HCHS/SOL Analysis Methods at Baseline

September 2016
Version 4.0

Prepared by the
HCHS/SOL Coordinating Center
Collaborative Studies Coordinating Center
UNC Department of Biostatistics

Daniela Sotres-Alvarez
Marston Youngblood
Franklyn Gonzalez II
Sonia Davis
Jianwen Cai

(Former contributors:
Lisa M LaVange, Diane Catellier,
Annie Green Howard, William Kalsbeek)

This document is CONFIDENTIAL and for EXCLUSIVE use by HCHS/SOL investigators and NHLBI-NIH. Its purpose is to illustrate methods not to report results. Please send questions, suggestions and comments to Marston.Youngblood@unc.edu and dsotres@unc.edu

Analysis Methods for HCHS/SOL Baseline Data

Table of Contents

i. FORWARD

Note to Users of these Analysis Methods Guidelines

- This Guide is for illustration purposes in working with the HCHS/SOL datasets and has been developed using baseline data for the full cohort (n = 16,415) restricted to waves 1 and 2 (n=11,815).
- Included on the HCHS/SOL baseline examination datasets beginning with INV4 are three sampling weight variables (weight_final_norm_overall, weight_final_norm_center, weight_final_expanded), which are described in sections 1.2 to 1.4. All weights were calibrated to the age, gender and Hispanic/Latino background distributions from the 2010 US Census for the four study field centers.
- The document is not intended for direct citation.
- Statistical program output used in the examples in this Guide has been modified and/or formatted for presentation and clarity.
- Additional documentation for SAS 9.4 can be found at https://support.sas.com/documentation/onlinedoc/stat/ and for SAS 9.2 at: http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#titlepage.htm and for SAS 9.3 at: http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#titlepage.htm

MAIN Updates in Version 4.0 (Sept 2016)

- 2010 US Census age distribution included in output 1.5
- SAS code added in sections 5.4. 6.4. 9.4.7 to estimate age-adjusted prevalences
- NEW Chapter 12 on missing data
- NEW Chapter 13 on multiple comparisons

MAIN Updates in Version 3.0 (Sept 2013)

- Chapter 9 “adjusted and standardized prevalences” has been updated to compare the use of survey linear models and logistic models to estimate prevalences
- NEW Chapter 11 discusses recommendations on how to adjust for field center

MAIN Updates in Version 2.0 (Dec 2012)

- Sampling weights ‘weight_final_norm_**overall**’ and ‘weight_final_norm_**center**’ are introduced. Sampling weight ‘weight_final_norm’ is removed and dropped from use.
    - Section 1.2 is updated accordingly, and new section 1.4 added
- All programming examples are updated to use ‘**weight_final_norm_overall**’

MAIN Updates in Version 1.2 (Jun 2012)

- Central and South American Hispanic Background groups are separated
- HCHS/SOL Database Version 3.1 (June 2012; N=16,415) with final sampling weight variable (weight_final_norm) is used rather than HCHS/SOL Database Version 2.2 (August 2011; N=11,405) with weight_norm derived for interim analysis
- Subpopulation and domain analysis are used to restrict analysis to only include participants in waves 1 and 2
- SUDAAN code is now provided for generalized logit models in Section 8.3
- Section 10 now includes methods to calculate p-values for correlations as well as point estimates for Pearson correlation coefficients in SAS.
- SAS code using the LSMEANS statement, which is only available in SAS 9.3, is provided to calculate adjusted means directly in Section 7.2.1.

1. INTRODUCTION

1.1. Sample design

The HCHS/SOL cohort was selected through a stratified multi-stage area probability sample design (LaVange, Kalsbeek, et al., 2010) which is briefly described below. The community areas in each of the four field centers (Bronx, Chicago, Miami, and San Diego) were delineated by census tracts from the 2000 decennial census. Field centers purposively selected the tracts to be targeted for recruitment, and the **target population** for the study was then defined as all non-institutionalized Hispanic/Latino adults aged 18-74 years residing in the defined community areas. HCHS/SOL participants were selected using a probability sample design within these areas to provide a representative sample of the target population.

At the ***first* stage of sample selection**, a stratified, simple random sample of census block groups (BGs), which served as the primary sampling units (PSUs), was selected for each field center. Four strata were formed for PSU selection by cross-classifying block groups by two census-derived variables: socioeconomic status (SES) as measured by the proportion of persons in the 2000 census aged 25 years and older with at least a high school education (2 levels), and proportion of the population that reported being Hispanic/Latino (2 levels). Block groups in the 'high' Hispanic/Latino concentration strata were oversampled relative to block groups in the 'low' strata. The block group selection probabilities did not differ between 'low' vs. 'high' SES strata. Special strata were created for a subset of field centers as needed to target specific neighborhoods. A fifth and sixth stratum were added in Miami for areas of high Central and South American concentration and high Cuban concentration (Hialeah), respectively. In the Bronx, a fifth stratum was defined as a portion of a high-rise housing complex (named Co-op City) to provide additional income diversity, and two additional strata were appended after the study started to increase coverage. Therefore, a total of 21 strata were defined across the four field centers; however, the analysis datasets contain 20 strata, due to the fact that one stratum (low SES/low Hispanic concentration in Miami) contained no block groups. All block groups in the special strata were selected at the first stage with no additional stratification. Overall, 670 (72.4%) of the 925 block groups in the targeted community areas were selected for inclusion in the study.

At the ***second* stage**, separate stratified samples of household addresses in each of the sample PSUs were selected from lists of postal addresses stratified by 'Hispanic/Latino surname' versus all other. Addresses in the Hispanic/Latino surname stratum were oversampled to increase screening efficiency for Hispanics in the household sample. Overall, 127,213 addresses (sample frame) were selected for inclusion in the study. Selected households were screened for eligibility, where eligibility is defined as having at least one self-identified Hispanic/Latino household member aged 18-74 years. Two methods were used for over-sampling adults aged 45-74 years within households. Method 1, implemented at the start of the study, was designed to keep all households intact, with no sub-sampling at the person level. With this method, households in which all Hispanic/Latinos fell in the 45-74 year age range were selected with certainty (probability of selection = 1), while all other households were selected with probability < 1. Method 2, incorporated at staggered times across the field centers, divided a household into two clusters, one of adults aged 45-74 and one of adults 18-44. The 45-74 year clusters were

selected with certainty, while the 18-44 year clusters were selected with a probability < 1. This household member selection algorithm was designed to provide the target age distribution for the study while minimizing the amount of screening information required for households that may not be selected. Once adopted, it was used for person-level sampling for the remainder of the study. The selection probabilities of the 18-44 year clusters were closely monitored and adjusted as needed to provide the target age distribution.

An additional modification was made to the sample design early in the study. Rather than screen only those apartments selected into the sample in a multi-unit building, field centers were given the option to screen all apartments, provided there were 30 or fewer in the building. This 'multi-unit screening option' improved the efficiency of the sample in neighborhoods where only a small fraction of apartments yielded eligible Hispanic households. The disadvantage is that it increased the clustering of the sample and created an additional step in the calculation of sampling weights, described below.

The sample of postal addresses in each field center was randomly sub-sampled to form three waves of addresses corresponding to three years of recruitment. Each wave provides a representative sample of the target community area in each center, thereby enabling interim analyses of study data to be conducted for valid inference to the target population.

1.2. Sampling weights

The overall HCHS/SOL target population is defined as all Hispanic/Latino adults aged 18-74 years and residing in the target areas (defined by census block groups) across the four participating sites. Over-sampling at both stages of sample selection was used to increase the likelihood that a selected address yielded an eligible household with adults aged 45-74 years. As a result, participants included in the final HCHS/SOL cohort were selected with unequal probabilities of selection, and these probabilities need to be taken into account during data analysis to appropriately represent the target population. The use of **sampling weights that reflect unequal probabilities of selection must be incorporated in all analyses to calculate appropriate estimates of population characteristics and their corresponding standard errors.** HCHS/SOL sampling weights are described in a Technical Report, and briefly summarized here**.** The sampling weights are the product of a 'base weight' and three adjustments: 1) multiplicative adjustments for differential non-response at the household and person level made relative to the sampling frame, 2) trimming to handle extreme values of weights and 3) a second multiplicative adjustment to calibrate the weights to known population distributions.

The base weights are calculated as the product of the reciprocals of the probabilities of selection at each stage of sampling, namely, selection of block groups, of households within block groups, and of individuals within households. These base weights are then adjusted for differential non-response at both the household and person-level. Non-response adjustment factors are defined as the reciprocal of an estimate of the probability that a sample household agrees to be screened and to participate in the study, and the probability that a person selected into the sample agrees to participate and completes the clinic exam. The non-response adjusted weights are then trimmed to reduce the variability of the weights as well as the impact of extremely large weights on estimation. The final sampling weights that accompany the release of the baseline examination data for the full

cohort (all three waves) are calibrated to the 2010 US Census Population[1] according to age, sex and Hispanic background. The non-response adjusted, trimmed, and calibrated sampling weight for each sample respondent is referred to as its expanded sampling weight. The sum of expanded sampling weights overall and for a subgroup of interest equals the estimated number of persons in the target population and subgroup thereof.

In addition to the expanded sampling weight, a sampling weight normalized to the overall HCHS/SOL cohort sample size is also calculated. The normalization procedure corresponds to multiplying all weights by a constant factor such that the sum of weights equals the cohort sample size (16,415), and the average weight equals 1. The normalized weight is primarily intended for use when analyzing data from all four field centers combined. Its use will avoid errors that can result if the statistical software package used for the analysis is not designed for probability sampling. In particular, incorrect standard errors and degrees of freedom for test statistics can result with use of the expanded weights in some software packages.

A sampling weight that is normalized to each field center sample size separately is also provided. The center-specific normalization procedure corresponds to multiplying each weight in the center by a constant factor such that the sum of weights in the center equals the center sample size, and the average weight across all persons in the center equals 1. With this procedure, the constant factors differ from center to center. Center-specific normalized weights are appropriate for use whenever data from single centers are being analyzed separately, or when centers are being compared with respect to certain characteristics. When used for analysis of pooled center data, the target population parameter being estimated corresponds to the average parameter across the four field center target populations. Further discussion of the three types of weights and an explanation for choosing the appropriate sampling weight for an analysis is provided in section 1.4.

To illustrate the importance of using sampling weights in estimating characteristics of the target population, consider the impact of oversampling individuals over the age of 45 at the last stage of sample selection. As a result of this oversampling, the distribution of the HCHS/SOL sample is older than the actual target population. Overall, 59.2% in the HCHS/SOL sample are aged 45-74 years, whereas the percentage in the target population aged 45-74 years is 40.2%, based on the 2010 US Census. The weighted estimate yields the population percentage exactly, due to the fact that the final weights were jointly calibrated to the age and gender distribution of the target population as estimated by the 2010 US Census.

[1] Note, that for interim analyses (waves 1 and 2) weights were calibrated to the US 2008 American Community Survey (ACS) because at the time 2010 US Census was not available. Because the calibration data source at interim (2008 ACS) only approximated the target population for the study, the sampling weights were calibrated to population percentages by age and gender and not to population total counts of persons. Consequently, the sampling weights were standardized to sum to the sample size in each field center.

Impact of weights in HCHS/SOL due to oversampling adults 45 and over

	Unweighted Frequency	*Unweighted Percentage*	*Weighted Percentage**
AGEGROUP_C2	*N*	*Percent*	*Percent*
Age 18-44	6701	40.82	59.78
Age 45+	9714	59.18	40.22
	16,415	100.00	100.00

*Based on expanded or normalized overall weight

1.3. Analysis accounting for sample design

Unequal weighting, stratification, and cluster sampling can all impact analysis of data arising from a complex sample design such as that employed for HCHS/SOL. The impact of sampling weights on point estimates of population characteristics was described in section 1.2 for cases in which sampling weights reflect differential probabilities of selection due to oversampling certain segments of the population (e.g., persons aged 45 – 74 years). In addition to point estimation, probability sample designs can impact other aspects of data analysis. Unequal weighting and cluster sampling (at both the BG and household levels) tend to increase the variability of population estimates and reduce the power available for statistical tests, while stratification has the reverse effect. For valid inference to the target population, all three aspects of the HCHS/SOL sample design need to be taken into account during data analysis. Failure to do so typically results in over-stating both the precision of estimates and the statistical significance of hypothesis tests. Several statistical software packages can accommodate complex survey data such as SAS, Stata, SPSS, SUDAAN, R, and Mplus. There are several references (Brogan DG, 1998, Encyclopedia of Biostatistics; Siller and Tompkins, SUGI 31) that have compared these packages in terms of estimates, capabilities, ease-of use, cost, etc.

The aim of this guide is to illustrate different statistical analysis procedures accommodating the sample design of HCHS/SOL using mainly SAS and SAS-callable SUDAAN and, when convenient, showing code for other statistical packages. The data used in this guide (n=11,815) are from participants from all four field centers in waves 1 and 2 (for confidentiality) of the study sample who provided informed consent with non-missing sampling weight.

In most statistical packages requesting weighted estimates while ignoring cluster sampling and stratification will give correct point estimates of means and percentages, regression model coefficients, and percentiles in the target population, but it will not give correct standard errors, confidence intervals, or p-values for hypothesis testing. Hence, to fully account for the HCHS/SOL sample design we need to use software intended for complex sample surveys in which weights, cluster sampling, and stratification are all considered in producing analysis findings. Special SAS procedures (e.g., SURVEYMEANS, or other SAS procedures that start with the word ‘survey’) and special software packages (e.g., SUDAAN) are available for this purpose. The design is specified in each of these procedures through specification of the sampling weight, the first-stage stratum id, the cluster id, and the ‘with replacement’ option for variance computations. Although the HCHS/SOL sample was selected without replacement at each stage (i.e., a BG or a

household address was not eligible to be selected more than once), computing variance estimates assuming with replacement sampling of BGs at the first stage of selection provides a conservative estimate of variances and covariances that incorporate correlations at all subsequent stages of selection (LaVange, Koch, and Schwartz, 2001). There may be some loss of efficiency with the use of with replacement estimates of variance, but this trade-off is usually acceptable given the lack of information about appropriate correlation structures that would need to be specified if without replacement variances were to be computed. This loss of efficiency yields conservative estimates in that variances and p-values are both somewhat larger under this assumption.

SAS statements to specify HCHS/SOL sample design

Statistical analysis using HCHS/SOL sample data must account for the complex sample design by specifying PSU strata (STRAT), primary sampling unit (PSU_ID) and sampling weights (e.g., WEIGHT_FINAL_NORM_OVERALL). In SAS, this is done by including three statements in all survey procedures: STRATA, CLUSTER and WEIGHT.

```
strata Strat;
cluster PSU_ID;
weight Weight_Final_Norm_Overall;
```

SAS-callable SUDAAN statements to specify HCHS/SOL sample design

In SUDAAN, a complex sample design is specified using two statements and two additional options. The PSU strata (STRAT) and primary sampling unit (PSU_ID) variables are included on the NEST statement, and all analyses must use sorted data by the NEST variables. Sampling weights (WEIGHT_FINAL_NORM_OVERALL) are specified in the WEIGHT statement. In addition, in the PROC statement the DESIGN=WR option must be used to indicate a stratified with replacement design. The FILETYPE option (e.g., FILETYPE=SAS) must also be specified in the PROC statement.

```
nest Strat PSU_ID;
weight Weight_Final_Norm_Overall;
```

1.4. Selecting the appropriate sampling weight in an analysis

There are three sampling weights in the baseline data release (INV4, December, 2012; n=16,415): weight_final_norm_overall, weight_final_expanded, and weight_final_norm_center.

Weight_final_norm_overall is the normalized weight that sums to the number of study participants from all 4 field centers (16,415). The analyst should use weight_final_norm_overall for almost all purposes when conducting analyses using the HCHS/SOL data from all 4 field centers combined, so that the degrees of freedom from the sum of the weights is *not inflated* when conducting statistical tests of significance. If it is important for the user to estimate the *number of people* with a particular characteristic or outcome from the target population areas, then **weight_final_expanded** should be used. These weights basically expand the study cohort to the 2010 Census based estimate of the number of people in the target area. When estimating means, proportions, and regression coefficients, both the expansion weights and the overall normalized weights will yield identical point estimates. However, incorrect standard error estimates and degrees of freedom values may result from the use of expansion weights in some statistical

software packages. Normalized weights will provide the correct degrees of freedom for tests with all software packages and are therefore preferred.

Weight_final_norm_center is the normalized weight that sums to the number of study participants within its field center. This sampling weight variable should be used when data from one field center are used for analysis, or in some ancillary studies where only one field center's sample is involved (e.g. Vision HCHS/SOL Ancillary Study conducted only at Miami field center). For example, a paper by researchers at the San Diego site who want to understand phenomena relevant to the San Diego target population alone should use the center-specific normalized weight for the San Diego center. Center-specific normalized weights may also be of interest when comparing estimates of population parameters across field centers. Use of these weights when analyzing data from the four centers combined will produce estimates of population characteristics that are averaged across the four target populations.

To understand the differences in the use of the two normalized weights, note that the design of the HCHS/SOL is different from the traditional survey sample design. In the traditional survey sample design, there is one target population (e.g., all U.S. Hispanics/Latinos) and the sample design is centered on drawing representative samples from this one target population. The HCHS/SOL has essentially four target populations, one for each field center, and the sample selection process varied somewhat from center to center. Use of the center-specific sampling weights to calculate a prevalence rate produces unbiased estimates of the prevalence *in each center's target population*. When data from the four locales are pooled for analysis, the resulting prevalence estimate is unbiased for the average prevalence across the four field centers. This population parameter in essence represents a simple average, as if each center represented exactly ¼ of the overall target population since the respondent sample sizes for each center were about the same. *This is not optimal for combined samples.* In contrast, use of the overall normalized weight provides estimates of prevalence in the overall target population, defined as the union of the four field center target populations. That is, use of the overall normalized sampling weight provides an estimated prevalence that in essence represents the weighted average of the prevalence in each of the four target populations.

Output 1.4 provides a comparison of informative statistics for the three sampling weights and illustrates the effect of the different normalization constants.

In summary, **weight_final_norm_overall** should be used when analyzing data from all four field centers combined and **weight_final_norm_center** should be used when analyzing data from only one field center or when comparing parameters across centers.

Output 1.4 Summary of the three sampling weights

Center-specific normalized (weight_final_norm_center)	Bronx	Chicago	Miami	San Diego	Overall
Normalization Constant	0.0217	0.0400	0.0213	0.0240	---
Unweighted N	4118	4134	4077	4086	16415
Sum of Weights	4118	4134	4077	4086	16415
Mean of Weights	1	1	1	1	1
Std. Dev. of Weights	1.1579	0.9233	0.7112	1.2962	1.0463
Overall normalized (weight_final_norm_overall)	Bronx	Chicago	Miami	San Diego	Overall
Normalization Constant	---	---	---	---	0.0250
Unweighted N	4118	4134	4077	4086	16415
Sum of Weights	4760	2591	4803	4261	16415
Mean of Weights	1.1560	0.6267	1.1781	1.0427	1
Std. Dev. of Weights	1.3386	0.5787	0.8380	1.3516	1.101
Expanded weights (weight_final_expanded)	Bronx	Chicago	Miami	San Diego	Overall
Normalization Constant	1	1	1	1	1
Unweighted N	4118	4134	4077	4086	16415
Sum of Weights	190079	103452	191794	170121	655446
Mean of Weights	46.1581	25.0247	47.0429	41.6351	39.9297
Std. Dev. of Weights	53.4485	23.1053	33.4592	53.9692	43.9639

1.5. Age distribution of HCHS/SOL

When we are interested in comparing outcomes that are related to age, it is important to note that there are substantial differences in the age distribution among Hispanic backgrounds. Cubans are on average older and Mexicans are on average younger. This can be seen in Output 1.5 which gives the weighted age proportion in each age group for each Hispanic background. Therefore, when comparing these outcomes across Hispanic background, it is very important to account (adjust) for age. Otherwise, outcome differences might be due solely to differences in the age distribution among backgrounds and not due to background.

Output 1.5 US 2010 and 2000 Census age distribution and HCHS/SOL estimated target population age distribution by Hispanic/Latino background

Age Group	***US 2010 (Standard Population)***	***US 2000 (Standard Population)***	*HCHS/SOL Target Population*	*Dominican*	*Central Americans*	*Cubans*	*Mexicans*	*Puerto-Ricans*	*South Americans*
18-29	**0.2396835934**	**0.235800444**	0.2719	0.20177	0.17447	0.10652	0.17331	0.14301	0.12605
30-39	**0.1858332065**	**0.222616766**	0.2106	0.11821	0.18024	0.10865	0.16636	0.10964	0.16433
40-49	**0.2018408995**	**0.225788538**	0.2180	0.25476	0.24668	0.27269	0.26592	0.23102	0.26704
50-59	**0.1942642657**	**0.162064749**	0.1609	0.26563	0.27210	0.28718	0.24861	0.29080	0.27544
60-69	**0.1354254551**	**0.107135420**	0.1065	0.13383	0.10687	0.18023	0.12059	0.18188	0.14379
70-74	**0.0429525799**	**0.046594083**	0.0321	0.02582	0.01964	0.04474	0.02520	0.04364	0.02334
Total	**1.00000**	**1.00000**	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

Because the age distributions in HCHS/SOL and NHANES are different, we need to use a standard age distribution to make comparisons between the two studies. External standardization involves applying a standard age distribution to one or more populations to eliminate differences in age distribution as a reason for differences in rates among populations. In sections 5.4, 6.4 and 9.4.7, we illustrate how to externally standardize the HCHS/SOL population to the US 2000 population age distribution (Klein RJ, Schoenborn CA, 2001), which is the age distribution used by current published NHANES results. **However, the United States 2010 Decennial Census age distribution data are now available so we recommend to standardize HCHS/SOL to this newer standard distribution (first column in output 1.5).** Note that HCHS/SOL sampling weights are calibrated (age, gender and Hispanic/Latino background) to the US 2010 Census within the specific HCHS/SOL target areas, whereas conducting external age standardization to the US 2010 Census refers to the United States age distribution. However, note that HCHS/SOL estimates, after external standardization to the US 2010 age-distribution, do not generalize to the entire US Hispanic/Latino population, but rather to the Hispanic/Latino population living in the target areas had they followed the same age-distribution as from the US 2010 Census. The choice of which population to standardize depends on the population that one wishes to compare to the HCHS/SOL population. Because the NHANES population and the standardized 2000 population distribution included individuals both under the age of 18 and over the age of 74 and HCHS/SOL did not, we modified the age distribution used in NHANES to match our protocol. We used published population distributions #6 and #10 in Table 2 of Klein and Schoenborn (2001) to create a set of age standardization ranges. Distribution #6 provided the estimate for the population size 75 years and older (n=16,574). We truncated the age range of distribution #10 using values for ages 70-79 (n=16,141), and 80 and above (n=9,159). By difference, our estimated population 2000 comparison age range of 70-74 is 8,726 thousand people (e.g. estimated people aged 70 to 74 is equal to 16,141 plus 9,159 minus 16,574).

1.6. Subpopulation or domain analysis

For an analysis of a particular subpopulation or *domain* (e.g., men and women) one must use the SUBPOPN statement in SUDAAN or the DOMAIN statement in SAS. SAS version 9.1 or later allows domain analyses for the SURVEYMEANS and SURVEYFREQ procedures, but domain analyses in modeling procedures (SURVEYREG and SURVEYLOGISTIC) are only available in SAS version 9.2 or later. **Simply using a subset of the data file where the observations you wish to exclude have been previously deleted or using 'BY' or 'WHERE' clause in a procedure will not produce the right estimates of standard errors.** The subpopulation and domain statements assume that even if there are no observations in a primary sampling unit in the sample there may be some in the target subpopulation and hence we need to include its appropriate contribution to the variance. Stata, Mplus, and R also have the capability of performing subpopulation analyses; example code has been provided in sections 5 and 6.

Rather than excluding subjects from the analysis population, when analyzing this data, one should use the subpopulation or domain statements to restrict the analysis to those who meet the inclusion criteria. This can be done by generating a flag to denote included subjects and then the use of this flag on the subpopulation or domain statement. By flagging these subjects and using subpopulation and domain statements, these 'excluded' participants are still assumed to be a part of the target population and therefore contribute to the variance calculations.

It is important to note that when using the DOMAIN statement in SAS, results are given for all values of the variable given. Therefore, when using a flag, results are given separately for both those who meet the inclusion criteria and for those who do not meet the inclusion criteria. One must take care to select the appropriate results.

For example, when analyzing sleep data, one need to restrict the analysis to subjects who have sleep monitor data and at least 30 minutes of valid sleep data. An example dataset used for this analysis is shown below. The variable KEEP_MS13 is used as a way of flagging subjects that should be included in the analysis.

```
DATA MS13;
 MERGE PART_DERV(in=in_der KEEP=ID)
       slpa (keep=id slpa12 slpa15 slpa30 slpa54 slpa97 slpa121 in=in_slpa);
  BY id;
  IF in_der;
  IF in_SLPA then FLAG_sleepstudy = 1;
  else FLAG_sleepstudy = 0;
  label FLAG_sleepstudy = 'Subject has Sleep Study Data (In_SLPA): 0=No,1=Yes';
  if slpa30 > .z then FLAG_VSHORT = (slpa30 < 0.5);
  label FLAG_VSHORT = "Very short study: recording time < 0.5 hr";
  *- KEEP_MS13: -*;
  KEEP_MS13 = (FLAG_SLEEPSTUDY=1 and FLAG_VSHORT=0);
  Label KEEP_MS13 = "FLAG MS#13 participants to include";
 run;
```

In this document to illustrate the use of the subpopulation analyses and, for confidentiality reasons, we are restricting the analysis to only subjects in Waves 1 and 2. As such, we have created a flag, KEEP_DATA, which is an indicator set to one if the participant is in Waves 1 or 2 and zero otherwise.

2. GENERAL METHODS SECTIONS FOR PAPERS

2.1. Guidelines for describing the HCHS/SOL sample design

Manuscripts should include the strategy used to take into account the complex sample design used to select participants for the HCHS/SOL cohort. In particular, statistical methods to incorporate stratification factors, sampling weights, and clustering (particularly of persons within the same household) in the data analysis should be described. The following description of the HCHS/SOL cohort sample design can be paraphrased for use in manuscripts.

The HCHS/SOL is a prospective study which enrolled 16,415 participants in four communities in the United States from diverse cultural and genetic origins who self-identify as Hispanic/Latino. Participants in the HCHS/SOL self-identify their background (or their families) as Cuban, Dominican, Puerto Rican, Mexican, Central American, or South American (with country specified). Recruitment was implemented through a two-stage area household probability design (LaVange, Kalsbeek, et al 2010). This article includes participants who attended the HCHS/SOL field center baseline examination and have sample weights and values for the variables analyzed.

2.2. Guidelines for using race variable

A high proportion of participants (40% based in waves 1 and 2) did not respond to the question about race. This variable should be avoided or used under very limited circumstances.

3. SAMPLE DATA SET

The dataset used in this document is HCHS/SOL Baseline Database (INV4, n=16,415) and analyses are restricted to only include participants in waves 1 and 2 (n=11,815) for confidentiality reasons. All the variables used to illustrate the analysis methods are in the participant derived file with the exception of sbpa5, female (a numeric version of gender variable), diabetes3_C2 (an indicator variable for diabetic vs. not diabetic, based on diabetes3), KEEP_DATA and hypert=100*hypertension2. KEEP_DATA is an indicator variable (one if the participant is in wave 1 or 2 and zero otherwise) that will be used in the subpopulation (domain) statement. The numeric version of gender was created since SUDAAN only allows for numeric variables. Below we present type and brief description for all variables, and corresponding format level values for categorical variables. All code and output in this guide used SAS-callable SUDAAN v. 10 and SAS v. 9.3.

Variable	*Type*	*Label*
ID	Num	Subject ID
CENTER	Char	Center (‘B’ = Bronx, ‘C’ = Chicago, ‘M’ = ‘Miami’, ‘S’ = ‘San Diego’)
CENTERNUM	Num	Numeric Center (1=Bronx, 2=Chicago, 3=‘Miami’, 4=‘San Diego’)
KEEP_DATA	Num	Flag used for inclusion in analysis document
STRAT	Num	Stratification Variable ID
PSU_ID	Num	Primary Sampling Unit ID
COHORT	Num	Cohort Year (from Household Address Wave Assignment)
WEIGHT_FINAL_NORM_OV	Num	Sampling weight normalized to the overall sample size (N=16,415)
AGE	Num	Age (Continuous)
AGEGROUP_C2	Num	2 Level Age Group (1 = 18-44, 2 = 45+)
AGEGROUP_C6_NHANES	Num	6-Level NHANES standardization Age Groups
Female	Num	Gender (1 = Female, 0 = Male)
GENDER	Char	Gender (F=Female, M=Male)
BKGRD1_C7	Num	7-level re-classification of Hispanic/Latino Background
site_bkgrd	Num	Center/Hispanic Background Cross-Classification (collapsed
Dominican	Num	Indicator variable for Dominican
Central	Num	Indicator variable for Central Americans
Cuban	Num	Indicator variable for Cubans
Mexican	Num	Indicator variable for Mexicans
Puerto_Rican	Num	Indicator variable for Puerto Ricans
South	Num	Indicator variable for South Americans
Other	Num	Indicator variable for Other Hispanic Background
BMI	Num	BMI (kg/m^2)
BMIGRP_C4	Num	4-level grouped Body Mass Index – WHO (1=Underweight,
CIGARETTE_USE	Num	Cigarette_Use (1=Never,2=Former,3=Current)
DIABETES3_C2	Num	Diabetes (1=Diabetic (Diabetes3=3), 0 = Not Diabetic)
DIABETES3	Num	3-level grouped Diabetes – includes self-report
HYPERTENSION2	Num	Hypertension using NHANES definition (0=No, 1=Yes)
Hypert	Num	Hypertension2*100
SASH_ALL	Num	Short acculturation scale
SBPA5	Num	Average Systolic (SBPA5)

NOTE: Variables for SAS analysis are identical to those used in SUDAAN v. 10. In SUDAAN 9 and later versions, independent categorical variables with levels coded as 0 may be placed on the CLASS statement without being considered as missing values. However, if you use any prior version of SUDAAN 9, you must use the SUBGROUP and LEVELS statements instead in which level 0 is considered as a missing value and excluded from the analysis. Therefore, in previous versions of SUDAAN when using the SUBGROUP statement the m levels of the variable had to be the consecutive integers 1, 2,…, m.

4. WEIGHTED QUANTILES

4.1. SBP Percentiles

4.1.1. SAS

Estimating quantiles for subpopulations, using the domain statement, is not available in SAS version 9.2, but it is in SUDAAN. For this reason, we suggest using alternative software to estimate percentiles.

4.1.2. SUDAAN

In SUDAAN, a complex sample design is specified using two statements and two additional options. The strata (strat) and primary sampling unit (psu_id) variables are included on the NEST statement, and all analyses must use sorted data by the NEST variables. Sampling weights (weight_final_norm_overall) are specified in the WEIGHT statement. In addition, in the PROC statement the DESIGN=WR option must be used to indicate a stratified with replacement design. The FILETYPE option (e.g., FILETYPE=SAS) must also be specified in the PROC statement. **WARNING:** The dataset must be ordered (sorted) by the variable(s) in NEST statement.

In this document we are restricting the analysis to include only participants in waves 1 and 2 for confidentiality reasons. Hence, to correctly specify the sample design we use the flag KEEP_DATA (indicator variable set to 1 if the participant is in Wave 1 and 2 and zero otherwise) in the subpopulation statement (see section 1.6).

The DESCRIPT procedure produces descriptive statistics for continuous and categorical variables. These statistics include means, geometric means, medians and other quantiles, percentages, and their standard errors. Estimating quantiles for subpopulations is available in SUDAAN using 1) the statement SUBPOPN directly, or 2) the TABLE statement with the subpopulation variable. The SUBPOPN is more useful when focus is one particular subpopulation whereas using the TABLE statement is useful when all subpopulations are of interest. For example, code to estimate deciles of SBP for women, men and overall is:

```
proc sort data=SOL; by strat psu_id; run;
proc descript data=sol filetype=sas design=wr;
    nest strat PSU_ID;
    weight weight_final_norm_overall;
    var sbpa5;
    percentile 10 20 30 40 50 60 70 80 90;
    subpopn female = 1 & KEEP_DATA = 1 / name="Female";
run;
/*Code below produces same results as PROC DESCRIPT above but for both genders*/
proc sort data=SOL; by strat psu_id; run;
proc descript data=sol filetype=sas design=wr;
    nest strat PSU_ID;
    weight weight_final_norm_overall;
    class female; /*VARIABLE IN TABLE NEEDS TO BE INCLUDED IN CLASS STATEMENT*/
    var sbpa5;
    table female; /*USE TABLE STATEMENT TO GET DECILES BY GENDER*/
    subpopn KEEP_DATA=1;
    percentile 10 20 30 40 50 60 70 80 90;
```

```
run;
```

Output 4.1.2. SBP percentiles by gender, PROC DESCRIPT with TABLE statement

1-female 0-male	Percentiles	N	Qtile	SE
1	10.00	7054	96.33	0.43
	20.00	7054	101.27	0.48
	30.00	7054	105.09	0.46
	40.00	7054	108.71	0.60
	50.00	7054	112.33	0.53
	60.00	7054	116.88	0.64
	70.00	7054	121.86	0.63
	80.00	7054	129.19	0.66
	90.00	7054	141.29	0.79

5. WEIGHTED MEANS

In this section we illustrate how to estimate 1) unadjusted means and other descriptive statistics accounting appropriately for the study design, and 2) standardized means to an external population. Section 7 illustrates how to estimate adjusted means using linear models.

5.1. SBP Mean

5.1.1. SAS

Statistical analysis using HCHS/SOL data must account for the complex sample design by specifying strata (strat), primary sampling unit (psu_id), and sampling weights (weight_final_norm_overall). In SAS, this is done by including three statements in all survey PROCEDURES: STRATA, CLUSTER and WEIGHT. **In this document, we are restricting the analysis to include only participants in waves 1 and 2 for confidentiality reasons. Hence, to correctly specify the sample design we use the flag KEEP_DATA (indicator variable set to 1 if the participant is in waves 1 and 2 and zero otherwise) in the domain statement (see section 1.6).**

The procedure SURVEYMEANS estimates means, standard errors, p-values, confidence limits, and other descriptive statistics that appropriately account for the study design.

```
proc surveymeans data = sol nobs mean stderr;
   strata strat; cluster PSU_ID; weight weight_final_norm_overall;
   domain KEEP_DATA;
   var sbpa5;
run;
```

Output 5.1.1. Overall SBP mean, PROC SURVEYMEANS

KEEP_DATA	*Variable*	*Label*	*N*	*Mean*	*Std Error of Mean*
1	*SBPA5*	Average Systolic (SBPA5)	11806	119.876203	0.257031

WARNING: SAS will output all levels of the variable specified in the DOMAIN statement (in this case for KEEP_DATA = 0 and 1) and one must make sure to select the correct output.

5.1.2. SUDAAN

The next group of SUDAAN statements invokes DESCRIPT to produce means, standard errors, p-values, confidence limits, and other descriptive statistics that appropriately account for the study design.

WARNING: The analysis dataset must be ordered (sorted) by the variable(s) in NEST statement.

```
proc descript data=sol filetype=sas design=wr;
   nest strat PSU_ID; weight weight_final_norm_overall;
   subpopn KEEP_DATA=1;
   var sbpa5;
run;
```

Output 5.1.2. Overall SBP mean, PROC DESCRIPT

Variable	SUDAAN Reserved Variable One	Sample Size	Mean	SE Mean
Average Systolic (SBPA5)	1	11806.000	119.876	0.257

5.2. Unadjusted SBP mean by background – WARNING

WARNING: When the outcome distribution differs by age and one is interested in comparing means by Hispanic background, we need to adjust for age to account for the difference in age distributions among Hispanic backgrounds (section 1.5).

5.2.1. SAS

The DOMAIN statement provides means within subpopulations. Variables in the DOMAIN statement can either be numeric (categorical or binary) or character. To request multiple subpopulation analyses, you simply list the multiple subpopulations separated by an asterisk (*). Output 5.2.1 shows the mean SBP by Hispanic background (subpopulations).

```
proc surveymeans data = sol nobs mean stderr;
   strata strat; cluster PSU_ID; weight weight_final_norm_overall;
   domain KEEP_DATA*bkgrd1_c7;
   var sbpa5;
   title "WARNING: These SBP means are not age-adjusted";
run;
```

Output 5.2.1. Unadjusted (by age) SBP mean by background, PROC SURVEYMEANS
WARNING: These SBP means are not age-adjusted.

KEEP_DATA	*7-level re-classification of Hispanic/Latino Background*	*Label*	*N*	*Mean*	*Std Error of Mean*
1	Dominican	Average Systolic (SBPA5)	1001	121.943918	0.663300
	Central Americans	Average Systolic (SBPA5)	1369	120.782512	0.674074
	Cuban	Average Systolic (SBPA5)	1668	123.826186	0.510905
	Mexican	Average Systolic (SBPA5)	4618	116.730523	0.396191
	Puerto Rican	Average Systolic (SBPA5)	1954	121.839646	0.580319
	South Americans	Average Systolic (SBPA5)	758	118.524935	0.835293
	Mixed/Other	Average Systolic (SBPA5)	370	117.070027	0.919016

5.2.2. SUDAAN

The TABLE statement provides means within subpopulations. The variable(s) used in the TABLE statement need to be declared in either the SUBGROUP (paired with the LEVELS statement) or CLASS statements. Variables in the SUBGROUP statement must be both numeric and non-zero and variables in the CLASS statement must be numeric. For example, the mean SBP (unadjusted by age) by Hispanic background is given by:

```
proc descript data=sol filetype=sas design=wr;
   nest strat PSU_ID; weight weight_final_norm_overall;
   subpopn KEEP_DATA=1;
   class bkgrd1_c7;
   tables bkgrd1_c7; var sbpa5;
   title "WARNING: These SBP means are not age-adjusted";
run;
```

Output 5.2.2. Unadjusted (by age) SBP mean by background, PROC DESCRIPT
WARNING: These SBP means are not age-adjusted.

Variable	**7-level re-classification of Hispanic/Latino Background**	**Sample Size**	**Mean**	**SE Mean**
Average Systolic (SBPA5)	Total	11738.000	119.839	0.256
	Dominican	1001.000	121.944	0.663
	Central American	1369.000	120.783	0.674
	Cuban	1668.000	123.826	0.511
	Mexican	4618.000	116.731	0.396
	Puerto Rican	1954.000	121.840	0.580
	South American	758.000	118.525	0.835
	Mixed/Other	370.000	117.070	0.919

NOTE: Because Hispanic background has missing data, the total SBP mean is for those with non-missing Hispanic background (n=11,738) instead of the mean for all participants with SBP data (n=11,806) as in Output 5.1.2.

To request specific mean differences, use the CLASS and CONTRAST statements. Variables included in the CONTRAST statement cannot be included in the TABLES statement.

```
proc descript data=sol filetype=sas design=wr;
   nest strat PSU_ID; weight weight_final_norm_overall;
   class bkgrd1_c7;
   subpopn KEEP_DATA=1;
   var sbpa5;
   contrast bkgrd1_c7=(0 1 -1  0 0 0 0)/ name="Central Am vs. Cuban";
   contrast bkgrd1_c7=(0 0  1 -1 0 0 0)/ name="Cuban vs. Mexican";
run;
```

NOTE: When using PROC DESCRIPT, the default order of all CLASS variables is based on the internal and unformatted version of the variables with ascending order. Hence, it is important to note the order of the CLASS variables when using CONTRAST and ESTIMATE statements.

Output 5.2.2.1. Mean SBP differences (unadjusted by age), PROC DESCRIPT
WARNING: These SBP mean differences are not age-adjusted.

SUDAAN Reserved Variable One	Sample Size	Cntrst Mean	SE Cntrst Mean
Central Am vs. Cuban	3037.000	-3.044	0.889
Cuban vs. Mexican	6286.000	7.096	0.645

To request all pairwise contrasts, simply use:

```
   pairwise bkgrd1_c7 / name ='Country of origin differences';
```

The variable included in the PAIRWISE statement, just as variables included in the CONTRAST statement, cannot be included in the TABLES statement.

Output 5.2.2.2. Mean SBP differences (unadjusted by age) for pairwise combinations with Central American, PROC DESCRIPT
WARNING: These SBP mean differences are not age-adjusted.

Country of origin differences: (Central American, Cuban)	3037.000	-3.044	0.889
Country of origin differences: (Central American, Mexican)	5987.000	4.052	0.781
Country of origin differences: (Central American, Puerto Rican)	3323.000	-1.057	0.875
Country of origin differences: (Central American, South American)	2127.000	2.258	1.157
Country of origin differences: (Central American, Mixed/Other)	1739.000	3.712	1.161

If you only want to display the results for a specific subpopulation, then you add that variable to the CLASS statement and that specific category in the SUBPOPN statement.

```
   subpopn KEEP_DATA=1 & agegroup_c2=1;
```

5.2.3. Stata

In STATA, we use the SVYSET statement to specify the sample design; we provide the primary sampling unit (psu_id), and specify the sampling weights (weight_final_norm_overall) and strata variable (strat).

```
use sol.dta

/* SET UP SURVEY DESIGN */
svyset psu_id [pweight=weight_final_norm_overall], strata(strat)
/* OVERALL POPULATION - USING SVY: MEAN */
svy, subpopn(keep_data): mean sbpa5, over(bkgrd1_c7)
```

5.2.4. R

In R, we use the SVYDESIGN function to specify the sample design; we use the options ID, STRATA and WEIGHTS to specify the primary sampling unit (psu_id), strata (strat), and sampling weights (weight_final_norm_overall), respectively. SVYBY function is used with SVYMEAN in the function call to request the mean by background.

```
# READ IN DATASET
sol <- read.csv("sol.csv", header = TRUE, sep = ",")
sol$BKGRD1_C7 <- as.factor(sol$BKGRD1_C7)

# SURVEY LIBRARY
library(survey)
# CREATE SURVEY DESIGN OBJECT
hchs.dsgn <- svydesign(id=~PSU_ID, strata=~STRAT,
weights=~WEIGHT_FINAL_NORM_OVERALL, data=sol, nest=TRUE)
# OVERALL POPULATION - USING SVYMEAN
svyby(~SBPA5, ~KEEP_DATA+BKGRD1_C7, design=hchs.dsgn, svymean, na.rm=TRUE
keep.var=TRUE)
```

5.3. SBP mean by background stratified by agegroup_c2

5.3.1. SAS

Stratifying by age groups using the DOMAIN statement is one way to account for differences in age between Hispanic backgrounds. In section 7.3, we illustrate how to adjust means for age using linear models.

```
proc surveymeans data = sol nobs mean stderr;
   strata strat; cluster PSU_ID; weight weight_final_norm_overall;
   var sbpa5;
   domain KEEP_DATA*agegroup_C2*bkgrd1_c7;
run;
```

Output 5.3.1. SBP mean by background stratified by age group, PROC SURVEYMEANS

KEEP_DATA	*1(18-44),2(45+)*	*7-level re-classification of Hispanic/Latino Background*	*N*	*Mean*	*Std Error of Mean*
1	Age 18-44	Dominican	412	115.922804	0.726072
		Central Americans	637	115.173853	0.698029
		Cuban	547	114.809834	0.514392
		Mexican	2096	112.570554	0.403279
		Puerto Rican	669	115.275058	0.662234
		South Americans	290	111.260535	0.740990
		Mixed/Other	226	113.613565	0.891610
	Age 45+	Dominican	589	131.295384	0.800108
		Central Americans	732	131.046092	1.117760
		Cuban	1121	131.693581	0.574789
		Mexican	2522	125.326849	0.708775
		Puerto Rican	1285	129.756613	0.855703
		South Americans	468	127.209814	1.094239
		Mixed/Other	144	128.377139	2.519813

5.3.2. SUDAAN

To estimate SBP by age group, we use the TABLE statement in PROC DESCRIPT.

```
proc descript data=sol filetype=sas design=wr;
   nest strat PSU_ID;
   weight weight_final_norm_overall;
   class bkgrd1_c7 agegroup_c2;
   subpopn KEEP_DATA=1;
   tables agegroup_c2*bkgrd1_c7;
   var sbpa5; run;
```

Output 5.3.2. SBP mean by background stratified by age group, PROC DESCRIPT

Variable	1(18-44),2(45+)	Sample Size	Weighted Size	Mean	SE Mean
Age 18-44	Total	4877.000	7348.242	113.832	0.233
	Dominican	412.000	679.052	115.923	0.726
	Central American	637.000	634.618	115.174	0.698
	Cuban	547.000	1113.487	114.810	0.514
	Mexican	2096.000	3139.436	112.571	0.403
	Puerto Rican	669.000	1061.246	115.275	0.662
	South American	290.000	328.840	111.261	0.741
	Mixed/Other	226.000	391.563	113.614	0.892
Age 45+	Total	6861.000	4854.081	128.932	0.334
	Dominican	589.000	437.220	131.295	0.800
	Central American	732.000	346.795	131.046	1.118
	Cuban	1121.000	1276.101	131.694	0.575
	Mexican	2522.000	1519.249	125.327	0.709
	Puerto Rican	1285.000	879.964	129.757	0.856
	South American	468.000	275.056	127.210	1.094
	Mixed/Other	144.000	119.696	128.377	2.520

5.3.3. Stata

The next group of statements uses the Stata software to calculate unadjusted SBP mean by background stratified by agegroup_c2.

```
/* SET UP SURVEY DESIGN */
svyset psu_id [pweight=weight_final_norm_overall], strata(strat)

/* STRATIFIED – USING SVY: MEAN */
svy: mean sbpa5, over(agegroup_c2 bkgrd1_c7)
```

5.3.4. R

The next group of statements uses the R software to calculate unadjusted SBP mean by background stratified by agegroup_c2. We use the SVYMEAN with SVYBY function to request the mean calculated for each background and agegroup_c2.

```
# READ IN DATASET
sol <- read.csv("sol.csv", header = TRUE, sep = ",")
sol$BKGRD1_C7 <- as.factor(sol$BKGRD1_C7)
sol$AGEGROUP_C2 <- as.factor(sol$AGEGROUP_C2)

# CREATE SURVEY DESIGN OBJECT
hchs.dsgn <- svydesign(id=~PSU_ID, strata=~STRAT,
weights=~WEIGHT_FINAL_NORM_OVERALL, data=sol, nest=TRUE)

# STRATIFIED - USING SVYMEAN
svyby(~SBPA5, ~BKGRD1_C7*AGEGROUP_C2, design=hchs.dsgn, svymean, na.rm=TRUE)
```

5.4. Age-standardized SBP mean by background

Because the age distributions in HCHS/SOL and NHANES are different, we need to use a standard age distribution to make comparisons between the two studies. Here we illustrate how to estimate means standardizing the HCHS/SOL population to the US 2000 population age distribution (Klein RJ, Schoenborn CA, 2001), which is the age distribution used by NHANES.

As explained in section 1.5, HCHS/SOL sampling weights are calibrated (age, gender and Hispanic/Latino background) to the US 2010 Census within the specific HCHS/SOL target areas, whereas conducting external age standardization to the US 2000 Census refers to the United States age distribution. However, note that HCHS/SOL estimates after external standardization to the US 2000 age-distribution do not generalize to the entire US Hispanic/Latino population, but rather to the Hispanic/Latino population living in the target areas had they followed the same age-distribution as from the US 2000 Census.

5.4.1. SAS

PROC SURVEYMEANS in SAS 9.4 does not have a specific statement that can easily do external age standardization like SUDAAN does. However, we can apply the contrast statement in the SURVEYREG to adjust for external age standardization.

```
proc surveyreg data = SOL order=internal;
   strata strat;
   cluster psu_id;
   weight weight_final_norm_overall;
   domain KEEP_DATA;
   class bkgrd1_c7 agegroup_c6_nhanes;
   model sbpa5 = bkgrd1_c7 agegroup_c6_nhanes bkgrd1_c7*agegroup_c6_nhanes /
                 solution noint;
   /* NOTE this is 2000 US CENSUS AGE DISTRIBUTION */
   estimate 'Dominican' bkgrd1_c7 1 0 0 0 0 0 0
                        agegroup_c6_nhanes 0.235800444 0.222616766 0.225788538
                        0.162064749 0.10713542 0.046594083
                        bkgrd1_c7*agegroup_c6_nhanes 0.235800444 0.222616766
                        0.225788538 0.162064749 0.10713542 0.046594083
                        0 0 0 0 0 0
                        0 0 0 0 0 0
                        0 0 0 0 0 0
                        0 0 0 0 0 0
                        0 0 0 0 0 0
                        0 0 0 0 0 0 / e;
 run;
```

Output 5.4.1 SBP mean age-standardized to US 2000 population in Dominican, PROC SURVEYREG

		Estimate Standard			
Label	*Estimate*	*Error*	*DF*	*t Value*	*Pr > |t|*
Dominican	123.22	0.5404	644	228.00	<.0001

5.4.2. SUDAAN

External standardization can easily be implemented in SUDAAN by including age group variable (e.g. agegroup_c6_NHANES) in the CLASS and STDVAR statements and specifying the external age distribution (e.g. US 2000 population from Output 1.5) in the STDWGT statement.

```
proc descript data=sol filetype=sas design=wr;
   nest strat PSU_ID;
   weight weight_final_norm_overall;
   class Bkgrd1_c7 AGEGROUP_C6_NHANES;
   subpopn KEEP_DATA=1;
   var sbpa5;
   stdvar AGEGROUP_C6_NHANES;
/* 2010 US CENSUS AGE DISTRIBUTION IS HERE INSRTED AS A COMMENT
stdwgt 0.2396835934 0.1858332065 0.2018408995 0.1942642657 0.13542545510.0429525799;*/
/* 2000 US CENSUS AGE DISTRIBUTION */
stdwgt 0.235800444 0.222616766 0.225788538 0.162064749 0.10713542 0.046594083;
run;
```

Output 5.4.2. SBP mean by background age-standardized to US 2000 population, PROC DESCRIPT

Variable	7-level re-classification of Hispanic/Latino Background	Sample Size	Mean	SE Mean
Average Systolic (SBPA5)	Total	11729.000	120.445	0.197
	Dominican	1000.000	123.219	0.540
	Central American	1368.000	122.785	0.624
	Cuban	1667.000	121.385	0.380
	Mexican	4614.000	118.560	0.375
	Puerto Rican	1953.000	121.431	0.511
	South American	757.000	118.326	0.621
	Mixed/Other	370.000	120.378	0.961

These standardized SBP means are the expected SBP had each Hispanic background had exactly the same age distribution (US 2000 age distribution). Note that, compared to unadjusted SBP means from Output 5.2.2, the age-standardized SBP mean for Cubans is lower and for Mexicans is higher. Hence, the unadjusted mean difference of 7.1 between Cubans and Mexicans is reduced to 2.8 once we use a standard age distribution.

NOTE: In Output 5.4.2, the sample sizes are slightly smaller than in Output 5.2.2 because there is some missing data for age.

6. WEIGHTED PROPORTIONS

In this section we illustrate how to estimate 1) unadjusted proportions accounting appropriately for the study design, and 2) standardized proportions to an external population. Section 9 illustrates how to estimate adjusted proportions using linear models (design-based estimates).

6.1. Diabetes prevalence

6.1.1. SAS

The next group of SAS statements invokes SURVEYFREQ to produce frequency distributions for categorical variables that appropriately account for the study design. To request frequency distributions within subpopulations, you simply list the subpopulation(s) before the response variable separated by an asterisk (*). The default in SAS is to provide overall frequencies. To get percentages within each specific subpopulation, specify the ROW option in the TABLES statement.

```
proc surveyfreq data = sol;
   strata strat; cluster PSU_ID; weight weight_final_norm_overall;
   tables KEEP_DATA*DIABETES3_C2 / row cl;
run;
```

Output 6.1.1. Diabetes prevalence, PROC SURVEYFREQ

KEEP_DATA	DIABETES3_C2	Frequency	Row	Std Err of	95% Confidence Limits	
1	0	9396	84.1093	0.5218	83.0846	85.1339
	1	2405	15.8907	0.5218	14.8661	16.9154
Frequency Missing = 20						

6.1.2. SUDAAN

The next group of SUDAAN statements invokes CROSSTAB to produce frequency distributions for categorical variables that appropriately account for the study design. The categorical variable of interest is specified in the CLASS and TABLES statements. Variables in the CLASS statement must be numeric.

```
proc crosstab data=sol filetype=sas design=wr;
   nest strat PSU_ID; weight weight_final_norm_overall;
   subpopn KEEP_DATA=1;   class DIABETES3_C2;   tables DIABETES3_C2;
run;
```

Output 6.1.2. Diabetes prevalence, PROC CROSSTAB

DIABETES3_C2	Sample Size	Weighted Size	Tot Percent	SE Tot Percent
Total	11801.00	12256.99	100.00	0.00
0	9396.00	10309.26	84.11	0.52
1	2405.00	1947.73	15.89	0.52

6.2. Unadjusted diabetes prevalence by background – WARNING

WARNING: When the prevalence of an outcome differs by age and one is interested in comparing prevalences by Hispanic background, we need to adjust for age to account for the difference in age distributions among Hispanic backgrounds (section 1.5).

6.2.1. SAS

The prevalence of diabetes by Hispanic background is specified in the TABLES statement by crossing bkgrd1_c7 and DIABETES3_C2. The default in SAS is to provide overall frequencies (the 'percent' column in Output 6.2.1). To get percentages within each specific background group, specify the ROW option in the TABLES statement.

```
proc surveyfreq data = sol;
   strata strat; cluster PSU_ID; weight weight_final_norm_overall;
   tables KEEP_DATA*bkgrd1_c7*DIABETES3_C2 / row;
   title "WARNING: These diabetes prevalence estimates are not age-adjusted";
run;
```

Output 6.2.1. Diabetes prevalence by background, PROC SURVEYFREQ
WARNING: These prevalence estimates are not age-adjusted.

BKGRD1_C7	DIABETES3_C2	Frequency	Percent	Std Err of Percent	Row Percent	Std Err of Row Percent
Dominican	0	812	7.7158	0.6183	84.3550	1.4790
	1	189	1.4310	0.1817	15.6450	1.4790
Central American	0	1119	6.8891	0.5957	85.6385	1.3281
	1	251	1.1553	0.1174	14.3615	1.3281
Cuban	0	1361	16.3127	1.5754	83.3264	1.1536
	1	306	3.2642	0.3501	16.6736	1.1536
Mexican	0	3667	32.3754	1.5886	84.8009	0.7694
	1	952	5.8027	0.3676	15.1991	0.7694
Puerto Rican	0	1418	12.8215	0.7532	80.5560	1.2558
	1	539	3.0948	0.2498	19.4440	1.2558
South American	0	671	4.4899	0.3532	90.7359	1.3436
	1	87	0.4584	0.0653	9.2641	1.3436

6.2.2. SUDAAN

To estimate proportions within subpopulations, simply list the subpopulation(s) before the response variable separated by an asterisk (*) in the TABLES statement. Make sure all variables in the TABLE statement are specified in the CLASS statement.

```
proc crosstab data=sol filetype=sas design=wr;
   nest strat PSU_ID; weight weight_final_norm_overall;
   subpopn KEEP_DATA=1;
   class  bkgrd1_c7 DIABETES3_C2;
   tables bkgrd1_c7*DIABETES3_C2;
   title "WARNING: These diabetes prevalence estimates are not age-adjusted";
run;
```

Output 6.2.2. Diabetes prevalence by Hispanic background, PROC CROSSTAB
WARNING: These diabetes prevalence estimates are not age-adjusted.

7-level re-classification of Hispanic/Latino Background	DIABETES3_C2	Sample Size	Tot Percent	SE Tot Percent	Row Percent	SE Row Percent
Dominican	0	812.00	7.72	0.62	84.36	1.48
	1	189.00	1.43	0.18	15.64	1.48
Central American	0	1119.00	6.89	0.60	85.64	1.33
	1	251.00	1.16	0.12	14.36	1.33
Cuban	0	1361.00	16.31	1.58	83.33	1.15
	1	306.00	3.26	0.35	16.67	1.15
Mexican	0	3667.00	32.38	1.59	84.80	0.77
	1	952.00	5.80	0.37	15.20	0.77
Puerto Rican	0	1418.00	12.82	0.75	80.56	1.26
	1	539.00	3.09	0.25	19.44	1.26
South American	0	671.00	4.49	0.35	90.74	1.34
	1	87.00	0.46	0.07	9.26	1.34

If you only want to display the results for a specific subpopulation, add that variable to the CLASS statement and to that specific category in the SUBPOPN statement.

```
subpopn KEEP_DATA=1 & agegroup_c2=1;
```

6.2.3. Stata

The next group of statements uses the Stata software to calculate unadjusted diabetes prevalence by background. We use SVY: PROP with the OVER statement to request the prevalence by background.

```
/* SET UP SURVEY DESIGN */
svyset psu_id [pweight=weight_final_norm_overall], strata(strat)

/* OVERALL POPULATION – USING SVY: PROP */
svy: prop DIABETES3_C2, over(bkgrd1_c7)
```

6.2.4. R

The next group of statements uses the R software to calculate unadjusted diabetes prevalence by background. We use SVYMEAN within SVYBY function was to request the prevalence by background.

```
# READ IN DATASET
sol <- read.csv("sol.csv", header = TRUE, sep = ",")
sol$BKGRD1_C7 <- as.factor(sol$BKGRD1_C7)
# CREATE SURVEY DESIGN OBJECT
hchs.dsgn <- svydesign(id=~PSU_ID, strata=~STRAT,
weights=~WEIGHT_FINAL_NORM_OVERALL, data=sol, nest=TRUE)
# OVERALL POPULATION - USING SVYMEAN
svyby(~DIABETES3_C2, ~BKGRD1_C7, design=hchs.dsgn, svymean, na.rm=TRUE)
```

6.3. Diabetes prevalence by background stratified by agegroup_c2

One way that prevalence estimates are comparable among Hispanic backgrounds is to stratify by age group.

6.3.1. SAS

The prevalence of diabetes can be estimated by Hispanic background within each age group by crossing these two variables with DIABETES3_C2 in the TABLES statement. We present output for participants 45 years and older only.

```
proc surveyfreq data = sol;
   strata strat; cluster PSU_ID;  weight weight_final_norm_overall;
   tables KEEP_DATA*agegroup_C2*bkgrd1_c7*DIABETES3_C2 / row;
run;
```

Output 6.3.1. Diabetes prevalence by background for participants 45 years and older, PROC SURVEYFREQ

BKGRD1_C7	DIABETES3_C2	Frequency	Percent	Std Err of Percent	Row Percent	Std Err of Row Percent
Dominican	*0*	424	6.2559	0.5842	69.4754	2.5764
	1	165	2.7486	0.3583	30.5246	2.5764
Central American	*0*	516	4.8302	0.3710	67.5677	2.6191
	1	217	2.3185	0.2493	32.4323	2.6191
Cuban	*0*	836	18.8303	1.6922	71.6744	1.7206
	1	284	7.4417	0.7070	28.3256	1.7206
Mexican	*0*	1750	21.5447	1.4158	68.8380	1.6693
	1	773	9.7530	0.7683	31.1620	1.6693
Puerto Rican	*0*	800	11.8463	0.8888	65.2782	2.3293
	1	488	6.3011	0.5325	34.7218	2.3293
South American	*0*	385	4.5683	0.3623	80.6442	2.4348
	1	83	1.0965	0.1601	19.3558	2.4348

6.3.2. SUDAAN

The prevalence of diabetes can be estimated by Hispanic background within each age group by crossing these two variables with DIABETES3_C2 in the TABLES statement. We present output for participants 45 years and older only.

```
proc crosstab data=sol filetype=sas design=wr;
   nest strat PSU_ID;
   weight weight_final_norm_overall;
   subpopn KEEP_DATA=1;
   class  agegroup_C2 bkgrd1_c7 DIABETES3_C2;
   tables agegroup_C2*bkgrd1_c7*DIABETES3_C2;
run;
```

Output 6.3.2. Diabetes prevalence by background for participants 45 years and older, PROC CROSSTAB

1(18-44),2(45+)	7-level re-classification of Hispanic/Latino Background	Sample Size	Tot Percent	SE Tot Percent	Row Percent	SE Row Percent
Dominican	Total	589.00	9.00	0.81	100.00	0.00
	0	424.00	6.26	0.58	69.48	2.58
	1	165.00	2.75	0.36	30.52	2.58
Central American	Total	733.00	7.15	0.49	100.00	0.00
	0	516.00	4.83	0.37	67.57	2.62
	1	217.00	2.32	0.25	32.43	2.62
Cuban	Total	1120.00	26.27	2.18	100.00	0.00
	0	836.00	18.83	1.69	71.67	1.72
	1	284.00	7.44	0.71	28.33	1.72
Mexican	Total	2523.00	31.30	1.88	100.00	0.00
	0	1750.00	21.54	1.42	68.84	1.67
	1	773.00	9.75	0.77	31.16	1.67
Puerto Rican	Total	1288.00	18.15	1.11	100.00	0.00
	0	800.00	11.85	0.89	65.28	2.33
	1	488.00	6.30	0.53	34.72	2.33
South American	Total	468.00	5.66	0.42	100.00	0.00
	0	385.00	4.57	0.36	80.64	2.43
	1	83.00	1.10	0.16	19.36	2.43

6.3.3. Stata

The next group of statements uses the Stata software to calculate unadjusted diabetes prevalence by background stratified by agegroup_c2. We use SVY: PROP with the OVER statement to estimate the mean by background and agegroup_c2.

```
use sol.dta

/* SET UP SURVEY DESIGN */
svyset psu_id [pweight=weight_final_norm_overall], strata(strat)

/* STRATIFIED  - USING SVY: PROP */
svy: prop DIABETES3_C2, over(agegroup_c2 bkgrd1_c7)
```

6.3.4. R

The next group of statements uses the R software to calculate unadjusted diabetes prevalence by background stratified by agegroup_c2. We use SVYMEAN with SVYBY function to estimate the prevalence by background and agegroup_c2.

```
# READ IN DATASET
sol <- read.csv("sol.csv", header = TRUE, sep = ",")
sol$BKGRD1_C7 <- as.factor(sol$BKGRD1_C7)
sol$AGEGROUP_C2 <- as.factor(sol$AGEGROUP_C2)

# CREATE SURVEY DESIGN OBJECT
hchs.dsgn <- svydesign(id=~PSU_ID, strata=~STRAT,
weights=~WEIGHT_FINAL_NORM_OVERALL, data=sol, nest=TRUE)

# STRATIFIED - USING SVYMEAN
svyby(~DIABETES3_C2, ~BKGRD1_C7*AGEGROUP_C2, design=hchs.dsgn, svymean,
na.rm=TRUE)
```

6.4. Age-standardized diabetes prevalence by background

Because the age distributions in HCHS/SOL and NHANES are different, we need to use a standard age distribution to make comparisons between the two studies. Here we illustrate how to estimate proportions standardizing the HCHS/SOL population to the US 2000 population age distribution (Klein RJ, Schoenborn CA, 2001), which is the age distribution used by NHANES.

6.4.1. SAS

Currently, SAS 9.3 does not allow for external age standardization in PROC SURVEYMEANS. However, we can apply the contrast statement in the SURVEYREG to adjust for external age standardization.

```
proc surveyreg data = sol order=internal;
   strata strat;
   cluster psu_id;
   weight &weight;
   domain KEEP_DATA;
   class Bkgrd1_c7 AGEGROUP_C6_NHANES DIABETES3_C2;
   model DIABETES3_C2 = Bkgrd1_c7 AGEGROUP_C6_NHANES
Bkgrd1_c7*AGEGROUP_C6_NHANES / solution noint;
   /* 2000 US CENSUS AGE DISTRIBUTION */
   estimate 'Dominican' bkgrd1_c7 1 0 0 0 0 0 0
                        agegroup_c6_nhanes 0.235800444 0.222616766 0.225788538
                        0.162064749 0.10713542 0.046594083
                        bkgrd1_c7*agegroup_c6_nhanes 0.235800444 0.222616766
                        0.225788538 0.162064749 0.10713542 0.046594083
                                                0 0 0 0 0 0
                                                0 0 0 0 0 0
                                                0 0 0 0 0 0
                                                0 0 0 0 0 0
                                                0 0 0 0 0 0
                                                0 0 0 0 0 0 / e;
 run;
```

Output 6.4.1. Diabetes prevalence in Dominican age-standardized to the US 2000 population, PROC SURVEYREG

		Estimate Standard			
Label	*Estimate*	*Error*	*DF*	*t Value*	*Pr > |t|*
Dominican	0.1828	0.01340	644	13.64	<.0001

6.4.2. SUDAAN

External standardization is easily done in SUDAAN using PROC DESCRIPT. Include the age group variable (e.g., agegroup_c6_NHANES) in the CLASS and STDVAR statements and specify the external age distribution (e.g., to the US 2000 population in Output 1.5) in the STDWGT statement.

```
proc descript data=sol filetype=sas design=wr;
   nest strat PSU_ID; weight weight_final_norm_overall;
   subpopn KEEP_DATA=1;
   class Bkgrd1_c7 AGEGROUP_C6_NHANES DIABETES3_C2;
   var DIABETES3_C2;
   catlevel 1; /* Output only % of diabetes (i.e. DIABETES3_C2=1) */
   stdvar AGEGROUP_C6_NHANES;
   /* 2000 US CENSUS AGE DISTRIBUTION */
   stdwgt 0.235800444 0.222616766 0.225788538 0.162064749 0.10713542
            0.046594083;
run;
```

Output 6.4.2. Diabetes prevalence by background age-standardized to the US 2000 population, PROC DESCRIPT

Variable	7-level re-classification of Hispanic/Latino Background	Sample Size	Percent	SE Percent
DIABETES3_C2: 1	Total	11733	16.74	0.48
	Dominican	1000	18.28	1.34
	Central American	1369	17.31	1.36
	Cuban	1666	12.76	0.83
	Mexican	4615	18.51	0.89
	Puerto Rican	1956	18.85	1.13
	South American	757	9.17	1.15
	Mixed/Other	370	20.23	2.98

7. LINEAR MODELS TO ESTIMATE EFFECTS AND ADJUSTED MEANS

In this section we use linear models to 1) estimate effects of risk factors and covariates on outcomes and to 2) adjust means for covariates.

7.1. Effects

7.1.1. SAS

The next group of SAS statements invokes SURVEYREG to produce linear regression estimates that appropriately account for the study design. The SOLUTION option in the MODEL statement outputs the parameter estimates for the regression coefficients as well as the t-values for those parameter estimates.

```
proc surveyreg data=sol; /* DEFAULT: order=formatted */
   strata strat; cluster PSU_ID; weight weight_final_norm_overall;
   domain KEEP_DATA;
   class cigarette_use Bkgrd1_c7;
   model sbpa5 = cigarette_use bkgrd1_c7 age female  / solution;
run;
```

WARNING: In PROC SURVEYREG the reference group for categorical independent variables is always the last level. By default, when variables are formatted the reference group is the last category according to the formatted order; this option can be explicitly specified in the PROC SURVEYREG statement with ORDER=FORMATTED. To use the internal order of the variable (thereby ignoring formatting), you need to specify explicitly the option ORDER=INTERNAL in the DATA statement. To change the reference category in PROC SURVEYREG to anything other than the last formatted value or the last numeric value, the actual variable must be recoded in a previous data step. This is not the case with PROC SURVEYLOGISTIC or SUDAAN.

For example, BKGRD1_C7 has internal values 0, 1, 2, 3, 4, 5 and 6 with corresponding formatted values ‘Dominican’, ‘Central American, ‘Cuban’, ‘Mexican’, ‘Puerto Rican’, ‘South American’ and ‘Mixed/Other’. If we specify ORDER=INTERNAL option, the reference group is 6 (‘Mixed/Other’). If we specify the ORDER=FORMATTED option, which is the default, ‘South American’ (5) is the reference group because it is the last formatted value alphabetically.

Statistically, any level can serve as the referent and is a matter of preference which one to use. If we want the Mexicans to be the reference, we need to create another variable for which the last value is assigned to Mexicans. To avoid recoding a variable, we will use ORDER=FORMATTED making the alphabetically last level, South Americans, the reference group for these SURVEYREG procedures. For all other sections, unless otherwise specified, Mexicans will be used as the reference group for the Hispanic Background variable, as Mexicans is the Hispanic Background group with the largest sample size.

Output 7.1.1a. Linear regression on SBP, PROC SURVEYREG

Tests of Model Effects

Effect	*Num DF*	*F Value*	*Pr > F*
Model	10	256.36	<.0001
Intercept	1	21230.5	<.0001
CIGARETTE_USE	2	3.86	0.0216
BKGRD1_C7	6	14.73	<.0001
AGE	1	1461.67	<.0001
female	1	437.96	<.0001

Estimated Regression Coefficients

Parameter	*Estimate*	*Standard Error*	*t Value*	*Pr > |t|*
Intercept	99.2782453	0.88404014	112.30	<.0001
CIGARETTE_USE Current Smoker of Cigarettes	-1.1013224	0.50163854	-2.20	0.0285
CIGARETTE_USE Formerly Smoked Cigarettes	-1.1344024	0.52746574	-2.15	0.0319
CIGARETTE_USE Never Smoked Cigarettes	0.0000000	0.00000000	.	.
BKGRD1_C7 Central American	3.9388347	0.89322031	4.41	<.0001
BKGRD1_C7 Cuban	2.9582240	0.71943533	4.11	<.0001
BKGRD1_C7 Dominican	5.3930972	0.79576866	6.78	<.0001
BKGRD1_C7 Mexican	0.6936168	0.72067716	0.96	0.3362
BKGRD1_C7 Mixed/Other	2.7577516	0.98089097	2.81	0.0051
BKGRD1_C7 Puerto Rican	3.3856041	0.80753290	4.19	<.0001
BKGRD1_C7 South American	0.0000000	0.00000000	.	.
AGE	0.5569241	0.01456704	38.23	<.0001
female	-7.8618026	0.37566794	-20.93	<.0001

NOTE: SURVEYREG in SAS version 9.1 does not allow for subpopulation analyses, but version 9.2 does. Due to the sample design of the HCHS/SOL study, the use of a WHERE statement is not equivalent to a subpopulation analysis. Further information about subpopulation analysis can be found in Section 1.6.

For example, to estimate the effects of cigarette use, background and age in systolic blood pressure by gender we add female in the DOMAIN statement. We present output for women only.

```
proc surveyreg data = sol; /* DEFAULT: order=formatted */
   strata strat;
   cluster PSU_ID;
   weight weight_final_norm_overall;
   class cigarette_use bkgrd1_c7;
   model sbpa5 = cigarette_use bkgrd1_c7 age / solution;
   domain KEEP_DATA*female;
```

```
run;
```

Output 7.1.1b. Linear regression on SBP for women, PROC SURVEYREG

Tests of Model Effects

Effect	*Num DF*	*F Value*	*Pr > F*
Model	9	203.66	<.0001
Intercept	1	10039.0	<.0001
CIGARETTE_USE	2	2.47	0.0855
BKGRD1_C7	6	14.23	<.0001
AGE	1	1498.90	<.0001

Estimated Regression Coefficients

Parameter	*Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Design Effect*
Intercept	85.4438078	1.04806013	81.53	<.0001	2.92
CIGARETTE_USE Current Smoker of Cigarettes	-1.6838053	0.80904837	-2.08	0.0378	6.38
CIGARETTE_USE Formerly Smoked Cigarettes	-0.9040283	0.76215449	-1.19	0.2360	4.77
CIGARETTE_USE Never Smoked Cigarettes	0.0000000	0.00000000	.	.	.
BKGRD1_C7 Central American	4.4021555	1.04787040	4.20	<.0001	2.64
BKGRD1_C7 Cuban	4.3667961	0.87037501	5.02	<.0001	2.28
BKGRD1_C7 Dominican	5.0869351	0.93577098	5.44	<.0001	2.29
BKGRD1_C7 Mexican	-0.0024186	0.87641717	-0.00	0.9978	2.63
BKGRD1_C7 Mixed/Other	2.9405146	1.34455137	2.19	0.0291	3.07
BKGRD1_C7 Puerto Rican	4.7448634	1.01535932	4.67	<.0001	2.94
BKGRD1_C7 South American	0.0000000	0.00000000	.	.	.
AGE	0.6974642	0.01801506	38.72	<.0001	5.39

7.1.2. SUDAAN

The next group of SUDAAN statements invokes REGRESS to produce linear regression estimates that appropriately account for the study design.

```
proc regress data=sol filetype=sas design=wr;
   nest strat PSU_ID;
   weight weight_final_norm_overall;
   class cigarette_use bkgrd1_c7;
   subpopn KEEP_DATA=1;
   model sbpa5 = cigarette_use bkgrd1_c7 age female;
   reflevel bkgrd1_c7=3; /* reference: Mexicans */
run;
```

NOTE: In SUDAAN, to specify the reference level for any categorical variable you use the REFLEVEL statement; simply specify the variable name equal to the internal numerical value for the reference level.

Output 7.1.2a. Linear regression on SBP, PROC REGRESS

Contrast	Degrees of Freedom	Wald ChiSq	P-value Wald ChiSq
OVERALL MODEL	11	524340.24	0.0000
MODEL MINUS INTERCEPT	10	2565.15	0.0000
INTERCEPT	.	.	.
CIGARETTE_USE	2	7.73	0.0210
BKGRD1_C7	6	88.42	0.0000
AGE	1	1462.57	0.0000
FEMALE	1	438.23	0.0000

Independent Variables and Effects		Beta Coeff.	SE Beta	T-Test B=0	P-value T-Test B=0
Intercept		98.8705	0.7500	131.8360	0.0000
Cigarette_Use (1=Never,2=Former- ,3=Current) Never Smoked	Cigarettes	1.1013	0.5015	2.1961	0.0284
	Formerly Smoked Cigarettes	-0.0331	0.6414	-0.0516	0.9589
	Current Smoker of Cigarettes	0.0000	0.0000	.	.
7-level re-classification of Hispanic/Latino Background	Dominican	4.6995	0.6034	7.7881	0.0000
	Central American	3.2452	0.6545	4.9585	0.0000
	Cuban	2.2646	0.5140	4.4062	0.0000
	Mexican	0.0000	0.0000	.	.
	Puerto Rican	2.6920	0.6225	4.3243	0.0000
	South American	-0.6936	0.7205	-0.9627	0.3360
	Mixed/Other	2.0641	0.8417	2.4523	0.0145
Age		0.5569	0.0146	38.2436	0.0000
1-female 0-male		-7.8618	0.3756	-20.9340	0.0000

The SUBPOPN statement provides regression estimates for a specified subpopulation. For example, the linear regression estimates on SBP for females are produced by:

```
proc regress data=sol filetype=sas design=wr;
   nest strat PSU_ID;
   weight weight_final_norm_overall;
   class cigarette_use bkgrd1_c7 ;
   model sbpa5 = cigarette_use bkgrd1_c7 age;
   reflevel bkgrd1_c7=3; /* reference: Mexicans */
   subpopn KEEP_DATA=1 & female=1;
run;
```

Output 7.1.2b. Linear regression on SBP for women, PROC REGRESS

Contrast	Degrees of Freedom	Wald ChiSq	P-value Wald ChiSq
OVERALL MODEL	10	325992.57	0.0000
MODEL MINUS INTERCEPT	9	1833.97	0.0000
INTERCEPT	.	.	.
CIGARETTE_USE	2	4.94	0.0846
BKGRD1_C7	6	85.45	0.0000
AGE	1	1499.73	0.0000

Independent Variables and Effects		Beta Coeff.	SE Beta	T-Test B=0	P-value T-Test B=0
Intercept		83.7576	1.1844	70.7162	0.0000
Cigarette_Use (1=Never,2=Former- ,3=Current) Never Smoked	Cigarettes	1.6838	0.8088	2.0818	0.0378
	Formerly Smoked Cigarettes	0.7798	0.9917	0.7863	0.4320
	Current Smoker of Cigarettes	0.0000	0.0000	.	.
7-level re-classification of Hispanic/Latino Background	Dominican	5.0894	0.7596	6.7001	0.0000
	Central American	4.4046	0.8199	5.3723	0.0000
	Cuban	4.3692	0.7095	6.1582	0.0000
	Mexican	0.0000	0.0000	.	.
	Puerto Rican	4.7473	0.8154	5.8222	0.0000
	South American	0.0024	0.8762	0.0028	1.0000
	Mixed/Other	2.9429	1.2254	2.4015	0.0166
Age		0.6975	0.0180	38.7263	0.0000

NOTE: Subpopulation statistics cannot be properly obtained using BY or WHERE statements, since this restricts the overall analyses data set to just those participants meeting the criterion defined in the BY or WHERE statements. These approaches use the incorrect denominators in computing the standard errors, confidence intervals, and test statistics. Further information about subpopulation analysis can be found in Section 1.5.

7.2. Unadjusted SBP mean by background – WARNING

7.2.1. SAS

If we only include background in the MODEL statement and do not include an intercept, the regression coefficients are the unadjusted SBP means by background. Because survey PROCEDURES in SAS 9.3 do not provide adjusted means directly, we need to explicitly and carefully specify the contrast matrix using the ESTIMATE statement (section 7.3.1). Here, we illustrate how to specify the ESTIMATE statements to provide unadjusted SBP means by background. The NOINT option requests to not include an intercept in the model. Alternatively, we can remove the NOINT option and include explicitly the word 'intercept' and the coefficient one in the ESTIMATE statement. Regression coefficients from PROC SURVEYREG are identical to the estimable functions specified in the ESTIMATE statements below (Output 7.2.1) and to the unadjusted means from PROC SURVEYMEANS (Output 5.2.1).

Output 7.2.1. Unadjusted (by age) SBP mean by background, PROC SURVEYREG
WARNING: These SBP means are not age-adjusted.

Estimated Regression Coefficients

Parameter	*Estimate*	*Standard Error*	*t Value*	*Pr > |t|*
BKGRD1_C7 Dominican	121.943918	0.66342182	183.81	<.0001
BKGRD1_C7 Central American	120.782512	0.67419770	179.15	<.0001
BKGRD1_C7 Cuban	123.826186	0.51099940	242.32	<.0001
BKGRD1_C7 Mexican	116.730523	0.39626341	294.58	<.0001
BKGRD1_C7 Puerto Rican	121.839646	0.58042564	209.91	<.0001
BKGRD1_C7 South American	118.524935	0.83544687	141.87	<.0001
BKGRD1_C7 Mixed/Other	117.070027	0.91918475	127.36	<.0001

Analysis of Estimable Functions

Parameter	*Estimate*	*Standard Error*	*t Value*	*Pr > |t|*
Dominican	121.943918	0.66342182	183.81	<.0001
Central Amer	120.782512	0.67419770	179.15	<.0001
Cuban	123.826186	0.51099940	242.32	<.0001
Mexican	116.730523	0.39626341	294.58	<.0001
Puerto-Rican	121.839646	0.58042564	209.91	<.0001
South Americ	118.524935	0.83544687	141.87	<.0001

WARNING: In SAS, the ORDER= option is very important when you use the ESTIMATE or CONTRAST statement because it specifies the order for the levels of classification variables. This ordering determines which parameters in the model correspond to each level in the data. When the variables have formats attached to them, the default ORDER=FORMATTED sorts the levels by the formatted values. For example, with the default ORDER=FORMATTED the first level is 'Central American' with internal value one and not 'Dominican' with internal value zero. The code above (using ORDER=INTERNAL) and the code below (using the default ORDER=FORMATTED) produce identical estimates (Output 7.2.1), but the ESTIMATE statements are different due to the different ORDER= option. In SAS, it is useful to specify the option E in the ESTIMATE statement after a slash (/) to display the entire contrast matrix (or vector) to confirm the parameters' order. **We strongly recommend using ORDER=INTERNAL.**

```
proc surveyreg data = sol;
  strata strat; cluster psu_id; weight weight_final_norm_overall;
  domain KEEP_DATA;
  class bkgrd1_c7 ;
  model sbpa5 = bkgrd1_c7 / solution noint;
  /* WARNING: Statements below follow formatted order which is used by default
because no ORDER=option was specified */
  estimate 'Dominican'    bkgrd1_c7 1 0 0 0 0 0 0 / e;
  estimate 'Central Amer' bkgrd1_c7 0 1 0 0 0 0 0 / e;
  estimate 'Cuban'        bkgrd1_c7 0 0 1 0 0 0 0  / e;
  estimate 'Mexican    '  bkgrd1_c7 0 0 0 1 0 0 0  / e;
  estimate 'Puerto-Rican' bkgrd1_c7 0 0 0 0 1 0 0  / e;
  estimate 'South Americ' bkgrd1_c7 0 0 0 0 0 1 0  / e;
  title "WARNING: These SBP means are not age-adjusted";
run;
```

NOTE: SAS 9.3 does provide adjusted means directly with the LSMEANS statement for variables in the CLASS statement. The 'AT MEANS' or 'AT AGE=' options must be used so that these estimates will be adjusted to the weighted means. The LSMEANS statement is not available in SAS 9.2. The 'AT MEANS' option provides the adjusted mean at the overall adjusted age mean, whereas we can specify any age with the 'AT THE AGE' option. Results identical to Output 7.2.1 are produced when using the LSMEANS statement with the AT MEANS options are shown below.

```
proc surveyreg data = sol order=internal;
  strata strat; cluster psu_id; weight weight_final_norm_overall;
  domain KEEP_DATA;
  class bkgrd1_c7 ;
  model sbpa5 = bkgrd1_c7 / solution noint;
  lsmeans bkgrd1_c7 / at means;
  title "WARNING: These SBP means are not age-adjusted";
run;
```

7.2.2. SUDAAN code

In PROC REGRESS, if we only include background in the model, the CONDMARG statement provides unadjusted SBP means by background, and these are identical to the regression coefficients. Therefore, Output 7.2.2 is identical to the unadjusted means using PROC DESCRIPT from Output 5.2.2. In SUDAAN, these unadjusted means can be estimated using linear models as follows:

NOTE: A conditional marginal mean is an estimate of the expected outcome for an individual conditional on belonging to a specific group (e.g., Hispanic background) and having covariate values equal to the weighted average covariates.

```
proc regress data=sol filetype=sas design=wr;
    nest strat PSU_ID;  weight weight_final_norm_overall;
    class bkgrd1_c7;   subpopn KEEP_DATA=1;
    reflevel bkgrd1_c7=3; /* reference: Mexicans */
    model sbpa5 = bkgrd1_c7;
    /* NOTE: Point estimates are the same for predmarg, condmarg and lsmeans
       for linear models but NOT for generalized linear models */
    condmarg bkgrd1_c7; /* SE identical to SAS */
   title "WARNING: These SBP means are not age-adjusted";
run;
```

Output 7.2.2. Unadjusted (by age) SBP mean by background, PROC REGRESS **WARNING:** These SBP means are not age-adjusted.

Conditional Marginal #1		Conditional Marginal	SE	T:Marg-=0	P-value
7-level re-classification of Hispanic/Latino Background	Dominican	121.944	0.663	183.844	0.000
	Central American	120.783	0.674	179.183	0.000
	Cuban	123.826	0.511	242.366	0.000
	Mexican	116.731	0.396	294.632	0.000
	Puerto Rican	121.840	0.580	209.953	0.000
	South American	118.525	0.835	141.896	0.000
	Mixed/Other	117.070	0.919	127.386	0.000

7.3. Age-adjusted SBP mean by background

Frequently in epidemiologic studies, we are interested in estimating the average outcome associated with different risk factors controlling for covariates. For example, we might be interested in reporting the average systolic blood pressure (SBP) by smoking status adjusting by age and gender. Adjustment can be internal (i.e. to the same population) or external (to a reference population).

Survey PROCEDURES in SAS 9.2 do not provide adjusted means directly as SUDAAN does, but SAS 9.3 does. However, you can obtain adjusted means in SAS PROC SURVEYREG by explicitly and carefully specifying the contrast matrix in the ESTIMATE statement. For example, to estimate the adjusted SBP mean by background to the overall mean age you specify the weighted mean age in the contrast matrix at the ESTIMATE statement. These adjusted means and standard errors are conditional on the individual having the average mean age. Similarly, when more covariates are being adjusted in the model you need to provide in the contrast matrix the weighted means and weighted frequencies previously estimated.

7.3.1. SAS

To estimate age-adjusted SBP mean by background we can use the LSMEANS statement as shown in 7.2.1. Here, we illustrate what SAS is doing 'back the scenes'. First, we estimate the weighted mean age in the subpopulation of interest and then we include age in the linear model and specify the weighted average age into the ESTIMATE statement. In HCHS/SOL the weighted mean age is 41 years.

The ‘long’ way of calculating it:

```
proc surveymeans data = sol;
   strata strat; cluster PSU_ID; weight weight_final_norm_overall;
   domain KEEP_DATA;   var age;
run;
```

Weighed Mean Age for HCHS/SOL Population (By KEEP_DATA variable)

KEEP_DATA	Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
1	*AGE*	Age	11815	40.885787	0.288874	40.3185381	41.4530359

```
%let meanage = 40.885787; /* Mean age in waves 1 and 2 */
proc surveyreg data = sol order=internal;
  strata strat; cluster psu_id; weight weight_final_norm_overall;
  domain KEEP_DATA;  class bkgrd1_c7;
  model sbpa5 = bkgrd1_c7 age / solution noint;
  /* WARNING: MAKE SURE ORDER=INTERNAL */
  estimate 'Dominican'    bkgrd1_c7 1 0 0 0 0 0 0 age &meanage / e;
  estimate 'Central Amer' bkgrd1_c7 0 1 0 0 0 0 0 age &meanage / e;
  estimate 'Cuban'        bkgrd1_c7 0 0 1 0 0 0 0 age &meanage / e;
  estimate 'Mexican   '   bkgrd1_c7 0 0 0 1 0 0 0 age &meanage / e;
  estimate 'Puerto-Rican' bkgrd1_c7 0 0 0 0 1 0 0 age &meanage / e;
  estimate 'South Americ' bkgrd1_c7 0 0 0 0 0 1 0 age &meanage / e;
run;
```

The ‘short’ way of calculating it:

```
proc surveyreg data = sol order=internal; /* EQUIVALENTLY */
  strata strat; cluster psu_id; weight weight_final_norm_overall;
  domain KEEP_DATA;  class bkgrd1_c7;
  model sbpa5 = bkgrd1_c7 age / solution noint;
  lsmeans bkgrd1_c7 / at means;
run;
```

Output 7.3.1. Age-adjusted SBP mean by background, PROC SURVEYREG

Analysis of Estimable Functions

Parameter	Age	Estimate	Standard Error	t Value	Pr > |t|
Dominican	40.88	122.521339	0.54034854	226.75	<.0001
Central Amer	40.88	121.478644	0.58238481	208.59	<.0001
Cuban	40.88	120.908882	0.38955561	310.38	<.0001
Mexican	40.88	118.087270	0.35176794	335.70	<.0001
Puerto-Rican	40.88	120.938073	0.50827097	237.94	<.0001
South Americ	40.88	117.558055	0.63758292	184.38	<.0001

Sometimes, we are interested in adjusting to a specific age (e.g. 60 y) instead of the overall mean. So, that specific age would be specified in the contrast matrix at the ESTIMATE statement.

7.3.2. SUDAAN

In SUDAAN, the age-adjusted SBP by background can be estimated by including age in the MODEL statement and requesting CONDMARG statement with background variable.

```
proc regress data=sol filetype=sas design=wr;
    nest strat PSU_ID;
    weight weight_final_norm_overall;
    class bkgrd1_c7;
    subpopn KEEP_DATA=1;
    reflevel bkgrd1_c7=3; /* reference: Mexicans */
    model sbpa5 = bkgrd1_c7 age;
    condmarg bkgrd1_c7;
run;
```

Output 7.3.2. Age-adjusted SBP mean by background, PROC REGRESS

Conditional Marginal #1		Conditional Margin-al	SE	T:Marg-=0	P-value
7-level re-classification of Hispanic/Latino Background	Dominican	122.519	0.540	226.787	0.000
	Central American	121.476	0.582	208.631	0.000
	Cuban	120.906	0.389	310.433	0.000
	Mexican	118.085	0.352	335.775	0.000
	Puerto Rican	120.936	0.508	237.986	0.000
	South American	117.556	0.637	184.415	0.000
	Mixed/Other	120.368	0.862	139.691	0.000

7.4. Age-sex adjusted SPB mean by background

If we are interested in SBP means by Hispanic background adjusted by age and gender, we need to include these two covariates in the MODEL statement.

7.4.1. SAS

Commonly, age and sex adjusted means are of interest, and, as gender is a categorical variable, we must also calculate the weighted sex frequencies in order to get age and sex adjusted mean SBP. For example, we define the macro variable female by 0.518 which is the weighted frequency of women in HCHS/SOL and we include it in the ESTIMATE statement.

```
%let female = 0.518192;
proc surveyreg data = sol order=internal;
  strata strat;  cluster psu_id;  weight weight_final_norm_overall;
  domain KEEP_DATA; class bkgrd1_c7 ;
  model sbpa5 = bkgrd1_c7 age female / solution noint;
  /* WARNING: MAKE SURE ORDER=INTERNAL */
  estimate 'Dominican'  bkgrd1_c7 1 0 0 0 0 0 0 age &meanage female &female / e;
  estimate 'Central Am' bkgrd1_c7 0 1 0 0 0 0 0 age &meanage female &female / e;
  estimate 'Cuban'      bkgrd1_c7 0 0 1 0 0 0 0 age &meanage female &female / e;
  estimate 'Mexican   ' bkgrd1_c7 0 0 0 1 0 0 0 age &meanage female &female / e;
  estimate 'Puerto-Ric' bkgrd1_c7 0 0 0 0 1 0 0 age &meanage female &female / e;
  estimate 'South Amer' bkgrd1_c7 0 0 0 0 0 1 0 age &meanage female &female / e;
run;
```

Output 7.4.1. Age and sex adjusted SBP mean by background, PROC SURVEYREG

Analysis of Estimable Functions

Parameter	*Estimate*	*Standard Error*	*t Value*	*Pr > |t|*
Dominican	123.126292	0.50289908	244.83	<.0001
Central Amer	121.573287	0.55230065	220.12	<.0001
Cuban	120.455546	0.36868647	326.72	<.0001
Mexican	118.255529	0.34079953	346.99	<.0001
Puerto-Rican	120.696458	0.50390374	239.52	<.0001
South Americ	117.578087	0.61627200	190.79	<.0001

7.4.2. SUDAAN

If we are interested in SBP means by Hispanic background adjusted by age and gender then we need to 1) include these two covariates in the MODEL statement and 2) include the CONDMARG with bkgrd1_c7.

```
proc regress data=sol filetype=sas design=wr;
    nest strat PSU_ID;
    weight weight_final_norm_overall;
    class bkgrd1_c7;
    subpopn KEEP_DATA=1;
    reflevel bkgrd1_c7=3; /* reference: Mexicans */
    model sbpa5 = bkgrd1_c7 age female;
    condmarg bkgrd1_c7;
run;
```

Output 7.4.2. Age and sex adjusted SBP mean by background, PROC REGRESS

Conditional Marginal #1		**Conditional Marginal**	**SE**	**T:Marg-=0**	**P-value**
7-level re-classification of Hispanic/Latino Background	Dominican	123.127	0.503	244.882	0.000
	Central American	121.574	0.552	220.176	0.000
	Cuban	120.456	0.369	326.793	0.000
	Mexican	118.256	0.341	347.095	0.000
	Puerto Rican	120.697	0.504	239.581	0.000
	South American	117.579	0.616	190.835	0.000
	Mixed/Other	120.269	0.768	156.541	0.000

Estimates in Output 7.4.2 are slightly different (in the 3rd decimal place) to those from Output 7.4.1 using explicitly the ESTIMATE statements, because SUDAAN is using the weighted mean age among those without missing values in sbpa5.

7.5. Age-adjusted SPB mean by background stratified by gender

Often, it is of interest to estimate adjusted means stratifying by a particular variable. There are two ways to calculate adjusted means stratifying for a variable. One way requires explicitly defining the stratified model by using interaction terms and calculating adjusted

means from explicitly writing the appropriate ESTIMATE statements. The second way is to use the subpopulation statement (DOMAIN in SAS or SUBPOPN in SUDAAN). However, you need to be very clear on what your question of interest is and whether you want to adjust to the overall mean or to the specific subpopulation mean.

7.5.1. SAS code

WARNING: In SAS, the DOMAIN statement provides output for all subpopulations; hence, the ESTIMATE statements apply to all subpopulations. This is very useful when we are interested in comparing means across all subpopulations because it adjusts to the overall mean not to the specific subpopulation mean.

In contrast, if we are interested in estimating adjusted means for 1) one specific subpopulation or for 2) each level of a stratification variable, an ESTIMATE statement for each level of the stratification variable must be specified using stratification level specific weighted means and percentages. For example, the age-adjusted SBP mean by background stratifying by gender are given by:

```
proc surveymeans data = sol;
   strata strat;   cluster PSU_ID;   weight weight_final_norm_overall;
   var age;
   domain KEEP_DATA*female;
run;

/* Age for each gender subpopulation output as macro variable */
%let meanage_female = 41.599911;
%let meanage_male   = 40.117736;
%let meanage        = 40.885787; /* Mean age in waves 1 and 2 */

proc surveyreg data = sol order=internal;
  strata strat; cluster psu_id; weight weight_final_norm_overall;
  class bkgrd1_c7;
  model sbpa5 = bkgrd1_c7 age / solution noint;
  domain KEEP_DATA*female;
  /* WARNING: WHEN FORMATS ARE USED THE ORDER IS ALPHABETICAL */
  estimate 'Dominican - M'    bkgrd1_c7  1 0 0 0 0 0 0 age &meanage_male  / e;
  estimate 'Central Am - M'   bkgrd1_c7  0 1 0 0 0 0 0 age &meanage_male  / e;
  estimate 'Cuban - M'        bkgrd1_c7  0 0 1 0 0 0 0 age &meanage_male  / e;
  estimate 'Mexican - M'      bkgrd1_c7  0 0 0 1 0 0 0 age &meanage_male  / e;
  estimate 'Puerto-Ric - M'   bkgrd1_c7  0 0 0 0 1 0 0 age &meanage_male  / e;
  estimate 'South Am - M'     bkgrd1_c7  0 0 0 0 0 1 0 age &meanage_male  / e;
  estimate 'Dominican - F'    bkgrd1_c7  1 0 0 0 0 0 0 age &meanage_female  / e;
  estimate 'Central Am - F'   bkgrd1_c7  0 1 0 0 0 0 0 age &meanage_female  / e;
  estimate 'Cuban - F'        bkgrd1_c7  0 0 1 0 0 0 0 age &meanage_female  / e;
  estimate 'Mexican - F'      bkgrd1_c7  0 0 0 1 0 0 0 age &meanage_female  / e;
  estimate 'Puerto-Rican - F' bkgrd1_c7  0 0 0 0 1 0 0 age &meanage_female  / e;
  estimate 'South Am - F'     bkgrd1_c7  0 0 0 0 0 1 0 age &meanage_female  / e;

  lsmeans BKGRD1_C7 / at age=&meanage_female; /* equivalent to estimate
  statements for females*/
  lsmeans BKGRD1_C7 / at means; /* Adjusted SBP to domain subpopulation */
  lsmeans BKGRD1_C7 / at age=&meanage;  /* Adjusted SBP to overall age */

run;
```

WARNING: Note that we are only interested in one set of values for the estimable functions in the output. Output 7.5.1a only displays for female = 1 so the bolded results below are the estimates for females that we are interested in, but the estimates for males would need to be selected from the female=0 output.

Output 7.5.1a. Age-adjusted mean SPB by background stratified by gender – ONLY OUTPUT FOR FEMALES, SURVEYREG with DOMAIN

Analysis of Estimable Functions

Parameter	*Estimate*	*Standard Error*	*t Value*	*Pr > |t|*
Dominican - Female	119.309791	0.59797185	199.52	<.0001
Central Am - Female	118.599541	0.66583940	178.12	<.0001
Cuban – Female	118.348843	0.48061825	246.24	<.0001
Mexican – Female	114.149287	0.46592364	245.00	<.0001
Puerto-Rican - Female	118.284851	0.63625137	185.91	<.0001
South Am – Female	114.146562	0.73907562	154.45	<.0001

Now we illustrate how to obtain these age-adjusted means by gender including the interaction between background and gender explicitly in the MODEL statement. Here are the ESTIMATE statements needed for estimating means for Dominicans by gender adjusted using 1) gender-specific age mean, or 2) overall age mean.

```
proc surveyreg data = sol order=internal;
  strata strat;  cluster psu_id;   weight weight_final_norm_overall;
  domain KEEP_DATA;
  class bkgrd1_c7;
  model sbpa5 = bkgrd1_c7 age female bkgrd1_c7*female age*female/solution noint;
  /*WARNING: ORDER=INTERNAL option must be specified to use unformatted values*/
  estimate 'Dom-M (M age mean)'    /* Gender-specific age mean */
  bkgrd1_c7        1 0 0 0 0 0 0 age &meanage_male female 0
  bkgrd1_c7*female 0 0 0 0 0 0 0 age*female 0 / e;
  estimate 'Dom-F (F age mean)'
  bkgrd1_c7        1 0 0 0 0 0 0 age &meanage_female female 1
  bkgrd1_c7*female 1 0 0 0 0 0 0 age*female &meanage_female / e;
  /* Overall age mean */
  estimate 'Dom-M (overall age mean)'
  bkgrd1_c7        1 0 0 0 0 0 0 age  &meanage female 0
  bkgrd1_c7*female 0 0 0 0 0 0 0 age*female 0 / e;
  estimate 'Dom-F (overall age mean)'
  bkgrd1_c7        1 0 0 0 0 0 0 age &meanage female 1
  bkgrd1_c7*female 1 0 0 0 0 0 0 age*female &meanage / e;
run;
```

Output 7.5.1b. Age-adjusted mean SPB for Dominicans stratified by gender – SURVEYREG with interaction

Analysis of Estimable Functions

Parameter	*Estimate*	*Standard Error*	*t Value*	*Pr > |t|*
Dom-M (M age mean)	127.351401	0.79873370	159.44	<.0001
Dom-F (F age mean)	119.309791	0.59811858	199.48	<.0001
Dom-M (overall age mean)	127.717015	0.79782360	160.08	<.0001
Dom-F (overall age mean)	118.932265	0.59908633	198.52	<.0001

Note, for example, that the SBP mean for Dominican females adjusted to the overall age mean is identical when we include the interaction terms in the model statement (Output 7.5.1b) or when we use the domain statement (Output 7.5.1a).

7.5.2. SUDAAN code

The SUBPOPN statement only allows for analysis of one subpopulation within each PROCEDURE. To calculate adjusted means by background for males and females, two PROC calls must be used each one with a different SUBPOPN statement. For example, code below provides the age-adjusted SBP mean for females using the mean age for women. Alternatively, add SEX and the interaction with background in the model.

WARNING: Because SUDAAN focuses on the analysis of one subpopulation at a time the mean age used to adjust internally is the weighted mean age for that specific subpopulation. Therefore, if you are interested in comparing means across all subpopulations, you need to include the interaction term in the MODEL statement.

```
proc regress data=sol filetype=sas design=wr;
    nest strat PSU_ID;
    weight weight_final_norm_overall;
    class bkgrd1_c7;
    model sbpa5 = bkgrd1_c7 age;
    subpopn KEEP_DATA=1 & female=1;
    condmarg bkgrd1_c7;
run;
```

Output 7.5.2a. Age-adjusted mean SPB by background for women, PROC REGRESS

Conditional Marginal #1		**Conditional Marginal**	**SE**	**T:Marg-=0**	**P-value**
7-level re-classification of Hispanic/Latino Background	Dominican	119.307	0.598	199.559	0.000
	Central American	118.597	0.666	178.155	0.000
	Cuban	118.346	0.481	246.290	0.000
	Mexican	114.146	0.466	245.046	0.000
	Puerto Rican	118.282	0.636	185.946	0.000
	South American	114.144	0.739	154.473	0.000
	Mixed/Other	116.916	1.161	100.685	0.000

These age-adjusted SBP for females (Output 7.5.2a) are almost identical to those obtained by explicitly writing the ESTIMATE statements in SAS with mean age for women (Output 7.5.1a). Estimates are slightly different since Output 7.5.2a use estimates of the weighted mean age among those without missing values in sbpa5.

In SUDAAN, we can easily obtain means adjusted to the overall age mean without having to write the ESTIMATE statements explicitly as in SAS. Simply specify the interactions in the MODEL statement and request ALL (or just the interaction term) in the CONDMARG statement.

```
proc regress data=sol filetype=sas design=wr;
    nest strat PSU_ID; weight weight_final_norm_overall;
    subpopn KEEP_DATA=1; class bkgrd1_c7 female;
    model sbpa5 = bkgrd1_c7 age female bkgrd1_c7*female age*female;
    condmarg  / all;
run;
```

Output 7.5.2b. Age-adjusted mean SPB by background stratified by gender – REGRESS with interaction.

Conditional Marginal #1		Conditional Marginal	SE	T:Marg-=0	P-value
7-level re-classification of Hispanic/Latino Background, 1-female 0-male	Dominican, 0	127.648	0.798	160.042	0.000
	Dominican, 1	118.809	0.599	198.292	0.000
	Central American, 0	124.756	0.841	148.402	0.000
	Central American, 1	118.098	0.665	177.597	0.000
	Cuban, 0	123.576	0.565	218.543	0.000
	Cuban, 1	117.848	0.481	245.239	0.000
	Mexican, 0	123.011	0.517	237.922	0.000
	Mexican, 1	113.648	0.465	244.597	0.000
	Puerto Rican, 0	123.636	0.684	180.756	0.000
	Puerto Rican, 1	117.784	0.635	185.431	0.000
	South American, 0	121.132	0.929	130.331	0.000
	South American, 1	113.645	0.740	153.647	0.000
	Mixed/Other, 0	124.038	1.030	120.477	0.000
	Mixed/Other, 1	116.418	1.161	100.281	0.000

NOTE: These age-adjusted SBP means by gender from Output 7.5.2b are adjusted to the overall age mean, and are almost identical to those obtained from explicitly stating the ESTIMATE contrast using the overall age mean (Output 7.5.1b). Estimates are slightly different, since Output 7.5.2b use estimates of the weighted mean age among those without missing values in sbpa5.

NOTE: For linear regression models, the least squares means and conditional marginal means yield equivalent results for estimated means and associated standard errors. The predicted marginal means are also the same, but are different for nonlinear models such as logistic regression. However, the standard errors for predicted marginal means differ slightly from those for conditional marginal means because, for predicted marginal means, the variability of the covariates (which are considered as random variables) is taken into account. In practice, LSMEANS or CONDMARG are the most commonly used.

8. LOGISTIC REGRESSION MODELS TO ESTIMATE EFFECTS

Logistic regression models are used to model categorical outcomes (binary, ordinal or nominal) and estimate adjusted odds ratios (OR).

8.1. Logistic regression model for a binary outcome

8.1.1. SAS

The next group of SAS statements invokes SURVEYLOGISTIC to produce logistic regression estimates that appropriately account for the study design. The response level ordering is important because, by default, PROC SURVEYLOGISTIC models the probabilities of response levels with lower ordered values. This means that for binary variables coded 0/1 with 1 denoting the event SURVEYLOGISTIC models the probability of nonevent (0). One way to change the default is to use the DESCENDING option. Note that in SAS 9.2 the DESCENDING option in SURVEYLOGISTIC is specified in parenthesis in the MODEL statement not in the PROC statement as it is in PROC LOGISTIC. Another way to change the default is to specify option REF= '0' in the MODEL statement as the nonevent category for the response variable.

```
proc surveylogistic data = sol;
   strata strat;
   cluster PSU_ID;
   weight weight_final_norm_overall;
   domain KEEP_DATA;
   class bkgrd1_c7(ref='Mexican');
   model DIABETES3_C2(descending) = bmi bkgrd1_c7 age female;
run;
```

NOTE: Unlike PROC SURVEYREG in PROC SURVEYLOGISTIC, the reference category for any categorical variable can be easily changed. By default, the levels are sorted by the formatted values, so the last variable according to the external formats is the reference category. The ORDER=INTERNAL statement can be used to establish the internally ordered variable as the reference category. Note that the ORDER= option is on the CLASS statement, not on the PROC statement as in the PROC SURVEYREG statement. If another reference category is desired, the formatted level is specified in the CLASS statement. Unlike SUDAAN, in PROC SURVEYLOGISTIC the actual formatted level is specified (e.g., 'Mexican') rather than the internal value (3).

Output 8.1.1 Association of BMI and demographic characteristics with Diabetes, PROC SURVEYLOGISTIC

Probability modeled is DIABETES3_C2=1.

Type 3 Analysis of Effects

Effect	*DF*	*Wald Chi-Square*	*Pr > ChiSq*
BMI	1	196.7382	<.0001
BKGRD1_C7	6	50.0921	<.0001
AGE	1	698.2752	<.0001
female	1	3.3819	0.0659

Odds Ratio Estimates			
Effect	*Point Estimate*	*95% Wald Confidence Limits*	
BMI	1.092	1.079	1.106
BKGRD1_C7 Central American vs Mexican	0.848	0.659	1.091
BKGRD1_C7 Cuban vs Mexican	0.558	0.447	0.698
BKGRD1_C7 Dominican vs Mexican	0.897	0.683	1.177
BKGRD1_C7 Mixed/Other vs Mexican	0.962	0.533	1.763
BKGRD1_C7 Puerto Rican vs Mexican	0.851	0.681	1.063
BKGRD1_C7 South American vs Mexican	0.379	0.269	0.534
AGE	1.088	1.081	1.094
FEMALE	0.874	0.756	1.010

NOTE: SURVEYLOGISTIC in SAS version 9.1 does not allow for subpopulation analyses, but version 9.2 does.

Code below models the association between BMI, Hispanic background and age with diabetes stratifying by gender using the DOMAIN statement.

```
proc surveylogistic data = sol;
   strata strat;
   cluster PSU_ID;
   weight weight_final_norm_overall;
   domain KEEP_DATA*female;
   class bkgrd1_c7(ref='Mexican');
   model DIABETES3_C2(descending) = bmi bkgrd1_c7 age;
run;
```

8.1.2. SUDAAN

RLOGIST produces logistic regression estimates that appropriately account for the study design.

```
proc rlogist data=sol filetype=sas design=wr;
   nest strat PSU_ID; weight weight_final_norm_overall;
   subpopn KEEP_DATA=1;
   class bkgrd1_c7;
   reflevel bkgrd1_c7 = 3; /* reference: Mexicans */
   model DIABETES3_C2 = bmi Bkgrd1_c7 age female;
run;
```

NOTE: To change the reference level for any categorical variable in SUDAAN, a REFLEVEL statement must be included followed by variable name, an equal sign and the unformatted numerical value of the variable you wish to distinguish as the reference level. The previous code requests Mexicans (3) as the reference level for bkgrd1_c7. Otherwise, the default reference category will be the highest internal value (5). Unlike SAS, in PROC RLOGIST the internal value is always used rather than the actual formatted level.

OUTPUT 8.1.2. Association of BMI and demographic characteristics with Diabetes, PROC RLOGIST

Contrast	Degrees of Freedom	Wald ChiSq	P-value Wald ChiSq
OVERALL MODEL	10	1955.90	0.0000
MODEL MINUS INTERCEPT	9	862.72	0.0000
INTERCEPT	.	.	.
BMI	1	196.58	0.0000
BKGRD1_C7	6	50.03	0.0000
AGE	1	698.53	0.0000
FEMALE	1	3.33	0.0658

Independent Variables and Effects		Odds Ratio	Lower 95% Limit OR	Upper 95% Limit OR
Intercept		0.0003	0.0002	0.0006
BMI (kg/m2)		1.0923	1.0789	1.1059
7-level re-classification of Hispanic/Latino Background	Dominican	0.8968	0.6832	1.1772
	Central American	0.8480	0.6591	1.0912
	Cuban	0.5583	0.4465	0.6981
	Mexican	1.0000	1.0000	1.0000
	Puerto Rican	0.8512	0.6811	1.0639
	South American	0.3792	0.2690	0.5345
	Mixed/Other	0.9622	0.5429	1.7374
Age		1.0875	1.0808	1.0943
1-female 0-male		0.8738	0.7558	1.0103

The SUBPOPN statement can be used to calculate estimates for a specified subpopulation.

```
proc rlogist data=sol filetype=sas design=wr;
   nest strat PSU_ID; weight weight_final_norm_overall;
   subpopn KEEP_DATA=1 & female=1;
   class bkgrd1_c7;
   reflevel bkgrd1_c7=3; /* reference: Mexicans */
   model DIABETES3_C2 = bmi bkgrd1_c7 age;
run;
```

8.2. Cumulative logit model (proportional odds model)

8.2.1. SAS

PROC SURVEYLOGISTIC also allows fitting cumulative logit models, which are appropriate for ordinal outcomes when the proportional odds assumption is met. This assumes that the odds, for all covariates, of being in an outcome category less than or equal to some value are the same regardless of the cutoff point. For example, diabetes3 is an ordinal variable with three levels: non-diabetic, pre-diabetic and diabetic. So, in a proportional odds model, we are modeling the logit of diabetic vs. (pre-diabetic or non-diabetic) simultaneously with the logit of (diabetic or pre-diabetic) vs. non-diabetic. The proportional odds assumption means that the association (OR) between a covariate and the outcome is the same for both logits.

NOTE: The cumulative logit model is the default in PROC SURVEYLOGISTIC when a multi-level outcome variable is specified. It can also be explicitly defined by using the LINK=CLOGIT option in the MODEL statement.

WARNING: It is very important to note how SAS orders the response variable. Unless the variables are in the correct order alphabetically, the ORDER=INTERNAL statement is necessary to order the response variable numerically (which is usually what is desired). This ORDER=INTERNAL statement must go on the CLASS statement. When using the ORDER=INTERNAL statement, this orders the variables from the lowest value to the highest. In the case of the diabetes3 variable (1='Non-diabetic', 2='Pre-diabetic' or 3='Diabetic'). Hence, the logits modeled will be non-diabetic vs. (pre-diabetic and diabetic) simultaneously with (non-diabetic or pre-diabetic) vs. diabetic. These are not the models we are interested in modeling and, therefore, the DESCENDING option can be used after the response variable in the model statement to switch the default order (from lowest to highest) to highest to lowest.

The order of the response variable should always be checked in the output. Output 8.2.1 shows the order specified as 'Diabetic,' 'Pre-diabetic' and then 'Non-diabetic.' This means we are modeling the logit of diabetic vs. (pre-diabetic or non-diabetic) simultaneously with the logit of (diabetic or pre-diabetic) vs. non-diabetic.

```
proc surveylogistic data=sol;
    strata strat;
    cluster PSU_ID;
    weight weight_final_norm_overall;
    domain KEEP_DATA;
    class diabetes3 bkgrd1_c7(ref='Mexican') / order=internal;
    model diabetes3 (descending) = bmi bkgrd1_c7 age female / link=clogit;
run;
```

Output 8.2.1. Cumulative logit model (proportional odds model) for diabetes 3, PROC SURVEYLOGISTIC

Response Profile

Ordered Value	*DIABETES3*	*Total Frequency*	*Total Weight*
1	Diabetic	2356	1908.4878
2	Pre-diabetic	4354	4190.1072
3	Non-diabetic	4991	6069.4364

Probabilities modeled are cumulated over the lower Ordered Values.

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
BMI	1	241.6796	<.0001
BKGRD1_C7	6	54.6350	<.0001
AGE	1	1179.9912	<.0001
female	1	44.7790	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
BMI	1.095	1.083	1.108
BKGRD1_C7 Dominican vs Mexican	0.770	0.614	0.965
BKGRD1_C7 Central American vs Mexican	0.858	0.694	1.059
BKGRD1_C7 Cuban vs Mexican	0.596	0.506	0.701
BKGRD1_C7 Puerto Rican vs Mexican	0.789	0.663	0.939
BKGRD1_C7 South American vs Mexican	0.521	0.417	0.651
BKGRD1_C7 Mixed/Other vs Mexican	0.925	0.664	1.288
AGE	1.082	1.077	1.087
female	0.688	0.616	0.767

Score Test for the Proportional Odds Assumption		
Chi-Square	DF	Pr > ChiSq
57.3696	9	<.0001

PROC SURVEYLOGISTIC tests the proportional odds assumption. Even if the assumption does not hold the output still includes parameter estimates. If the proportional odds assumption does not hold, individual indicator variables will need to be created for each level of the outcome variable and multiple logistic regression models would need to be fitted using the same reference group. This latter model is commonly referred to as the generalized logit model which is used for nominal outcomes as well as ordinal outcomes that violate the proportion odds assumption. The proportional odds test in Output 8.2.1 is highly significant (p-value < 0.0001), indicating that the assumption of proportional odds does not hold; therefore, the previous estimates should not be used.

8.2.2. SUDAAN

Unlike SAS, which handles ordinal multi-level outcomes in the same procedure as binary outcomes, SUDAAN utilizes a different procedure, PROC MULTILOG, to fit ordinal or

nominal outcomes. The cumulative logit model (proportional odds model) is specified by the CUMLOGIT option the MODEL statement.

```
proc multilog data=sol filetype=sas design=wr;
   nest strat PSU_ID;
   weight weight_final_norm_overall;
   subpopn KEEP_DATA=1;
   class diabetes3 bkgrd1_c7 / dir=descending;
   reflevel bkgrd1_c7=3; /* reference: Mexicans */
   model diabetes3 = bmi bkgrd1_c7 age female / cumlogit;
run;
```

The default with MULTILOG is to make the highest internal value as the lowest level, which is always included in the reference category. If the lowest internal value is actually the lowest level, the option DIR=DESCENDING can be specified on the CLASS statement. This will make all CLASS variables by descending order; so, if another reference group is desired for these, it must be specified on the REFLEVEL statement. The order of the response variables can be seen below. Probabilities modeled are cumulated over the higher ordered values.

```
-------------------------------------------------------------------------
  3-level grouped Diabetes
  includes self-report              Frequency            Value
-------------------------------------------------------------------------
Ordered Position: 1                 2405                 Diabetic
Ordered Position: 2                 4378                 Pre-diabetic
Ordered Position: 3                 5018                 Non-diabetic
-------------------------------------------------------------------------
```

Output 8.2.2. Cumulative logit model (proportional odds model) for diabetes3, PROC MULTILOG

Contrast	Degrees of Freedom	Wald ChiSq	P-value Wald ChiSq
OVERALL MODEL	11	3442.12	0.0000
MODEL MINUS INTERCEPT	9	1401.05	0.0000
BMI	1	233.98	0.0000
BKGRD1_C7	6	55.43	0.0000
AGE	1	1169.85	0.0000
FEMALE	1	44.65	0.0000

DIABETES3 (cum- logit), Independent Variables and Effects		Odds Ratio	Lower 95% Limit OR	Upper 95% Limit OR
DIABETES3 (cum- logit) Intercept 1:	Diabetic	0.0005	0.0003	0.0007
	Intercept 2: Pre-diabetic	0.0043	0.0027	0.0067
BMI (kg/m2)		1.0952	1.0825	1.1080
7-level re-classification of Hispanic/Latino Background	Mixed/Other	0.9249	0.6650	1.2863
	South American	0.5208	0.4181	0.6488
	Puerto Rican	0.7889	0.6617	0.9406
	Mexican	1.0000	1.0000	1.0000
	Cuban	0.5956	0.5060	0.7011
	Central American	0.8575	0.6938	1.0600
	Dominican	0.7698	0.6131	0.9666
Age		1.0819	1.0770	1.0868
1-female 0-male		0.6876	0.6159	0.7676

8.3. Generalized logit model

8.3.1. SAS code

If the proportional odds assumption does not hold, individual indicator variables will need to be created for each level of the outcome variable and multiple logistic regression models would need to be fitted using the same reference group. This generalized logit model is available in PROC SURVEYLOGISTIC by specifying the LINK=GLOGIT in the MODEL statement.

By default, SAS uses the largest value (in this case diabetes3=3) as the reference group. This can be changed by including the DESCENDING option in the MODEL statement.

```
proc surveylogistic data=sol;
    strata strat;
    cluster PSU_ID;
    weight weight_final_norm_overall;
    domain KEEP_DATA;
    class diabetes3 bkgrd1_c7(ref='Mexican') / order=internal param=ref;
    model diabetes3 (descending)   = bmi bkgrd1_c7 age female / link=glogit;
run;
```

Output 8.3.1. Generalized logit model for diabetes3, PROC SURVEYLOGISTIC

Response Profile

Ordered Value	*DIABETES3*	*Total Frequency*	*Total Weight*
1	Diabetic	2356	1908.4878
2	Pre-diabetic	4354	4190.1072
3	Non-diabetic	4991	6069.4364

Logits modeled use DIABETES3='Non-diabetic' as the reference category.

Type 3 Analysis of Effects

Effect	*DF*	*Wald Chi-Square*	*Pr > ChiSq*
BMI	2	248.8300	<.0001
BKGRD1_C7	12	71.0933	<.0001
AGE	2	1061.9938	<.0001
female	2	58.3951	<.0001

Odds Ratio Estimates

Effect	*DIABETES3*	*Point Estimate*	*95% Wald Confidence Limits*	
BMI	*Diabetic*	1.147	1.128	1.167
BMI	*Pre-diabetic*	1.085	1.070	1.101
BKGRD1_C7 Dominican vs Mexican	*Diabetic*	0.729	0.522	1.019
BKGRD1_C7 Dominican vs Mexican	*Pre-diabetic*	0.713	0.554	0.918
BKGRD1_C7 Central American vs Mexican	*Diabetic*	0.770	0.559	1.061
BKGRD1_C7 Central American vs Mexican	*Pre-diabetic*	0.854	0.683	1.069
BKGRD1_C7 Cuban vs Mexican	*Diabetic*	0.435	0.338	0.560
BKGRD1_C7 Cuban vs Mexican	*Pre-diabetic*	0.657	0.542	0.796
BKGRD1_C7 Puerto Rican vs Mexican	*Diabetic*	0.704	0.540	0.919
BKGRD1_C7 Puerto Rican vs Mexican	*Pre-diabetic*	0.733	0.595	0.904
BKGRD1_C7 South American vs Mexican	*Diabetic*	0.297	0.204	0.434
BKGRD1_C7 South American vs Mexican	*Pre-diabetic*	0.674	0.522	0.869
BKGRD1_C7 Mixed/Other vs Mexican	*Diabetic*	0.894	0.487	1.639
BKGRD1_C7 Mixed/Other vs Mexican	*Pre-diabetic*	0.866	0.602	1.245
AGE	*Diabetic*	1.129	1.120	1.137
AGE	*Pre-diabetic*	1.066	1.060	1.072
female	*Diabetic*	0.656	0.553	0.780
female	*Pre-diabetic*	0.619	0.547	0.700

8.3.2. SUDAAN code

If the proportional odds assumption does not hold or the outcome is nominal, we can fit a generalized logit model by specifying the GENLOGIT option in the MODEL statement.

```
proc multilog data=sol filetype=sas design=wr;
   nest strat PSU_ID; weight weight_final_norm_overall;
   subpopn KEEP_DATA=1;
   class diabetes3 bkgrd1_c7 / dir=descending;
   reflevel bkgrd1_c7=3; /* reference: Mexicans */
   model diabetes3 = bmi bkgrd1_c7 age female / genlogit;
run;
```

Output 8.3.2. Generalized logit model for diabetes3, PROC MULTILOG

Contrast	Degrees of Freedom	Wald ChiSq	P-value Wald ChiSq
OVERALL MODEL	20	1599.26	0.0000
MODEL MINUS INTERCEPT	18	1295.53	0.0000
INTERCEPT	.	.	.
BMI	2	249.13	0.0000
BKGRD1_C7	12	71.18	0.0000
AGE	2	1063.21	0.0000
FEMALE	2	58.46	0.0000

The output gives parameter estimates and odds ratios separately for each level of the response variable, compared to the reference response level (in this case 'Non-diabetic'). For brevity, we have only included the OR for the model comparing pre-diabetic to non-diabetic.

DIABETES3 (log-odds)	Independent Variables and Effects		Odds Ratio	Lower 95% Limit OR	Upper 95% Limit OR
Pre-diabetic vs Non- diabetic	Intercept		0.0081	0.0047	0.0138
		BMI (kg/m2)	1.0850	1.0697	1.1005
		7-level re-classification of Hispanic/Latino Background Mixed/Other	0.8659	0.6021	1.2451
		South American	0.6737	0.5221	0.8695
		Puerto Rican	0.7334	0.5945	0.9047
		Mexican	1.0000	1.0000	1.0000
		Cuban	0.6567	0.5418	0.7961
		Central American	0.8542	0.6824	1.0691
		Dominican	0.7130	0.5534	0.9186
		Age	1.0662	1.0601	1.0722
		1-female 0-male	0.6188	0.5468	0.7003

9. ADJUSTED AND STANDARDIZED PREVALENCES WITH LINEAR OR LOGISTIC REGRESSION

9.1. Introduction

When evaluating categorical outcomes from a survey sample such as HCHS/SOL, there are two specific goals, which may require different statistical methods.

- **Prevalence Estimation** provides **descriptive statistics** used to describe the rate or distribution of the outcome in the target population or population subgroups.
- **Identifying predictors and quantifying differences between levels of categorical predictors** requires fitting statistical models to calculate odds ratios and adjust for potential covariates**.**

When the purpose of the analysis is to ***identify predictors*** for a categorical outcome, or ***quantify differences between levels of categorical predictors with odds ratios (ORs), logistic regression*** models should be used, especially when adjusting for a number of additional covariates. Survey linear models are not used in this situation. See section 8 for examples. **Prevalence estimates** may be obtained from a variety of ways, each with advantages and disadvantages. **Regardless of the method used, we recommend avoiding statistical comparison of subgroup prevalence with p-values, and instead limit presentations to standard errors or confidence intervals**. P-values are more appropriately presented when the goal is to identify predictors or quantify differences after adjusting for all relevant covariates.

In this section, we show that we can use survey linear or logistic models to estimate 1) prevalence, 2) adjusted prevalence (internally), and 3) adjusted prevalence to an external population (standardization).

9.2. Methods for estimating prevalence

Prevalence estimates can be obtained in a variety of ways. Estimates of interest include the following:

1) Prevalence in the population or population subgroups,
2) Adjusted prevalence (internally),
3) Adjusted prevalence to an external population (standardization),
4) Differences in adjusted prevalence between population subgroups.

9.2.1. Use of survey linear regression

Although not regularly considered for binary outcomes, survey linear regression models are helpful when the purpose of the analysis is to ***estimate the prevalence of a binary outcome in the target population, or describe the variation in prevalence across subgroups of the target population*** (Hellevik, 2009).

There is often an interest in quantifying the difference in adjusted prevalence between population subgroups, and the survey linear model can be used to calculate the estimate, SE and confidence interval of pairwise subgroup prevalence differences via contrast statements.

In complex survey designs, the estimated **population prevalence from survey linear models** is an ***assumption-free design-based estimate***. The survey linear model is equivalent to ***weighted least squares*** (WLS). The only assumption for WLS is that the data comes from a probability random sample. There is no assumption of normality of residuals, and hence the variances and p-values are valid, even if the outcome is dichotomous, although sufficient sample size is required within each subgroup for the large sample approximation to the normal distribution to apply (Koch, Gillings, Stokes, 1980). See examples in section 9.4.

9.2.2. Use of survey logistic regression

Survey logistic regression models are also helpful when the purpose of the analysis is to estimate the prevalence of a categorical outcome in the target population, or describe the variation in prevalence across subgroups of the target population.

Logistic regression provides odds ratio comparisons of subgroups (on a relative scale), which may be less intuitive or relevant than differences (on an absolute scale, obtained by survey linear regression) for the comparisons of prevalence in population subgroups. However, survey logistic regression also calculates prevalence estimates (also called risk estimates or marginals), as well as differences of these estimates.

Two marginal prevalence estimates are available from survey logistic regression.

> **Conditional marginals** provide an estimate of the subgroup prevalence for a hypothetical subject assuming that the subject is at the weighted average value of all continuous covariates in the model, or a weighted average of the categorical covariates (similar to a least squares mean). Other values can be selected other than the weighted average if there was interest in estimating the prevalence for one specific value of a covariate. For example, a conditional marginal adjusting for age is calculated at the weighted mean age of the sample; only this one numeric age value is used in the calculations***. Conditional marginals can be helpful when the goal is to adjust estimates to a particular mean value*** in order to compare prevalence ***externally,*** such as to a different population or a different study.
>
> **Predicted marginals** provide an estimate of the subgroup prevalence that is a weighted average of all participants in the sample. For example, predicted marginals adjusting for age are based on the weighted average of the observed age for all participants in the sample, which is then applied to the marginal calculation within each subgroup. Predicted marginals ***adjust to the distribution of the target population***, and are ***most helpful for internal*** comparisons of subgroups within the target population.

Specifically, let *A* be a categorical variable and *a* represent a given level of *A*. An estimate of the **conditional marginal** for category A=a is equivalent to the estimated probability for ***a person*** with *A=a* and with all other characteristics equal to the weighted average for those variables in the observed population. An estimate of the **predicted marginal** for a

given level of A is equivalent to the estimated prevalence for ***a hypothetical population*** (or standardized population) of individuals such that every individual has A=a with all other characteristics the same as those in the observed population.

As we will see in our example, adjusting for age in the next section, the **predicted marginal tends to be closer to the prevalence estimate from the linear regression**.

When planning and reporting prevalence estimates using logistic regression, authors should clearly specify which of the two marginal estimates (conditional or predicted) will be reported, and clearly describe their interpretation. This is not necessary for linear regression models where the least squares means, conditional marginal means, and predicted marginal means are the same.

Figure 9.1. Formulas for conditional and predicted marginals from the SUDAAN manual (SUDAAN Release 11 Language Manual, section 19.9.3)

Conditional and Predicted Marginals (Model-Adjusted Risks)

The estimated ***conditional marginal*** from the logistic model is:

$$\mathbf{M}_C = \frac{\exp(\bar{\mathbf{x}}^{*\prime}\hat{\boldsymbol{\beta}})}{1+\exp(\bar{\mathbf{x}}^{*\prime}\hat{\boldsymbol{\beta}})}, \text{ where } \bar{\mathbf{x}}^{*\prime} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{I_h}\sum_{j=1}^{J_{hi}}\sum_{k=1}^{K_{hij}} w_{hijk}\mathbf{x}'^{*}_{hijk}}{\sum_{h=1}^{H}\sum_{i=1}^{I_h}\sum_{j=1}^{J_{hi}}\sum_{k=1}^{K_{hij}} w_{hijk}},$$

and the estimated ***predicted marginal*** is:

$$\mathbf{M}_P = \frac{1}{w_{++++}}\sum_{h=1}^{H}\sum_{i=1}^{I_h}\sum_{j=1}^{J_{hi}}\sum_{k=1}^{K_{hij}} w_{hijk}\frac{\exp(\mathbf{x}'^{*}_{hijk}\hat{\boldsymbol{\beta}})}{1+\exp(\mathbf{x}'^{*}_{hijk}\hat{\boldsymbol{\beta}})}.$$

The vector $\mathbf{M}_P$ contains the estimates of the predicted marginal for each level of a categorical covariate or each user-specified value for a continuous covariate. The vector of predicted marginals is written as:

$$\mathbf{M}'_P = \left[M_{P,X=\mathbf{x}^*_1} \quad M_{P,X=\mathbf{x}^*_2} \quad \cdots \quad M_{P,X=\mathbf{x}^*_m}\right],$$

where $M_{P,X=\mathbf{x}^*_1}$ corresponds to the predicted marginal for the first level of the *m*-level covariate, and so on. An analogous argument applies to the vector of conditional marginals, $\mathbf{M}_C$. The estimated standard error of the vector of marginals is determined using design-based variance estimation methods specified by the user.

9.2.3. Comparison between survey linear and logistic regression

Often, linear and logistic models are equally effective for investigating the variation of prevalence among population subgroups. The choice among them may be a matter of personal taste or computational convenience. On the other hand, for some applications, one of these models may be preferable because of more parsimonious structure or easier interpretation.

- **Linear models**, which directly estimate prevalence and difference of prevalence, have the technical advantage of estimating the population prevalence with an expected value. However, linear models may yield predicted values outside the range of 0-100. In the linear model, the conditional and predicted marginal are identical.
- **Logistic models**, which directly estimate odds ratios, calculate estimates of the prevalence with marginals. Logistic models have the technical advantage of always yielding predicted marginal estimates of prevalence in the range of 0-100. Similarly, multi-level categorical outcomes can be estimated by multi-log models. Because the model is non-linear, standard errors may tend to change based on the number of covariates in the model, and conditional and predicted marginals are not the same.

For large samples, as in most cases for HCHS/SOL, provided we limit prevalence estimates to large subgroups, linear models rarely yield predicted values outside the range of 0-100 (Hellevik 2009). We can assess the goodness of fit of either model via the Wald goodness of fit chi-square test. Generally a model with a reasonable goodness of fit test will not have predicted estimates outside 0-100 (Koch, Gillings, Stokes, 1980).

For rare events (for example, < 10%), use of continuous predictors in the survey linear regression might produce prevalence estimates that are negative. **To avoid negative estimates, continuous predictors should be avoided in survey linear regression estimation of prevalence**. For example, with survey linear regression, adjusting for age groups is preferable to adjusting for continuous age to obtain age-adjusted prevalence. For this reason, it may be preferable to use survey logistic regression to estimate prevalence when event rates are small.

9.3. Recommended wording for manuscript methods sections

Linear regression:

[Continuous or categorical] age-adjusted prevalence estimates for the target population of Hispanic/Latinos in the 4 HCHS/SOL communities were calculated using survey regression weighted least squares, which adjusts each subgroup to the age distribution of the target population.

Logistic regression with predicted marginals:

[Continuous or categorical] age-adjusted prevalence estimates for the target population of Hispanic/Latinos in the 4 HCHS/SOL communities were calculated using survey logistic

regression predicted marginals, which adjust each subgroup to the age distribution of the target population.

Logistic regression with conditional marginals:

[Continuous or categorical] age-adjusted prevalence estimates for the target population of Hispanic/Latinos in the 4 HCHS/SOL communities were calculated using survey logistic regression conditional marginals, which adjust each subgroup to the [weighted mean age of the target population xx.x, see Table x] or [age of 60].

9.4. Hypertension prevalence estimation examples

In this section, we show that we can use survey linear or logistic regression models to estimate 1) prevalence, 2) adjusted prevalence (internally), and 3) adjusted prevalence to an external population (standardization). Lastly, we illustrate how to adjust for site and, when sample size is appropriate, how to compare prevalence between sites for a particular Hispanic background of interest.

Here we show that we obtain the same prevalence estimates (and corresponding SE) using SUDAAN procedures CROSSTAB, DESCRIPT and REGRESS. For example, the unadjusted prevalence of hypertension for males and females using PROC CROSSTAB is given by:

```
proc crosstab data=sol filetype=sas design=wr;
   nest strat PSU_ID;
   weight weight_final_norm_overall;
   subpopn KEEP_DATA=1;
   class hypertension2 female;
   tables female*hypertension2;
run;
```

Output 9.4a Hypertension prevalence by gender, PROC CROSSTAB

1-female 0-male	Hypertension using NHANES definition	Sample Size	Tot Percent	Row Percent	SE Row Percent
Total	Total	11814	100.00	100.00	0.00
	0	8285	75.87	75.87	0.67
	1	3529	24.13	24.13	0.67
0	Total	4755	48.19	100.00	0.00
	0	3404	36.82	76.41	0.93
	1	1351	11.37	23.59	0.93
1	Total	7059	51.81	100.00	0.00
	0	4881	39.06	75.38	0.79
	1	2178	12.76	24.62	0.79

PROC DESCRIPT produces descriptive statistics for continuous and categorical variables; hence, we can use PROC DESCRIPT to estimate proportions (e.g. prevalence) instead of PROC CROSSTAB (section 6). For convenience, we use the CATLEVEL statement in PROC DESCRIPT to specify that we are dealing with a categorical variable and that we are interested in a percentage rather than a mean (proportion). Hence, for a dichotomous variable coded 0/1 (e.g. hypertension2), the output is the percentage 24.13 for overall hypertension prevalence rather than the proportion 0.2413. Similarly, for multilevel categorical variables, the CATLEVEL statement specifies which level of the variable one is interested in. Only the proportion for one level at a time can be calculated.

```
proc descript data=sol filetype=sas design=wr;
   nest strat PSU_ID; weight weight_final_norm_overall;
   subpopn KEEP_DATA=1;
   class hypertension2 female;
   var hypertension2;
   catlevel 1; /* Output only % of hypertension (i.e. hypertension2=1) */
run;
```

The prevalence of hypertension by gender in Output 9.4a specified with PROC CROSSTAB is identical to Output 9.4b below estimated using PROC DESCRIPT.

Output 9.4b. Hypertension prevalence by gender, PROC DESCRIPT

Variable	1-female 0-male	Sample Size	Percent	SE Percent
Hypertension using NHANES definition: 1	Total	11814	24.13	0.67
	0	4755	23.59	0.93
	1	7059	24.62	0.79

Now, we use PROC REGRESS to estimate the prevalence of hypertension by gender. In the MODEL statement we specify 1) the binary variable hypertension2 as the outcome, 2) female indicator as the only covariate and 3) NOINT option to not include an intercept in the model. The regression coefficients from this linear regression model are the prevalence of hypertension by gender. Because there are no other covariates in the model, they match the conditional marginals exactly.

```
proc regress data=SOL filetype=sas design=wr;
    nest strat PSU_ID; weight weight_final_norm_overall;
    subpopn KEEP_DATA=1;
    class female;
    model hypertension2 = female / noint;
    condmarg / all;
run;
```

The prevalence of hypertension by gender in Output 9.4c estimated from the linear model (specified with PROC REGRESS) is identical to estimates from Output 9.4a and 9.4b which are from specific procedures for categorical outcomes.

Output 9.4c Hypertension prevalence by gender, PROC REGRESS using hypertension coded 0-1

Independent Variables and Effects		Beta Coeff.	SE Beta	T-Test B=0	P-value T-Test B=0
1-female 0-male	0	0.236	0.009	25.329	0.000
	1	0.246	0.008	31.028	0.000
Conditional Marginal #1		**Conditional Marginal**	**SE**	**T:Marg-=0**	**P-value**
1-female 0-male	0	0.236	0.009	25.329	0.000
	1	0.246	0.008	31.028	0.000

NOTE: If we use the trick of creating and using an outcome coded 0 and 100 (e.g. hypert) instead of 0 and 1 (e.g. hypertension2), we get percentages and not proportions.

```
proc regress data=SOL filetype=sas design=wr;
    nest strat PSU_ID; weight weight_final_norm_overall;
    subpopn KEEP_DATA=1;
    class female;
    model hypert = female / noint;
    condmarg / all;
run;
```

Output 9.4d Hypertension prevalence by gender, PROC REGRESS using hypertension coded 0-100

Independent Variables and Effects		Beta Coeff.	SE Beta	T-Test B=0	P-value T-Test B=0
1-female 0-male	0	23.589	0.931	25.329	0.000
	1	24.624	0.794	31.028	0.000

Conditional Marginal #1		Conditional Marginal	SE	T:Marg-=0	P-value
1-female 0-male	0	23.589	0.931	25.329	0.000
	1	24.624	0.794	31.028	0.000

The following SUDAAN code using PROC DESCRIPT estimates the weighted hypertension prevalence by Hispanic Background and 10-year NHANES age groups.

```
proc descript data=sol filetype=sas design=wr;
   nest strat PSU_ID;
   weight weight_final_norm_overall;
   subpopn KEEP_DATA=1;
   class hypertension2 agegroup_c6_nhanes;
   var hypertension2;
   catlevel 1; /* Output only % of hypertension (i.e. hypertension2=1) */
run;
```

Output 9.4e Hypertension prevalence by age group, PROC DESCRIPT

Variable	6-Level Age Groups (NHANES standardization)	Sample Size	Percent	SE Percent
Hypertension using NHANES definition: 1	Total	11805	24.09	0.67
	18-29 yrs.	1985	2.35	0.38
	30-39 yrs.	1731	10.05	1.05
	40-49 yrs.	3020	21.57	1.01
	50-59 yrs.	3061	45.45	1.32
	60-69 yrs.	1641	65.50	1.79
	70-74 yrs.	367	78.77	3.21

The following SUDAAN code using PROC DESCRIPT estimates the weighted hypertension prevalence by Hispanic/Latino background.

```
proc descript data=sol filetype=sas design=wr;
   nest strat PSU_ID;
   weight weight_final_norm_overall;
   subpopn KEEP_DATA=1;
   class hypertension2 bkgrd1_c7 agegroup_c6_nhanes;
   var hypertension2;
   catlevel 1; /* Output only % of hypertension (i.e. hypertension2=1) */
run;
```

Output 9.4f Hypertension prevalence by background, PROC DESCRIPT
WARNING: Unadjusted prevalence.

Variable	7-level re-classification of Hispanic/Latino Background	Sample Size	Percent	SE Percent
Hypertension using NHANES definition: 1	Total	11746	24.09	0.67
	Dominican	1001	28.15	1.85
	Central American	1370	20.63	1.33
	Cuban	1668	35.64	1.51
	Mexican	4621	16.71	1.09
	Puerto Rican	1958	30.67	1.38
	South American	758	19.57	2.03
	Mixed/Other	370	15.44	2.99

9.4.1 Unadjusted prevalence by Hispanic/Latino background

The following SUDAAN code using PROC REGRESS and PROC RLOGISTIC also estimates the weighted hypertension prevalence by Hispanic/Latino background. Note that **ALL prevalence estimates are the same** from DESCRIPT, survey linear regression, and survey logistic regression conditional and predicted marginals.

```
proc regress data=sol filetype=sas design=wr;
      nest strat PSU_ID ;
      weight weight_final_norm_overall;
      subpopn KEEP_DATA=1;
      class bkgrd1_c7;
      model hypert = bkgrd1_c7;
      condmarg / all;
run;
```

Output 9.4.1a Hypertension prevalence by background, PROC REGRESS
WARNING: Unadjusted prevalence

Conditional Marginal		Conditional Marginal	SE	T:Marg-=0	P-value
Intercept		24.09	0.64	37.68	0.00
7-level re-classification of Hispanic/Latino Background	Dominican	28.15	1.85	15.23	0.00
	Central American	20.63	1.33	15.49	0.00
	Cuban	35.64	1.51	23.68	0.00
	Mexican	16.71	1.09	15.30	0.00
	Puerto Rican	30.67	1.38	22.28	0.00
	South American	19.57	2.03	9.67	0.00
	Mixed/Other	15.44	2.99	5.17	0.00

```
proc rlogist data=sol filetype=sas design=wr deft4;
   nest strat PSU_ID / NOSORTCK; weight weight_final_norm_overall;
   class bkgrd1_c7;
   subpopn KEEP_DATA=1;
   reflevel bkgrd1_c7=3;
   model hypertension2 =  bkgrd1_C7;
   setenv decwidth=4;
   CONDMARG bkgrd1_C7;
   PREDMARG bkgrd1_C7;
run;
```

Output 9.4.1b Hypertension prevalence by background, PROC RLOGIST
WARNING: Unadjusted prevalence

Predicted Marginal		Predicted Marginal	SE	Lower 95% Limit	Upper 95% Limit	T:Marg=0	P-value
7-level re-classification of Hispanic/Latino Background	Dominican	0.2815	0.0185	0.2467	0.3192	15.2328	0.0000
	Central American	0.2063	0.0133	0.1814	0.2337	15.4945	0.0000
	Cuban	0.3564	0.0151	0.3274	0.3865	23.6815	0.0000
	Mexican	0.1671	0.0109	0.1467	0.1897	15.2980	0.0000
	Puerto Rican	0.3067	0.0138	0.2804	0.3344	22.2760	0.0000
	South American	0.1957	0.0203	0.1590	0.2386	9.6655	0.0000
	Mixed/Other	0.1544	0.0299	0.1043	0.2224	5.1680	0.0000

Conditional Marginal		Conditional Marginal	SE	Lower 95% Limit	Upper 95% Limit	T:Marg=0	P-value
7-level re-classification of Hispanic/Latino Background	Dominican	0.2815	0.0185	0.2467	0.3192	15.2328	0.0000
	Central American	0.2063	0.0133	0.1814	0.2337	15.4945	0.0000
	Cuban	0.3564	0.0151	0.3274	0.3865	23.6815	0.0000
	Mexican	0.1671	0.0109	0.1467	0.1897	15.2980	0.0000
	Puerto Rican	0.3067	0.0138	0.2804	0.3344	22.2760	0.0000
	South American	0.1957	0.0203	0.1590	0.2386	9.6655	0.0000
	Mixed/Other	0.1544	0.0299	0.1043	0.2224	5.1680	0.0000

It is misleading to compare these hypertension prevalences between Hispanic backgrounds because the age distributions are different and the hypertension prevalence is different by age group. Hence, to appropriately compare backgrounds, we should either stratify by age group or adjust to a common age for all backgrounds.

9.4.2. Age (category) adjusted estimates

We can adjust prevalence for covariates using linear models. In particular, when the outcome depends on age and we are interested in comparing Hispanic/Latino backgrounds within HCHS/SOL it is important to adjust for age because each background has a different age distribution. One way to do this is to stratify by age group (section 6.3). Another is to adjust statistically for age using either age groups or age as a continuous variable. When the association is not linear, using age groups is preferred.

```
proc regress data=SOL filetype=sas design=wr;
    nest strat PSU_ID;  weight weight_final_norm_overall;
    subpopn KEEP_DATA=1;
    class bkgrd1_c7 agegroup_c6_NHANES;
    model hypert = bkgrd1_c7 agegroup_c6_NHANES;
    condmarg / all;
run;
```

Output 9.4.2a Age-adjusted hypertension prevalence by background using age groups, PROC REGRESS

Conditional Marginal		Conditional Marginal	SE	T:Marg-=0	P-value
7-level re-classification of Hispanic/Latino Background	Dominican	29.434	1.511	19.483	0.000
	Central American	23.189	1.139	20.352	0.000
	Cuban	27.834	1.038	26.821	0.000
	Mexican	20.363	0.872	23.339	0.000
	Puerto Rican	27.978	1.136	24.618	0.000
	South American	17.294	1.523	11.355	0.000
	Mixed/Other	22.977	2.163	10.622	0.000

Note that these age-adjusted hypertension prevalence estimates are different from the unadjusted hypertension prevalence estimates (Output 9.4f, 9.4.1a and 9.4.1b) because of the differences in age distributions among Hispanic/Latino background. Recall that Cubans are on average much older and hence their unadjusted prevalence is 35.64 whereas the age-adjusted hypertension prevalence is reduced to 27.83.

The following SUDAAN code uses PROC RLOGISTIC to estimate the same age-adjusted hypertension prevalence by Hispanic/Latino background**. Due to the different methods used in logistic regression compared to linear regression, the prevalence estimates from logistic regression are not the same as the linear estimates. In addition, the conditional and predicted marginals are not the same. In this example, the predicted marginals are very similar to the linear regression estimates, but the conditional marginal are substantially lower**. This has to do with the different calculation methods of the marginals. Standard errors are different between the 3 methods, but there is no clear pattern of one consistently being higher or lower than another.

```
proc rlogist data=sol filetype=sas design=wr deft4;
   nest strat PSU_ID / NOSORTCK; weight weight_final_norm_overall;
   class bkgrd1_c7 agegroup_c6_nhanes;
   subpopn KEEP_DATA=1;
   reflevel bkgrd1_c7=3 agegroup_c6_nhanes=6;
   model hypertension2 =  bkgrd1_C7 agegroup_c6_nhanes;
   CONDMARG bkgrd1_C7;
   PREDMARG bkgrd1_C7;
run;
```

Output 9.4.2b Age-adjusted hypertension prevalence by background using age groups, PROC RLOGIST

Predicted Marginal		Predicted Marginal	SE	Lower 95% Limit	Upper 95% Limit	T:Marg=0	P-value
7-level re-classification of Hispanic/Latino Background	Dominican	0.2958	0.0159	0.2655	0.3281	18.5480	0.0000
	Central American	0.2318	0.0127	0.2078	0.2575	18.3063	0.0000
	Cuban	0.2718	0.0102	0.2522	0.2923	26.6138	0.0000
	Mexican	0.1999	0.0105	0.1802	0.2212	19.1209	0.0000
	Puerto Rican	0.2768	0.0115	0.2548	0.2999	24.1171	0.0000
	South American	0.1785	0.0147	0.1515	0.2092	12.1573	0.0000
	Mixed/Other	0.2275	0.0310	0.1725	0.2939	7.3485	0.0000

Conditional Marginal		Conditional Marginal	SE	Lower 95% Limit	Upper 95% Limit	T:Marg=0	P-value
7-level re-classification of Hispanic/Latino Background	Dominican	0.2163	0.0201	0.1794	0.2584	10.7515	0.0000
	Central American	0.1474	0.0132	0.1232	0.1753	11.1337	0.0000
	Cuban	0.1889	0.0119	0.1666	0.2135	15.8222	0.0000
	Mexican	0.1179	0.0094	0.1006	0.1377	12.4978	0.0000
	Puerto Rican	0.1945	0.0144	0.1677	0.2244	13.4765	0.0000
	South American	0.0999	0.0123	0.0782	0.1267	8.1341	0.0000
	Mixed/Other	0.1433	0.0286	0.0956	0.2091	5.0039	0.0000

9.4.3. Age (continuous) adjusted estimate, using default weighted sample mean

The following SUDAAN code uses PROC REGRESS and PROC RLOGISTIC to obtain age-adjusted hypertension prevalence by Hispanic/Latino background, with adjustment for continuous age. The SAS code uses PROC SURVEYLOGISTIC, which could obtain identical outputs for the conditional marginal. However, it does not have the function to produce predicted marginal for now.

As we saw above for age-group adjustment, the prevalence estimates are different between the 3 methods**. Logistic regression predicted marginal are similar to the linear regression, and logistic conditional marginal are substantially less**.

```
proc regress data=sol filetype=sas design=wr;
    nest strat PSU_ID;  weight weight_final_norm_overall;
    subpopn KEEP_DATA=1;
    class bkgrd1_c7 ;
    model hypert = bkgrd1_c7 age;
    condmarg / all;
run;
```

Output 9.4.3a Age-adjusted hypertension prevalence by background using continuous age, Adjusted to the internal sample age PROC REGRESS

Conditional Marginal		Conditional Marginal	SE	T:Marg-=0	P-value
7-level re-classification of Hispanic/Latino Background	Dominican	29.70	1.54	19.27	0.00
	Central American	22.50	1.14	19.68	0.00
	Cuban	27.79	1.11	25.03	0.00
	Mexican	20.36	0.91	22.49	0.00
	Puerto Rican	28.22	1.15	24.47	0.00
	South American	16.97	1.58	10.75	0.00
	Mixed/Other	24.31	2.23	10.90	0.00

```
proc rlogist data=sol filetype=sas design=wr deft4;
   nest strat PSU_ID / NOSORTCK; weight weight_final_norm_overall;
   class bkgrd1_c7;
   subpopn KEEP_DATA=1;
   reflevel bkgrd1_c7=3;
   model hypertension2 =  bkgrd1_C7  age;
   CONDMARG bkgrd1_C7;
   PREDMARG bkgrd1_C7;
run;
```

Output 9.4.3b Age-adjusted hypertension prevalence by background using continuous age, adjusted to the internal sample age PROC RLOGIST

Predicted Marginal		Predicted Marginal	SE	Lower 95% Limit	Upper 95% Limit	T:Marg=0	P-value
7-level re-classification of Hispanic/Latino Background	Dominican	0.2955	0.0161	0.2649	0.3280	18.3741	0.0000
	Central American	0.2310	0.0125	0.2073	0.2564	18.4637	0.0000
	Cuban	0.2689	0.0103	0.2491	0.2897	26.0179	0.0000
	Mexican	0.2023	0.0106	0.1822	0.2240	19.0316	0.0000
	Puerto Rican	0.2780	0.0115	0.2561	0.3011	24.2467	0.0000
	South American	0.1799	0.0149	0.1524	0.2111	12.0514	0.0000
	Mixed/Other	0.2307	0.0335	0.1715	0.3029	6.8863	0.0000

Conditional Marginal		Conditional Marginal	SE	Lower 95% Limit	Upper 95% Limit	T:Marg=0	P-value
7-level re-classification of Hispanic/Latino Background	Dominican	0.2259	0.0197	0.1895	0.2670	11.4407	0.0000
	Central American	0.1535	0.0129	0.1299	0.1805	11.9234	0.0000
	Cuban	0.1946	0.0113	0.1734	0.2178	17.2423	0.0000
	Mexican	0.1254	0.0099	0.1073	0.1462	12.6920	0.0000
	Puerto Rican	0.2052	0.0137	0.1795	0.2335	14.9378	0.0000
	South American	0.1053	0.0130	0.0824	0.1337	8.0921	0.0000
	Mixed/Other	0.1533	0.0325	0.0997	0.2283	4.7214	0.0000

In SAS, the LSMEANS command requires the variables to be declared in the CLASS statement and also the GLM parameterization.

```
proc surveylogistic  data = SOL order=internal;
   strata strat; cluster psu_id;weight weight_final_norm_overall;
   domain KEEP_DATA;
   class bkgrd1_c7 / param=glm;
   model hypert(descending) = bkgrd1_c7 age;
   lsmeans bkgrd1_c7 / at means ilink cl;
run;
```

Output 9.4.3c Age-adjusted hypertension prevalence by background using continuous age, adjusted to the internal sample age PROC SURVEYLOGISTIC

7-level re-classification of Hispanic/Latino Background	Mean	Standard Error of Mean	Lower Mean	Upper Mean
Central American	0.1535	0.01288	0.1299	0.1805
Cuban	0.1946	0.01129	0.1734	0.2177
Dominican	0.2259	0.01975	0.1895	0.2669
Mexican	0.1254	0.00989	0.1073	0.1461
Mixed/Other	0.1532	0.03246	0.0998	0.2281
Puerto Rican	0.2051	0.01374	0.1795	0.2334
South American	0.1053	0.01301	0.08237	0.1337

9.4.4. Age (continuous) adjusted estimate, using a specified age

The following SUDAAN code uses PROC REGRESS and PROC RLOGISTIC to obtain age-adjusted hypertension prevalence by Hispanic Background, with adjustment for continuous age, specifying a specific age rather than using the default sample age. The SAS code uses PROC SURVEYLOGISTIC to produce the same output for the conditional marginal.

Note that since a ***specific value for age*** is specified, ***only to the conditional marginal is applicable*** in this case, and not the predicted marginal. Recall the predicted marginal is based on the entire age distribution of the sample. SUDAAN will produce both the conditional and predicted marginal in the provided output and no error is generated, but the predicted marginal in the output are identical to the conditional marginal. Avoid referring to this output as a predicted marginal, since that is an incorrect interpretation.

```
proc regress data=SOL filetype=sas design=wr;
    nest strat PSU_ID;   weight weight_final_norm_overall;
    subpopn KEEP_DATA=1;
    class bkgrd1_c7;
    model hypert = bkgrd1_c7 age bkgrd1_c7*age;
    CONDMARG bkgrd1_c7*age / age=(60);
    COND_EFF bkgrd1_c7=(1 0 0 0 0 0 0)*age=(1) / age=(60)
          NAME="Dominican at age=60";
run;
```

Output 9.4.4a Age-adjusted hypertension prevalence by background using continuous age, Adjusted to AGE=60, PROC REGRESS

Conditional Marginal #1		Conditional Marginal	SE	Lower 95% Limit	Upper 95% Limit	T:Marg=0	P-value
7-level re-classification of Hispanic/Latino Background, Age	Dominican, 60	59.63	2.71	54.30	64.96	21.97	0.0000
	Central American, 60	49.41	2.25	44.99	53.82	21.97	0.0000
	Cuban, 60	58.93	1.78	55.43	62.43	33.08	0.0000
	Mexican, 60	43.40	2.22	39.04	47.76	19.56	0.0000
	Puerto Rican, 60	58.18	1.86	54.53	61.84	31.23	0.0000
	South American, 60	41.44	3.27	35.03	47.86	12.69	0.0000
	Mixed/Other, 60	50.36	6.80	37.00	63.72	7.40	0.0000

```
proc rlogist data=sol filetype=sas design=wr;
   nest strat PSU_ID / NOSORTCK;
   weight &weight;
   class bkgrd1_c7;
   subpopn KEEP_DATA=1;
   reflevel bkgrd1_c7=3;
   model hypertension2 =  bkgrd1_C7 age bkgrd1_c7*AGE;
   PREDMARG bkgrd1_c7*AGE / AGE=(60);
   CONDMARG bkgrd1_c7*AGE / AGE=(60);
   COND_EFF bkgrd1_c7=(1 0 0 0 0 0 0)*age=(1) / age=(60)
      NAME="Dominican at age=60";
```

```
run;
```

Output 9.4.4b Age-adjusted hypertension prevalence by background using continuous age, Adjusted to AGE=60, PROC RLOGIST

Conditional Marginal #1		Condition-al Marginal	SE	Lower 95% Limit	Upper 95% Limit	T:Marg=0	P-value
7-level re-classification of Hispanic/Latino Background, Age	Dominican, 60	0.6500	0.0340	0.5806	0.7136	19.0969	0.0000
	Central American, 60	0.5498	0.0267	0.4969	0.6015	20.5569	0.0000
	Cuban, 60	0.6182	0.0231	0.5720	0.6624	26.7714	0.0000
	Mexican, 60	0.4922	0.0272	0.4391	0.5456	18.0841	0.0000
	Puerto Rican, 60	0.6230	0.0226	0.5776	0.6663	27.5174	0.0000
	South American, 60	0.4440	0.0400	0.3674	0.5234	11.0865	0.0000
	Mixed/Other, 60	0.5923	0.0995	0.3929	0.7654	5.9558	0.0000

```
proc surveylogistic  data = SOL order=internal;
   strata strat; cluster psu_id;weight weight_final_norm_overall;
   domain KEEP_DATA;
   class bkgrd1_c7 / param=glm;
   model hypert(descending) = bkgrd1_c7 age bkgrd1_c7*age;
   lsmeans bkgrd1_c7 / at age=60 ilink cl;
run;
```

Output 9.4.4c Age-adjusted hypertension prevalence by background using continuous age, Adjusted to AGE=60, PROC SURVEYLOGISTIC

7-level re-classification of Hispanic/Latino Background	AGE	Mean	Standard Error of Mean	Lower Mean	Upper Mean
Central American	60.00	0.5498	0.02675	0.4970	0.6015
Cuban	60.00	0.6182	0.02310	0.5720	0.6624
Dominican	60.00	0.6500	0.03405	0.5807	0.7135
Mexican	60.00	0.4922	0.02723	0.4392	0.5455
Mixed/Other	60.00	0.5923	0.09947	0.3932	0.7651
Puerto Rican	60.00	0.6230	0.02265	0.5777	0.6662
South American	60.00	0.4440	0.04006	0.3675	0.5232

9.4.5 Age, gender and site adjusted hypertension prevalence by background

The distribution of background is very different by site. For example in Output 9.4.5 one can see that a majority of the San Diego participants are of Mexican background, whereas Miami has very few participants of Mexican background. Adjusting for site is good practice to account in the point estimates for the differences between sites. We recommended adjusting for site specifically when comparisons of Hispanic backgrounds are of interest.

Output 9.4.5. Number of participants by Hispanic background and site

CENTER (Center)	*BKGRD1_C7(7-level re-classification of Hispanic/Latino Background)*							
	Dominican	*Central American*	*Cuban*	*Mexican*	*Puerto Rican*	*South American*	*Mixed/ Other*	*Total*
Bronx	1380	219	45	208	1837	187	200	4076
Chicago	27	418	25	2409	770	374	100	4123
Miami	64	1034	2269	38	82	468	112	4067
San Diego	2	61	9	3817	39	43	91	4062
Total	1473	1732	2348	6472	2728	1072	503	16328

Frequency Missing = 87

13 Grey shaded boxes indicate adequate sample size (n>=100) for comparisons between backgrounds within sites or vice versa. All backgrounds with < 100 within each site are pooled with the Mixed/other category for that site.

```
proc regress data=sol filetype=sas design=wr;
    nest strat PSU_ID;
    weight weight_final_norm_overall;
    subpopn KEEP_DATA=1;
    class bkgrd1_c7 site;
    model hypert = bkgrd1_c7 age female SITE;
    condmarg bkgrd1_c7;
run;
```

Output 9.4.5a Age, gender and site adjusted hypertension prevalence by Hispanic background, PROC REGRESS

Conditional Marginal #1		**Conditional Marginal**	**SE**	**T:Marg-=0**	**P-value**
7-level re-classification of Hispanic/Latino Background	Dominican	30.450	1.820	16.734	0.000
	Central American	22.980	1.200	19.144	0.000
	Cuban	28.026	1.462	19.172	0.000
	Mexican	19.525	0.912	21.404	0.000
	Puerto Rican	29.033	1.323	21.938	0.000
	South American	17.624	1.594	11.056	0.000
	Mixed/Other	24.361	2.256	10.798	0.000

Note that the parameter estimates from Output 9.4.5a are not that different from those in Output 9.4.3a, which did not adjust for site. The p-value for the 3 df global test for site is 0.0386. See chapter 11 for general recommendations on how to adjust for field center and Hispanic/Latino background.

9.4.6 Age-adjusted hypertension prevalence by site and background

If we are interested in comparing prevalence by Hispanic background between sites, then, typically in statistics, we include an interaction term in the model for site and Hispanic background. However, there are some combinations of site and Hispanic background that do not have adequate sample sizes (Output 9.4.5). Hence, the CC created a 17–level nominal variable for Hispanic background and site combination (called site_bkgrd) in which each background within a site with n < 100 is pooled with the Mixed/other category for that site.

```
proc regress data=SOL filetype=sas design=wr;
    nest strat PSU_ID;  weight weight_final_norm_overall;
    subpopn KEEP_DATA=1;
    class site_bkgrd;
    model hypert = site_bkgrd age;
    condmarg / all;
run;
```

Output 9.4.6a Age-Adjusted hypertension prevalence by site and background, PROC REGRESS

Conditional Marginal #1			Conditional Marginal	SE	T:Marg-=0	P-value
Center/Hispanic Background Cross-Classification (collapsed categories)	Dominicans	Bronx	29.845	1.579	18.900	0.000
	Central Americans	Bronx	24.385	3.069	7.945	0.000
		Chicago	17.430	1.827	9.542	0.000
		Miami	23.295	1.463	15.920	0.000
	Cubans	Miami	27.675	1.128	24.525	0.000
	Mexicans	Bronx	13.133	1.429	9.189	0.000
		Chicago	17.299	0.979	17.664	0.000
		San Diego	22.294	1.215	18.350	0.000
	Puerto Ricans	Bronx	28.523	1.409	20.240	0.000
		Chicago	29.291	2.127	13.769	0.000
	South Americans	Bronx	17.916	3.228	5.551	0.000
		Chicago	12.779	2.109	6.060	0.000
		Miami	18.891	2.446	7.723	0.000
	Other	Bronx	28.613	3.878	7.377	0.000
		Chicago	28.224	4.146	6.807	0.000
		Miami	23.864	2.381	10.023	0.000
		San Diego	17.055	2.793	6.107	0.000

To make field center contrasts within each Hispanic background, it is easiest to use PROC SURVEYREG with CONTRAST statements.

```
proc surveyreg data = sol order=internal;
  strata strat; cluster psu_id; weight weight_final_norm_overall;
  domain KEEP_DATA;
  class site_bkgrd ;
  model hypert = site_bkgrd age female/ solution noint;
  /* WARNING: MAKE SURE ORDER=INTERNAL */
  contrast "C/A - B vs. C"   site_bkgrd 0 1 -1  0 0 0 0 0 0 0 0 0 0 0 0 0 0;
  contrast "C/A - B vs. M"   site_bkgrd 0 1  0 -1 0 0 0 0 0 0 0 0 0 0 0 0 0;
  contrast "C/A - C vs. M"   site_bkgrd 0 0  1 -1 0 0 0 0 0 0 0 0 0 0 0 0 0;
  contrast "C/A - Overall"   site_bkgrd 0 1  0 -1 0 0 0 0 0 0 0 0 0 0 0 0 0,
                             site_bkgrd 0 0  1 -1 0 0 0 0 0 0 0 0 0 0 0 0 0;

  contrast "Mex - B vs. C"   site_bkgrd 0 0 0 0 0 1 -1  0 0 0 0 0 0 0 0 0 0;
  contrast "Mex - B vs. S"   site_bkgrd 0 0 0 0 0 1  0 -1 0 0 0 0 0 0 0 0 0;
  contrast "Mex - C vs. S"   site_bkgrd 0 0 0 0 0 0  1 -1 0 0 0 0 0 0 0 0 0;
  contrast "Mex - Overall"   site_bkgrd 0 0 0 0 0 1  0 -1 0 0 0 0 0 0 0 0 0,
                             site_bkgrd 0 0 0 0 0 0  1 -1 0 0 0 0 0 0 0 0 0;

  contrast "PR  - B vs. C"   site_bkgrd 0 0 0 0 0 0  0 0 1 -1 0 0 0 0 0 0 0;

  contrast "S - B vs. C"     site_bkgrd 0 0 0 0 0 0 0 0 0 0 1 -1  0 0 0 0 0;
  contrast "S - B vs. M"     site_bkgrd 0 0 0 0 0 0 0 0 0 0 1  0 -1 0 0 0 0;
  contrast "S - C vs. M"     site_bkgrd 0 0 0 0 0 0 0 0 0 0 0  1 -1 0 0 0 0;
  contrast "S - Overall"     site_bkgrd 0 0 0 0 0 0 0 0 0 0 1  0 -1 0 0 0 0,
                             site_bkgrd 0 0 0 0 0 0 0 0 0 0 0  1 -1 0 0 0 0;
run;
```

Output 9.4.6b Tests for pairwise hypertension comparisons between sites within background, PROC SURVEYREG

Hispanic Origin	*Site Comparisons*	*D.O.F*	*Wald*	*P-value*
Hispanic Background Specific Contrasts				
Central Americans	Bronx. vs. Chicago	1	3.81	0.0514
	Bronx vs. Miami	1	0.10	0.7535
	Chicago vs. Miami	1	6.50	0.0110
	Overall	2	3.76	0.0238
Mexicans	Bronx. vs. Chicago	1	5.96	0.0149
	Bronx vs. San Diego	1	25.76	<.0001
	Chicago vs. San Diego	1	11.10	0.0009
	Overall	2	13.32	<.0001
Puerto Ricans	Bronx vs. Chicago	1	0.09	0.7702
South Americans	Bronx. vs. Chicago	1	1.75	0.1858
	Bronx vs. Miami	1	0.07	0.7898
	Chicago vs. Miami	1	3.69	0.0551
	Overall	2	2.07	0.1267

9.4.7. Age standardized estimates to the US 2000 census

In the previous example, we calculated adjusted prevalence estimates to a one specific age value. To account for differences in ***age distribution*** between two populations and to present ***estimates for a standard age distribution***, we can use external standardization with categorical age, in 10-year age groups. Here we illustrate how to externally standardize the HCHS/SOL population to the US 2000 population age distribution (Klein RJ, Schoenborn CA, 2001).

In SUDAAN we can estimate standardized prevalence using PROC DESCRIPT by including age group (e.g. agegroup_c6_NHANES) in the CLASS and STDVAR statements and specifying the external age distribution (e.g., to the US 2000 population in Output 1.5) in the STDWGT statement. The output provides the standardized prevalence of the outcome specified in the VAR statement for each level of variables included in the TABLES statement.

```
proc descript data=sol design=wr;
  nest strat PSU_ID; weight weight_final_norm_overall;
  subpopn KEEP_DATA=1; class hypertension2 bkgrd1_c7 agegroup_c6_nhanes;
  tables bkgrd1_c7; var hypertension2;
  catlevel 1; /* Output only % of hypertension (i.e. hypertension2=1) */
  stdvar agegroup_c6_nhanes;
  stdwgt 0.235800444 0.222616766 0.225788538 0.162064749 0.10713542 0.046594083;
run;
```

Output 9.4.7 Age -standardized hypertension prevalence by Hispanic/Latino Background, PROC DESCRIPT

Variable	7-level re-classification of Hispanic/Latino Background	Sample Size	Percent	SE Percent	Lower 95% Limit Percent	Upper 95% Limit Percent
Hypertension using NHANES definition: 1	Total	11737	25.69	0.49	24.74	26.67
	Dominican	1000	31.58	1.54	28.64	34.68
	Central American	1369	24.85	1.23	22.52	27.34
	Cuban	1667	28.77	0.95	26.94	30.68
	Mexican	4617	21.36	1.04	19.38	23.48
	Puerto Rican	1957	29.36	1.13	27.19	31.62
	South American	757	19.13	1.52	16.33	22.29
	Mixed/Other	370	25.09	3.09	19.53	31.62

SAS SURVEYREG does not have a statement that easily allows for external standardization. Hence we need to explicitly fit the linear model and specify the coefficients of the external population we want to standardize to. For example, if we want to estimate the hypertension prevalence by Hispanic background standardizing to the US 2000 population, then in the MODEL statement we include main effects for Hispanic background and age group and their interaction. Then, in the ESTIMATE statement we specify the contrast matrix (see section 7.2 for details on the ESTIMATE statement). In particular, below we illustrate how to write the contrast matrix to estimate the age-standardized hypertension prevalence for Dominicans. This example is also shown in Output 9.4.7a using SUDAAN.

Note that this follows the same general programming format as we used in section 9.4.4 to obtain conditional marginal at one specified age. Also note that, since we are specifying specific values for age, the predicted marginals from logistic regression are not applicable.

```
proc regress data=sol filetype=sas design=wr;
    nest strat PSU_ID ;     weight weight_final_norm_overall;
    subpopn KEEP_DATA=1;    class bkgrd1_c7 agegroup_c6_nhanes;
    model hypert = bkgrd1_c7 agegroup_c6_nhanes bkgrd1_c7*agegroup_c6_nhanes;
    cond_eff bkgrd1_c7=(1 0 0 0 0 0 0)*agegroup_c6_nhanes=(0.235800444
    0.222616766 0.225788538 0.162064749 0.10713542 0.046594083)/
    NAME=" Dominican";
run;
```

Output 9.4.7a Age-standardized hypertension prevalence for Dominicans – PROC REGRESS

Contrasted Conditional Marginal	CONDMARG Contrast	SE	T-Stat	P-value
Dominican	31.58	1.54	20.53	0.0000

```
proc rlogist data=sol filetype=sas design=wr deft4;
   nest strat PSU_ID / NOSORTCK;  weight weight_final_norm_overall;
   class bkgrd1_c7 agegroup_c6_nhanes;   subpopn KEEP_DATA=1;
   model hypertension2 =  bkgrd1_c7 agegroup_c6_nhanes
   bkgrd1_c7*agegroup_c6_nhanes;
      cond_eff bkgrd1_c7=(1 0 0 0 0 0 0)*agegroup_c6_nhanes=(0.235800444
   0.222616766 0.225788538 0.162064749 0.10713542 0.046594083)/
   NAME="Dominican";
   run;
```

Output 9.4.7b Age-standardized hypertension prevalence for Dominicans – PROC RLOGIST

Contrasted Conditional Marginal	CONDMARG Contrast	SE	T-Stat	P-value
Dominican	0.3158	0.0154	20.5261	0.0000

```
proc surveyreg data = SOL order=internal;
   strata strat;   cluster psu_id;   weight &weight;
   domain KEEP_DATA;   class bkgrd1_c7 agegroup_c6_nhanes;
   model hypert = bkgrd1_c7 agegroup_c6_nhanes bkgrd1_c7*agegroup_c6_nhanes /
                  solution noint;
   /* 2000 US Census age-distribution */
   estimate 'Dominican' bkgrd1_c7 1 0 0 0 0 0 0
                        agegroup_c6_nhanes 0.235800444 0.222616766 0.225788538
                        0.162064749 0.10713542 0.046594083
                        bkgrd1_c7*agegroup_c6_nhanes 0.235800444 0.222616766
                        0.225788538 0.162064749 0.10713542 0.046594083
                                                0 0 0 0 0 0
                                                0 0 0 0 0 0
                                                0 0 0 0 0 0
                                                0 0 0 0 0 0
                                                0 0 0 0 0 0
                                                0 0 0 0 0 0 / e;
 run;
```

Output 9.4.7c Age-standardized hypertension prevalence for Dominicans – PROC SURVEYREG

		Estimate Standard			
Label	*Estimate*	*Error*	*DF*	*t Value*	*Pr > |t|*
Dominican	31.5799	1.5405	644	20.50	<.0001

Note that the hypertension prevalence for Dominicans standardized to the US 2000 population using PROC DESCRIPT (Output 9.4.7) is identical to using a linear regression model and specifying the contrast matrix (Output 9.4.7a) or the RLOGIST model with conditional marginal (Output 9.4.7b) or the SURVEYREG model (Output 9.4.7c).

Summary

This section has shown examples of using survey linear and logistic regression to estimate subgroup prevalence.

- These methods vary somewhat in their interpretation, and none of them are wrong.
- **The choice of method should be specified a prior in the analysis plan and in the manuscript methods section.**
- Because of the differences in the way that marginals (conditional or predicted) are calculated, and the interpretation of these marginals**, when using logistic regression, we recommend the use of predicted marginal (for internal adjustment**) unless the author is interested in adjusting to a pre-specified value or distribution (external standardization).

10. WEIGHTED CORRELATIONS

In this section, we illustrate how to obtain Pearson correlation coefficients and how to test whether this correlation is significant using methods which appropriately account for the study design. The estimated correlation coefficients should be calculated using one procedure (PROC CORR) and the p-values should be calculated using another procedure (PROC REGRESS or PROC SURVEYREG).

10.1. SAS code

The next group of SAS statements invokes PROC CORR to obtain Pearson correlation coefficients that appropriately account for the study design. You must specify the VARDEF=WEIGHT option in the PROC statement and specify the WEIGHT variable.

NOTE: Do not use any standard errors, confidence intervals, or statistical tests based on this procedure – they will not be correct. This procedure should only be used to get estimates of correlation coefficients. You can use the noprob option in the PROC CORR statement so that SAS does not display the p-values.

As there is no DOMAIN statement in this procedure, since we are not using this procedure to calculate standard errors, confidence intervals or statistical tests, the WHERE statement can be used to calculate point estimates.

```
proc corr data = sol nosimple noprob vardef=weight;
   weight weight_final_norm_overall;
   where KEEP_DATA=1;
   var sbpa5;
   with age;
run;
```

Output 10.1. PROC CORR weighted correlation of age and sbpa5

Pearson Correlation Coefficients
Number of Observations

	SBPA5
AGE	0.48259
Age	11806

While PROC CORR should be used to estimate Pearson correlation coefficients, we use regression models to generate p-values for these correlations. Using either PROC SURVEYREG or PROC REGRESS, regress the first variable on the second and then regress the second variable on the first. Take the largest p-value from these two models to produce a conservative p-value for the test. As seen in the output below, both p-values are less than 0.0001, and therefore this is the p-value for the test for correlation between the age and SBP5 variable.

```
proc surveyreg data=sol;
    strata strat;
    cluster psu_id;
    domain KEEP_DATA;
    weight weight_final_norm_overall;
    model  age = sbpa5;
run;

proc surveyreg data=sol;
    strata strat;
    cluster psu_id;
    domain KEEP_DATA;
    weight weight_final_norm_overall;
    model  sbpa5 = age;
run;
```

Output 10.2. PROC SURVEYREG regressing sbpa5 on age and age on sbpa5

Tests of Model Effects

Effect	*Num DF*	*F Value*	*Pr > F*
Model	1	1808.28	<.0001
Intercept	1	69.50	<.0001
SBPA5	1	1808.28	<.0001

Tests of Model Effects

Effect	*Num DF*	*F Value*	*Pr > F*
Model	1	1486.30	<.0001
Intercept	1	27411.3	<.0001
AGE	1	1486.30	<.0001

11. ACCOUNTING FOR CENTER EFFECT IN HCHS/SOL ANALYSES

In this section, we provide recommendations on how to adjust for Hispanic/Latino background and field center given that these two variables are highly collinear in HCHS/SOL (table 9.4.5). **Hence, including both as main effects in the model could be problematic in some situations.**

Note that:

- Background and field center collinearity is due to (1) some Hispanic/Latino background groups are more concentrated in some places (due to immigration differences) and (2) study design (RFP called to field centers to provide predominantly specific backgrounds);
- Field center is one of the stratification factors of the study- and sample- design, and as such it has to be taken into account when conducting statistical analyses (see sections 1.1 and 1.3). In HCHS/SOL, the variable 'Strat', which has 21 strata, has the field center embedded. Its inclusion under the strata statement in software that accommodates complex survey makes the variances to be correctly estimated.

Appropriate statistical methods and their correct specification can help answer some research questions but not all. In general, there are two overall goals: (1) Report population estimates and (2) study the association between exposure and health outcomes. Hence, these two distinct analytic objectives motivated the approach to sample selection in HCHS/SOL (LaVange et al, Ann Epi 2010):

1. The study sample must support estimates of prevalence of baseline risk factors, both overall and by Hispanic/Latino background and other demographic subgroups.
2. The study sample must support the evaluation of the relationships between various risk factors and disease outcomes.

11.1. Report population estimates

Inference of population estimates is to the target population (4 field centers pooled), and prevalences reflect the health status of the target population *per se*. Prevalence differences among backgrounds are due to different reasons, and understanding the differences is a separate aim. Age and gender adjustment (internal to HCHS/SOL or to an external population) is done to compare Hispanic/Latino background, had they had the same age distribution, because there are substantial differences in the age distribution among Hispanic backgrounds. For example, Cubans are on average older and Mexicans are on average younger.

Recommendations:

- Estimate prevalence WITHOUT adjusting for site and any other covariates except possibly for age and gender. Prevalence is a population quantity and should be reported as it is. Examining whether prevalences are different across field centers or among Hispanic/Latino background is a different goal. See section 11.2 for ways to account for possible field center effect when studying the associations.
- Do not perform statistical testing (i.e., do not report p-values) across Hispanic/Latino background groups. Instead, report confidence intervals or standard errors (SE).

11.2. Study the association between exposure and health outcomes

How to account for a field center effect when assessing the association between exposure and health outcomes depends on whether Hispanic/Latino background is the exposure of interest or not.

11.2.1. An exposure other than Hispanic/Latino background

Ideally, we would like to adjust for both Hispanic/Latino background and field center as main effects. By including field center, the exposure effect estimates are adjusted for potential differences across field centers (e.g. weather, measurement error due to differences in technicians, etc) which are not explained by the covariates in the model. One way to include both and avoid the collinearity is by using the reduced interaction term (site_bkrd with 17 levels; see output 9.4.5 and 9.4.6 for its description).

Recommendation:

Adjust for (1) both Hispanic/Latino background and fielding center as main effects, and separately (2) for the reduced interaction term (site_bkrd with 17 levels). If including site_bkgrd leads to convergence issues, use background only. In such cases, the methods section of the manuscript should explain that the reduced interaction between site and background (because of their collinearity) could not be adjusted for due to small cell sizes relative to the model.

For example, in manuscript #13 'Sleep Disordered Breathing' there was interest in assessing the effect of sleep apnea on diabetes after controlling for age, gender, BMI, waist circumference, cigarette use and alcohol use. There was no difference in the adjusted odds ratio between sleep apnea and diabetes2 either using both main effects or using site_bkgrd (output 11.2.1).

```
proc sort data=work.ms13; by strat PSU_ID; run;
proc multilog data=MS13 filetype=sas design=wr;
 nest strat PSU_ID; weight weight_final_norm_overall;
 class ahi3p_ge15 BKGRD1_C7 male agegrp2_c5 BMIGRP_C3 cigarette_use
       alcohol_use centernum diabetes2 / dir=descending;
 model diabetes2 = ahi3p_ge15 agegrp2_c5 male cigarette_use alcohol_use
       BKGRD1_C7 centernum BMIGRP_C3 waistcircumf / genlogit;
 subpopn keep_ms13= 1 and cohort in (1,2);
run;
```

Output 11.2.1. Adjusted OR (95% CI) between sleep apnea (AHI3p>=15) and diabetes2

	Normal glucose regulation	Impaired glucose tolerance	Diabetes
Covariates + center + Hispanic/Latino background	1	1.39 (1.05, 1.83)	1.82 (1.36, 2.45)
Covariates + SITE_BKGRD	1	1.41 (1.07, 1.85)	1.84 (1.37, 2.46)

Covariates are age, gender,, cigarette use, alcohol use, BMI group and waist circumference.

11.2.2. Main effect of interest is Hispanic/Latino background

When the effect of Hispanic/Latino background is of interest, the analysis is more involved. Depending on the results of the initial analysis, further analyses could be needed. In general, we recommend the following steps:

Step 1. Specify the statistical model of interest

Fit a model with Hispanic/Latino background and important covariates without field center;

Step 2. Assess the effect of field center in confounding Hispanic/Latino background effect.

Add field center to the model fitted in step 1. If either of the following criteria is satisfied, we recommend exploring the effect further.

a) The effect of field center (3 df Wald test) is significant at the 0.15 level. We recommend inflating the significance level to be more conservative since some of the effect for center can be absorbed by the estimates for background resulting lower power. In addition, we recommend exploring the effect for borderline cases.

b) The estimates for background change in a meaningful way when field center is added to the model. One criterion for meaningful change is a 10% change on the parameter estimate and estimated outcome or effect measure of interest (e.g. estimated model-based prevalence, estimated means for continuous outcomes, or estimated log-odds for OR).

If neither a) nor b) is satisfied, we recommend using the Hispanic/Latino background main effect estimate from the model fitted in step 1 to compare among Hispanic/Latino background groups after adjusting for important covariates.

If either a) or b) is satisfied, further analyses should be conducted to help understanding the difference across Hispanic/Latino background and field center (step 3).

Step 3. Understand the effect of field center (only if either 2a or 2b is satisfied)

The following are some suggestions for further analyses.

- As an exploratory tool, provide a summary table of the main measure of interest by center and Hispanic/Latino background (BKGRD1_C7). For example, provide the age-adjusted prevalence estimates by center and BKGRD1_C7 in a similar format as Table 9.4.5. This can provide preliminary information on the potential interaction between Hispanic/Latino background and field center.
- When the analysis is for the full HCHS/SOL cohort, fit a model with 17-level nominal variable SITE_BKGRD as a covariate and depending on the research question focus on comparing backgrounds within a field center or vice versa (see sections 9.4.6, 11.3 and 11.4 for examples, and consider using Bonferroni correction to p-values to account for multiple comparisons). This 17-level nominal variable is a way to include the interaction between Hispanic/Latino background and field center without having sparse cells (see tables 11.5.1. and 11.5.2 as an example). Also see manuscript #20 Hypertension Awareness, treatment and control.
- Conduct in-depth analyses to understand differences in health outcomes between field centers within each Hispanic/Latino background (e.g. among Mexicans), whenever possible.
- Conduct in-depth analyses to examine the differences between Hispanic/Latino backgrounds within each field center (e.g. Bronx), whenever possible.
- Ideally, try to come up with a clinical hypothesis about an objective exposure which could explain all or part of the center effect (such as weather, air pollution, distance from typical port of entry to the US for each background, and so on), which can be incorporated in the Discussion section of a manuscript.
- Discussion should be provided to address the potential difference across Hispanic/Latino background and field center.

11.3 Example: center does not confound the effect of Hispanic/Latino background

One of the aims in HCHS/SOL is to study whether there are prevalence differences between Hispanic/Latino background groups after controlling for important covariates. For example, in manuscript #13 ‘Sleep Disordered Breathing’ authors were interested in whether sleep apnea (AHI≥15) prevalences are different among Hispanic/Latino background groups after adjusting for age and BMI and stratifying by gender. The p-value for field center is 0.4498, but there are two model-based prevalences that change more than 10% (Cubans and Mexicans females; see output 11.3). Further analysis by center and Hispanic/Latino background is suggested.

Model 1: Model of interest

Logit { Sleep Apnea } = bkgrd1_c7 age bmi male male*(bkgrd1_c7 age bmi)

Model 2: Add field center to model 1.

Logit { Sleep Apnea } = same covariates as in model 1 + centernum

```
proc sort data=ms13; by strat PSU_ID; run;
proc regress data=ms13 filetype=sas design=wr;
 nest strat PSU_ID; weight weight_final_norm_overall;
 class bkgrd1_c7 male centernum;
 model
   ahi3p_ge15_100 = bkgrd1_c7 age bmi male male*(bkgrd1_c7 age bmi) centernum;
 subpopn keep_ms13 = 1 and wave in (1,2);
 condmarg male*bkgrd1_c7;
run;
```

Output 11.3 Age and BMI adjusted sleep apnea model-based prevalence by Hispanic/Latino background

	Female				Male		
Background	WITHOUT center	WITH center	% change		WITHOUT center	WITH center	% change
Dominican	5.44	5.97	9		15.45	15.97	3
CA	4.57	5.04	9		15.13	15.49	2
Cuban	5.12	5.85	12		17.48	18.20	4
Mexican	5.59	4.94	-13		13.69	13.06	-5
PR	6.41	6.66	4		12.37	12.58	2
SA	5.54	5.86	5		13.03	13.25	2
Other/Mixed	4.61	4.81	4		15.11	15.17	0
p-value for centernum		0.4498					

NOTE: These sleep apnea model-based prevalences are adjusted to the overall age (40.9 yrs) and overall BMI mean (29.2 kg/m^2), and to the weighted proportion of center when the latter included.

11.4 Example: effect of sleep apnea on hypertension differs by background

Model (1)
Logit { Hypertension } = sleep apnea + bkgrd1_c7 + sleep apnea* bkgrd1_c7 + age + male + BMI + education + marital status + cigarette use + alcohol use

Model (2): Logit { Hypertension } = same covariates in model 1 + centernum

Model (3): Logit { Hypertension } = same covariates in model 1 + site_bkgrd

```
proc rlogist data=MS13 filetype=sas design=wr;
 nest strat PSU_ID; weight weight_final_norm_overall;
 class  ahi3p_ge15 site_bkgrd male agegrp2_c5 cigarette_use alcohol_use BMIGRP_C3;
 model hypertension = ahi3p_ge15 site_bkgrd agegrp2_c5 male cigarette_use
                      alcohol_use ahi3p_ge15* site_bkgrd BMIGRP_C3 WAISTCIRCUMF;
 effects ahi3p_ge15=(0 1) / site_bkgrd=1 exp name="AOR Dom - B";
 effects ahi3p_ge15=(0 1) / site_bkgrd=2 exp name="AOR CA - B";
 effects ahi3p_ge15=(0 1) / site_bkgrd=3 exp name="AOR CA - C";
 effects ahi3p_ge15=(0 1) / site_bkgrd=4 exp name="AOR CA - M";
 effects ahi3p_ge15=(0 1) / site_bkgrd=5 exp name="AOR Cub - M";
 effects ahi3p_ge15=(0 1) / site_bkgrd=6 exp name="AOR Mex - B";
 effects ahi3p_ge15=(0 1) / site_bkgrd=7 exp name="AOR Mex - C";
 effects ahi3p_ge15=(0 1) / site_bkgrd=8 exp name="AOR Mex - SD";
 effects ahi3p_ge15=(0 1) / site_bkgrd=9 exp name="AOR PR - B";
 effects ahi3p_ge15=(0 1) / site_bkgrd=10 exp name="AOR PR - C";
 effects ahi3p_ge15=(0 1) / site_bkgrd=11 exp name="AOR SA - B";
 effects ahi3p_ge15=(0 1) / site_bkgrd=12 exp name="AOR SA - C";
 effects ahi3p_ge15=(0 1) / site_bkgrd=13 exp name="AOR SA - M";
 effects ahi3p_ge15=(0 1) / site_bkgrd=14 exp name="AOR Other - B";
 effects ahi3p_ge15=(0 1) / site_bkgrd=15 exp name="AOR Other - C";
 effects ahi3p_ge15=(0 1) / site_bkgrd=16 exp name="AOR Other - M";
 effects ahi3p_ge15=(0 1) / site_bkgrd=17 exp name="AOR Other - SD";
 reflevel ahi3p_ge15=0 site_bkgrd =1 agegrp2_c5=1 male=0 cigarette_use=1 alcohol_use=1
       BMIGRP_C3=2;
 subpopn keep_ms13 = 1 and wave in (1,2);
run;
```

Output 11.4. Adjusted OR of sleep apnea on hypertension by Hispanic/Latino background and field center.

	Adjusted OR (95% CI)
Dominicans – Bronx	0.47(0.21, 1.04)
Central American - Bronx	1.03(0.11, 9.41)
Central American -	0.99(0.38, 2.56)
Central American - Miami	2.11(0.84, 5.31)
Cubans – Miami	1.26(0.82, 1.95)
Mexicans – Bronx	6.26(1.07, 36.49)
Mexicans – Chicago	1.62(0.82, 3.18)
Mexicans - San Diego	1.05(0.66, 1.69)
Puerto Ricans – Bronx	2.36(1.28, 4.37)
Puerto Ricans - Chicago	3.23(1.32, 7.90)
South American - Bronx	0.34(0.08, 1.48)
South American - Chicago	0.97(0.32, 2.97)
South American - Miami	1.63(0.62, 4.28)
Other – Bronx	0.62(0.12, 3.22)
Other – Chicago	0.25(0.06, 1.17)
Other – Miami	1.06(0.18, 6.37)
Other - San Diego	0.91(0.23, 3.61)

12. Missing Data

In this chapter, we provide

- guidelines on how to address missing outcomes and covariates for HCHS/SOL baseline data.
- implications for missing data in the complex survey design.
- examples of applying missing data methods using SAS and Mplus in HCHS/SOL.

12.1 Types of missing data

Missing data are an unavoidable and common problem in epidemiological and clinical research. It may lead to biased estimates and reduced precision. In general, there are three types of missing mechanisms (Rubin 1987): missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). For MCAR, the missingness does not depend on any other variables. It is the result of random events and produces a subsample that is representative of the total sample. For MAR, the missingness only depends on observed information. Conditional on that observed information, the resulting missingness is random. For MNAR, the missingness depends on unobserved information, which could include uncollected variables or collected variables that are missing; as a result, the source of the missingness cannot be controlled. For example, if income has missing data but the missingness does not depend on any variables (collected or uncollected), then it is MCAR. If the missingness of the income variable depends only on observed covariates (e.g. gender), then it is MAR. If the missingness of the income variable depends on whether they have a high or low income, which is unobserved for the subjects with missing income variables, then it is MNAR. In practice, MNAR is difficult to verify and the way to address it is very different from the other missing types. **In this chapter, we will not discuss MNAR further, and all the recommendations and examples are based on the assumption of MCAR or MAR.**

12.2 Evaluate missing data

The extent of missing data is a problem that can be quantified by the percent of missing data for the outcome and each covariate and the overall missing rate, where the **overall missing rate** is the percentage of subjects with at least one missing variable.
The **missing data pattern** is the combination of missing and non-missing variables observed in the analytic data. The missing pattern is important because it shows both the amount and structure of the missingness. The missing pattern is called **monotone missing pattern,** if the variables can be appropriately ordered such that the event that a variable is missing for a particular individual implies that all subsequent variables are missing for that individual. Otherwise, it is called the **non-monotone missing pattern**. For example, in HCHS/SOL there are two spirometry measures: before and after bronchodilation. Whenever the first measure is missing, the second measure is also missing (see **output 12.2.1** below where “X” indicates that the variable is observed, and “.” indicates the variable is missing). In the output, "O" means all the variables are missing

and would not be imputed. Physical activity data from accelerometer also has a monotone missing pattern (**output 12.2.2**). The first level of missing is based on whether the participants returned the accelerometer or not (ACTICALYN). Only the subjects who returned the accelerometer can be further categorized as adherent or not based on the HCHS/SOL protocol (ADHERENTYN; adherence is defined as at least 3 days of at least 10 hours of wear time each). This missing data pattern displayed in output 12.2.2 can be obtained by creating a cross tabulation of the indicator variables ACTICALYN and ADHERENTYN; 0 corresponds to missing.

```
ODS SELECT MISSPATTERN;
PROC MI DATA = WORK.CH12_DATA NIMPUTE=0;
      VAR VALID_SPIROMETRY VALID_SPIROMETRY_POSTBD;
RUN;
```

Output 12.2.1. Missing data pattern for spirometry measures

1st Spirometry (before dilation)	*2nd Spirometry (after dilation)*	*Freq*	*Percent (%)*
X	*X*	*1178*	*7.18*
X	*.*	*14431*	*87.80*
O	*O*	*806*	*5.03*

```
proc freq data = work.ch12_data;
      table ACTICALYN*ADHERENTYN / list;
run;
```

Output 12.2.2. Missing data pattern in objectively-measured physical activity measures

ACTICALYN (Actical returned)	*ADHERENTYN (Adherent Participant)*	*Freq*	*Percent (%)*
0 (No)	*0 (No)*	*1502*	*9.15*
1 (Yes)	*0 (No)*	*2163*	*13.18*
1 (Yes)	*1 (Yes)*	*12750*	*77.67*

In SAS, we can use PROC MI as a quick way to check the extent of missing data and the missing patterns without doing multiple imputation. **Output 12.2.3** shows PROC MI output for a non-monotone missing pattern by including variables BMI, EDUCATION_C3 and US_BORN in the HCHS/SOL full cohort (n=16,415). “Group 1” shows 16,270 participants (99.12 %) with no missing data for all three variables. “Group 2-6” show all the possible combinations of missingness among these three variables. We specify the option “nimpute=0” to avoid the imputation steps, and the ODS to only display the missing patterns in the output.

```
ODS SELECT MISSPATTERN;
PROC MI DATA=WORK.SOL NIMPUTE=0;
  VAR BMI EDUCATION_C3 US_BORN;
RUN;
```

Output 12.2.3. Missing data pattern for BMI, EDUCATION_C3, and US_BORN

Missing Data Patterns

Group	*BMI*	*EDUCATION_C3*	*US_BORN*	*Freq*	*Percent*
1	X	X	X	16270	99.12
2	X	X	.	1	0.01
3	X	.	X	19	0.12
4	X	.	.	54	0.33
5	.	X	X	53	0.32
6	O	O	O	18	0.11

Recommendations:

- Evaluate the extent of the missing for the outcome and covariates and identify the missing data patterns.
- Consider mechanisms that might generate the missing data.

12.3 Approaches for handling missing data

In this section, we summarize four statistical approaches to handle missing data: (1) complete case analyses (CCA), (2) multiple imputation (MI), (3) likelihood-based methods, and (4) inverse probability weighting (IPW). In general, if the missing data in the regression model is MCAR and does not depend on any measured or unmeasured variables, then regression coefficients are unbiased given the complete case analysis. **If the missingness of the covariates is MAR, then we should address the missing data problem using MI, likelihood-based methods, or IPW.** We will explain these four methods in the following subsections.

12.3.1 Complete Case Analysis (CCA)

The complete case analysis is based on the subset with no missing data in the outcome and the covariates. This method is easy to implement and it is the default method in most statistical software. CCA is valid when (1) the missingness is MCAR or (2) the overall missing rate is small, such as **less than 5% of the total sample.** Under MCAR, CCA still provides unbiased estimates, but CCA causes a loss of efficiency (larger standard errors) to some extent depending on the amount of missing data. Even if the MCAR assumption is violated but the overall missing rate is small, the impact of the bias is likely to be small.

12.3.2 Multiple Imputation (MI)

Multiple imputation (MI) is a commonly used approach to handle missing data under the MAR assumption. It is applicable to the missingness in both outcomes and covariates, and it can be used for both monotone and non-monotone missing patterns. **It has three steps:**

1. Generate m (typically m=5 to 10) imputed values for each missing observation using joint **imputation models**. This will result in m complete datasets with no missing data. Several (m) imputed values are required to reflect uncertainty about the missing value.
2. Fit the statistical model (**analysis model**) in each complete dataset.
3. Combine the results of m separate analyses using Rubin's rule (Rubin 1987), accounting for uncertainty in the imputation.

A key point in MI is to appropriately specify the covariates and interactions in the imputation model. Misspecified imputation models lead to biased estimates. The imputation model must be a richer model and the analysis model must be nested in the imputation model in order to reduce bias. **In other words, the variables used in the analysis model should always be included in the imputation model.**

The standard error of the final estimates is based on both the between- and within-imputation variability. Rubin (1987, p. 114) shows that the efficiency of an estimate based on m imputations is approximately $\left(1+\frac{r}{m}\right)^{-1}$, where r is the rate of missing information for the variable being imputed. Unless the rate of missing information is very high, the standard error is thus slightly reduced as the number of imputation datasets increases. In most situations, there is simply little advantage to producing and analyzing more than a few imputed datasets. Thus, we recommend generating 5-10 datasets if the extent of overall missing data is not too large (< 20%) to achieve > 96% of the maximum statistical efficiency. The advantage of using a higher number of imputed data sets is better coverage of MI confidence intervals or power levels for MI hypothesis tests (Berglund and Heeringa, 2014).

There are various methods to impute missing data in step 1 depending on the types of variables (continuous, nominal, or ordinal variables) and the missing data pattern. For monotone missing data patterns, missing data can be imputed by linear and logistic regression based on fully observed variables. When the number of variables in the imputation model is large, monotone missing patterns are less common. For non-monotone missing patterns, we can apply the conditional Gaussian approach, Markov Chain Monte Carlo (MCMC) methods, chained equations (aka fully conditional, or predictive mean matching. See section 2.4 of Horton and Kleinman (JASA 2007) for more details and references.

MI is available in many commonly used statistical packages. **In SAS 9.3, PROC MI** imputes the datasets and **PROC MIANALYZE** combines the analysis results from the multiple imputed complete datasets using Rubin's rule. See SUGI paper 265-2010 "An introduction to MI of complex survey data using SAS" by P. Berglund. In particular, PROC

MI has various imputation methods including predictive mean matching, regression, MCMC, discriminant models, and fully conditional specification (FCS also known as MICE acronym for multiple imputations chained equations). This latter method is described in Raghunathan et al (2001) and van Buuren (2007). The FCS models currently implemented in SAS MI procedure are:

1) Linear regression for continuous outcomes.
2) Linear regression with mean matching for continuous outcomes:
   - Similar to linear regression, but the imputed value is chosen from a set of closest real values observed in the data, to ensure that all the imputed values are feasible;
3) Discriminant method for categorical outcomes (nominal)
   - Default for categorical outcomes.
   - **Note: FCS method assumes normality of all model covariates in the imputation model**. By default, classification variables are not used as model covariates in the imputation model with this method, unless a special option is requested (CLASSEFFECTS=INCLUDE).
   - To reduce bias, a normalizing transformation should be applied to non-normal (skewed) continuous variables as each variable is assumed to have a normal marginal distribution. Fewer biases have been observed with the log or *lnskew* (STATA) transformations versus using non-transformed (skewed) variables (Berglund, 2015).
4) Logistic regression for categorical outcomes:
   - Can be requested with the LOGIT link for binary outcomes and the proportional odds model, or with the GLOGIT link for the generalized logit model with nominal outcomes.

In **Sudaan 11,** the **PROC IMPUTE** offers four imputation methods under the survey sampling framework. **In R and S-Plus**, the MICE package implements the chained equations method, and the Hmisc package supports the predictive mean matching method. **In STATA**, the ICE package implements the chained equations method. There are other standalone software packages for MI such as IVEware and LogXact. However, none of those methods account for the complex survey design. The complex survey design needs to be considered when applying any missing data approach. Ignoring the design factors can cause bias in the MI. There is evidence based on simulations in the literature (Reiter et al, 2006) that in the complex survey design settings the imputation model should account for stratification and clustering with respect to sampling strata and primary sampling units (PSU). Reiter et al (2006) suggest reasonably simple strategies for incorporating the sampling design into MI. See APPENDIX for a brief summary.

12.3.3 Likelihood-Based Approaches

Full-information maximum likelihood (FIML) is a model-based missing data procedure in which subjects with complete and partially complete data are analyzed together, and model parameters are estimated using all ("full") of the information available. The only observations excluded are those with all outcomes and covariates with missing values. A likelihood-based approach can be implemented for monotone and non-monotone missing patterns for baseline covariates. The likelihood method follows the same steps as the expectation-maximization (EM) algorithm. This approach is based on the assumption of MAR, and requires the knowledge of nuisance distributions of the covariates. For example, assume we have one outcome variable Y, and two covariates X_1 and X_2. If X_2 is MAR with Y and X_1 fully observed, we need to generate the conditional distribution of X_2 given the observed value of Y and X_1. This conditional distribution is used in the expectation step with the current parameter estimates. In the maximization step, the parameters of interests are estimated based on the expectation step. The procedure is repeated until the algorithm converges. When the missing pattern is complicated, the joint distribution of all the covariates and outcome variables is hard to generate. Assumptions need to be made in order to simplify the calculation. **Mplus allows implementing FIML to handle non-monotone missing data under MAR or MCAR.**

12.3.4 Inverse Probability Weighting (IPW)

Inverse probability weighting (IPW) is one method which allows correcting for the bias of the estimates obtained by complete-case analyses and can be implemented in complex survey designs. It assumes MAR and it is applicable to both non-monotone and monotone missing patterns. In HCHS/SOL, IPW has been implemented for the analysis of physical activity data, as measured by an accelerometer, and pulmonary outcomes. In these examples, we did not expect other baseline covariates to highly predict the specific missing values, as would be done in MI, yet we did expect that baseline covariates might reasonably predict the missing status of these outcomes.

To compute the IPW in a survey sample with missing data, we fit an unweighted logistic regression (i.e. not weighted by sampling weight) on the missing status of the participant, but include the design variables (strata and PSU) and the sampling weight as covariates to capture the potential effect of design factors (e.g. social-economic status, high or low concentration of Hispanics in the strata, and survey non-response) on the probability of missingness. IPW for individuals with complete data is then calculated as the inverse of the predicted probability of being complete. The subsequent analyses should then adjust for the product of IPW and sampling weight to obtain unbiased estimates accounting for missing data and survey sampling design. For further background, **Seaman and White (2013) provide a review of the implementation and advantages and disadvantages of using IPW to handle missing data in epidemiological research.**

Similar to MI, IPW is sensitive to misspecification of the logistic model used to predict missingness. Since IPW is based on the predicted probability of being complete, one should make the logistic regression model as rich as feasible to increase the prediction accuracy. Note that the missingness model can contain different covariates from the subsequent

analysis models but must contain all participants to be included in the analysis model. One problem that can occur while building a rich IPW model is the loss of records due to sporadic missing values across the covariates included in the IPW model. In this situation, MI can be done on the covariates to make m imputed datasets with complete covariates, then the inverse probability weight can be calculated in each of the imputed datasets, and a final IPW is obtained through Rubin's Rule. This method was used to calculate the IPW of the actical physical activity and some pulmonary outcomes.

Compared to MI, IPW is generally less efficient. In other words, IPW will produce a larger standard error for the estimates compared to MI. This is because IPW is based on complete cases and multiple imputation makes use of all observed data. IPW may be most applicable when there is a large amount of missing outcome data which might not be strongly modeled by covariates through MI.

12.3.5 Recommendations

- In cases where the overall missing rate is small, e.g. less than 5% of the total sample, we can conduct complete case analysis (CCA) regardless of the missing data mechanism. The analytic sample is comprised of those subjects with non-missing outcomes and covariates. In the methods section of the manuscript, make sure to report the number of participants that were excluded.
- IPW may be most applicable when there is a large amount of missing outcome data which might not be strongly modeled by covariates through MI.
- MI is preferable when the missing data are mostly in the covariates and less in the outcome because it makes use of the partially observed data.
- When the imputation model and the analysis model are the same (i.e., the same variables are available for analysis), MI and FIML yield similar results.

12.4 EXAMPLE using SAS and MPlus

As a motivating example, we will use MS135 "Associations of objectively measured physical activity (PA) and sedentary behavior with depressive symptoms: Results from HCHS/SOL". In this manuscript, the authors were interested in estimating the association between PA (light, moderate and vigorous physical activity) and depression symptoms score (CESD10) separately by PA assessment (accelerometer or self-report questionnaire). **For illustration purposes**, we will estimate the association between moderate to vigorous PA (MV_DAY) and CESD10 adjusted for potential confounders using three missing data methods: MI, IPW, and FIML. MI and IPW are implemented using SAS and FIML using Mplus.

Let us assume we are interested in estimating the association between moderate to vigorous PA (MV_DAY) and CESD10 adjusted for age, sex, BMI, education level, income, physical health (assessed by the SF-12), and SASH social and language subscales. In addition to fitting a full model we are interested in fitting several reduced models (nested within each other); for example one only adjusting for age, sex, BMI, background and education. We will not impute Hispanic/Latino background. Instead, we combine the missing background (n=87) with those with Mixed/Other background (n=503).

FULL MODEL:
CESD10 = MVPA + AGE + MALE + BMI + BKGRD1_C7NOMISS + EDUCATION_C3 + INCOME + AGG_PHYS + SASH_SOC + SASH_LANG + e

REDUCED MODEL (nested within FULL model):
CESD10 = MVPA + AGE + MALE + BMI+ BKGRD1_C7NOMISS + EDUCATION_C3 +e

The first step for all three methods (MI, IPW, and FIML) is to assess the extent of missing data and the missing data patterns coming from variables in the analysis model (outcome, exposure and covariates) plus all the variables needed to account for missing data. We know variables AGE and MALE don't have any missing, and from output 12.2.2 that 12,750 participants have accelerometer data. There is very little missing data in CESD10 and hence there are 12,577 observations with no missing data in the outcome (CESD10) and exposure (MV_DAY). However, the reduced model has 12,533 observations (output 12.4) whereas the full model has only 11,007 observations.

```
ODS SELECT MISSPATTERN;
PROC MI DATA=WORK.CH12 NIMPUTE=0;
  VAR CESD10 MV_DAY MALE AGE BMI EDUCATION_C3;
RUN;
```

Output 12.4 Missing Data patterns for CESD10, MV_DAY, MALE, AGE, BMI and EDUCATION C_3

Missing Data Patterns

GROUP	CESD10	MV_DAY	MALE	AGE	BMI	EDUCATION_C3	FREQ	PERCENT
1	**X**	**X**	**X**	**X**	**X**	**X**	**12533**	**76.35**
2	X	X	X	X	X	.	22	0.13
3	X	X	X	X	.	X	22	0.13
4	X	.	X	X	X	X	3452	21.03
5	X	.	X	X	X	.	9	0.05
6	X	.	X	X	.	X	21	0.13
7	.	X	X	X	X	X	168	1.02
8	.	X	X	X	X	.	4	0.02
9	.	X	X	X	.	X	1	0.01
10	.	.	X	X	X	X	118	0.72
11	.	.	X	X	X	.	38	0.23
12	.	.	X	X	.	X	9	0.05
13	.	.	X	X	.	.	18	0.11

12.4.1 Complete Case Analysis (CCA)

The complete case analysis (CCA) of the reduced model is carried out in SAS and is shown for reference (output 12.4.2).

```
PROC SURVEYREG DATA = WORK.CH12_DATA;
      STRATA STRAT;
      CLUSTER PSU_ID;
      CLASS EDUCATION_C3 MALE;
      DOMAIN ADHERENTYN;
      WEIGHT WEIGHT_FINAL_NORM_OVERALL;
      MODEL CESD10 = MV_DAY AGE MALE BMI EDUCATION_C3 / SOLUTION;
RUN;
```

Output 12.4.1 Linear regression of CESD10 using CCA (N=12,533), PROCSURVEYREG with WEIGHT_FINAL_NORM_OVERALL

Estimated Regression Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	4.4712386	0.45893805	9.74	<.0001
MV_DAY	0.0025763	0.00215964	1.19	0.2333
AGE	0.0182471	0.00508416	3.59	0.0004
MALE 0	-1.8059081	0.14981329	-12.05	<.0001
BMI	0.0600064	0.01340920	4.48	<.0001
EDUCATION_C3 1	1.4224329	0.18708192	7.60	<.0001
EDUCATION_C3 2	0.7176376	0.18500626	3.88	0.0001

12.4.2. Multiple Imputation (MI) using SAS

12.4.2.1 Only MV_DAY imputed from accelerometer variables

As explained in section 12.3.2 **multiple imputation (MI) has three general steps** and a key point is to appropriately specify the covariates and interactions in the imputation model. Misspecified imputation models lead to biased estimates. The imputation model must be a richer model and the analysis model must be nested in the imputation model in order to reduce bias. In other words, the variables used in the analysis model should always be included in the imputation model. **The three steps for implementing MI are:**

Step 1. Fit the imputation model to generate m complete datasets (typically 5)

a. Include all of the following variables:

- Outcome: CESD10
- Main variable of interest (exposure in epidemiology): MV_DAY

- Confounders for the analysis model: MALE, INCOME_C5_NOMISS, AGE, BMI, AGG_PHYS, SASH_SOC, SASH_LANG, EDUCATION_C3
- HCHS/SOL study design variables: STRAT, WEIGHT_FINAL_NORM_OVERALL. We were not able to include PSU_ID because the models do not converge.
- Variables associated with the probability of being a missing case
- Include interaction terms as appropriate

b. Choose a method to impute missing data that suits the type and pattern of missing data (see section 3.2).
c. Carefully determine the model specification for imputing each variable: its scale and which variables could help to predict it.
d. Add constraints (e.g. min and max) to obtain only plausible imputed values

Output 12.4.2 Missing data per variable

Variable	N	N Miss	Minimum	Maximum
STRAT	16415	0	10.00	29.00
WEIGHT_FINAL_NORM_OVERALL	16415	0	0.08	20.78
MALE	16415	0	0.00	1.00
AGE	16415	0	18.00	76.00
BKGRD1_C7NOMISS	16415	0	0.00	6.00
BMI	16344	71	13.82	70.35
EDUCATION_C3	16324	91	1.00	3.00
SASH_LANG	16313	102	1.00	5.00
AGG_PHYS	16176	239	5.47	76.01
CESD10	16059	356	0.00	30.00
SASH_SOC	15686	729	1.00	5.00
INCOME	14927	1488	1.00	10.00
SED_DAY	12750	3665	17.17	1357.83
LIGHT_DAY	12750	3665	4.00	1325.67
MOD_DAY	12750	3665	0.00	643.50
VIG_DAY	12750	3665	0.00	304.17
MV_DAY	12750	3665	0.00	643.50

See SAS online documentation for the different methods for multiple imputation that are available in the MI procedure, and for details on the statements and options. Briefly, FCS statement specifies a multivariate imputation by fully conditional specification methods. DISCRIM specifies the discriminant function method for nominal categorical variables. LOGISTIC specifies the logistic regression method for binary or ordinal categorical variables. REG specifies the regression method for continuous covariates, and variables

are modeled as continuous using REG unless specified otherwise. By default, PROC MI imputes the variables following the order in the VAR statement and hence we have to order the variables from the least amount of missing to the most. The MIN and MAX values can be specified to obtain imputed values within the range of what was observed for a particular variable. These values must be listed in the same order as the variables are listed in VAR the statement. For example, AGE is the 4th variable listed in the VAR statement. The corresponding values for the minimum and maximum AGE should be listed in the 4th position in the MIN and MAX statements; in this case 18 and 76 respectively. A period can be used in the place of a number if no min and/or max needs to be specified. Min and max values can't be specified for CLASS variables and thus a period should be used in the corresponding positions. For an imputed variable that uses the discriminant function method, if no covariates are specified, then all other variables in the VAR statement are used as the covariates with the CLASSEFFECTS = INCLUDE option. SEED is specified in order to replicate the results. Because we specified NIMPUTE=**10,** the five imputed datasets will be stacked and identified with the variable _imputation_ and saved in one single dataset named `WORK.MI10`.

```
PROC MI DATA = WORK.CH12_DATA SEED=1645 NIMPUTE=10 OUT = WORK.MI10
      MIN = . . . . . 13 . 1 5 0 1 . 17 4 0 0
      MAX = . . . . . 71 . 5 77 30 5 . 1358 1326 644 305;
      CLASS STRAT MALE EDUCATION_C3 INCOME BKGRD1_C7NOMISS;
      VAR STRAT WEIGHT_FINAL_NORM_OVERALL MALE AGE BKGRD1_C7NOMISS BMI
          EDUCATION_C3 SASH_LANG AGG_PHYS CESD10 SASH_SOC INCOME MV_DAY;
      FCS   DISCRIM (STRAT BKGRD1_C7NOMISS / CLASSEFFECTS=INCLUDE)
            LOGISTIC(MALE EDUCATION_C3 INCOME)
      REG /* BY DEFAULT, ALL CONTINUOUS VARIABLES */ PLOTS=TRACE(MEAN);
RUN; ODS GRAPHICS OFF;
```

Information on within- and between-imputation variance is part of the standard output of PROC MI (**output 12.4.3**). Within-imputation variance is the average of the sampling variances from the 10 imputed datasets. Between-imputation variance is a measure of the variability in the parameter estimates obtained from the 10 imputations. A descriptive procedure can be used to confirm that variables of interest have zero missing values (**output 12.4.4**). For example, pre-imputation, CESD10 had 356 missing values and post-imputation there are 0 missing values. Similarly, SED_DAY had 3665 missing values pre-imputation and 0 post-imputation.

Output 12.4.3 Within and between variance information (10 imputations)

Variable	Variance			DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
	Between	Within	Total				
BMI	0.000002524	0.002246	0.002249	16346	0.001236	0.001235	0.999877
SASH_LANG	0.000000209	0.000071391	0.000071621	16059	0.003217	0.003209	0.999679
AGG_PHYS	0.000074281	0.005806	0.005888	12021	0.014072	0.013919	0.998610
CESD10	0.000045086	0.002262	0.002311	8816.6	0.021927	0.021557	0.997849
SASH_SOC	0.000000759	0.000021812	0.000022647	4663.6	0.038297	0.037175	0.996296
MV_DAY	0.006246	0.055401	0.062272	703.65	0.124020	0.112733	0.988852

Output 12.4.4 Asses missing data AFTER imputation (10 imputations)

Variable	N	N Miss	Minimum	Maximum
STRAT	164150	0	10	29
WEIGHT_FINAL_NORM_OVERALL	164150	0	0.08	20.78
MALE	164150	0	0	1
AGE	164150	0	18	76
BKGRD1_C7NOMISS	164150	0	0	6
BMI	164150	0	13.34	70.35
EDUCATION_C3	164150	0	1	3
SASH_LANG	164150	0	1	5
AGG_PHYS	164150	0	5.47	76.93
CESD10	164150	0	0	30
SASH_SOC	164150	0	1	5
INCOME	164150	0	1	10
MV_DAY	164150	0	0	947.22

Step 2. Run the analysis model using the imputed data sets in PROC SURVEYREG

The output data from PROC MI has the ten complete datasets (all missing values are imputed) stacked and identified with the variable _imputation_. In other words, each of the ten datasets has 16,415 observations and no missing data in any variable. We can use the statement “by” to simultaneously analyze all five datasets in one single call of PROC SURVEYREG.

```
PROC SURVEYREG DATA = WORK.MI10;
      BY _IMPUTATION_;
      STRATA STRAT;
      CLUSTER PSU_ID;
      CLASS EDUCATION_C3 MALE;
      WEIGHT WEIGHT_FINAL_NORM_OVERALL;
      MODEL CESD10 = MV_DAY AGE MALE BMI EDUCATION_C3 / SOLUTION;
      ODS OUTPUT PARAMETERESTIMATES = WORK.OUTREGEX2;

DATA OUTREGEX2;
      SET OUTREGEX2;
      PARAMETER=COMPRESS(PARAMETER);
      WHERE PARAMETER^="EDUCATION_C3";
RUN;
```

Step 3. Combine the results of m separate analyses using Rubin’s rule accounting for uncertainty in the imputation using PROC MIANALYZE.

The data set produced by the ODS OUTPUT statement of PROC SURVEYREG requires the use of the COMPRESS option to format the data such that PROC MIANALYZE can correctly process the parameter estimates. The syntax below illustrates how to correctly remove the blanks in the variable called PARAMETER in the output data set “outregex2”:

```
PROC MIANALYZE PARMS=OUTREGEX2;
      MODELEFFECTS INTERCEPT MV_DAY AGE MALE BMI EDUCATION_C31 EDUCATION_C32;
RUN;
```

Output 12.4.5 Linear regression of CESD10 using multiple imputation and only MV_DAY included from accelerometer variables (N=16,415), PROC MIANALYZE

Parameter	Estimate	Std Error	95% Confidence Limits		DF	t for H0	Pr > |t|
INTERCEPT	4.192533	0.430504	3.34868	5.03639	12528	9.74	<.0001
MV_DAY	0.003948	0.002342	-0.00067	0.00856	234.63	1.69	0.0932
AGE	0.020501	0.004622	0.01144	0.02956	13039	4.44	<.0001
MALE0	-1.862460	0.131463	-2.12013	-1.60479	40669	-14.17	<.0001
BMI	0.071545	0.012597	0.04685	0.09624	18549	5.68	<.0001
EDUCATION_C31	1.555730	0.185752	1.19165	1.91981	33597	8.38	<.0001
EDUCATION_C32	0.713161	0.165894	0.38801	1.03831	57044	4.30	<.0001

12.4.2.1 All 4 intensities (sedentary, light, moderate and vigorous) imputed

Because MV_DAY is the sum of minutes in moderate and in vigorous activity we will impute these two variables directly and then sum them up to impute MV_DAY. Further, we will impute SED_DAY and LIGHT_DAY as all four intensities jointly can help better impute MV_DAY. TOT_HRS (which we will not use in this analysis) would be created after imputing directly the four intensities: sedentary, light, moderate and vigorous.

Given that MV_DAY is the sum of minutes from moderate and vigorous (i.e. MOD_DAY, VIG_DAY) it was not included in the imputation. Hence, after imputing missing values, TOT_HRS is computed as the sum of SED_DAY, LIGHT_DAY, MOD_DAY, and VIG_DAY divided by 60.

```
DATA WORK.MI10;
      SET WORK.MI10;
      MV_DAY = SUM (OF MOD_DAY VIG_DAY);
RUN;
```

Output 12.4.6 Linear regression of CESD10 using multiple imputation and all four activity intensities included (sedentary, light, moderate & vigorous; N=16415), PROC MIANALYZE

Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		DF	t for H0:	Pr > |t|
INTERCEPT	4.238067	0.424776	3.40549	5.07064	32499	9.98	<.0001
MV_DAY	0.004989	0.002355	0.00036	0.00962	499.87	2.12	0.0346
AGE	0.021038	0.004709	0.01181	0.03027	3820.5	4.47	<.0001
MALE	-1.874621	0.132402	-2.13413	-1.61511	76028	-14.16	<.0001
BMI	0.068607	0.012424	0.04425	0.09297	3502.8	5.52	<.0001
EDUCATION_C31	1.529852	0.184958	1.16733	1.89237	57052	8.27	<.0001
EDUCATION_C32	0.719868	0.168122	0.39029	1.04944	6515.8	4.28	<.0001

Note that all parameter estimates and standard errors are very similar to those from IPW and MI with only MV_DAY imputed. However, the regression coefficient for MV_DAY is now significantly different to zero.

12.4.3. FIML (Full Information Maximum Likelihood) using Mplus

SAS code to create data for Mplus

```
data dataforMplus (keep = PSU_ID STRAT WEIGHT_PA_IPW_OVERALL SASH_LANG AGG_PHYS
SASH_SOC AGE BMI CESD10 MV_DAY  ADHERENTYN MALE EDU1 EDU2
BMI AGG_PHYS SASH_LANG SASH_SOC);
  set work.CH12_DATA;
  /* CREATE DUMMYS FOR MPLUS */
  if EDUCATION_C3 ne . then do;
     EDU1 = (EDUCATION_C3=1);
     EDU2 = (EDUCATION_C3=2);
  end;
  ID = _n_;
run;
```

Mplus code to run the REDUCED MODEL and auxiliary data

```
DATA:
FILE = 'Ch12ForMplus.dat';

VARIABLE:
!VARIABLES IN THE SAME ORDER OF AS CREATED IN THE DATASET;
NAMES = PSU_ID STRAT IPW AGE BMI CESD10 MV_DAY
        ADHERENTYN MALE EDU1 EDU2 ID;
MISSING = .;
USEVARIABLES = PSU_ID STRAT IPW AGE BMI CESD10 MV_DAY MALE EDU1 EDU2
```

```
               BKGRD0 BKGRD1 BKGRD2 BKGRD4 BKGRD5 BKGRD6 SASH_LANG
               AGG_PHYS SASH_SOC INCOME;
CLUSTER = PSU_ID;
STRAT = strat;
WEIGHT = IPW; !IPW is WEIGHT_PA_IPW_OVERALL
AUXILIARY = (m) BKGRD0 BKGRD1 BKGRD2 BKGRD4 BKGRD5 BKGRD6 SASH_LANG
                AGG_PHYS SASH_SOC INCOME;

ANALYSIS:
TYPE = COMPLEX;

MODEL:
CESD10 on MV_DAY AGE MALE BMI EDU1 EDU2;


SUMMARY OF ANALYSIS

Number of groups                                                 1
Number of observations                                       12533

Number of dependent variables                                    1
Number of independent variables                                  6
Number of continuous latent variables                            0

Observed dependent variables

  Continuous
   CESD10

Observed independent variables
   AGE         BMI         MV_DAY      MALE        EDU1        EDU2

Variables with special functions

  Stratification        STRAT
  Cluster variable      PSU_ID
  Weight variable       IPW

Estimator                                                      MLR
Information matrix                                        OBSERVED
Maximum number of iterations                                 1000
Convergence criterion                                   0.500D-04
Maximum number of steepest descent iterations                  20
Maximum number of iterations for H1                          2000
Convergence criterion for H1                            0.100D-03

Input data file(s)
  Ch12ForMplus.dat
Input data format  FREE


SUMMARY OF DATA

                     Number of missing data patterns        1
    Number of strata                        20
    Number of clusters                     661
```

THE MODEL ESTIMATION TERMINATED NORMALLY

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
CESD10 ON				
MV_DAY	0.003	0.002	1.210	0.226
AGE	0.021	0.005	3.972	0.000
MALE	-1.790	0.164	-10.891	0.000
BMI	0.060	0.014	4.188	0.000
EDU1	1.497	0.202	7.420	0.000
EDU2	0.684	0.212	3.231	0.001
Intercepts				
CESD10	4.382	0.512	8.562	0.000
Residual Variances				
CESD10	32.605	0.858	38.020	0.000

MODEL FIT INFORMATION

Number of Free Parameters 8

Loglikelihood Including the Auxiliary Part

H0 Value	-322035.072
H0 Scaling Correction Factor for MLR	4.0565
H1 Value	-322035.072
H1 Scaling Correction Factor for MLR	4.0565

Information Criteria Including the Auxiliary Part

Number of Free Parameters	170
Akaike (AIC)	644410.144
Bayesian (BIC)	645676.548
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	645136.306

12.4.4 IPW (Inverse probability weighting) using SAS

First, we use logistic regression to predict missing status and get the IPW. See section 6 in HCHS/SOL Physical Activity Data Overview, Methods and Guidelines for a detailed description on how WEIGHT_PA_IPW_OVERALL was calculated to be used for analyzing objectively-measured physical activity derived variables. The weights obtained from the IPW procedure can be used in the `WEIGHT` statement in any of the SAS survey procedures. This is exemplified using PROC SURVEYREG below.

```
PROC SURVEYREG DATA = CH12_DATA;
      STRATA STRAT; CLUSTER PSU_ID; WEIGHT WEIGHT_PA_IPW_OVERALL;
      CLASS EDUCATION_C3 MALE;
      DOMAIN ADHERENTYN;
      MODEL CESD10 = MV_DAY AGE MALE BMI EDUCATION_C3 / SOLUTION;
RUN;
```

Output 12.4.7. Linear regression of CESD10 using IPW (N=12,533), PROC SURVEYREG with WEIGHT_PA_IPW_OVERALL

Estimated Regression Coefficients

Parameter	Estimate	Std Error	t Value	Pr > |t|
INTERCEPT	4.4072511	0.51159112	8.61	<.0001
MV_DAY	0.0027639	0.00227422	1.22	0.2247
AGE	0.0216528	0.00537874	4.03	<.0001
MALE	-1.7754484	0.16447125	-10.79	<.0001
BMI	0.0582915	0.01432972	4.07	<.0001
EDUCATION_C3 1	1.4954092	0.20174778	7.41	<.0001
EDUCATION_C3 2	0.6849468	0.21166351	3.24	0.0013

12.5 Recommendations on reporting missing data

See the three proposed guidelines for reporting missing covariate data given in Figure 3 in Horton and Kleinman (JASA, 2007) which is a reprint of Burton and Altman (British J of Cancer, 2004).

Proposed guidelines for reporting missing covariate data
(Figure 3 from Burton and Altman 2004)

1. quantification of completeness of covariate data
(a) if availability of data is an exclusion criterion, specify the number of cases excluded for this reason,
(b) provide the total number of eligible cases and the number with complete data,
(c) report the frequency of missing data for every variable considered. If there is only a small amount of overall missingness (e.g. $> 90\%$ of cases with complete data), then the number of incomplete variables and the maximum amount of missingness in any variable are sufficient

2. approaches for handling missing covariate data
(a) provide sufficient details of the methods adopted to handle missing covariate data for all incomplete covariates
(b) give appropriate references for any imputation method used
(c) for each analysis, specify the number of cases included and the associated number of events

3. exploration of the missing data
(a) discuss any known reasons for missing covariate data
(b) present the results of any comparisons of characteristics between the cases with or without missing data

References

Berglund PA. An Introduction to Multiple Imputation of Complex Sample Data using SAS® v9.2. SAS Global Forum 2010. Paper 265-2010.

Berglund PA Multiple Imputation Using the Fully Conditional Specification Method: A Comparison of SAS®, Stata, IVEware, and R . SAS Global Forum 2010. Paper 2081-2015.

Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol. 1995 Dec 15;142(12):1255-64.

Haukoos JS, Newgard CD. Advanced statistics: missing data in clinical research--part 1: an introduction and conceptual framework. Acad Emerg Med. 2007 Jul;14(7):662-8.

Haziza D and Picard F (2012). Doubly robust point and variance estimation in the presence of imputed survey data. Can J Stat 40(2): 259-281.

Horton & Kleinman (2007). "Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models." *The American Statistician* 61(1):79-90.

Horton, Lipsitz, & Parzen (2003). "A potential for bias when rounding in MI." *The American Statistician* 57(4):229-232.

Moore CG, Lipsitz SR, Addy CL, Hussey JR, Fitzmaurice G, Natarajan S. Logistic regression with incomplete covariate data in complex survey sampling: application of reweighted estimating equations. Epidemiology. 2009;20(3):382-90.

Newgard CD, Haukoos JS. Advanced statistics: missing data in clinical research--part 2: multiple imputation. Acad Emerg Med. 2007 Jul;14(7):669-78.

Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P. (2001) A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. Survey Methodology, 27(1), 85-95.

Rathouz PJ, Preisser JS. Missing Data: Weighting and Imputation, in Encyclopedia of Health Economics: Third Edition, Revised and Expanded, ed. A. J. Culyer. Elsevier, Inc, (in press).

Reiter JP, Raghunathan TE and Kinney SK (2006). The importance of modeling the sampling design in multiple imputation for missing data. Survey Methodology 32(2): 143-149.

Seaman SR and White IR (2011). Review of inverse probability weighting for dealing with missing data. Statistical Methods in Medical Research 0(0): 1-18.

Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009 Jun 29;338:b2393.

Books

Little, RJA and Rubin, DB. (2002). *Statistical Analysis with Missing Data.* Rubin, DB. 2nd ed. Hoboken, NJ: Wiley.

Rubin, DB (1987) *Multiple imputation for nonresponse in surveys.* New York, NY: Wiley.

Heeringa SG, West BT and Berglund PA (2010). Applied Survey Data Analysis. Chapman & Hall / CRC Press.

Other Resources

Dr Rod Little website: http://sitemaker.umich.edu/rlittle/missing_data

Carpenter J: http://missingdata.lshtm.ac.uk/

Van Buuren (software for Multiple Imputation): http://multiple-imputation.com/

Useful Websites for Examples of Software and Statistical Analysis:

NHANES Sample Code & Datasets
http://www.cdc.gov/nchs/tutorials/nhanes/Downloads/intro.htm
Program code used in this tutorial follow procedures available in SAS 9.1, SAS 8.0 and SUDAAN 9.0 or SUDAAN 8.0, and Stata/SE 10.0.
In particular section “Clean & Recode Data”:
Task 1: How to Identify and Recode Missing Data
Step 1: Identify Missing and Unavailable Values
Step 2: Recode Unavailable Values as Missing
Step 3: Evaluate Extent of Missing Data

UCLA Academic Technology Services (Statistical Computing)

SUDAAN FAQ
How can I use multiply imputed data sets in SUDAAN?
http://www.ats.ucla.edu/stat/sudaan/faq/mi.htm

Past Classes and Workshops Available Online
In particular Multiple Imputation using Stata or SAS
http://www.ats.ucla.edu/stat/seminars/default.htm

13. MULTIPLE COMPARISONS

Multiple testing is recognized as a common problem in epidemiological research but test procedures are underutilized because the problem is complex. In HCHS/SOL, as in any large epidemiologic study, there are several levels of multiplicity: several group comparisons (Hispanic/Latino background), more than one endpoint and repeated measurements of each endpoint. Bender and Lange (2001) provide a nontechnical overview on situations and methods for multiple hypothesis tests adjustment. In this section, we summarize those that are relevant for HCHS/SOL baseline analyses and provide some recommendations and examples.

Conducting multiple comparisons creates a type I error rate for the collection of tests that may be higher than the nominal test size (common alpha level of 0.05). Hence, adjustment for multiple testing is needed to reduce the probability of incorrectly declaring there are group differences when in reality there are not.

13.1 General procedure based on p-value

The well-known Bonferroni method is the simplest multiple test procedure. With k test significant at alpha level, the Bonferrroni method accepts those as statistically significant if their individual unadjusted p values are smaller than alpha/k. The adjusted p values are calculated by multiplying the individual unadjusted p-values by the number of tests. This methods is applicable in any multiple test situation but should be only used when the number of tests is small (<5) and the correlation among the tests statistics are quite small. Its main advantage is its applicability to different types of data (continuous, nominal, ordinal) and different tests statistics. The downside is that because of its generality it has low statistical power.

13.2 Special procedures for multiple test adjustments

13.2.1 Group comparisons

One of the primary aims in HCHS/SOL is to compare means (prevalences) among Hispanic/Latino background groups. In general, we do not recommend conducting statistical tests across Hispanic/Latino backgrounds when reporting population estimates (see section 11). However, when there is interest in testing which means (prevalences) are different after adjusting for important covariates multiple testing should be done. For example, if we want to compare the age-BMI adjusted prevalence of sleep apnea across all seven Hispanic/Latino backgrounds then there are 21 pairwise comparisons; 15 if the mixed/other group is excluded. **We recommend doing an overall test first. Only if the overall test is significant proceed to conducting pairwise comparisons adjusting for multiple comparisons. In particular Tukey-Kramer adjustment is appropriate because the design is unbalanced and the variances across groups are usually different.** See Sleep-disordered breathing in HCHS/SOL by Redline et al (*Am J Respir Crit Care Med* 2014) and an example using SAS complex survey procedures at the end.

13.2.2 Multiple endpoints

When there are multiple endpoints reported in a manuscript we recommend specifying one primary outcome. If this is not possible do a Bonferroni adjustment for the overall tests.

13.2.3 Subgroup analyses

If there is interest in testing the difference in effect by subgroup then first test the interaction. If significant, then conduct pairwise comparisons adjusting for multiple comparisons. If there is interest in stratified analyses (e.g. gender) we also recommend conducting pairwise comparisons adjusting for multiple comparisons within each stratum.

13.2.4 Example. Age-BMI adjusted prevalence by Hispanic/Latino background group; overall test and pairwise comparisons.

Age-BMI adjusted sleep apnea (AHI 3% desaturation ≥ 15) prevalence by HCHS/SOL background groups among men. From Table 4 of Redline, Sotres-Alvarez, Loredo *et al* (*Am J Respir Crit Care Med* 2014).

```
proc surveymeans data = ms13; /* MEAN age and BMI */
  strata strat; cluster PSU_ID; weight &weight; domain keep_ms13;
  var age bmi;
run;

%let meanage = 41.078370;
%let meanBMI = 29.317093;

proc surveyreg data = ms13 order=internal;
 format bkgrd1_c7 BKG1_C7F.;
 strata strat; cluster psu_id; weight &weight;
 domain keep_ms13m;  class bkgrd1_c7;
 model ahi3p_ge15_100 = bkgrd1_c7 age bmi / solution;
 lsmeans bkgrd1_c7 / at age=&meanage at bmi=&meanBMI CL lines adjust=Tukey;
run;
```

Output 13.2 Age-BMI adjusted sleep apnea prevalence by Hispanic/Latino background among men, PROC SURVEYREG

Tests of Model Effects

Effect	*Num DF*	*F Value*	*Pr > F*
Model	8	73.63	<.0001
Intercept	1	332.48	<.0001
BKGRD1_C7	**6**	**2.19**	**0.0419**
AGE	1	225.38	<.0001
BMI	1	205.77	<.0001

Tukey-Kramer Grouping for BKGRD1_C7 Least Squares Means (Alpha=0.05)

LS-means with the same letter are not significantly different.

BKGRD1_C7	*AGE*	*BMI*	*Estimate*		
Cuban	**41.08**	**29.30**	**17.7790**		**A**
	41.08	29.30			A
More than one	41.08	29.30	16.4186	B	A
	41.08	29.30		B	A
Dominican	41.08	29.30	15.9919	B	A
	41.08	29.30		B	A
Central American	41.08	29.30	15.1250	B	A
	41.08	29.30		B	A
Mexican	41.08	29.30	14.3663	B	A
	41.08	29.30		B	A
South American	41.08	29.30	12.9622	B	A
	41.08	29.30		B	
Puerto Rican	**41.08**	**29.30**	**11.7231**	**B**	

References

Bender R and Lange S. Adjusting for multiple testing--when and how? *J Clin Epidemiol.* 2001 Apr;54(4):343-9.

Redline S, Sotres-Alvarez D, Loredo J, Hall M, Patel SR, Ramos A, Shah N, Ries A, Arens R, Barnhart J, Youngblood M, Zee P, Daviglus ML. Sleep-disordered breathing in Hispanic/Latino individuals of diverse backgrounds. The Hispanic Community Health Study/Study of Latinos. *Am J Respir Crit Care Med*. 2014 Feb 1;189(3):335-44.

14. REFERENCES

Berglund PA. An Introduction to Multiple Imputation of Complex Sample Data using SAS® v9.2. SAS Global Forum 2010. Paper 265-2010.

Berglund PA Multiple Imputation Using the Fully Conditional Specification Method: A Comparison of SAS®, Stata, IVEware, and R . SAS Global Forum 2010. Paper 2081-2015.

Bender R and Lange S. Adjusting for multiple testing--when and how? *J Clin Epidemiol.* 2001 Apr;54(4):343-9.

Bieler GS, Brown GG, Williams RL, and Brogan DJ. (2010) Estimating Model-Adjusted Risks, Risk Differences, and Risk Ratios From Complex Survey Data Am. *J. Epidemiol*; 171(5): 618-623.

Brogan DG. (1998) Software for Sample Survey Data: Misuse of Standard Packages. *Encyclopedia of Biostatistics*. New York: John Wiley, Volume 5, pages 4167-4174.

Hellevik O. (2009) Linear versus logistic regression when the dependent variable is a dichotomy. *Quality and Quantity*; 43: 59-74.

Kish, L. (1965) *Survey sampling*. New York: John Wiley & Sons.

Klein RJ, Schoenborn CA. (2001) Age adjustment using the 2000 projected U.S. population. *Healthy People 2010 Stat Notes*; (20):1-10.

Koch GG, Gillings DB, and Stokes ME. (1980) Biostatistical implications of design, sampling, and measurement to health science data analysis. *Ann. Review Public Health;* 1:163-225.

Korn EL and Graubard BI. (1999) Analysis of Health Surveys. New York: John Wiley & Sons.

LaVange LM, & Kalsbeek W, et. Al. (2010) Sample design and cohort selection in the Hispanic Community Health Study/Study of Lations. *Annals of Epidemiology*; 20(8): 642-649.

LaVange LM, Koch GG , and Schwartz TA. (2001) Applying sample survey methods to clinical trials data. *Statist Med*; 20:2609–2623.

Siller AB, & Tompkins L. (2006) The big four: analyzing complex sample survey data using SAS, SPSS, STATA, and SUDAAN. *Proceedings of the Thirty-First SAS Users Group International; 172*-31.

Stokes ME, Davis CS, Koch GG (2000). Categorical Data Analysis Using the SAS System, 2nd edition. Cary, NC, SAS Institute.