Paper 140-31

U.S. Health and Nutrition: SAS® Survey Procedures and NHANES

Jeff Gossett, University of Arkansas for Medical Sciences, Little Rock, AR
Chan-Hee Jo, University of Arkansas for Medical Sciences, Little Rock, AR
Pippa Simpson, University of Arkansas for Medical Sciences, Little Rock, AR

ABSTRACT

The National Health and Nutrition Examination Survey (NHANES) is used to evaluate the health and nutrition of the United States (CDC 2005). It is composed of cross-sectional, nationally representative health examination surveys of the U.S. civilian, non-institutionalized population. A complex, stratified, multistage probability cluster sampling design is used in the surveys. Analyzing survey data such as the National Health and Nutrition Survey requires software that accounts for the complex design. You can use SAS® survey procedures to analyze data from NHANES, including the SURVEYMEANS, SURVEYFREQ, SURVEYREG, and SURVEYLOGISTIC procedures. We discuss the use of these procedures and contrast them with the procedures available in SAS callable SUDAAN®(RTI 2001).

INTRODUCTION

The NHANES of the National Center for Health Statistics, Centers for Disease Control and Prevention, is composed of a sequential series of cross-sectional, nationally representative health examination surveys of the US civilian non-institutionalized population. A complex, stratified, multistage probability cluster sampling design is used in the surveys to make the data collection feasible and to be able to address questions about sub populations. In order to make valid statistical inferences, you must account for the complex design. According to the December 2005 update of the NHANES Analytic Guidelines, software such as SUDAAN, STATA and SAS can be used to estimate appropriate sampling errors by the Taylor series (linearization) method. The stratum variable is WTMVSTRA and the PSU variable is WTMVPSU. Typically, the data set should first be sorted by WTMVSTRA and WTMVPSU (CDC 2005). SAS offers the SURVEYMEANS, SURVEYFREQ, SURVEYREG, and SURVEYLOGISTIC procedures for the analysis of survey data. For a full description of each procedures capabilities, please refer to the SAS documentation.

We will combine data from the NHANES 1999-2000 and 2001-2002 releases. For completeness we should note that when the NHANES 1999-2000 data were released, the recommended variance estimation method was the jackknife replication method rather than Taylor linearization, and sets of jackknife replicate weights were included in the release. Although that data could not be analyzed with the SAS Survey Procedures directly, SAS DATA step programming may be used to calculate the estimates (Gossett et al 2004.)

METHODS

OBJECTIVES

A recent paper examined US demographic trends in mid-arm circumference and recommended blood pressure cuffs using the NHANES (Ostchega et al. 2005). The analyses were done with SUDAAN software (RTI 2001). If there is an increasing trend in mid-arm circumference, larger blood pressure cuffs may be needed to make accurate measurements. We will use SAS to duplicate some of the same calculations. The population of interest is adults ages 20 and older. :

1. We will estimate the population distribution of recommended blood pressure cuff sizes, which are based on mid-arm circumference, for subpopulations based on race, sex, and age group using PROC SURVEYFREQ.
2. We will estimate mean mid-arm circumference (cm) for domains defined by gender, age group, and race/ethnicity using PROC SURVEYMEANS.
3. Since the authors were interested in comparing data from 1999-2002 to data from previous NHANES, they employed an age adjustment using the direct method in SUDAAN, employing US Census population projected estimates for the year 2000. We will calculate age adjusted population estimates of mean mid-arm circumference by race and gender using Estimate statements with PROC SURVEYREG.
4. We will then compare age-adjusted mid-arm circumference subpopulation means using means comparisons implemented using PROC SURVEYREG with Estimate statements.

Pharmaceutical companies may estimate disease prevalence to better understand the need for future drug treatments. High blood cholesterol is a major health concern, and total serum blood cholesterol was measured on each respondent.

1. Within subpopulations defined by race or sex, we will estimate the population proportion having high cholesterol (cholesterol>240 mg/dl) using PROC SURVEYMEANS.
2. We will calculate age-adjusted estimates of the population proportion using PROC SURVEYREG with Estimate statements.

3. We will use PROC SURVEYLOGISTIC to estimate model parameter estimates and odd ratios for a simple logistic regression model for the response high blood cholesterol.

NHANES DATA

Starting with the 1999-2000 data release, the National Center for Health Statistics (NCHS) has released public use data sets in two-year groupings. The data for the interview, laboratory and examination components of the survey are released in numerous SAS transport files for the demographics, examination, laboratory, and questionnaire data files. Some measures are only recorded on subsets of the population. Some sample code used to build the data sets is found in the Appendix.

In order to meet our objectives, for each cycle of NHANES (1999-2000 and 2001-2002), we need variables from 4 data sets:

1. Demographics – RIAGENDR (gender), RIDAGEYR (age in years), RIDRETH1 (race), WTMVSTRA (stratum), WTMVPSU (primary sampling unit), and WTMEC4YR (4 year Examination sample weights).
2. Blood Pressure – BPACSZ (blood pressure cuff size)
3. Body Measures – BMXARMC (mid-arm circumference (cm))
4. Lab 13 – LBXTC (serum total cholesterol).

The respondent identification variable, SEQN, is the key variable for merging the files. The data set should first be sorted by the strata variable (WTMVSTRA) and primary sampling unit (WTMVPSU) prior to using any of the Survey data analysis procedures. After merging the data from each cycle and concatenating the resulting two data sets, to form a dataset named "ALL." Two age group variables and an adult indicator (age>19) were defined using DATASTEP programming. The variable AGE1 has levels 20 to 39, 40 to 59, and 60 and older. The variable AGE2 has levels 20 to 39, 40 to 59, 60 to 74, and 75 and older. Prior to the 1999 to 2000 NHANES, NHANES did not include persons ages 75 and older. These variables will also be helpful for the age-standardization calculations.

EXAMPLE 1: DISTRIBUTION OF RECOMMENDED BLOOD PRESSURE CUFF SIZE BY VARIOUS DEMOGRAPHICS

In SAS we use PROC SURVEYFREQ to obtain a distribution of blood pressure cuff size by sex, race and age group for adults ages 20 and older. By default, our dataset contains both adults and children. We only want the adults ages 20 and older. Whereas SUDAAN offers the SUBPOPN statement to restrict the population, we either have to use a dummy category for the ages that aren't of interest or subset the dataset (e.g. with a where statement or DATASTEP deletion.) With a large national data set, we have good representation within each cluster, so we could get away with deleting unwanted respondents. Alternatively, we could include a dummy category for those children under 20. One advantage of SAS over SUDAAN is that we can use ODS to create an RTF file with nice tables. The following PROC SURVEYFREQ statements request a two-way table sex, race, and age group by blood pressure cuff size that includes row percentages.

```
TITLE CROSSTABULATION OF BLOOD PRESSURE CUFF SIZE BY SEX, RACE, AND AGE GROUP FOR
ADULTS AGES 20 AND OLDER;
PROC SURVEYFREQ DATA=ALL(WHERE=(ADULT=1));
  CLUSTER SDMVPSU;
  STRATA SDMVSTRA;
  TABLES SEX*BPACSZ RACE*BPACSZ R_AGE*BPACSZ /ROW;
  WEIGHT WTMEC4YR;
RUN;
```

Partial output of cross tabulation of Sex and Blood Pressure cuff size.

Table of sex by BPACSZ

sex	BPACSZ	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent
F	Infant(6x12)	2	8892	8078	0.0046	0.0042
	Child(9x17)	220	4641896	715587	2.4212	0.3727
	Adult(12x22)	2065	43518514	1720764	22.6990	0.8006
	Large(15x32)	2002	40919177	2199263	21.3432	0.8115
	Thigh(18x35)	493	10419490	879465	5.4347	0.4071
	Total	4782	99507969	3092459	51.9027	0.4361
Total	Total	9094	191720299	5723938	100.000	

Frequency Missing = 377

EXAMPLE 2: MEAN MID ARM CIRCUMFERENCE BY VARIOUS DEMOGRAPHICS

To calculate the mean mid arm circumference by sex and race, we use the following SAS SURVEYMEANS code. The NHANES documentation indicates that you should use the examination weight variable WTMEC4YR. The DOMAIN statement indicates that we want summary tables by RACE and by SEX. If you want summaries by RACE and SEX, you can use RACE*SEX.

```
PROC SURVEYMEANS DATA=ALL(WHERE=(ADULT=1));
  CLUSTER SDMVPSU;
  STRATA SDMVSTRA;
  DOMAIN SEX RACE;
  VAR BMXARMC;
  WEIGHT WTMEC4YR;
RUN;
```

Mean Arm Circumference by Race

Domain Analysis: Linked NH3 Race/Ethnicity - Recode					
Race	N	Mean Arm Circumference (cm)	Std Error Mean	95%CL for Mean	
Non-Hispanic White	4473	32.79	0.10	32.59	33.01
Non-Hispanic Black	1749	34.22	0.15	33.93	34.53
Mexican American	2188	32.51	0.11	32.28	32.75
Other	243	31.16	0.64	29.85	32.50
Other Hispanic	480	32.70	0.23	32.24	33.18
Total	9133	32.87	0.0876	32.69	33.05

If you wanted to compare the mean mid arm circumferences of adult males and females, you can use PROC SURVEYREG using an estimate statement.

```
PROC SURVEYREG DATA=ALL(WHERE=(ADULT=1));
  CLUSTER SDMVPSU;
  STRATA SDMVSTRA;
  CLASS SEX ;
  MODEL BMXARMC=SEX;
  WEIGHT WTMEC4YR;
  ESTIMATE 'M-VS-F' SEX -1 1;
   RUN;
```

Results: Males have an average arm circumference that is 2 cm larger (just under 1 inch).

Analysis of Estimable Functions				
Parameter	Estimate	Standard Error	t Value	Pr > |t|
M-vs-F	1.895	0.124	15.27	<.0001

AGE ADJUSTMENT

Age-adjustment is commonly used for trend analyses between NHANES surveys and for comparisons between subgroups within NHANES in an attempt to equalize the populations on age. If a statistic of interest varies substantially by age within race-ethnic categories, then you should consider age-standardized estimates. If we are interested in trends in arm circumference size between NHANES samples, you might want to adjust the estimates to account for possible differences in the age distributions. It is common practice to standardize to an age distribution based on the 2000 Census. The NHANES Analytic Guidelines for 1999-2000 provide the following guidance on the age adjustment:

The following standard proportions are based on the 2000 standard population and should be used in NHANES 1999-2000 analyses when using 20 year age groups for 20 years and older.

Age Group (AGE1)	*Proportion*
20-39	*0.3966*
40-59	*0.3718*
60+	*0.2316*

Past NHANES surveys did not have sample persons at ages 75 years and over. To compare age-adjusted (ages 20-74 years only) statistics for NHANES 1999-2002 with past NHANES surveys, the following standard proportions should be used:

Age Group (AGE2)	*Proportion*
20-39	*0.4332*
40-59	*0.4062*
60-74	*0.1606*

In the SUDAAN software, these proportions are used with statements STDVAR and STDWGT, where STDVAR lists the name of the variable with age categories used in standardization and STDWGT lists the corresponding proportions from the year 2000 Census.

There are no corresponding statements in SAS for standardization. However, SAS does provide the tools to do the calculations. Age adjusted estimates may be calculated using Estimate statements in PROC SURVEYREG. Let's assume that you want to estimate age-adjusted average arm circumference by race. For each race, you need to estimate the average arm circumference for each age group. The overall estimate X_A is a weighted average of the age domain means:

$$X_A = p_1 X_1 + p_2 X_2 + p_3 X_3$$

Where p_1, p_2, and p_3 are the proportions within each age range, and X_1, X_2, and X_3 are the domain mean estimates corresponding to the age ranges. To demonstrate the calculation, use the age distribution: 39.66% (ages 20-39), 37.18% (ages 40-59), and 23.16% (ages 60+). Use SURVEYMEANS to calculate unadjusted mean arm circumferences by age group and race.

EXAMPLE 3: AGE-ADJUSTED MEAN MID ARM CIRCUMFERENCE BY RACE 1

```
PROC SURVEYMEANS DATA=ALL(WHERE=(ADULT=1));
  CLUSTER SDMVPSU;
  STRATA SDMVSTRA;
  DOMAINS AGE1*RACE;
  VAR BMXARMC;
  WEIGHT WTMEC4YR;
RUN;
```

You use the resulting estimates, reweight the marginals and sum them to get totals. For example, for the Non-Hispanic Whites, you multiply the 20-39 age group mean (32.47) by 0.3966, the 40-50 mean (33.51) by 0.3718, and the 60+ mean (32.17) by 0.2316. The mean for Non-Hispanic whites, is then the sum of the three: Age Adjusted Mean Arm Circumference NH Whites= 32.47*0.3966 + 33.51*0.3718 + 32.17*0.2316 = 32.79. Similarly, you can calculate the age adjusted mean for Non-Hispanic blacks to be 34.22.

Marginal Means and Age Standardized Means by Race

Race	Age Group	Mean (M)	Std Age Proportion (P)	M*P	Age Std Mean
NH White	20-39	32.472076	0.3966	12.878	32.79
	40-59	33.510145	0.3718	12.459	
	60+	32.171736	0.2316	7.4510	
NH Black	20-39	33.947845	0.3966	13.4637	34.22
	40-59	34.746154	0.3718	12.9186	
	60+	33.844167	0.2316	7.8383	

EXAMPLE 4: AGE-ADJUSTED MEAN MID ARM CIRCUMFERENCES BY RACE 2

You can use PROC SURVEYREG with Estimate statements to obtain the standardized estimates. Depending on your comfort level with writing estimate statements, you may choose between a cell means model and an effects model. We will illustrate the calculations with an effects model (overall mean plus effects of each factor). You may want to use the SOLUTION option on the model statement to obtain an ordered list of model parameters. The ordered Race groups are: Mexican American, Non-Hispanic Black, Non-Hispanic White, Other, and Other Hispanic. The ordered age groups are: 20 to 39, 40 to 59, and 60 and older. The order of model parameters depends on order on class statement), the estimate statements would look like:

```
PROC SURVEYREG DATA=ALL(WHERE=(ADULT=1));
  CLUSTER SDMVPSU;
  STRATA SDMVSTRA;
```

```
  CLASS RACE AGE1;
  WEIGHT WTMEC4YR;
  MODEL BMXARMC = AGE1 RACE;
ESTIMATE "MEX AMER" INTERCEPT 1 RACE 1 0 0 0 0 AGE1 0.3966 0.3718 0.2316 AGE1*RACE
0.3966 0.3718 0.2316 0 0 0 0 0 0 0 0 0 0 0;
ESTIMATE "NH BLACK"  INTERCEPT 1 RACE 0 1 0 0 0 AGE1 0.3966 0.3718 0.2316 AGE1*RACE
0 0 0 0.3966 0.3718 0.2316 0 0 0 0 0 0 0 0;
ESTIMATE "NH WHITE"  INTERCEPT 1 RACE 0 0 1 0 0 AGE1 0.3966 0.3718 0.2316 AGE1*RACE
0 0 0 0 0 0 0.3966 0.3718 0.2316 0 0 0 0 0 0;
ESTIMATE "OTHER"  INTERCEPT 1 RACE 0 0 0 1 0 AGE1 0.3966 0.3718 0.2316
AGE1*RACE 0 0 0 0 0 0 0 0 0 0 0.3966 0.3718 0.2316 0 0 0;
ESTIMATE "OTHER HISPANIC"  INTERCEPT 1 RACE 0 0 0 0 1 AGE1 0.3966 0.3718 0.2316
AGE1*RACE 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.3966 0.3718 0.2316;
RUN;
```

If you want to compare the age standardized mean estimates you can use ESTIMATE statements. For example, suppose you wanted to compare age-adjusted mean estimates of Non-Hispanic Whites to Non-Hispanic Blacks. It is a simple matter of subtracting coefficients of the estimates of NH Blacks and NH Whites:

```
ESTIMATE “NH BLACK – NH WHITE” RACE 0 1 -1 0 0 AGE1*RACE  0 0 0 0.3966 0.3718 0.2316
-0.3966 -0.3718 -0.2316 0 0 0 0 0 0;
```

Age-Adjusted Mean Arm Circumference (cm) Estimates from SAS SURVEYREG

Parameter	Estimate	Std Error	T value	Pr>|t|
Mex Amer	32.62	0.12	261.9	<.0001
NH Black	34.22	0.15	231.4	<.0001
NH White	32.79	0.11	307.4	<.0001
Other	31.04	0.60	51.8	<.0001
Other Hispanic	32.63	0.20	161.7	<.0001
NH Black - NH White	1.43	0.18	7.9	<.0001

EXAMPLE 5: POPULATION PROPORTION OF CHOLESTEROL PREVALENCE – ADULTS 20 TO 74.

```
TITLE “UNADJUSTED MEANS FOR PERCENTAGE WITH HIGH TOTAL CHOLESTEROL”;
PROC SURVEYMEANS DATA=ALL(WHERE=(20<=RIDAGEYR<=74));
  CLUSTER SDMVPSU;
  STRATA SDMVSTRA;
  DOMAIN SEX RACE AGE2;
  VAR HIGHTC;
  WEIGHT WTMEC4YR;
RUN;
```

Unadjusted Estimate of Percentage of Adults 20 to 74 with High Cholesterol by Race

Race	N	Mean	Std Err
Total	7740	17.06	0.67
Non-Hispanic White	3596	17.88	0.85
Non-Hispanic Black	1500	13.90	1.04
Mexican American	2000	12.80	0.87
Other	217	19.21	3.00
Other Hispanic	427	16.96	1.80

EXAMPLE 6: POPULATION PROPORTION OF CHOLESTEROL PREVALENCE – AGE STANDARDIZED ESTIMATES.

Now compute age-adjusted population proportion of adults with high cholesterol (at least 240 mg/dl) by race. Pretend that you would like to compare to a previous NHANES, so you will use age distribution 2 for the adjustment. You will need to include only the adults ages 20 to 74. To estimate the Age standardized prevalence of high serum total cholesterol of adults 20-74 years of age: United States, 1999-2002 you can use PROC SURVEYREG.

```
PROC SURVEYREG DATA=ALL(WHERE=(20<=RIDAGEYR<=74));
  CLUSTER SDMVPSU;
  STRATA SDMVSTRA;
  CLASS RACE AGE1;
  WEIGHT WTMEC4YR;
  MODEL HIGHTC = RACE AGE1 AGE1*RACE /SOLUTION;
ESTIMATE "MEX AMER" INTERCEPT 1 RACE 1 0 0 0 0 AGE1  0.4332 0.4062 0.1606
    AGE1*RACE  0.4332 0.4062 0.1606 0 0 0 0 0 0 0 0 0 0 0;
ESTIMATE "NH BLACK"  INTERCEPT 1 RACE 0 1 0 0 0 AGE1  0.4332 0.4062 0.1606
    AGE1*RACE  0 0 0  0.4332 0.4062 0.1606 0 0 0 0 0 0 0 0 0;
ESTIMATE "NH WHITE"  INTERCEPT 1 RACE 0 0 1 0 0 AGE1  0.4332 0.4062 0.1606
    AGE1*RACE 0 0 0 0 0 0  0.4332 0.4062 0.1606 0 0 0 0 0 0;
ESTIMATE "OTHER"  INTERCEPT 1 RACE 0 0 0 1 0 AGE1  0.4332 0.4062 0.1606
    AGE1*RACE 0 0 0 0 0 0 0 0 0  0.4332 0.4062 0.1606 0 0 0;
ESTIMATE "OTHER HISPANIC"  INTERCEPT 1 RACE 0 0 0 0 1 AGE1  0.4332 0.4062 0.1606
    AGE1*RACE 0 0 0 0 0 0 0 0 0 0 0 0 0.4332 0.4062 0.1606;
ESTIMATE "NH BLACK - NH WHITE" RACE 0 1 -1 0 0
    AGE1*RACE  0 0 0  0.4332 0.4062 0.1606 -0.4332 -0.4062 -0.1606  0 0 0 0 0 0;
RUN;
```

Resulting Age Adjusted Estimates

Parameter	Estimate	Error	t Value	Pr > |t|
Mex Amer	14.91	0.93	15.99	<.0001
NH Black	14.57	1.11	13.15	<.0001
NH White	17.42	0.81	21.4	<.0001
Other	19.22	2.85	6.75	<.0001
Other Hispanic	18.67	2.25	8.3	<.0001
NH Black - NH White	-2.85	1.38	-2.06	0.0486

The age-adjusted estimates for high cholesterol seem to be higher.

Parameter	Unadjusted Percentage	Age-Adjusted Percentage
Mex Amer	12.80%	14.91%
NH Black	13.90%	14.57%
NH White	17.88%	17.42%
Other	19.21%	19.22%
Other Hispanic	16.96%	18.67%

EXAMPLE 7: LOGISTIC REGRESSION ANALYSIS OF HIGH CHOLESTEROL

You can use PROC SURVEYLOGISTIC to fit a logistic model to the response High Cholesterol. The response, High Cholesterol (HighTC), is coded as 0 for those not having high cholesterol, and 1 for those having high cholesterol. By default, the event "HighTC=0" is modeled (i.e. low cholesterol). We use the syntax "(event="1")" to indicate that we want to model high cholesterol (hightc=1). Suppose you want to use as predictors: sex, age group, and race. By default, the reference category is the last level. To specify the reference categories, you can use the "ref=" option on the class statement. If you want to use Males as the reference for sex, the lowest/first age category for age group, and Non-Hispanic Whites for race, you can use the following call.

```
PROC SURVEYLOGISTIC DATA=ALL(WHERE=(20<=RIDAGEYR)) ;
  CLUSTER SDMVPSU; STRATA SDMVSTRA;
  CLASS SEX(ref='M') AGE1(ref=FIRST) RACE(ref='Non-Hispanic White');
  WEIGHT WTMEC4YR;
  MODEL HIGHTC(EVENT='1') = SEX AGE1 RACE ;
RUN;
```

You can see that only AGE1 is significant at the 5% level.

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
sex	1	1.5123	0.2188
age1	2	92.8665	<.0001
Race	4	8.3592	0.0793

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
sex F vs M	1.091	0.950	1.253
age1 2 vs 1	2.257	1.834	2.778
age1 3 vs 1	2.348	1.968	2.803
Race Mexican American vs Non-Hispanic White	0.824	0.687	0.990
Race Non-Hispanic Black vs Non-Hispanic White	0.813	0.657	1.006
Race Other vs Non-Hispanic White	1.174	0.798	1.728
Race Other Hispanic vs Non-Hispanic White	1.036	0.776	1.384

A more detailed discussion of PROC SURVEYLOGISTIC may be found in An (2002).

CONCLUSION

We have shown that using the new Survey Analysis procedures in SAS, we can analyze complex sample surveys and national datasets like NHANES. Furthermore, we can calculate age-adjusted estimates, calculate population prevalence, and perform modeling such as linear or logistic regression.

REFERENCES

Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, [2001-2002][http://www.cdc.gov/nchs/about/major/nhanes/nhanes01-02.htm].

Research Triangle Institute (2001). SUDAAN User's Manual, Release 8.0 Research Triangle Park, NC: Research Triangle Institute.

SAS Institute, Inc. (2003), *SAS OnlineDoc, Version 9.1,HTML Format*

US demographic trends in mid-arm circumference and recommended blood pressure cuffs: 1988–2002. Y Ostchega, C Dillon, M Carroll, R J Prineas, M McDowell. Journal of Human Hypertension 19, 885-891 (01 Nov 2005)

JM Gossett, P Simpson, JG Parker, and WL Simon. *How Complex Can Complex Survey Analysis be with SAS®?* SUGI 27, Paper 266-27.

Klein RJ, Schoenborn CA. *Age adjustment using the 2000 projected U.S. population. Healthy People Statistical Notes*, no. 20. Hyattsville, Maryland: National Center for Health Statistics. January 2001. (http://www.cdc.gov/nchs/products/pubs/pubd/hp2k/statnt/20-11.htm)

Age-adjustment and trend analyses (from NHANES 1999-2000 Addendum to the NHANES III Analytic Guidelines, Last Update 8/30/02. (http://www.cdc.gov/nchs/data/nhanes/guidelines1.pdf)

NHANES 1999-2000 Addendum to the NHANES III Analytic Guidelines. Last Update 8/30/02 (http://www.cdc.gov/nchs/data/nhanes/guidelines1.pdf)

NHANES Analytic Guidelines, June 2004 Version. (http://www.cdc.gov/nchs/data/nhanes/nhanes_general_guidelines_june_04.pdf)

The National Health and Nutrition Examination Survey (NHANES) ANALYTIC AND REPORTING GUIDELINES, Last Update: December, 2005. National Center for Health Statistics
Centers for Disease Control and Prevention Hyattsville, Maryland
(http://www.cdc.gov/nchs/data/nhanes/nhanes_03_04/nhanes_analytic_guidelines_dec_2005.pdf)

Anthony B. An. *Performing Logistic Regression on Survey Data with the New SURVEYLOGISTIC Procedure.* PSUG 2002, Paper SAS05. (http://www.lexjansen.com/pharmasug/2002/proceed/sas/sas05.pdf)

ACKNOWLEDGMENTS

The authors would like to thank Dr. Margaret Bogle and the Agricultural Research Service, United States Department of Agriculture, Project No. 6251-53000-002-00D, the Delta NIRI project, for their support The authors would also like to recognize the support and encouragement of Janice Gossett.

CONTACT INFORMATION

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

Jeff Gossett
University of Arkansas for Medical Sciences, Dept. of Pediatrics
1120 Marshall St
Little Rock, AR 72202
Work Phone: (501) 364-4960
Fax: (501) 364-1431
E-mail: gossettjeffreym@uams.edu

APPENDIX – BUILDING NHANES DATASETS 1999-2000 AND 2001-2002

1. Download data from the NHANES web site: http://www.cdc.gov/nchs/nhanes.htm. The 1999-2000 data and 2001-2002 are downloaded separately as SAS transport files. The 2001-2002 data release page: http://www.cdc.gov/nchs/about/major/nhanes/NHANES01_02.htm contains further details.
2. Extract the data sets and merge as needed. See instructions with SAS code: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHANES/NHANES2001-2002/examrgcd.txt.
3. Merge data sets by sequence number (SEQN).
4. Sort data by STRATA (SDMVSTRA) AND PRIMARY SAMPLING UNIT (SDMVPSU) prior to analysis;

```
** 1999 TO 2000 DATA **;
LIBNAME A1 XPORT 'C:\NHANES9900\DATA\BMX.XPT';
LIBNAME B1 XPORT 'C:\NHANES9900\DATA\BPX.XPT';
LIBNAME C1 XPORT 'C:\NHANES9900\DATA\DEMO.XPT';
LIBNAME D1 XPORT 'C:\NHANES9900\DATA\LAB13.XPT';
DATA N9900;
  MERGE A1.BMX B1.BPX C1.DEMO D1.LAB13;
  BY SEQN;
  KEEP BMXARMC BPAARM BPACSZ RIAGENDR RIDAGEEX RIDAGEMN RIDAGEYR RIDRETH1 RIDRETH2
RIDSTATR SDDSRVYR SDMVPSU SDMVSTRA SEQN WTMEC4YR LBDHDL LBDHDLSI LBDTCSI LBXTC;
RUN;
** 2001 TO 2002 DATA **;
LIBNAME A2 XPORT 'C:\NHANES0102\DATA\BMX_B.XPT';
LIBNAME B2 XPORT 'C:\NHANES0102\DATA\BPX_B.XPT';
LIBNAME C2 XPORT 'C:\NHANES0102\DATA\DEMO_B.XPT';
LIBNAME D2 XPORT 'C:\NHANES9900\DATA\L13_B.XPT';

DATA N0102;
  MERGE A2.BMP_B_R B2.BPX_B F2.DEMO_B D2.L13_B;
  BY SEQN;
KEEP BMXARMC BPAARM BPACSZ RIAGENDR RIDAGEEX RIDAGEMN RIDAGEYR RIDRETH1 RIDRETH2
RIDSTATR SDDSRVYR SDMVPSU SDMVSTRA SEQN WTMEC4YR LBDHDL LBDHDLSI LBDTCSI LBXTC;
  RUN;
PROC FORMAT;
  VALUE BPACSZF 1="INFANT(6X12)" 2="CHILD(9X17)" 3="ADULT(12X22)" 4="LARGE(15X32)"
5="THIGH(18X35)";
  VALUE RACEF 0="TOTAL" 1="NON-HISPANIC WHITE" 2="NON-HISPANIC BLACK" 3="MEXICAN
AMERICAN" 4="OTHER" 5="OTHER HISPANIC";
  VALUE SEXF 1='M' 2='F' 0="TOTAL";
  VALUE AGEF 1=”0-19” 2="20-39" 3="40-59" 4="60+";
RUN;
**CONCATENATE THE 1999-2000 AND 2001-2002 DATA SETS.
DATA ALL;
  SET N9900(IN=I) N0102(IN=J);
  ADULT=2-(RIDAGEYR>19);  ** 1=ADULT, 2=CHILD **;
  RENAME RIDRETH2=RACE RIAGENDR=SEX;
  AGE1=1+(RIDAGEYR>39)+(RIDAGEYR>59);
  AGE2=1+(RIDAGEYR>19)+(RIDAGEYR>39)+(RIDAGEYR>59);
LABEL LBXTC='SERUM TOTAL CHOLESTEROL'
      AGE1='AGE GROUP'
      HIGHTC='>=240 MG/DL';
IF LBXTC>=240 THEN HIGHTC=100;
ELSE IF LBXTC^=. THEN HIGHTC=0;
HIGHCHOL=HIGHTC/100;
  FORMAT AGE2 AGEF. BPACSZ BPACSZF. RIDRETH2 RACEF. RIAGENDR SEXF.;
RUN;
```