SAS ® Procedures for Analysis of Sample Survey Data

Anthony An and Donna Watts, SAS Institute Inc., Cary, NC

Anthony An, SAS Institute Inc., SAS Campus Drive, R5243, Cary, NC 27513

KEY WORDS

Statistical software, sample surveys, complex sample design, probability-based sampling, variance estimation, data analysis

ABSTRACT

Researchers use sample surveys to obtain information on a wide variety of issues. Many surveys are based on probability-based complex sample designs, which may include stratification, clustering, and unequal weighting. To make statistically valid inferences from the sample to the study population, researchers must analyze the data taking into account the sample design. In Version 8 of the SAS® System, new procedures are available for sample selection and for analysis of data from complex sample surveys. These procedures use input describing the sample design to produce the appropriate statistical analyses.

The SURVEYSELECT procedure selects probability samples using various sample designs, including stratified sampling and sampling with probability proportional to size. The SURVEYMEANS procedure computes descriptive statistics for sample survey data, including means, totals, proportions, ratios, and their standard errors. The SURVEYREG procedure fits linear regression models and produces hypothesis tests and estimates for survey data. This paper describes the capabilities of these procedures and illustrates their use.

INTRODUCTION

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

Traditional SAS procedures, such as the MEANS procedure and the GLM procedure, compute statistics under the assumption that the sample is drawn from an infinite population by simple random sampling. These procedures generally do not correctly estimate the variance of an estimate if they are applied to a sample drawn by a complex sample design. Therefore SAS users have requested procedures that analyze data from complex sample surveys. In response to this request, SAS Institute has developed the SURVEYSELECT, SURVEYMEANS, and SURVEYREG procedures. These new procedures are available in Version 8 of the SAS System. As part of SAS/STAT® software, these procedures are fully integrated with the data access, management and presentation tools of the SAS system.

The SURVEYSELECT procedure provides a variety of methods for selecting probability-based random samples. The section "Sample Selection" describes PROC SURVEYSELECT in detail. PROC SURVEYMEANS and PROC SURVEYREG analyze survey data collected according to a complex survey design. The section "Descriptive Statistics" shows the use of PROC SURVEYMEANS to obtain population parameter estimates. The section "Regression Analysis" describes PROC SURVEYREG and illustrates how to perform regression analysis for survey data.

Complete documentation describing the syntax and statistical methodology for these new procedures is available in SAS Institute (1999). You can also obtain information on these procedures at the SAS Institute Research and Development web site: http://www.sas.com/rnd/ Enhancements to these three procedures, as well as additional procedures for the analysis of survey data, are currently under development and will be available in future releases of the SAS System.

SAMPLE SELECTION

The SURVEYSELECT procedure provides a variety of methods for selecting probability-based random samples. The procedure can select a simple random sample or a sample according to a complex multi-stage sample design that includes stratification, clustering, and unequal probabilities of selection. The procedure uses fast, efficient algorithms for sample selection. Thus, it performs well even for very large input data sets or sampling frames, which may occur in practice for large-scale sample surveys.

To select a sample with PROC SURVEYSELECT, you input a SAS data set that contains the sampling frame, or list of units from which the sample is to be selected. You also specify the selection method, the desired sample size or sampling rate, and other selection parameters. The SURVEYSELECT procedure selects the sample, producing an output data set that contains the selected units, their selection probabilities, and sampling weights. When you are selecting a sample in multiple stages, you invoke the procedure separately for each stage of selection, inputting the frame and selection parameters for each current stage.

Capabilities

The SURVEYSELECT procedure provides methods for both equal probability sampling and probability proportional to size (PPS) sampling. In equal probability sampling, each unit in the sampling frame, or in a stratum, has the same probability of being selected for the sample. In PPS sampling, a unit's selection probability is proportional to its size measure. For details on probability sampling methods, refer to Cochran (1977), Kish (1965), and Kalton (1983).

The SURVEYSELECT procedure provides the following equal probability sampling methods:

- simple random sampling
- unrestricted random sampling (with replacement)
- systematic random sampling
- sequential random sampling

This procedure also provides the following probability proportional to size (PPS) methods:

- PPS without replacement
- PPS with replacement
- PPS systematic
- various PPS algorithms for selecting two units per stratum
- sequential PPS with minimum replacement

The SURVEYSELECT procedure can perform stratified sampling, selecting samples independently within the specified strata, or non-overlapping subgroups of the survey population. Stratification controls the distribution of the sample size in the strata and is widely used in practice towards meeting a variety of survey objectives. When you are using a sequential selection method, the SURVEYSELECT procedure also can sort by control variables within strata for the additional control of implicit stratification.

The SURVEYSELECT procedure provides replicated sampling, where the total sample is composed of a set of replicates, each selected in the same way. You can use replicated sampling to study variable nonsampling errors, such as variability in the results obtained by different interviewers. You can also use replication to compute standard errors for the combined sample estimates.

Syntax

The following statements control the SURVEYSELECT procedure. Items within the <> are optional.

PROC SURVEYSELECT *<options>*;
SIZE *variable*;
STRATA *variables*;
CONTROL *variables*;
ID *variables*;

The PROC SURVEYSELECT statement invokes the procedure. The options in this statement name the input and output data sets and specify the sample selection method, the sample size or sampling rate, and other sampling parameters. The SIZE statement specifies the variable that contains the size measure and is required whenever the sample selection method is probability proportional to size. All other statements are optional. The STRATA statement names one or more stratification variables. The CONTROL statement, which you can use with sequential sampling methods, specifies one or more variables for ordering units within strata. The ID statement identifies variables to copy from the input data set to the output data set of selected sampling units.

Input

In the PROC SURVEYSELECT statement, you identify the sampling frame and specify sample selection parameters. The DATA= option names the sampling frame, or input data set. This data set should contain any variables identified in the STRATA, CONTROL, and SIZE statements, and it should be sorted by the STRATA variables. Use the METHOD= option to specify the selection method. If you omit this option, by default the procedure uses simple random sampling if there is no SIZE statement or PPS sampling if there is a SIZE statement.

You must specify the desired sample size or sampling rate for sample selection. If you are not using stratified selection, or if the sample size or sampling rate is the same for all strata, you can use the N= option to specify the sample size or the R= option to specify the sampling rate. To specify sample sizes or rates by strata, you can use an input data set that contains the STRATA variables and a sample size or rate variable. Alternatively, you can use the N=$(n_1, n_2, ..., n_s)$ syntax in the PROC SURVEYSELECT statement, listing the stratum sample sizes $n_1, n_2, ..., n_s$, in the order in which the strata appear in the input data set. Similar syntax is available for the R= option.

You can specify other sample selection options in the PROC SURVEYSELECT statement. The SEED= option specifies the initial seed for the random number generator. You can use the REP= option to specify the number of replicates to be selected. The procedure selects replicates independently, each with the specified sample size or rate. For sequential selection methods using CONTROL variables, you can specify the type of sorting, nested or serpentine, with the SORT= option. There are also options available to specify minimum, maximum, and certainty size measures when using PPS selection. Other options request additional sample selection statistics for the output data set.

Output

The SURVEYSELECT procedure produces an output data set that contains the sample of selected units plus selection information for use in sampling weight construction and survey data analysis. This output data set has one observation for each unit in the sample. It contains any STRATUM, CONTROL, and SIZE variables specified for sample selection. It also contains the selection probability, or expected number of hits for methods that allow multiple selections per sampling unit, and the sampling weight component for each selected unit. Optionally, joint probabilities of selection are available for certain PPS selection methods. Other output variables include the number of hits and the replicate number for replicated sampling.

Example

This example shows the use of PROC SURVEYSELECT to select a stratified random sample. The data were constructed solely for illustrative purposes. In this example, an Internet service provider wants to conduct a customer satisfaction survey. The survey population consists of the company's current subscribers, and the company plans to select a sample of customers and make inferences for the survey population from the sample data.

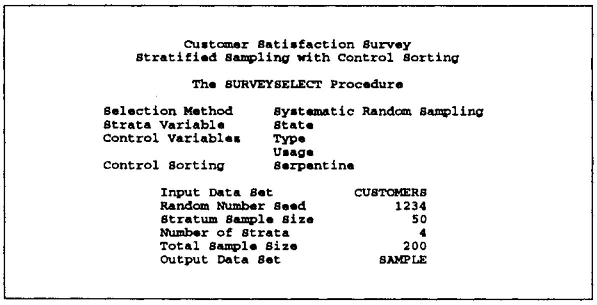
The SAS data set Customers contains the sampling frame, or list of units in the survey population. The data set Customers contains an observation for each customer, with a total of 13,471 observations. In this data set, the variable CustomerID uniquely identifies each customer. The variable State contains the state of the customer's address. The company has customers in the following four states: Georgia (GA), Alabama (AL), Florida (FL), and South Carolina (SC). The variable Type equals 'Old' if the customer has subscribed to the service for more than one year; otherwise, the variable Type equals 'New'. The variable Usage contains the customer's average monthly service usage, in minutes.

The sample design for the customer satisfaction survey is stratified by State. Within these strata, the design specifies sorting by Type and Usage, to provide additional control over the distribution of the sample. Customers are then selected by systematic random sampling within strata. The following PROC SURVEYSELECT statements select a probability sample of customers from the Customers data set using this design:

```
title1 'Customer Satisfaction Survey';
title2 'Stratified Sampling
        with Control Sorting';
proc surveyselect  data=Customers
       method=sys  seed=1234  n=50
       out=Sample;
   strata   State;
   control  Type  Usage;
run;
```

The STRATA statement names the stratification variable State. The CONTROL statement names the control variables Type and Usage. In the PROC SURVEYSELECT statement, the METHOD=SYS option requests systematic random sampling. The SEED=1234 option specifies the initial seed for random number generation. The N=50 option specifies a sample size of 50 customers for each stratum. To specify different sample sizes for different strata, use the N=*SAS-data-set* option to name a data set that contains the stratum sample sizes, or use the N=(*values*) option to list the desired sample sizes in the PROC statement.

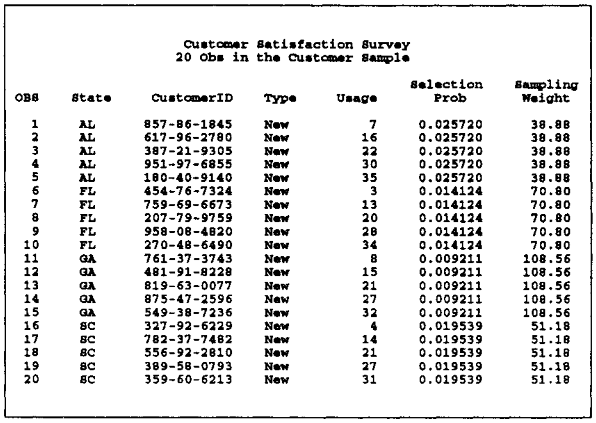
Figure 1 displays the output from PROC SURVEYSELECT, which summarizes the sample selection. A sample of 200 customers is selected, using systematic random sampling within strata determined by State. The sampling frame Customers is sorted by control variables Type and Usage within strata. The type of sorting is serpentine, which is used by default since SORT=NEST is not specified. For information on serpentine sorting, refer to Chromy (1979) and Williams and Chromy (1980).



```
                 Customer Satisfaction Survey
            Stratified Sampling with Control Sorting

                  The SURVEYSELECT Procedure

      Selection Method      Systematic Random Sampling
      Strata Variable       State
      Control Variables     Type
                            Usage
      Control Sorting       Serpentine

            Input Data Set              CUSTOMERS
            Random Number Seed               1234
            Stratum Sample Size                50
            Number of Strata                    4
            Total Sample Size                 200
            Output Data Set                SAMPLE
```

Figure 1. Sample Selection Summary

The output data set Sample contains the sample of 200 customers. This data set includes all the variables from the DATA= input data set Customers. It also contains the variable SelectionProb, which stores the selection probability for each customer in the sample. The variable SamplingWeight contains the sampling weights, which are computed as inverse selection probabilities. Figure 2 displays 20 observations of the output data set Sample.



```
                       Customer Satisfaction Survey
                      20 Obs in the Customer Sample

                                                          Selection    Sampling
OBS    State     CustomerID     Type      Usage       Prob      Weight

  1     AL      857-86-1845     New          7      0.025720       38.88
  2     AL      617-96-2780     New         16      0.025720       38.88
  3     AL      387-21-9305     New         22      0.025720       38.88
  4     AL      951-97-6855     New         30      0.025720       38.88
  5     AL      180-40-9140     New         35      0.025720       38.88
  6     FL      454-76-7324     New          3      0.014124       70.80
  7     FL      759-69-6673     New         13      0.014124       70.80
  8     FL      207-79-9759     New         20      0.014124       70.80
  9     FL      958-08-4820     New         28      0.014124       70.80
 10     FL      270-48-6490     New         34      0.014124       70.80
 11     GA      761-37-3743     New          8      0.009211      108.56
 12     GA      481-91-8228     New         15      0.009211      108.56
 13     GA      819-63-0077     New         21      0.009211      108.56
 14     GA      875-47-2596     New         27      0.009211      108.56
 15     GA      549-38-7236     New         32      0.009211      108.56
 16     SC      327-92-6229     New          4      0.019539       51.18
 17     SC      782-37-7482     New         14      0.019539       51.18
 18     SC      556-92-2810     New         21      0.019539       51.18
 19     SC      389-58-0793     New         27      0.019539       51.18
 20     SC      359-60-6213     New         31      0.019539       51.18
```

Figure 2. 20 Observations of the Customer Sample

DESCRIPTIVE STATISTICS

The SURVEYMEANS procedure produces estimates of the following survey population parameters:

- means
- totals
- proportions
- ratios

For these estimates, the procedure can produce standard errors, variances, confidence limits, coefficients of variation, and *t*-tests. In addition to estimates for the entire survey population, the procedure can compute estimates for population subgroups or domains. The procedure also provides data summary and design summary information.

Computational Method

The SURVEYMEANS procedure uses the Taylor expansion method to estimate sampling errors of estimators based on complex sample designs. This method obtains a linear approximation for the estimator and then uses the variance estimate for this approximation to estimate the variance of the estimate itself (Woodruff 1971, Fuller 1975). PROC SURVEYMEANS uses Taylor expansion to estimate the variance of the population total. When there are clusters, or primary sampling units (PSUs), in the sample design, the procedure estimates the variance from the variation among PSU totals. When the design is stratified, the procedure pools stratum variance estimates to compute the variance estimate for the population total.

The variance estimates for the mean and the mean PSU total are based on the variance estimate for the population total. For *t*-tests of the estimates, the degrees of freedom equals the number of clusters minus the number of strata in the sample design.

For a multistage sample design, the variance estimation method depends only on the first stage of the sample design. So, the required input includes only first-stage cluster (PSU) and first-stage stratum identification. You do not need to input design information for any additional stages of sampling. This variance estimation method assumes that the first-stage sample is drawn with replacement, as it often is in practice, or that the first-stage sampling fraction is small. This assumption may result in an overestimate of the variance, but this should be fairly small if the first-stage sampling fraction is small.

For more information on the analysis of sample survey data, refer to Binder and Roberts (1999), Lee, Forthoffer, and Lorimor (1989), Cochran (1977), Kish (1965), and Hansen, Hurwitz, and Madow (1953).

Syntax

The following statements control the SURVEYMEANS procedure. Items within the <> are optional.

PROC SURVEYMEANS *<options>*
    <statistic-keywords>;
VAR *variables*;
CLASS *variables*;
RATIO *<'label'> variables / variables*;
DOMAIN *variables < variable∗variable*
    variable∗variable∗variable . . .>;
STRATA *variables / <options>*;
CLUSTER *variables*;
WEIGHT *variable*;
BY *variables*;

The PROC SURVEYMEANS statement invokes the procedure. It optionally names the input data sets and specifies statistics for the procedure to compute. The VAR statement identifies the variables to be analyzed. The CLASS statement identifies those numeric variables that are to be analyzed as categorical variables.

The RATIO statement requests ratio analysis for means or proportions of analysis variables. The statement names the variables whose means will be used as numerators or denominators in the ratio, the numerator variables appearing before the slash "/" and the denominator variables after the slash. The RATIO statement is new for the SURVEYMEANS procedure in Release 8.2 of the SAS System.

The DOMAIN statement lists variables that define domains for subpopulation analyses. The DOMAIN statement is new in Release 8.1 of the SAS System.

The STRATA statement lists the variables that form the strata in a stratified sample design. The CLUSTER statement specifies cluster identification variables in a clustered sample design. The WEIGHT statement names the sampling weight variable. You can use a BY statement with PROC SURVEYMEANS to obtain separate, independent analyses for groups defined by the BY variables.

Input

In the PROC statement, you identify the data set to be analyzed and specify sample design information. The DATA= option names the input data set to be analyzed. If your analysis include a finite population correction factor, you can input either the sampling rate or the population total using the R= or N= option. If your design is stratified, with different sampling rates or totals for different strata, then you can input these rates or totals in a SAS data set containing the STRATA variables. You provide other sample design information to PROC SURVEYMEANS in the STRATA, CLUSTER, and WEIGHT statements.

In the PROC SURVEYMEANS statement, you also specify statistics for the procedure to compute. Available statistics include the population mean (or proportion), the population total, and ratios specified in the RATIO statement, together with the corresponding variance estimates, confidence limits, and *t*-tests. You can also request data set summary information and sample design information, such as the number of PSUs, the sampling rates, and the sum of the sampling weights. You use the LIST option in the STRATA statement to request stratum-level information, including the number of observations, number of PSUs, and sampling rate for each stratum.

Output

PROC SURVEYMEANS displays the information and statistics you request, as described above. Available tables include "Data Summary", "Statistics", "Class Level Information", "Stratum Information", and "Domain Analysis". If you do not specify statistics to compute, by default the procedure provides the estimate of the population mean, its standard error, and 95% confidence limits for each analysis variable. You can save any displayed output from the SURVEYMEANS procedure in a SAS data set using the Output Delivery System.

Example

This example shows the use of PROC SURVEYMEANS to estimate population totals from survey data. The data were constructed solely for illustrative purposes. In this example, a marketing research firm has a household database with information on 235 households in North Carolina and South Carolina. The firm wants to estimate the total income and the total basic living expense of these households for the past year. A probability sample of households is selected from this database, or survey population. The sample design is a one-stage stratified design. The sampling frame, or list of households in the study population, is stratified by geographical region within state. Within each stratum, a sample of households is selected using simple random sampling. The total sample size is 19 households.

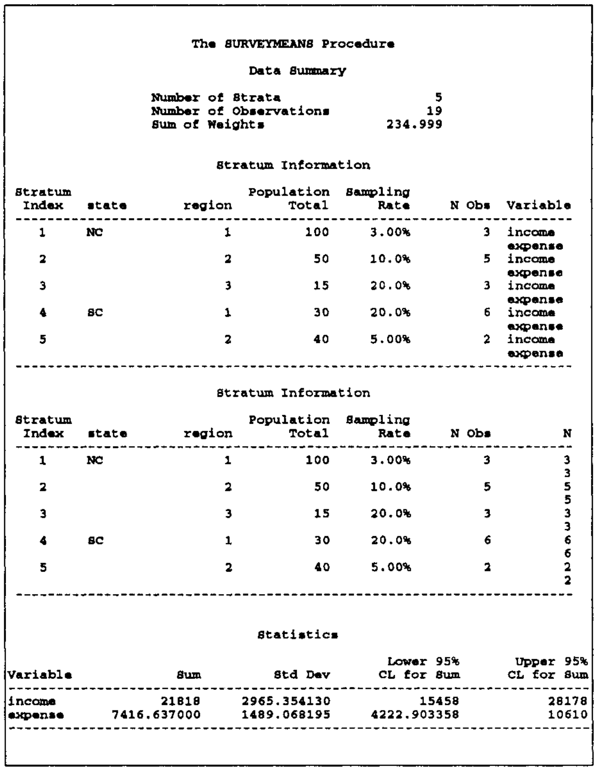
The data set HHSample contains an observation for each sample household and the following variables: ID, the household identification number; STATE; REGION; INCOME, the income for the past year; EXPENSE, the basic living expense; and WEIGHT, the sampling weight. The data set StrataTotals is used to provide stratum size information to PROC SURVEYMEANS. It contains the stratum variables STATE and REGION and the variable _TOTAL_, which gives the total number of households in the population for each stratum.

The following PROC SURVEYMEANS statements estimate the total income and total basic living expenses for this survey population.

```
proc surveymeans  data=HHSample
      N=StrataTotals
      sum  clsum  fraction;
   var     income  expense;
   strata  state  region  /  list;
   weight  weight;
run;
```

The PROC statement invokes the procedure and names the input data sets. The data set HHSample contains the survey data to be analyzed. The data set StrataTotals provides the population total (number of households) for each stratum. The SUM option requests estimates of population totals and their standard deviations for the analysis variables. The CLSUM option requests confidence limits for the estimates.

The VAR statement specifies the two analysis variables, INCOME and EXPENSE. The STRATA statement names STATE and REGION as the stratification variables in the sample design. The LIST option in the STRATA statement requests stratum-level data summary and design information. The WEIGHT statement names WEIGHT as the sampling weight variable.



```
                      The SURVEYMEANS Procedure

                             Data Summary

                 Number of Strata                      5
                 Number of Observations               19
                 Sum of Weights                  234.999


                         Stratum Information

Stratum                           Population  Sampling
 Index   state           region        Total      Rate      N Obs  Variable
--------------------------------------------------------------------------
   1     NC                   1          100     3.00%          3  income
                                                                   expense
   2                          2           50     10.0%          5  income
                                                                   expense
   3                          3           15     20.0%          3  income
                                                                   expense
   4     SC                   1           30     20.0%          6  income
                                                                   expense
   5                          2           40     5.00%          2  income
                                                                   expense
--------------------------------------------------------------------------

                         Stratum Information

Stratum                           Population  Sampling
 Index   state           region        Total      Rate      N Obs        N
--------------------------------------------------------------------------
   1     NC                   1          100     3.00%          3        3
                                                                         3
   2                          2           50     10.0%          5        5
                                                                         5
   3                          3           15     20.0%          3        3
                                                                         3
   4     SC                   1           30     20.0%          6        6
                                                                         6
   5                          2           40     5.00%          2        2
                                                                         2
--------------------------------------------------------------------------


                             Statistics

                                                Lower 95%        Upper 95%
Variable           Sum          Std Dev        CL for Sum       CL for Sum
--------------------------------------------------------------------------
income           21818      2965.354130             15458            28178
expense    7416.637000      1489.068195       4222.903358            10610
--------------------------------------------------------------------------
```

Figure 3. Output from PROC SURVEYMEANS

Figure 3 shows the data summary and statistics from PROC SURVEYMEANS. There are 5 strata and 19 observations in the sample. The "Stratum Information" table shows the population total, sampling rate, and sample size (column "N Obs") for each stratum. Also for each stratum, this table gives the number of observations included in the analysis for each variable (column "N").

The "Statistics" table displays the estimated population totals and standard deviations for the variables INCOME and EXPENSE. This table also shows the 95% confidence limits for these estimates. For the survey population of 235 households, the estimated total income is $21,818 (in thousands) with standard deviation $2,965 (in thousands). The estimated total living expenses of these households is $7,417 (in thousands) with standard deviation $1,489 (in thousands).

REGRESSION ANALYSIS

The SURVEYREG procedure performs regression analysis for sample survey data. 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.

Computational Method

The SURVEYREG procedure 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 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 *et al.* 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.

When there are clusters, or PSUs, in the sample design, PROC SURVEYREG estimates the covariance matrix from the variation among PSU totals. When the design is stratified, the procedure pools stratum variance estimates to compute the covariance matrix. Wald's *F*-test and the *t*-test for estimators and effects are based on the estimated covariance matrix of the regression coefficients. For these tests, if you do not provide the denominator degrees of freedom using the DF= option in the PROC statement, by default the denominator degrees of freedom equals the number of clusters minus the number of strata in the sample design. This variance estimation method assumes that first-stage sampling is with replacement and does not require input information on any additional stages of sampling. See "Computational Method" in the section "Descriptive Statistics."

Syntax

The following statements control the SURVEYREG procedure. Items within the <> are optional.

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

The PROC statement invokes the procedure. You can use options in this statement to name the input data set to be analyzed and specify the sample design information.

The CLASS statement identifies those variables that are to be treated as categorical variables in the MODEL statement, and it must appear before the MODEL statement. The MODEL statement, which is required, specifies the dependent (response) variable and the independent 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 by a variable name or by a special notation using variable names and operators, as described in "The GLM Procedure" in SAS Institute (1999). You can use only one numerical variable as the dependent variable in the MODEL statement.

You can use an ESTIMATE statement to estimate a linear function of the regression parameters by giving the coefficients for each effect in the model. You can use a CONTRAST statement to obtain custom hypothesis tests for linear combinations of the regression parameters.

The STRATA statement lists the variables that form the strata in a stratified sample design. The CLUSTER statement specifies cluster identification variables in a clustered sample design. The WEIGHT statement names the sampling weight variable. You can use a BY statement to perform separate, independent regression analyses for population subgroups.

Input

In the PROC statement, you identify the data set to be analyzed and specify sample design information. The DATA= option names the input data set to be analyzed. If your analysis includes a finite population correction factor, you can input either the sampling rate or the population total using the R= or N= option. If your design is stratified with different sampling rates or totals for different strata, then you can input these rates or totals in a SAS data set containing the STRATA variables. You can provide other sample design information in the STRATA, CLUSTER, and WEIGHT statements.

In the MODEL statement, you specify the model to be fitted and request statistics for that model. To estimate an estimable linear function of the regression parameters, you specify the coefficients for each effect parameter in the ESTIMATE statement. To test custom hypotheses for linear combinations of the regression parameters, you provide the coefficients for each linear function in the CONTRAST statement.

You can use the LIST option in the STRATA statement to request stratum-level information, including the number of observations, number of PSUs, and sampling rate for each stratum. You can use the NOCOLLAPSE option to control strata collapsing for the variance estimation when there are empty strata or single unit strata. By default, the procedure collapses those strata that contain fewer than two sampling units into a pooled stratum, computes the sampling rate in the pooled stratum, and adjusts the degrees of freedom in the variance estimation.

Output

PROC SURVEYREG presents the regression analysis results in several tables, depending on the options you request. Available output includes the following:

- data summary and design summary information
- a one-way analysis of variance for the dependent variable
- Wald's *F*-test for all effects in the model
- regression fit statistics
- estimates of regression coefficients, their standard errors, and *t*-tests
- analysis of contrasts
- analysis of estimable functions

You can save any displayed output from the SURVEYREG procedure in a SAS data set using the Output Delivery System.

Example

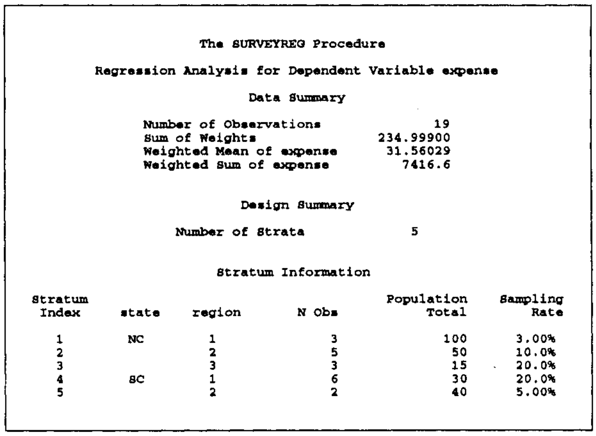
This example illustrates the use of PROC SURVEYREG with the household income survey data from the previous example. Suppose it is of interest to examine the relationship between total household income and total household living expenses in the survey population. The following linear function can be used to model the relationship:

$$\text{expense} = \alpha + \beta * \text{income} + \text{error}$$

The following statements fit this linear model using the household income survey data:

```
proc surveyreg  data=HHSample
      N=StrataTotals;
   model   expense = income;
   strata  state  region  /  list;
   weight  weight;
run;
```

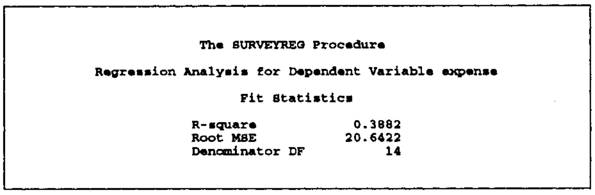
In the PROC statement, the option DATA=HHSample specifies that the input sample survey data set is HHSample, and the data set StrataTotals contains the population totals for the strata. The MODEL statement specifies the MODEL, with EXPENSE as the dependent variable and INCOME as the independent variable. The STRATA statement identifies the stratification variables as STATE and REGION. The LIST option requests a table of stratum-level information.



```
                     The SURVEYREG Procedure

        Regression Analysis for Dependent Variable expense

                          Data Summary

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


                         Design Summary

               Number of Strata               5


                      Stratum Information

Stratum                                       Population    Sampling
 Index     state    region          N Obs          Total        Rate

   1        NC        1                 3            100       3.00%
   2                  2                 5             50       10.0%
   3                  3                 3             15       20.0%
   4        SC        1                 6             30       20.0%
   5                  2                 2             40       5.00%
```

Figure 4. Summary Information

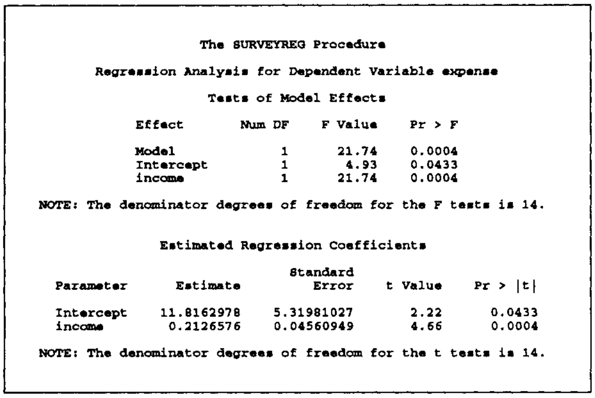
Figure 4 shows the data and design summary information produced by PROC SURVEYREG. The procedure calculates each stratum's sampling rate using the population total given in the input data set StrataTotals.



```
                    The SURVEYREG Procedure

       Regression Analysis for Dependent Variable expense

                        Fit Statistics

                  R-square              0.3882
                  Root MSE             20.6422
                  Denominator DF            14
```

Figure 5. Regression Fit Statistics

Figure 5 shows the regression fit statistics, including the R-square.



```
                    The SURVEYREG Procedure

       Regression Analysis for Dependent Variable expense

                    Tests of Model Effects

            Effect        Num DF     F Value     Pr > F

            Model              1       21.74     0.0004
            Intercept          1        4.93     0.0433
            income             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
  income          0.2126576       0.04560949        4.66       0.0004

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

Figure 6. Regression Tests and Estimates

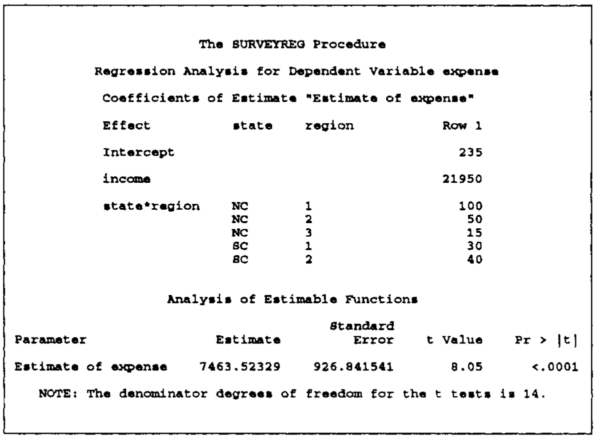
Figure 6 presents Wald's *F*-tests for the regression effects in the model. For the household income and expense study, both the INTERCEPT and the INCOME effects are significant at the 5% level. Figure 6 also gives the regression coefficient estimates with their standard errors and their *t*-tests.

Assume that the researchers obtain the amount of total income for all households in the survey population as $21,950 (in thousands). A regression estimate for the total basic living expenses in all households can then be obtained using the preceding linear model and an ESTIMATE statement. The following SURVEYREG statements compute the regression estimate:

```
proc surveyreg  data=HHSample
      N=StrataTotals;
   strata  state  region;
   class   state  region;
   model   expense = income state*region;
   weight  weight;
   estimate 'Estimate of expense'
      intercept 235  income 21950
      state*region 100 50 15 30 40 / e;
run;
```

To obtain a regression estimate with a stratified sample design, stratum should be used as a main effect in the model (Fuller et al. 1989, p. 99). The stratum effect is the STATE*REGION effect in the MODEL statement. The CLASS statement precedes the MODEL statement and lists the stratification variables as categorical.

The ESTIMATE statement defines a linear function of the regression parameters to produce the regression estimate. The coefficient for the INTERCEPT effect is 235, which is the total number of households in the survey population. The coefficients for the stratum effect are the total number of households in each stratum, which are 100, 50, 15, 30, and 40, respectively. The coefficient for the effect INCOME is 21950, the total income over all households in the survey population. To list the coefficients specified for the linear function, use the E option in the ESTIMATE statement.



```
                    The SURVEYREG Procedure

       Regression Analysis for Dependent Variable expense

        Coefficients of Estimate "Estimate of expense"

        Effect            state    region             Row 1

        Intercept                                       235

        income                                        21950

        state*region      NC       1                    100
                          NC       2                     50
                          NC       3                     15
                          SC       1                     30
                          SC       2                     40

                 Analysis of Estimable Functions

                                           Standard
Parameter                   Estimate          Error    t Value    Pr > |t|

Estimate of expense       7463.52329     926.841541       8.05      <.0001

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

Figure 7. Regression Estimate of Living Expenses

The ESTIMATE statement produces Figure 7. The table "Coefficients for ESTIMATE 'Estimate of expense'" lists the coefficients of the function specified in the ESTIMATE statement. The table "ESTIMATE Statement Results" presents the regression estimate of the total living expenses: $7,464 (in thousands) with an estimated standard error $927 (in thousands). This table also provides the *t*-test for the regression estimate.

References

Binder, D. and Roberts, G. (1999), "Design-based and Model-based Methods for Estimating Model Parameters", *Analysis of Survey Data*, eds. C. Skinner and R. Chambers, New York: John Wiley & Sons, Inc.

Chromy, J. R. (1979), "Sequential Sample Selection Methods," *Proceedings of the American Statistical Association, Survey Research Methods Section*, 401–406.

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.

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

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953), *Sample Survey Methods and Theory*, Volumes I and II, New York: John Wiley & Sons, Inc.

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

Kalton, G. (1983), *Introduction to Survey Sampling*, Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-035, Beverly Hills and London: Sage Publications, Inc.

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

Lee, E.S., Forthoffer, R.N., and Lorimor, R.J. (1989), *Analyzing Complex Survey Data*, Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-071, Beverly Hills, CA, and London: Sage Publications, Inc.

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

SAS Institute Inc. (1999), *SAS/STAT User's Guide, Version 8-1*, Cary, NC: SAS Institute Inc.

Williams, R.L. and Chromy, J.R. (1980), "SAS Sample Selection Macros," *Proceedings of the Fifth Annual SAS Users Group International Conference*, 5, 392–396.

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

Contact Information

Your comments and questions are valued and encouraged. Please contact the authors at

Anthony B. An, SAS Institute Inc., SAS Campus Drive, R5243, Cary, NC 27513. Phone (919)531-5879. FAX (919)677-4444. Email Anthony.An@sas.com

Donna L. Watts, SAS Institute Inc., Atlanta Plaza, Suite 3200, 950 E. Paces Ferry Road N.E., Atlanta, GA 30326. Phone (404)814-2560 ext 238. FAX (404)814-2556. Email Donna.Watts@sas.com

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

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