ANALYTICS RECOMMENDATIONS FOR HINTS 5, CYCLE 2 DATA

November 2018

CONTENTS

Overview of HINTS

The Health Information National Trends Survey (HINTS) is a nationally representative survey that 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), HINTS 4, Cycle 1 (2011/2012); HINTS 4, Cycle 2 (2012/2013); HINTS 4, Cycle 3 (late 2013); HINTS 4, Cycle 4 (2014); HINTS 5, Cycle 1; and HINTS-FDA, Cycle 1 (2015).

HINTS 5

The HINTS 5 administration includes four data collection cycles over four years, starting in 2017. The first of these cycles (HINTS 5, Cycle 1) was conducted from January through May 2017. The focus of this report is HINTS 5, Cycle 2. HINTS 5 draws upon the lessons learned from prior iterations of HINTS. A single-mode mail survey was implemented for HINTS 5, Cycle 2. 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 5 began in January 2018 and concluded in May 2018. HINTS 5, Cycle 2 was a self-administered mailed questionnaire. The sampling frame of addresses, provided by Marketing Systems G
roup (MSG), was grouped into three strata: 1) addresses in areas with high concentrations of minority populations; 2) addresses in areas with low concentrations of minority populations; and 3) addresses located in counties comprising Central Appalachia, regardless of minority population. All non-vacant residential addresses in the United States present in the MSG database, including post office (P.O.) boxes, throwbacks (i.e., street addresses for which mail is redirected by the U.S. 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 for 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). The second stage of sampling consisted of selecting one adult within each sampled household using 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. Refer to the HINTS 5, Cycle 2 Methodology Report for more extensive information about the sampling procedures.

Sample Size and Response Rates

The final HINTS 5, Cycle 2 sample consists of 3,504 respondents. Note that 70 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%–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 32.39%.

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 do not 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.

With the release of HINTS 5, Cycle 2, the SPSS dataset will contain variance codes that will allow for inferential statistical testing using Taylor Series Linearization along with the Complex Samples module. Please see the “Important Analytic Variables in the Database” section for more information about the variance codes, and the “Variance Estimation Methods: Replicate vs. Taylor Linearization” section for more information about the two variance estimation methods.

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.

Important Analytic Variables in the Database

Refer to the HINTS 5, 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 2016 American Community Survey (ACS) of the U.S. Census Bureau were used to calibrate the HINTS 5, Cycle 2 control totals with the following variables: age, gender, education, marital status, race, ethnicity, and census region. In addition, variables from the 2017 National Health Interview Survey (NHIS) were used to calibrate HINTS 5, 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 5, Cycle 2 Methodology Report” included in the data download, or see Korn and Graubard (1999).

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 two values: 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.

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).

HISP_HH: This variable codes for households identified as Hispanic by either being in a high linguistically isolated strata, or having a Hispanic surname match, or both.

APP_REGION: This variable codes for Appalachia subregion.

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

LANGUAGE_FLAG: This variable codes for the 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 partially 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 148 partially complete questionnaires. Fifty-one 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 5, 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)

Replicate Weights: The HINTS 5, 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 5, 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.

Taylor Series: The HINTS 5, Cycle 2 database contains two variables that can be used to calculate standard errors using the Taylor series, namely VAR_STRATUM and VAR_CLUSTER (see

VAR_STRATUM and VAR_CLUSTER variables in the previous section for strata definitions.). The degrees of freedom for the Taylor series, 98, is based on 50 PSUs in each of the two sampling strata (#psus - #strata = 50*2 – 2 = 98).

Statistical Software Example Code

This section provides some coding examples using SAS, SPSS, and STATA for common types of statistical analyses using HINTS 5, Cycle 2 data. For SAS and STATA, you'll see two sets of code: one when using replicate methods for variance estimation, and one for Taylor Series linearization. For replicate methods, 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. For Taylor Series, the code will incorporate the final sample weight and the two variance codes to compute variance estimates. Although these examples specifically use HINTS 5, 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 5, Cycle 2 data: seekcancerinfo & generalhealth.

Analyzing Data Using SAS

Prior to using the HINTS 5, Cycle 2 SAS data, it is important to apply the SAS formats. To do this, follow the steps below.

1. Download all HINTS 5, Cycle 2 documents to a folder on your computer. This should be the same folder where you create the SAS library in step 3.
2. Open the SAS program "HINTS 5 Cycle 2 Public Formats.sas."
3. Change the file location specification in the "library" statement at the top of the program to the location where you want the format library to be stored before you run this program.
4. Run the program "HINTS 5 Cycle 2 Public Formats.sas" to create a permanent SAS format library that is used to analyze the HINTS dataset.
5. Open the SAS program "HINTS 5 Cycle 2 Public Format Assignments.sas."
6. Change the file location specification in the OPTIONS statement at the top of the program to the name of the library where you placed the formats. Also insert the library name for the SET and DATA statements and assign a name to the formatted data in the DATA statement.
7. Run the program "HINTS 5 Cycle 2 Public Format Assignments.sas" to create the formatted SAS dataset.

Note the following:

a. Make sure to run the program "HINTS 5 Cycle 2 Public Formats.sas" BEFORE you run "HINTS 5 Cycle 2 Public Format Assignments.sas" to create the formatted HINTS dataset.
b. If you are getting an error statement saying that SAS is unable to find the formats, make sure you run the OPTIONS statement that includes the correct library name where the formats can be found.

This section gives some SAS (Version 9.3 and higher) coding examples for common types of statistical analyses using HINTS 5, Cycle 2 data. Subsection 1 shows how to complete common analyses using replicate weights, and subsection 2 shows analyses using the Taylor series linearization approach. For either approach, we begin by doing data management of the HINTS 5 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, 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=(hints5c2); *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 hints5cycle2;
set hints5c2.hints5_cycle2_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;
```

Replicate Weights Variance Estimation Method

Frequency Table and Chi-Square Test

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_FINWT0, and those of the jackknife replicate weights, PERSON_FINWT1—PERSON_FINWT50. The jackknife adjustment factor for each replicate weight is 0.98. This syntax is consistent for all procedures. Other datasets that incorporate replicate weight jackknife designs will follow a similar syntax.

```
proc surveyfreq data = hints5cycle2 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. Standalone 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 HINTS5-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	hints5cycle2
Number of Replicates	50

(continued on next page)

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		4.9105	0.7284	55.2312	4.7562	10.0395	1.479
	Female	170	3.9803	0.4157	44.7688	4.7562	7.7909	0.8149
	Total	267	8.8907	0.7996	100			
12 years or completed high school	Male	244	10.8815	0.6847	48.9365	2.0492	22.2474	1.4228
	Female	377	11.3545	0.5031	51.0635	2.0492	22.225	0.973
	Total	621	22.236	0.7892	100			
Some college	Male	421	19.3355	0.6663	48.3366	0.917	39.5317	1.3424
	Female	607	20.6662	0.4288	51.6634	0.917	40.4517	0.8426
	Total	1028	40.0017	0.8591	100			
College graduate or higher	Male	621	13.7838	0.0663	47.7419	0.2724	28.1813	0.1521
	Female	876	15.0877	0.138	52.2581	0.2724	29.5324	0.2349
	Total	1497	28.8716	0.1441	100			
Total	Male	1383	48.9113	0.1805			100	
	Female	2030	51.0887	0.1805			100	
	Total	3413	100					

Frequency Missing =91

Wald Chi-Square Test	
Chi-Square	26.7018
F Value	8.9006
Num DF	3
Den DF	49
Pr > F	<.0001
Adj F Value	8.5373
Num DF	3
Den DF	47
Pr > Adj F	0.0001
Sample Size = 3413	

The row percentages above show that a higher weighted proportion of college graduates in the sample are women (52%) than men (48%). Respondents with less than a high school diploma include more men (55%) than women (45%). The statistic for the Chi-square test of independence and its associated p-value indicate that the distributions of educational attainment between men and women are significantly different.

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= hints5cycle2 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 of the equal sign in the model statement, while all covariates should be listed on the right-hand side. 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	hints5cycle2
Number of Replicates	50

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.1104	0.2858	0.1492	0.6993
Female	1	0.3782	0.1342	7.9429	0.0048
HighSchool	1	-0.52	0.2839	3.355	0.067
SomeCollege	1	-0.0819	0.3012	0.074	0.7856
CollegeorMore	1	0.3312	0.2756	1.4434	0.2296

(continued on next page)

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Female	1.46	1.122	1.899
HighSchool	0.595	0.341	1.037
SomeCollege	0.921	0.511	1.662
CollegeorMore	1.393	0.811	2.39

Contrast Test Results

Contrast	DF	Wald Chi-Square	Pr > ChiSq
Overall model	5	59.1028	<.0001
Overall model minus intercept	4	48.3452	<.0001
Gender	1	7.9429	0.0048
Education overall	3	31.8942	<.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 "Analysis of Maximum Likelihood Estimates" table above). According to this model, women appear to be 1.46 times as likely as men to have searched for cancer information.

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= hints5cycle2 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;
```

(output on next page)

Variance Estimation	
Method	Jackknife
Replicate Weights	hints5cycle2
Number of Replicates	50

Estimated Regression of Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	3.0286336	0.18051496	16.78	<.0001
Female	0.0529823	0.06285156	0.84	0.4033
HighSchool	-0.3499955	0.17780967	-1.97	0.0547
SomeCollege	-0.484692	0.17464474	-2.78	0.0078
CollegeorMore	-0.9064306	0.1719162	-5.27	<.0001

The table labeled Estimated Regression of Coefficients shows that respondents with some college reported better general health than those with less than a high school education (p=0.0078) when controlling for all other variables in the model. Keep in mind that the outcome, general health, is coded such that lower scores correspond to better health. This table also shows that this association applies to those with a college degree or higher (coefficient=-0.91, p<.0001) when comparing to respondents with less than a high school education. However, there's no significant difference in health score between males and females (p=0.4) and those with only a high school diploma and those without a high school diploma (p=0.0547).

Analysis of Contrasts

Contrast	Num DF	F Value	Pr > F
Overall model	5	3786.15	<.0001
Overall model minus intercept	4	30.65	<.0001
Gender	1	0.71	0.4033
Education overall	3	35.08	<.0001

The table labeled Analysis of Contrasts also shows that the association between gender and general health is not significant, but the association between education and general health is significant.

Taylor Series Linearization Variance Estimation Method

Frequency Table and Chi-Square Test

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 strata VAR_STRATUM, cluster VAR_CLUSTER, and overall sample weight PERSON_FINWT0. This syntax is consistent for all procedures. Other analyses that use Taylor Series approximation will follow a similar syntax.

```
proc surveyfreq data = hints5cycle2 varmethod = TAYLOR;
strata VAR_STRATUM; cluster VAR_CLUSTER;
weight person_finwt0;
tables edu*gender / row col wchisq;
run;
```

The *tables* statement defines the frequencies that should be generated. Standalone 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.

Data Summary	
Number of Strata	2
Number of Clusters	100
Number of Observations	3504
Sum of Weights	249489772

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	97	4.9105	0.8402	55.2312	4.9848	10.0395	1.6933
	Female	170	3.9803	0.4144	44.7688	4.9848	7.7909	0.7949
	Total	267	8.8907	0.9313	100			
12 years or completed high school	Male	244	10.8815	0.9488	48.9365	2.9476	22.2474	1.8954
	Female	377	11.3545	0.763	51.0635	2.9476	22.225	1.3664
	Total	621	22.236	1.1228	100			
Some college	Male	421	19.3355	1.6647	48.3366	3.0152	39.5317	2.7403
	Female	607	20.6662	1.2318	51.6634	3.0152	40.4517	1.9135
	Total	1028	40.0017	1.6867	100			
College graduate or higher	Male	621	13.7838	0.7455	47.7419	1.8108	28.1813	1.7421
	Female	876	15.0877	0.7308	52.2581	1.8108	29.5324	1.2883
	Total	1497	28.8716	1.0444	100			
Total	Male	1383	48.9113	1.4982			100	
	Female	2030	51.0887	1.4982			100	
	Total	3413	100					

Frequency Missing =91

Wald Chi-Square Test	
Chi-Square	1.5554
F Value	0.5185
Num DF	3
Den DF	98
Pr > F	0.6705
Adj F Value	0.5079
Num DF	3
Den DF	96
Pr > Adj F	0.6778
Sample Size = 3413	

The row percentages above show that a higher weighted proportion of college graduates in the sample are women (52%) than men (48%). Respondents with less than a high school diploma include more men (55%) than women (45%). The Chi-squared test of independence statistic and associated p value suggest that one should accept the null hypothesis that the two variables are not associated, which indicates that there is not a significant difference between the distributions of educational attainment for these two groups.

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= hints5cycle2 varmethod=TAYLOR;
strata VAR_STRATUM; cluster VAR_CLUSTER;
weight person_finwt0;
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.

(output on next page)

Variance Estimation	
Method	Taylor Series
Variance Adjustment	Degrees of Freedom (DF)

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.1104	0.2675	-0.41	0.6807
Female	1	0.3782	0.1194	3.17	0.0021
HighSchool	1	-0.52	0.294	-1.77	0.0801
SomeCollege	1	-0.0819	0.2661	-0.31	0.7588
CollegeorMore	1	0.3312	0.2465	1.34	0.1822

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Female	1.46	1.152	1.85
HighSchool	0.595	0.332	1.066
SomeCollege	0.921	0.543	1.562
CollegeorMore	1.393	0.854	2.271

Contrast Test Results

Contrast	DF	Wald Chi-Square	Pr > ChiSq
Overall model	5	9.84	<.0001
Overall model minus intercept	4	10.3	<.0001
Gender	1	10.03	0.0021
Education overall	3	9.59	<.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 “Analysis of Maximum Likelihood Estimates” table above). According to this model, women appear to be statistically more likely than men to have searched for cancer information.

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= hints5cycle2 varmethod=TAYLOR;
strata VAR_STRATUM; cluster VAR_CLUSTER;
weight person_finwt0;
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	Taylor Series
Variance Adjustment	Degrees of Freedom (DF)

Estimated Regression of Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	3.0286336	0.15576305	19.44	<.0001
Female	0.0529823	0.06465077	0.82	0.4145
HighSchool	-0.3499955	0.15994325	-2.19	0.031
SomeCollege	-0.484692	0.15023805	-3.23	0.0017
CollegeorMore	-0.9064306	0.14986866	-6.05	<.0001

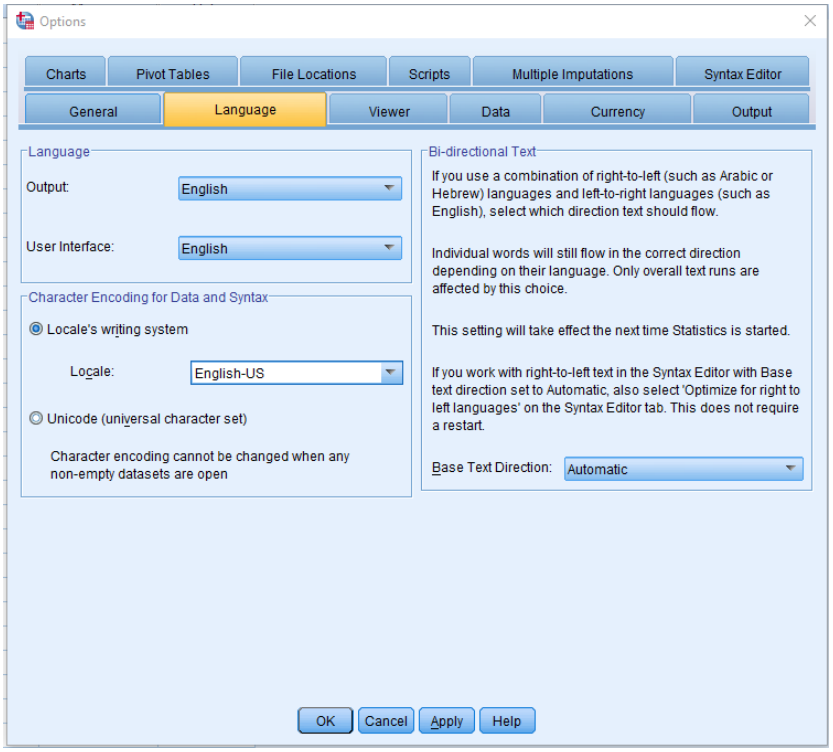
Compared to those respondents with less than a high school education, those who completed some college on average reported significantly better general health (i.e., the negative beta coefficient indicates that the average health score is lower among those with some college, and the health variable is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

Analysis of Contrasts

Contrast	Num DF	F Value	Pr > F
Overall model	5	2789.52	<.0001
Overall model minus intercept	4	24.91	<.0001
Gender	1	0.67	0.4145
Education overall	3	31.06	<.0001

From the above table, we can see that gender is **not** significantly associated with general health, but education is significantly associated with general health, adjusting for all variables in the model.

Prior to opening the HINTS 5, 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 5, Cycle 2 data. 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, 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 WOR (without 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 5 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 always uses the first level of category as the reference category, while SAS uses the last level of category as the reference by default. Users in SPSS cannot define the reference category by themselves. To make SPSS results comparable with SAS, we reverse coded the variables in SPSS. It is better to use dummy variables instead of categorical variables in SPSS complex survey procedures, such as CSLOGISTIC. We use dummy variables for gender and education level in both CSLOGISTIC and CSGLM procedures. 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 negative values to missing.
DATASET ACTIVATE DataSet1.
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".

*Create dummy variables for CSLOGISTIC and CSGLM procedures.
RECODE edu (1=0) (2 thru 4=1) (ELSE=Copy) INTO flippedLessthanHS.
VARIABLE LABELS  flippedLessthanHS 'flippedLessthanHS'.
EXECUTE.
RECODE edu (1=1) (2=0) (3 thru 4=1) (ELSE=Copy) INTO flippedHighSchool.
VARIABLE LABELS  flippedHighSchool 'flippedHighSchool'.
EXECUTE.
RECODE edu (3=0) (4=1) (1 thru 2=1) (ELSE=Copy) INTO flippedSomeCollege.
VARIABLE LABELS  flippedSomeCollege 'flippedSomeCollege'.
EXECUTE.
RECODE edu (4=0) (1 thru 3=1) (ELSE=Copy) INTO flippedCollegeorMore.
VARIABLE LABELS  flippedCollegeorMore 'flippedCollegeorMore'.
EXECUTE.
RECODE gender (2=1) (1=0) (ELSE=Copy) INTO female.
VARIABLE LABELS  female 'female'.
EXECUTE.
```

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='(sample.csaplan)'
  /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		
			Male	Female	Total
Less than high school	**Population Size**	Estimate	11961323.729	9695491.352	21656815.081
		Standard Error	2144365.143	1020647.983	2437037.731
		Unweighted Count	97	170	267
	% within edu	Estimate	55.2%	44.8%	100.0%
		Standard Error	5.0%	5.0%	0.0%
		Unweighted Count	97	170	267
	% within gender	Estimate	10.0%	7.8%	8.9%
		Standard Error	1.7%	0.8%	0.9%
		Unweighted Count	97	170	267
	% of Total	Estimate	4.9%	4.0%	8.9%
		Standard Error	0.8%	0.4%	0.9%
		Unweighted Count	97	170	267
12 years or completed high school	**Population Size**	Estimate	26506066.212	27658154.723	54164220.934
		Standard Error	2498453.406	1927258.559	3135226.633
		Unweighted Count	244	377	621
	% within edu	Estimate	48.9%	51.1%	100.0%
		Standard Error	2.9%	2.9%	0.0%
		Unweighted Count	244	377	621
	% within gender	Estimate	22.2%	22.2%	22.2%
		Standard Error	1.9%	1.4%	1.1%
		Unweighted Count	244	377	621
	% of Total	Estimate	10.9%	11.4%	22.2%
		Standard Error	0.9%	0.8%	1.1%
		Unweighted Count	244	377	621
Some college	**Population Size**	Estimate	47098937.896	50340477.613	97439415.509
		Standard Error	4632315.166	3063645.943	5291127.557
		Unweighted Count	421	607	1028
	% within edu	Estimate	48.3%	51.7%	100.0%
		Standard Error	3.0%	3.0%	0.0%
		Unweighted Count	421	607	1028

Edu			Gender		
			Male	Female	Total
	% within gender	Estimate	39.5%	40.5%	40.0%
		Standard Error	2.7%	1.9%	1.7%
		Unweighted Count	421	607	1028
	% of Total	Estimate	19.3%	20.7%	40.0%
		Standard Error	1.7%	1.2%	1.7%
		Unweighted Count	421	607	1028
College graduate or higher	Population Size	Estimate	33575807.255	36751913.819	70327721.074
		Standard Error	1614210.025	1948229.642	2494137.924
		Unweighted Count	621	876	1497
	% within edu	Estimate	47.7%	52.3%	100.0%
		Standard Error	1.8%	1.8%	0.0%
		Unweighted Count	621	876	1497
	% within gender	Estimate	28.2%	29.5%	28.9%
		Standard Error	1.7%	1.3%	1.0%
		Unweighted Count	621	876	1497
	% of Total	Estimate	13.8%	15.1%	28.9%
		Standard Error	0.7%	0.7%	1.0%
		Unweighted Count	621	876	1497
Total	Population Size	Estimate	119142135.091	124446037.508	243588172.599
		Standard Error	5628235.815	4224600.435	6811953.919
		Unweighted Count	1383	2030	3413
	% within edu	Estimate	48.9%	51.1%	100.0%
		Standard Error	1.5%	1.5%	0.0%
		Unweighted Count	1383	2030	3413
	% within gender	Estimate	100.0%	100.0%	100.0%
		Standard Error	0.0%	0.0%	0.0%
		Unweighted Count	1383	2030	3413
	% of Total	Estimate	48.9%	51.1%	100.0%
		Standard Error	1.5%	1.5%	0.0%
		Unweighted Count	1383	2030	3413

The row percentages above show that a higher weighted proportion of college graduates in the sample are women (52%) than men (48%). Respondents with less than a high school diploma include more men (55%) than women (45%).

		Chi-Square	Adjusted F	df1	df2	Significance
edu * gender	Pearson	5.570	0.626	2.754	269.927	0.585
	Likelihood Ratio	5.574	0.626	2.754	269.927	0.585

Pearson chi-square test statistic and Likelihood Ratio test statistic and their associated p-values suggest that one should accept the null hypothesis that the two variables are not associated, which indicates that there is not a significant difference between the distributions of educational attainment for men and women.

The results of these tests conducted in SPSS based on Taylor Series linearization contradict the results conducted in SAS using replication shown in the "Analyzing Data Using SAS" section. (In SAS, the distributions of education 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(0) BY flippedgender flippedHighSchool flippedSomeCollege
flippedCollegeorMore
  /PLAN FILE='(sample.csaplan)'
  /MODEL flippedgender flippedHighSchool flippedSomeCollege flippedCollegeorMore
  /CUSTOM Label = 'Overall model minus intercept'
   LMATRIX = flippedgender -1/2 1/2;  flippedHighSchool -1/2 1/2; flippedSomeCollege -1/2 1/2;
flippedCollegeorMore -1/2 1/2
   /CUSTOM Label = 'Gender'
  LMATRIX = flippedgender -1/2 1/2
   /CUSTOM Label = 'Education overall'
    LMATRIX = flippedHighSchool -1/2 1/2; flippedSomeCollege -1/2 1/2; flippedCollegeorMore -1/2 1/2
  /INTERCEPT INCLUDE=YES SHOW=YES
  /STATISTICS PARAMETER TTEST EXP SE CINTERVAL DEFF DEFFSQRT
  /TEST TYPE=CHISQUARE PADJUST=LSD
 /ODDSRATIOS FACTOR=[flippedgender(HIGH)]
 /ODDSRATIOS FACTOR=[flippedHighSchool(HIGH)]
 /ODDSRATIOS FACTOR=[flippedSomeCollege(HIGH)]
 /ODDSRATIOS FACTOR=[flippedCollegeorMore(HIGH)]
  /MISSING CLASSMISSING=EXCLUDE
```

```
  /CRITERIA MXITER=100 MXSTEP=5 PCONVERGE=[1e-008 RELATIVE] LCONVERGE=[0]
CHKSEP=20 CILEVEL=95
  /PRINT SUMMARY CLASSTABLE 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 (0) 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 subcommand MODEL specifies the covariates 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, DEFF indicates the design effect for each coefficient estimate, and DEFFSQRT indicates the square root of the design effect 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**	2717
	Invalid	787
	Total	3504
Population Size		192765100.411
Stage 1	**Strata**	2
	Units	100
Sampling Design Degrees of Freedom		98

Parameter Estimates

seekcancerinfo_recode		Estimate	Std. Error	95% Confidence Interval		Hypothesis Test		
				Lower	Upper	t	df	Sig.
Yes	**(Intercept)**	-0.110	0.267	-0.641	0.420	-0.413	98.000	0.680
	[flippedgender=Female]	0.378	0.119	0.141	0.615	3.170	98.000	0.002
	[flippedgender=Male]	.000						
	[flippededu=CollegeorMore]	0.331	0.246	-0.158	0.820	1.344	98.000	0.182
	[flippededu=SomeCollege]	-0.082	0.266	-0.610	0.446	-0.308	98.000	0.759
	[flippededu=HighSchool]	-0.520	0.294	-1.103	0.063	-1.770	98.000	0.080
	[flippededu=LessThanHighSchool]	.000						

Odds Ratios

seekcancerinfo_recode			Odds Ratio	95% Confidence Interval	
				Lower	Upper
flippedgender	Male vs. Female	Yes	1.460	1.152	1.850
flippededu	CollegeorMore vs **LessThanHighSchool**	Yes	1.393	0.854	2.270
	SomeCollege vs. **LessThanHighSchool**	Yes	0.921	0.544	1.562
	HighSchool vs. **LessThanHighSchool**	Yes	0.595	0.332	1.065

Overall Model Minus Intercept

df	Wald Chi-Square	Sig.
4.000	42.563	0.000

Gender

df	Wald Chi-Square	Sig.
1.000	10.048	0.002

Education Overall

df	Wald Chi-Square	Sig.
3.000	29.405	0.000

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 appear to be statistically more likely than men to have searched for cancer information.

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 flippedHighSchool flippedSomeCollege
flippedCollegeorMore
  /PLAN FILE='(sample.csaplan)'
  /MODEL flippedgender flippedHighSchool flippedSomeCollege flippedCollegeorMore
  /CUSTOM   Label = 'Overall model minus intercept'
   LMATRIX = flippedgender -1/2 1/2; flippedHighSchool -1/2 1/2; flippedSomeCollege -1/2 1/2;
flippedCollegeorMore -1/2 1/2
   /CUSTOM    Label = 'Gender'
  LMATRIX = flippedgender -1/2 1/2
   /CUSTOM    Label = 'Education overall'
    LMATRIX = flippedHighSchool -1/2 1/2; flippedSomeCollege -1/2 1/2; flippedCollegeorMore -1/2 1/2
  /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**	3384
	Invalid	120
	Total	3504
Population Size		242248876.895
Stage 1	**Strata**	2
	Units	100
Sampling Design Degrees of Freedom		98

Parameter Estimates

Parameter	Estimate	Std. Error	95% Confidence Interval		Hypothesis Test		
			Lower	Upper	t	df	Sig.
(Intercept)	3.029	0.156	2.720	3.338	19.455	98.000	0.000
[flippededu=CollegeorMore]	-0.906	0.150	-1.204	-0.609	-6.052	98.000	0.000
[flippededu=SomeCollege]	-0.485	0.150	-0.783	-0.187	-3.228	98.000	0.002
[flippededu=HighSchool]	-0.350	0.160	-0.667	-0.033	-2.190	98.000	0.031
[flippededu=LessthanHighSchool]	.000[b]						
[flippedgender=Female]	0.053	0.065	-0.075	0.181	0.820	98.000	0.414
[flippedgender=Male]	.000[b]						

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 who completed high school and those with a college degree or higher. We do not interpret the estimates for female because the corresponding p-value is greater than .05.

Overall Model Minus Intercept

df1	df2	Wald F	Sig.
4.000	95.000	24.180	0.000

Gender

df1	df2	Wald F	Sig.
1.000	98.000	0.672	0.414

Education Overall

df1	df2	Wald F	Sig.
3.000	96.000	30.458	0.000

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

Analyzing Data Using Stata

This section gives some Stata (Version 10.0 and higher) coding examples for common types of statistical analyses using HINTS 5, Cycle 2 data. Subsection 1 shows how to complete common analyses using replicate weights, and subsection 2 shows analyses using the Taylor Series linearization approach. For either approach, we begin by doing data management of the HINTS 5 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. 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 Stata **tabulate** command to verify proper coding.

```
use "file path\hints5_cycle2_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 == -6 | 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
```

Replicate Weights Variance Estimation Method

Declare survey design

Stata requires that the survey design be declared for the dataset globally before any analysis. The declared survey design will be applied to all future survey commands unless another survey design is declared. Other datasets 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 and are displayed in percentages. The options wald and noadjust together request the unadjusted Wald test for independence. Stata recommends the default Pearson test for independence. Other tests and statistics are also available; see the Stata website for more information: http://www.stata.com.

```
Jknife *: for cell counts

Number of strata   =         1              Number of obs      =         3,413
                                            Population size    =   243,588,173
                                            Replications       =            50
                                            Design df          =            49
```

Education	Gender: Male	Female	Total
Less tha	55.23122	44.76878	1.0e+02
	10.03954	7.79092	8.89075
12 years	48.93649	51.06351	1.0e+02
	22.24743	22.22502	22.23598
Some col	48.33664	51.66336	1.0e+02
	39.53172	40.45165	40.00170
College	47.74192	52.25808	1.0e+02
	28.18130	29.53241	28.87157
Total	48.91130	51.08870	1.0e+02
	1.0e+02	1.0e+02	1.0e+02

```
  Key:  row percentage
        column percentage

  Wald (Pearson):
    Unadjusted    chi2(3)          =   26.7021
    Unadjusted    F(3, 49)         =    8.9007     P = 0.0001
    Adjusted      F(3, 47)         =    8.5374     P = 0.0001
```

For the purposes of computing appropriate degrees of freedom for the estimator of the HINTS 5, 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 the user to specify the denominator degrees of freedom.

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 the categorical variable with the 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 redefines 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 the **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     =        2,717
                                            Population size   =  192,765,100
                                            Replications      =           50
                                            Design df         =           49
                                            F(   4,     46)   =        11.35
                                            Prob > F          =       0.0000
```

.

seekcancerinfo	Coef.	Jknife * Std. Err.	t	P>|t|	[95% Conf.	Interval]
_Igender_2	.3781961	.1341927	2.82	0.007	.1085258	.6478664
_Iedu_2	-.5199813	.2838837	-1.83	0.073	-1.090467	.0505044
_Iedu_3	-.0819317	.301154	-0.27	0.787	-.6871232	.5232599
_Iedu_4	.3311621	.2756464	1.20	0.235	-.22277	.8850943
_cons	-.1104163	.2858201	-0.39	0.701	-.6847932	.4639607

Unadjusted Wald test

```
 ( 1)  [seekcancerinfo]_Igender_2 = 0
 ( 2)  [seekcancerinfo]_Iedu_2 = 0
 ( 3)  [seekcancerinfo]_Iedu_3 = 0
 ( 4)  [seekcancerinfo]_Iedu_4 = 0
 ( 5)  [seekcancerinfo]_cons = 0

       F(  5,    49) =   11.82
            Prob > F =    0.0000
```

Unadjusted Wald test

```
 ( 1)  [seekcancerinfo]_Igender_2 = 0
 ( 2)  [seekcancerinfo]_Iedu_2 = 0
 ( 3)  [seekcancerinfo]_Iedu_3 = 0
 ( 4)  [seekcancerinfo]_Iedu_4 = 0

       F(  4,    49) =   12.09
            Prob > F =    0.0000
```

Unadjusted Wald test

```
 ( 1)  [seekcancerinfo]_Igender_2 = 0

       F(  1,    49) =    7.94
            Prob > F =    0.0069
```

Unadjusted Wald test

```
 ( 1)  [seekcancerinfo]_Iedu_2 = 0
 ( 2)  [seekcancerinfo]_Iedu_3 = 0
 ( 3)  [seekcancerinfo]_Iedu_4 = 0

       F(  3,    49) =   10.63
            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     =        2,717
                                              Population size   =  192,765,100
                                              Replications      =           50
                                              Design df         =           49
                                              F(   4,     46)   =        11.35
                                              Prob > F          =       0.0000
```

seekcancerinfo	Odds Ratio	Jknife * Std. Err.	t	P>|t|	[95% Conf.	Interval]
_Igender_2	1.459649	.1958742	2.82	0.007	1.114634	1.911458
_Iedu_2	.5945317	.1687779	-1.83	0.073	.3360595	1.051802
_Iedu_3	.9213349	.2774637	-0.27	0.787	.5030211	1.68752
_Iedu_4	1.392586	.3838612	1.20	0.235	.8002989	2.423213
_cons	.8954613	.2559408	-0.39	0.701	.5041945	1.59036

Note: _cons estimates baseline odds.

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, females appear to be statistically more inclined to search for cancer information compared with males.

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 Igender2 Iedu2 Iedu3 Iedu4 _cons, nosvyadjust test

Igender2 Iedu2 Iedu3 Iedu4, nosvyadjust

test Igender2, nosvyadjust
```

```
test Iedu2 Iedu3 Iedu4, 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     =        3,384
                                               Population size   =  242,248,877
                                               Replications      =           50
                                               Design df         =           49
                                               F(   4,     46)   =        28.77
                                               Prob > F          =       0.0000
                                               R-squared         =       0.0770

------------------------------------------------------------------------------
             |             Jknife *
generalhea~h |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Igender_2 |   .0529823   .0628516     0.84   0.403    -.0733226    .1792873
     _Iedu_2 |  -.3499955   .1778097    -1.97   0.055    -.7073174    .0073264
     _Iedu_3 |   -.484692   .1746447    -2.78   0.008    -.8356537   -.1337303
     _Iedu_4 |  -.9064306   .1719162    -5.27   0.000    -1.251909   -.5609521
       _cons |   3.028634   .1805151    16.78   0.000     2.665875    3.391392
------------------------------------------------------------------------------

   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) = 3786.16
               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) =   30.65
             Prob > F =    0.0000
```

```
Unadjusted Wald test

 ( 1)  _Igender_2 = 0

       F(  1,    49) =    0.71
            Prob > F =    0.4033

 Unadjusted Wald test

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

        F(  3,    49) =   35.08
             Prob > F =    0.0000
```

From the above table, it can be seen that, compared to those respondents with less than a high school education, those with some college or a college degree or higher have a significantly negative linear association with the outcome (i.e., better reported health), controlling for all variables in the model. We do not interpret the gender variable or those with a high school education because they are non-significant.

Taylor Series Linearization Variance Estimation Method

Declare survey design

Stata requires that the survey design be declared for the dataset globally before any analysis. The declared survey design will be applied to all future survey commands unless another survey design is declared. Other datasets that incorporate the final sample weight and stratum and cluster variables will utilize the same code.

```
* Declare survey design for the data set
  svyset var_cluster [pw=person_finwt0], strata(var_stratum)
```

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 and are displayed in percentages. The options wald and noadjust together request the unadjusted Wald test for independence. Stata recommends the default Pearson test for independence. Other tests and statistics are also available; see the Stata website for more information: http://www.stata.com.

```
(running tabulate on estimation sample)

Number of strata   =         2                  Number of obs     =         3,413
Number of PSUs     =       100                  Population size   =   243,588,173
                                                Design df         =            98
```

Education	Gender Male	Female	Total
Less tha	55.23122	44.76878	1.0e+02
	10.03954	7.79092	8.89075
12 years	48.93649	51.06351	1.0e+02
	22.24743	22.22502	22.23598
Some col	48.33664	51.66336	1.0e+02
	39.53172	40.45165	40.00170
College	47.74192	52.25808	1.0e+02
	28.18130	29.53241	28.87157
Total	48.91130	51.08870	1.0e+02
	1.0e+02	1.0e+02	1.0e+02

```
  Key:  row percentage
        column percentage

  Wald (Pearson):
    Unadjusted    chi2(3)         =    1.5554
    Unadjusted    F(3, 98)        =    0.5185     P = 0.6705
    Adjusted      F(3, 96)        =    0.5079     P = 0.6778
```

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 redefines 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)

Survey: Logistic regression

Number of strata   =         2                  Number of obs     =        2,717
Number of PSUs     =       100                  Population size   =  192,765,100
                                                Design df         =           98
                                                F(   4,     95)   =        10.31
                                                Prob > F          =      0.0000
```

seekcancerinfo	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
_Igender_2	.3781961	.1193114	3.17	0.002	.1414265	.6149658
_Iedu_2	-.5199813	.2938028	-1.77	0.080	-1.103023	.0630609
_Iedu_3	-.0819317	.2658775	-0.31	0.759	-.6095569	.4456935
_Iedu_4	.3311621	.2463198	1.34	0.182	-.1576515	.8199758
_cons	-.1104163	.2673143	-0.41	0.680	-.6408927	.4200602

```
Unadjusted Wald test

 ( 1)  [seekcancerinfo]_Igender_2 = 0
 ( 2)  [seekcancerinfo]_Iedu_2 = 0
 ( 3)  [seekcancerinfo]_Iedu_3 = 0
 ( 4)  [seekcancerinfo]_Iedu_4 = 0

       F(  4,    98) =   10.64
            Prob > F =    0.0000
```

```
   Unadjusted Wald test

     ( 1)  [seekcancerinfo]_Igender_2 = 0
     ( 2)  [seekcancerinfo]_Iedu_2 = 0
     ( 3)  [seekcancerinfo]_Iedu_3 = 0
     ( 4)  [seekcancerinfo]_Iedu_4 = 0
     ( 5)  [seekcancerinfo]_cons = 0

           F(  5,    98) =   10.28
                Prob > F =    0.0000

   Jnadjusted Wald test

    ( 1)  [seekcancerinfo]_Igender_2 = 0

          F(  1,    98) =   10.05
               Prob > F =    0.0020

 Unadjusted Wald test

  ( 1)  [seekcancerinfo]_Iedu_2 = 0
  ( 2)  [seekcancerinfo]_Iedu_3 = 0
  ( 3)  [seekcancerinfo]_Iedu_4 = 0

        F(  3,    98) =    9.80
             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)

Survey: Logistic regression

Number of strata   =         2                Number of obs     =        2,717
Number of PSUs     =       100                Population size   =  192,765,100
                                              Design df         =           98
                                              F(   4,     95)   =        10.31
                                              Prob > F          =       0.0000
```

seekcancerinfo	Odds Ratio	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
_Igender_2	1.459649	.1741528	3.17	0.002	1.151916	1.849593
_Iedu_2	.5945317	.1746751	-1.77	0.080	.3318662	1.065092
_Iedu_3	.9213349	.2449622	-0.31	0.759	.5435917	1.561573
_Iedu_4	1.392586	.3430214	1.34	0.182	.8541474	2.270445
_cons	.8954613	.2393696	-0.41	0.680	.5268219	1.522053

```
Note: _cons estimates baseline odds.
```

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, females appear to be statistically more inclined to search for cancer information compared with males.

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)

    Survey: Linear regression

    Number of strata   =         2                Number of obs     =        3,384
    Number of PSUs     =       100                Population size   =  242,248,877
                                                  Design df         =           98
                                                  F(   4,     95)   =        24.18
                                                  Prob > F          =       0.0000
                                                  R-squared         =       0.0770
```

generalhea~h	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
_Igender_2	.0529823	.0646125	0.82	0.414	-.0752392	.1812038
_Iedu_2	-.3499955	.1598487	-2.19	0.031	-.66721	-.032781
_Iedu_3	-.484692	.1501492	-3.23	0.002	-.7826582	-.1867258
_Iedu_4	-.9064306	.14978	-6.05	0.000	-1.203664	-.609197
_cons	3.028634	.1556709	19.46	0.000	2.71971	3.337558

```
Unadjusted Wald test

 ( 1)  _Igender_2 = 0
 ( 2)  _Iedu_2 = 0
 ( 3)  _Iedu_3 = 0
 ( 4)  _Iedu_4 = 0
 ( 5)  _cons = 0

       F(  5,    98) = 2792.82
            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,    98) =   24.94
            Prob > F =    0.0000


Unadjusted Wald test

 ( 1)  _Igender_2 = 0

       F(  1,    98) =    0.67
            Prob > F =    0.4142


Unadjusted Wald test

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

       F(  3,    98) =   31.09
            Prob > F =    0.0000
```

From the above table, it can be seen that, compared to those respondents with less than a high school education, those with a high school education, some college, or a college degree or higher have a significantly negative linear association with the outcome (i.e., better reported health), controlling for all variables in the model. We don't interpret the gender variable because it is non-significant.

Merging HINTS Survey Iterations

This section provides SAS and SPSS code respectively, to combine HINTS 5, Cycle 1 and HINTS 5, Cycle 2 survey iterations. The provided code will generate one final sample weight for population point estimates and 100 replicate weights to compute standard errors.

Merging HINTS 5, Cycle 1 and HINTS 5, Cycle 2 using SAS

This section provides SAS (Version 9.3 and higher) code for merging the HINTS 5, Cycle 1 and HINTS 5, Cycle 2 iterations. It first creates a temporary format for a new "survey" variable that will distinguish between the two iterations. The code then creates two temporary data files and adds the new "survey" variable to each dataset. Next, the two files are merged into one. It will match up variables that have the same name and format and create a merged data file (n = 6,789) that contains one final sample weight (for population point estimates) and 100 replicate weights (NWGT1 TO NWGT100; to compute standard errors).

```
/*FIRST CREATE THE FORMAT FOR THE SURVEY VARIABLE*/
proc format;
value survey
1="HINTS 5 CYCLE 1"

2="HINTS 5 CYCLE 2"

;
run;

/*************************************************************************/

/*CREATE TWO SEPARATE TEMPORARY DATA FILES THAT CONTAIN THE NEW 'SURVEY'
VARIABLE.

options fmtsearch=(HINTS5C1); /*PUT NAME OF LIBRARY WHERE HINTS 5
CYCLE 1 FORMATS ARE STORED*/

data tempHINTS5CYCLE1;

set HINTS5C1.hints5_cycle1_public; /*PUT NAME OF LIBRARY AND NAME OF EXISTING
HINTS 5 CYCLE 1 DATA FILE*/

survey=1;
format survey survey.;

run;

options fmtsearch=(HINTS5C2); /* PUT NAME OF LIBRARY WHERE HINTS 5 CYCLE 2
FORMATS ARE STORED*/

data tempHINTS5CYCLE2;

set HINTS5C2.hints5_cycle2_public; /*PUT NAME OF LIBRARY AND NAME OF EXISTING
HINTS 5 CYCLE 2 DATA FILE*/
```

```
survey=2;
format survey survey.;
run;

/*****************************************************************/
```

SAS Code to Set Up Final and Replicate Weights for the Replicate Variance Estimation Method

```
/*THIS CODE MERGES THE TWO TEMPORARY DATA SETS CREATED ABOVE. IT ALSO CREATES
ONE FINAL SAMPLE WEIGHT (NWGT0) AND 100 REPLICATE WEIGHTS (NWGT1 THRU
NWGT100)*/

data mergeHINTS5C1_HINTS5C2;
set tempHINTS5CYCLE1 tempHINTS5CYCLE2;

/*Create Replicate Weights for trend tests*/

**Replicate Weights;
array hints51wgts[50] person_finwt1-person_finwt50;
array hints52wgts[50] person_finwt1-person_finwt50;
array newWghts[100] nwgt1-nwgt100;

**Adjust Final And Replicate Weights;
if survey eq 1 then do i=1 to 50;   *HINTS 5 CYCLE 1;

nwgt0=person_finwt0;
newWghts[i]=hints51wgts[i];

newWghts[50+i]=person_finwt0;
end;

else if survey eq 2 then do i=1 to 50; *HINTS 5 CYCLE 2;

nwgt0=person_finwt0;
newWghts[50+i]=hints52wgts[i];

newWghts[i]=person_finwt0;
end;

run;
/*****************************************************************/

/*YOU CAN USE THE CODE BELOW TO RUN SIMPLE FREQUENCIES ON TWO COMMON
VARIABLES, 'SEEKHEALTHINFO' AND 'CHANGASKQUESTIONS'*/

/*SAS CODE*/
proc surveyfreq data = mergehints5c1_hints5c2 varmethod = jackknife;

weight nwgt0;
repweights nwgt1-nwgt100 / df = 98 jkcoefs = 0.98;

tables seekhealthinfo chanceaskquestions; run;
```

SAS Code to Merge HINTS 5, Cycle 1 and HINTS 5, Cycle 2 for the Taylor Series

Linearization Method

```
/*THIS CODE MERGES THE TWO HINTS DATA SETS CREATED ABOVE USING THE TAYLOR
SERIES LINEARZATION METHOD. PLEASE NOTE, THIS CODE IS BASED ON THE ASSUMPTION
THAT THE DATA SETS HAVE THE CORRECT VARIANCE CODES AND HHID VARIABLES MATCH*/
Data MergeHints5C1_Hints5c2;
set hints5c1.hints5_cycle1_public hints5c2.Hints5_cycle2_public;
run;

proc surveyfreq data = MergeHints5C1_Hints5c2 varmethod = TAYLOR;
strata VAR_STRATUM; cluster VAR_CLUSTER;
weight person_finwt0;
tables seekhealthinfo chanceaskquestions / row col;
run;
```

Merging HINTS 5, Cycle 1 and HINTS 5, Cycle 2 using SPSS

This section provides SPSS (Version 25) code for merging the HINTS 5, Cycle 1 and HINTS 5, Cycle 2 iterations. It first creates a temporary format for a new "survey" variable that will distinguish between the two iterations. The code then creates two temporary data files and adds the new "survey" variable to each dataset. Next, the two files are merged into one. It will match up variables that have the same name and format and create a merged data file (n = 6,789) that contains one final sample weight (NGWT0; for population point estimates) and 100 replicate weights (NWGT1 TO NWGT100; to compute standard errors).

For merging HINTS 5 Cycle 1 and HINTS 5 Cycle 2 on SPSS, you would need to follow similar steps already discussed earlier.

Note that a plan file is required to conduct analyses in SPSS. To create a plan file and subsequently conduct analyses, open the dataset "hints5_cycle2_public" and paste the follow syntax in the SPSS Syntax Editor:

```
* Encoding: UTF-8.
* Analysis Preparation Wizard.
* INSERT DATH OF PATH TO SAMPLE DESIGN FILE IN /PLAN FILE=.
CSPLAN ANALYSIS
  /PLAN FILE= 'H:\Hints5 Cycle 2 Data\HINTS 5 Cycle 2\SPSS
Data\Sample.csaplan'
  /PLANVARS ANALYSISWEIGHT=PERSON_FINWT0
  /SRSESTIMATOR TYPE=WOR
  /PRINT PLAN
  /DESIGN STRATA=VAR_STRATUM CLUSTER=VAR_CLUSTER
  /ESTIMATOR TYPE=WR.
```

Once you have your plan file, you can begin the merging process. You will want to open the two datasets "hints5_cycle1_public" and "hints5_cycle2_public" with SPSS. On one of the datasets you will navigate to the "Data" dropdown and select "Merge Files". You will be given the option to merge by cases or variables. Because we are merging two different cycles with mostly the same variables, we will want to select merge by "Add Cases". You will then select the other dataset that is open from the window that pops up and click continue. Ensure that the variables you need in the new merged dataset you are creating is in the "Variables in New Active Dataset" box. Once you have verified all your desired variables are in that box, click "OK".

```
DATASET ACTIVATE DataSet1.
ADD FILES /FILE=*
  /RENAME (AccessedFamRec_MyPwd AccessedFamRec_TheirPwd AccessFamilyMedRec AlcoholConditions_Cancer
    AlcoholConditions_Cholesterol AlcoholConditions_Diabetes AlcoholConditions_HeartDisease
    AlcoholConditions_LiverDisease AlcoholConditions_Overweight AlcoholIncreaseCancer
    AlcoholReduceHeart APP_REGION CancerAbilityToWork CancerDeniedCoverage CancerFatal
    CancerHurtFinances CancerMoreCommon CancerTx_Chemo CancerTx_Other CancerTx_Radiation
    CancerTx_Surgery CancerTxSummary CaOther_OS Caregiving_Family Caregiving_HoursPerWeek
    Caregiving_Other_OS CellPhone ChanceGetCancer ClinicalTrialCancerTx ConcernedQuality
    ConfidentGetHealthInf ConsiderQuit DiscussedClinicalTrial DiscussHPVVaccination12m DrTalkLungTest
    ElectInfoSafe Electronic_CompletedForms Electronic_HCPSearch Electronic_MadeAppts EmotionalSupport
    EverHadPSATest FreqWorryCancer Fruit Frustrated GeneticTestUse_Cat GeneticTestUse_DetermineMed
    GeneticTestUse_DeterminePass GeneticTestUse_DetermineRisk GeneticTestUse_DetermineTx
    HadTest_Ancestry HadTest_BRCA HadTest_Cat HadTest_CFCarrier HadTest_DNAFing HadTest_Lynch
    HadTest_None HadTest_NotSure HadTest_Other HadTest_Other_OS HadTest_Paternity
HCPEncourageOnlineRec
    HealthIns_Other_OS HeardDNATest HelpDailyChores HookahLessHarm HowLongFinishTreatment_Cat
    HowLongModerateExerciseHr HowLongModerateExerciseMn HPVMedicalTreatment HPVShotPrevent HPVSTD
    LotOfEffort MailSurveyTime_Hrs MailSurveyTime_Min NotAccessed_Other_OS OccupationStatus_OS
    OfferedAccessHCP2 OfferedAccessInsurer2 PersonID PhoneInHome RatherNotKnowChance RecordsOnline_Labs
    RecordsOnline_MakeAppt RecordsOnline_Meds RecordsOnline_MonitorHealth RecordsOnline_ViewResults
    SexualOrientation_OS SkinCancerHPExam SkinCancerSelfCheck SmokelessLessHarm Stratum
    StrongNeedHealthInfo StrongNeedHealthInfo_OS TalkHealthFriends TanningBed TooHardUnderstand
    TriedQuit TrustCharities TrustDoctor TrustFamily TrustGov TrustInternet TrustNewsMag TrustRadio
    TrustReligiousOrgs TrustTelevision UndergoCancerTreatment UnderstandOnlineMedRec UseMenuCalorieInfo
    Vegetables WhereUseInternet_GamingDevice WhereUseInternet_School=d0 d1 d2 d3 d4 d5 d6 d7 d8 d9 d10
    d11 d12 d13 d14 d15 d16 d17 d18 d19 d20 d21 d22 d23 d24 d25 d26 d27 d28 d29 d30 d31 d32 d33 d34 d35
    d36 d37 d38 d39 d40 d41 d42 d43 d44 d45 d46 d47 d48 d49 d50 d51 d52 d53 d54 d55 d56 d57 d58 d59 d60
    d61 d62 d63 d64 d65 d66 d67 d68 d69 d70 d71 d72 d73 d74 d75 d76 d77 d78 d79 d80 d81 d82 d83 d84 d85
    d86 d87 d88 d89 d90 d91 d92 d93 d94 d95 d96 d97 d98 d99 d100 d101 d102 d103 d104 d105 d106 d107
    d108 d109 d110 d111)
```

```
/FILE='DataSet2'
/RENAME (APP_REGION AverageCaloriesPerDay AverageCaloriesPerDay_DK AverageTimeSitting
  AvgDrinksPerWeek CalorieInfo_FewerCalories CalorieInfo_FewerItems CalorieInfo_LargerSizes
  CalorieInfo_MoreCalories CalorieInfo_MoreItems CalorieInfo_SmallerSizes CancerConcernedQuality
  CancerConfidentGetHealthInf CancerFrustrated CancerLotOfEffort CancerTooHardUnderstand
  CancerTrustCharities CancerTrustDoctor CancerTrustFamily CancerTrustGov CancerTrustInternet
  CancerTrustNewsMag CancerTrustRadio CancerTrustReligiousOrgs CancerTrustTelevision CaOther_OS
  Caregiver_AccessHelp Caregiver_Counseling Caregiver_MedTrain Caregiver_RespiteCare
  Caregiver_SupportGroup CaregiverTraining_Cat CaregiverTraining_Hotline CaregiverTraining_InPerson
  CaregiverTraining_OnlineVideo CaregiverTraining_ReadingMat CaregiverTraining_Virtual
  Caregiving_AccessMedRec Caregiving_AnotherFam Caregiving_ArrangeSvcs Caregiving_Bathing
  Caregiving_BedsChairs Caregiving_CommunicateHCP Caregiving_Dressing Caregiving_Feeding
  Caregiving_Finances Caregiving_HoursPerWeek2 Caregiving_Housework Caregiving_HowLong
  Caregiving_Incontinence Caregiving_MealPrep Caregiving_MedTasks Caregiving_Other_OS
  Caregiving_Professional Caregiving_Reside Caregiving_Shopping Caregiving_SpendTime
  Caregiving_Toilet Caregiving_Transportation CaregivingActivities_Cat CaregivingMedAct_Cat
  ConfidentFamilyHistory DrinkDaysPerWeek DrinksPerDay Electronic_LookedAssistance EmotionalSupport2
  EverOfferedAccessRec EverTestedColonCa FamiliarFamilyCancer FamilyCancer_Brother FamilyCancer_Cat
  FamilyCancer_Children FamilyCancer_Father FamilyCancer_HCP FamilyCancer_Mother FamilyCancer_None
  FamilyCancer_OthFam FamilyCancer_Sister FORM_NAME FreqWorryCancerAgain FreqWorryCancerNoDx
  HaveDevice_Cat HaveDevice_None HCPAdvisedLimitingSun HealthIns_Other_OS HelpDailyChores2
  HelpPreparingMeals HelpRunErrands HelpTransportDoctor HowLongModerateExerciseMinutes ImagineCancer
  ImagineCancerAgain InfluenceCancer_EatingHealthy InfluenceCancer_Obesity
  InfluenceCancer_RegExercise KnowledgePalliativeCare MailSurveyTimeHrs MailSurveyTimeMin
  NotAccessed_Other_OS NoticeCalorieInfoOnMenu OccupationStatus_OS PCGoal_HelpFamCope
  PCGoal_ManageSymptoms PCGoal_MoreTime PCGoal_SocEmotSupport PCHospiceCare PCMeansGivingUp
  PCObligatedToInform PCStopTreatments PCStrongNeedInfo PCThinkDeath PCTrustInfo PersonID
  SeenFederalCourtTobaccoMessages SexualOrientation_OS SpendTimeInSunTanning Stratum
  StrongNeedCancerInfo StrongNeedCancerInfo_OS SunEffectAfter1Hour TalkHealthFriends2
  TimesUsedTanningBed TobaccoMessages_Addictiveness TobaccoMessages_Cat
  TobaccoMessages_EnhanceDelivery TobaccoMessages_HESecondhand TobaccoMessages_HESmoking
  TobaccoMessages_LowTarLight UnderstandCalorieInfo VAR_CLUSTER VAR_STRATUM WhoOffered_Cat
```

```
  WhoOffered_HCP WhoOffered_Insurer WhoOffered_Other WhoOffered_Other_OS=d112 d113 d114 d115 d116
  d117 d118 d119 d120 d121 d122 d123 d124 d125 d126 d127 d128 d129 d130 d131 d132 d133 d134 d135 d136
  d137 d138 d139 d140 d141 d142 d143 d144 d145 d146 d147 d148 d149 d150 d151 d152 d153 d154 d155 d156
  d157 d158 d159 d160 d161 d162 d163 d164 d165 d166 d167 d168 d169 d170 d171 d172 d173 d174 d175 d176
  d177 d178 d179 d180 d181 d182 d183 d184 d185 d186 d187 d188 d189 d190 d191 d192 d193 d194 d195 d196
  d197 d198 d199 d200 d201 d202 d203 d204 d205 d206 d207 d208 d209 d210 d211 d212 d213 d214 d215 d216
  d217 d218 d219 d220 d221 d222 d223 d224 d225 d226 d227 d228 d229 d230 d231 d232 d233 d234 d235 d236
  d237 d238 d239 d240 d241 d242 d243 d244 d245 d246 d247)
 /DROP=d0 d1 d2 d3 d4 d5 d6 d7 d8 d9 d10 d11 d12 d13 d14 d15 d16 d17 d18 d19 d20 d21 d22 d23 d24
  d25 d26 d27 d28 d29 d30 d31 d32 d33 d34 d35 d36 d37 d38 d39 d40 d41 d42 d43 d44 d45 d46 d47 d48 d49
  d50 d51 d52 d53 d54 d55 d56 d57 d58 d59 d60 d61 d62 d63 d64 d65 d66 d67 d68 d69 d70 d71 d72 d73 d74
  d75 d76 d77 d78 d79 d80 d81 d82 d83 d84 d85 d86 d87 d88 d89 d90 d91 d92 d93 d94 d95 d96 d97 d98 d99
  d100 d101 d102 d103 d104 d105 d106 d107 d108 d109 d110 d111 d112 d113 d114 d115 d116 d117 d118 d119
  d120 d121 d122 d123 d124 d125 d126 d127 d128 d129 d130 d131 d132 d133 d134 d135 d136 d137 d138 d139
  d140 d141 d142 d143 d144 d145 d146 d147 d148 d149 d150 d151 d152 d153 d154 d155 d156 d157 d158 d159
  d160 d161 d162 d163 d164 d165 d166 d167 d168 d169 d170 d171 d172 d173 d174 d175 d176 d177 d178 d179
  d180 d181 d182 d183 d184 d185 d186 d187 d188 d189 d190 d191 d192 d193 d194 d195 d196 d197 d198 d199
  d200 d201 d202 d203 d204 d205 d206 d207 d208 d209 d210 d211 d212 d213 d214 d215 d216 d217 d218 d219
  d220 d221 d222 d223 d224 d225 d226 d227 d228 d229 d230 d231 d232 d233 d234 d235 d236 d237 d238 d239
  d240 d241 d242 d243 d244 d245 d246 d247.
EXECUTE.
/*******************************************************/
/*YOU CAN USE THE CODE BELOW TO RUN SIMPLE FREQUENCIES ON TWO COMMON
VARIABLES, 'LOTOFEFFORT' AND 'TRUSTDOCTOR'*/
/*SPSS CODE*/
CSTABULATE
  /PLAN FILE='H:\Hints5 Cycle 2 Data\HINTS 5 Cycle 2\SPSS
Data\Sample.csaplan' /*INSERT PATH OF DATA SET HERE*/
  /TABLES VARIABLES=TrustDoctor LotOfEffort
  /CELLS POPSIZE TABLEPCT
  /STATISTICS SE COUNT
  /MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

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 Wiley & Sons.

Kott, P.S. (2009). Calibration Weighting: Combining Probability Samples and Linear Prediction Models. Chapter 25 in Pfeffermann, D. and Rao, C.R. (eds.) *Handbook of Statistics Vol. 29B: Sample Surveys: Inference and Analysis*. Elsevier: Amsterdam

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.

Wolter, K. (2007). *Introduction to Variance Estimation*. 2nd edition. Springer-Verlag: New York