Paper 348-2012

Simplifying the Analysis of Complex Survey Data Using the SAS® Survey Analysis Procedures

Varma Nadimpalli, Katie Hubbell, Westat, Rockville, Maryland

ABSTRACT

Large sample-based surveys often have complex sample designs, with design features including stratification, clustering, multi-stage sampling, and unequal probability of selection of observations. The calculation of the associated sampling weights often involves nonresponse adjustments and raking to external control totals. The analysis usually includes descriptive statistics such as frequencies, means, totals and their standard errors. Using standard statistical software modules such as PROC SUMMARY, PROC FREQ and PROC MEANS to analyze such data results in underestimation of variance, as these routines assume that the data is from a simple random sample and do not take into account the complex nature of the sample. Now, however, the survey analysis procedures such as PROC SURVEYMEANS and PROC SURVEYFREQ that have been added to SAS/STAT® software can compute variances that accurately reflect complex sample design and estimation procedures. This paper compares the complexity of the variance estimation code used in earlier projects with the simplicity of the code that is possible using the survey analysis procedures.

KEYWORDS: Complex survey, Survey procedures, SURVEYFREQ, SURVEYMEANS, SURVEYREG

INTRODUCTION

Household surveys and other large sample-based surveys utilize complex sample designs to collect data to control survey costs. The most common components of complex survey designs are multi-frame sampling, multiple response modes, multiple stages of sampling, stratification, clustering, unequal weights or sampling rates, and nonresponse adjustment or raking. All these design components have been shown to increase efficiency in survey after survey. But these standard survey techniques can complicate analyses due to selection with varying probabilities and non-independent selections. Investigations have shown that ignoring the complex sample design would lead to bias estimates and misleading estimates of standard errors. So, the complex design must be taken into account in the survey analysis, weights need to be used in analyzing the survey data, and the variances of survey estimates need to be computed in a manner that reflects the design. Incorporating the design features into analysis requires choosing a special method of variance estimation. A class of techniques called replication methods provides one way to estimate variances for a complex sample design.

VARIANCE ESTIMATION IN A COMPLEX SURVEY

The most common variance estimation methods are balanced repeated replication (BRR), Fay's modified balanced repeated replication, Jackknife repeated replication (JK), and the Taylor series approximation. Studies have shown that none of these methods perform consistently better or worse, and the choice may depend in most cases on the design and on the relative costs of computing and the availability of resources.

SAS/STAT® software now provides specialized SAS procedures to analyze complex survey data with unequal weights: PROC SURVEYMEANS, PROC SURVEYFREQ, PROC SURVEYREG, and PROC SURVEYLOGISTIC can now be used to compute means, frequencies, regression, and logistic analysis, respectively. When these routines were first introduced in SAS 7, they used only Taylor linearization for estimating variances and were limited in their capabilities. Because there are many large surveys that use BRR or Jackknife for estimating variance many analysts, continued to use other specialized survey software programs for estimating variance, or used SAS routines like PROC SUMMARY, PROC FREQ and PROC MEANS with DATA step programming to calculate the variance and standard errors.

The SURVEYMEANS, SURVEYFREQ, SURVEYREG, and SURVEYLOGISTIC procedures now provide a choice of variance estimation methods, including the BRR, Fay's modified balanced repeated replication,

Jackknife repeated replication (JK), and the Taylor series approximation. They also use ODS graphics to create graphs as a part of the output.

This paper revisits the complexity of the variance estimation code used in earlier projects with the simplicity of the code that is possible using the survey analysis procedures. We then compare our estimates to those produced by WesVar. (WesVar is a software program developed by Westat for computing estimates and variances of complex survey data*).* Particularly when analyzing data from multiple time points, special programming efforts were needed to compute the variances and standard errors of the difference or change. Now we are able to take advantage of the simple statements SAS has provided to compute the variance, standard errors, and confidence intervals of the difference/change between time points.

SAS PROC FREQ/MEANS vs. PROC SURVEYFREQ/SURVEYMEANS

The FREQ and MEANS procedures have been part of SAS for over 30 years and are probably two of the most used SAS procedures for data exploration and statistical analysis for categorical and continuous variables. The FREQ procedure produces one-way to n-way frequency and contingency tables for categorical data while PROC MEANS produces means, medians, standard errors and other descriptive statistics for the continuous data. These procedures assume that the underlying sample is a simple random sample of an infinite population and the all units were selected with equal probability. Using a weight statement, we can adjust for the differences between the sample and the targeted population. Both these procedures produce an unbiased estimate, but the standard errors will necessarily be incorrect when a complex sample design is involved, as they do not have any way to reflect the nature of the design..

The SURVEYFREQ procedure, like PROC FREQ, produces one-way to n-way frequency and cross tabulation tables from sample survey data for categorical variables, but also computes variance estimates based on the sample design used to obtain the survey data. Confidence limits, coefficients of variation, and design effects are also available. The procedure provides a variety of options to customize the table display. The design can be a complex multistage survey design with stratification, clustering, and unequal weighting. The table request syntax for PROC SURVEYFREQ is very similar to the table request syntax for PROC FREQ. As in PROC FREQ, you can request more than one table in the same TABLES statement, and you can use multiple TABLES statements in the same invocation of the procedure.

The SURVEYMEANS procedure, like the MEANS procedure, produces means, medians, standard errors and other descriptive statistics for continuous variables, but also estimates variances and confidence limits and performs t tests for these statistics based on the sample design.

PROC SURVEYFREQ and PROC SURVEYMEANS now provide a choice of variance estimation methods, which include Taylor series linearization, balanced repeated replication (BRR), Fay's modified balanced repeated replication and the jackknife, by using the **PROC** statement options **VARMETHOD = TAYLOR**, **VARMETHOD=BRR**, **VARMETHOD=BRR (fay=c),** where c is correction factor, and **VARMETHOD=JACKKNIFE** respectively. The WEIGHT statement names the sampling weight variable. The REPWEIGHTS statement names replicate weight variables for BRR or jackknife variance estimation. You can use a BY statement with PROC SURVEYFREQ or SURVEYMEANS to obtain separate analyses for groups defined by the BY variables. All statements can appear multiple times, except for the PROC SURVEYFREQ/SURVEYMEANS statement and the WEIGHT statement, each of which can appear only once.

REVISITING EARLIER VARIANCE ESTIMATION CODE

For revisiting our earlier analysis, we are using datasets that have multiple years of data stacked from a complex survey with 62 replicate weights calculated using jackknife repeated replication method. We are using the new SURVEYFREQ, SURVEYMEANS and SURVEYREG procedures and comparing them to a complex set of code using PROC SUMMARY, PROC FREQ, PROC MEANS, and DATA step programming. The DATA step programming will be different for different procedures. "Original code" refers to the actual code that was used to obtain the estimate and standard deviations using PROC SUMMARY / FREQ / MEANS. The original code is provided on the left hand side while the new code using the survey analysis procedures is on the right.

We compared the results from these two methods with those produced by using the WesVar software. We are using a **BY** statement to perform the analysis by year. **WEIGHT** is the full sample weight and **REPL1** through

REPL62 are replicate weights created using the jackknife replication method. The **JKNFACT** dataset has a set of 62 JKN factors. First we show the steps to analyze categorical variables, then steps to analyze continuous variables. We include the code to obtain an overall estimate by years, then the code to obtain the standard deviations of change for the overall estimate. Finally, we show the analysis by certain categories or parameters. (We are not providing the code for the change on these as it is more complex.)

Analysis of Categorical Variables: Overall Estimate

Here we used PROC SUMMARY for obtaining the estimate and standard deviations for categorical variables. PROC FREQ can also be used, but the data step programming is different. So we used PROC FREQ to obtain the estimate and standard deviation of the change.

Original Code

Compute the weighted summary on the variable of interest by year using PROC SUMMARY and output the dataset:

```
proc summary data = paper;
 class Var1;
 by year;
 var weight repl1-repl62;
 output out=actual sum=;
run;
```

Compute the denominator for the full sample weight and the replicate weights using the output from the previous step:

```
%macro denom;
data denom(drop=weight repl1-repl62 Var1 _type_
_freq_);
  set actual;
  if _type_ = 0;
  den  a=weight;
  %do _i_=1 %to 62;
    dena&_i_=repl&_i_;
  %end;
run;
%mend denom;
%denom
```

Compute the numerator for the sample weight and the replicate weights using _type_ and the categorical variable value 1. We assumed the value of the variable to be 0 and 1. So we subset to 1:

```
%macro numer;
data numer (drop= weight repl1-repl62 Var1 _type_
                  _freq_);
 set actual;
 if (_type_=1 and Var1=1);
 num_a=weight;
 %do _i_=1 %to 62;
   numa&_i_=repl&_i_;
%end;
run;
%mend numer;
```

New PROC SURVEYFREQ procedure

Compared to the original code on the left, a simple PROC SURVEYFREQ can now be used to produce the estimate and standard deviations of the categorical variables. We used the 'NOSUMMARY' option in the PROC statement to suppress the 'Data Summary' table, since we didn't want to show all of that output.

```
ods output CrossTabs=CrossTabs1;
proc surveyfreq data = paper nosummary
 varmethod = jk;
 tables year*var1;
 weight weight;
 repweight repl1-repl62 /jkcoefs= jknfact;
run;
```

%***numer***
Merge the two datasets and the jackknife factors to calculate the estimate, variance and standard deviation.

```
%macro calc;
data calc1 (keep=year var class est sd);
  merge denom numer jknfact1;
   by year;
  Esta= (num_a/den_a);
  vara=0;
  %do _i_=1 %to 62;
   vara+jkn&_i_*(numa&_i_ /
              dena&_i_-esta)**2;
  %end;
  Est = esta*100;
  SD = sqrt(vara)*100;
run;
%mend calc;
%calc
```

Difference between two time points

This section shows the code to compute the estimate and standard deviation of the change between two years/time points. Here we are providing just 2 years, but multiple time points or years could also be used

Original Code

We are using PROC FREQ to compute the difference in estimate and standard error between year 1 and year 2. PROC SUMMARY can also be used.

Compute the estimate using the full sample weight

```
proc freq data = paper;
  tables Var1/out=full;
  by year;
  weight weight;
 run;

data full(keep=year percent);
 set full;
 if Var1 = 1;
run;
```

Compute the replicate the estimates using the 62 replicate weights

```
%macro temp;
%do _i_=1 %to 62;

 proc freq data= paper;
  tables var1 / out=repl&_i_;
```

New SURVEYFREQ Procedure

*The new option '**RISK**' in PROC SURVEYFREQ produces the estimate, standard error and confidence interval of the change. This is a great development in the survey procedures as computation of the standard errors of change gets messier if more and more variables are involved. Without modifying the dataset or modifying the original code, in one single PROC SURVEYFREQ procedure we can now get the estimate and standard errors of the time points and the change*

```
ods output CrossTabs=CrossTabs1 Risk2=Diff1;
 proc surveyfreq data = paper nosummary
                varmethod = jk;
  tables year*var1/risk;
  weight weight;
  repweight repl1-repl62 / jkcoefs = jknfact;
run;
```

```
 weight repl&_i_;
 by year;
run;

 data repl&_i_(keep=year repno
              percent);
 set repl&_i_;
 if Var1 = 1;
 Repno = &_i_;
run;

 proc append base=outstat out= repl&_i_;
 run;

%end;
%mend temp;
%temp

proc sort data = outstat;
 by year;
run;

data temp;
 merge outstat full(rename=(percent=Est));
  by year;
run;

proc sort data = temp;
 by repno year;
run;
```

Since we need to compute the covariance between year 1 and year 2, subset the data sets by years.

```
data Yr1(rename = (Est = EstYr1
 Percent = RepEstYr1) drop=year)
     Yr2(rename = (Est = EstYr2
 Percent = RepEstYr2) drop=year);
 set temp;
  by repno;
 if Year = 1 then output Yr1;
 if Year = 2 then output Yr2;
run;
```

Merge both the datasets created above and compute the variance and covariance.

```
data temp1;
 merge Yr1 Yr2 jknfact;
  by repno;
   r1 = jkcoefficient*(repestyr1-
                    estyr1)**2;
   r2 = jkcoefficient*(repestyr2-
                   estyr2)**2;
   cv = jkcoefficient*(repestyr1-
        estyr1)*(repestyr2-estyr2);
   dr = r1 + r2 - 2*cv;
run;
```

Summarize the variance and covariance variables
computed above.

```
proc summary data = temp1;
 var r1 r2 dr;
 output out=temp2 sum=;
run;

data temp2;
 set temp2;
 SD1 = sqrt(r1);
 SD2 = sqrt(r2);
 SDD = sqrt(dr);
run;

proc print data = temp2;
run;
```

Output from Original Method

The standard error for the two time points and the change were bolded and colored.

Obs	SD1	SD2	SDD
1	***1.27061***	***1.10335***	***1.00603***

.

Output from Proc Surveyfreq for the difference using the 'RISK' statement

The estimate and standard errors for individual year and the change were bolded and colored.

The SURVEYFREQ Procedure

Variance Estimation

Method	Jackknife
Replicate Weights	PAPER
Number of Replicates	62

Table of YEAR by Var1

YEAR	Var1	Weighted Frequency	Std Dev of Frequency	Wgt Freq	Percent	Std Err of Percent
XXXX	1	124148	15001656	1031425	43.6451	1.2858
	2	23123	3053361	295766	8.8833	0.7319
	Total	147271	18055018	1193943	52.5284	1.4113
YYYY	1	108275	13721067	856849	39.9194	1.2956
	2	18877	2595811	175252	7.5521	0.5711
	Total	127152	16316878	907741	47.4716	1.4113
Total	1	232423	28722723	1710638	83.5646	1.0842
	2	42000	5649172	408563	16.4354	1.0842
	Total	274423	34371895	1873812	100.000	

```
                    Column 1 Risk Estimates

                                  Standard
                Risk              Error    95% Confidence Limits

Row 1           0.8309            0.0127   0.8055      0.8563
Row 2           0.8409            0.0110   0.8189      0.8630
Total           0.8356            0.0108   0.8140      0.8573

Difference     -0.0100            0.0101   -0.0301     0.0101

                Difference is (Row 1 - Row 2)

                    Sample Size = 274423
                    Column 2 Risk Estimates

                                  Standard
                Risk              Error    95% Confidence Limits

Row 1           0.1691            0.0127   0.1437      0.1945
Row 2           0.1591            0.0110   0.1370      0.1811
Total           0.1644            0.0108   0.1427      0.1860

Difference      0.0100            0.0101   -0.0101     0.0301

                Difference is (Row 1 - Row 2)
                    Sample Size = 274423
```

The numbers from the original code and PROC SURVEYFREQ are then compared with numbers obtained using Wesvar. Here we are providing the standard deviations of the change only. The numbers were multiplied by 100 to express in percentage points.

	WesVar			PROC SURVEYFREQ			Original Method		
	Year1	Year2	Change	Year1	Year2	Change	Year1	Year2	Change
	SE	SE	SE	SE	SE	SE	SE	SE	SE
Overall	1.2706%	1.1034%	1.0060%	1.2706	1.1034	1.0060	1.2706	1.1034	1.0060

Estimation by parameters

Many times the analysis involves subcategories of variables and we are to compute the estimate and standard deviation on them. The following code demonstrates how the estimate and standard error were computed using PROC SUMMARY, and how easily it can be computed using PROC SURVEYFREQ. Since the calculation of the change is messy using the original code, the calculation of change is not provided here. But for PROC SURVEYFREQ all that is required is a simple change in the TABLES statement. Use the RISK statement in the tables. In this case avoid using the ***BY*** *statement and use the variable in the* ***TABLES*** *statement.*

Original Code using PROC SUMMARY

```
%macro Paper(Cat=);
 proc sort data = paper;
  by year &cat;
 run;

 proc summary data = paper;
  class Var1;
```

PROC SURVEYFREQ

```
 ods output CrossTabs= CrossTabs;
proc surveyfreq data = paper varmethod = jk;
  tables (Cat1 Cat2 Cat3 Cat4 Cat5)*Var1/row;
  weight weight;
  by year;
  repweight repl1-repl62/jkcoefs=jknfact;
run;
```

```
 by year &cat;
 var weight repl1-repl62;
 output out=actual sum=;
run;

%macro denom;
 data denom(drop=weight repl1-
  repl62 Var1 _type_ _freq_);
   set actual;
   by year &cat;
  if _type_ = 0;
  den_a=weight;
  %do _i_=1 %to 62;
   dena&_i_=repl&_i_;
  %end;
 run;
%mend denom;
%denom

%macro numer;
data numer (drop= weight repl1-
  repl62 Var1 _type_ _freq_);
 set actual;
  by year &cat;
 if (_type_=1 and Var1=1);
 num_a=weight;
 %do _i_=1 %to 62;
   numa&_i_=repl&_i_;
 %end;
run;
%mend numer;
%numer

data numer;
 merge numer jknfact1;
  by year;
run;

%macro calc;
 data &cat (keep=year var class
           est sd);
  merge denom numer;
   by year &cat;
  length Var Class $8.;
  Esta= (num_a/den_a);
  vara=0;
  %do _i_=1 %to 62;
   Vara + jkn&_i_ * (numa&_i_ /
   dena&_i_-esta)**2;
  %end;
  Est = esta*100;
  SD = sqrt(vara)*100;
 run;

%mend calc;
%calc
```

```
%mend Paper;
%Paper(cat=Cat1)
%Paper(cat=Cat2)
%Paper(cat=Cat3)
%Paper(cat=Cat4)
```

Comparing Results:

Here we are presenting the results using the original method, PROC SURVEYFREQ, and results obtained using Wesvar. Here too the numbers were multiplied with 100 to express them in percentage points.

	WesVar				PROC SURVEYFREQ				PROC SUMMARY Macro			
	Year1		Year2		Year1		Year2		Year1		Year2	
	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE
Overall	83.0886%	1.2706%	84.0913%	1.1034%	83.0886	1.2706	84.0913	1.1034	83.0886	1.2706	84.0913	1.1034
Cat1												
Subcat1	87.5448%	1.6769%	88.2940%	1.2381%	87.5448	1.6769	88.2940	1.2381	87.5448	1.6769	88.2940	1.2381
Subcat2	75.4084%	1.9779%	77.3910%	2.0869%	75.4084	1.9779	77.3910	2.0869	75.4084	1.9779	77.3910	2.0869
Cat2												
Subcat1	90.0424%	1.1090%	89.4061%	0.9800%	90.0424	1.109	89.4061	0.9800	90.0424	1.109	89.4061	0.9800
Subcat2	80.0064%	1.3423%	81.4020%	1.2913%	80.0064	1.3423	81.4020	1.2913	80.0064	1.3423	81.4020	1.2913
Cat3												
Subcat1	86.5188%	1.4555%	87.5912%	1.1517%	86.5188	1.4555	87.5912	1.1517	86.5188	1.4555	87.5912	1.1517
Subcat2	82.9689%	1.4362%	83.4429%	1.2579%	82.9689	1.4362	83.4429	1.2579	82.9689	1.4362	83.4429	1.2579
Subcat3	78.8602%	1.9508%	78.1985%	2.4570%	78.8602	1.9508	78.1985	2.4570	78.8602	1.9508	78.1985	2.4570
Cat4												
Subcat1	96.7663%	0.9588%	91.5817%	-	96.7663	0.9588	91.5817	-	96.7663	0.9588	91.5817	-
Subcat2	85.2669%	4.9129%	82.8421%	5.3110%	85.2669	4.9129	82.8421	5.3110	85.2669	4.9129	82.8421	5.3110
Subcat3	82.9759%	1.2872%	84.1028%	1.0910%	82.9759	1.2872	84.1028	1.0910	82.9759	1.2872	84.1028	1.0910
Cat5												
Subcat1	80.9780%	3.9484%	82.8349%	1.7596%	80.9780	3.9484	82.8349	1.7596	80.9780	3.9484	82.8349	1.7596
Subcat2	80.2366%	9.4231%	78.3089%	4.7095%	80.2366	9.4231	78.3089	4.7095	80.2366	9.4231	78.3089	4.7095
Subcat3	83.2934%	1.1764%	84.3353%	1.1634%	83.2934	1.1764	84.3353	1.1634	83.2934	1.1764	84.3353	1.1634

Analysis of Continuous Variables:
Overall Estimate

Original Code

Compute the weighted values on the variables of interest by multiplying the variable with full sample weight. Repeat the same with 62 replicate weights:

```
%macro cont;
data temp;
 set paper;
 if Cont1 ^= . then ContWt0 = Cont1*weight;
 %do _i_ = 1 %to 62;
  ContWt&_i_ = cont1*repl&_i_;
 %end;
run;
%mend cont;
%cont
```

PROC SURVEYMEANS

In contrast, using a single PROC SURVEYMEANS we can get the means and standard deviations for the continuous variables:

```
proc surveymeans data =paper
            varmethod = jackknife;
var cont1;
 weight Weight;
  class year;
  repweights Repl1-Repl62 / jkcoefs = jknfact;
 run;
```

Summarize the full sample weight, 62 replicate weights and the newly computed weighted variables using a PROC SUMMARY or PROC MEANS procedure:

```
proc summary data=temp;
 var Weight repl1-repl62 Contwt0 contwt1-contwt62;
 by year;
 output out=counter sum=;
run;

%macro calc;
data Est;
 merge counter jknfact;
  by year;
 Est = contwt0 / weight;
 Vara = 0;
 %do _i_ = 1 %to 62;
  vara+jkn&_i_*((contwt&_i_/repl&_i_)-est)**2;
 %end;
run;
%mend calc;
%calc
```

Computing Difference for Continuous Variables

Original Code

Here we are using PROC SUMMARY to compute the estimate and standard errors of the change. First compute the weighted variable by multiplying the continuous variable with weight for each time point:

```
%macro cfp;
data Yr1;
 set paper;
 if year = 1;
 cfpwt = cont1*weight;
 %do _i_ = 1 %to 62;
  cfpwt&_i_ = volume*repl&_i_;
 %end;
run;
%mend cfp;
%cfp
```

Summarize the actual weights and computed weighted variables for year 1:

```
proc summary data=Yr1;
  var weight repl1-repl62
      cfpwt cfpwt1-cfpwt62;
  output out=counter sum=;
run;
```

PROC SURVEYREG

PROC SURVEYREG can be used to compute the standard deviation on continuous variables. PROC SURVEYMEANS can't be used to compute the estimate and standard errors of the change. The new LSMEANS statement with a DIFF can be used and are relatively new in SAS 9.3.

```
ods graphics on;
proc surveyreg data=paper varmethod = jk;
  class year;
  model cont1 = year;
  lsmeans year / diff plots=(diff meanplot(cl));
  weight Weight;
  repweight repl1-repl62/jkcoefs= jknfact;
run;
ods graphics off;
```

Repeat the above two steps for year 2:

```
%macro wt;
data Yr2;
 set paper;
 if year = 2;
 recdwt = weight*cont1;
 Wt2    = weight;
 %do _i_ = 1 %to 62;
  recwt&_i_ = volume*repl&_i_;
  repwt&_i_ = repl&_i_;
 %end;
run;
%mend wt;
%wt;

proc summary data=Yr2;
  var wt2 repwt1-repwt62
      recdwt recwt1-recwt62;
  output out=actual sum=;
run;
```

Merge the dataset and compute the ratios:

```
%macro both;
data both(drop= weight repl1-repl62
               cfpwt cfpwt1-cfpwt62
               wt2 repwt1-repwt62
               recdwt recwt1-recwt62);
 merge counter actual;
  by _type_;
 Ratio1 = cfpwt  / weight;
 Ratio2 = recdwt / wt2;
 %do _i_ = 1 %to 62;
  Ra1&_i_ = cfpwt&_i_ / repl&_i_;
  Ra2&_i_ = recwt&_i_ / repwt&_i_;
 %end;
run;
%mend both;
%both;
```

Merge the ratio computed dataset and jkn factors and compute the variance and standard deviations:

```
%macro calc;
data calc;
 merge both jknfact;
 Est1= ratio1;
 Est2= ratio2;
 covar=0; var1=0; var2=0;
 %do _i_=1 %to 62;
   covar+jkn&_i_*(ra1&_i_-est1)*
              (ra2&_i_-est2);
   var2+jkn&_i_*(ra2&_i_-est2)**2;
   var1+jkn&_i_*(ra1&_i_-est1)**2;
 %end;
 vardiff=var1+var2-2*covar;
 sddiff = sqrt(vardiff);
 sd08 = sqrt(var1);
```

```
 sd09 = sqrt(var2);
 diff=est1-est2;
run;
%mend calc;
%calc

proc print data = calc;
 var Est1 SD2 Est2 SD2 Diff sddiff;
run;
```

OUTPUT using Original Code:

The estimate and standard error of individual years and change we bolded and colored.

Obs	Est1	sd1	Est2	sd2	diff	sddiff
1	**9.24612**	**0.50959**	**8.50929**	**0.56863**	**0.73683**	**0.43624**

Output from PROC SURVEYREG

The SURVEYREG Procedure

Regression Analysis for Dependent Variable VOLUME

Data Summary

Number of Observations	274423
Sum of Weights	34371895
Weighted Mean of VOLUME	8.89634
Weighted Sum of VOLUME	305783957

Fit Statistics

R-square	0.002480
Root MSE	7.3796
Denominator DF	62

Variance Estimation

Method	Jackknife
Replicate Weights	PAPER
Number of Replicates	62

Class Level Information

Variable	Label	Class Levels	Values
YEAR	YEAR OF DATA	2	Yr1 Yr2

Tests of Model Effects

Effect	Num DF	F Value	Pr > F
Model	1	2.85	0.0962
Intercept	1	323.09	<.0001
YEAR	1	2.85	0.0962

NOTE: The denominator degrees of freedom for the F tests is 62.

YEAR Least Squares Means

Year	Estimate	Standard Error	DF	t Value	Pr > |t|
Yr1	9.2461	0.5096	62	18.14	<.0001
Yr2	8.5093	0.5686	62	14.96	<.0001

The SURVEYREG Procedure

Regression Analysis for Dependent Variable VOLUME

Differences of YEAR Least Squares Means

YEAR OF DATA	YEAR OF DATA	Estimate	Standard Error	DF	t Value	Pr > |t|
Yr1	Yr2	0.7368	0.4362	62	1.69	0.0962

CONCLUSION

The most commonly used variance estimation methods for survey data are now available in the SURVEYFREQ, SURVEYMEANS, SURVEYREG and SURVEYLOGISTIC procedures. These can now be used instead of a series of procedures and DATA step programming. Flexible implementation of replication variance estimation methods in SAS can now be used for a wide variety of estimators. We used a large dataset with nearly 300,000 observations and 100 variables and noted that PROC SURVEYMEANS and PROC SURVEYFREQ are computationally intensive. If you are generating an analysis without standard errors or confidence intervals, we would recommend staying with PROC MEANS or PROC FREQ for familiarity and efficiency. Both support a WEIGHT statement and would produce results identical to their SURVEY counterparts, but if you are using a complex design and using replicate weights to compute the variance and confidence intervals, we strongly recommend using the PROC SURVEY* procedures.

ACKNOWLEDGMENTS

The authors thank Mike Rhoads, Michael Raithel, Rick Mitchell and Fran Bents for all the help they received.

DISCLAIMER

The contents of this paper is the work of the authors and do not necessarily represent the opinions, recommendations, or practices of Westat.

REFERENCES

WesVar 5.0 User's Guide,(2007) Westat Inc. Rockville, MD.

SAS/STAT® 9.22 User's Guide, (2010). SAS Institute Inc, Cary, NC.

Lohr, S. L. (1999), Sampling: Design and Analysis, Pacific Grove, CA: Duxbury Press.

Euan Sul Lee, Ranald . Forthofer and Ronald J. Lorimar (1999), Analyzing Complex Survey Data, Sage Publications Inc.

Pushpal K Mukhopadhyay, Anthony B. An, Randall D. Tobias, and Donna L. Watts (2008), Try, Try Again: Replication-Based Variance Estimation Methods for Survey Data Analysis in SAS® 9.2, SAS Global Forum.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Varma Nadimpalli
Westat
1600 Research Boulevard, WB272
Rockville, MD 20850
240-453-2799
VarmaNadimpalli@westat.com

Katie Hubbell
Westat
1600 Research Boulevard, RA1464
Rockville, MD 20850
301-294-2020
KatieHubbell@westat.com

TRADEMARKS