Analytics Recommendations for HINTS4 – Cycle 1 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 were conducted in 2003, 2005, and 2007/2008.

HINTS 4

The most recent version of HINTS administration (referred to as HINTS 4) includes four mail-mode data collection cycles over the course of three years. The first of these cycles (Cycle 1) was conducted from October 2011 through February 2012 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 2007-2008, a single-mode mail survey was implemented for HINTS 4. HINTS 4 involves 4 separate data collection cycles over a three year field period, which began late in 2011 and will extend into 2014. For more extensive background about the HINTS program and previous data collection efforts, see Finney Rutten et al. (in press).

Methodology

Data collection for Cycle 1 of HINTS 4 was initiated in October 2011 and concluded in February of 2012. HINTS 4 Cycle 1 was a self-administered mailed questionnaire. The sampling consisted of a two-stage stratified sample of addresses used by Marketing Systems Group (MSG). 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 card mailing, and up to three additional mailings of the questionnaire as needed by non-responding households.

As with prior HINTS administrations, methodological experiments were embedded to explore the potential impact of varying methodological strategies on response rates. Two methods of respondent selection were used for Cycle 1: the “All Adult” method and the “Next Birthday” method. In the All Adult method, two questionnaires were sent with each mailing, where all adults residing in a sampled household were asked to complete the questionnaire. In the next birthday method, one questionnaire was sent with each mailing so that the adult who would have the next birthday in the sampled household was asked to complete the questionnaire. All households were sent a full version of the survey; those households that did not respond to the first two mailings were then sent a reduced version of the survey. The reduced version instrument had 134 items, compared to the Full version, which had 205 items.

The survey was also administered in both English and Spanish. Spanish instruments were sent to households that were flagged as living in a linguistically-isolated area, having a potentially Spanish surname, or called to request materials in Spanish. Refer to the HINTS 4 Cycle 1 Methodology Report for more extensive information about the sampling procedures.

Sample Size and Response Rates

The final HINTS 4 Cycle 1 sample consisted of 3,959 respondents. Note that 52 of these respondents were considered partial completers who did not answer the entire survey. Household response rates were calculated separately for each respondent selection method using the American Association for Public Opinion Research response rate 2 (RR2) formula (2011). The household response rate for the next birthday method was 37.9%. The household response rate for the all adult method was 35.3%. For the all adult method, a within household response rate was calculated at 84.6%. The final response rate was determined by combining response rates across respondent selection method in proportion to the sample allocated to each method. For Cycle 1 of HINTS 4, the final response rate was 36.7%,

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 probably 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.

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 hints4cycle1_09062017_public.sas7bdat data file in a folder on your computer.

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

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

Important Analytic Variables in the Database

Note: Refer to the HINTS 4 Cycle 1 Methodology Report for more information regarding the weighting and stratification variables listed below.

PERSON_FINWT0: Final sample weight used to get population estimates. Note that estimates from the 2010 American Community Survey (ACS) of the US Census Bureau were used to calibrate the HINTS 4 Cycle 1 control totals with the following variables: Age, gender, education, marital status, race, ethnicity, and census region. In addition, variables from the 2010 National Health Interview Survey (NHIS) were used to calibrate HINTS 4 data control totals regarding: Percent with health insurance and percent ever had cancer.

PERSON_FINWT1 THRU 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 1 Methodology Report” included with this document or in Korn and Graubard (1999).

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

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 It has three values: Central Appalachia regardless of minority population (CA), 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

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: This variable codes for selection method (All Adult vs. Next Birthday). **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 52 partially complete questionnaires. The 22 questionnaires with fewer than 50% of the required questions answered in Sections A and B were coded as incompletely-filledout 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, Race/Ethnicity (RaceEthn), Do you rent or own your house? (K14), How comfortable do you feel speaking English? (K8), and W ere you born in the United States? (K6).

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 1 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 1 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, and 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 1 data, should use 49 ddf in their statistical models. HINTS statistical analyses that involve more than one iteration of data will typically utillize 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.

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. (in press) Picking Up the Pace: Changes in Method and Frame for the Health Information National Trends Survey (2011 – 2014). Journal of Health Communication.

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 Wiley & 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 1 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 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 text gives some SAS (Version 9.2 and higher) coding examples for common types of statistical analyses using HINTS4-Cycle 1 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. 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 Generalf
1 = "Excellent"
2 = "Very good"
3 = "Good"
4 = "Fair"
5 = "Poor";

run;


data hints4cycle1;
set hints4.hints4cycle1_09062017_public;

/*Recode negative values to missing*/
if genderc = 1 then gender = 1;
if genderc = 2 then gender = 2;
if genderc = -9 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, and negative values to missing for
proc surveylogistic procedure*/
if seekcancerinfo = 2 then seekcancerinfo = 0;
if seekcancerinfo = -9 then seekcancerinfo = .;

/*Recode negative values to missing for proc surveyreg 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;
```

Crosstabs procedure

This syntax is consistent for all procedures. Other data sets that incorporate the final sample weight and the 50 jackknife replicate weights will utilize the same three lines of code.

```
proc surveyfreq data=hints4cycle1 varmethod=jackknife;
      weight person_finwt0;
      repweights person_finwt1-person_finwt50 / df=49 jkcoefs=0.98;
      tables gender edu;
      tables edu*gender / row col nowt 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 option wchisq requests Wald chi-square test for independence. Other tests and statistics are also available; see the SAS site link at the end of this document for more information.

For the purposes of computing appropriate degrees of freedom for the estimator of the HINTS4-Cycle 1 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      HINTS4CYCLE1
Number of Replicates 50
```

Table of edu 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	145	6.3297	0.2447	51.8841	1.2504	13.0234	0.4661
	Female	223	5.8700	0.1546	48.1159	1.2504	11.4208	0.2870
	Total	368	12.1997	0.2666	100.000			
12 years or completed high school	Male	296	11.7210	0.6497	51.0491	1.8562	24.1160	1.3446
	Female	462	11.2392	0.4871	48.9509	1.8562	21.8672	0.9333
	Total	758	22.9602	0.7609	100.000			
Some	Male	468	14.8951	0.6365	47.2323	1.3061	30.6468	1.3015

college	Female	679	16.6407	0.4719	52.7677	1.3061	32.3765	0.9163
	Total	1147	31.5359	0.7773	100.000			
College graduate or higher	Male	616	15.6567	0.1706	47.0109	0.3380	32.2137	0.3636
	Female	891	17.6476	0.1087	52.9891	0.3380	34.3355	0.2077
	Total	1507	33.3043	0.1822	100.000			
Total	Male	1525	48.6024	0.2343			100.000	
	Female	2255	51.3976	0.2343			100.000	
	Total	3780	100.000					

Frequency Missing = 179

Wald Chi-Square Test

Chi-Square	23.8394
F Value	7.9465
Num DF	3
Den DF	49
Pr > F	0.0002
Adj F Value	7.6221
Num DF	3
Den DF	47
Pr > Adj F	0.0003

Sample Size = 3780

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=hints4cycle1 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 HINTS4CYCLE1
Number of Replicates 50

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.0009	0.1906	27.5727	<.0001
Female	1	0.7034	0.1073	43.0055	<.0001
HighSchool	1	0.2517	0.2176	1.3373	0.2475
SomeCollege	1	0.8074	0.2193	13.5538	0.0002
CollegeorMore	1	1.2586	0.2072	36.8964	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Female	2.021	1.638	2.493
HighSchool	1.286	0.840	1.970
SomeCollege	2.242	1.459	3.446
CollegeorMore	3.521	2.346	5.284

Contrast Test Results

Contrast	DF	Wald Chi-Square	Pr > ChiSq
Overall model	5	147.1441	<.0001
Overall model minus intercept	4	109.3120	<.0001
Gender	1	43.0055	<.0001
Education overall	3	72.2950	<.0001

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=hints4cycle1 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       HINTS4CYCLE1
Number of Replicates 50
```

Analysis of Contrasts

Contrast	Num DF	F Value	Pr > F
Overall model	5	2956.25	<.0001
Overall model minus intercept	4	24.39	<.0001
Gender	1	0.42	0.5212
Education overall	3	32.28	<.0001

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

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.

Estimated Regression Coefficients

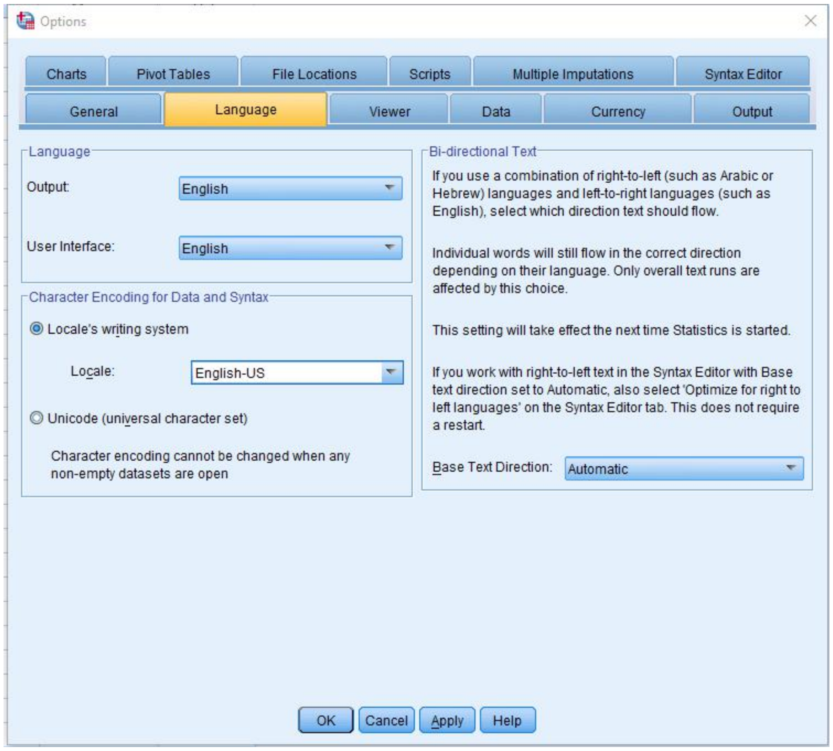
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	2.8929636	0.09492330	30.48	<.0001
Female	-0.0320630	0.04961695	-0.65	0.5212
HighSchool	-0.1666242	0.11081877	-1.50	0.1391
SomeCollege	-0.3323270	0.10708879	-3.10	0.0032
CollegeorMore	-0.6350249	0.09011497	-7.05	<.0001

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 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 B: Analyzing data using SPSS

Prior to opening the HINTS 4, Cycle 1 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.



Language

Output: English

User Interface: English

Character Encoding for Data and Syntax

● Locale's writing system

Locale: English-US

○ Unicode (universal character set)

Character encoding cannot be changed when any non-empty datasets are open

Bi-directional Text

If you use a combination of right-to-left (such as Arabic or Hebrew) languages and left-to-right languages (such as English), select which direction text should flow.

Individual words will still flow in the correct direction depending on their language. Only overall text runs are affected by this choice.

This setting will take effect the next time Statistics is started.

If you work with right-to-left text in the Syntax Editor with Base text direction set to Automatic, also select 'Optimize for right to left languages' on the Syntax Editor tab. This does not require a restart.

Base Text Direction: Automatic

This section gives some SPSS (Version 25 and higher) coding examples for common types of statistical analyses using HINTS 4 Cycle 1 data. These examples will incorporate the stratum variable, VAR_STRATUM, and the cluster variable VAR_CLUSTER. Although these examples specifically use HINTS 4, Cycle 1 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 1 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 1 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	14284377.426	13246958.335	27531335.762
		Standard Error	2261701.539	1427065.058	2552147.189
		Unweighted Count	145	223	368
	% within edu	Estimate	51.9%	48.1%	100.0%
		Standard Error	5.0%	5.0%	0.0%
		Unweighted Count	145	223	368
	% within gender	Estimate	13.0%	11.4%	12.2%
		Standard Error	2.0%	1.2%	1.1%
		Unweighted Count	145	223	368
	% of Total	Estimate	6.3%	5.9%	12.2%
		Standard Error	1.0%	0.6%	1.1%
		Unweighted Count	145	223	368
12 years or completed high school	Population Size	Estimate	26450941.144	25363740.770	51814681.914
		Standard Error	2773518.761	2021332.927	3273453.911
		Unweighted Count	296	462	758
	% within edu	Estimate	51.0%	49.0%	100.0%
		Standard Error	3.4%	3.4%	0.0%
		Unweighted Count	296	462	758
	% within gender	Estimate	24.1%	21.9%	23.0%
		Standard Error	2.4%	1.6%	1.4%
		Unweighted Count	296	462	758
	% of Total	Estimate	11.7%	11.2%	23.0%
		Standard Error	1.2%	0.9%	1.4%
		Unweighted Count	296	462	758
Some college	Population Size	Estimate	33614103.078	37553530.456	71167633.534

		Standard Error	3358898.865	2139739.931	4227964.957
		Unweighted Count	468	679	1147
	% within edu	Estimate	47.2%	52.8%	100.0%
		Standard Error	2.7%	2.7%	0.0%
		Unweighted Count	468	679	1147
	% within gender	Estimate	30.6%	32.4%	31.5%
		Standard Error	2.5%	1.6%	1.5%
		Unweighted Count	468	679	1147
	% of Total	Estimate	14.9%	16.6%	31.5%
		Standard Error	1.3%	0.9%	1.5%
		Unweighted Count	468	679	1147
College graduate or higher	Population Size	Estimate	35332710.264	39825796.489	75158506.753
		Standard Error	2069881.825	1333095.345	2418841.232
		Unweighted Count	616	891	1507
	% within edu	Estimate	47.0%	53.0%	100.0%
		Standard Error	1.7%	1.7%	0.0%
		Unweighted Count	616	891	1507
	% within gender	Estimate	32.2%	34.3%	33.3%
		Standard Error	1.7%	1.1%	1.0%
		Unweighted Count	616	891	1507
	% of Total	Estimate	15.7%	17.6%	33.3%
		Standard Error	0.9%	0.6%	1.0%
		Unweighted Count	616	891	1507
Total	Population Size	Estimate	109682131.912	115990026.051	225672157.963
		Standard Error	4672303.032	3090321.594	5637629.824
		Unweighted Count	1525	2255	3780
	% within edu	Estimate	48.6%	51.4%	100.0%
		Standard Error	1.2%	1.2%	0.0%
		Unweighted Count	1525	2255	3780
	% within gender	Estimate	100.0%	100.0%	100.0%
		Standard Error	0.0%	0.0%	0.0%
		Unweighted Count	1525	2255	3780
	% of Total	Estimate	48.6%	51.4%	100.0%
		Standard Error	1.2%	1.2%	0.0%
		Unweighted Count	1525	2255	3780

Tests of Independence

		Chi-Square	Adjusted F	df1	df2	Sig.
edu * gender	Pearson	6.240	.553	2.636	345.353	.624
	Likelihood Ratio	6.240	.553	2.636	345.353	.624

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, 17.6% vs 15.7%. 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).

The results of these tests conducted in SPSS based on Taylor Series linearization contradict the results conducted in SAS using replication shown in Appendix A. (In SAS, the distributions of educational attainment between men and women were determined to be statistically different.) This is a good example of how the variance estimation method used can affect the outcome of a statistical test. Both education and gender are variables used in the raking process as part of the HINTS weighting procedure. As a result, the standard errors based on replication are much smaller than those based on Taylor Series linearization, which in turn results in significant differences in SAS but not in SPSS.

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	3749
	Invalid	210
	Total	3959
Population Size		224501470.172
Stage 1	Strata	3
	Units	134
Sampling Design Degrees of Freedom		131

Parameter Estimates

seekcancerinfo_recode	Parameter	B	Std. Error	95% Confidence Interval		Hypothesis Test			Exp(B)	95% Confidence Interval for Exp(B)	
				Lower	Upper	t	df	Sig.		Lower	Upper
Yes	(Intercept)	-1.001	.200	-1.396	-.606	-5.009	131.000	.000	.368	.248	.546
	Female	.703	.104	.498	.909	6.781	131.000	.000	2.021	1.646	2.481
	College Graduate or Higher	1.259	.213	.838	1.679	5.921	131.000	.000	3.521	2.312	5.361
	Some College	.807	.214	.385	1.230	3.780	131.000	.000	2.242	1.469	3.421
	12 Years or Completed High School	.252	.216	-.175	.678	1.167	131.000	.245	1.286	.840	1.970

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	Upper
flippedgender	Female vs. Male	2.021	1.646	2.481
flippededu	College graduate or higher vs. Less than high school	3.521	2.312	5.361
	Some college vs. Less than high school	2.242	1.469	3.421
	12 years or completed high school vs. Less than high school	1.286	.840	1.970

Overall Model Minus Intercept

df	Wald Chi-Square	Sig.
4.000	91.778	.000

Gender

df	Wald Chi-Square	Sig.
1.000	45.973	.000

Education Overall

df	Wald Chi-Square	Sig.
3.000	61.124	.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	3752

	Invalid	207
	Total	3959
Population Size		222913453.733
Stage 1	Strata	3
	Units	134
Sampling Design Degrees of Freedom		131

Parameter Estimates[a]

			95% Confidence Interval		Hypothesis Test		
Parameter	Estimate	Std. Error	Lower	Upper	t	df	Sig.
(Intercept)	2.893	.089	2.717	3.069	32.536	131.000	.000
College Graduate or Higher	-.635	.084	-.801	-.469	-7.551	131.000	.000
Some College	-.332	.102	-.533	-.131	-3.270	131.000	.001
12 Years or Completed High School	-.167	.099	-.363	.030	-1.676	131.000	.096
Female	-.032	.048	-.128	.064	-.663	131.000	.508

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 some college 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 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	128.000	25.334	.000

Gender

df1	df2	Wald F	Sig.
1.000	131.000	.440	.508

Education

df1	df2	Wald F	Sig.
3.000	129.000	34.021	.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 document gives some SUDAAN (Version 10.0.1 and higher) coding examples for common types of statistical analyses using HINTS4-Cycle 1 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 hints4cycle1;
set hints4.hints4cycle1_09062017_public;

/*Recode negative values to missing*/
if genderc = 1 then gender = 1;
if genderc = 2 then gender = 2;
if genderc = -9 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, and negative values to missing for
proc rlogist procedure*/
if seekcancerinfo = 2 then seekcancerinfo = 0;
if seekcancerinfo = -9 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;
```

Crosstabs procedure

This syntax is consistent for all procedures. Other data sets that incorporate the final sample weight and the 50 jackknife replicate weights will utilize the same three lines of code.

```
proc crosstab data=hints4cycle1 design=jackknife ddf =49;
weight person_finwt0;
jackwgts person_finwt1-person_finwt50 / adjjack=.98;
class gender edu;
tables gender edu gender*edu;
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 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 at the end of this document 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 1 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

What is the highest grade or level of schooling you completed?		Are you male or female?		
		Total	**Male**	**Female**
Total	**Sample Size**	3780	1525	2255
	Col Percent	100.00	100.00	100.00
	Row Percent	100.00	48.60	51.40
Less than HS	**Sample Size**	368	145	223
	Col Percent	12.20	13.02	11.42
	Row Percent	100.00	51.88	48.12
12 years or completed HS	**Sample Size**	758	296	462
	Col Percent	22.96	24.12	21.87
	Row Percent	100.00	51.05	48.95
Some college	**Sample Size**	1147	468	679
	Col Percent	31.54	30.65	32.38
	Row Percent	100.00	47.23	52.77
College graduate or higher	**Sample Size**	1507	616	891
	Col Percent	33.30	32.21	34.34
	Row Percent	100.00	47.01	52.99

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

ChiSq	7.95
P-value for ChiSq	0.0002
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 = hints4cycle1 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 *reflevel* statement to explicitly define another reference level.

Variance Estimation Method: Replicate Weight Jackknife
Working Correlations: Independent
Link Function: Logit
Response variable SEEKCANCERINFO: A9. 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	-1.00	0.19	-5.25	0.0000
Gender				
Male	0.00	0.00	.	.
Female	0.70	0.11	6.56	0.0000
Education Level				
Less than HS	0.00	0.00	.	.
12 years or HS completed	0.25	0.22	1.16	0.2531
Some College	0.81	0.22	3.68	0.0006
College graduate or higher	1.26	0.21	6.07	0.0000

Contrast	**Wald F**	**P-value Wald Chi-Sq**
Overall Model	29.43	0.0000
Model minus intercept	27.33	0.0000
Intercept	.	.
Gender	43.01	0.0000
Edu	24.1	0.0000

Variance Estimation Method: Replicate Weight Jackknife
Working Correlations: Independent
Link Function: Logit
Response variable SEEKCANCERINFO: A9. Have you ever looked for information about cancer from any source?

by: Independent Variables and Effects.

Independent variables and effects	Odds Ratio	Lower 95% Limit OR	Upper 95% Limit OR
Intercept	0.37	0.25	0.54
Gender			
Male	1.00	1.00	1.00
Female	2.02	1.63	2.51
Education Level			
Less than HS	1.00	1.00	1.00
12 years or HS completed	1.29	0.83	1.99
Some College	2.24	1.44	3.48
College graduate or higher	3.52	2.32	5.34

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 = hints4cycle1 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: D1. In general, would you say your health is…
by: Contrast.

Contrast	Wald F	P-value Wald F
Overall Model	2956.25	0.0000
Model minus intercept	24.39	0.0000
Intercept	.	.
Gender	0.42	0.5212
Edu	32.28	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: D1. 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	2.89	0.09	30.48	0.0000
Gender				
Male	0.00	0.00	.	.
Female	-0.03	0.05	-0.65	0.5212
Education Level				
Less than HS	0.00	0.00	.	.
12 years or HS completed	-0.17	0.11	-1.50	0.1391
Some College	-0.33	0.11	-3.10	0.0032
College graduate or higher	-0.64	0.09	-7.05	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 text gives some Stata (Version 10.0 and higher) coding examples for common types of statistical analyses using HINTS4-Cycle 1 data. We begin by doing data management of the HINTS 4 data. 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 (.), 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. 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 Stata **tabulate** command to verify proper coding.

```
set memory 512m

use "file path name\ hints4cycle1_09062017_public.dta", clear

* 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 == -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. Stand-alone variable 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 site link at the end of this document for more information.

For the purposes of computing appropriate degrees of freedom for the estimator of the HINTS4-Cycle 1 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     =      3780
                                          Population size   =  225672158
                                          Replications      =       50
                                          Design df         =       49
```

Education	Gender		
	Male	Female	Total
Less tha	51.88407	48.11593	1.0e+02
	13.02343	11.42077	12.19970
12 years	51.04912	48.95088	1.0e+02
	24.11600	21.86717	22.96016
Some col	47.23229	52.76771	1.0e+02
	30.64684	32.37652	31.53585
College	47.01093	52.98907	1.0e+02

	32.21373	34.33554	33.30429
Total	48.60242	51.39758	1.0e+02
	1.0e+02	1.0e+02	1.0e+02

```
Key: row percentages
     column percentages

Wald (Pearson):
  Unadjusted   chi2(3)     =  23.8396
  Unadjusted   F(3, 49)    =   7.9465    P = 0.0002
  Adjusted     F(3, 47)    =   7.6222    P = 0.0003
```

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 = 3749
Population size = 224501470
Replications = 50
Design df = 49
F(4, 46) = 25.65
Prob > F = 0.0000

<table>
<tr><th>seekcancer~o</th><th>Coef.</th><th>Jknife *
Std. Err.</th><th>t</th><th>P>|t|</th><th colspan="2">[95% Conf. Interval]</th></tr>
<tr><td>_Igender_2</td><td>.7034254</td><td>.1072645</td><td>6.56</td><td>0.000</td><td>.4878693</td><td>.9189816</td></tr>
<tr><td>_Iedu_2</td><td>.2516506</td><td>.2176112</td><td>1.16</td><td>0.253</td><td>-.1856555</td><td>.6889567</td></tr>
<tr><td>_Iedu_3</td><td>.8074176</td><td>.2193146</td><td>3.68</td><td>0.001</td><td>.3666884</td><td>1.248147</td></tr>
<tr><td>_Iedu_4</td><td>1.258633</td><td>.2072081</td><td>6.07</td><td>0.000</td><td>.8422324</td><td>1.675033</td></tr>
<tr><td>_cons</td><td>-1.000879</td><td>.1906082</td><td>-5.25</td><td>0.000</td><td>-1.383921</td><td>-.6178375</td></tr>
</table>

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) = 29.43
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) = 27.33
Prob > F = 0.0000

```
Unadjusted Wald test

 ( 1) _Igender_2 = 0

       F( 1, 49) =  43.01
            Prob > F = 0.0000


Unadjusted Wald test

 ( 1) _Iedu_2 =  0
 ( 2)  _Iedu_3 =  0
 ( 3) _Iedu_4 = 0

       F( 3, 49) =  24.10
            Prob > F = 0.0000


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      =       3749
                                            Population size    = 224501470
                                            Replications       =         50
                                            Design df          =         49
                                            F(   4,     46)    =      25.65
                                            Prob > F           =     0.0000
```

seekcancer~o	Odds Ratio	Jknife * Std. Err.	t	P>|t|	[95% Conf. Interval]	
_Igender_2	2.020663	.2167454	6.56	0.000	1.628842	2.506736
_Iedu_2	1.286147	.2798799	1.16	0.253	.8305597	1.991637

_Iedu_3	2.242111	.4917277	3.68	0.001	1.442948	3.483881
_Iedu_4	3.520604	.7294977	6.07	0.000	2.321544	5.338971

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      =       3752
                                           Population size    = 222913454
                                           Replications       =         50
```

Design df = 49
F(4, 46) = 22.90
Prob > F = 0.0000
R-squared = 0.0553

generalhea~h	Coef.	Jknife * Std. Err.	t	P>|t|	[95% Conf. Interval]	
_Igender_2	-.032063	.049617	-0.65	0.521	-.131772	.0676461
_Iedu_2	-.1666242	.1108188	-1.50	0.139	-.3893229	.0560744
_Iedu_3	-.332327	.1070888	-3.10	0.003	-.5475299	-.117124
_Iedu_4	-.6350249	.090115	-7.05	0.000	-.8161177	-.4539322
_cons	2.892964	.0949233	30.48	0.000	2.702208	3.083719

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) = 2956.27
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) = 24.39
Prob > F = 0.0000

```
Unadjusted Wald test

 ( 1) _Igender_2 = 0

       F(  1,  49) =   0.42
            Prob > F = 0.5212


Unadjusted Wald test

 ( 1) _Iedu_2 =  0
 ( 2)  _Iedu_3 =  0
 ( 3) _Iedu_4 = 0

       F( 3, 49) =  32.28
            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.