Analytics Recommendations for HINTS 4 – Cycle 2 Data

October 2020

Table of Contents

Overview of HINTS

The Health Information National Trends Survey (HINTS) is a nationally-representative survey which has been administered every few years by the National Cancer Institute since 2003. The HINTS target population is all adults aged 18 or older in the civilian non-institutionalized population of the United States. The HINTS program collects data on the American public's need for, access to, and use of health-related information and health-related behaviors, perceptions and knowledge. (Hesse, et al., 2006; Nelson, et al., 2004). Previous iterations include HINTS 1 (2003), HINTS 2 (2005), HINTS 3 (2007/2008) and HINTS 4 Cycle 1 (2011/2012).

HINTS 4

The HINTS 4 administration includes four mail-mode data collection cycles over approximately three years starting in 2011. The second of these cycles (HINTS 4 Cycle 2) was conducted from October 2012 through January 2013, and is the focus of this report. HINTS 4 draws upon the lessons learned from prior iterations of HINTS while employing some new strategies (Link, 2005). Based on the higher response rates for the mail survey (over the RDD survey) in HINTS 3, a single-mode mail survey was implemented for all HINTS 4 cycles. For more extensive background about the HINTS program and previous data collection efforts, see Finney Rutten et al. (2012).

Methodology

Data collection for Cycle 2 of HINTS 4 was initiated in October 2012 and concluded in January of 2013. HINTS 4 Cycle 2 was a self-administered mailed questionnaire. The sampling frame of addresses, provided by Marketing Systems Group (MSG), was grouped into three strata: 1) addresses in areas with high concentrations of minority population; 2) addresses in areas with low concentrations of minority population; 3) addresses located in counties comprising Central Appalachia regardless of minority population. All non-vacant residential addresses in the United States present on the MSG database, including post office (P.O.) boxes, throwbacks (i.e., street addresses for which mail is redirected by the United States Postal Service to a specified P.O. box), and seasonal addresses, were subject to sampling. The protocol for mailing the questionnaires involved an initial mailing of the questionnaire, followed by a reminder postcard, and up to two additional mailings of the questionnaire as needed by non-responding households. Most households received one survey per mailing (in English), while households that were potentially Spanish-speaking received two surveys per mailing (one in English and one in Spanish). Refer to the HINTS 4 Cycle 2 Methodology Report for more extensive information about the sampling procedures.

A methodological experiment was embedded in Cycle 2 in an attempt to increase the participation of Spanish-speaking respondents and consisted of two levels: 1) Mailing both an English and Spanish questionnaire only to Spanish surname and Linguistically Isolated households; 2) Mail both an English and a Spanish questionnaire to all households. See the Methodology Report for more information. The second-stage of sampling consisted of selecting one adult within each sampled household using only the Next Birthday Method. In this method, the adult who would have the next birthday in the sampled household was asked to complete the questionnaire. A $2 monetary incentive was included with the survey to encourage participation.

Sample Size and Response Rates

The final HINTS 4 Cycle 2 sample consists of 3,630 respondents. Note that 48 of these respondents were considered partial completers who did not answer the entire survey. A questionnaire was considered to be complete if at least 80% of Sections A and B were answered. A questionnaire was considered to be partially complete if 50% to 79% of the questions were answered in Sections A and B. Household response rates were calculated using the American Association for Public Opinion Research response rate 2 (RR2) formula. The overall household response rate using the Next Birthday method was 39.97%.

Analyzing HINTS Data

If you are solely interested in calculating point estimates (means, proportions etc.), either weighted or unweighted, you can use programs including SAS, SPSS, STATA and Systat. If you plan on doing inferential statistical testing using the data (i.e., anything that involves calculating a p value or confidence interval), it is important that you utilize a statistical program that can incorporate the replicate weights that are included in the HINTS database. The issue is that the standard errors in your analyses will most likely be underestimated if you don't incorporate the jackknife replicate weights; therefore, your p-values will be smaller than they "should" be, your tests will be more liberal, and you are more likely to make a type I error. Statistical programs like SUDAAN, STATA, SAS and Wesvar can incorporate the replicate weights found in the HINTS database.

Note that analyses of HINTS variables that contain a large number of valid responses usually produce reliable estimates, but analyses of variables with a small number of valid responses may yield unreliable estimates, as indicated by their large variances. The analyst should pay particular attention to the standard error and coefficient of variation (relative standard error) for estimates of means, proportions, and totals, and the analyst should report these when writing up results. It is important that the analyst realizes that small sample sizes for particular analyses will tend to result in unstable estimates.

If you are using the SAS data file and wish to format it, you should do the following:

1. Save the hints4cycle2_09062017_public.sas7bdat and the hints4cycle2formats_09062017.sas7bdat data files in a folder on your computer.

2. Create a SAS library that refers to this folder.

3. Open and run all of the syntax in the hints4cycle2formats.sas file to link the formats to variables in the dataset. Note that there is a libname statement, and a proc format statement.

Important Analytic Variables in the Database

Note: Refer to the HINTS 4 Cycle 2 Methodology Report for more information regarding the weighting and stratification variables listed below.

PERSON_FINWT0: Final sample weight used to calculate population estimates. Note that estimates from the 2011 American Community Survey (ACS) of the US Census Bureau were used to calibrate the HINTS 4 Cycle 2 control totals with the following variables: Age, gender, education, marital status, race, ethnicity, and census region. In addition, variables from the 2011 National Health Interview Survey (NHIS) were used to calibrate HINTS 4 Cycle 2 data control totals regarding: Percent with health insurance and percent ever had cancer.

PERSON_FINWT1 THROUGH PERSON_FINWT50: Fifty replicate weights that can be used to calculate accurate standard error of estimates using the jackknife replication method. More information about how these weights were created can be found in the “HINTS 4 Cycle 2 Methodology Report” included in the data download, or see Korn and Graubard (1999).

STRATUM/CLUSTER VARIABLES FOR TAYLOR LINEARIZATION METHODS

VAR_STRATUM: This variable identifies the first-stage sampling stratum of a HINTS sample for a given data collection cycle. It is the variable assigned to the STRATA parameter when specifying the sample design to compute variances using the Taylor Series Linearization method. It has three values: Central Appalachia regardless of minority population (CA), high minority (HM), and low minority (LM).

VAR_CLUSTER: This variable identifies the cluster of sampling units of a HINTS sample for a given data collection cycle used for estimating variances. It is the variable assigned to the CLUSTER parameter when specifying the sample design to compute variances using the Taylor Series Linearization method. It has values ranging from 1 to 50.

OTHER VARIABLES

STRATUM: This variable codes for whether the respondent was in the Low or High Minority Area sampling stratum.

HIGHSPANLI: This variable codes for whether the respondent was in the High Spanish Linguistically Isolated stratum (Yes or No).

HISPSURNAME: This variable codes for whether there was a Hispanic surname match for this respondent (Yes or No).

TREATMENT_C2: This variable codes for the Spanish mailing protocol for the embedded experiment (see Methodology Report for more information). Each respondent was coded as living in a household where: 1) The household received both English and Spanish materials only if it was considered a Spanish surname or Linguistically Isolated household; or 2) The household was sent both English and Spanish questionnaires regardless of surname or being linguistically isolated.

FORM TYPE: This variable codes for the type of survey completed (Long or Short form).

LANGUAGE_FLAG: This variable codes for language the survey was completed in (English or Spanish).

QDISP: This variable codes for whether the survey returned by the respondent was considered Complete or Partial Complete. A complete questionnaire was defined as any questionnaire with at least 80% of the required questions answered in Sections A and B. A partial complete was defined as when between 50% and 79% of the questions were answered in Sections A and B. There were 48 partially complete questionnaires. The 55 questionnaires with fewer than 50% of the required questions answered in Sections A and B were coded as incompletely-filled out and discarded.

INCOMERANGES_IMP: This is the income variable (INCOMERANGES) imputed for missing data. To impute for missing items, PROC HOTDECK from the SUDAAN statistical software was used. PROC HOTDECK uses the Cox-Iannacchione Weighted Sequential Hot Deck imputation method as described by Cox (1980). The following variables were used as imputation classes given their strong association with the income variable: Education (O6), Race/Ethnicity (RaceEthn), Do you currently rent or own your house? (O15), How well do you speak English? (O9), and Were you born in the United States? (O7).

Variance Estimation Methods: Replicate vs. Taylor Linearization

Variance estimation procedures have been developed to account for complex sample designs. Taylor series (linear approximation) and replication (including jackknife and balanced repeated replication, BRR) are the most widely used approaches for variance estimation. Either of these techniques allow the analyst to appropriately reflect factors such as the selection of the sample, differential sampling rates to subsample a subpopulation, and nonresponse adjustments in estimating sampling error of survey statistics. Both procedures have good large sample statistical properties, and under most conditions, these procedures are statistically equivalent. Wolter (2007) is a useful reference on the theory and applications of these methods.

The HINTS 4, Cycle 2 datasets include variance codes and replicate weights so analysts can use either Taylor Series or replication methods for variance estimation. The following points may provide some guidance regarding which method will best reflect the HINTS sample design in your analysis.

TAYLOR SERIES	REPLICATION METHODS
• Most appropriate for simple statistics, such as means and proportions, since the approach linearizes the estimator of a statistic and then uses standard variance estimation methods.	• Useful for simple statistics such as means and proportions, as well as nonlinear functions. • Easy to use with a large number of variables. • Better accounts for variance reduction procedures such as raking and post-stratification. However, the variance reduction obtained with these procedures depends on the type of statistic and the correlation between the item of interest and the dimensions used in raking and post-stratification. Depending on your analysis, this may or may not be an advantage.

The Taylor Series variance estimation procedure is based on a mathematical approach that linearizes the estimator of a statistic using a Taylor Series expansion and then uses standard variance methods to estimate the variance of the linearized statistic.

The replication procedure, on the other hand, is based on a repeated sampling approach. The procedure uses estimators computed on subsets of the sample, where subsets are selected in a way that reflect the sample design. By providing weights for each subset of the sample, called replicate weights, end users can estimate the variance of a variety of estimators using standard weighted sums. The variability among the replicates is used to estimate the sampling variance of the point estimator.

An important advantage of replication is that it provides a simple way to account for adjustments made in weighting, particularly those with variance-reducing properties, such as weight calibration procedures. (See Kott, 2009, for a discussion of calibration methods, including raking, and their effects on variance estimation). The survey weights for HINTS were raked to control totals in the final step of the weighting process. However, the magnitude of the reduction generally depends on the type of estimate (i.e., total, proportion) and the correlation between the variable being analyzed and the dimensions used in raking.

Although SPSS's estimates of variance based on linearization take into account the sample design of the survey, they do not properly reflect the variance reduction due to raking. Thus, when comparing across Taylor series and replicate methods, analyses with Taylor series tend to have larger standard errors and generally provide more conservative tests of significance. The difference in the magnitude of standard errors between the two methods, however, will be smaller when using analysis variables that have little to no relationship with the raking variables.

Denominator Degrees of Freedom (DDF)

The HINTS 4 Cycle 2 database contains a set of 50 replicate weights to compute accurate standard errors for statistical testing procedures. These replicate weights were created using a jackknife minus one replication method; when analyzing one iteration of HINTS data, the proper denominator degrees of freedom (ddf) is 49. Thus, analysts who are only using the HINTS 4 Cycle 2 data should use 49 ddf in their statistical models. HINTS statistical analyses that involve more than one iteration of data will typically utilize a set of 50*k replicate weights, where they can be viewed as being created using a stratified jackknife method with k as the number of strata, and 49*k as the appropriate ddf. Analysts who were merging two iterations of data and making comparisons should adjust the ddf to be 98 (49*2) etc.

References

Cox, B. G. (1980). “The Weighted Sequential Hot Deck Imputation Procedure”. Proceedings of the American Statistical Association, Section on Survey Research Methods.

Finney Rutten, L. J., Davis, T., Beckjord, E. B., Blake, K., Moser, R. P., & Moser, R. P. (2012) Picking Up the Pace: Changes in Method and Frame for the Health Information National Trends Survey (2011 – 2014). Journal of Health Communication, 17 (8), 979-989..

Hesse, B. W., Moser, R. P., Rutten, L. J., & Kreps, G. L. (2006). The health information national trends survey: research from the baseline.*J Health Commun, 11 Suppl 1*, vii-xvi.

Korn, E. L., & Graubard, B. I. (1999). Analysis of health surveys. New York: John W iley & Sons.

Nelson, D. E., Kreps, G. L., Hesse, B. W., Croyle, R. T., Willis, G., Arora, N. K., et al. (2004). The Health Information National Trends Survey (HINTS): development, design, and dissemination. *J Health Commun, 9*(5), 443-460; discussion 481-444.

Appendix

The following appendices provide some coding examples using SAS, SUDAAN, and STATA for common types of statistical analyses using HINTS 4 Cycle 2 data. These examples will incorporate both the final sample weight (to get population estimates) and the set of 50 jackknife replicate weights to get the proper standard error, using the replication variance estimation method. The appendices also provide a coding example using SPSS, which incorporates the final sample weight and the variance codes for use with Taylor Series Linearization. Although these examples specifically use HINTS 4 Cycle 2 data, the concepts used here are generally applicable to other types of analyses. We will consider an analysis that includes gender, education level (edu) and two questions that are specific to the HINTS 4 data: seekcancerinfo & generalhealth.

- **Appendix A:** Analyzing data using SAS
- **Appendix B:** Analyzing data using SPSS
- **Appendix C:** Analyzing data using SUDAAN
- **Appendix D:** Analyzing data using STATA

Appendix A: Analyzing data using SAS

This section gives some SAS (Version 9.3 and higher) coding examples for common types of statistical analyses using HINTS 4 Cycle 2 data. We begin by doing data management of the HINTS 4 data in a SAS DATA step. We first decided to exclude all “Missing data (Not Ascertained)” and “Multiple responses selected in error” responses from the analyses. By setting these values to missing (.), SAS will exclude these responses from procedures where these variables are specifically accessed. For logistic regression modeling within the PROC SURVEYLOGISTIC procedure, SAS expects the response variable to be dichotomous with values (0, 1), so this variable will also be recoded at this point. It is better to use dummy variables instead of categorical variables in SAS survey procedures, such as PROC SURVEYREG. We use dummy variables for gender and education level in both PROC SURVEYLOGISTIC and PROC SURVEYREG procedures. W hen recoding existing variables, it is generally recommended to create new variables, rather than over-writing the existing variables. Note: New variables should always be compared to original source variables in a SAS PROC FREQ procedure to verify proper coding.

```
options fmtsearch=(hints4c2);  *This is used to call up the formats;
substitute your library name in the parentheses;

proc format;  *First create some temporary formats;

Value Genderf
1 = "Male"
2 = "Female";

Value Educationf
1 = "Less than high school"
2 = "12 years or completed high school"
3 = "Some college"
4 = "College graduate or higher";

value seekcancerinfof
1 = "Yes"
0 = "No";

Value Generalf
1 = "Excellent"
2 = "Very good"
3 = "Good"
4 = "Fair"
5 = "Poor";

run;


data hints4cycle2;
set hints4c2.hints4cycle2_09062017_public;

/*Recode negative values to missing*/
if genderc = 1 then gender = 1;
if genderc = 2 then gender = 2;
```

```
if genderc in (-9, -6) then gender = .;

/*Recode education into four levels, and negative values to missing*/
if education in (1, 2) then edu = 1;
if education = 3 then edu = 2;
if education in (4, 5) then edu = 3;
if education in (6, 7) then edu = 4;
if education = -9 then edu = .;

/*Recode seekcancerinfo to 0-1 format for proc rlogist procedure, and
negative values to missing */
if seekcancerinfo = 2 then seekcancerinfo = 0;
if seekcancerinfo in (-9, -6, -2, -1) then seekcancerinfo = .;

/*Recode negative values to missing for proc regress procedure*/
if generalhealth in (-5, -9) then generalhealth = .;


/*Create dummy variables for proc surveylogistic and proc surveyreg
procedures*/
if gender = 1 then
      Female = 0;
else if gender = 2 then
      Female = 1;

if edu = 1 then
      do;
            HighSchool = 0;
            SomeCollege = 0;
            CollegeorMore = 0;
      end;
else if edu = 2 then
      do;
            HighSchool = 1;
            SomeCollege = 0;
            CollegeorMore = 0;
      end;
else if edu = 3 then
      do;
            HighSchool = 0;
            SomeCollege = 1;
            CollegeorMore = 0;
      end;
else if edu = 4 then
      do;
            HighSchool = 0;
            SomeCollege = 0;
            CollegeorMore = 1;
      end;

/*Apply formats to recoded variables */
format gender genderf. edu educationf. seekcancerinfo seekcancerinfof.
generalhealth generalf.;

run;
```

Proc Surveyfreq procedure

We are now ready to begin using SAS 9.3 to examine the relationships among these variables. Using **PROC SURVEYFREQ**, we will first generate a cross-frequency table of education by gender, along with a (Wald) Chi-squared test of independence. Note the syntax of the overall sample weight, PERSON_FINW T0, and those of the jackknife replicate weights, PERSON_FINWT1—PERSON_FINW T50. The jackknife adjustment factor for each replicate weight is 0.98. This syntax is consistent for all procedures. Other data sets that incorporate replicate weight jackknife designs will follow a similar syntax.

```
proc surveyfreq data = hints4cycle2 varmethod =jackknife;
      weight person_finwt0;
      repweights person_finwt1-person_finwt50 / df = 49 jkcoefs =0.98;
      tables edu*gender / row col wchisq;
run;
```

The *tables* statement defines the frequencies that should be generated. Stand-alone variables listed here result in one-way frequencies, while a "*" between variables will define cross-frequencies. The *row* option produces row percentages and standard errors, allowing us to view stratified percentages. Similarly, the *col* option produces column percentages and standard errors, allowing us to view stratified percentages. The option *wchisq* requests Wald chi-square test for independence. Other tests and statistics are also available; see the SAS 9.3 Product Documentation Site for more information.

For the purposes of computing appropriate degrees of freedom for the estimator of the HINTS4-Cycle 2 differences, we can assume, as an approximation, that the sample is a simple random sample of size 50 (corresponding to the 50 replicates: each replicate provides a 'pseudo sample unit') from a normal distribution. The denominator degrees of freedom (df) is equal to 49*k, where k is the number of iterations of data used in this analysis.

Variance Estimation	
Method	Jackknife
Replicate Weights	H4C2
Number of Replicates	50

Table Education by Gender

edu	gender	Frequency	Percent	Std Err of Percent	Row Percent	Std Err of Row Percent	Column Percent	Std Err of Col Percent
Less than high school	Male	139	6.97	0.47	52.43	1.51	14.28	0.96
	Female	185	6.32	0.28	47.57	1.51	12.33	0.55
	Total	324	13.28	0.66	100.00			
12 years or completed high school	Male	280	10.10	0.68	49.75	2.08	20.70	1.39
	Female	478	10.20	0.58	50.25	2.08	19.90	1.13
	Total	758	20.29	0.95	100.00			
Some college	Male	403	17.75	0.82	47.00	1.39	36.39	1.67
	Female	641	20.01	0.68	53.00	1.39	39.06	1.27
	Total	1044	37.76	1.10	100.00			
College graduate or	Male	547	13.96	0.52	48.69	1.08	28.62	1.08
	Female	816	14.71	0.42	51.31	1.08	28.70	0.84

higher	Total	1363	28.66	0.72	100.00			
Total	Male	1369	48.76	0.25			100.00	
	Female	2120	51.24	0.25			100.00	
	Total	3489	100.00					

Frequency Missing = 141

Wald Chi-Square Test	
Chi-Square	6.4479
F Value	2.1493
Num DF	3
Den DF	49
Pr > F	0.106
Adj F Value	2.0616
Num DF	3
Den DF	47
Pr > Adj F	0.1181

Sample Size = 3489

The weighted percentages above show that a greater proportion of women have at least a college degree compared to men, 14.71% vs. 13.96%. However, the Chi-squared test of independence indicates that there is no significant difference between these two proportions (p-value < 0.05).

Logistic Regression

This example demonstrates a multivariable logistic regression model using **PROC SURVEYLOGISTIC**; recall that the response should be a dichotomous 0-1 variable.

```
/*Multivariable logistic regression of gender and education on
SeekCancerInfo*/
proc surveylogistic data= hints4cycle2 varmethod=jackknife;
      weight person_finwt0;
      repweights person_finwt1-person_finwt50 / df=49 jkcoefs=0.98;
      model seekcancerinfo (descending) = Female HighSchool SomeCollege
CollegeorMore / tech=newton xconv=1e-8;
      contrast 'Overall model' intercept 1,
                              Female 1,
                              HighSchool 1,
                              SomeCollege 1,
                              CollegeorMore 1;
      contrast 'Overall model minus intercept' Female 1,
                              HighSchool 1,
                              SomeCollege 1,
                              CollegeorMore 1;
      contrast 'Gender' Female 1;
      contrast 'Education overall' HighSchool 1,
                              SomeCollege 1,
                              CollegeorMore 1;
run;
```

The response variable should be on the left hand side (LHS) of the equal sign in the model statement, while all covariates should be listed on the right hand side (RHS). The *descending* option requests the probability of seekcancerinfo="Yes" to be modeled. The "Male" is the reference group for gender effect while "Less than high school" is the reference group for education level effect. The option *tech=newton* requests the Newton-Raphson algorithm. The option xconv=1e-8 helps to avoid early termination of the iteration.

Variance Estimation	
Method	Jackknife
Replicate Weights	H4C2
Number of Replicates	50

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.7285	0.1965	13.7451	0.0002
Female	1	0.3626	0.1195	9.205	0.0024
HighSchool	1	0.1184	0.2208	0.2875	0.5918
SomeCollege	1	0.4149	0.1929	4.6261	0.0315
CollegeorMore	1	0.7414	0.1998	13.77	0.0002

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Female	1.437	1.137	1.816
HighSchool	1.126	0.73	1.735
SomeCollege	1.514	1.038	2.21
CollegeorMore	2.099	1.419	3.105

Contrast Test Results

Contrast	DF	Wald Chi-Square	Pr > ChiSq
Overall model	5	25.1347	0.0001
Overall model minus intercept	4	24.9693	<.0001
Gender	1	9.205	0.0024
Education overall	3	19.2597	0.0002

To identify levels/variables that display a significant difference in response, the rule of thumb is to examine odds ratios where the confidence interval does not contain 1 (by default, SAS will use alpha=.05 to determine statistical significance; this value can be changed by the user using code). However, significance may also be garnered from the test of whether the associated beta parameter is equal to 0 (see first regression table above). According to this model, women and college students appear to be statistically more inclined to search for cancer information (compared with men and those who did not graduate from high school, respectively).

Linear Regression

This example demonstrates a multivariable linear regression model using **PROC SURVEYREG**; recall that the response should be a continuous variable. For the purposes of this example, we decided to use an outcome with five levels as a continuous variable (GENERALHEALTH). Note that higher values on GENERALHEALTH indicate poorer self-reported health status.

```
/*Multivariable linear regression of gender and education on GeneralHealth*/
proc surveyreg data= hints4cycle2 varmethod=jackknife;
      weight person_finwt0;
      repweights person_finwt1-person_finwt50 / df=49 jkcoefs=0.98;
      model generalhealth = Female HighSchool SomeCollege CollegeorMore;
      contrast 'Overall model' intercept 1,
                               Female 1,
                               HighSchool 1,
                               SomeCollege 1,
                               CollegeorMore 1;
      contrast 'Overall model minus intercept' Female 1,
                               HighSchool 1,
                               SomeCollege 1,
                               CollegeorMore 1;
      contrast 'Gender' Female 1;
      contrast 'Education overall' HighSchool 1,
                               SomeCollege 1,
                               CollegeorMore 1;
run;
```

Variance Estimation	
Method	Jackknife
Replicate Weights	H4C2
Number of Replicates	50

Analysis of Contrasts

Contrast	Num DF	F Value	Pr > F
Overall model	5	2294	<.0001
Overall model minus intercept	4	51.54	<.0001
Gender	1	0.35	0.5558
Education overall	3	54.2	<.0001

NOTE: The denominator degrees of freedom for the F tests is 49.

From the above table, we can see that Gender is not associated with general health, but Edu is associated, adjusting for all variables in the model.

Estimated Regression of Coefficients

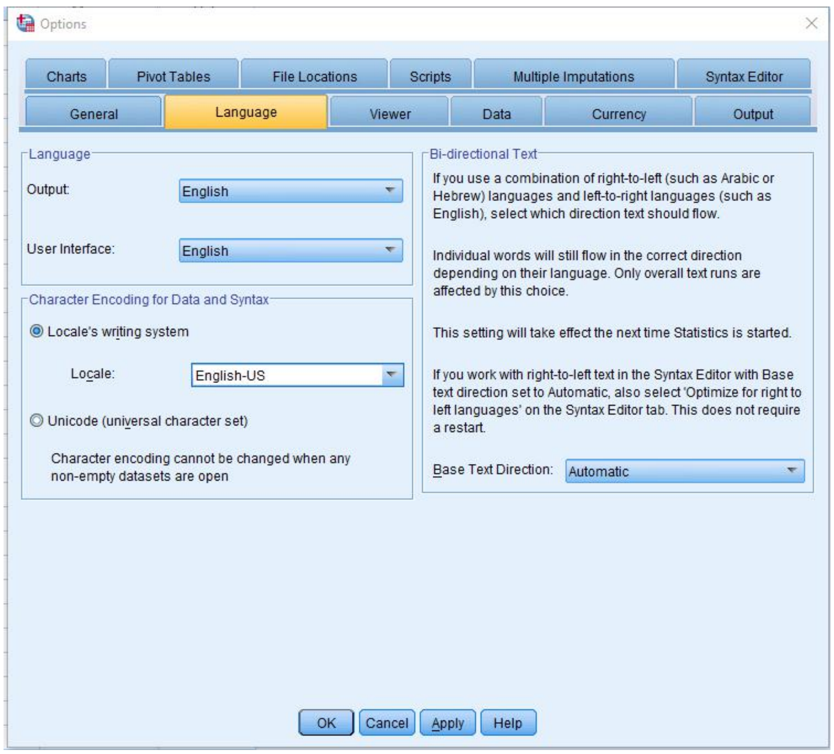
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	3.1883428	0.0911525	34.98	<.0001
Female	0.0325468	0.05487372	0.59	0.5558
HighSchool	-0.5355744	0.11310305	-4.74	<.0001
SomeCollege	-0.6466411	0.10123915	-6.39	<.0001
CollegeorMore	-0.9360366	0.08140692	-11.5	<.0001

NOTE: The denominator degrees of freedom for the t-tests is 49.

From the above table, it can be seen that, compared to those respondents with Less than a High School education, those with Some College have a significantly negative linear association with general health (i.e., better reported health), controlling for all variables in the model. This association also applies to those with a College Degree or Higher. We don’t interpret the Gender variable because it is non-significant.

Appendix B: Analyzing data using SPSS

Prior to opening the HINTS 4, Cycle 2 SPSS data, it is important to ensure that your SPSS environment is set up to be compatible with the dataset. Specifically, the language encoding (i.e., the way that character data are stored and accessed) must match between your environment and the dataset. We recommend locale encoding in U.S. English over Unicode encoding. To ensure compatibility, you must update the language encoding manually through the graphic user interface (GUI). In a new SPSS session, from the empty dataset window, select “Edit” > “Options…” from the menu bar. In the pop-up box, select the “Language” tab. In this tab, look for the “Character Encoding for Data and Syntax” section. Select the “Locale’s writing system” option and English-US or en-US from the “Locale:” dropdown list. “English-US” and “en-US” from the drop down are the common aliases used by SPSS to describe U.S. English encoding; if you do not see these specific aliases verbatim, choose the English alias that is most similar. Click “OK” to save your changes. You may now open the HINTS SPSS data without compatibility issues.



This section gives some SPSS (Version 25 and higher) coding examples for common types of statistical analyses using HINTS 4 Cycle 2 data. These examples will incorporate the stratum variable, VAR_STRATUM, and the cluster variable VAR_CLUSTER. Although these examples specifically use HINTS 4, Cycle 2 data, the concepts used here are generally applicable to other types of analyses. We will consider an analysis that includes gender, education level (edu) and two questions that are specific to the HINTS 4, Cycle 2 data: seekcancerinfo & generalhealth.

We begin by creating an analysis plan using the Complex Samples analysis procedures to specify the sample design; PERSON_FINWT0 is the sample weight variable (the final weight for the composite sample, no group differences found), VAR_STRATUM is the stratum variable, and VAR_CLUSTER is the cluster variable. The subcommand SRSESTIMATOR specifies the variance estimator under the simple random sampling assumption. The default value is WR (with replacement), and it includes the finite population correction in the variance computation. The subcommand PRINT is used to control output from CSPLAN, and the syntax PLAN means to display a summary of plan specifications. The subcommand DESIGN with keyword STRATA identifies the sampling stratification variable, and the keyword cluster CLUSTER identifies the grouping of sampling units for variance estimation. The subcommand ESTIMATOR specifies the variance estimation method used in the analysis. The syntax TYPE=WR requires the estimation method of selection with replacement.

```
* Analysis Preparation Wizard.
*substitute your library name in the parentheses of /PLAN FILE=.
CSPLAN ANALYSIS
 /PLAN FILE='(sample.csaplan)'
 /PLANVARS ANALYSISWEIGHT=PERSON_FINWT0
 /SRSESTIMATOR TYPE=WOR
 /PRINT PLAN
 /DESIGN STRATA=VAR_STRATUM CLUSTER=VAR_CLUSTER
 /ESTIMATOR TYPE=WR.
```

We completed data management of the HINTS 4 Cycle 2 data in a SPSS RECODE step. We first decided to exclude all “Missing data (Not Ascertained)” and “Multiple responses selected in error” responses from the analyses. By setting these values to missing (SYSMIS), SPSS will exclude these responses from procedures where these variables are specifically accessed. For logistic regression modeling in the CSLOGISTIC procedure, SPSS by default always uses the last (highest) level of category of the covariates as the reference, similar to SAS. Users in SPSS cannot define the reference category by themselves unless they reorder the categories to create the desired value as the reference, such as using reverse coding (see example below). To make SPSS results comparable with SAS, we reverse coded the variables in SPSS. When recoding existing variables, it is generally recommended to create new variables, rather than over-writing the existing variables. Note: New variables should always be compared to original source variables in a SPSS CROSSTABS procedure to verify proper coding.

```
RECODE GenderC (1=1) (2=2) (ELSE=SYSMIS) INTO gender.
VARIABLE LABELS gender 'gender'.
EXECUTE.

*Recode education into four levels, and negative values to missing.
RECODE Education (3=2) (1 thru 2=1) (4 thru 5=3) (6 thru 7=4) (ELSE=SYSMIS) INTO edu.
VARIABLE LABELS edu 'edu'.
EXECUTE.

*Recode seekcancerinfo to 0- 1 format for CSLOGISTIC procedure, and negative values to missing.
RECODE SeekCancerInfo (2=0) (1=1) (ELSE=SYSMIS) INTO seekcancerinfo_recode.
VARIABLE LABELS seekcancerinfo_recode 'seekcancerinfo_recode'.
EXECUTE.

*Recode negative values to missing for CSGLM procedure.
RECODE GeneralHealth (1 thru 5=Copy) (ELSE=SYSMIS) INTO genhealth_recode.
VARIABLE LABELS genhealth_recode 'genhealth_recode'.
EXECUTE.

*Reverse coding.
RECODE gender (1=2) (2=1) (ELSE=Copy) INTO flippedgender.
VARIABLE LABELS flippedgender 'flippedgender'.
EXECUTE.

*Reverse coding.
RECODE edu (1=4) (2=3) (3=2) (4=1) (ELSE=Copy) INTO flippededu.
VARIABLE LABELS flippededu 'flippededu'.
```

```
EXECUTE.

*Add value labels to recoded variables.
VALUE LABELS gender 1 "Male" 2 "Female".
VALUE LABELS flippedgender 2 "Male" 1 "Female".
VALUE LABELS edu 1 "Less than high school" 2 "12 years or completed high school" 3 "Some college" 4
"College graduate or higher".
VALUE LABELS flippededu 4 "Less than high school" 3 "12 years or completed high school" 2 "Some
college" 1 "College graduate or higher".
VALUE LABELS seekcancerinfo_recode 1 "Yes" 0 "No".
VALUE LABELS genhealth_recode 1 "Excellent" 2 "Very good" 3 "Good" 4 "Fair" 5 "Poor".
```

Frequency Table and Chi-Square Test

We are now ready to begin using SPSS v25 to examine the relationships among these variables. Using **CSTABULATE**, we will first generate a cross-frequency table of education by gender. Note that we specify the file that contains the sample design specification using the subcommand PLAN. This syntax is consistent for all procedures. Other analyses using the same sample design will follow a similar syntax.

```
* Complex Samples Crosstabs.
CSTABULATE
/PLAN FILE=”(plan filename)”
/TABLES VARIABLES=edu BY gender
/CELLS POPSIZE ROWPCT COLPCT TABLEPCT
/STATISTICS SE COUNT
/TEST INDEPENDENCE
/MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

The TABLES subcommand defines the tabulation variables, where the syntax “BY” indicates the two-way crosstabulation. The CELLS subcommand specifies the summary value estimates to be displayed in the table. The *POPSIZE* option produces population size estimates for each cell and marginal. The *ROWPCT* option produces row percentages and standard errors. Similarly, the *COLPCT* option produces column percentages and standard errors. The *TABLEPCT* option produces table percentages and standard errors for each cell. The STATISTICS subcommand specifies the statistics to be displayed with the summary value estimates. The *SE* option produces the standard error for each summary value, and the *COUNT* option produces unweighted counts. The TEST subcommand specifies tests for the table. The *INDEPENDENCE* option produces the test of independence for the two-way crosstabulations. The MISSING subcommand specifies how missing values are handled. The *SCOPE* statement specifies which cases are used in the analyses. The *TABLE* option specifies that cases with all valid data for the tabulation variables are used in the analyses. The *CLASSMISSING* statement specifies whether user-defined missing values are included or excluded. The *EXCLUDE* option specifies user-defined missing values to be excluded in the analysis.

edu * gender

			Gender		
Education			Male	Female	Total
Less than high school	Population Size	Estimate	16001011.395	14516946.599	30517957.994
		Standard Error	1946630.239	1351449.050	2405049.762
		Unweighted Count	139	185	324
	% within edu	Estimate	52.4%	47.6%	100.0%
		Standard Error	3.8%	3.8%	0.0%
		Unweighted Count	139	185	324

	% within gender	Estimate	14.3%	12.3%	13.3%
		Standard Error	1.6%	1.1%	1.0%
		Unweighted Count	139	185	324
	% of Total	Estimate	7.0%	6.3%	13.3%
		Standard Error	0.8%	0.6%	1.0%
		Unweighted Count	139	185	324
12 years or completed high school	Population Size	Estimate	23192925.636	23429412.615	46622338.251
		Standard Error	2666726.765	1690802.790	3087092.015
		Unweighted Count	280	478	758
	% within edu	Estimate	49.7%	50.3%	100.0%
		Standard Error	3.5%	3.5%	0.0%
		Unweighted Count	280	478	758
	% within gender	Estimate	20.7%	19.9%	20.3%
		Standard Error	1.9%	1.3%	1.2%
		Unweighted Count	280	478	758
	% of Total	Estimate	10.1%	10.2%	20.3%
		Standard Error	1.1%	0.7%	1.2%
		Unweighted Count	280	478	758
Some college	Population Size	Estimate	40767268.647	45973484.121	86740752.769
		Standard Error	3125924.965	2554649.194	3935430.157
		Unweighted Count	403	641	1044
	% within edu	Estimate	47.0%	53.0%	100.0%
		Standard Error	2.4%	2.4%	0.0%
		Unweighted Count	403	641	1044
	% within gender	Estimate	36.4%	39.1%	37.8%
		Standard Error	2.4%	1.4%	1.4%
		Unweighted Count	403	641	1044
	% of Total	Estimate	17.7%	20.0%	37.8%
		Standard Error	1.2%	1.1%	1.4%
		Unweighted Count	403	641	1044
College graduate or higher	Population Size	Estimate	32062482.132	33787469.188	65849951.320
		Standard Error	1911627.710	1519028.788	2477869.482
		Unweighted Count	547	816	1363
	% within edu	Estimate	48.7%	51.3%	100.0%
		Standard Error	1.8%	1.8%	0.0%
		Unweighted Count	547	816	1363
	% within gender	Estimate	28.6%	28.7%	28.7%
		Standard Error	1.5%	1.2%	1.0%
		Unweighted Count	547	816	1363

	% of Total	Estimate	14.0%	14.7%	28.7%
		Standard Error	0.8%	0.7%	1.0%
		Unweighted Count	547	816	1363
Total	Population Size	Estimate	112023687.810	117707312.523	229731000.334
		Standard Error	5180924.693	3707931.492	5778262.328
		Unweighted Count	1369	2120	3489
	% within edu	Estimate	48.8%	51.2%	100.0%
		Standard Error	1.5%	1.5%	0.0%
		Unweighted Count	1369	2120	3489
	% within gender	Estimate	100.0%	100.0%	100.0%
		Standard Error	0.0%	0.0%	0.0%
		Unweighted Count	1369	2120	3489
	% of Total	Estimate	48.8%	51.2%	100.0%
		Standard Error	1.5%	1.5%	0.0%
		Unweighted Count	1369	2120	3489

Tests of Independence

		Chi-Square	Adjusted F	df1	df2	Sig.
edu * gender	Pearson	4.413	.603	2.760	369.799	.600
	Likelihood Ratio	4.414	.603	2.760	369.799	.600

The adjusted F is a variant of the second-order Rao-Scott adjusted chi-square statistic. Significance is based on the adjusted F and its degrees of freedom.

The weighted percentages above show that a greater proportion of women have at least a college degree compared to men, 14.7% vs 14.0%. The Chi-squared test of independence indicates that there is not a significant difference between the educational distribution in these two groups (p-value > .05).

Note that the CSTABULATE procedure provides results for the Pearson Chi-square and Likelihood Ratio tests, but not for the Wald Chi-square test of independence. To get the results for the Wald Chi-square test of independence, users can conduct a logistic regression model in the CSLOGISTIC procedure in which the type of Chi-square test can be specified.

Logistic Regression

This example demonstrates a multivariable logistic regression model using **CSLOGISTIC**; recall that the response should be a categorical variable.

```
*Multivariable logistic regression of gender and education on SeekCancerInfo.
CSLOGISTIC  seekcancerinfo_recode (LOW) BY flippedgender flippededu
 /PLAN FILE='(sample.csaplan)'
 /MODEL flippedgender flippededu
 /CUSTOM  Label = 'Overall model minus intercept'
  LMATRIX = flippedgender 1/2 -1/2;
          flippededu 1/3 1/3 1/3 -1;
```

```
        flippededu 1/3 1/3 -1 1/3 ;
        flippededu 1/3 -1 1/3 1/3;
        flippededu -1 1/3 1/3 1/3
 /CUSTOM  Label = 'Gender'
 LMATRIX =  flippedgender 1/2 -1/2
 /CUSTOM  Label = 'Education overall'
 LMATRIX = flippededu 1/3 1/3 1/3 -1;
        flippededu 1/3 1/3 -1 1/3 ;
        flippededu 1/3 -1 1/3 1/3;
        flippededu -1 1/3 1/3 1/3
 /INTERCEPT INCLUDE=YES SHOW=YES
 /STATISTICS PARAMETER SE CINTERVAL TTEST EXP
 /TEST TYPE=CHISQUARE PADJUST=LSD
 /ODDSRATIOS FACTOR=[flippedgender(HIGH)]
 /ODDSRATIOS FACTOR=[flippededu(HIGH)]
 /MISSING CLASSMISSING=EXCLUDE
 /CRITERIA MXITER=100 MXSTEP=50 PCONVERGE=[1e-008 RELATIVE] LCONVERGE=[0] CHKSEP=20
CILEVEL=95
 /PRINT SUMMARY COVB CORB VARIABLEINFO SAMPLEINFO.
```

The response variable should be on the left-hand side of the BY statement, while all covariates should be listed on the right-hand side. The (LOW) option indicates that the lowest category is the reference category, thus requests the probability of seekcancerinfo = “Yes” to be modeled. The “Male” is the reference group for gender effect, while “Less than high school” is the reference group for education level effect. The subcommand MODEL specifies all variables in the model. The CUSTOM subcommand allows users to define custom hypothesis tests. The LMATRIX statement specifies coefficients of contrasts, which are used for studying the effects in the model. The INTERCEPT subcommand specifies whether to include or show the intercept in the final estimates. The STATISTICS subcommand specifies the statistics to be estimated and shown in the final result, where the syntax PARAMETER indicates the coefficient estimates, EXP indicates the exponentiated coefficient estimates, SE indicates the standard error for each coefficient estimate, CINTERVAL indicates the confidence interval for each coefficient estimate. The TEST subcommand specifies the type of test statistic and the method of adjusting the significance level to be used for hypothesis tests that are requested on the MODEL and CUSTOM subcommands, where the syntax CHISQUARE indicates the Wald chi-square test, and LSD indicates the least significant difference. The ODDSRATIOS subcommand estimates odds ratios for certain factors. The subcommand MISSING specifies how to handle missing data. The subcommand CRITERIA offers controls on the iterative algorithm that is used for estimations. The option PCONVERGE= [1e-008 RELATIVE] helps to avoid early termination of the iteration. The subcommand PRINT is used to display optional output.

Sample Design Information

		N
Unweighted Cases	Valid	2879
	Invalid	751
	Total	3630
Population Size		187658584.422
Stage 1	Strata	3
	Units	131
Sampling Design Degrees of Freedom		128

Parameter Estimates

seekcancerinfo_recode	Parameter	B	Std. Error	95% Confidence Interval Lower	95% Confidence Interval Upper	Hypothesis Test t	Hypothesis Test df	Hypothesis Test Sig.	Exp(B)	95% Confidence Interval for Exp(B) Lower	95% Confidence Interval for Exp(B) Upper
Yes	(Intercept)	-.728	.206	-1.136	-.321	-3.537	128.000	.001	.483	.321	.725
	Female	.363	.121	.123	.602	2.992	128.000	.003	1.437	1.131	1.827
	College Graduate or Higher	.741	.202	.342	1.140	3.677	128.000	.000	2.099	1.408	3.128
	Some College	.415	.212	-.004	.834	1.959	128.000	.052	1.514	.996	2.302
	12 Years of Completed High School	.118	.244	-.364	.600	.486	128.000	.628	1.126	.695	1.823

Dependent Variable: seekcancerinfo_recode (reference category = No)

Model: (Intercept), flippedgender, flippededu

a. Set to zero because this parameter is redundant.

Odds Ratios

		Odds Ratio	95% Confidence Interval Lower	95% Confidence Interval Upper
flippedgender	Female vs. Male	1.437	1.131	1.827
flippededu	College graduate or higher vs. Less than high school	2.099	1.408	3.128
	Some college vs. Less than high school	1.514	.996	2.302
	12 years or completed high school vs. Less than high school	1.126	.695	1.823

Overall Model Minus Intercept

df	Wald Chi-Square	Sig.
4.000	26.508	.000

Gender

df	Wald Chi-Square	Sig.
1.000	8.951	.003

Education Overall

df	Wald Chi-Square	Sig.
3.000	22.205	.000

To identify levels/variables that display a significant difference in response, the rule of thumb is to examine odds ratios where the confidence interval does not contain 1 (by default, SPSS will use alpha=.05 to determine statistical significance; this value can be changed by the user using code). However, significance may also be garnered from the test of whether the associated beta parameter is equal to 0 (see "Parameter Estimates" table above). According to this model, women and those with at least a high school degree appear to be statistically more inclined to search for cancer information (compared with men and those who did not graduate from high school, respectively).

Note that in SPSS we cannot get the overall model effect, even if we used the CUSTOM subcommand to conduct custom hypothesis tests.

Linear Regression

This example demonstrates a multivariable linear regression model using **CSGLM**; recall that the response should be a continuous variable. For the purposes of this example, we decided to use an outcome with five levels as a continuous variable (GENERALHEALTH). Note that higher values on GENERALHEALTH indicate poorer self-reported health status.

```
* Multivariable linear regression of gender and education on GeneralHealth.
CSGLM genhealth_recode BY flippedgender flippededu
 /PLAN FILE='(sample.csaplan)'
 /MODEL flippededu flippedgender
 /CUSTOM  Label = 'Overall model minus intercept'
  LMATRIX = flippedgender 1/2 -1/2;
           flippededu 1/3 1/3 1/3 -1;
           flippededu 1/3 1/3 -1 1/3 ;
           flippededu 1/3 -1 1/3 1/3;
           flippededu -1 1/3 1/3 1/3
  /CUSTOM  Label = 'Gender'
 LMATRIX =  flippedgender 1/2 -1/2
  /CUSTOM  Label = 'Education overall'
  LMATRIX =  flippededu 1/3 1/3 1/3 -1;
             flippededu 1/3 1/3 -1 1/3 ;
             flippededu 1/3 -1 1/3 1/3;
             flippededu -1 1/3 1/3 1/3
 /INTERCEPT INCLUDE=YES SHOW=YES
 /STATISTICS PARAMETER SE CINTERVAL TTEST
 /PRINT SUMMARY VARIABLEINFO SAMPLEINFO
 /TEST TYPE=F PADJUST=LSD
 /MISSING CLASSMISSING=EXCLUDE
 /CRITERIA CILEVEL=95.
```

Sample Design Information

		N
Unweighted Cases	Valid	3401
	Invalid	229

	Total	3630
Population Size		223433147.369
Stage 1	Strata	3
	Units	135
Sampling Design Degrees of Freedom		132

Parameter Estimates[a]

Parameter	Estimate	Std. Error	95% Confidence Interval Lower	95% Confidence Interval Upper	Hypothesis Test t	Hypothesis Test df	Hypothesis Test Sig.
(Intercept)	3.188	.087	3.016	3.361	36.569	132.000	.000
College Graduate or Higher	-.936	.077	-1.088	-.784	-12.155	132.000	.000
Some College	-.647	.096	-.836	-.458	-6.767	132.000	.000
12 Years or Completed High School	-.536	.113	-.759	-.313	-4.751	132.000	.000
Female	.033	.055	-.076	.142	.591	132.000	.556

a. Model: genhealth_recode = (Intercept) + flippededu + flippedgender

b. Set to zero because this parameter is redundant.

Compared to those respondents with less than a high school education, those who completed 12 years of school or completed high school on average reported significantly better general health (i.e., the negative beta coefficient indicates that the average health score is lower among those with some college, and the health variable is coded such that lower scores correspond to better health), controlling for all variables in the model. This association also applies to those who have completed some college and those with a college degree or higher. We do not interpret the estimates for the Gender variable because the corresponding p-value is greater than .05.

Overall Model Minus Intercept

df1	df2	Wald F	Sig.
4.000	129.000	60.343	.000

Gender

df1	df2	Wald F	Sig.
1.000	132.000	.349	.556

Education

df1	df2	Wald F	Sig.
3.000	130.000	64.144	.000

From the above table, we can see that education, but not gender, is significantly associated with general health.

Appendix C: Analyzing data using SUDAAN

This section gives some SUDAAN (Version 10.0.1 and higher) coding examples for common types of statistical analyses using HINTS 4 Cycle 2 data. We begin by doing data management of the HINTS 4 data in a SAS DATA step. We first decided to exclude all “Missing data (Not Ascertained)” and “Multiple responses selected in error” responses from the analyses. By setting these values to missing (.), SAS will exclude these responses from procedures where these variables are specifically accessed. For logistic regression modeling within the PROC RLOGIST procedure, SUDAAN expects the response variable to be dichotomous with values (0, 1), so this variable will also be recoded at this point. When recoding existing variables, it is generally recommended to create new variables of rather than over-writing the existing variables. Note: New variables should always be compared to original source variables in a SAS PROC FREQ procedure to verify proper coding.

```
proc format;  *First create some temporary formats;

Value Genderf
1 = "Male"
2 = "Female";

Value Educationf
1 = "Less than high school"
2 = "12 years or completed high school"
3 = "Some college"
4 = "College graduate or higher";

value seekcancerinfof
1 = "Yes"
0 = "No";

Value General
1 = "Excellent"
2 = "Very good"
3 = "Good"
4 = "Fair"
5 = "Poor";

run;


data hints4cycle2;
set hints4c2.hints4cycle2_09062017_public;

/*Recode negative values to missing*/
if genderc = 1 then gender = 1;
if genderc = 2 then gender = 2;
if genderc in (-9, -6) then gender = .;

/*Recode education into four levels, and negative values to missing*/
if education in (1, 2) then edu = 1;
if education = 3 then edu = 2;
if education in (4, 5) then edu = 3;
if education in (6, 7) then edu = 4;
if  education  =  -9  then  edu  =  .;
```

```
/*Recode seekcancerinfo to 0-1 format for proc rlogist procedure, and
negative values to missing */
if seekcancerinfo = 2 then seekcancerinfo = 0;
if seekcancerinfo in (-9, -6, -2, -1) then seekcancerinfo = .;

/*Recode negative values to missing for proc regress procedure*/
if generalhealth in (-5, -9) then generalhealth = .;

/*Apply formats to recoded variables */
format gender genderf. edu educationf. seekcancerinfo seekcancerinfof.
generalhealth general.;

run;
```

We are now ready to begin using SUDAAN to examine the relationships among these variables. Using **proc crosstab**, we will first generate a cross-frequency table of education and gender, along with a (Wald) Chi-squared test of independence. Note the syntax of the overall sample weight, PERSON_FINW T0, and those of the jackknife replicate weights, PERSON_FINWT1—PERSONFINW T50. The jackknife adjustment factor for each replicate weight is 0.98. This syntax is consistent for all procedures. Other data sets that incorporate replicate weight jackknife designs will follow a similar syntax.

```
proc crosstab data= hints4cycle2 design=jackknife ddf =49;
weight person_finwt0;
jackwgts person_finwt1-person_finwt50 / adjjack=.98;
class gender edu;
tables edu*gender;
test chisq;
run;
```

Since this procedure is mainly for categorical variables, each variable should be specified as such by inclusion in the class statement (which is ubiquitous in all SUDAAN procedures). The *tables* statement defines the frequencies that should be generated. Stand-alone variables listed here result in one-way frequencies, while a “*” between variables will define cross-frequencies. In general, the PROC CROSSTAB procedure may be used to investigate n-way variable frequencies, along with their relationships. This is accomplished by the *test* statement, which defines various types of independence tests: here a Chi-Squared test is implemented. Other tests and statistics are also available; see the SUDAAN site link for more information.

The HINTS 4 database for a single iteration contains a set of 50 replicate weights to compute accurate standard errors for statistical testing procedures. These replicate weights were created using a jackknife minus one replication method. Thus, the proper denominator degrees of freedom (ddf) should be 49 when one iteration of HINTS data is being analyzed. Thus, analysts who are only using the HINTS 4 Cycle 2 data should use 49 ddf in their statistical models.

HINTS 4 databases with more than one iteration of data will contain a set of 50*k replicate weights, where they can be viewed as being created using a stratified jackknife method with k as the number of strata and 49*k as the appropriate ddf. Analysts who were merging two iterations of data and making comparisons these should adjust the ddf to be 98 (49*2) etc.

Variance Estimation Method: Replicate Weight Jackknife
By: EDU, GENDER

Are you male or female?

What is the highest grade or level of schooling you completed?		Total	Male	Female
Total	**Sample Size**	3489	1369	2120
	Col Percent	100.00%	100.00%	100.00%
	Row Percent	100.00%	48.76%	51.24%
Less than HS	**Sample Size**	324	139	185
	Col Percent	13.28%	14.28%	12.33%
	Row Percent	100.00%	52.43%	47.57%
12 years or completed HS	**Sample Size**	758	280	478
	Col Percent	20.29%	20.70%	19.90%
	Row Percent	100.00%	49.75%	50.25%
Some college	**Sample Size**	1044	403	641
	Col Percent	37.76%	36.39%	39.06%
	Row Percent	100.00%	47.00%	53.00%
College graduate or higher	**Sample Size**	1363	547	816
	Col Percent	28.66%	28.62%	28.70%
	Row Percent	100.00%	48.69%	51.31%

Variance Estimation Method: Replicate Weight Jackknife
Chi Square Test of Independence for EDU and GENDER

ChiSq	2.15
P-value for ChiSq	0.106
Degress of Freedom ChiSq	3

Logistic Regression

This example demonstrates a multivariable logistic regression model using **PROC RLOGIST** (*RLOGIST* is used to differentiate it from the SAS procedure, PROC LOGISTIC, and is used with SAS-callable SUDAAN); recall that the response should be a dichotomous 0-1 variable.

```
/*Multivariable logistic regression of gender and education on
SeekCancerInfo*/
proc rlogist data = hints4cycle2 design = jackknife ddf = 49;
weight person_finwt0;
jackwgts person_finwt1-person_finwt50 / adjjack = 0.98;
class gender edu;
model seekcancerinfo = gender edu;
reflev gender=1 edu=1;
run;
```

The response variable should be on the left hand side (LHS) of the equal sign in the model statement, while all covariates should be listed on the right hand side (RHS). Categorical variables should also be included in the class statement. By default, the reference level of each categorical variable is that of the highest numeric level. This may be changed by using the reflev statement to explicitly define another reference level.

Variance Estimation Method: Replicate Weight Jackknife
Working Correlations: Independent
Link Function: Logit
Response variable SEEKCANCERINFO: A5. Have you ever looked for information about cancer from any source?
by: Independent Variables and Effects.

Independent variables and effects	Beta Coeff.	SE Beta	T-test B=0	P-value T-Test B=0
Intercept	-0.73	0.20	-3.71	0.0005
Gender				
Male	0.00	0.00	.	.
Female	0.36	0.12	3.03	0.0039
Education Level				
Less than HS	0.00	0.00	.	.
12 years or HS completed	0.12	0.22	0.54	0.5942
Some College	0.41	0.19	2.15	0.0364
College graduate or higher	0.74	0.20	3.71	0.0005

Contrast Test Results

Contrast	Degrees of Freedom	Wald F	P-value Wald Chi-Sq
Overall Model	5	5.03	0.0009
Model minus intercept	4	6.24	0.0004
Intercept	.	.	.
Gender	1	9.2	0.0039
Edu	3	6.42	0.0009

Odds Ratio Estimates

Independent variables and effects	Odds Ratio	Lower 95% Limit OR	Upper 95% Limit OR
Intercept	0.48	0.33	0.72
Gender			
Male	1.00	1.00	1.00
Female	1.44	1.13	1.83
Education Level			
Less than HS	1.00	1.00	1.00

12 years or HS completed	1.13	0.72	1.75
Some College	1.51	1.03	2.23
College graduate or higher	2.10	1.40	3.14

To identify levels/variables that display a significant difference in response, the rule of thumb is to examine odds ratios where the confidence interval does not contain 1 (by default, SUDAAN will use alpha=.05 to determine statistical significance; this value can be changed by the user using code). However, significance may also be garnered from the test of whether the associated beta parameter is equal to 0 (see first regression table above). According to this model, women and college students appear to be statistically more inclined to search for cancer information (compared with men and those who did not graduate from high school, respectively).

Linear Regression

This example demonstrates a multivariable linear regression model using **PROC REGRESS** (REGRESS is used to differentiate it from the SAS procedure, PROC REG, and is used with SAS-callable SUDAAN); recall that the response should be a continuous variable. For the purposes of this example, we decided to use an outcome with five levels as a continuous variable (GENERALHEALTH). Note that higher values on GENERALHEALTH indicate poorer self-reported health status.

```
/*Multivariable linear regression of gender and education on GeneralHealth*/
proc regress data = hints4cycle2 design = jackknife ddf = 49;
weight person_finwt0;
jackwgts person_finwt1-person_finwt50 / adjjack = 0.98;
class gender edu;
model generalhealth = gender edu ;
reflev gender=1 edu=1;
run;
```

Variance Estimation Method: Replicate Weight Jackknife
Working Correlations: Independent
Link Function: Identity
Response variable GENERALHEALTH: F1. In general, would you say your health is…
by: Contrast.

Contrast	Degrees of Freedom	Wald F	P-value Wald F
Overall Model	5	2294.00	0.0000
Model minus intercept	4	51.54	0.0000
Intercept	.	.	.
Gender	1	0.35	0.5558
Edu	3	54.20	0.0000

From the above table, we can see that Gender is not associated with the outcome, but Edu is associated, adjusting for all variables in the model.

Variance Estimation Method: Replicate Weight Jackknife
Working Correlations: Independent
Link Function: Identity
Response variable GENERALHEALTH: F1. In general, would you say your health is…
by: Independent Variables and Effects.

Independent variables and effects	Beta Coeff.	SE Beta	T-test B=0	P-value T-Test B=0
Intercept	3.19	0.09	34.98	0.0000
Gender				
Male	0.00	0.00	.	.
Female	0.03	0.05	0.59	0.5558
Education Level				
Less than HS	0.00	0.00	.	.
12 years or HS completed	-0.54	0.11	-4.74	0.0000
Some College	-0.65	0.10	-6.39	0.0000
College graduate or higher	-0.94	0.08	-11.50	0.0000

From the above table, it can be seen that, compared to those respondents with Less than High School education, those with Some College have a significantly negative linear association with the outcome (i.e., better reported health), controlling for all variables in the model. This association also applies to those with a College Degree or Higher. We don't interpret the Gender variable because it is non-significant.

Appendix D: Analyzing data using STATA

This section gives some Stata (Version 10.0 and higher) coding examples for common types of statistical analyses using HINTS 4 Cycle 2 data. We begin by doing data management of the HINTS 4 data. We first decided to exclude all “Missing data (Not Ascertained)”, “Multiple responses selected in error”, “Question answered in error (Commission Error)” and “Inapplicable, coded 2 in SeekHealthInfo” responses from the analyses. By setting these values to missing (.), Stata will exclude these responses from analysis commands where these variables are specifically accessed. For logistic regression modeling within the **svy: logit** command, Stata expects the response variable to be dichotomous with values (0, 1), so this variable will also be recoded at this point. W hen recoding existing variables, it is generally recommended to create new variables of rather than over-writing the existing variables. Note: New variables should always be compared to original source variables in a Stata **tabulate** command to verify proper coding.

```
use “file path\hints4cycle2_09062017_public.dta”
* Recode negative values to missing

recode genderc (1=1 "Male") (2=2 "Female") (nonmissing=.), generate(gender)

label variable gender "Gender"

* Recode education into four levels, and negative values to missing

recode education (1/2=1 "Less than high school") (3=2 "12 years or completed
high school") (4/5=3 "Some college") (6/7=4 "College graduate or higher")
(nonmissing=.), generate(edu)

label variable edu "Education"



* Recode seekcancerinfo to 0-1 format, and negative values to missing for
svy: logit

replace seekcancerinfo = 0 if seekcancerinfo == 2

replace seekcancerinfo = . if seekcancerinfo == -1 | seekcancerinfo == -2 |
seekcancerinfo == -9

label define seekcancerinfo 0 "No" 1 "Yes"

label val seekcancerinfo seekcancerinfo



* Recode negative values to missing for svy: regress

replace generalhealth = . if generalhealth == -5 | generalhealth == -9
```

Declare survey design

Stata requires declaring the survey design for the data set globally before any analysis. The declared survey design will be applied to all future survey commands unless another survey design is declared. Other data sets that incorporate the final sample weight and the 50 jackknife replicate weights will utilize the same code.

```
* Declare survey design for the data set

svyset [pw=person_finwt0], jkrw(person_finwt1-person_finwt50,
multiplier(0.98)) vce(jack) mse
```

Cross-tabulation

```
* cross-tabulation

svy: tabulate edu gender, column row format(%8.5f) percent wald noadjust
```

The **svy: tabulate** command defines the frequencies that should be generated. Single variables listed in **svy: tabulate** results in one-way frequencies, while two variables will define cross-frequencies. The options **column** and **row** request column and row frequencies, respectively. The option **percent** requests the frequencies are displayed in percentage. The options **wald** and **noadjust** together request unadjusted Wald test for independence. Stata recommends default pearson test for independence. Other tests and statistics are also available; see the Stata website for more information: http://www.stata.com/

For the purposes of computing appropriate degrees of freedom for the estimator of the HINTS 4 Cycle 2 differences, we can assume as an approximation that the sample is a simple random sample of size 50 (corresponding to the 50 replicates: each replicate provides a 'pseudo sample unit') from a normal distribution. The denominator degrees of freedom (df) is equal to 49*k, where k is the number of iterations of data used in this analysis. Stata uses the number of replicates minus one as the denominator degrees of freedom and does not provide the option for user to specify the denominator degrees of freedom.

Jknife *: for cell counts

Number of strata = 1 Number of obs = 3489

Population size = 229731000

Replications = 50

Design df = 49

Education	Gender		Total
	Male	Female	
Less than HS	52.43146	47.56854	1.0e+02
	14.28360	12.33309	13.28421

12 years or HS completed	49.74638	50.25362	1.0e+02
	20.70359	19.90481	20.29432
Some college	46.99898	53.00102	1.0e+02
	36.39165	39.05746	37.75753
College grad or higher	48.69021	51.30979	1.0e+02
	28.62116	28.70465	28.66394
Total	48.76298	51.23702	1.0e+02
	1.0e+02	1.0e+02	1.0e+02

Key: row percentages

column percentages

Wald (Pearson):

Unadjusted chi2(3) = 6.4479

Unadjusted F(3, 49) = 2.1493 P = 0.1060

Adjusted F(3, 47) = 2.0616 P = 0.1181

Logistic Regression

This example demonstrates a multivariable logistic regression model using **svy: logit** (to get parameters) and **svy, or: logit** (to get odds ratios); recall that the response should be a dichotomous 0-1 variable.

```
* Define reference group for categorical variables for both svy: logit and svy: regress

char gender [omit] 1

char edu [omit] 1


* Multivariable logistic regression of gender and education on seekcancerinfo

xi: svy: logit seekcancerinfo i.gender i.edu

test _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4 _cons, nosvyadjust

test _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust
```

```
test _Igender_2, nosvyadjust

test _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust

xi: svy, or: logit seekcancerinfo i.gender i.edu
```

The **char** command defines categorical variable with reference group. The "Male" is the reference group for gender effect while the "Less than high school" is the reference group for education level effect. These definitions will be applied to future commands until another **char** command re-defines the reference group. The xi command will create proper dummy variables for i.gender and i.edu variables in the analysis commands. The response variable should be the first variable in **svy: logit** command and be followed by all covariates. The **test** command tests the hypotheses about estimated parameters.

i.gender _Igender_1-2 (naturally coded; _Igender_1 omitted)

i.edu _Iedu_1-4 (naturally coded; _Iedu_1 omitted)

(running logit on estimation sample)

Jackknife replications (50)

----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5

.... 50

Survey: Logistic regression

Number of strata = 1 Number of obs = 2879

Population size = 187658584

Replications = 50

Design df = 49

F(4, 46) = 5.86

Prob > F = 0.0007

seekcancer~o	Coef.	Jknife * Std. Err.	t	P>|t|	[95% Conf. Interval]	
_Igender_2	.3626414	.1195269	3.030	0.004	.1224431	.6028396
_Iedu_2	.1183938	.2207892	0.54	0.594	-.3252986	.5620862

_Iedu_3	.4148765	.1928915	2.15	0.036	.0272465	.8025065
_Iedu_4	.7414217	.1998014	3.71	0.001	.3399059	1.142938
_cons	-.7284504	.1964834	-3.71	0.001	-1.123298	-.3336023

Unadjusted Wald test

(1) _Igender_2 = 0

(2) _Iedu_2 = 0

(3) _Iedu_3 = 0

(4) _Iedu_4 = 0

(5) _cons = 0

F(5, 49) = 5.03

Prob > F = 0.0009

Unadjusted Wald test

(1) _Igender_2 = 0

(2) _Iedu_2 = 0

(3) _Iedu_3 = 0

(4) _Iedu_4 = 0

F(4, 49) = 6.24

Prob > F = 0.0004

Unadjusted Wald test

(1) _Igender_2 = 0

F(1, 49) = 9.20

Prob > F = 0.0039

Unadjusted Wald test

(1) _Iedu_2 = 0

(2) _Iedu_3 = 0

(3) _Iedu_4 = 0

F(3, 49) = 6.42

Prob > F = 0.0009

i.gender _Igender_1-2 (naturally coded; _Igender_1 omitted)

i.edu _Iedu_1-4 (naturally coded; _Iedu_1 omitted)

(running logit on estimation sample)

Jackknife replications (50)

----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5

.... 50

Survey: Logistic regression

Number of strata = 1 Number of obs = 2879

Population size = 187658584

Replications = 50

Design df = 49

F(4, 46) = 5.86

Prob > F = 0.0007

seekcancer~o	Odds Ratio	Jknife *	t	P>|t|	[95% Conf. Interval]

		Std. Err.				
_Igender_2	1.43712	.1717745	3.030	0.004	1.130255	1.8273
_Iedu_2	1.125687	.2485396	0.54	0.594	.7223116	1.754329
_Iedu_3	1.514184	.2920732	2.15	0.036	1.027621	2.231126
_Iedu_4	2.098918	.4193666	3.71	0.001	1.404815	3.135967

To identify levels/variables that display a significant difference in response, the rule of thumb is to examine odds ratios where the confidence interval does not contain 1 (by default, Stata will use alpha=.05 to determine statistical significance; this value can be changed by the user using code). However, significance may also be garnered from the test of whether the associated beta parameter is equal to 0 (see first regression table above). According to this model, women and college students appear to be statistically more inclined to search for cancer information (compared with men and those who did not graduate from high school, respectively).

Linear Regression

This example demonstrates a multivariable linear regression model using **svy: regress**; recall that the response should be a continuous variable. For the purposes of this example, we decided to use an outcome with five levels as a continuous variable (generalhealth). Note that higher values on generalhealth indicate poorer self-reported health status.

```
* Multivariable linear regression of gender and education on generalhealth

xi: svy: regress generalhealth i.gender i.edu

test _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4 _cons, nosvyadjust

test _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust

test _Igender_2, nosvyadjust

test _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust
```

```
i.gender      _Igender_1-2      (naturally coded; _Igender_1 omitted)

i.edu         _Iedu_1-4         (naturally coded; _Iedu_1 omitted)

(running regress on estimation sample)

Jackknife replications (50)

----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5

.................................................. 50
```

Survey: Linear regression

Number of strata = 1   Number of obs = 3401

Population size = 223433147

Replications = 50

Design df = 49

F(4, 46) = 48.38

Prob > F = 0.0000

R-squared = 0.0880

generalhea~h	Coef.	Jknife * Std. Err.	t	P>|t|	[95% Conf. Interval]	
_Igender_2	.0325468	.0548737	0.59	0.556	-.077726	.1428197
_Iedu_2	-.5355744	.1131031	-4.74	0.000	-.7628635	-.3082853
_Iedu_3	-.6466411	.1012391	-6.39	0.000	-.8500887	-.4431934
_Iedu_4	-.9360366	.0814069	-11.50	0.000	-1.09963	-.7724433
_cons	3.188343	.0911525	34.98	0.000	3.005165	3.371521

Unadjusted Wald test

(1) _Igender_2 = 0

(2) _Iedu_2 = 0

(3) _Iedu_3 = 0

(4) _Iedu_4 = 0

(5) _cons = 0

F(5, 49) = 2294.03

Prob > F =   0.0000

Unadjusted Wald test

(1) _Igender_2 = 0

(2) _Iedu_2 = 0

(3) _Iedu_3 = 0

(4) _Iedu_4 = 0

F(4,   49) = 51.54

Prob > F =   0.0000

Unadjusted Wald test

(1) _Igender_2 = 0

F(1,   49) =   0.35

Prob > F =   0.5558

Unadjusted Wald test

(1) _Iedu_2 = 0

(2) _Iedu_3 = 0

(3) _Iedu_4 = 0

F(3,   49) = 54.20

Prob > F =   0.0000

From the above table, it can be seen that, compared to those respondents with Less than High School education, those with Some College have a significantly negative linear association with the outcome (i.e., better reported health), controlling for all variables in the model. This association also applies to those with a College Degree or Higher. We don’t interpret the Gender variable because it is non- significant.