Paper 1644-2015

Analysis of Survey Data Using the SAS® SURVEY Procedures: A Primer

Patricia A. Berglund, University of Michigan-Institute for Social Research

ABSTRACT

This paper provides an overview of analysis of data derived from complex sample designs. General discussion of how and why analysis of complex sample data differs from standard analysis is included. In addition, a variety of applications are presented using PROC SURVEYMEANS, PROC SURVEYFREQ, PROC SURVEYREG, PROC SURVEYLOGISTIC, and PROC SURVEYPHREG, with an emphasis on correct usage and interpretation of results.

INTRODUCTION

Paper 1644-2015 presents a primer on analysis of complex sample design data and how it differs from analysis of standard or simple random sample (SRS) data. Applications focus on key SAS SURVEY procedures including PROC SURVEYMEANS, PROC SURVEYFREQ, PROC SURVEYREG, PROC SURVEYLOGISTIC, and PROC SURVEYPHREG. Each analysis application emphasizes procedure selection, syntax generation, and interpretation of results. Commonly used techniques such as subpopulation analyses and hypothesis tests are included. Comparison to results from standard SAS procedures that lack the ability to correctly incorporate complex sample design features are included to illustrate the potential pitfalls of not using the appropriate SURVEY procedures.

Analysts will gain a basic understanding of how to correctly analyze data derived from complex sample designs with the appropriate SURVEY commands. Although the level is suited to those getting started with survey data analysis, it is assumed that the analyst has an understanding of basic statistical techniques for standard data analysis and a desire to apply these concepts to survey data analysis. SAS 9.4 is used in all applications.

FEATURES OF COMPLEX SAMPLE DESIGN DATA

Complex surveys are comprised of data derived from sample designs that adjust for non-response and differing probabilities of selection. Complex samples differ from standard or simple random samples in that they assume independence of observations while complex samples do not. Most SAS procedures assume that data used is derived from a simple random sample and under-estimate variances when analyzing data from complex samples. Therefore, analysis of data from complex surveys should include methods of variance estimation that account for these sample design features (Kish, 1965 and Rust, 1985).

The SURVEY procedures (PROC SURVEYSELECT, PROC SURVEYMEANS, PROC SURVEYFREQ, PROC SURVEYREG, PROC SURVEYLOGISTIC, and PROC SURVEYPHREG) allow the analyst to perform sampling, incorporate the correct complex sample design features and produce appropriate design-based variances. The focus of this presentation is data analysis rather than sampling therefore use of PROC SURVEYSELECT is omitted.

By default, variance estimation in the SURVEY procedures is done using the Taylor Series Linearization (TSL) method but Repeated Replication approaches such as the Jackknife Repeated Replication (JRR) and Balanced Repeated Replication (BRR) methods are also readily available. Some data producers prefer to release replicate weights due to confidentiality concerns about strata and cluster variables. For the most part, these two methods will produce very similar results and can be used interchangeably. One exception is for data sets published with only replicate weights. In this situation, the analyst must use a Repeated Replication method for variance estimation. For more on choice of variance estimation methods, see the SAS SURVEY documentation, Rust (1985) or Lohr (2010).

APPLICATIONS

The applications demonstrate use of the SURVEY procedures to perform typical descriptive and regression analyses for continuous and categorical variables and time to event indicators used in survival analysis. The applications are broadly outlined in Table 1. Each application presents relevant syntax with selected options, tabular output and graphics from the ODS GRAPHICS system, and interpretation of selected results.

Variable Type	Analytic Technique	SURVEY Procedure
Continuous	Means, Totals	PROC SURVEYMEANS
	Linear Contrasts	PROC SURVEYMEANS/PROC SURVEYREG
	Linear Regression	PROC SURVEYREG
Categorical	Frequency Tables	PROC SURVEYFREQ
	Logistic Regression	PROC SURVEYLOGISTIC
Time to Event of Interest (Survival Analysis)	Proportional Hazards Regression	PROC SURVEYPHREG

Table 1. Outline of Survey Data Analysis Applications

APPLICATION DATA SETS

Data from the National Comorbidity Survey-Replication, a nationally representative sample based on a stratified, multi-stage area probability sample of the United States population (Kessler et al, 2004) and the NHANES 2005-2006 data set (http://wwwn.cdc.gov/nchs/nhanes/search/nhanes05_06.aspx) are used in the applications. Both data sets are based on complex sample designs and contain key variables representing the design features along with weights that adjust for non-response, differing probabilities of selection and post-stratification to a given population. For details on the sample design and weights for each data set, please refer to the data documentation.

VARIABLE LISTS

Figures 1 and 2 include lists of selected variables from the NHANES 2005-2006 data set and the NCS-R data set. The variables are either public release versions used "as is" or variables that have been prepared for use in these analyses. For variable values/ meanings, refer to either the public release codebook (NHANES 2005-2006 data set) or the variable labels (NCS-R data set).

The CONTENTS Procedure

Data Set Name	WORK.ONE	**Observations**	10348
Member Type	DATA	**Variables**	30
Engine	V9	**Indexes**	0
Created	05/06/2014 12:02:42	**Observation Length**	240
Last Modified	05/06/2014 12:02:42	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	WINDOWS_32		
Encoding	wlatin1 Western (Windows)		

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Label
20	BMXBMI	Num	8	Body Mass Index (kg/m**2)
13	BPXDI1	Num	8	Diastolic: Blood pres (1st rdg) mm Hg
15	BPXDI2	Num	8	Diastolic: Blood pres (2nd rdg) mm Hg
17	BPXDI3	Num	8	Diastolic: Blood pres (3rd rdg) mm Hg
19	BPXDI4	Num	8	Diastolic: Blood pres (4th rdg) mm Hg
10	BPXPLS	Num	8	60 sec. pulse (30 sec. pulse * 2):
11	BPXPULS	Num	8	Pulse regular or irregular?
12	BPXSY1	Num	8	Systolic: Blood pres (1st rdg) mm Hg
14	BPXSY2	Num	8	Systolic: Blood pres (2nd rdg) mm Hg
16	BPXSY3	Num	8	Systolic: Blood pres (3rd rdg) mm Hg
18	BPXSY4	Num	8	Systolic: Blood pres (4th rdg) mm Hg
5	INDFMPIR	Num	8	Family PIR
22	LBDHDD	Num	8	Direct HDL-Cholesterol (mg/dL)
21	LBXTC	Num	8	Total Cholesterol(mg/dL)
2	RIAGENDR	Num	8	Gender - Adjudicated
3	RIDAGEYR	Num	8	Age at Screening Adjudicated - Recode
4	RIDRETH1	Num	8	1=mex 2=oth hisp 3=white 4=black 5=other
8	SDMVPSU	Num	8	Masked Variance Pseudo-PSU
9	SDMVSTRA	Num	8	Masked Variance Pseudo-Stratum
1	SEQN	Num	8	Respondent sequence number
6	WTINT2YR	Num	8	Full Sample 2 Year Interview Weight
7	WTMEC2YR	Num	8	Full Sample 2 Year MEC Exam Weight
28	age51	Num	8	1=Age >=51 0 = Age < 51
27	age18p	Num	8	1=Age >= 18 0=Age < 18
26	bpxdi1_1	Num	8	Diastolic Blood Pressure with 0 set to Missing
24	edcat	Num	8	1=0-11 2=12 3=13-15 4=16+ Years of Education
23	irregular	Num	8	1=Irregular Heart Beat 0=Not Irregular Heart Beat
25	marcat	Num	8	1=Married 2=Previously Married 3=Never Married
29	obese	Num	8	Indicator of Being Obese 1=BMI >=30 0=BMI <30 and not missing

Figure 1. Contents Listing and Alphabetic List of Variables from the NHANES 2005-2006 Data Set

The CONTENTS Procedure

Data Set Name	D.CHAPTER_EXERCISES_NCSR	Observations	9282
Member Type	DATA	Variables	22
Engine	V9	Indexes	0
Created	07/06/2010 07:45:24	Observation Length	176
Last Modified	07/06/2010 07:45:24	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Label
1	CASEID	Num	8	CASE IDENTIFICATION NUMBER
2	DSM_GAD	Num	8	1=DSM GAD 5=No DSM GAD
6	ED4CAT	Num	8	1=0-11 Years 2=12 Years 3=13-15 Years 4=16+ Years
3	GAD_OND	Num	8	GAD Age of Onset
11	HHINC	Num	8	Household Income : Topcode
5	MAR3CAT	Num	8	1=Married 2=Sep/Div/Widow 3=Never Married
9	NCSRWTLG	Num	8	NCSR sample part 2 weight
8	NCSRWTSH	Num	8	NCSR sample part 1 weight
7	OBESE6CA	Num	8	1=<18.5 2=18.5-24.9 3=25-29.9 4=30-34.9 5=35-39.9 6=40+
4	REGION	Num	8	1=North East 2=North Central 3=South 4=West
14	SECLUSTR	Num	8	SAMPLING ERROR CLUSTER
13	SESTRAT	Num	8	SAMPLING ERROR STRATUM
10	SEX	Num	8	1=Male 2=Female
12	WKSTAT3C	Num	8	1=Employed 2=Unemployed 3=Out of Labor Force
21	ag4cat	Num	8	1=17-29 2=30-44 3=45-59 4=60+
19	ald	Num	8	1=Alcohol Dependence 0=No Alcohol Dependence
15	bmi	Num	8	Body Mass Index
22	intwage	Num	8	Age at Interview
16	mde	Num	8	1=Major Depressive Episode 0=No Major Depressive Episode
20	racecat	Num	8	1=Asian/Other 2=Hispanic 3=Black 4=White
17	sexf	Num	8	1=Female 0=Not Female
18	sexm	Num	8	1=Male 0=Not Male

Figure 2. Contents Listing and Alphabetic List of Variables from the NCS-R Data Set

COMPLEX SAMPLE DESIGN VARIABLES

The analyst should have a clear understanding of the complex sample design variables and weights to ensure that they are correctly used with the SURVEY commands. Implementation in the selected procedure will ensure production of unbiased point estimates with correct design-based variances.

The risk of not accounting for the complex sample design is that estimates of standard errors/variances and hypothesis tests will likely be under-estimated and may result in erroneous conclusions about analytic results.

The public-release versions of the NHANES and NCS-R data sets provide strata and cluster/primary sampling unit (PSU) variables along with final weight(s) to be used in the SURVEY procedures. The characteristics of each design and weight variable are presented in Table 2.

Data Set	Strata Variable	Cluster Variable	Weight Variables
NHANES 2005-2006	SDMVSTRA-Masked Strata with 15 values	SDMVPSU-Masked Primary Sampling Unit with values of 1 and 2	WTMEC2YR-Interviewed/MEC Examination Weight WTINT2YR-Interviewed only Weight
NCS-R	SESTRAT-Strata Variable with 42 values	SECLUSTR-Sampling Error Computing Unit with values of 1 or 2	NCSRWTSH-Part 1 or "Short" Part of the Interview, n=9282 NCSRWTLG-Part 2 or "Long" Part of the Interview, n=5692

Table 2. Characteristics of the Complex Sample Design Variables and Weights, NHANES 2005-2006 and NCS-R Data Sets

DESCRIPTIVE ANALYSIS OF CONTINUOUS VARIABLES

Application 1. Means Analysis of Body Mass Index

The first application demonstrates analysis of Body Mass Index (BMXBMI) using NHANES 2005-2006 data. The application compares weighted mean BMI from PROC SURVEYMEANS and PROC MEANS with the goal of illustrating how the variances are affected by inclusion of the complex sample design features and weights. The two year Medical Examination Component weight is used in both examples and as a reminder, this is the correct weight for those that completed the interview and MEC portions of the NHANES 2005-2006 survey.

The first block of code uses PROC MEANS with the selected weight (WTMEC2YR) but omits the complex sample variables. The statistics n, number missing, mean and standard error are requested on the PROC MEANS statement and the VAR statement lists BMXBMI as the analysis variable:

```
proc means n nmiss mean stderr;
 weight wtmec2yr;
 var bmxbmi;
run;
```

The second block of code adds the STRATA (SDMVSTRA) and CLUSTER (SDMVPSU) statements along with the WEIGHT (WTMEC2YR) statement and employs the SURVEYMEANS procedure rather than PROC MEANS:

```
proc surveymeans;
     weight wtmec2yr; strata sdmvstra; cluster sdmvpsu;
     var bmxbmi;
    run;
```

Means Analysis of BMI with PROC MEANS

The MEANS Procedure

Analysis Variable : BMXBMI Body Mass Index (kg/m**2)			
N	N Miss	Mean	Std Error
8949	1399	26.4005009	0.0782413

Figure 3. Weighted Mean BMI With SRS Standard Error from PROC MEANS, NHANES 2005-2006 Data

Figure 3 presents results from a weighted PROC MEANS analysis with standard errors based on a simple random sample assumption. The estimated mean BMI is 26.4 with a standard error of 0.08.

Means Analysis of BMI with PROC SURVEYMEANS

The SURVEYMEANS Procedure

Data Summary	
Number of Strata	15
Number of Clusters	30
Number of Observations	10348
Number of Observations Used	9950
Number of Obs with Nonpositive Weights	398
Sum of Weights	291616892

Statistics

Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
BMXBMI	Body Mass Index (kg/m**2)	8949	26.400501	0.218710	25.9343309	26.8666709

Figure 4. Weighted Mean BMI With Design-Based Standard Error from PROC SURVEYMEANS, NHANES 2005-2006 Data

Figure 4 includes a Data Summary table with counts for the strata and cluster variables and information about the weight used. The Statistics table shows the same estimate of mean BMI (26.4) but the standard error from PROC SURVEYMEANS is much higher (0.22) due to incorporation of the complex sample design features through use of the STRATA, CLUSTER, and WEIGHT statements. By default, PROC SURVEYMEANS utilizes the TSL method for variance estimation. This simple comparison highlights the impact of the inclusion of complex sample design features on variance estimates and also implies that ignoring these features can have a major impact on overall conclusions.

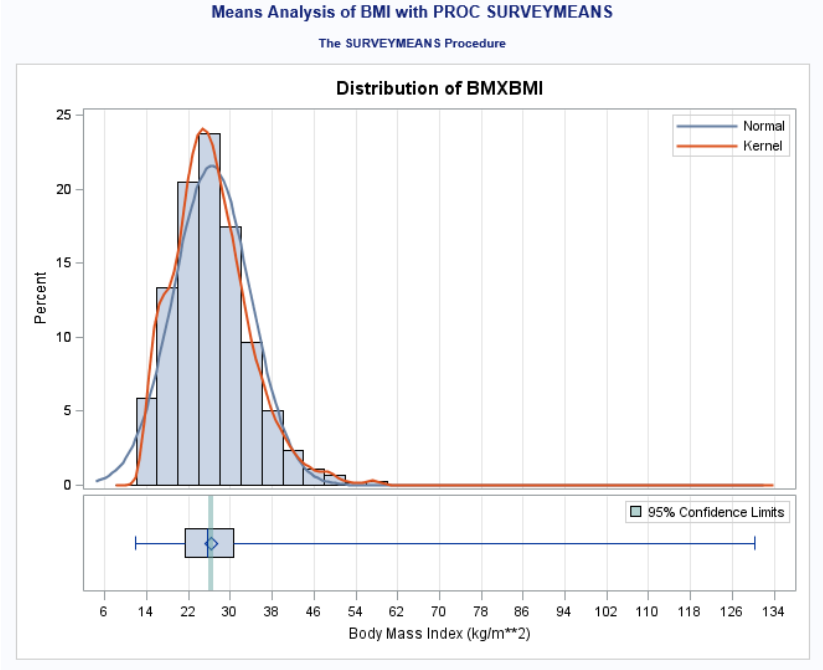


Figure 5. ODS GRAPHICS Histogram of Body Mass Index from PROC SURVEYMEANS, NHANES 2005-2006 Data

Figure 5. is produced by default by ODS GRAPHICS and includes a histogram with superimposed normal and kernel curves along with a box plot, both for the BMI variable. The histogram indicates a relatively normal distribution along with the mean, median, IQR, 95% confidence limits, and range of BMI presented in the box plot.

The following code illustrates use of PROC SURVEYMEANS with an optional Repeated Replication variance estimation method rather than the default Taylor Series Linearization approach (Rust, 1985). The selected method is Jackknife Repeated Replication (VARMETHOD=JK) with all other syntax unchanged from the TSL:

```
proc surveymeans varmethod=jk;
 weight wtmec2yr;  strata sdmvstra;  cluster sdmvpsu;
 var bmxbmi;
run;
```

The SURVEYMEANS Procedure

Data Summary	
Number of Strata	15
Number of Clusters	30
Number of Observations	10348
Number of Observations Used	9950
Number of Obs with Nonpositive Weights	398
Sum of Weights	291616892

Variance Estimation	
Method	Jackknife
Number of Replicates	30

Statistics

Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
BMXBMI	Body Mass Index (kg/m**2)	8949	26.400501	0.218782	25.9341778	26.8668240

Figure 6. Mean BMI with Jackknife Repeated Replication Variance Method from PROC SURVEYMEANS, NHANES 2005-2006 Data

A comparison of standard errors between the two estimation methods shows that the TSL standard error is 0.2187 while the JRR standard error is 0.2188. As expected, the standard errors are very comparable. In general, the TSL method is appropriate for the majority of analyses except when only replicate weights are provided by the data producer or for certain non-linear statistical techniques. For more information on selection of variance estimation methods, see Rust (1985).

Application 2. Totals for Obesity Status

Application 2 demonstrates how to obtain the estimated total number of people considered obese (coded as a binary yes/no variable) in the United States during 2005-2006. Because the NHANES two year interviewed/MEC weight sums to the US population size during these years, no additional weight scaling is needed. Had the weights been normalized to sum to the sample size instead of the inferred population, rescaling would be needed to produce correct population totals.

The following code requests the desired statistics of SUMWGT, SUM, and STD on the PROC SURVEYMEANS statement and produces a sum of weights used in this analysis, an estimated weighted sum (or total) plus the standard deviation of the sum. There are many more statistics available upon request in PROC SURVEYMEANS, see the documentation for a full list:

```
proc surveymeans sumwgt sum std;
 weight wtmec2yr;  strata sdmvstra;  cluster sdmvpsu;
 var obese;
run;
```

Statistics

Variable	Label	Sum of Weights	Sum	Std Dev
obese	Indicator of Being Obese 1=BMI >=30 0=BMI <30 and not missing	279011426	75837426	6205352

Figure 7. Population Total of Persons Considered Obese in US Population, NHANES 2005-2006 Data

The output in Figure 7 indicates that an estimated 75,837,426 (standard deviation=6,205,352) children and adults in the US were considered obese during 2005-2006. The total and standard deviation are based on the sum of weights used in this analysis, 279,011,426.

Application 3. Subpopulation Analysis of BMI By Gender

Application 3 performs a subpopulation analysis of BMI by gender domains. Analysts doing subpopulation analyses are often tempted to simply subset the data using a WHERE or BY statement to include only the observations of interest (a conditional approach) but in the context of survey data analysis, a subset approach will nearly always produce incorrect results. The reason is that subpopulation domains are often unrelated to the original sample design and have sample sizes with variability that should be taken into account.

This issue is addressed by the PROC SURVEYMEANS documentation (SAS/STAT 13.1) as follows: "The formation of these domains might be unrelated to the sample design. Therefore, the sample sizes for the domains are random variables. Use a DOMAIN statement to incorporate this variability into the variance estimation. Note that a DOMAIN statement is different from a BY statement. In a BY statement, you treat the sample sizes as fixed in each subpopulation, and you perform analysis within each BY group independently."

Although this particular application uses PROC SURVEYMEANS, these concepts apply to all types of subpopulation analyses and each SAS SURVEY procedure includes a way to correctly analyze subpopulations.

The code for a correct domain analysis of BMI by gender is presented below. Note the use of the DOMAIN statement with RIAGENDR along with the usual STRATA, CLUSTER, and WEIGHT statements.

The VAR statement defines BMXBMI as the analysis variable while the FORMAT statement applies user-defined value labels for gender where 1=Male and 2=Female. Box plots from ODS GRAPHICS are also presented:

```
proc surveymeans;
 weight wtmec2yr; strata sdmvstra; cluster sdmvpsu;
 var bmxbmi;
 domain riagendr;
 format riagendr sexf.;
run;
```

BMI by Gender Domains

The SURVEYMEANS Procedure

Domain Statistics in RIAGENDR

Gender - Adjudicated	Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
Male	BMXBMI	Body Mass Index (kg/m**2)	4359	26.281119	0.226400	25.7985601	26.7636785
Female	BMXBMI	Body Mass Index (kg/m**2)	4590	26.514639	0.248692	25.9845648	27.0447136

BMI by Gender Domains

The SURVEYMEANS Procedure

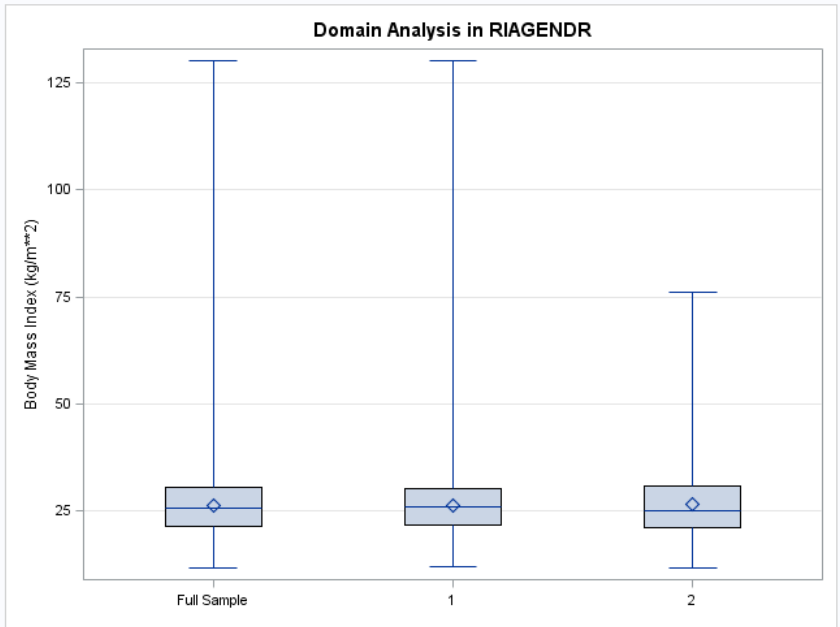


Figure 8. Mean BMI by Gender: Tabular Output and Box Plots, NHANES 2005-2006 Data

Figure 8 presents tabular output including estimated means, standard errors, and confidence limits for BMI separately by gender along with a set of box plots for the full sample and each gender domain. For men, the estimated mean BMI and standard error is 26.28 (0.23) while for women, 26.51 (0.25).

The box plots in the lower section of Figure 8 consist of separate plots for the full sample, men (code=1), and women (code=2). In these plots, the median (horizontal line), the mean (diamond in box), IQR (blue box), and range (whiskers) are presented. Note the differences in the range of BMI for men versus women; the range is much larger for men with a maximum value of about 127 for males versus about 76 for females.

Application 4. Linear Contrast of Means

Another common analytic technique is to perform linear contrasts among means. While PROC SURVEYMEANS does not include the ability to directly perform a design-based linear contrast, this can

be done by PROC SURVEYREG with a CONTRAST statement. This technique is demonstrated in Application 4. In addition, linear contrasts can also be performed with use of the LSMEANS statement and DIFF option in PROC SURVEYREG. This approach will be covered in a later application.

The current application considers the difference in mean BMI for those currently married versus previously married and tests if this difference is statistically significant at the alpha=0.05 level. The following code uses PROC SURVEYREG with a CONTRAST statement to perform a custom hypothesis test, that is, category 1 of MARCAT (currently married) versus category 2 (previously married) with category 3 (never married) omitted. The MODEL statement with BMXBMI predicted by marital status with the SOLUTION option produces a fixed effects solution for the linear outcome of body mass index predicted by *k*-1 levels (3-1=2) of marital status.

As explained above, the CONTRAST statement specifies a linear contrast of marital status category 1 (married) - category 2 (previously married). Note that the estimated parameters of BMI by marital status (when combined with the model intercept) will match the mean values from PROC SURVEYMEANS. This test incorporates the design-based standard errors to produce a survey adjusted test:

```
proc surveyreg;
 weight wtmec2yr;  strata sdmvstra;  cluster sdmvpsu;
 class marcat;
 model bmxbmi=marcat / solution;
 contrast 'Mean Married BMI-Mean Previously Married BMI' marcat 1 -1;
run;
```

Note: The denominator degrees of freedom for the F tests is 15.

Estimated Regression Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	26.1606650	0.29495677	88.69	<.0001
marcat 1	2.3685911	0.24363278	9.72	<.0001
marcat 2	2.9900595	0.23083327	12.95	<.0001
marcat 3	0.0000000	0.00000000	.	.

Analysis of Contrasts

Contrast	Num DF	F Value	Pr > F
Mean Married BMI-Mean Previously Married BMI	1	4.66	0.0474

Note: The denominator degrees of freedom for the F tests is 15.

Figure 9. Estimated Regression Coefficients and Analysis of Contrasts of Mean BMI by Marital Status, NHANES 2005-2006 Data

Regarding Figure 9, the Estimated Regression Coefficients table displays estimated parameters and standard errors for BMI by marital status. The Analysis of Contrasts table presents a test of the null hypothesis that there is no difference in mean BMI for currently married versus previously married respondents. Specifically, the results indicate the difference of married mean BMI (2.36) - previously married mean BMI (2.99) is equal to -0.63 with a design-adjusted test (F=4.66, 1 df, p=0.0474) significant at alpha=0.05 level. Therefore, the null hypothesis is rejected. More complex custom contrasts can be coded via the CONTRAST statement in PROC SURVEYREG, see the SAS/STAT documentation for details.

DESCRIPTIVE ANALYSIS OF CLASSIFICATION VARIABLES

PROC SURVEYFREQ produces design-based frequency tables with correctly estimated variances and hypothesis tests for one-way and n-way tables. The next set of applications present basic techniques for analysis of classification variables. Options included are subpopulation analyses using an implied DOMAIN statement and tests of independence with the CHISQ option of the TABLES statement.

Application 5. Frequency Table Analysis

Again using the NHANES 2005-2006 data set, Application 5 demonstrates use of PROC SURVEYFREQ to analyze marital status. Along with the usual STRATA, CLUSTER, and WEIGHT statements, the TABLES statement identifies the marital status variable (MARCAT) and the FORMAT statement applies user-defined formats to marital status (1=Married, 2=Previously Married, 3=Never Married):

```
proc surveyfreq;
 weight wtmec2yr; strata sdmvstra; cluster sdmvpsu;
 tables marcat;
 format marcat marf.;
run;
```

SURVEYFREQ analysis of Marital Status

The SURVEYFREQ Procedure

Data Summary	
Number of Strata	15
Number of Clusters	30
Number of Observations	10348
Number of Observations Used	9950
Number of Obs with Nonpositive Weights	398
Sum of Weights	291616892

1=Married 2=Previously Married 3=Never Married

marcat	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent
Married	3074	139130779	10560131	59.1774	1.4090
Previously Married	1037	39748009	2694561	16.9063	0.6701
Never Married	2312	56229114	2809688	23.9163	1.1178
Total	6423	235107903	14223022	100.000	

Frequency Missing = 3527

Figure 10. Frequency Table Analysis of Marital Status, NHANES 2005-2006 Data

From Figure 10, the Data Summary table shows 15 strata with 2 clusters per strata or 30 overall clusters in the NHANES data set. The number of observations used in the analysis is 9,950 of 10,348 with 398 observations with non-positive weights. The weights sum to the reference population of about 291 million people in the US in 2005-2006.

The frequency table results suggest that an estimated 59.18% (1.41) of the US adult population were married in 2005-2006, 16.91% (0.67) previously married and 23.92% (1.12) never married. The default output includes unweighted and weighted frequency counts, weighted overall percentages, standard deviation and standard error of the weighted frequency counts and 3,527 missing on the marital status variable (likely children). The missing data is excluded from this analysis but could be separately analyzed using the NOMCAR option on the PROC SURVEYFREQ statement, see the SAS/STAT documentation for details.

Application 6. Cross-tabulation of Obesity Status and Gender

Application 6 examines the relationship of gender and obesity status using the NHANES 2005-2006 data set. Recall that the indicator of being obese (OBESE) is coded as 1 if BMI is >=30 and set to 0 if BMI < 30. Use of gender (RIAGENDR) in the first position of the TABLES statement requests a cross-tabulation of gender and obesity status where gender serves as an implied domain variable. The implied DOMAIN variable is needed since PROC SURVEYFREQ lacks a separate DOMAIN statement like the other SURVEY procedures. This concept can be extended to n-way tables by listing the DOMAIN variable(s) in the first position(s) of the TABLES statement.

The ROW and CHISQ(SECONDORDER) options on the TABLES statement control the output to produce row percentages and a second order Rao-Scott Chi-Square test of independence between gender and obesity status. Previously defined formats are applied via use of the FORMAT statement:

```
proc surveyfreq;
  weight wtmec2yr; strata sdmvstra; cluster sdmvpsu;
  tables riagendr*obese/row chisq(secondorder);
  format riagendr sexf. obese obesef.;
run;
```

Table of RIAGENDR by obese

RIAGENDR	obese	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent	Row Percent	Std Err of Row Percent
Male	No	3467	101161430	5949257	36.2571	0.6652	74.1798	1.6395
	Yes	892	35211930	3840015	12.6202	0.8710	25.8202	1.6395
	Total	4359	136373359	8872636	48.8773	0.4270	100.000	
Female	No	3430	102012571	6513634	36.5621	0.7333	71.5185	1.2595
	Yes	1160	40625496	2663937	14.5605	0.6461	28.4815	1.2595
	Total	4590	142638067	8270162	51.1227	0.4270	100.000	
Total	No	6897	203174000	12353903	72.8192	1.2564		
	Yes	2052	75837426	6205352	27.1808	1.2564		
	Total	8949	279011426	17005007	100.000			

Frequency Missing = 1001

Rao-Scott Chi-Square Test	
Pearson Chi-Square	8.0015
Design Correction	2.4143
First-Order Chi-Square	3.3141
Second-Order Chi-Square	3.3141
DF	1
Pr > ChiSq	0.0687
F Value	3.3141
Num DF	1
Den DF	15
Pr > F	0.0887
Sample Size = 8949	

Figure 11. Cross-Tabulation of Obesity Status and Gender, NHANES 2005-2006 Data

The results presented in Figure 11 indicate that among males, an estimated 25.8% (1.6) are considered obese while 28.5% (1.3) of females are obese. These numbers are contained in the columns named "Row Percent" and "Std Err of Row Percent".

The Rao Scott Chi-Square Test section of Figure 11 presents Chi-Square tests including the requested Rao-Scott second-order Chi-Square test (Rao and Scott, 1984). The second-order test indicates a nearly

significant result with Chi-Square=3.31, df=1, *p*=0.0687. The Pearson Chi-Square and *F* Test are included by default and also suggest acceptance of the null hypothesis of no independence between obesity status and gender.

As a reminder, these hypothesis tests incorporate design-based variances and as such, are correctly calculated for complex sample design data. Had these tests been performed using a SRS assumption with PROC FREQ, they would likely be significant and lead the analyst to make incorrect conclusions about independence between these variables.

REGRESSION ANALYSIS

LINEAR REGRESSION

PROC SURVEYREG is the survey data analysis analog to PROC REG and other standard linear modeling procedures such as PROC MIXED, PROC GLM, and PROC GENMOD. This procedure performs linear regression with many optional statements such as CLASS, CONTRAST, DOMAIN, LSMEANS, among others. As with other SURVEY procedures, use of the STRATA, CLUSTER, and WEIGHT statements incorporates the complex sample design features and weights for use with the Taylor Series Linearization variance estimation method. Repeated Replication methods (JRR and BRR) are also available in this procedure and can be requested on the PROC statement with the VARMETHOD option. The optional DOMAIN statement enables a correctly subpopulation analysis in PROC SURVEYREG and ODS GRAPHICS are available in this procedure.

Application 7. Linear Regression of Systolic Blood Pressure

Application 7 focuses on use of PROC SURVEYREG to perform linear regression of the first of four systolic blood pressure measurements (BPXSY1) regressed on obesity status (OBESE), gender (RIAGENDR), and education (EDCAT). Because the analytic goal is to examine these relationships within the subpopulation of those 40 years of age and older, use of a DOMAIN statement is employed to perform this analysis.

Prior to running the regression, the variables EDCAT and a subpopulation indicator of being age 40+ are created in the DATA STEP. Education is coded in four categories with 1=0-11 years of education, 2=12 years, 3=13-15 years, and 4=16+ years. The AGE40P indicator is coded as 1 if age >=40 and 0 if age < 40.

The syntax for the regression analysis uses the now familiar design variables and weight along with a MODEL statement with the SOLUTION option to obtain a fixed effects solution of parameter estimates, standard errors, and *t*-tests. As a reminder, the standard errors are design-based and are used in all associated tests of significance.

This application highlights use of the LSMEANS statement with a DIFF option to request least squares adjusted means and tests of differences among means for each educational group. As a comparison, one custom test using the CONTRAST statement to contrast the first and second categories of the education groups, i.e. education 0-11 yrs - education 12 years is included. As expected, the results from the LSMEANS/DIFF and CONTRAST statements are identical.

The DOMAIN statement requests a separate analysis for each level of the variable AGE40P, i.e., analyses for those under age 40 and those age 40+. By default, a full sample analysis is included in the output:

```
proc surveyreg;
 weight wtmec2yr; strata sdmvstra; cluster sdmvpsu;
 class marcat riagendr edcat;
 model bpxsy1=riagendr obese edcat / solution;
 lsmeans edcat / diff;
 domain age40p;
 contrast 'Education 0-11 Yrs v. Education 12 Yrs' edcat 1 -1 0 0 ;
 format riagendr sexf. obese obesef. edcat edf.;
```

```
run ;
```

The SURVEYREG Procedure

Age 40+=1, <40 =0=1

Domain Regression Analysis for Variable BPXSY1

Domain Summary	
Number of Observations	5473
Number of Observations in Domain	2490
Number of Observations Not in Domain	2983
Sum of Weights in Domain	112967812
Weighted Mean of BPXSY1	128.39633
Weighted Sum of BPXSY1	1.45047E10

Fit Statistics	
R-Square	0.03512
Root MSE	19.5763
Denominator DF	15

Tests of Model Effects			
Effect	Num DF	F Value	Pr > F
Model	5	20.03	<.0001
Intercept	1	24892.9	<.0001
RIAGENDR	1	2.00	0.1780
obese	1	10.80	0.0050
edcat	3	15.84	<.0001

Note: The denominator degrees of freedom for the F tests is 15.

Estimated Regression Coefficients				
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	122.796403	0.74595362	164.62	<.0001
RIAGENDR Female	1.276632	0.90344036	1.41	0.1780
RIAGENDR Male	0.000000	0.00000000	.	.
obese	3.217007	0.97901498	3.29	0.0050
edcat 0-11 Yrs	9.516801	1.40585101	6.77	<.0001
edcat 12 Yrs	5.048720	0.99811817	5.06	0.0001
edcat 13-15 Yrs	2.667799	1.01861739	2.62	0.0194
edcat 16+ Yrs	0.000000	0.00000000	.	.

Figure 12. Linear Regression of Blood Pressure Among Those Age 40+, NHANES 2005-2006 Data

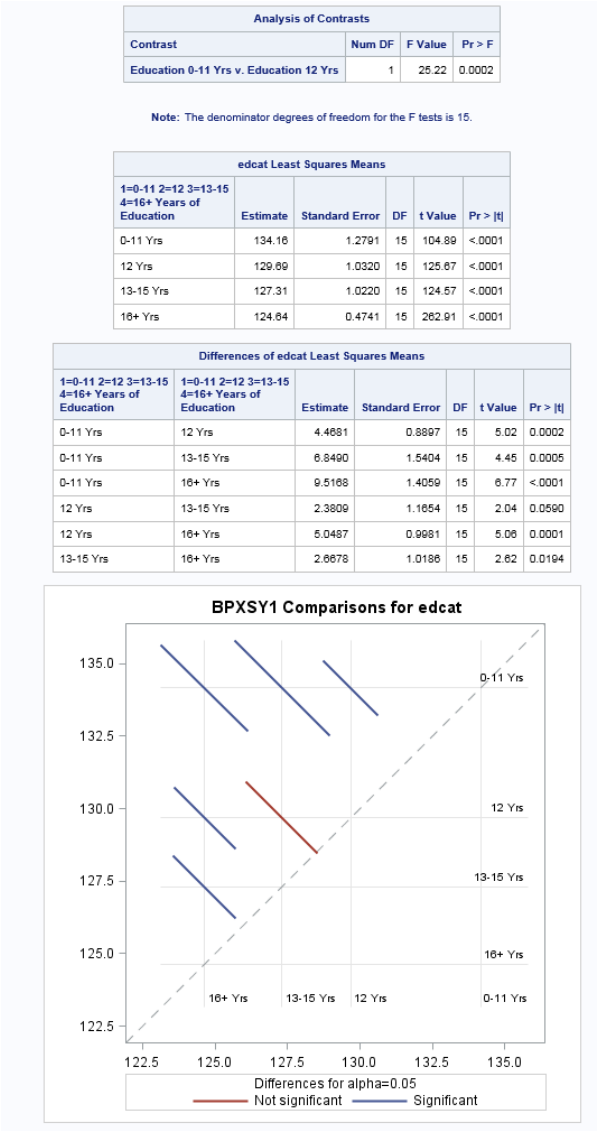
The selected results in Figure 12 are for just the domain of interest: respondents age 40+ at the time of interview. The information at the top of the output states: "**Age 40+=1, <40=0=1**" which means the results are for those age 40 and older. It is important to carefully examine the statements above the Domain Summary to verify that the correct output is being used and interpreted.

The Domain Summary table details the number of observations in the full analysis (5,473), the number of observations in the domain of interest (2,490) and number not in the domain of interest (2,983) along with the weighted mean and sum of the dependent variable, BPXSY1 (1st measurement of systolic blood pressure).

The Fit Statistics indicate a very low R-Square value (0.04) suggesting more predictors are needed to explain the variability of the dependent variable and the Tests of Model Effects results suggest that gender does not significantly differ from zero contribution to the model but obesity status and education in 4 groups do.

The Estimated Regression Coefficients output provides parameter estimates and design-based standard errors from the regression model. These results indicate that among respondents age 40+: as compared to men, being female results in non-significantly higher estimated systolic blood pressure while being obese or in one of the three lowest education groups results in significantly higher estimated systolic

blood pressure as compared to the non-obese and those with 16+ years of education, always holding all other model covariates to a fixed value.



Analysis of Contrasts

Contrast	Num DF	F Value	Pr > F
Education 0-11 Yrs v. Education 12 Yrs	1	25.22	0.0002

Note: The denominator degrees of freedom for the F tests is 15.

edcat Least Squares Means

1=0-11 2=12 3=13-15 4=16+ Years of Education	Estimate	Standard Error	DF	t Value	Pr > |t|
0-11 Yrs	134.16	1.2791	15	104.89	<.0001
12 Yrs	129.69	1.0320	15	125.67	<.0001
13-15 Yrs	127.31	1.0220	15	124.57	<.0001
16+ Yrs	124.64	0.4741	15	262.91	<.0001

Differences of edcat Least Squares Means

1=0-11 2=12 3=13-15 4=16+ Years of Education	1=0-11 2=12 3=13-15 4=16+ Years of Education	Estimate	Standard Error	DF	t Value	Pr > |t|
0-11 Yrs	12 Yrs	4.4681	0.8897	15	5.02	0.0002
0-11 Yrs	13-15 Yrs	6.8490	1.5404	15	4.45	0.0005
0-11 Yrs	16+ Yrs	9.5168	1.4059	15	6.77	<.0001
12 Yrs	13-15 Yrs	2.3809	1.1654	15	2.04	0.0590
12 Yrs	16+ Yrs	5.0487	0.9981	15	5.06	0.0001
13-15 Yrs	16+ Yrs	2.6678	1.0186	15	2.62	0.0194

Figure 13. Contrasts from Linear Regression of Blood Pressure Among Those Age 40 Plus, NHANES 2005-2006 Data

Figure 13 presents Least Squares Means and Contrasts results in tabular and graphic formats within the domain of interest. The Analysis of Contrasts table includes results from the custom contrast of 0-11 years of education versus 12 years of education, F_{1df}=25.22, p=00002. This contrast is highly significant at the alpha=0.05 level.

The Least Squares Means table provides adjusted means for each level of education while the Differences of Edcat Least Squares Means table presents design-based differences for all possible education contrasts. Examination of the p values in far right column indicates that all p values are significant except for the contrast of 12 years of education versus 13-15 years of education (p=0.06).

The plot displays these same findings with blue lines indicating significant differences and the sole red line touching the dotted line indicating a non-significant contrast. The ODS GRAPHICS plot enables easy identification of significant and non-significant contrasts with no additional coding required by the analyst.

LOGISTIC REGRESSION

PROC SURVEYLOGISTIC is the tool of choice for design-based logistic regression. A variety of logistic regression models with binary, ordinal, or nominal outcomes are available in this procedure. As with other SURVEY procedures, variance estimation can be done with Taylor Series Linearization or one of the Repeated Replication (JRR or BRR) methods. The type of logistic regression is specified with the LINK option on the MODEL statement while other optional statements are CLASS, DOMAIN, TEST, CONTRAST, and LSMEANS, among others. See the PROC SURVEYLOGISTIC documentation for more information and examples.

Application 8. Logistic Regression of Major Depressive Episode

Application 8 uses data from the National Comorbidity Survey-Replication (NCS-R), Part 1 (n=9,282). This data set is a nationally representative data set focused on mental health diagnoses and related issues and is rich with classification variables related to mental health (Kessler et al, 2004).

The example uses logistic regression with a binary outcome variable indicating a diagnosis of DSM-IV lifetime Major Depressive Episode (MDE: 1=Yes, 0=No). The regression predicts the probability of having MDE using an indicator of being female (SEXF), a classification variable representing years of education attained in four categories (ED4CAT: 1=0-11 years, 2=12 years, 3=13-15 years, 4=16+ years), and an indicator of having DSM-IV lifetime Generalized Anxiety Disorder (DSM_GAD: 1=Yes GAD, 5=No GAD).

The CLASS statement requests reference group parameterization (PARAM=REF) along with custom specification of omitted categories for ED4CAT and DSM_GAD (REF='0-11 Yrs' and REF='5').

The MODEL statement requests the predicted probability of a diagnosis of MDE (event='1') and the TEST statement specifies a multivariate test of GAD level=1 and being female (sexf=1) and their joint contribution to the overall logistic model (TEST DSM_GAD1, SEXF):

```
proc surveylogistic data=ncsr;
  strata sestrat; cluster seclustr; weight ncsrwtsh;
  class ed4cat (ref='0-11 Yrs') dsm_gad (ref='5')/param=ref;
  model mde (event='1')=dsm_gad sexf ed4cat; format ed4cat edf.;
  testgad_sexf: test dsm_gad1, sexf;
 run;
```

Response Profile

Ordered Value	mde	Total Frequency	Total Weight
1	0	7453	7502.5364
2	1	1829	1779.4637

Probability modeled is mde=1.

Class Level Information

Class	Value	Design Variables		
ED4CAT	0-11 Yrs	0	0	0
	12 Yrs	1	0	0
	13-15 Yrs	0	1	0
	16+ Yrs	0	0	1
DSM_GAD	1	1		
	5	0		

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
DSM_GAD	1	421.3460	<.0001
sexf	1	40.3023	<.0001
ED4CAT	3	12.4111	0.0061

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.0628	0.0785	689.7219	<.0001
DSM_GAD	1	1	2.2812	0.1111	421.3460	<.0001
sexf		1	0.4131	0.0651	40.3023	<.0001
ED4CAT	12 Yrs	1	0.1146	0.0644	3.1624	0.0754
ED4CAT	13-15 Yrs	1	0.2365	0.0702	11.3306	0.0008
ED4CAT	16+ Yrs	1	0.0969	0.0730	1.7629	0.1843

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
DSM_GAD 1 vs 5	9.789	7.873	12.171
sexf	1.512	1.331	1.717
ED4CAT 12 Yrs vs 0-11 Yrs	1.121	0.988	1.272
ED4CAT 13-15 Yrs vs 0-11 Yrs	1.267	1.104	1.454
ED4CAT 16+ Yrs vs 0-11 Yrs	1.102	0.955	1.271

Linear Hypotheses Testing Results

Label	Wald Chi-Square	DF	Pr > ChiSq
testgad_sexf	658.3859	2	<.0001

Figure 14. Selected Output from Logistic Regression of Major Depressive Episode, NCS-R Data

Regarding Figure 14, the Response Profile table details that 1,829 (1,779 weighted) of the 9,282 respondents were diagnosed with lifetime Major Depressive Episode. The Class Level information shows that the omitted categories are 0-11 years of education and no lifetime GAD.

Results presented in the Analysis of Maximum Likelihood Estimates table and Odds Ratio Estimates table indicate that those with higher levels of education are more likely to have an MDE diagnosis but only the 13-15 years of education group is significant, as compared to those with 0-11 years of education. Women are significantly more likely than men to have a diagnosis of MDE and those with GAD are significantly more likely than those without GAD to be diagnosed with MDE, always holding all else equal.

The Linear Hypotheses Testing Results indicate that the joint contribution of GAD and gender are significantly different than zero contribution to the model (Wald Chi-Square=638.39, 2 df, p <.0001). All results are correctly adjusted for the complex sample design by use of PROC SURVEYLOGISTIC with weights and the Taylor Series Linearization method of variance estimation.

Due to space considerations, many additional features such as ordinal logistic regression (outcome > 2 categories with order), nominal logistic regression (outcome with > 2 categories without order), along with the LSMEANS, LSMESTIMATES, DOMAIN, UNIT, ODS GRAPHICS, CONTRAST, and EFFECT statements are not covered. In addition, the NOMCAR (Not Missing Completely at Random) statement which allows analysis of a separate domain comprised of cases with missing data is also excluded.

SURVIVAL ANALYSIS

Survival analysis is broadly focused on time to an event of interest. Common events might be time to disease onset, death, machine failure, or other analyses where survival/failure is studied. Time is treated as continuous (seconds, days or some other small unit) or in discrete units (years, quarters, decades). With either approach, censoring is a related concept used to identify a point in time where information or knowledge of event status is no longer available. In survey data analysis, censoring is often defined by point of interview since no additional information can be gathered after this point. This does not mean, however, that the event of interest cannot occur after censoring, just that knowledge of the event is no longer available.

PROC SURVEYPHREG performs design-based survival analysis using the Proportional Hazards model (Cox, D.R., 1972). This type of model is considered semi-parametric and assumes continuous time with proportional hazards among covariates. It is also commonly called a Cox model.

Application 9. Survival Analysis of General Anxiety Disorder

Application 9 covers use of PROC SURVEYPHREG to perform a Proportional Hazards model using NCS-R Part 1 data (n=9,282).

Prior to analysis, a variable representing time between the event of interest or being censored is constructed in the DATA STEP. For example, AGEEVENT is set to the age of onset of lifetime DSM-IV Generalized Anxiety Disorder where the diagnosis occurred (coded as DSM_GAD=1) or age censored represented by age at interview (INTWAGE) where no GAD onset occurred. Therefore, every respondent receives an value for age on the AGEEVENT variable.

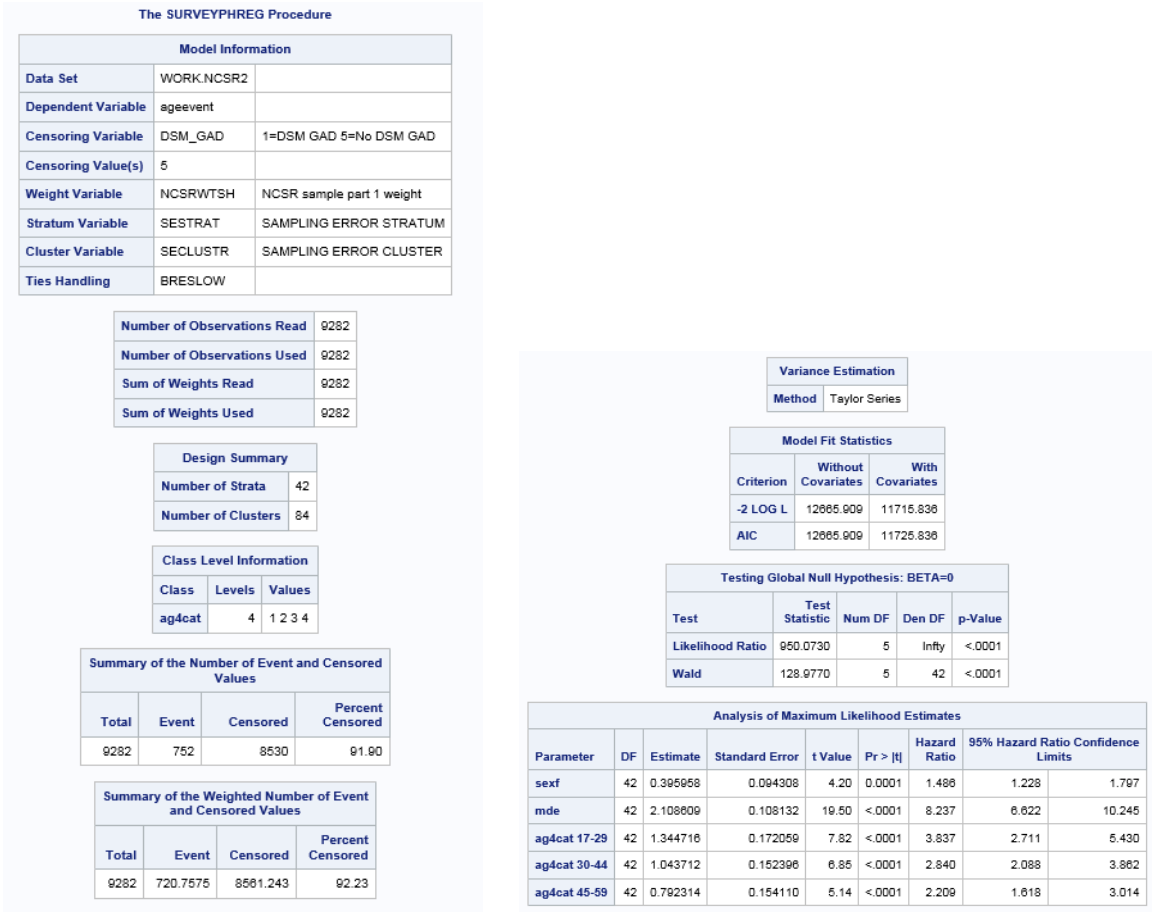
The PROC statement invokes the SURVEYPHREG command and reads the NCSR2 data set as input. The usual STRATA, CLUSTER, and WEIGHT statements use the SESTRAT, SECLUSTR, and NCSRWTSH variables to set the strata, cluster, and weight for the analysis. The age variable (AG4CAT) is declared as classification with reference group parameterizations (CLASS AG4CAT / PARAM=REF).

The MODEL statement merits further explanation. In this code, AGEEVENT*DSM_GAD(5) defines the dependent variable as the product of AGEEVENT and DSM_GAD(5) where the (5) represents censored observations. Thus, the dependent variable evaluates time to the onset of GAD or being censored and is regressed on an indicator of being female (SEXF), an indicator of having Major Depressive Episode (coded as MDE=1=Yes or 0=No), and age in years expressed as four categories (coded as AG4CAT:1=17-29, 2=30-44, 3=45-59, 4=60+ years of age). Use of the RISKLIMITS option on the

MODEL statement requests confidence limits for the hazard ratios produced by the SURVEYPHREG procedure:

```
data ncsr2;
  set ncsr;
  if dsm_gad=1 then ageevent=gad_ond;
  else if dsm_gad=5 then ageevent=intwage;
run;

proc surveyphreg data=ncsr2;
  strata sestrat; cluster seclustr; weight ncsrwtsh;
  class ag4cat/param=ref;
  model ageevent*dsm_gad(5)= sexf mde ag4cat/risklimits;
run ;
```



The SURVEYPHREG Procedure

Model Information		
Data Set	WORK.NCSR2	
Dependent Variable	ageevent	
Censoring Variable	DSM_GAD	1=DSM GAD 5=No DSM GAD
Censoring Value(s)	5	
Weight Variable	NCSRWTSH	NCSR sample part 1 weight
Stratum Variable	SESTRAT	SAMPLING ERROR STRATUM
Cluster Variable	SECLUSTR	SAMPLING ERROR CLUSTER
Ties Handling	BRESLOW	

Number of Observations Read	9282
Number of Observations Used	9282
Sum of Weights Read	9282
Sum of Weights Used	9282

Design Summary	
Number of Strata	42
Number of Clusters	84

Class Level Information		
Class	Levels	Values
ag4cat	4	1 2 3 4

Summary of the Number of Event and Censored Values			
Total	Event	Censored	Percent Censored
9282	752	8530	91.90

Summary of the Weighted Number of Event and Censored Values			
Total	Event	Censored	Percent Censored
9282	720.7575	8561.243	92.23

Variance Estimation	
Method	Taylor Series

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	12665.909	11715.836
AIC	12665.909	11725.836

Testing Global Null Hypothesis: BETA=0				
Test	Test Statistic	Num DF	Den DF	p-Value
Likelihood Ratio	950.0730	5	Infty	<.0001
Wald	128.9770	5	42	<.0001

Analysis of Maximum Likelihood Estimates								
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Hazard Ratio	95% Hazard Ratio Confidence Limits	
sexf	42	0.395958	0.094308	4.20	0.0001	1.486	1.228	1.797
mde	42	2.108609	0.108132	19.50	<.0001	8.237	6.622	10.245
ag4cat 17-29	42	1.344716	0.172059	7.82	<.0001	3.837	2.711	5.430
ag4cat 30-44	42	1.043712	0.152396	6.85	<.0001	2.840	2.088	3.862
ag4cat 45-59	42	0.792314	0.154110	5.14	<.0001	2.209	1.618	3.014

Figure 15. Selected Output from Proportional Hazards Regression of Time to Onset to GAD, NCS-R Data

Based on Figure 15, the results indicate that 752 respondents have an onset of lifetime DSM-IV GAD with 8,530 censored at age of interview age (un-weighted counts). When weighted with the Part 1 weight, roughly 7.8% of the sample have an onset of lifetime GAD with 92.2% censored at the age of interview.

The Variance Estimation method used is the default Taylor Series Linearization method and the Model Fit Statistics indicate a much lower AIC with covariates versus without covariates (11725.84 versus 12665.91) while the Global Tests indicate a good overall model fit with Wald=128.98, 5 df, p <.0001.

The Analysis of Maximum Estimates table presents estimated hazard ratios by default. Hazard ratios are interpreted as the probability that an event will occur at time t, given that it has not yet occurred (a

conditional probability). The ratio for a given predictor represents the impact that a one unit change in that predictor will have on the expected hazard.

The results from Figure 15 suggest that holding all else equal, being female, having MDE, and being in younger age groups at interview have significantly elevated hazards of GAD onset, as compared to omitted categories of male, those without MDE, and those age 60+, respectively. Standard errors and 95% confidence limits are design-adjusted and the degrees of freedom used in the analysis are equal to the number of clusters (2 clusters* 42 strata=84) - number of strata (42)=42. Additional features and options including TEST, LSMEANS, and NOMCAR statements are available but not included here due to space limitations.

CONCLUSION

This presentation has covered the basics of survey data analysis using the SAS SURVEY procedures. Applications using PROC SURVEYMEANS, PROC SURVEYFREQ, PROC SURVEYREG, PROC SURVEYLOGISTIC and PROC SURVEYPHREG have been presented along with a variety of optional statements and features such as the DOMAIN, TEST, LSMEANS, and CONTRAST statements and ODS GRAPHICS.

REFERENCES AND RESOURCES

Cox, D.R., Regression models and life tables, *Journal of the Royal Statistical Society-B*, 34, 187-220, 1972.

Heeringa, West, and Berglund (2010), "Applied Survey Data Analysis", Chapman Hall.

Kessler, R.C., Berglund, P., Chiu, W.T., Demler, O., Heeringa, S., Hiripi, E., et al., The US National Comorbidity Survey Replication (NCS-R): Design and field procedures, *International Journal of Methods in Psychiatric Research,* 13(2), 69-92, 2004.

Kish, L., *Survey Sampling*, John Wiley & Sons, New York, 1965

Korn, E. L. and Graubard, B. I. (1999), Analysis of Health Surveys, New York: John Wiley & Sons.

Lee, E. S., Forthofer, R. N., and Lorimor, R. J. (1989), Analyzing Complex Survey Data, Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-071, Beverly Hills, CA: Sage Publications.

Lohr, S. L. (2010), *Sampling: Design and Analysis*, 2nd Edition, Boston: Brooks/Cole.

Rao, J.N.K. and Scott, A.J., On chi-squared test for multiway contingency tables with cell proportions estimated from survey data, *The Annals of Statistics*, 12, 46-60, 1984

Rust, K. (1985), "Variance Estimation for Complex Estimators in Sample Surveys," Journal of Official Statistics, 1, 381–397.

CONTACT INFORMATION

Your comments and questions are valued and encouraged!

Contact the author at:

Patricia A. Berglund

University of Michigan-Institute for Social Research

pberg@umich.edu