
           FarmArea          1      21.74    0.0004

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


               Estimated Regression Coefficients

                                Standard
  Parameter      Estimate          Error    t Value    Pr > |t|

  Intercept    11.8162978    5.31981027       2.22      0.0433
  FarmArea      0.2126576    0.04560949       4.66      0.0004

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


                   Covariance of Estimated
                   Regression Coefficients

                          Intercept          FarmArea

         Intercept      28.300381277      -0.146471538
         FarmArea       -0.146471538      0.0020802259
```

Output 62.4.2 displays tests of model effects and the estimated regression coefficients and their covariance matrix.

Alternatively, you can assume that the linear relationship between corn yield (CornYield) and farm area (FarmArea) is different among the states. Therefore, you consider fitting Model II.

In order to analyze the data using Model II, you create auxiliary variables FarmAreaNE and FarmAreaIA to represent farm area in different states:

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

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

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

```
title1 'Analysis of Farm Area and Corn Yield';
title2 'Model II: Same Intercept, Different Slopes';
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 using the new data set FarmsByState. The analysis uses the auxilary variables FarmAreaIA and FarmAreaNE as the regressors.

```
proc SURVEYREG data=FarmsByState total=TotalInStrata;
   strata State Region;
   model CornYield = FarmAreaIA FarmAreaNE / covb;
   weight Weight;
run;
```

Output 62.4.3. Regression Results from Fitting Model II

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

                    The SURVEYREG Procedure

        Regression Analysis for Dependent Variable CornYield

                         Data Summary

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


                         Design Summary

                   Number of Strata             5


                         Fit Statistics

                     R-square            0.8158
                     Root MSE           11.6759
                     Denominator DF          14


                 Estimated Regression Coefficients

                                 Standard
 Parameter       Estimate           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

                     Intercept        FarmAreaIA        FarmAreaNE

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

Output 62.4.3 displays the data summary, design information, fit summary, and parameter estimates and their covariance matrix. 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.

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

```
title1 'Analysis of Farm Area and Corn Yield';
title2 'Model III: Different Intercepts and Slopes';
proc SURVEYREG data=FarmsByState total=TotalInStrata;
   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 will presents the intercepts in two states.

Output 62.4.4. 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

                           Data Summary

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


                           Design Summary

                    Number of Strata            5


                           Fit Statistics

                     R-square           0.9300
                     Root MSE          11.9810
                     Denominator DF         14


                  Estimated Regression Coefficients

                                      Standard
    Parameter            Estimate         Error    t Value    Pr > |t|

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

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


           Covariance of Estimated Regression Coefficients

                                      State
                  State Iowa        Nebraska      FarmAreaIA      FarmAreaNE

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

Output 62.4.4 displays the regression results for fitting Model III, including the data summary, 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.

Example 62.5. Regression Estimator for Stratified Sample

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

Similar to Example 62.3 on page 3251, by incorporating auxilary 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 sampling design. The auxilary information is the total farm areas in regions of each state, as displayed in Table 62.4. You want to estimate the total corn yield using this information under the three linear models given in Example 62.4.

Table 62.4. Information for Each Stratum

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

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

```
title1 'Estimate Corn Yield from Farm Size';
title2 'Model I: Same Intercept and Slope';
proc surveyreg data=Farms total=TotalInStrata;
   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 contraint in each stratum that the weighted total number of farms equals to the total number of farms in the stratum, you can include the strata as an effect in the MODEL statement, effect State*Region. Thus, the CLASS statement must list

the STRATA variables, State and Region, as classification variables. The following ESTIMATE statement specifies the regression estimator, which is a linear function of the regression parameters.

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

This linear function contains the total for each explanatory variable in the model. Because the sampling units are farms in this example, the coefficient for Intercept in the ESTIMATE statement is the total number of farms (235); the coefficient for FarmArea is the total farm area listed in Table 62.4 (21950); and the coefficients for effect State*Region are the total number of farms in each strata (as displayed in Table 62.4).

Output 62.5.1. Regression Estimator for the Total of CornYield under Model I

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

                        The SURVEYREG Procedure

            Regression Analysis for Dependent Variable CornYield

                    Analysis of Estimable Functions

                                                 Standard
Parameter                              Estimate       Error  t Value  Pr > |t|

Estimate of CornYield under Model I  7463.52329  926.841541     8.05    <.0001

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

Output 62.5.1 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.

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

```
title1 'Estimate Corn Yield from Farm Size';
title2 'Model II: Same Intercept, Different Slopes';
proc surveyreg data=FarmsByState total=TotalInStrata;
   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 62.4). 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 62.4. The total number of farms in each strata are the coefficients for the strata effect.

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

Output 62.5.2. Regression Estimator for the Total of CornYield under Model II

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

                          The SURVEYREG Procedure

              Regression Analysis for Dependent Variable CornYield

                          Analysis of Estimable Functions

                                                  Standard
Parameter                              Estimate        Error   t Value   Pr > |t|

Total of CornYield under Model II   7580.48657   859.180439      8.82     <.0001

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

Output 62.5.2 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.

```
title1 'Estimate Corn Yield from Farm Size';
title2 'Model III: Different Intercepts and Slopes';
proc SURVEYREG data=FarmsByState total=TotalInStrata;
   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, using the NOINT option. Among the ESTIMATE statements, the coefficients for explanatory variables are different depending on which regression estimator is estimated. For example, in the ESTIMATE statement

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

the coefficients for the effect State are 165 and 0, respectively. This indicates that the total number of farms in Iowa is 165 and the total number of farms in Nebraska is 0, because the estimation is the total corn yield in Iowa only. Similarly, the total numbers of farms in three regions in Iowa are used for the coefficients of the strata effect State*Region, as displayed in Table 62.4.

Output 62.5.3. Regression Estimator for the Total of CornYield under Model III

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

                         The SURVEYREG Procedure

             Regression Analysis for Dependent Variable CornYield

                      Analysis of Estimable Functions

                                                          Standard
Parameter                                    Estimate        Error  t Value

Total CornYield in Iowa under Model III    6246.10697  851.272372     7.34
Total CornYield in Nebraska under Model III 1334.37961  116.302948    11.47
Total CornYield in both states under Model III 7580.48657  859.180439     8.82

                      Analysis of Estimable Functions

           Parameter                                    Pr > |t|

           Total CornYield in Iowa under Model III        <.0001
           Total CornYield in Nebraska under Model III    <.0001
           Total CornYield in both states under Model III <.0001

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

Output 62.5.3 displays the results from the three regression estimators 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.

Example 62.6. Stratum Collapse

In a stratified sample, it is possible that some strata will have only one sampling unit. When this happens, PROC SURVEYREG collapses these strata that contain single sampling unit into a pooled stratum. For more detailed information on stratum collapse, see the section "Stratum Collapse" on page 3236.

Suppose that you have the following data.

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

The variable Stratum is 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 StratumTotal.

```
data StratumTotal;
   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 StratumTotal contains more strata than the data set Sample. Also in the data set StratumTotal, 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 the strata that are not present in the data set Sample; for the multiple entries of a stratum, the procedure uses the first observation. In this example, Stratum=33 has the stratum total _TOTAL_=40.

The following SAS statements perform the regression analysis.

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

Output 62.6.1. Summary of Data and Regression

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

                  The SURVEYREG Procedure

          Regression Analysis for Dependent Variable Y

                       Data Summary

              Number of Observations            8
              Mean of Y                   2.75000
              Sum of Y                   22.00000


                       Design Summary

            Number of Strata                     5
            Number of Strata Collapsed           2


                       Fit Statistics

                  R-square           0.9555
                  Root MSE           0.5129
                  Denominator DF          4
```

Output 62.6.1 displays that there are a total of 5 strata in the input data set, and 2 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 3237).

Output 62.6.2. Stratification Information

Stratified Sample with Single Sampling Unit in Strata
With Stratum Collapse

The SURVEYREG Procedure

Regression Analysis for Dependent Variable Y

Stratum Information

Stratum Index	Collapsed	Stratum	N Obs	Population Total	Sampling Rate
1		10	2	10	0.20
2		11	2	20	0.10
3		12	2	32	0.06
4	Yes	14	1	50	0.02
5	Yes	33	1	40	0.03
0	Pooled		2	90	0.02

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

Output 62.6.2 displays the stratification information, including stratum collapse. Under the column Collapsed, the fourth (Stratum Index=4) stratum and the fifth (Stratum Index=5) stratum 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%, which combined from the 4th stratum and the 5th stratum (see the section "Sampling Rate of the Pooled Stratum from Collapse" on page 3241).

Output 62.6.3. Parameter Estimates and Effect Tests

Stratified Sample with Single Sampling Unit in Strata
With Stratum Collapse

The SURVEYREG Procedure

Regression Analysis for Dependent Variable Y

Tests of Model Effects

Effect	Num DF	F Value	Pr > F
Model	1	155.62	0.0002
Intercept	1	0.24	0.6503
X	1	155.62	0.0002

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

Estimated Regression Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.13004484	0.26578532	0.49	0.6503
X	1.10313901	0.08842825	12.47	0.0002

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

Output 62.6.3 displays the parameter estimates and the tests of the significance of the model effects.

Alternatively, if you prefer not to collapse the strata that have single sampling unit, you can specify the NOCOLLAPSE option in the STRATA statement.

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

Output 62.6.4. Summary of Data and Regression

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

                   The SURVEYREG Procedure

         Regression Analysis for Dependent Variable Y

                        Data Summary

             Number of Observations             8
             Mean of Y                    2.75000
             Sum of Y                    22.00000


                        Design Summary

                Number of Strata             5


                        Fit Statistics

                  R-square            0.9555
                  Root MSE            0.5129
                  Denominator DF           3
```

Output 62.6.4 does not contain stratum collapse information as compared to Output 62.6.1. The denominator degrees of freedom is 3 instead of 4 as in Output 62.6.1.

Output 62.6.5. Stratification Information

Stratified Sample with Single Sampling Unit in Strata
Without Stratum Collapse

The SURVEYREG Procedure

Regression Analysis for Dependent Variable Y

Stratum Information

Stratum Index	Stratum	N Obs	Population Total	Sampling Rate
1	10	2	10	0.20
2	11	2	20	0.10
3	12	2	32	0.06
4	14	1	50	0.02
5	33	1	40	0.03

In Output 62.6.5, although the fourth stratum and the fifth stratum contain only one observation, no stratum collapse occurs as in Output 62.6.2.

Output 62.6.6. Parameter Estimates and Effect Tests

Stratified Sample with Single Sampling Unit in Strata
Without Stratum Collapse

The SURVEYREG Procedure

Regression Analysis for Dependent Variable Y

Tests of Model Effects

Effect	Num DF	F Value	Pr > F
Model	1	391.94	0.0003
Intercept	1	0.25	0.6508
X	1	391.94	0.0003

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

Estimated Regression Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.13004484	0.25957741	0.50	0.6508
X	1.10313901	0.05572135	19.80	0.0003

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

As a result of not collapsing strata, the standard error estimates of the parameters are different from those in Output 62.6.3, the tests of the significance of model effects are different as well.

Example 62.7. Domain Analysis

Recall that in the section "Getting Started" on page 3217, you collected 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 also use the same sample to estimate the average weekly spending among male and female students, respectively. This is often called domain analysis (subgroup analysis). You can use PROC SURVEYREG to perform domain analysis as in the following example.

```
data IceCreamData;
   input Grade Spending Income Gender$ @@;
   if Gender='M' then Male=1; else Male=0;
   if Gender='F' then Female=1; else Female=0;
   datalines;
 7   7  39  M   7   7  38  F   8  12  47  F
 9  10  47  M   7   1  34  M   7  10  43  M
 7   3  44  M   8  20  60  F   8  19  57  M
 7   2  35  M   7   2  36  F   9  15  51  F
 8  16  53  F   7   6  37  F   7   6  41  M
 7   6  39  M   9  15  50  M   8  17  57  F
 8  14  46  M   9   8  41  M   9   8  41  F
 9   7  47  F   7   3  39  F   7  12  50  M
 7   4  43  M   9  14  46  F   8  18  58  M
 9   9  44  F   7   2  37  F   7   1  37  M
 7   4  44  M   7  11  42  M   9   8  41  M
 8  10  42  M   8  13  46  F   7   2  40  F
 9   6  45  F   9  11  45  M   7   2  36  F
 7   9  46  F
 ;
```

In the data set IceCreamData, 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. Male and Female are two indicator variables that identify the subgroups of male and female students, respectively.

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

In the data set StudentTotal, 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 estimate the average spending in the subgroup of male students.

```
title1 'Ice Cream Spending Analysis';
title2 'Domain Analysis for Subgroup: Male Students';
proc surveyreg data=IceCreamData total=StudentTotal;
   strata Grade;
   model Spending = Male / noint;
   ods select ParameterEstimates;
run;
```

Output 62.7.1. Domain Analysis for Male Students

```
                  Ice Cream Spending Analysis
          Domain Analysis for Subgroup: Male Students

                    The SURVEYREG Procedure

       Regression Analysis for Dependent Variable Spending

                Estimated Regression Coefficients

                                 Standard
   Parameter      Estimate         Error    t Value    Pr > |t|

   Male         8.57142857    0.97971846       8.75      <.0001

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

Output 62.7.1 shows that average spending for the subgroup of male students is $8.57 with a standard error of $.99.

Similarly, you can obtain a domain analysis for the subgroup of female students with the following statements.

```
title1 'Ice Cream Spending Analysis';
title2 'Domain Analysis for Subgroup: Female Students';
proc surveyreg data=IceCreamData total=StudentTotal;
   strata Grade /list;
   model Spending = Female / noint;
run;
```

Output 62.7.2. Domain Analysis for Female Students

```
                Ice Cream Spending Analysis
          Domain Analysis for Subgroup: Female Students

                The SURVEYREG Procedure

      Regression Analysis for Dependent Variable Spending

              Estimated Regression Coefficients

                                Standard
  Parameter      Estimate          Error    t Value    Pr > |t|

  Female       8.94736842    1.06370643       8.41      <.0001

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

Output 62.7.2 shows that average spending for the subgroup of female students is $8.95 with a standard error of $1.06.

Note that you would not obtain the same results by using a subset of your sample, for example, by restricting the analysis to male students using a WHERE clause or a BY statement. This is because the domain sample size is not fixed in the original sample design, but is actually a random variable. The variance estimation for the domain mean must include this variability of the sample size. Refer to Cochran (1977) for more details.

References

Cochran, W. G. (1977), *Sampling Techniques*, Third Edition, New York: John Wiley & Sons, Inc.

Foreman, E. K. (1991), *Survey Sampling Principles*, New York: Marcel Dekker, Inc.

Fuller, W. A. (1975), "Regression Analysis for Sample Survey," *Sankhyā*, 37 (3), Series C, 117–132.

Kish, L. (1965), *Survey Sampling*, New York: John Wiley & Sons, Inc.

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

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

Statistical Laboratory (1989), *PC CARP*, Ames, IA: Statistical Laboratory, Iowa State University.

Woodruff, R. S. (1971), "A Simple Method for Approximating the Variance of a Complicated Estimate," *Journal of the American Statistical Association*, 66, 411–414.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/STAT® User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999.

SAS/STAT® User's Guide, Version 8

ISBN 1–58025–494–2

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

1st printing, October 1999