New SAS® Procedures for Analysis of Sample Survey Data

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

Abstract

Researchers use sample surveys to obtain information on a wide variety of issues. Many surveys are based on probability-based complex sample designs, including stratified selection, 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 7 of the SAS System, These new procedures are being added for the 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, 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 procedures will be available as experimental procedures in the initial release of Version 7 of the SAS System. Complete documentation describing the syntax and statistical methodology for these procedures will be available in a technical report at the time of the release. You can obtain more information on these procedures at the SAS Institute Research and Development web site: http://www.sas.com/rnd/

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. The section "Comparison among Procedures" discusses the differences in estimation methodology and results between traditional statistical procedures and the new survey procedures.

Income and Expenditure Survey

This paper uses the following example to illustrate the new procedures. The data in this example are artificially constructed solely for this purpose.

Suppose a marketing research firm has a household database containing information on 235 households in North Carolina and South Carolina, as shown in Table 1. The firm wants to obtain information about the potential economic impact of these households. Specifically, the firm wants to estimate the total income and the total basic living expenses of these households for the past year, where basic living expenses include items such as food, housing, transportation, and so forth. Also, the firm wants to investigate the relationship between total income and basic

living expenses among these households.

Table 1. Sampling Frame

Household ID	State	Region
1	NC	1
2	NC	1
⋮	⋮	⋮
101	NC	2
⋮	⋮	⋮
151	NC	3
⋮	⋮	⋮
166	SC	1
⋮	⋮	⋮
196	SC	2
⋮	⋮	⋮
235	SC	2

To accomplish these objectives, the firm selects a probability sample of households from its data base, or survey population. The sample design is a one-stage stratified design, with households as the sampling units. 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.

Table 2. Number of Households in Each Stratum

State	Region	Number of Households in Population	Sample
NC	1	100	3
	2	50	5
	3	15	3
SC	1	30	6
	2	40	2
Total		235	19

Table 2 lists the number of households in each stratum of the survey population and the number of sample households selected from each stratum. The total sample size is 19 households.

The SURVEYSELECT procedure, described in the section "Sample Selection," selects a probability sample of 19 households from the survey population. Table 1 shows the observations in the sampling frame. The households are identified only by numbers to protect confidentiality. The following SAS statements create the SAS data set FRAME, which will be input to PROC SURVEYSELECT.

```
data frame;
   input id state$ region;
   datalines;
1   NC     1
2   NC     1
...
;
```

The variable ID contains the household identification numbers. The variables STATE and REGION store the state and geographical region for each household.

Table 3. Sample of Income and Expenditure Survey

HH* ID	State	Region	Total Income ($**)	Living Expenses ($**)	Weight
11	NC	1	100	54	33.333
93	NC	1	83	25	33.333
36	NC	1	25	10	33.333
48	NC	2	120	83	10.000
123	NC	2	50	35	10.000
104	NC	2	110	65	10.000
131	NC	2	60	35	10.000
141	NC	2	45	20	10.000
157	NC	3	23	5	5.000
161	NC	3	10	8	5.000
152	NC	3	350	125	5.000
189	SC	1	130	20	5.000
173	SC	1	245	25	5.000
169	SC	1	150	33	5.000
182	SC	1	263	50	5.000
177	SC	1	320	47	5.000
194	SC	1	204	25	5.000
201	SC	2	80	11	20.000
228	SC	2	48	8	20.000

* HH=Household

** dollars are in thousands

Table 3 displays the income and expenditure sample, which contains the 19 sample households, the sampling weights, and the survey data collected from the sample households. A household's sampling weight is the reciprocal of its selection probability. PROC SURVEYSELECT provides the sampling weights, which are needed for the data analysis. The following statements create the data set HHSample for the sample in Table 3.

```
data HHSample;
   input ID state$ region income
         expense weight;
   datalines;
11  NC 1 100  54 33.333
93  NC 1  83  25 33.333
36  NC 1  25  10 33.333
...
228 SC 2  48   8 20.000
;
```

The variable ID contains the household identification numbers. The variable INCOME is the household's income for the past year, and the variable EXPENSE is the household's basic living expenses for the past year. The variable WEIGHT contains the sampling weight.

To provide stratum size information to the survey data analysis procedures, the following statements create the data set StrataTotals that contains the population stratum sizes, as shown in Table 2.

```
data StrataTotals;
   input state$ region _TOTAL_;
   datalines;
NC 1 100
NC 2  50
NC 3  15
SC 1  30
SC 2  40
;
```

The variable _TOTAL_ contains the total number of households in each stratum.

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. With probability sampling, each unit in the survey population has a known, positive probability of selection. This property of probability sampling avoids selection bias and enables you to use statistical theory to make valid inferences from the sample to the survey population.

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. PPS selection is often used in cluster sampling, where you select clusters (or groups of sampling units) of varying size in the first stage of selection. For example, clusters may be schools, hospitals, or geographical areas, and the final sampling units may be students, patients, or citizens. Cluster sampling can provide efficiencies in frame construction and other survey operations. For details on probability sampling methods, refer to Cochran (1977), Kalton (1983), and Chromy (1979).

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 procedure uses fast, efficient algorithms for these sample selection methods. Thus, it performs well even for very large input data sets or sampling frames, which may occur in practice for large-scale sample surveys.

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. It is widely used in practice towards meeting a variety of survey objectives. For example, with stratification you can ensure adequate sample sizes for subgroups of interest, including small subgroups, or you can use stratification towards improving the precision of the overall estimates. 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=($n_1, n_2, \ldots, n_s$) syntax in the PROC SURVEYSELECT statement, listing the stratum sample sizes $n_1, n_2, \ldots, n_s$, 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

For the example described in the section ‘‘Income and Expenditure Survey,’’ the sampling frame is the list of households in the research firm’s database saved in the SAS data set FRAME.

The following SURVEYSELECT statements select a probability sample of households from the data set FRAME. The METHOD=SRS option specifies that simple random sampling is to be used for sample selection. In simple random sampling, units are selected with equal probability and without replacement. The N = (3, 5, 3, 6, 2) option specifies the sample sizes for the strata — a sample of 3 households from the first stratum, 5 households from the second stratum, and so on. The OUT=SAMPLE option names the output data set that contains the selected sample. The STRATA statement identifies STATE and REGION as the stratification variables. The input data set FRAME is sorted by these stratification variables.

```
proc surveyselect data=frame out=sample
   method=srs n=(3, 5, 3, 6, 2);
   strata state region; run;
```

The SURVEYSELECT procedure selects a stratified random sample of households from the sampling frame using simple random sampling and the specified stratum sizes. PROC SURVEYSELECT produces the output data set SAMPLE, which contains the selected observations and their selection probabilities and sampling weights. The data set SAMPLE contains the sampling unit identification variable ID and the stratification variables STATE and REGION from the data set FRAME. The data set SAMPLE also contains the selection probabilities in the variable SelectionProb and the sampling weights in the variable Weight. In this example, a household's selection probability equals the number of selected households divided by the total number of households in its stratum. A household's sampling weight is the reciprocal of its selection probability. Table 3 in the section ''Income and Expenditure Survey'' shows the selected sample of households and the data collected on income and expenses.

Descriptive Statistics

The SURVEYMEANS procedure produces estimates of survey population totals and means, estimates of their variances, confidence limits, and other descriptive statistics. When computing these estimates, the procedure takes into account the sample design used to select the survey sample. The sample design can be a complex survey sample design with stratification, clustering, and unequal weighting. In addition to estimates for the entire survey population, the procedure can compute estimates for population subgroups.

Computational Method

The SURVEYMEANS procedure uses the Taylor expansion method for estimating 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 PSUs, in the sample design, the procedure estimates 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.

This variance estimation method assumes that first-stage sampling is with replacement, although often in practice it is not. This assumption may result in an overestimate of the variance, but this should be very small if the first-stage sampling fraction is small. Additionally, this 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 about any additional stages of sampling.

Capabilities

The SURVEYMEANS procedure can compute the following statistics:

- population total estimate and its standard deviation and corresponding *t*-test
- PSU-level mean estimate and its standard error and corresponding *t*-test
- 95% confidence limits for the population total and for the PSU-level mean estimates
- degrees of freedom for the variance estimation
- mean-per-element estimate and its standard error
- data summary information
- combined sampling fraction over strata and the total number of primary sampling units (PSUs)

Syntax

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

PROC SURVEYMEANS <*options*>
<*statistic-keywords*>;
CLASS *variables*;
VAR *variables*;
STRATA *variables* / <*options*>;
CLUSTER *variables*;
WEIGHT *variable*;
BY *variables*;

The PROC SURVEYMEANS statement invokes the procedure. You use this statement to name the input data set to be analyzed, specify sample design information, and request statistical computations. The CLASS statement identifies those numerical variables that are to be treated as categorical variables by the procedure. The VAR statement identifies the variables to be analyzed. 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 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 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 and population total, together with their variance estimates and confidence limits. The procedure can also compute PSU-level totals and means. You can 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 produces the information and statistics you request. If you do not specify statistics to compute, by default the procedure provides the estimate of the population total, its standard deviation, and its 95% confidence limits. You can save any printed output from the procedure to a SAS data set using the Output Delivery System.

Example

This example uses the survey data described in the section ‘‘Income and Expenditure Survey.’’ You can use PROC SURVEYMEANS to estimate the total income and total basic living expenses for the households in the survey population. The following statements invoke PROC SURVEYMEANS to compute these estimates and their standard deviations.

```
proc surveymeans data=HHSample N=StrataTotals
   sum df clm 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 CLM option requests confidence limits for the estimates, the DF option requests the associated degrees of freedom, and the FRACTION option requests the combined sampling rates over all strata.

The VAR statement specifies the two analysis variables, INCOME and EXPENSE. The STRATA statement identifies 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

Stratum Information

Stratum ID	state	region	Population Total	Sampling Rate	N Obs	Variable	N
1	FL	1	30	0.2	3	income	3
						expense	3
2		2	40	0.05	5	income	5
						expense	5
3	NC	1	100	0.03	3	income	3
						expense	3
4		2	50	0.1	6	income	6
						expense	6
5		3	15	0.2	2	income	2
						expense	2

Statistics

Variable	Sampling Fraction	DF	Sum	Std Dev	Lower 95% CL for Sum	Upper 95% CL for Sum
income	0.081	14	21818	2965.354	15458	28178
expense	0.081	14	7416.637	1489.068	4222.903	10610

Figure 1. Output from PROC SURVEYMEANS

Figure 1 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 with 14 degrees of freedom. The combined sampling fraction is 8.1% of the households in the survey population. Over all 235 households in the survey

population in North Carolina and South Carolina, 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 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.

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 for these tests 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*>;
STRATA *variables* / <*options*>;
CLUSTER *variables*;
CLASS *variables*;
MODEL *dependent* = <*effects*> / <*options*>;
WEIGHT *variable*;
ESTIMATE '*label*' *effect values* / <*options*>;
CONTRAST '*label*' *effect values* / <*options*>;
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 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 CLASS statement identifies those variables that are to be treated as categorical variables in the MODEL statement. The CLASS statement 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 a special notation using variable names and operators, as described in Chapter 24, ''The GLM Procedure,'' of the *SAS/STAT User's Guide, Version 6, Fourth Edition, Volumn 2.* You can use only one numerical variable as the dependent variable in the MODEL statement. The WEIGHT statement names the sampling weight variable. 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. You can use a BY statement to perform separate 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.

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.

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.

Output

PROC SURVEYREG presents the regression analysis results in several tables:

- summary information including the number of strata and observations in the analysis, the R-square for the regression, and the estimated population mean and total for the dependent variable
- stratum-level information (if you specify the LIST option in the STRATA statement), including the number of observations, the number of PSUs, the sampling rate, and the stratum collapsing information for each stratum
- a one-way analysis of variance for the dependent variable and Wald's *F*-test for all effects in the model
- estimates of regression coefficients, their standard errors, and *t*-tests
- estimates and corresponding tests for estimable linear functions of the regression parameters

You can save any printed output from the procedure to a SAS data set using the Output Delivery System.

Example

Consider the household income survey data in the section ''Income and Expenditure Survey.'' The firm wants to explore the relationship between the total income and the total basic living expenses of a household in the survey population. The researchers use the following linear function to model this relationship:

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

The firm predicts the total basic living expenses of all the households in the survey population by the household income.

The following statements fit this linear model using the household data in Table 3:

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

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

The SURVEYREG Procedure
Regression Analysis for Dependent Variable expense

Data Summary

Number of Strata	5
Number of Observations	19
R-square	0.388210
Root of MSE	20.642226
Denominator DF	14
Sum of Weights	234.999000
Weighted Mean of expense	31.560292
Weighted Sum of expense	7416.637000

Figure 2. Summary Information

Figure 2 summarizes the regression and the data set information.

The SURVEYREG Procedure

Regression Analysis for Dependent Variable expense

Stratum Information

Stratum ID	state	region	Number of Observations	Population Total	Sampling Rate
1	FL	1	6	30	0.2
2		2	2	40	0.05
3	NC	1	3	100	0.03
4		2	5	50	0.1
5		3	3	15	0.2

Figure 3. Stratum-level Information

Figure 3 shows the stratum-level information about the survey design. 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

                 Testing Effects in the Model

            Source       Num DF    F Value    Pr > F

            Intercept         1       4.93    0.0433

            income            1      21.74    0.0004

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

       Regression Analysis for Dependent Variable expense

               Estimated Regression Coefficients

                                 Standard
Parameter        Estimate           Error    t Value    Pr > |t|

Intercept       11.816298        5.319810       2.22      0.0433
income           0.212658        0.045609       4.66      0.0004

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

Figure 4. Regression Analysis for Living Expenses

Figure 4 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 4 also gives the regression coefficient estimates with their standard errors and their *t*-tests.

Assume that the firm obtains the amount of total income over all households in the survey population as $21,950 (in thousands). Researchers can obtain a regression estimate for the total basic living expenses in all households using the preceding linear model and an ESTIMATE statement. The following statements illustrate the computation of 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, you need to use stratum as a main effect in the model (Statistical Laboratory 1989, p. 99). The stratum effect is the STATE*REGION effect in the MODEL statement. Therefore, the CLASS statement must list stratification variables before the MODEL statement. An ESTIMATE statement labeled ''Estimate of expense'' defines a linear function of the regression parameters in the model to produce the regression estimate for the total living expenses. From Table 2, the coefficient for the INTERCEPT effect is 235, 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 get a coefficient list for the linear function specified, you use the E option in the ESTIMATE statement.

```
                    The SURVEYREG Procedure

      Regression Analysis for Dependent Variable expense

       Coefficients for 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

        Regression Analysis for Dependent Variable expense

                    ESTIMATE Statement Results

                                        Standard
ESTIMATE Label           Estimate          Error  t Value  Pr > |t|

Estimate of expense   7463.523285     926.841541     8.05    <.0001

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

Figure 5. Regression Estimate of Living Expenses

The procedure produces Figure 5 for the ESTIMATE statement. The table ''Coefficients for ESTIMATE 'Estimate of expense'' lists the coefficients of the estimable 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.

Comparison among Procedures

The SURVEYMEANS and SURVEYREG procedures analyze sample survey data taking into account the sample design. Therefore, the computation methods are different from those of traditional statistical procedures. This section compares the different estimates that these procedures compute for total basic living expenses using the ''Income and Expenditure Survey'' data.

Table 4 lists six estimators used by these procedures based on different sample design assumptions. The notation $T(\textit{procedure}|\textit{sample design})$ indicates the procedure and the sample design used to compute the total estimate.

Table 4. Estimators of Total Living Expenses

Estimator	Procedure	Sample Design
$T(SM\|STR)$	SURVEYMEANS	Stratified
$T(SM\|SRS)$	SURVEYMEANS	Simple Random
$T(M\|SRS)$	MEANS	Simple Random
$T(M\|WT)$	MEANS	Simple Random Unequal Weights
$T(SR\|STR)$	SURVEYREG	Stratified
$T(GLM\|SRS)$	GLM	Simple Random

$T(SM|STR)$ is a total estimator produced by PROC SURVEYMEANS. The variance estimator of $T(SM|STR)$ uses stratification (STR) from the sample design.

$T(SM|SRS)$ is also a total estimator calculated by PROC SURVEYMEANS, but $T(SM|SRS)$ and its variance estimator assume a simple random sampling design (SRS). $T(SM|SRS)$ ignores stratification and assumes equal probabilities of selection for sampling units.

$T(M|SRS)$ is a total estimator produced by PROC MEANS. $T(SM|SRS)$ ignores stratification and assumes that data are collected from a simple random sampling design. Since PROC MEANS does not provide the population total estimator directly, you multiply the estimated mean by the population size to obtain the estimator of the total.

$T(M|WT)$ is a total estimator similar to $T(M|SRS)$ produced by PROC MEANS. However, $T(M|WT)$ uses the same sampling weights as in $T(SM|STR)$. The variance estimation of $T(M|WT)$ also ignores stratification and assumes that data are collected from a simple random sampling design. To obtain $T(M|WT)$ from the PROC MEANS output, you multiply the weighted mean by the population size.

$T(SR|STR)$ is a regression estimator computed by PROC SURVEYREG. $T(SR|STR)$ takes into account stratification information and the auxiliary information from the independent variables.

$T(GLM|SRS)$ is a regression estimator generated by PROC GLM. $T(GLM|SRS)$ assumes a simple random sampling design ignoring the stratification. $T(GLM|SRS)$ uses auxiliary information from the independent variables to produce the regression estimator.

For a simple random sample design, each sampling unit is selected with equal probability without replacement. Therefore, sampling weights for each unit are the same, and they are equal to the reciprocals of the sampling rates. A data set SRSSample contains these new weights created from the data set HHSample.

```
data SRSSample;
   set HHSample;
   weight=235/19;
```

The following statements compute the different estimates for total living expenses.

```
title 'T(SM|STR)';
proc surveymeans data=HHSample N=StrataTotals sum;
   strata state region;
   var    expense;
   weight weight;

title 'T(M|WT)';
proc means data=HHSample mean stderr;
   var    expense;
   weight weight;

title 'T(SM|SRS)';
proc surveymeans data=SRSSample N=235 sum;
   var    expense;
   weight weight;

title 'T(M|SRS)';
proc means data=SRSSample mean stderr;
   var    expense;

title 'T(SR|STR)';
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;

title 'T(GLM|SRS)';
proc glm data=HHSample;
   class    state region;
   model    expense = income state*region;
   estimate 'Estimate of expense'
            INTERCEPT 235 income 21950
            state*region 100 50 15 30 40;
run;
```

Table 5 displays the six estimates from Table 4 and their standard deviations.

Table 5. Comparison among Expense Estimators

Estimator	Estimate($*)	Standard Deviation ($*)
$T(SM\|STR)$	7,417	1,489
$T(M\|WT)$	7,417	1,383
$T(SM\|SRS)$	8,460	1,557
$T(M\|SRS)$	8,460	1,624
$T(SR\|STR)$	7,464	927
$T(GLM\|SRS)$	7,459	1,004

* dollars are in thousands

PROC SURVEYMEANS produces different estimates, $T(SM|STR) = \$7,417K$ and $T(SM|SRS) = \$8,460K$. The standard deviation of $T(SM|SRS)$ is \$1,557K, which is slightly bigger than the standard deviation of $T(SM|STR)$. Stratification improves the estimation precision of the variance estimation.

PROC MEANS also produces different estimates, $T(M|SRS) = \$8,460K$ and $T(M|WT) = \$7,417K$. And $T(M|WT)$ has a smaller variance estimate than $T(M|SRS)$.

When assuming a simple random sampling design, PROC SURVEYMEANS and PROC MEANS produce the same point estimates $T(SM|SRS)$ and $T(M|SRS)$, but with different variance estimates. The estimated standard deviation of $T(SM|SRS)$ is \$1,557K, which is about 95.9% of the estimated standard deviation of $T(M|SRS)$, \$1,624K. This is because PROC SURVEYMEANS takes into account the finite population correction in the variance estimation. $T(SM|SRS)$ and $T(M|SRS)$ summarize the sample and apply only to a theoretical population with the same compsition as the sample. When the sample is drawn from a complex survey, these estimators produce biased estimates. In this example, $T(SM|SRS)$ and $T(M|SRS)$ are used only for the comparison.

Both PROC MEANS and PROC SURVEYMEANS use the same set of sampling weights to calculate $T(SM|STR)$ and $T(M|WT)$. Although $T(SM|STR)$ and $T(M|WT)$ have the same point estimates, they have different variance estimates. Because PROC SURVEYMEANS uses Taylor's approximation theory and the stratification in the variance estimation, the estimated standard deviation of $T(SM|STR)$ is different from the calculation by PROC MEANS. In this example, the estimated standard deviation of $T(SM|STR)$, which is \$1,489K, is slightly larger than the estimated standard deviation of $T(M|WT)$. Both $T(SM|STR)$ and $T(M|WT)$ are estimators applied to the survey population.

PROC SURVEYREG and PROC GLM produce different regression estimates, $T(SR|STR)$ and $T(GLM|SRS)$. Using stratification, $T(SR|STR)$ has a slightly smaller estimated standard deviation, \$927K, than the estimated standard deviation of $T(GLM|SRS)$.

In comparison to the estimator $T(SM|STR)$, the regression estimator $T(SR|STR)$ improves the estimation precision by reducing the standard deviation estimate from \$1,489K to \$927K, a reduction of 37.7%.

The lesson from this simple example is that the analysis of data from a complex survey should use the sample design information in order to produce statistically valid inferences.

Acknowledgments

We are grateful to Robert N. Rodriguez, Maura E. Stokes, and Donna M. Sawyer of the Applications Division at SAS Institute for their valuable assistance in the preparation of this manuscript.

References

Chromy, J. R. (1979), "Sequential Sample Selection Methods," *American Statistical Association 1979 Proceedings of the 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.

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

SAS Institute Inc. (1989), *SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 2*, Cary, NC: SAS Institute 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.

Authors

Anthony B. An, SAS Institute Inc., SAS Campus Drive, R5243, Cary, NC 27513. Phone (919)677-8000 ext 5879. FAX (919)677-4444. Email sasaba@unx.sas.com

Donna L. Watts, SAS Institute Inc., Atlanta Plaza, Suite 3390, 950 E. Paces Ferry Rd N.E., Atlanta, GA 30326. Phone (404)814-2560 ext 238. FAX (919)814-2556. Email sasdlw@unx.sas.com