

# Analysis of Complex Sample Survey Data Using the SURVEY PROCEDURES and Macro Coding

Patricia A. Berglund, Institute For Social Research-University of Michigan, Ann Arbor, Michigan

## ABSTRACT

The paper presents the defining characteristics of complex sample surveys and demonstrates the use of PROC SURVEYMEANS, PROC SURVEYFREQ, PROC SURVEYREG, PROC SURVEYLOGISTIC and SAS® macro coding to correctly analyze these data. Means, Frequency Tables, Linear regression, and Logistic regression are programmed and run assuming a simple random sample and a complex sample design. The analytic techniques presented can be used on any operating system and are intended for an intermediate level audience.

Keywords: complex sample survey data; PROC SURVEYFREQ, PROC SURVEYREG, PROC SURVEYMEANS, PROC SURVEYLOGISTIC, macro language; adjusted variance estimates; simple random samples.

## INTRODUCTION

The paper presents the defining characteristics of complex sample surveys and demonstrates the use of PROC SURVEYMEANS, PROC SURVEYFREQ, PROC SURVEYREG, PROC SURVEYLOGISTIC, and macro coding to correctly analyze these data. Programming techniques and results from simple random sample and complex design corrected analyses are demonstrated and compared.

## BACKGROUND INFORMATION ON COMPLEX SAMPLE SURVEYS

Complex surveys are comprised of data that originate with sample designs that adjust for non-response and differing probabilities of selection. Complex samples differ from simple random samples (SRS) in that SRS designs assume independence of observations while complex samples do not. Statistics produced by most SAS procedures assume a simple random sample and result in under-estimation of variances when analyzing data from complex samples. Therefore, analysis of data from complex surveys should include specific calculation of variance estimates that account for these sample characteristics.

The analyses in this paper use data from the National Comorbidity Survey, a nationally representative sample based on a stratified, multi-stage area probability sample of the United States population (Heeringa, 1996). Weights that adjust for non-response and differing probabilities of selection are routinely used in analyses.

The NCS data file also includes two variables that allow analysts to incorporate the complex survey design into variance estimation computations: the stratum and SECU (Sampling Error Computing Unit). The stratum variable (strata variable) in the NCS reflects the primary sample units while the SECU (cluster variable) represents random half samples of the stratum variable. In the NCS dataset, there are 42 strata with 2 half-samples within each strata.

Use of the SAS survey procedures with the correct weight, strata, and cluster variables will produce variance estimates that incorporate the complex sample design and reflect the inflation of the estimate due to the design effect.

## THE TAYLOR SERIES APPROACH

The Taylor Series Linearization approach (Rust, 1985) is based on a method that derives a linear approximation of variance estimates that are in turn used to develop corrected standard errors and confidence intervals for statistics of interest. A major advantage of the Taylor Series is that it is very efficient computationally as individual replicate models do not have to be calculated. The SAS SURVEYMEANS, SURVEYREG, and SURVEYLOGISTIC procedures use the Taylor Series method.

## RESAMPLING APPROACHES

Resampling approaches follow a specified method of selecting observations defined as probability sub-samples or replicates from which variance estimates are derived. Because the formulation of the probability samples is based upon the complex design, unbiased, design-corrected variance estimates can be derived.

Commonly used methods of resampling include Balanced Repeated Replication (BRR) and Jackknife Repeated Replication (JRR), (Wolter, 1985). Balanced Repeated Replication is a method that reduces the number of sub-samples needed by dividing each stratum into halves. Once the statistic of interest is derived from the half samples (or replicates) the design corrected variance estimations can be developed by using the usual formulas for variance and standard errors.

The Jackknife Repeated Replication method is similar to the BRR in that it performs replicate calculations of interest after developing replicates by deletion of a small and different portion of the total sample for each of the sample subsets.

## PRESENTATION OF THE SAS PROGRAMS AND RESULTS

Four common analysis techniques are demonstrated: Means, Frequency Tables, Linear Regression, and Logistic Regression. For each of these analytical techniques, simple random sample standard errors along with complex design corrected standard errors are re-calculated by the SAS procedure or macro and are compared. All examples use data from the National Comorbidity Survey.

### MEANS

Here are code and output from PROC SURVEYMEANS with and without the strata and cluster variable specifications. Note that using PROC SURVEYMEANS without the strata and cluster variables amounts to using PROC MEANS because a simple random sample is assumed. Note use of weight variable in analysis. (Table 1)

```
PROC SURVEYMEANS data=d.ncsdxdm3 ;
title1 "Using PROC SURVEYMEANS without strata or cluster variables" ;
title2 "Simple Random Sample" ;
var incpers ;
weight plfwt ;
run ;
```

**Table 1: USING PROC SURVEYMEANS WITHOUT STRATA OR CLUSTER VARIABLES  
SIMPLE RANDOM SAMPLE**

#### The SURVEYMEANS Procedure

| Data Summary           |      |         |                |
|------------------------|------|---------|----------------|
| Number of Observations |      | 8098    |                |
| Sum of Weights         |      | 8097.99 |                |
| Variable               | N    | Mean    | Standard Error |
| INCPERS                | 8098 | 19293   | 287.67         |

Next, by specifying the strata and cluster variables in the second PROC SURVEYMEANS run, the complex nature of the design is accounted for and the resultant variance estimates (standard errors) are properly adjusted. As expected, the use of the strata and cluster variables do not affect the estimated means or other statistics calculated, only the standard errors. (Table 2)

```
PROC SURVEYMEANS data=d.ncsdxdm3 ;
title1 "Using PROC SURVEYMEANS with strata and cluster specified" ;
title2 "Complex Sample" ;
strata str ;
cluster secu ;
var incpers ;
weight plfwt ;
run ;
```

**Table 2: USING PROC SURVEYMEANS WITH STRATA AND CLUSTER SPECIFIED  
COMPLEX SAMPLE**

#### The SURVEYMEANS Procedure

| Data Summary           |      |         |                |
|------------------------|------|---------|----------------|
| Number of Strata       |      | 42      |                |
| Number of Clusters     |      | 84      |                |
| Number of Observations |      | 8098    |                |
| Sum of Weights         |      | 8097.99 |                |
| Variable               | N    | Mean    | Standard Error |
| INCPERS                | 8098 | 19293   | 526.253015     |

## FREQUENCY TABLES

For frequency table analysis with categorical variables, PROC SURVEYFREQ (new in SAS v9.1) is demonstrated with and without the complex design variables. Lifetime major depression is coded as a 0/1 dummy variable and SRS and complex design adjusted standard errors are produced. (Table 3).

```
PROC SURVEYFREQ ;
table deplt1 ;
weight plfwt ;
run ;

PROC SURVEYFREQ ;
strata str ;
cluster secu ;
table deplt1 ;
weight plfwt ;
run ;
```

**Table 3: USING PROC SURVEYFREQ WITHOUT AND WITH STRATA OR CLUSTER VARIABLES  
SIMPLE RANDOM SAMPLE AND DESIGN ADJUSTED STANDARD ERRORS**

### The SURVEYFREQ Procedure

#### Major Depression in Lifetime

| Value | WgtN | Wgt %   | SRS Std.Error | Design Corrected SE |
|-------|------|---------|---------------|---------------------|
| 0     | 6716 | 82.9347 | .5814         | .6726               |
| 1     | 1382 | 17.0653 | .5814         | .6726               |

## LINEAR REGRESSION

For regression with a continuous dependent variable, PROC SURVEYREG is demonstrated both with and without the strata and cluster variables. In the following regressions, the dependent variable is personal income predicted by age and sex. Like the preceding means example, the results from the first invocation of PROC SURVEYREG are equivalent to a PROC REG analysis. This is due to the omission of the strata and cluster variables. Note the use of a weight variable for both approaches. (Table 4)

```
PROC SURVEYREG data=d.ncsdxdm3 ;
title1 "Using PROC SURVEYREG without strata or cluster variables" ;
title2 "Simple Random Sample" ;
model incpers=sexm age ;
weight plfwt ;
run ;
```

**Table 4: USING PROC SURVEYREG WITHOUT STRATA OR CLUSTER VARIABLES  
SIMPLE RANDOM SAMPLE**

### The SURVEYREG Procedure

#### Regression Analysis for Dependent Variable INCPERS (PERSONAL INCOME)

#### Estimated Regression Coefficients

| Parameter | Estimate   | Standard Error |
|-----------|------------|----------------|
| Intercept | -10928.832 | 729.725        |
| SEXM      | 10594.816  | 495.422        |
| AGE       | 751.194    | 22.492         |

NOTE: The denominator degrees of freedom for the t tests is 8097.

The second PROC SURVEYREG analysis accounts for the complex design through the use of the strata and cluster variables and standard errors are thus adjusted via the Taylor Series Linearization method. (Table 5)

```
PROC SURVEYREG data=d.ncsdxdm3 ;
title1 "Using PROC SURVEYREG with strata and cluster variables" ;
title2 "Complex Sample" ;
strata str ;
cluster secu ;
model incpers=sexm age ;
weight plfwt ;
run ;
```

**Table 5: USING PROC SURVEYREG WITH STRATA AND CLUSTER VARIABLES  
COMPLEX SAMPLE**

#### **The SURVEYREG Procedure**

#### **Regression Analysis for Dependent Variable**

#### **INCPERS (PERSONAL INCOME)**

#### **Estimated Regression Coefficients**

| Parameter | Estimate   | Standard Error |
|-----------|------------|----------------|
| Intercept | -10928.832 | 757.452        |
| SEXM      | 10594.816  | 584.125        |
| AGE       | 751.194    | 26.143         |

## **LOGISTIC REGRESSION**

The fourth analytic technique presented is logistic regression with a binary dependent variable. SAS v9.1 offers the new SURVEYLOGISTIC procedure while earlier versions of SAS such as v8.2 offers no procedure for logistic regression analysis of complex sample survey data. Use of PROC SURVEYLOGISTIC and SAS macro coding are demonstrated. The commonly used jackknife repeated replication method can be efficiently programmed using SAS macro language.

Results from PROC LOGISTIC and SURVEYLOGISTIC and the invocation of the %jacklog macro follow. The PROC SURVEYLOGISTIC standard errors are based on the Taylor Series method while the results from the %jacklog macro are recalculated with the strata and cluster variables utilized within the JRR framework. (Table 6 for LOGISTIC and SURVEYLOGISTIC output and Table 7 for results from the %jacklog macro)

```
PROC LOGISTIC descending data=d.ncsdxdm3 ;
weight plfwt ;
model deplt1 =sexf ;
run ;

PROC SURVEYLOGISTIC data=d.ncsdxdm3 ;
strata str ;
cluster secu ;
weight plfwt ;
model deplt1 (event='1') =sexf ;
run ;
```

**Table 6: USING PROC LOGISTIC AND SURVEY LOGISTIC FOR SRS AND COMPLEX SAMPLES****The LOGISTIC Procedure**

| Parameter | DF | Estimate | Std.Error |
|-----------|----|----------|-----------|
| SEXF      | 1  | 0.6293   | 0.0610    |

| Odds Ratio | Confidence Limits |
|------------|-------------------|
| 1.876      | 1.665 2.114       |

**The SURVEY LOGISTIC Procedure**

| Parameter | DF | Estimate | Std.Error |
|-----------|----|----------|-----------|
| SEXF      | 1  | 0.629    | 0.0929    |

| Odds Ratio | Confidence Limits |
|------------|-------------------|
| 1.876      | 1.564-2.251       |

Presented next is a macro called %jacklog (Jackknife Repeated Replication for logistic regression). This macro performs logistic regression on the entire sample and then repeats the regression for each of the replicate sub-samples. Each replicate represents approximately 83/84ths of the NCS sample.

The jackknife repeated replication approach is selected for logistic regression as it is the most flexible of the repeated replication methods. One drawback of the balanced repeated replication method is that it requires the strata and cluster variables to be structured as multiples of four.

(Table 7)

```

/*****
program recalculates variance estimates using
jackknife repeated replication / logistic regression
to account for complex sample design
authors Pat Berglund and Ellen Walters
*****/

```

```

%macro jacklog(ncluster,weight,depend,preds,
npred=1,indata=one);

```

```

  *evaluate number of predictors ;
  %do i=1 %to &npred;
    %let pred&i=%scan(&preds,&i,' ');
  %end;

```

```

%let nclust=%eval(&ncluster);

```

```

data one;
  set &indata;

```

```

  *create jackknife replicates by creating replicate weights ;

```

```

%macro wgtcal ;
  %do i=1 %to &nclust ;
    pwt&i=&weight;
    if str=&i and secu=1 then pwt&i=pwt&i*2 ;
    if str=&i and secu=2 then pwt&i=0 ;
  %end;
%mend;
%wgtcal ;

```

```

*run full sample model ;
PROC LOGISTIC data=ONE des
  outest=parms(rename=(%do i=1 %to &npred;
    &&pred&i=b&i %end;));
  model &depend=&preds ;
  weight &weight ;
run ;

```

**\*run replicate models using jackknife weights created in wgtcal macro ;**

```
%macro reps ;
  %do j=1 %to &nclust ;
    PROC LOGISTIC data=ONE des noprint
      outest=parms&j(rename=( %do i=1 %to &npred;
        &&pred&i=p&i._&j %end; ));
      model &depend=&preds ;
      weight pwt&j ;
      run ;
    %end ;
  %mend ;
%reps ;
```

```
data rep ;
  merge parms
    %do k=1 %to &nclust;
      parms&k
    %end;;
```

```
proc datasets;
  delete parms
    %do k=1 %to &nclust;
      parms&k
    %end;;
```

```
data two ;
  set rep ;
  drop _link__type__ intercept _lnlike__name_ ;
```

**\*calculate squared difference between full sample estimates and replicate estimates ;**

```
%macro it ;
  %do j=1 %to &nclust ;
    %do i=1 %to &npred;
      sb&i._&j=(b&i - p&i._&j)**2;
    %end;
  %end;
%mend ;
%it;
```

**\*calculate odd ratios, variance, standard error, and confidence limits using corrected variance/standard error ;**

```
%do i=1 %to &npred;
  orb&i=exp(b&i);
  sb&i=sum(of %do m=1 %to &nclust;

    sb&i._&m %end;);
  seb&i=sqrt(sb&i);
  lcib&i=exp(b&i-(1.96*seb&i)) ;
  ucib&i=exp(b&i+(1.96*seb&i)) ;
%end;
```

```
data three ;
  set two ;
```

```
%do i=1 %to &npred;
  drop b&i orb&i seb&i lcib&i ucib&i;
  b=b&i;
  or=orb&i;
  stderr=seb&i;
  l_95ci=lcib&i;
  u_95ci=ucib&i;
  name=&i;
  output;
%end;
data three ;
  set three ;
  %do i=1 %to &npred;
```

```

        if name=&i then _name_=
        "%upcase(%scan(&preds, &i, ' '));"
    %end;

```

**\*print key variables with corrected statistics ;**

```

proc print noobs ;
    var _name_ b stderr or l_95ci u_95ci;
run ;

```

```
%mend jacklog;
```

**\*invoke macro with relevant macro parameters ;**

```
%jacklog(42,p1fwt,depl1,sexf,npred=1,indata=d.ncsdxm3) ;
```

## OUTPUT FROM THE %JACKLOG MACRO

Here is the output from the %jacklog macro. Note that the Odds Ratio and Point Estimates are the same as from the previous SRS logistic regression but the standard errors and confidence intervals are corrected to account for the complex design. In general, variance estimates and standard errors are larger when design corrections are executed. (Table 7)

**Table 7: USING %JACKLOG macro UTILIZING JACKKNIFE REPEATED REPLICATION FOR COMPLEX DESIGN CORRECTIONS**

| Variable | Estimate | Std.Error |
|----------|----------|-----------|
| SEXF     | 0.629    | 0.092     |

| Odds Ratio | Confidence Limits |
|------------|-------------------|
| 1.876      | 1.564-2.249       |

## OVERVIEW OF COMPLEX DESIGN SOFTWARE

In addition to the SAS Survey procedures presented in this paper, numerous other options exist for analysis of complex survey data. Some of these tools are Sudaan, Stata (svy series of procedures), SPSS, IVEware (ISR freeware), and Westat products. These are all good options but with the inclusion of these procedures within SAS, the analyst can now correctly analyze data directly without using additional software.

## CONCLUSION

With the development of the SURVEYMEANS, SURVEYREG, and SURVEYLOGISTIC procedures, the user can conveniently and correctly analyze data from complex sample surveys. For other analytic techniques not yet included the survey procedures, SAS macro coding offers an efficient and powerful alternative option.

## REFERENCES

- Kalton, G. (1977). "Practical Methods for Estimating Survey Sampling Errors," Bulletin of the International Statistical Institute, Vol 47, 3, pp. 495-514.
- Rust, K. (1985). Variance Estimation for Complex Estimation in Sample Surveys. Journal of Official Statistics, Vol 1, 381-397. (CP)
- Wolter, K.M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag
- Heeringa, S. (1996) "National Comorbidity Survey (NCS): Procedures for Sampling Error Estimation".

## CONTACT INFORMATION

Your comments are welcome.

Contact the author at:

Patricia Berglund

Institute for Social Research

University of Michigan

426 Thompson St.

Ann Arbor, MI 48106

734-936-0551 Voice

[pberg@umich.edu](mailto:pberg@umich.edu)

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