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

Patricia A. Berglund, Institute for Social Research - University of Michigan
Wisconsin and Illinois SAS User's Group
June 25, 2014

Overview of Presentation

- Primer on use of the analytic SURVEY procedures:
  - PROC SURVEYMEANS- continuous variables
  - PROC SURVEYFREQ-classification/categorical variables
  - PROC SURVEYREG-linear regression
  - PROC SURVEYLOGISTIC-logistic regression for binary, nominal, ordinal outcomes
  - PROC SURVEYPHREG-proportional hazards survival model for continous outcome

- Focus on applications of each procedure using the NHANES 2005-2006 and NCS-R 2001-2002 data sets, both derived from a complex sample design
  - How use of SURVEY procedures correctly accounts for complex sample design and how use of standard (SRS) procedure underestimates variance, can lead to incorrect conclusions about analyses

Background on Complex Sample Design Data

Analysis of Complex Sample Design Data

- How to analyze?
  - Incorporate weights, stratification, and clustering through use of variables provided by data producer, generally 3 separate variables but sometimes provided as replicate weights
  - SURVEY procedures allow for correct estimation of variances/standard errors from complex samples
    - Variance estimation by Taylor Series Linearization (default), Jackknife Repeated Replication, or Balanced Repeated Replication (optional, using replicate weights)
- SURVEY procedures cover main analytic techniques:
    - Means/Totals
    - Frequency tables
    - Linear regression
    - Logistic regression
    - Survival models using Proportional Hazards regression

Why are SURVEY procedures needed?

- Use of complex sample design requires variance estimation that accounts for features such as stratification, clustering, and weights
- Most SAS procedures assume that data is from a simple random sample, assumes independence among respondents
- This is clearly not the case when using data based on a complex sample design

Complex Sample Survey Data: Probability Samples

- **Probability sample design**:
  - Each population element has a known, non-zero selection probability
  - Properly weighted, sample estimates are unbiased or nearly unbiased for the corresponding population statistic
  - Variance of sample statistics can be estimated from the sample data (measurability)

- **Simple random sample (SRS):**
  - A probability sample in which each element has an independent and equal chance of being selected for observation
  - Closest population sampling analog to independently and identically distributed (iid) data.

Complex Sample Survey Data: “Complex” Designs

- **Complex sample:**
  - A probability sample developed using sampling procedures such as stratification, clustering and weighting designed to improve statistical efficiency, reduce costs or improve precision for subgroup analyses relative to SRS
  - Unbiased estimates with measurable sampling error are still possible
  - Independence of observations, (iid), equal probabilities of selection may no longer hold

Analysis of Continuous Variables

PROC SURVEYMEANS

Survey Data Analysis-Continuous Variables

- Typical analyses:
  - Means
  - Totals
  - Ratios, quantiles (not shown here)
  - Use PROC SURVEYMEANS for each type of analysis
    - Variance estimation via TSL, JRR, or BRR method
    - Use of STRATA, CLUSTER, and WEIGHT statements (or replicate weights if supplied by data producer)
    - Replicate weights often used when data producer seeks to avoid confidentiality issues (NHANES 1999-2000)

Analysis of Body Mass Index

- This application uses the NHANES 2005-2006 data set:
  - The National Health and Nutrition Examination Survey is an ongoing health survey:
    - based on a complex sample design
    - produced by the NCHS, public release, see http://wwwn.cdc.gov/nchs/nhanes/ for details
    - data set has 15 strata with 2 clusters per strata (SDMVSTR, SDMVPSU)
    - weights:
      - interviewed but no medical exam (WTINT2YR)
      - interviewed and also participated in the medical examination (WTMEC2YR)
- The analysis focuses on estimated mean BMI among those that completed the interview and medical exam plus within selected subpopulations (domains) such as gender and marital status

NHANES 2005-2006 Subset

- Contents Listing:

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

SAS Code for Means Analysis of BMI-PROC MEANS v. PROC SURVEYMEANS

- Weighted means analysis of BMXBMI (BMI) using PROC MEANS
  - (no complex sample adjustment, just 2 year MEC weight):

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

- Design-adjusted, weighted means analysis of BMXBMI (BMI) using PROC SURVEYMEANS with STRATA, CLUSTER, WEIGHT statements:

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

Comparison of Results from PROC MEANS and PROC SURVEYMEANS

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

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

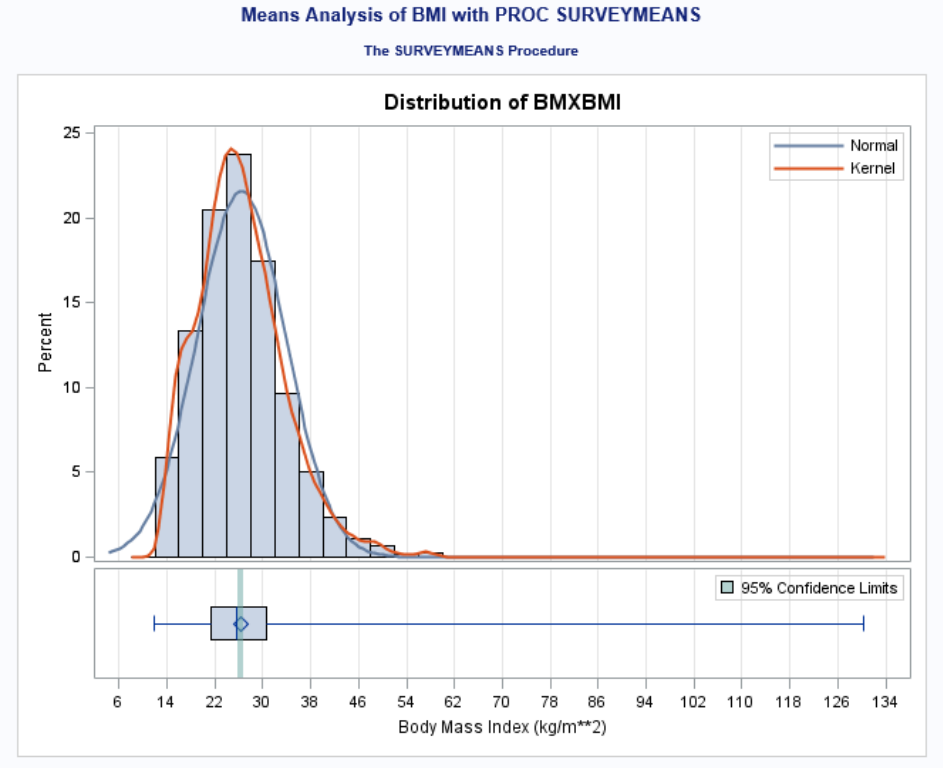
Statistics						
Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
BMXBMI	Body Mass Index (kg/m**2)	8949	26.400501	0.218710	25.9343309	26.8666709

Though the estimated mean of BMI = 26.400 for both analyses, the standard errors are 0.078 (PROC MEANS) and 0.218 (PROC SURVEYMEANS). This is expected due to the impact of the complex sample design on variance estimates.

PROC SURVEYMEANS correctly incorporates the stratification, clustering and weighting in this estimation with use of the Taylor Series Linearization method (TSL).

ODS GRAPHICS from PROC SURVEYMEANS

- ODS GRAPHICS are automatically produced unless you “turn off” these features (ODS GRAPHICS OFF;)
  - Built-in graphics appropriate for the particular procedure you are using
  - Easy way to produce high quality graphics for “free”, no coding required
  - The plot below is automatically produced by PROC SURVEYMEANS



The plot shows that BMI has a relatively normal distribution. It includes both the normal and kernel distributions imposed on the empirical distributions. A boxplot is included below the histogram.

Means Analysis with Jackknife Repeated Replication (JRR) Variance Method

- Jackknife Repeated Replication (JRR) is an alternative variance estimation method based on repeated replication (BRR is another RR option, see documentation for details)

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

Comparison of standard errors:
TSL=.2187
JRR=.2188
As expected, very similar results for this example.

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

Total Analysis from PROC SURVEYMEANS

- Totals are appropriate for binary variables such as being obese or having depression, typically coded yes/no or similar
- This example shows how to obtain the total number of people considered obese using the SUM option on the PROC SURVEYMEANS statement
- NHANES weights sum to population size therefore no scaling is needed, if weights are normalized to sample size then rescaling is needed for correct totals

```
proc surveymeans mean sum stderr ;
 weight wtmec2yr ;  strata sdmvstra ;  cluster sdmvpsu ;
 var obese ;
run ;
```

Statistics

Variable	Label	Mean	Std Error of Mean	Sum	Std Dev
obese	Indicator of Being Obese 1=BMI >=30 0=BMI <30 and not missing	0.271808	0.012564	75837426	6205352

Results suggest that an estimated 27.28% of the US population (2005-2006) had BMI >=30 (obese), this represents 75,837,426 people with this condition. This is based on the weight WTMEC2YR which sums to US population at that time.

Domain Analysis of BMI

- A common analytic task is estimation of a statistic among subpopulations or domains
- Subpopulation analyses must be done with a DOMAIN statement rather than a BY/WHERE statement
- Why?
  - From the SAS PROC SURVEYMEANS documentation (SAS/STAT 13.1):
  - "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."

- SAS code for a correct domain analysis of BMI by gender:

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

Output from BMI by Gender Analysis

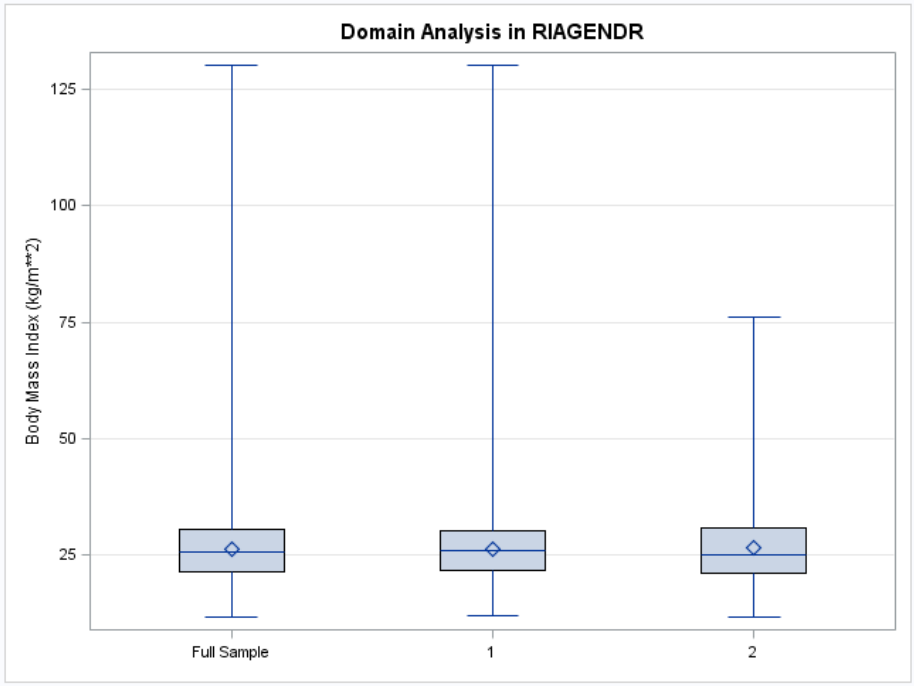
BMI by Gender Domains

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



Results show that estimated mean BMI for males=26.28 and for females=26.51.

The boxplots show slight differences in mean BMI by gender.

The full sample plot is provided by default.

Linear Contrasts of Mean BMI by Marital Status

- PROC SURVEYMEANS does not offer a built-in command to perform a linear contrast or difference in means, therefore use of PROC SURVEYREG with a CONTRAST statement is demonstrated for a test of significant differences in mean BMI by marital status
- This test can also be done with LSMEANS/DIFF in PROC SURVEYREG (more on this in the next section)
- Another slightly out of date but good option is the SAS Institute macro called %smsub (support.sas.com)
  - This provides a macro which produces contrasts much like the PROC SURVEYREG method demonstrated here

PROC SURVEYREG for Linear Contrasts

- Difference in mean BMI for those married v. previously married, is this statistically significant?
- Use PROC SURVEYREG with contrast statement to perform a custom hypothesis test, here category 1 (married) v. category 2 (previously married) with category 3 (never married omitted)

```
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 ;
```

Output from PROC SURVEYREG with CONTRAST

- The linear contrast tests the null hypothesis that there is no difference between 2 levels
- The contrast results show (married mean BMI (2.36)- previously married mean BMI (2.99))= -0.63 with a design-adjusted F=4.66 , 1 df, p=0.0474, significant at alpha 0.05
- This simple example serves as a starting point, more complicated tests can be coded into the CONTRAST statement if desired
- Check SAS documentation for details on use of the CONSTRAST statement,
- Alternatively, use LSMEANS statement which automatically does all differences, more to come on this option!

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.

Analysis of Classification Variables

PROC SURVEYFREQ

Frequency Tables and PROC SURVEYFREQ

- PROC SURVEYFREQ produces complex sample design adjusted variance estimates and hypothesis tests for one-way and multi-way tables
- Subpopulation analyses are done with an “implied” domain variable approach by listing the domain variable FIRST in TABLES statement
- Demonstrations include tables analysis of marital status, gender and obesity

Frequency Table of Marital Status

- Again using the NHANES 2005-2006 data, a one-way table of marital status is produced from PROC SURVEYFREQ

```
title "SURVEYFREQ analysis of Marital Status" ;
proc surveyfreq ;
 weight wtmec2yr ; strata sdmvstra ; cluster sdmvpsu ;
 tables marcat  ;
 format marcat marf. ;
run ;
```

Output from Analysis of Marital Status

Based on the SURVEYFREQ results, an estimated 59.18% (se=1.4) of the US adult population were married in 2005-2006, 16.91% (0.67) previously married and 23.92% (1.12) never married.

Also, 3527 respondents are missing on marital status (likely children)

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

Two-Way Frequency Table of Gender and Marital Status

- Use of RIAGENDR in first position on tables statement requests a crosstabulation of marital status for each level of gender or RIAGENDR (implied domain), concept can be extended to n-way tables
- Use of chisq(secondorder) on tables statement requests a 2nd order correction for the design-adjusted Rao-Scott ChiSq test, considered more accurate than the first order RS test, see documentation for details

```
title "SURVEYFREQ Analysis of Gender * Marital Status" ;
proc surveyfreq ;
 weight wtmec2yr ; strata sdmvstra ; cluster sdmvpsu ;
 tables riagendr*marcat/ row chisq(secondorder) ;
 format riagendr sexf. marcat marf. ;
run ;
```

Output for Gender * Marital Status Frequency Table

SURVEYFREQ Analysis of Gender * 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

Table of RIAGENDR by marcat

RIAGENDR	marcat	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent	Row Percent	Std Err of Row Percent
Male	Married	1553	70697547	5522798	30.0703	0.8370	62.2144	1.5100
	Previously Married	381	13724337	1487461	5.8375	0.4103	12.0775	0.8357
	Never Married	1164	29213498	1640107	12.4256	0.7266	25.7081	1.5441
	Total	3098	113635383	7323542	48.3333	0.4471	100.000	
Female	Married	1521	68433232	5124852	29.1072	0.6841	56.3364	1.5880
	Previously Married	656	26023672	1537451	11.0688	0.5490	21.4235	0.9437
	Never Married	1148	27015616	1672464	11.4907	0.6181	22.2401	1.0995
	Total	3325	121472521	7033598	51.6667	0.4471	100.000	
Total	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

Rao-Scott Chi-Square Test	
Pearson Chi-Square	100.3007
Design Correction	1.6420
First-Order Chi-Square	61.0852
Second-Order Chi-Square	56.7892
DF	1.86
Pr > ChiSq	<.0001
F Value	30.5426
Num DF	1.86
Den DF	27.89
Pr > F	<.0001

Sample Size = 6423

Row percentages suggest that 62.2% of males are currently married while 56.3% of women are married while 12.0% of men and 21.4% of women are previously married. 25.7% of men never marry and 22.2% of women never marry.

The Second-Order Chi-Square test suggests that men and women have significantly different estimated marital status, F=56.7, 1.86 df, p <.0001.

Obesity by Gender

- The next example examines gender differences for being considered obese (BMI >= 30)
- A two-way table from PROC SURVEYFREQ with a design-adjusted F or Chi-Square test allows us to correctly run this analysis

```
title "SURVEYFREQ Analysis of Gender * Obese Indicator " ;
proc surveyfreq ;
 weight wtmec2yr ; strata sdmvstra ; cluster sdmvpsu ;
 tables riagendr*obese/ row chisq(secondorder) ;
 format riagendr sexf. obese obesef. ;
run ;
```

Results for Two-Way Table of Obesity by Gender

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	

Among males, an estimated 25.8% (1.6) are obese while 28.5% (1.3) of females are obese.

The Rao-Scott 2nd order Chi-Square test is close to significant (at the alpha = 0.05 level) with Chi-Square=3.31, df=1, p=0.0687.

Linear Regression Analysis

PROC SURVEYREG

Linear Regression with PROC SURVEYREG

- PROC SURVEYREG is the survey data analysis analog to PROC REG and other linear modeling procedures (PROC MIXED, PROC GLM, PROC GENMOD)
- This tool provides the ability to perform linear regression with many optional statements such as CLASS, CONTRAST, DOMAIN, LSMEANS, and so on (see documentation for details)
- As with other SURVEY procedures, use of the STRATA, CLUSTER, WEIGHT statements incorporates the complex sample design stratification, clustering, and weights for use with TSL
- For repeated replication, replicate weights and the probability weights would be used
- For sub-population or domain analysis, use of the DOMAIN statement correctly performs a subpopulation analysis as well as a full sample analysis

Linear Regression Analysis of Systolic Blood Pressure

- This example focuses on a linear regression of systolic blood pressure regressed on obesity status and education
- Use of PROC SURVEYREG with selected optional statements
- The analytic goal is to examine relationships between blood pressure and obesity/education within the subpopulation of those 40 and older, therefore use of a DOMAIN statement is required
- In data step prior to regression, an indicator of age 40+ is created for use in DOMAIN statement:

  - (note: no missing data on age)

```
if ridageyr >= 40 then age40p=1;
  else age40p=0;
```

Linear Regression with PROC SURVEYREG

- Example shows use of PROC SURVEYREG with LSMEANS, CLASS, DOMAIN, and CONTRAST statements
- LSMEANS with DIFF option provide test of significance of differences between all levels of EDCAT (education in categories)
- CONTRAST statement allows custom specification of desired contrast, provide same results as LSMEANS / DIFF
- DOMAIN produces separate analyses in total sample, <40 years of age, and 40+ years of age (domain of interest)

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

PROC SURVEYREG Output for Subpopulation of Those Age 40+

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	112987812
Weighted Mean of BPXSY1	128.39633
Weighted Sum of BPXSY1	1.45047E10

Fit Statistics	
R-Square	0.03512
Root MSE	19.5783
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	.	.

Estimated Regression Coefficients output provides parameter estimates and correct standard errors from the regression model.

Results suggest that among those 40 and older, compared to men, females have non-significantly higher estimated systolic blood pressure. However, being obese or in lower education groups results in significantly higher estimated systolic blood pressure (compared to non-obese and the highest education group), always holding all else equal.

PROC SURVEYREG Output for Subpopulation of Those Age 40+

Analysis of Contrasts

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

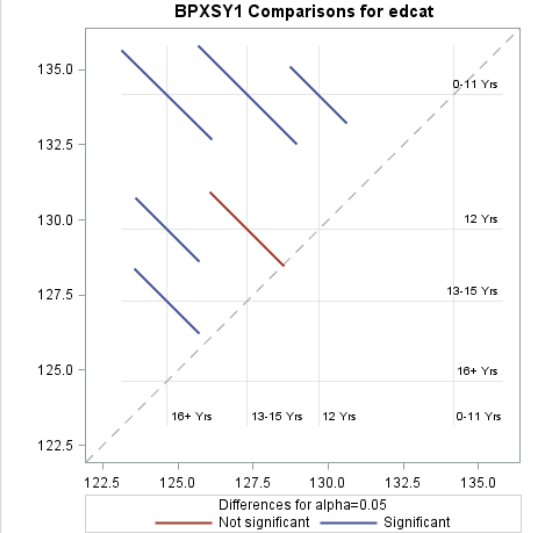
Note: The denominator degrees of freedom for the F tests is 15.

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



Analysis of Contrasts show that Ed 0-11 yrs v. Ed 12 yrs is significant with p value=0.0002.

Differences of LS Means shows estimated differences (minus intercept) for each level of education.

Comparisons Plot indicates which education differences are significant (blue) and not significant (red). If the slanted line touches the vertical dotted line, the difference is non-significant.

PROC SURVEYREG Output, continued

- Output displayed in previous slide was just for those in subpopulation of interest, age 40+ but the same output is displayed for total sample and also those < 40 years of age, check statement at the top of the output to determine which population is used
- What if we had not used PROC SURVEYREG but used PROC MIXED instead, would this alter our overall conclusions?

Linear Regression with PROC MIXED

Solution for Fixed Effects

Effect	Gender - Adjudicated	1=0-11 2=12 3=13-15 4=16+ Years of Education	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept			122.80	0.8735	2484	140.59	<.0001
RIAGENDR	Female		1.2766	0.7867	2484	1.62	0.1048
RIAGENDR	Male		0	.	.	.	.
obese			3.2170	0.8190	2484	3.93	<.0001
edcat		0-11 Yrs	9.5168	1.2064	2484	7.89	<.0001
edcat		12 Yrs	5.0487	1.0822	2484	4.67	<.0001
edcat		13-15 Yrs	2.6678	1.0507	2484	2.54	0.0112
edcat		16+ Yrs	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
RIAGENDR	1	2484	2.63	0.1048
obese	1	2484	15.43	<.0001
edcat	3	2484	22.42	<.0001

Contrasts

Label	Num DF	Den DF	F Value	Pr > F
Education 0-11 Yrs v. Education 12 Yrs	1	2484	13.52	0.0002

Least Squares Means

Effect	1=0-11 2=12 3=13-15 4=16+ Years of Education	Estimate	Standard Error	DF	t Value	Pr > |t|
edcat	0-11 Yrs	134.13	0.9399	2484	142.71	<.0001
edcat	12 Yrs	129.66	0.7705	2484	168.29	<.0001
edcat	13-15 Yrs	127.28	0.7270	2484	175.08	<.0001
edcat	16+ Yrs	124.62	0.7556	2484	164.91	<.0001

Differences of Least Squares Means

Effect	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|
edcat	0-11 Yrs	12 Yrs	4.4681	1.2150	2484	3.68	0.0002
edcat	0-11 Yrs	13-15 Yrs	6.8490	1.1880	2484	5.77	<.0001
edcat	0-11 Yrs	16+ Yrs	9.5168	1.2064	2484	7.89	<.0001
edcat	12 Yrs	13-15 Yrs	2.3809	1.0577	2484	2.25	0.0245
edcat	12 Yrs	16+ Yrs	5.0487	1.0822	2484	4.67	<.0001
edcat	13-15 Yrs	16+ Yrs	2.6678	1.0507	2484	2.54	0.0112

```
proc mixed ;
 weight wtmec2yr ;   class marcat riagendr edcat ;
 model bpxsy1=riagendr obese edcat / solution ;
 lsmeans edcat / diff ;  where age40p=1  ;
 format riagendr sexf. obese obesef. edcat edf. ;
 contrast 'Education 0-11 Yrs v. Education 12 Yrs' edcat 1 -1 0 0 ;
run ;
```

Overall conclusions remain the same except the difference between education 12 v. education 13-15 yrs; this is listed here as significant but when the complex sample is accounted for, this contrast becomes non-significant.

Often, conclusions will differ when using the correct design-based procedure.

Logistic Regression

PROC SURVEYLOGISTIC

Logistic Regression

- PROC SURVEYLOGISTIC is the tool of choice for a variety of logistic regression with outcomes such as:
  - Binary
  - Ordinal
  - Nominal
  - The different types of logistic regression can be requested through use of the LINK option on the MODEL statement
- Other optional statements are:
  - CLASS
  - DOMAIN
  - TEST
  - CONTRAST
  - LSMEAN and so on

PROC LOGISTIC with Binary Outcome

- Logistic regression with a binary outcome is a common use of PROC LOGISTIC/SURVEYLOGISTIC
- This example uses the NCS-R data:
  - National Comorbidity Survey-Replication, 2001-2003, Dr. Ronald Kessler PI, is a nationally representative data set focused on mental health diagnoses, treatment, and other socio-demographic issues, see http://www.hcp.med.harvard.edu/ncs/ for more information

NCS-R Data Subset Contents Lising

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

Analysis of Major Depressive Episode with PROC SURVEYLOGISTIC – Binary Outcome

- This analysis uses a binary outcome variable (MDE) coded 1=Yes has MDE and 0=No MDE predicted by a dummy variable for female (SEXF) and a categorical variable representing education (ED4CAT), and an indicator of having Generalized Anxiety Disorder (DSM_GAD)

- Other features used are:
  - reference parameterization for education (4 levels) and GAD (2 levels)
  - custom specification of reference groups and specification of the probability of having MDE (event='1') as the event being predicted
  - TEST statement to test GAD and sex for their joint contribution to the model

```
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 ;
```

Selected Output from PROC SURVEYLOGISTIC

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		

The Response Profile table details that 1829 (1779 weighted) of 9282 respondents were diagnosed with MDE.

The Class Level information shows that 0-11 yrs is the omitted category for education and 5 is omitted for DSM_GAD.

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

The estimates indicate that all predictors except 16+ yrs of education are significant predictors of the probability of having an MDE diagnosis, holding all else equal and comparing to education 0-11 yrs.

Having GAD, being female, and in educational categories 0-11 or 12 yrs. all significantly predict a diagnosis of MDE. All variance estimates are correctly design-adjusted.

The linear hypothesis test is testing the joint contribution of having GAD and being female equal to 0 contribution to the model. This test indicates that these two variables are jointly significantly different from 0. ($p < .0001$).

Analysis of Marital Status (Nominal Outcome) with PROC SURVEYLOGISTIC

- With marital status as a nominal outcome variable, use of the LINK=GLOGIT option on the MODEL statement is required to produce a multinomial logistic regression
  - default is LINK=LOGIT for PROC SURVEYLOGISTIC
- This example uses the same basic setup as the previous example but adds the correct link option to predict marital status category with education (4 categories) and uses the highest education level as the reference group

```
proc surveylogistic data=ncsr ;
 strata sestrat ; cluster seclustr ; weight ncsrwtsh ;
 class ed4cat / param=ref ;
 model mar3cat =ed4cat / link=glogit ;
 format ed4cat edf.  Format mar3cat marf. ;
 run ;
```

Selected Output from PROC SURVEYLOGISTIC

Response Profile

Ordered Value	MAR3CAT	Total Frequency	Total Weight
1	Married	5322	5182.4763
2	Never Married	1943	2202.2124
3	Previously Married	2017	1897.3115

Logits modeled use MAR3CAT='Previously Married' as the reference category.

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
ED4CAT	6	159.0946	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		MAR3CAT	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		Married	1	1.4605	0.0741	388.8883	<.0001
Intercept		Never Married	1	0.3646	0.0925	15.5482	<.0001
ED4CAT	0-11 Yrs	Married	1	-1.0283	0.1032	99.3724	<.0001
ED4CAT	0-11 Yrs	Never Married	1	-0.7217	0.1263	32.6485	<.0001
ED4CAT	12 Yrs	Married	1	-0.5658	0.0870	42.2957	<.0001
ED4CAT	12 Yrs	Never Married	1	-0.3236	0.1217	7.0662	0.0079
ED4CAT	13-15 Yrs	Married	1	-0.2865	0.0942	9.2387	0.0024
ED4CAT	13-15 Yrs	Never Married	1	0.1162	0.1359	0.7311	0.3925

Odds Ratio Estimates

Effect	MAR3CAT	Point Estimate	95% Wald Confidence Limits	
ED4CAT 0-11 Yrs vs 16+ Yrs	Married	0.358	0.292	0.438
ED4CAT 0-11 Yrs vs 16+ Yrs	Never Married	0.486	0.379	0.622
ED4CAT 12 Yrs vs 16+ Yrs	Married	0.568	0.479	0.673
ED4CAT 12 Yrs vs 16+ Yrs	Never Married	0.724	0.570	0.919
ED4CAT 13-15 Yrs vs 16+ Yrs	Married	0.751	0.624	0.903
ED4CAT 13-15 Yrs vs 16+ Yrs	Never Married	1.123	0.861	1.466

The Response Profile shows 3 values for marital status nominal variable, Married, Never Married, and Previously Married (Omitted).

The Type 3 test shows that education is a significant predictor of marital status (3 levels *2 outcomes = 6 df, p <.0001).

The estimates and odds ratios tables present results separately for each level of the response variable. They suggest that all equal being equal, those with lower education levels, compared to the highest education level, are significantly less likely to be married or never married, compared to those previously married. (The exception is education 13-15 yrs. predicting never married, p=.3925).

Additional Features of PROC SURVEYLOGISTIC

- Many other features are available but not presented here:
  - Ordinal logistic regression (outcome > 2 categories with order)
  - LSMEANS, LSMESTIMATES, DOMAIN, UNIT, ODS GRAPHICS, CONTRAST, and EFFECT statements
  - Another important option is the NOMCAR (Not Missing Completely at Random), allows creating a separate domain of the cases with missing data, enables comparisons with complete cases and analyzes missing data as a domain of its own
  - See documentation for more examples and details

Survival Analysis

PROC SURVEYPHREG

Features of Survival Analysis

- Survival analysis is focused on time and censoring
  - Time to event of interest
    - Disease onset
    - Death
    - Engine failure
  - Measurement of time
    - Continuous time (seconds, days)
    - Discrete time units (2 year periods, decades)
  - Censoring
    - No event of interest during time observed, considered censored (lost to follow-up)
    - Left and right censoring

Event History Data

- Longitudinal data
  - Prospectively collected on individuals followed over time (Panel Study for Income Dynamics)
- Administrative follow-up data
  - Administrative records used to link to additional survey data, prospectively follows those individuals to a key event such as death (NHANES III linked mortality file: http://cdc.gov/nchs/data/datalinkage)
- Retrospective data
  - Respondents asked to recall details about an event of interest which occurred at some point in the past (NCS-R)

Cox Proportional Hazards Model

- Cox PH models are considered semi-parametric, assume continuous time with proportional hazards among covariates
- PROC SURVEYPHREG for Cox model fitting with complex sample survey data is demonstrated
- Data used is NCS-R, requires a few special variables measuring time intervals between events of interest

PROC SURVEYPHREG

- Data step used to create AGEEVENT, set to age of onset of GAD (if DSM_GAD=1) or age at censor represented by INTWAGE or age at interview
- For the model, we use ageevent*dsm_gad(5) as the dependent variable where ageevent * GAD indicator with values of 5 representing those censored, meaning no GAD, covariates are female indicator, MDE indicator, and age in categories
- Use of RISKLIMITS on MODEL statement requests confidence limits for the hazard ratios

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

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

Output from PROC SURVEYPHREG

- Default output from PROC SURVEYPHREG includes the hazard ratio
  - hazard ratio is the probability that an event will occur at time t, given that it has not yet occurred (a conditional probability)
- What does it mean?
  - Hazard ratio for a given predictor represents the impact that a one unit change in that predictor will have on the expected hazard
  - For categorical predictors, the one unit change in a predictor is compared to the omitted reference category

Selected Output from PROC SURVEYPHREG, Outcome is Generalized Anxiety Disorder

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

See SAS code on slide 51.

Results indicate 752 respondents have GAD and 8530 censored at interview age (un-weighted). When weighted with the Part 2 weight, about 8% have GAD with 92% censored.

The Estimates table suggests that holding all else equal, being female, having MDE, and being in younger age groups at interview have significant and increased hazards of GAD onset, as compared to males, those without MDE, and oldest age group. Standard errors and CIs are design-adjusted by PROC SURVEYPHREG.

Associations of Generalized Anxiety Disorder and Age at Interview by Gender

- The next analysis focuses on a survival model predicting time to onset of GAD regressed on age at interview in categories among gender domains
- Use of LSMEANS with a DIFF option and a DOMAIN statement provides tests of differences in age means by gender, along with an ODS GRAPHICS plot, with this option PARAM=GLM must be specified

```
proc surveyphreg ;
  strata sestrat ; cluster seclustr ; weight ncsrwtsh ;
  class ag4cat (ref='4') / param=glm ;
  model ageevent*dsm_gad(5) = ag4cat / risklimits;
  lsmeans ag4cat / diff ;
  domain sexf ; format sexf sf. ;
 run ;
```

PROC SURVEYPHREG Output, Female

The SURVEYPHREG Procedure

Domain Analysis for domain sexf=Female

Number of Observations Read	9282
Number of Observations Used	5143
Sum of Weights Read	4837.294
Sum of Weights Used	4837.294

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
5143	531	4612	89.68

Summary of the Weighted Number of Event and Censored Values			
Total	Event	Censored	Percent Censored
4837.294	481.1094	4356.185	90.05

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	7808.002	7699.933
AIC	7808.002	7705.933

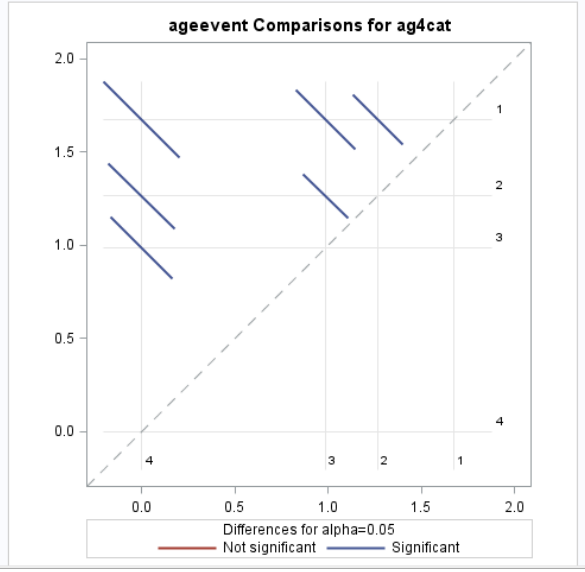
Testing Global Null Hypothesis: BETA=0				
Test	Test Statistic	Num DF	Den DF	p-Value
Likelihood Ratio	108.0691	3	Infty	<.0001
Wald	22.2805	3	42	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Hazard Ratio	95% Hazard Ratio Confidence Limits	
ag4cat 1	42	1.674974	0.203151	8.24	<.0001	5.339	3.543	8.044
ag4cat 2	42	1.264135	0.174229	7.26	<.0001	3.540	2.491	5.032
ag4cat 3	42	0.988026	0.165544	5.97	<.0001	2.686	1.923	3.751
ag4cat 4	42	0	.	.	.	1.000	.	.

Differences of ag4cat Least Squares Means

1=17-29 2=30-44 3=45-59 4=60+	1=17-29 2=30-44 3=45-59 4=60+	Estimate	Standard Error	DF	t Value	Pr > |t|
1	2	0.4108	0.1325	42	3.10	0.0034
1	3	0.6869	0.1560	42	4.40	<.0001
1	4	1.6750	0.2032	42	8.24	<.0001
2	3	0.2761	0.1194	42	2.31	0.0257
2	4	1.2641	0.1742	42	7.26	<.0001
3	4	0.9880	0.1655	42	5.97	<.0001



Among females, all 3 age categories each have a positive and significant impact on the hazard of GAD. In summary, being in a younger age group results in higher estimated hazards, compared to the oldest group.

The LSMEANS comparisons show all differences with CI's (blue lines) are positive and significant.

PROC SURVEYPHREG Output, Male

The SURVEYPHREG Procedure

Domain Analysis for domain sexf=Male

Number of Observations Read	9282
Number of Observations Used	4139
Sum of Weights Read	4444.706
Sum of Weights Used	4444.706

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
4139	221	3918	94.66

Summary of the Weighted Number of Event and Censored Values			
Total	Event	Censored	Percent Censored
4444.706	239.6481	4205.058	94.61

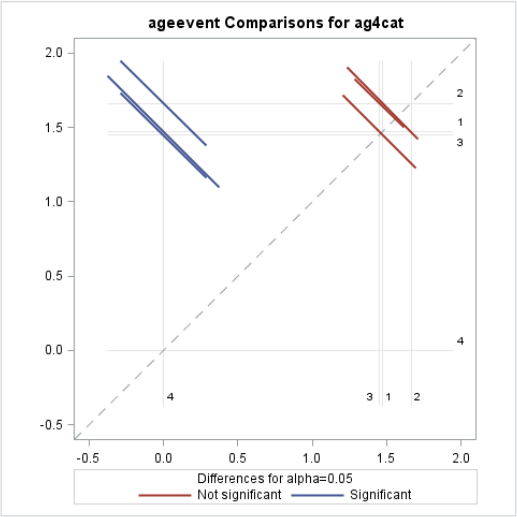
Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	3883.943	3824.757
AIC	3883.943	3830.757

Testing Global Null Hypothesis: BETA=0				
Test	Test Statistic	Num DF	Den DF	p-Value
Likelihood Ratio	59.1860	3	Infty	<.0001
Wald	11.2206	3	42	<.0001

Testing Global Null Hypothesis: BETA=0				
Test	Test Statistic	Num DF	Den DF	p-Value
Likelihood Ratio	59.1860	3	Infty	<.0001
Wald	11.2206	3	42	<.0001

Analysis of Maximum Likelihood Estimates								
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Hazard Ratio	95% Hazard Ratio Confidence Limits	
ag4cat 1	42	1.473901	0.370191	3.98	0.0003	4.366	2.069	9.216
ag4cat 2	42	1.663854	0.284279	5.85	<.0001	5.280	2.975	9.370
ag4cat 3	42	1.449252	0.283771	5.11	<.0001	4.260	2.403	7.553
ag4cat 4	42	0	.	.	.	1.000	.	.

Differences of ag4cat Least Squares Means						
1=17-29 2=30-44 3=45-59 4=60+	1=17-29 2=30-44 3=45-59 4=60+	Estimate	Standard Error	DF	t Value	Pr > |t|
1	2	-0.1900	0.2379	42	-0.80	0.4290
1	3	0.02465	0.2433	42	0.10	0.9198
1	4	1.4739	0.3702	42	3.98	0.0003
2	3	0.2146	0.1632	42	1.32	0.1956
2	4	1.6639	0.2843	42	5.85	<.0001
3	4	1.4493	0.2838	42	5.11	<.0001



Among males, all 3 age categories each have a positive and significant impact on the hazard of GAD, as compared to the omitted oldest age group.

The LSMEANS comparisons show only about half (blue lines) of the differences are significant with the each age group v. the oldest age group significant but the other comparisons non-significant (red lines that cross the 45 degree imposed line).

The DOMAIN analysis reveals differing patterns among gender of age at interview predicting the hazard of a GAD diagnosis.

Presentation Summary

- This presentation has covered the main analytic procedures in the SURVEY group:
  - PROC SURVEYMEANS
  - PROC SURVEYFREQ
  - PROC SURVEYREG
  - PROC SURVEYLOGISTIC
  - PROC SURVEYPHREG

- A variety of optional statements/features have been covered:
  - DOMAIN
  - TEST
  - LSMEANS
  - CONTRAST
  - ODS GRAPHICS
  - Comparison of results to Simple Random Sample based results

- Much more can be done with the SURVEY procedures, see SAS/STAT documentation and additional resources

Additional Resources and References

- SAS/STAT documentation and conference papers
- “Applied Survey Data Analysis” Heeringa, West, and Berglund (2010)
- Website for “Applied Survey Data Analysis” http://www.isr.umich.edu/src/smp/asda/
- IDRE/UCLA https://idre.ucla.edu/stats
- Korn, E. L. and Graubard, B. I. (1999), *Analysis of Health Surveys*, New York: John Wiley & Sons.
- Rust, K. (1985), “Variance Estimation for Complex Estimators in Sample Surveys,” *Journal of Official Statistics*, 1, 381–397.
- 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.

Author Contact Information

- Your comments and feedback are welcome and thank you for attending today!

- Patricia Berglund
- pberg@umich.edu