OVERVIEW OF THE HINTS 7 SURVEY (2024) AND DATA ANALYSIS RECOMMENDATIONS

May 2025

CONTENTS

Overview of HINTS

The Health Information National Trends Survey (HINTS) is a nationally representative, cross-sectional 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); HINTS 4, Cycle 2 (2012); HINTS 4, Cycle 3 (2013); HINTS 4, Cycle 4 (2014); HINTS-FDA, Cycle 1 (2015); HINTS-FDA, Cycle 2 (2017); HINTS 5, Cycle 1 (2017); and HINTS 5, Cycle 2 (2018); HINTS 5, Cycle 3 (2019); HINTS 5 Cycle 4 (2020); and HINTS 6 (2022).

HINTS 7

Starting with HINTS 6, data is collected on a biennial basis. For HINTS 7, a multi-mode (with push to web) survey was implemented using paper and web modes. For more extensive background about the HINTS program and previous data collection efforts, see Finney Rutten, et al. (2012).

Methodology

Data collection for HINTS 7 started on March 25, 2024, and concluded on September 16, 2024. HINTS 7 included two embedded methodological experiments intended to improve response rates and data quality. For an incentive experiment, a portion of respondents in the High Minority strata (the treatment group) were offered an incentive regardless of mode: they were offered a $10 incentive to complete the survey by paper or a $20 incentive to complete the survey online. The aim was to assess whether these additional incentive offerings would increase the response rate for racial and ethnic minoritized respondents. Respondents in the control group were offered $10 to complete the survey online and no additional incentive to complete the survey by paper. In the respondent commitment experiment, a portion of sampled households (n=7,200) were randomized to receive a statement at the beginning of their survey asking them to make a commitment to provide complete and accurate information. Based on recent literature (Vanette 2016, Conrad et al 2017, Hibben et al 2020) this study examined whether the inclusion of the statement would result in higher quality data without affecting the response rate. Results will be shared in future publications. For more information, see the HINTS 7 Methodology Report found here.

Although HINTS 7 included the standard $2 pre-incentive sent to all households, the incentive was shown in the envelope window for the HINTS 7 administration rather than being hidden from view in a solid envelope. Based on recent literature (DeBell, 2022; Sherr and Wells, 2021; Zhang et al., 2023), it was anticipated that having the cash incentive visible would increase the chances of the envelope being opened and therefore increase the response rate. Because of an unexpectedly low response to the first three HINTS mailings (per the mailing protocol further described below), HINTS 7 included a fourth mailing to and an increased incentive for a subsample of non-respondents (n=13,055), selected using a systematic sampling approach. This extra mailing, sent out on August 5, 2024, included an increased incentive payment of $30 for completion in any mode. The protocol for this additional mailing is described in Chapter 3 of the Methodology Report.

Both conditions used the same sampling frame provided by Marketing Systems Group (MSG) of addresses in the United States. All addresses were grouped into one of four strata; high and low minority (similar to previous HINTS iterations) by rural and urban area.

The mailing protocol for HINTS 7 followed a modified Dillman approach (Dillman, et al., 2009) with all selected households receiving a total of four mailings: an initial mailing, a reminder postcard, and two follow-up mailings. Potential Spanish households received contact materials in English and Spanish and both English and Spanish surveys. Respondents were able to toggle the web survey to complete it in either English or Spanish. English-only households that requested a Spanish survey received a Spanish paper survey in subsequent mailings.

One adult within each sampled household was selected 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. Refer to the HINTS 7 Methodology Report for more extensive information about the sampling and weighting procedures.

Sample Size and Response Rates

The final HINTS 7 sample consists of 7,278 respondents. Note that 70 of these respondents were considered partial completers who did not answer the entire survey. A questionnaire was 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 4 (RR4) formula. The overall household response rate using the next-birthday method was 27.3%. See the Methodology Report for more information.

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, R 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 compute the correct variance estimates when analyzing survey data that employ a complex sampling method, such as HINTS. The issue is that the standard errors in your analyses will most likely be inaccurate if you do not take into account the sampling procedure; your p-values may be smaller (or larger) than they "should" be and you are more likely to make an error in interpretation. HINTS data contain jackknife replicate weights to compute the correct variance estimates. Statistical programs like SAS, Stata, R, and Mplus can incorporate the replicate weights found in the HINTS database.

Note that the SPSS dataset will contain variance codes that will allow for inferential statistical testing using Taylor Series Linearization along with the Complex Samples module found in SPSS. 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 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 certain analyses will tend to result in unprecise estimates. Methods for obtaining confidence intervals for small proportions or limited degrees of freedom for small populations are described in Korn and Graubard’s Analysis of Health Surveys (1999; pp. 64-68). Related to this, beginning with HINTS 5 Cycle 4 (2020), the HINTS program has implemented data suppression thresholds wherein some variables with cells that have <25 responses are either collapsed, recoded to missing, or deleted/suppressed entirely. Thresholds were determined based solely on respondent disclosure risk, but small cell sizes also have implications for precision.

Preferred Reporting Items for Complex Sample Survey Analysis (PRICSSA)

HINTS encourages data analysts to use the PRICSSA checklist (Seidenberg et.al., 2023) when reporting on their analyses. PRICSSA is modeled after checklists like Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) and was created to increase consistency in reporting, reduce analytic errors, and increase analytic transparency and reproducibility. For more information when using HINTS, see: https://hints.cancer.gov/data/pricssa.aspx.

Important Analytic Variables in the Database

Refer to the HINTS 7 Methodology Report for more information regarding the weighting and stratification variables listed below.

Note that estimates from the 2023 American Community Survey (ACS) of the U.S. Census Bureau were used to calibrate the HINTS 7 control totals with the following variables: age, birth sex, education, marital status, race, ethnicity, and census region. In addition, the 2023 National Health Interview Survey (NHIS) was used to calibrate HINTS 7 data control totals regarding percent with health insurance and the 2023 National Center for Health Statistics (National Center for Health Statistics, Interactive Summary Health Statistics for Adults-2023) was used for percent ever been diagnosed with cancer.

Final Sample and Replicate Weights for Jackknife Replication

Included with the data are statistical weights. Below we have provided a brief description of these different weights, both final sample weights (to calculate population-level point estimates), and replicate weights (to calculate variance estimates).

PERSON_FINWT0: Final sample weight used to calculate population estimates for the combined sample.

PERSON_FINWT1 through PERSON_FINTW50: Fifty replicate weights that can be used to calculate accurate standard error of estimates using the jackknife replication method for the combined sample.

Stratum/Cluster Variables and Final Sample Weights for Taylor Series Linearization Methods

VAR_STRATUM: This variable identifies the first-stage sampling stratum of a HINTS sample for a given data collection cycle. For HINTS 7, this variable incorporates the two sets of strata used for sampling. 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 four values: high and low minority by rural and urban area.

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

Other Variables

FORMTYPE: This variable codes for whether the respondent completed the survey using the self-administered paper survey or on the web.

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

LANGUAGE_FLAG: This variable codes for the language the survey was completed in (English, Spanish, or Mixed (web cases only)).

INCOMERANGES_IMP: This is the income variable (INCOMERANGES) imputed for missing data. To impute 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 (R8), Race/Ethnicity (RaceEthn) (standard recode from R9 and R10), and current occupational status (R5).

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 7 dataset includes 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 reflects 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 7 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 or group of HINTS data, the proper denominator degrees of

freedom (ddf) is 49. 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 or groups, and 49*k as the appropriate ddf. Analysts who are merging two iterations of data and making comparisons should adjust the ddf to be 98 (49*2), etc.

Taylor Series: The HINTS 7 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, 196, is based on 50 PSUs in each of the four sampling strata (#psus - #strata = 50*4 – 4 = 196).

Statistical Software Example Code

This section provides some coding examples using SAS, SPSS, Stata, and R for common types of statistical analyses using HINTS 7 data.

For SAS, Stata, and R, 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-level point 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 7 data, the concepts used here are generally applicable to other types of analyses. We will consider an analysis that includes sex, education level (edu as a new variable) and two questions that are specific to the HINTS data: seekcancerinfo & generalhealth.

Analyzing Data Using SAS

Prior to using the HINTS 7 SAS data, it is important to apply the SAS formats. To do this, see the "How to Format the HINTS 7 SAS Dataset" document included in the data download.

1. Download all HINTS 7 documents to a folder on your computer. This should be the same folder where you create the SAS library in step #2.

2. Using SAS, create a permanent library to point to the folder where your data has been downloaded to (if you use the New Library icon, be sure to select, "enable at startup").

3. Open the SAS program *"HINTS7_Formats.sas"*

4. Change the file location specification in the "library" statement to be the name of the library created in step 2.

5. Run the program *"HINTS7_Formats.sas"* to create a permanent SAS format library that is used to analyze the HINTS dataset.

6. Open the SAS program "*HINTS7_Format_Assignments.sas"*

7. 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.

8. Run the program "*HINTS7_Format_Assignments.sas"* to create the formatted SAS data set.

Note:

1) Make sure to run the program *"HINTS7_Formats.sas"* BEFORE you run "*HINTS7_Format_Assignments.sas*" to create the formatted HINTS dataset.

2) If you are getting an error statement saying that SAS is unable to find the formats, make sure you have run the OPTIONS statement that includes the correct library name where the formats can be found.

This section gives some SAS (Version 9.4 and higher) coding examples for common types of statistical analyses using HINTS 7 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 7 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. 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.

SAS Data Management Code: Recoding Variables and Creating and Applying New Formats

```
*This is used to call up the formats, substitute your library name in the
parentheses;
options fmtsearch=(hints7);

*First create some temporary formats;

proc format;

        Value Sexf
        1 = "Female"
        2 = "Male"
        3 = "Don't Know";

        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 hints7;
      set hints7.hints7_public;

    /*Recode negative values to missing*/
    if birthsex = 1 then sex = 1;
    if birthsex = 2 then sex = 2;
    if birthsex = 3 then sex = 3;
    if birthsex in (-9, -7) then sex = . ;

    /*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 in (-9, -7) then edu = .;

    /*Recode seekcancerinfo to 0- 1 format for proc surveylogistic
    procedure, and negative values to missing */
    if seekcancerinfo = 2 then seekcancerinfo = 0;
    if seekcancerinfo in (-9, -7, -6, -2, -1) then seekcancerinfo = .;

    /*Recode negative values to missing for proc surveyreg procedure*/

    if generalhealth in (-5, -9, -7) then generalhealth = .;

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

    run;
```

SAS Replicate Weights Variance Estimation Method

Frequency Table and Chi-Square Test

We are now ready to begin using SAS 9.4 to examine the relationships among these variables. Using **PROC SURVEYFREQ**, we will first generate a cross-frequency table of education by sex, 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 = hints7 varmethod = jackknife;
      weight person_finwt0;
      repweights person_FINWT1-person_FINWT50 / df = 49 jkcoefs = 0.98;
      tables edu*sex / row col chisq(secondorder);
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 chisq requests Rao-Scott chi-square test for independence and the (secondorder) requests the second order effects. Other tests and statistics are also available; see the SAS Product Documentation Site for more information.

For the purposes of computing appropriate degrees of freedom for the estimator of the HINTS 7 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	HINTS7
Number of Replicates	50

Edu	sex	Frequency	Percent	Std Err of Percent	Row Percent	Std Err of Row Percent	Column Percent	Std Err of Col Percent
Less than high school	Female	267	2.6676	0.2689	43.6038	3.3087	5.5058	0.5567
	Male	161	3.3304	0.3410	54.4371	3.4000	6.5517	0.6644
	Don't Know	7	0.1199	0.0650	1.9591	1.0664	16.7139	10.5682
	Total	435	6.1178	0.4531	100			
12 years or completed high school	Female	702	12.0401	0.5389	56.3760	1.8378	24.8500	1.0828
	Male	408	9.1735	0.4793	42.9539	1.7724	18.0468	0.9120
	Don't Know	8	0.1431	0.0730	0.6702	0.3431	19.9597	11.8449
	Total	1118	21.3567	0.6623	100			
Some college	Female	1145	18.0797	0.4943	46.7855	0.9716	37.3155	0.9964
	Male	778	20.2069	0.5214	52.2901	0.9786	39.7524	0.9914
	Don't Know	10	0.3572	0.2065	0.9244	0.5327	49.8189	20.3534
	Total	1933	38.6439	0.7374	100			
College graduate or higher	Female	1879	15.6636	0.1782	46.2304	0.3946	32.3287	0.3490
	Male	1285	18.1211	0.1935	53.4837	0.4111	35.6491	0.3902
	Don't Know	16	0.0969	0.0376	0.2859	0.1109	13.5076	6.9279
	Total	3180	33.8816	0.2528	100			
Total	Female	3993	48.4510	0.3055			100	
	Male	2632	50.8319	0.3912			100	
	Don't Know	41	0.7171	0.2203			100	
	Total	6666	100					

Frequency Missing = 612

Rao-Scott Chi-Square Test	
Pearson Chi-Square	63.0493
Design Correction	3.2251
First-Order Chi-Square	19.5493
Second-Order Chi-Square	11.0634
DF	3.40
Pr > ChiSq	0.0161
F Value	3.2582
Num DF	3.40
Den DF	166.38
Pr > F	0.0185
Sample Size = 6666	

The row percentages above show that a higher weighted proportion of college graduates in the sample are male (53.5%) than female (46.2%). Respondents with less than a high school diploma include fewer females (44%) than males (54%). 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 sex and education on
SeekCancerInfo*/
proc surveylogistic data= hints7 varmethod=jackknife;
      weight person_FINWT0;
      repweights person_FINWT1-person_FINWT50 / df=49 jkcoefs=0.98;
      class edu (ref="Less than high school")
            sex (ref="Female")/param=REF;
      model seekcancerinfo (descending) = sex edu /tech=newton
      xconv=1e-8 CLPARM EXPB;
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 “Female” is the reference group for sex 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	HINTS7
Number of Replicates	50

Type 3 Analysis of Effects				
Effect	**F Value**	**Num DF**	**Den DF**	**Pr> F**
Sex	6.49	2	49	0.0032
Education	34.99	3	49	<.0001

Analysis of Maximum Likelihood Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**T value**	**Pr > |t|**	**95% confidence limits**	
Intercept	49	-0.9576	0.2539	-3.77	0.0004	-1.4677	-0.4475
Don't Know	49	-0.9062	0.8549	-1.06	0.2943	-2.6242	0.8117
Male	49	-0.2826	0.0848	-3.33	0.0016	-0.4530	-0.1122
12 years or completed high school	49	0.7790	0.2926	2.66	0.0105	0.1909	1.3670
Some College	49	1.0618	0.2590	4.10	0.0002	0.5412	1.5824
College graduate or higher	49	1.7123	0.2661	6.44	<.0001	1.1776	2.2470

Odds Ratio Estimates

Effect	Point Estimate	95% Confidence Limits	
Don't Know vs. Female	0.404	0.072	2.252
Male vs. Female	0.754	0.636	0.894
12 years or completed high school vs Less than high school	2.179	1.210	3.924
Some College vs Less than high school	2.892	1.718	4.867
College graduate or higher vs Less than high school	5.542	3.247	9.460

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, "Estimate" column). According to this model, males appear to have 0.75 times lower odds than females 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 sex and education on
GeneralHealth*/

proc surveyreg data= hints7 varmethod=jackknife;
      weight PERSON_FINWT0;
      repweights PERSON_FINWT1-PERSON_FINWT50 / df=49 jkcoefs=0.98;
      class edu (ref="Less than high school")
            sex (ref="Female");
      model generalhealth = edu sex /solution;
run;
```

Variance Estimation	
Method	Jackknife
Replicate Weights	HINTS7
Number of Replicates	50

Estimated Regression of Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	3.1479480	0.10652325	29.55	<.0001
12 years or completed high school	-0.2787725	0.11940504	-2.33	0.0237
Some College	-0.3935888	0.12287118	-3.20	0.0024
College graduate or higher	-0.7044893	0.12233856	-5.76	<.0001
Don't Know	0.0418274	0.35868316	0.12	0.9076
Male	-0.1192809	0.04440035	-2.69	0.0098

The table labeled Estimated Regression of Coefficients shows that respondents with a high school education, some college, and completed college reported better general health than those with less than a high school education 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. Compared to females, males have significantly better health.

Tests of Model Effects

Contrast	Num DF	F Value	Pr > F
Model	5	21.90	<.0001
Intercept	1	519.62	<.0001
Education	3	26.22	<.0001
Sex	2	3.61	0.0344

The table labeled Test of Model Effects also shows that the association between sex and general health and the association between education and general health are both significant.

SAS Taylor Series Linearization Variance Estimation Method

Frequency Table and Chi-Square Test

We are now ready to begin using SAS 9.4 to examine the relationships among these variables. Using **PROC SURVEYFREQ**, we will first generate a cross-frequency table of education by sex, 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 = hints7 varmethod = TAYLOR;
      strata VAR_STRATUM;
      cluster VAR_CLUSTER;
      weight person_finwt0;
      tables edu*sex / row col chisq(secondorder);
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 *chisq* requests Rao-Scott chi-square test for independence and the (*secondorder*) requests the second order effects. Other tests and statistics are also available; see the SAS Product Documentation Site for more information.

Data Summary	
Number of Strata	4
Number of Clusters	200
Number of Observations	7278
Sum of Weights	262266460

edu	sex	Frequency	Percent	Std Err of Percent	Row Percent	Std Err of Row Percent	Column Percent	Std Err of Col Percent
Less than high school	Female	267	2.6676	0.2866	43.6038	4.2699	5.5058	0.5955
	Male	161	3.3304	0.4197	54.4371	4.3851	6.5517	0.8238
	Don't Know	7	0.1199	0.0659	1.9591	1.0765	16.7139	9.2718
	Total	435	6.1178	0.4726	100			
12 years or completed high school	Female	702	12.0401	0.8046	56.3760	2.7750	24.8500	1.5013
	Male	408	9.1735	0.7269	42.9539	2.7893	18.0468	1.3300
	Don't Know	8	0.1431	0.0748	0.6702	0.3528	19.9597	10.3022
	Total	1118	21.3567	0.9578	100			
Some college	Female	1145	18.0797	0.8704	46.7855	1.9409	37.3155	1.4055
	Male	778	20.2069	0.9658	52.2901	1.9547	39.7524	1.5935
	Don't Know	10	1.5935	0.1978	0.9244	0.5100	49.8189	15.9125
	Total	1933	38.6439	1.0674	100			
College graduate or higher	Female	1879	15.6636	0.6689	46.2304	1.5287	32.3287	1.3755
	Male	1285	18.1211	0.6946	53.4837	1.5229	35.6491	1.2682
	Don't Know	16	0.0969	0.0363	0.2859	0.1066	13.5076	6.0807
	Total	3180	33.8816	0.8921	100			
Total	Female	3993	48.4510	1.1197			100	
	Male	2632	50.8319	1.1359			100	
	Don't Know	41	0.7171	0.2197				
	Total	6666	100					

Frequency Missing = 612

Rao-Scott Chi-Square Test	
Pearson Chi-Square	63.0493
Design Correction	4.1084
First-Order Chi-Square	15.3463
Second-Order Chi-Square	12.2270
DF	4.78
Pr > ChiSq	0.0276
F Value	2.5577
Num DF	4.78
Den DF	936.96
Pr > F	0.0282
Sample Size = **6666**	

The row percentages above show that a higher weighted proportion of college graduates in the sample are males (53.5%) than females (46.2%). Respondents with less than a high school diploma include fewer females (43.6%) than males (54.4%). The Chi-squared test of independence statistic and associated p value suggest that there is a significant difference between the distributions of educational attainment for these two groups.

In some cases, the results of these tests based on Taylor Series linearization may contradict the results using replication shown in the previous section (in this case, the distributions of educational attainment between males and females were determined to be statistically different using the replication method as well). Since both education and sex are variables used in the raking process as part of the HINTS weighting procedure, the variance estimation method used can affect the outcome of a statistical test. As a result, the standard errors based on replication are much smaller than those based on Taylor Series linearization, which in turn may result in significant differences using the replication method but not in the Taylor Series linearization method.

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 sex and education on
SeekCancerInfo*/
proc surveylogistic data= hints7 varmethod=TAYLOR;
      strata VAR_STRATUM;
      cluster VAR_CLUSTER;
      weight person_FINWT0;
      class edu (ref="Less than high school") sex
            (ref="Female")/param=REF;
      model seekcancerinfo (descending) = sex edu /tech=newton
      xconv=1e-8 CLPARM EXPB;
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 sex 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	
Methods	Taylor Series
Variance Adjustment	Degrees of Freedom (DF)

Type 3 Analysis of Effects				
Effect	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Sex	6.08	2	195	0.0027
Education	35.81	3	194	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	196	-0.9576	0.2419	-3.96	0.0001
Don't Know	196	-0.9062	0.6313	-1.44	0.1527
Male	196	-0.2826	0.0879	-3.22	0.0015
12 years or completed high school	196	0.7790	0.2658	2.93	0.0038
Some College	196	1.0618	0.2527	4.20	<.0001
College graduate or higher	196	1.7123	0.2500	6.85	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Confidence Limits	
Don't Know vs. Female	0.404	0.116	1.403
Male vs. Female	0.754	0.634	0.897
12 years or completed high school vs Less than High School	2.179	1.290	3.681
Some College vs Less than High School	2.892	1.757	4.760
College graduate or higher vs Less than High School	5.542	3.385	9.073

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, males appear to have statistically lower odds than females 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 sex and education on
GeneralHealth*/
proc surveyreg data= hints7 varmethod=TAYLOR;
      strata VAR_STRATUM;
      cluster VAR_CLUSTER;
      weight person_FINWT0;
      class edu (ref="Less than high school") sex
      (ref="Female");
      model generalhealth = edu sex/solution;
run;
```

Estimated Regression of Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	3.1479480	0.10088919	31.20	<.0001
Don't Know	0.0418274	0.27356300	0.15	0.8786
Male	-0.1192809	0.04349699	-2.74	0.0067
12 years or completed high school	-0.2787725	0.11221682	-2.48	0.0138
Some College	-0.3935888	0.11588572	-3.40	0.0008
College graduate or higher	-0.7044893	0.11677164	-6.03	<.0001

Compared to those respondents with less than a high school education, those who have a high school education, completed some college, and are college graduates on average reported significantly better general health (i.e., the negative beta coefficient indicates that the average health score is lower among those with a high school education, some college, and college graduates because the health variable is coded such that lower scores correspond to better health), controlling for all variables in the model. In this model, males also have significantly better health than females.

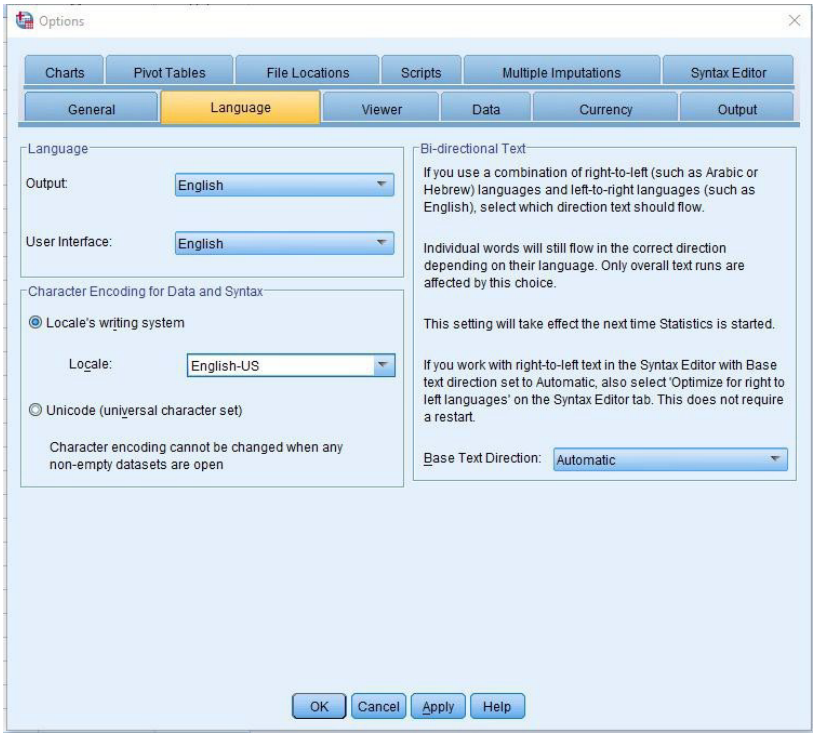
Tests of Model Effects

Contrast	Num DF	F Value	Pr > F
Overall model	5	21.23	<.0001
Intercept	1	840.40	<.0001
Sex	2	3.77	0.0249
Education	3	25.68	<.0001

From the above table, we can see that both sex and education are significantly associated with general health, adjusting for all variables in the model.

Analyzing Data Using SPSS—Taylor Series

Prior to opening the HINTS 7 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 22 and higher) coding examples for common types of statistical analyses using HINTS 7 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 (the final weight for the composite sample, no group differences found), VAR_STRATUM is the stratum variable, and VAR_CLUSTER is the cluster variable.

The subcommand SRSESTIMATOR specifies the variance estimator under the simple random sampling assumption. The default value is WR (with replacement), and it includes the finite population correction in the variance computation. The subcommand PRINT is used to control output from CSPLAN, and the syntax PLAN means to display a summary of plan specifications. The subcommand DESIGN with keyword STRATA identifies the sampling stratification variable, and the keyword CLUSTER 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 7 data in a SPSS RECODE step. We first decided to exclude all “Missing data (Not Ascertained)” and “Multiple responses selected in error” responses from the analyses. By setting these values to missing (SYSMIS), SPSS will exclude these responses from procedures where these variables are specifically accessed. For logistic regression modeling in the CSLOGISTIC procedure, SPSS by default always uses the last (highest) level of category of the covariates as the reference, similar to SAS. Users in SPSS cannot define the reference category by themselves unless they reorder the categories to create the desired value as the reference, such as using reverse coding (see example below). To make SPSS results comparable with SAS, we reverse coded the variables in SPSS. When recoding existing variables, it is generally recommended to create new variables, rather than over-writing the existing variables. Note: New variables should always be compared to original source variables in a SPSS CROSSTABS procedure to verify proper coding.

```
*Recode negative values to missing.
DATASET ACTIVATE DataSet1.
RECODE BirthSex (1=1) (2=2) (3=3) (ELSE=SYSMIS) INTO sex.
VARIABLE LABELS sex 'sex'. 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 sex (1=3) (2=2) (3=1)  (ELSE=Copy) INTO flippedsex.
VARIABLE LABELS flippedsex 'flippedsex'.
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 sex 1 "Female" 2 Male" 3
“Don’t Know”.
VALUE LABELS flippedsex 1 "Don’t Know" 2 "Male" 3 “Female”.
VALUE LABELS edu 1 "Less than high school" 2 "12 years or completed high school" 3 "Some college" 4
"College graduate or higher".
VALUE LABELS flippededu 4 "Less than high school" 3 "12 years or completed high school" 2 "Some college" 1
"College graduate or higher".
VALUE LABELS seekcancerinfo_recode 1 "Yes" 0 "No".
VALUE LABELS genhealth_recode 1 "Excellent" 2 "Very good" 3 "Good" 4 "Fair" 5 "Poor".
```

Frequency Table and Chi-Square Test

We are now ready to begin using SPSS v22 to examine the relationships among these variables. Using **CSTABULATE**, we will first generate a cross-frequency table of education by sex. Note that we specify the file that contains the sample design specification using the subcommand PLAN. This syntax is consistent for all procedures. Other analyses using the same sample design will follow a similar syntax.

```
* Complex Samples Crosstabs.
CSTABULATE
/PLAN FILE=”(plan filename)”
/TABLES VARIABLES=edu BY sex
/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.

			Sex			
Edu			Female	Male	Don't Know	Total
Less than high school	Population Size	Estimate	6487616.216	8099459.901	291482.165	14878558.283
		Standard Error	694767.519	1044969.844	160112.620	1185099.604
		Unweighted Count	267	161	7	435
	% within edu	Estimate	43.6%	54.4%	2.0%	100.0%
		Standard Error	4.3%	4.4%	1.1%	0.0%
		Unweighted Count	267	161	7	435
	% within sex	Estimate	5.5%	6.6%	16.7%	6.1%
		Standard Error	0.6%	0.8%	9.3%	0.5%
		Unweighted Count	267	161	7	435
	% of Total	Estimate	2.7%	3.3%	0.1%	6.1%
		Standard Error	0.3%	0.4%	0.1%	0.5%
		Unweighted Count	267	161	7	435
12 years or completed high school	Population Size	Estimate	29281610.785	22310185.831	348086.715	51939883.331
		Standard Error	2104999.303	1981590.955	180939.915	2882035.408
		Unweighted Count	702	408	8	1118
	% within edu	Estimate	56.4%	43.0%	0.7%	100.0%
		Standard Error	2.8%	2.8%	0.4%	0.0%
		Unweighted Count	702	408	8	1118
	% within sex	Estimate	24.9%	18.0%	20.0%	21.4%

		Standard Error	1.5%	1.3%	10.3%	1.0%
		Unweighted Count	702	408	8	1118
	% of Total	Estimate	12.0%	9.2%	0.1%	21.4%
		Standard Error	0.8%	0.7%	0.1%	1.0%
		Unweighted Count	702	408	8	1118
Some college	Population Size	Estimate	43970102.087	49143452.595	868816.207	93982370.890
		Standard Error	2092722.944	2808989.005	479396.543	3300482.607
		Unweighted Count	114	778	10	1933
	% within edu	Estimate	46.8%	52.3%	0.9%	100.0%

		Standard Error	1.9%	2.0%	0.5%	0.0%
		Unweighted Count	1145	778	10	1933
	% within sex	Estimate	37.3%	39.8%	49.8%	38.6%
		Standard Error	1.4%	1.6%	15.9%	1.1%
		Unweighted Count	1145	778	10	1933
	% of Total	Estimate	18.1%	20.2%	0.4%	38.6%
		Standard Error	0.9%	1.0%	0.2%	1.1%
		Unweighted Count	1145	778	10	1933
College graduate or higher	Population Size	Estimate	38094052.880	44070822.434	235565.384	82400440.698
		Standard Error	1631985.426	1662324.288	88136.246	2137142.748

		Unweighted Count	1879	1285	16	3180
	% within edu	Estimate	46.2%	53.5%	0.3%	100.0%
		Standard Error	1.5%	1.5%	0.1%	0.0%
		Unweighted Count	1879	1285	16	3180
	% within sex	Estimate	32.3%	35.6%	13.5%	33.9%
		Standard Error	1.4%	1.3%	6.1%	0.9%
		Unweighted Count	1879	1285	16	3180
	% of Total	Estimate	15.7%	18.1%	0.1%	33.9%
		Standard Error	0.7%	0.7%	0.0%	0.9%
		Unweighted Count	1879	1285	16	3180
Total	Population Size	Estimate	117833381.967	123623920.762	1743950.472	243201253.201
		Standard Error	3092258.920	4307981.214	529383.452	5018543.497
		Unweighted Count	3993	2632	41	6666
	% within edu	Estimate	48.5%	50.8%	0.7%	100.0%
		Standard Error	1.1%	1.1%	0.2%	0.0%
		Unweighted Count	3993	2632	41	6666
	% within sex	Estimate	100.0%	100.0%	100.0%	100.0%
		Standard Error	0.0%	0.0%	0.0%	0.0%
		Unweighted Count	3993	2632	41	6666
	% of Total	Estimate	48.5%	50.8%	0.7%	100.0%

		Standard Error	1.1%	1.1%	0.2%	0.0%
		Unweighted Count	3993	2632	41	6666

The row percentages above show that a higher weighted proportion of college graduates in the sample are males (53.5%) than females (46.2%). Respondents with less than a high school diploma include more males (55.5%) than females (44.5%).

Tests of Independence

		Chi-Square	Adjusted F	df1	df2	Significance
edu * sex	Pearson	63.049	3.315	5.053	990.417	.005
	Likelihood Ratio	61.902	3.255	5.053	990.417	.006

Pearson chi-square test statistic and Likelihood Ratio test statistic and their associated p-values suggest that one may reject the null hypothesis that the two variables are not associated, which indicates that there is a significant difference between the distributions of educational attainment for males and females. The Pearson and Likelihood Ratio tests are more liberal than the design adjusted Rao-Scott approximation available in SAS, which accounts for the difference in results between the tests of independence in SPSS and SAS using the Taylor Series approach. SPSS does not have an option to specify the more accurate Rao-Scott test at this time.

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 sex and education on SeekCancerInfo.
CSLOGISTIC seekcancerinfo_recode (LOW) BY flippedsex flippededu
/PLAN FILE='(sample.csaplan)'
/MODEL flippedsex flippededu
/CUSTOM Label = 'Overall model minus intercept'
LMATRIX = flippedsex 1/2 1/2 -1;
          flippededu   1/3
          1/3    1/3    -1;
          flippededu   1/3
          1/3   -1   1/3   ;
          flippededu 1/3 -
          1     1/3     1/3;
          flippededu    -1
          1/3 1/3 1/3
/CUSTOM Label = 'Sex'
LMATRIX = flippedsex1/2
1/2 -1
/CUSTOM Label = 'Education
overall' LMATRIX = flippededu 1/3
1/3 1/3 -1;
          flippededu   1/3
          1/3   -1   1/3   ;
          flippededu 1/3 -
```

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

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

Sample Design Information

		N
Unweighted Cases	**Valid**	6662
	Invalid	616
	Total	7278
Population Size		242996104.688
Stage 1	**Strata**	4
	Units	200
Sampling Design Degrees of Freedom		196

Parameter Estimates

seekcancerinfo_recode		B	Std. Error	95% Confidence Interval		Hypothesis Test			Exp(B)	95% Confidence Interval for Exp(B)	
				Lower	Upper	t	df	Sig.		Lower	Upper
Yes	**(Intercept)**	-.958	.242	-1.435	-.481	-3.960	196	0.000	.384	.238	.618
	Don't Know	-.906	.631	-2.151	.338	-1.436	196	.153	.404	.116	1.402
	Male	-.283	.088	-.456	-.109	-3.216	196	.002	.754	.634	.896
	College graduate or higher	1.712	.250	1.220	2.205	6.853	196	0.000	5.542	3.386	9.071
	Some college	1.062	.253	.564	1.560	4.203	196	0.000	2.892	1.757	4.759
	12 years or completed high school	.779	.266	.255	1.303	2.931	196	.004	2.179	1.290	3.680

Odds Ratios

		seekcancerinfo_recode	Odds Ratio	95% Confidence Interval	
				Lower	Upper
Sex	Don't Know vs. Female	Yes	.404	.116	1.402
	Male vs. Female	Yes	.754	.634	.896
Education	College graduate or higher vs. Less than high school	Yes	5.542	3.386	9.071
	Some college vs. Less than high school	Yes	2.892	1.757	4.759
	12 years or completed high school vs. Less than high school	Yes	2.179	1.290	3.680

Overall Model Minus Intercept

df	Wald Chi-Square	Sig.
4.000	111.101	0.000

Sex

df	Wald Chi-Square	Sig.
1.000	3.464	0.000

Education Overall

df	Wald Chi-Square	Sig.
3.000	108.612	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, males appear to be statistically less likely than females 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 sex and education on
GeneralHealth. CSGLM genhealth_recode BY flippedsex flippededu
/PLAN FILE='(sample.csaplan)'
/MODEL flippededu flippedsex
/CUSTOM Label = 'Overall model minus intercept'
LMATRIX = flippedsex 1/2 1/2 -1;
flippededu    1/3
1/3      1/3      -1;
flippededu    1/3
1/3    -1    1/3    ;
flippededu 1/3 -
1      1/3      1/3;
flippededu      -1
1/3 1/3 1/3
/CUSTOM Label = 'Sex'
LMATRIX = flippedsex 1/2
1/2 -1
/CUSTOM Label = 'Education
overall' LMATRIX = flippededu 1/3
1/3 1/3 -1; flippededu 1/3 1/3 -1 1/3 ;
flippededu 1/3 -1 1/3
1/3; flippededu -1
1/3 1/3 1/3
/INTERCEPT INCLUDE=YES SHOW=YES
/STATISTICS PARAMETER SE CINTERVAL TTEST
/PRINT SUMMARY VARIABLEINFO SAMPLEINFO
/TEST TYPE=F PADJUST=LSD
/MISSING CLASSMISSING=EXCLUDE
/CRITERIA CILEVEL=95.
```

Sample Design Information

		N
Unweighted Cases	**Valid**	6639
	Invalid	639
	Total	7278
Population Size		242393463.149
Stage 1	**Strata**	4
	Units	200
Sampling Design Degrees of Freedom		196

Parameter Estimates

			95% Confidence Interval		Hypothesis Test		
Parameter	**Estimate**	**Std. Error**	**Lower**	**Upper**	**t**	**df**	**Sig.**
(Intercept)	3.148	.101	2.949	3.347	31.214	196	0.000
College graduate or higher	-.704	.117	-.935	-.474	-6.035	196	0.000
Some college	-.394	.116	-.622	-.165	-3.398	196	.001
12 years or completed high school	-.279	.112	-.500	-.058	-2.485	196	.014
Don't Know	.042	.273	-.497	.581	.153	196	.879
Male	-.119	.043	-.205	-.034	-2.743	196	.007

Compared to those respondents with less than a high school education, those who have a high school education, completed some college, are a college graduate 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. Compared to females, being male is negatively associated with general health (i.e., males have significantly better health than females).

Overall Model Minus Intercept

df1	df2	Wald F	Sig.
4	193	26.050	0.000

Sex

df1	df2	Wald F	Sig.
1	196	7.519	0.007

Education Overall

df1	df2	Wald F	Sig.
3	194	25.293	0.000

From the above table, we can see that both education and sex are significantly associated with general health.

Analyzing Data Using Stata

This section gives some Stata (Version 10.0 and higher) coding examples for common types of statistical analyses using HINTS 7 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 7 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 SeekCancerInfo" 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\hints7_public.dta"
* Recode negative values to missing
recode BirthSex (1=1 "Female") (2=2 "Male") (3=3 "Don't Know") (nonmissing=.),
generate(sex)
label variable sex "Sex"

* 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 == -7 | SeekCancerInfo == -9
label define seekcancerinfo2 0 "No" 1 "Yes"

label values SeekCancerInfo seekcancerinfo2
```

```
* Recode negative values to missing for svy: regress

replace GeneralHealth = . if GeneralHealth == -5 | GeneralHealth == -7 |
GeneralHealth == -9
```

Stata 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. In this example and declared design we are using PERSON_FINWT0 and its associated replicate weights (PERSON_FINWT1through PERSON_FINWT50) for the composite sample with no group differences. 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: to obtain standard errors for total, row, and column
you must separately request each under different tabulate statements

svy: tabulate edu sex, cell format(%8.5f) percent se wald noadjust

svy: tabulate edu sex, row format(%8.5f) percent se wald noadjust

svy: tabulate edu sex, column format(%8.5f) percent se 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 cell, column, row request total cell, column, and row frequencies, respectively. These options must be individually run. 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      =         6,666
                                              Population size    =   243,201,253
                                              Replications       =            50
                                              Design df          =            49
```

Education	Sex			
	Female	Male	Don't Know	Total
Less than high school	2.66759	3.33035	0.11985	6.11780
	(0.26895)	(0.34104)	(0.06504)	(0.45307)
12 years	12.04007	9.17355	0.14313	21.35675
	(0.53890)	(0.47926)	(0.07301)	(0.66233)
Some college	18.07972	20.20691	0.35724	38.64387
	(0.49430)	(0.52137)	(0.20655)	(0.73739)
College	15.66359	18.12113	0.09686	33.88159
	(0.17820)	(0.19355)	(0.03757)	(0.25276)
Total	48.45098	50.83194	0.71708	1.0e+02
	(0.30552)	(0.39125)		

Key: cell percentage
(jackknife standard error of cell percentage)

Wald (Pearson):
Unadjusted chi2(6) = 67.7141
Unadjusted F(6, 49) = 11.2857 P = 0.0000
Adjusted F(6, 44) = 10.1341 P = 0.0000

Jknife *: for rows

Number of strata = 1 Number of obs = 6,666
Population size = 243,201,253
Replications = 50
Design df = 49

Education	Sex			
	Female	Male	Don't Know	Total
Less than high school	43.60380	54.43713	1.95908	1.0e+02
	(3.30870)	(3.39999)	(1.06638)	
12 years	56.37597	42.95386	0.67017	1.0e+02
	(1.83784)	(1.77242)	(0.34309)	
Some college	46.78548	52.29008	0.92445	1.0e+02

	(0.97156)	(0.97859)	(0.53269)	
College	46.23040	53.48372	0.28588	1.0e+02
	(0.39463)	(0.41106)	(0.11091)	
Total	48.45098	50.83194	0.71708	1.0e+02
	(0.30552)	(0.39125)	(0.22031)	

Key: row percentage
(jackknife standard error of row percentage)

Wald (Pearson):
Unadjusted chi2(6) = 67.7141
Unadjusted F(6, 49) = 11.2857 P = 0.0000
Adjusted F(6, 44) = 10.1341 P = 0.0000

Jknife *: for columns

Number of strata = 1 Number of obs = 6,666
Population size = 243,201,253
Replications = 50
Design df = 49

	Sex			
Education	Female	Male	Don't Know	Total
Less than high school	5.50575	6.55169	16.71390	6.11780
	(0.55673)	(0.66437)	(10.56824)	(0.45307)
12 years	24.85001	18.04682	19.95967	21.35675
	(1.08278)	(0.91203)	(11.84489)	(0.66233)
Some college	37.31549	39.75238	49.81886	38.64387
	(0.99638)	(0.99144)	(20.35336)	(0.73739)
College	32.32874	35.64911	13.50757	33.88159
	(0.34901)	(0.39020)	(6.92794)	(0.25276)
Total	1.0e+02	1.0e+02		1.0e+02

Key: column percentage
(jackknife standard error of column percentage)

Wald (Pearson):
Unadjusted chi2(6) = 67.7141
Unadjusted F(6, 49) = 11.2857 P = 0.0000
Adjusted F(6, 44) = 10.1341 P = 0.0000

For the purposes of computing appropriate degrees of freedom for the estimator of the HINTS 7 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 sex [omit] 1
char edu [omit] 1

* Multivariable logistic regression of sex and education on SeekCancerInfo

xi: svy: logit SeekCancerInfo i.sex i.edu
test _Isex_2 _Isex_3 _Iedu_2 _Iedu_3 _Iedu_4 _cons, nosvyadjust
test _Isex_2 _Isex_3 _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust
test _Isex_2 _Isex_3, nosvyadjust
test _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust
xi: svy, or: logit SeekCancerInfo i.sex i.edu
```

The **char** command defines the categorical variable with the reference group. The “Male” is the reference group for sex 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.sex 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.

```
. xi: svy: logit SeekCancerInfo i.sex  i.edu
     i.sex      _Isex_1-3         (naturally coded; _Isex_1 omitted)
     i.edu      _Iedu_1-4         (naturally coded; _Iedu_1 omitted)
     (running logit on estimation sample)

Survey: Logistic regression

Number of  strata       =      1              Number of obs            =             6,662
                                              Population size          =       242,996,105
                                              Replications             =                50
                                              Design df                =                49
                                              F(    5,       45)       =             19.99
```

Prob > F = 0.0000

seekcancerinfo	Coef.	Jknife * Std. Err.	t	P>|t|	[95% Conf.	Interval]
_Isex_2	-.2825765	.0847825	-3.33	0.002	-.4529532	-.1121997
_Isex_3	-.9062476	.8549039	-1.06	0.294	-2.624241	.811746
_Iedu_2	.7789734	.2926201	2.66	0.010	.1909312	1.367016
_Iedu_3	1.061824	.2590491	4.10	0.000	.5412449	1.582402
_Iedu_4	1.712338	.2660771	6.44	0.000	1.177636	2.24704
_cons	-.9575964	.2538529	-3.77	0.000	-1.467733	-.4474599

Unadjusted Wald test

(1) [seekcancerinfo]_Isex_2 = 0
(2) [SeekCancerInfo]_Isex_3 = 0
(3) [SeekCancerInfo]_Iedu_2 = 0
(4) [SeekCancerInfo]_Iedu_3 = 0
(5) [SeekCancerInfo]_Iedu_4 = 0
(6) [SeekCancerInfo]_cons = 0

F(6, 49) = 24.83
Prob > F = 0.0000

Unadjusted Wald test

(1) [SeekCancerInfo]_Isex_2 = 0
(2) [SeekCancerInfo]_Isex_3 = 0
(3) [SeekCancerInfo]_Iedu_2 = 0
(4) [SeekCancerInfo]_Iedu_3 = 0
(5) [SeekCancerInfo]_Iedu_4 = 0

F(5, 49) = 21.77
Prob > F = 0.0000

Unadjusted Wald test

(1) [SeekCancerInfo]_Isex_2 = 0
(2) [SeekCancerInfo]_Isex_3 = 0

F(2,49) = 6.49
Prob > F = 0.0032

```
Unadjusted Wald test

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

       F(  3,49)    =     34.99
       Prob > F     =     0.0000

i.sex      _Isex_1-3  (naturally coded; _Isex_1 omitted)
i.edu      _Iedu_1-4     (naturally coded; _Iedu_1 omitted)
(running logit on estimation sample) Survey: Logistic regression

Number of strata      =      1             Number of obs        =          6,662
                                           Population size      =    242,996,105
                                           Replications         =             50
                                           Design df            =             49
                                           F(    5,      45)    =          19.99
                                           Prob > F             =         0.0000
```

seekcancerinfo	Odds Ratio	Jknife * Std. Err.	t	P>|t|	[95% Conf.	Interval]
_Isex_2	.753839	.0639123	-3.33	0.002	.6357479	.8938657
_Isex_3	.4040375	.3454132	-1.06	0.294	.0724947	2.251836
_Iedu_2	2.179234	.6376877	2.66	0.010	1.210376	3.923623
_Iedu_3	2.891639	.7490765	4.10	0.000	1.718145	4.866632
_Iedu_4	5.541904	1.474574	6.44	0.000	3.24669	9.459696
_cons	.3838143	.0974324	-3.77	0.000	.2304474	.6392498

```
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, males appear to have 0.75 the odds as females to have searched for cancer information.

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 sex and education on GeneralHealth

xi: svy: regress GeneralHealth i.sex i.edu

test _Isex_2 _Isex_3  _Iedu_2 _Iedu_3 _Iedu_4 _cons, nosvyadjust
test _Isex_2 _Isex_3  _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust
test _Isex_2 _Isex_3, nosvyadjust
test _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust


i.sex        _Isex_1-3          (naturally coded; _Isex_1 omitted)
i.edu        _Iedu_1-4          (naturally coded; _Iedu_1 omitted)

(running regress on estimation sample)

Survey: Linear regression

    Number of strata     =          1                 Number of obs       =         6,639
                                                      Population size     =   242,393,463
                                                      Replications        =            50
                                                      Design df           =            49
                                                      F(    5,      45)   =         20.11
                                                      Prob > F            =        0.0000
                                                      R-squared           =        0.0510

------------------------------------------------------------------------------
             |             Jknife *
generalhealth|      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      _Isex_2|  -.1192809   .0444003    -2.69   0.010    -.2085068   -.0300551
      _Isex_3|   .0418274   .3586832     0.12   0.908    -.6789734    .7626283
      _Iedu_2|  -.2787725    .119405    -2.33   0.024    -.5187259   -.0388191
      _Iedu_3|  -.3935888   .1228712    -3.20   0.002    -.6405077    -.14667
      _Iedu_4|  -.7044893   .1223386    -5.76   0.000    -.9503378   -.4586407
        _cons|   3.147948   .1065234    29.55   0.000     2.933881    3.362015
------------------------------------------------------------------------------

    Unadjusted

    Wald test
    ( 1)  _Isex_2 = 0
    ( 2)  _Isex_3 = 0
    ( 3)  _Iedu_2 = 0
    ( 4)  _Iedu_3 = 0
    ( 5)  _Iedu_4 = 0
    ( 6)  _cons = 0

            F(  6,    49)    =      468.88
            Prob > F         =      0.0000
```

```
Unadjusted

Wald test
( 1)  _Isex_2 = 0
( 2)  _Isex_3 = 0
( 3)  _Iedu_2 = 0
( 4)  _Iedu_3 = 0
( 5)  _Iedu_4 = 0

       F( 5,  49)     =      21.90
       Prob > F       =      0.0000


Unadjusted Wald test
( 1)  _Isex_2 = 0
( 2)  _Isex_3 = 0

       F(  2, 49)     =      3.61
       Prob > F       =      0.0344

 Unadjusted Wald test ( 1) _Iedu_2 = 0
( 2) _Iedu_3 = 0
( 3) _Iedu_4 = 0

       F(  3, 49)     =      26.22
       Prob > F       =      0.0000
```

From the above table, compared to those respondents with less than a high school education, those with a high school education, those with some college or those with 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. Compared to females, being male is negatively associated with general health.

Stata 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. In this example and declared design we are using PERSON_FINWT0 for the composite sample with no group differences. 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 (Taylor series)
svyset VAR_CLUSTER [pw= PERSON_FINWT0], strata(VAR_STRATUM)
```

Cross-tabulation

```
* cross-tabulation
svy: tabulate edu sex, cell format(%8.5f) percent se wald noadjust
svy: tabulate edu sex, row format(%8.5f) percent se wald noadjust
svy: tabulate edu sex, column format(%8.5f) percent se 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 cell, column, row request total cell, column, and row frequencies, respectively. These options must be individually run. 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   =         4                Number of obs     =         6,666
Number of PSUs     =       200                Population size   =   243,201,253
                                              Design df         =           196
```

Education	Sex Female	Male	Don't Know	Total
Less than high school	2.66759	3.33035	0.11985	6.11780
	(0.28657)	(0.41967)	(0.06589)	(0.47262)
12 years	12.04007	9.17355	0.14313	21.35675
	(0.80465)	(0.72689)	(0.07480)	(0.95779)
Some college	18.07972	20.20691	0.35724	38.64387
	(0.87036)	(0.96578)	(0.19779)	(1.06739)
College	15.66359	18.12113	0.09686	33.88159
	(0.66893)	(0.69456)	(0.03629)	(0.89210)
Total	48.45098	50.83194	0.71708	1.0e+02
	(1.11971)	(1.13593)	(0.21973)	

```
  Key:  cell percentage
        (linearized standard error of cell percentage)

  Wald (Pearson):
    Unadjusted       chi2(6)     =      19.3958
    Unadjusted     F(6, 196)     =       3.2326   P = 0.0047
    Adjusted       F(6, 191)     =       3.1502   P = 0.0057

(running tabulate on estimation sample)

Number of strata   =         4                Number of obs     =         6,666
Number of PSUs     =       200                Population size   =   243,201,253
                                              Design df         =           196
```

Education	Sex Female	Male	Don't Know	Total
Less than high school	43.60380	54.43713	1.95908	1.0e+02
	(4.26992)	(4.38514)	(1.07649)	
12 years	56.37597	42.95386	0.67017	1.0e+02
	(2.77500)	(2.78934)	(0.35275)	
Some college	46.78548	52.29008	0.92445	1.0e+02
	(1.94090)	(1.95471)	(0.51000)	
College	46.23040	53.48372	0.28588	1.0e+02
	(1.52874)	(1.52293)	(0.10664)	
Total	48.45098	50.83194	0.71708	1.0e+02
	(1.11971)	(1.13593)	(0.21973)	

Key: row percentage
(linearized standard error of row percentage)

Wald (Pearson):
Unadjusted chi2(6) = 19.3958
Unadjusted F(6, 196) = 3.2326 P = 0.0047
Adjusted F(6, 191) = 3.1502 P = 0.0057

(running tabulate on estimation sample)

Number of strata = 4 Number of obs = 6,666
Number of PSUs = 200 Population size = 243,201,253
Design df = 196

Education	Sex Female	Male	Don't Know	Total
Less than high school	5.50575	6.55169	16.71390	6.11780
	(0.59547)	(0.82379)	(9.27177)	(0.47262)
12 years	24.85001	18.04682	19.95967	21.35675
	(1.50127)	(1.33000)	(10.30224)	(0.95779)
Some college	37.31549	39.75238	49.81886	38.64387
	(1.40550)	(1.59348)	(15.91246)	(1.06739)
College	32.32874	35.64911	13.50757	33.88159
	(1.37548)	(1.26820)	(6.08065)	(0.89210)
Total	1.0e+02	1.0e+02		1.0e+02

Key: column percentage

```
          (linearized standard error of column percentage)

Wald (Pearson):
  Unadjusted    chi2(6)        =      19.3958
  Unadjusted    F(6, 196)      =       3.2326  P = 0.0047
  Adjusted      F(6, 191)      =       3.1502  P = 0.0057
```

The results of these tests based on Taylor Series linearization are consistent with the results conducted using replication shown in the previous section. (In the previous section, the distributions of educational attainment between males and females were determined to be statistically different.) However, since both education and sex are variables used in the raking process as part of the HINTS weighting procedure, the standard errors based on replication are much smaller than those based on Taylor Series linearization. In some cases, this difference could potentially result in significant differences using the replication method but not the Taylor Series linearization method.

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 sex [omit] 1
char edu [omit] 1
* Multivariable logistic regression of sex and education on seekcancerinfo
xi: svy: logit SeekCancerInfo i.sex i.edu

test _Isex_2 _Isex_3 _Iedu_2 _Iedu_3 _Iedu_4 _cons, nosvyadjust
test _Isex_2 _Isex_3 _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust
test _Isex_2 _Isex_3, nosvyadjust
test _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust

xi: svy, or: logit SeekCancerInfo i.sex i.edu
```

The **char** command defines categorical variable with reference group. The “Male” is the reference group for sex 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.sex 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.sex      _Isex_1-3         (naturally coded; _Isex_1 omitted)
i.edu      _Iedu_1-4         (naturally coded; _Iedu_1 omitted)

(running logit on estimation sample) Survey: Logistic regression

Number of strata   =                 4         Number of obs      =          6,662
Number of PSUs     =               200         Population size    =    242,996,105
                                               Design df          =            196
                                               F(   5,     192)   =          21.79
```

Prob > F = 0.0000

seekcancerinfo	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
_Isex_2	-.2825765	.0878578	-3.22	0.002	-.4558444	-.1093086
_Isex_3	-.9062476	.6310147	-1.44	0.153	-2.150698	.3382026
_Iedu_2	.7789734	.2657364	2.93	0.004	.2549037	1.303043
_Iedu_3	1.061824	.2526392	4.20	0.000	.5635833	1.560064
_Iedu_4	1.712338	.2498755	6.85	0.000	1.219548	2.205128
_cons	-.9575964	.2418447	-3.96	0.000	-1.434548	-.4806445

Unadjusted Wald test

(1) [SeekCancerInfo]_Isex_2 = 0
(2) [SeekCancerInfo]_Isex_3 = 0
(3) [SeekCancerInfo]_Iedu_2 = 0
(4) [SeekCancerInfo]_Iedu_3 = 0
(5) [SeekCancerInfo]_Iedu_4 = 0

F(6,196) = 22.71
Prob > F = 0.0000

Unadjusted Wald test

(1) [SeekCancerInfo]_Isex_2 = 0
(2) [SeekCancerInfo]_Isex_3 = 0
(3) [SeekCancerInfo]_Iedu_2 = 0
(4) [SeekCancerInfo]_Iedu_3 = 0
(5) [SeekCancerInfo]_Iedu_4 = 0

F(5, 196) = 22.24
Prob > F = 0.0000

Unadjusted Wald test

(1) [SeekCancerInfo]_Isex_2 = 0
(2) [SeekCancerInfo]_Isex_3 = 0

F(2, 196) = 6.12
Prob > F = 0.0026

Unadjusted Wald test

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

F(3, 196) = 36.20
Prob > F = 0.0000

i.sex _Isex_1-3 (naturally coded; _Isex_1 omitted)

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

```
(running logit on estimation sample) Survey: Logistic regression

Number of strata      =               4       Number of obs   =          6,662
Number of PSUs        =             200     Population size   =    242,996,105
                                                Design df       =            196
                                                F(5, 192)       =          21.79
                                                Prob > F        =         0.0000
```

seekcancerinfo	Odds Ratio	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
_Isex_2	.753839	.0662306	-3.22	0.002	.6339125	.8964538
_Isex_3	.4040375	.2549536	-1.44	0.153	.1164029	1.402425
_Iedu_2	2.179234	.5791018	2.93	0.004	1.290337	3.68048
_Iedu_3	2.891639	.7305415	4.20	0.000	1.756957	4.759125
_Iedu_4	5.541904	1.384786	6.85	0.000	3.385659	9.071412
_cons	.3838143	.0928234	-3.96	0.000	.238223	.6183847

```
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, men appear to have 0.75 lower odds than women to have searched for cancer information.

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 sex and education on generalhealth

 xi: svy: regress GeneralHealth i.sex i.edu

test _Isex_2 _Isex_3 _Iedu_2 _Iedu_3 _Iedu_4 _cons, nosvyadjust
test _Isex_2 _Isex_3 _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust
test _Isex_2 _Isex_3, nosvyadjust
test _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust

i.sex  _Isex_1-3          (naturally coded; _Isex_1 omitted)

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

(running regress on estimation sample)

Survey: Linear regression

Number of strata    =          4                Number of obs     =         6,639
Number of PSUs      =        200                Population size   =   242,393,463
                                                Design df         =           196
```

```
                                                F(   5,     192)   =        20.81
                                                Prob > F           =       0.0000
                                                R-squared          =       0.0510
```

generalhealth	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
_Isex_2	-.1192809	.0434806	-2.74	0.007	-.2050308	-.033531
_Isex_3	.0418274	.27346	0.15	0.879	-.4974742	.5811291
_Iedu_2	-.2787725	.1121746	-2.49	0.014	-.4999966	-.0575485
_Iedu_3	-.3935888	.1158421	-3.40	0.001	-.6220457	-.1651319
_Iedu_4	-.7044893	.1167277	-6.04	0.000	-.9346927	-.4742858
_cons	3.147948	.1019607	31.21	0.000	2.949055	3.346841

```
Unadjusted Wald test
( 1)  _Isex_2 = 0
( 2)  _Isex_3 = 0
( 3)  _Iedu_2 = 0
( 4)  _Iedu_3 = 0
( 5)  _Iedu_4 = 0
( 6)  _cons = 0

           F( 6,  196)   =      4958.35
           Prob > F      =      0.0000

Unadjusted Wald test

( 1)  _Isex_2 = 0
( 2)  _Isex_3 = 0
( 3)  _Iedu_2 = 0
( 4)  _Iedu_3 = 0
( 5)  _Iedu_4 = 0

           F( 5,  196)   =      21.25
           Prob > F      =      0.0000

Unadjusted Wald test
( 1)  _Isex_2 = 0
( 2)  _Isex_3 = 0

           F( 2,  196)   =      3.77
           Prob > F      =      0.0248


Unadjusted Wald test
( 1)  _Iedu_2 = 0
( 2)  _Iedu_3 = 0
( 3)  _Iedu_4 = 0

           F( 3,  196)   =      25.70
           Prob > F      =      0.0000
```

From the above table, compared to those respondents with less than a high school education, those with a high school education, some college education, 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. Compared to females, being male is negatively associated with general health.

Analyzing Data Using R

This section gives some R (v 4.43.0) coding examples for common types of statistical analyses using HINTS 7 data. Subsection 1 shows how to complete common analyses using replicate weights, and subsection 2 shows analyses using the Taylor series linearization approach. R has many packages and libraries for data processing, statistical analysis, and other programming usages that must be loaded into R prior to use. Packages that have not been previously installed to the R library can be added using the install.packages("packagename") command prior to loading them from the library. This code provides the required packages and libraries that must be loaded into R prior to reading in the data, conducting data management, and running the example statistical analyses on the HINTS 7 data.

It is important to note that loading data into R using the haven package does not preserve variable label formats, except in the case of Stata data. Users who wish to import SAS or SPSS files and preserve variable label formats may use other packages for importing data, such as foreign. HINTS datasets in SAS, SPSS or Stata formats can also be imported into RStudio by clicking on the "file" tab, followed by "import dataset", and selecting the tab matching the file type.

```
library(haven) # For loading data from SAS, SPSS, or STATA into R
library(dplyr) # For data manipulation
library(survey) # For analyzing complex survey data
library(srvyr) # For manipulating survey objects with dplyr
library(broom) # For presenting tidy data tables

# Setting the working directory to file location
setwd('[WORKING DIRECTORY HERE]')

# Load data
df = haven::read_sas("hints7_public.sas7bdat")
```

Once the necessary libraries are loaded and the SAS dataset has been read into R, data management can be conducted using the dplyr library to create new variables or recode existing variables. 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, R will exclude these responses from procedures where these variables are specifically accessed. 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 to verify proper coding.

```
df = df |>
  dplyr::mutate(sex = case_match(factor(BirthSex),
                                 '1' ~ 'Female',
                                 '2' ~ 'Male',
                                 '3' ~ 'Dont Know')) |>

  dplyr::mutate(edu = case_match(factor(Education),
                                 c('1', '2') ~ 'Less than high school',
                                 '3' ~ '12 years or completed high school',
                                 c('4', '5') ~ 'Some college',
                                 c('6', '7') ~ 'College graduate or higher'))
```

```
|>

  dplyr::mutate(SeekCancerInfo = case_match(SeekCancerInfo,
                                            1 ~ 1,
                                            2 ~ 0))

# Setting the reference level for categorical variables
df$sex = relevel(factor(df$sex, ordered = F),
                 ref = 'Female')

df$edu = relevel(factor(df$edu, ordered = F),
                ref = 'Less than high school')
```

R Replicate Weights Variance Estimation Method

R package 'srvyr' requires that the survey object (svy_obj_rep) be created before any analysis using the as_survey_rep command. The survey object created will be called in subsequent analyses. In this example and declared design we are using PERSON_FINWT0 and its associated replicate weights (PERSON_FINWT1through PERSON_FINWT50). The code below creates a survey design object to account for replicate weights when running statistical analyses.

```
svy_obj_rep = as_survey_rep(.data = df,
                            weights = PERSON_FINWT0,
                            repweights = num_range(prefix = "PERSON_FINWT",
                                                   range = 1:50),
                            type = "JKn",
                            scale = 0.98,
                            rscales = rep(1, times = 50))
```

Crosstabulation and chi-square test:

```
# Crosstab
svy_obj_rep |>
  dplyr::filter(is.na(edu) == F,
                is.na(sex) == F) |>
  dplyr::group_by(edu, sex) |>
  dplyr::summarize(n = n(),
                   total = survey_total(),
                   pct = survey_prop())

## When `proportion` is unspecified, `survey_prop()` now defaults to `proportion = TR
UE`.
## ℹ This should improve confidence interval coverage.
## This message is displayed once per session.

## # A tibble: 12 × 7
## # Groups:   edu [4]
##    edu                             sex       n  total total_se     pct  pct_se
##    <fct>                           <fct> <int>  <dbl>    <dbl>   <dbl>   <dbl>
##  1 Less than high school           Fema…   267 6.49e6  649299. 0.436   0.0331
##  2 Less than high school           Dont…     7 2.91e5  157843. 0.0196  0.0107
```

```
##  3 Less than high school             Male    161 8.10e6  830769. 0.544   0.0340
##  4 12 years or completed high school Fema…   702 2.93e7 1353087. 0.564   0.0184
##  5 12 years or completed high school Dont…     8 3.48e5  177687. 0.00670 0.00343
##  6 12 years or completed high school Male    408 2.23e7 1165558. 0.430   0.0177
##  7 College graduate or higher        Fema…  1879 3.81e7  440771. 0.462   0.00395
##  8 College graduate or higher        Dont…    16 2.36e5   91301. 0.00286 0.00111
##  9 College graduate or higher        Male   1285 4.41e7  407132. 0.535   0.00411
## 10 Some college                      Fema…  1145 4.40e7 1222451. 0.468   0.00972
## 11 Some college                      Dont…    10 8.69e5  501879. 0.00924 0.00533
## 12 Some college                      Male    778 4.91e7 1384398. 0.523   0.00979
```

```r
# Chi-square test
svy_obj_rep |>
  svychisq(formula = ~ sex + edu,
           statistic = "F")
```

```
##
##  Pearson's X^2: Rao & Scott adjustment
##
## data:  NextMethod()
## F = 4.6031, ndf = 3.6893, ddf = 180.7756, p-value = 0.001977
```

The row percentages above show that a higher weighted proportion of college graduates in the sample are male (53.5%) than female (46.2%). Respondents with less than a high school diploma include fewer females (43.6%) than males (54.4%). The remaining percentage of college graduates (0.3%) and those with less than a high school diploma (2%) represent those who reported their birth sex as “don’t know”.The statistic for the Chi-square test of independence and its associated p-value indicate that the distributions of educational attainment between males and females are significantly different.

Logistic Regression

This example demonstrates a multivariable logistic regression model using **svyglm** and the survey object created in the first step (svy_obj_rep); recall that the response should be a dichotomous 0-1 variable. The response variable should be on the left-hand side of the tilde in the formula statement, while all covariates should be listed on the right-hand side. The “Female” is the reference group for sex effect, while “Less than high school” is the reference group for education level effect.

Computing a logistic regression:

```r
logistic_model = svy_obj_rep |>
  svyglm(formula = SeekCancerInfo ~ edu + sex,
         family = quasibinomial())

# For displaying general summary statistics
summary(logistic_model)
```

```
##
## Call:
## svyglm(svy_obj_rep, formula = SeekCancerInfo ~ edu + sex, family = quasibinomial()
)
##
```

```
## Survey design:
## Called via srvyr
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          -0.95760    0.25380  -3.773 0.000479 ***
## edu12 years or completed high school  0.77897    0.29260   2.662 0.010799 *
## eduCollege graduate or higher         1.71234    0.26602   6.437 7.68e-08 ***
## eduSome college                       1.06182    0.25899   4.100 0.000175 ***
## sexDont Know                         -0.90625    0.85483  -1.060 0.294860
## sexMale                              -0.28258    0.08478  -3.333 0.001749 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 1.012172)
##
## Number of Fisher Scoring iterations: 4
```

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. However, significance may also be garnered from the test of whether the associated beta parameter is equal to 0. According to this model, males appear to have 0.75 times lower odds than females to have searched for cancer information.

```
# For displaying odds ratios and 95% confidence intervals
tidy(logistic_model,
     conf.int = T,
     conf.level = 0.95,
     exponentiate = T)
```

```
## # A tibble: 6 × 7
##   term                   estimate std.error statistic p.value conf.low conf.high
##   <chr>                     <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 (Intercept)               0.384    0.254      -3.77 4.79e-4   0.230      0.640
## 2 edu12 years or comple…    2.18     0.293       2.66 1.08e-2   1.21       3.93
## 3 eduCollege graduate o…    5.54     0.266       6.44 7.68e-8   3.24       9.47
## 4 eduSome college           2.89     0.259       4.10 1.75e-4   1.72       4.87
## 5 sexDont Know              0.404    0.855      -1.06 2.95e-1   0.0721     2.26
## 6 sexMale                   0.754    0.0848     -3.33 1.75e-3   0.635      0.894
```

Linear Regression

This example demonstrates a multivariable linear regression model using svyglm and the survey object created in the first step (svy_obj_rep); 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.

Computing a linear regression:

```
linear_model = svy_obj_rep |>
  svyglm(formula = GeneralHealth ~ edu + sex,
         family = gaussian())

summary(linear_model)
```

```
##
## Call:
## svyglm(svy_obj_rep, formula = GeneralHealth ~ edu + sex, family = gaussian())
##
## Survey design:
## Called via srvyr
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           3.13412    0.10794  29.036  < 2e-16 ***
## edu12 years or completed high school -0.34828    0.12700  -2.742  0.00879 **
## eduCollege graduate or higher        -0.74569    0.12460  -5.985 3.55e-07 ***
## eduSome college                      -0.39689    0.12448  -3.188  0.00263 **
## sexDont Know                          0.01793    0.35492   0.051  0.95993
## sexMale                              -0.10295    0.05064  -2.033  0.04810 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.267925)
##
## Number of Fisher Scoring iterations: 2
```

The summary results show that respondents with a high school education, some college, and completed college reported better general health than those with less than a high school education 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. Also, while there was significant difference in reported general health between those who identified as male compared to females (p = 0.05), there's no significant difference in reported general health between females and individuals who report their birth sex as "don't know" (p = 0.96).

R Taylor Series Linearization Variance Estimation

The code below creates a survey design object (svy_obj_linear) to account for Taylor Series linearization sample weights when running statistical analyses.

```
svy_obj_linear = as_survey_design(.data = df,
                                  ids = VAR_CLUSTER,
                                  strata = VAR_STRATUM,
                                  weights = PERSON_FINWT0,
                                  nest = T)
```

Computing a crosstab and chi-square test

```
# Crosstab
svy_obj_linear |>
  dplyr::filter(is.na(edu) == F,
                is.na(sex) == F) |>
  dplyr::group_by(edu, sex) |>
  dplyr::summarize(n = n(),
                   total = survey_total(),
                   pct = survey_prop())
```

```
## # A tibble: 12 × 7
## # Groups:   edu [4]
##    edu                                sex       n  total total_se     pct  pct_se
##    <fct>                              <fct> <int>  <dbl>    <dbl>   <dbl>   <dbl>
##  1 Less than high school              Fema…   267 6.49e6  694768. 0.436   0.0427
##  2 Less than high school              Dont…     7 2.91e5  160113. 0.0196  0.0108
##  3 Less than high school              Male    161 8.10e6 1044970. 0.544   0.0439
##  4 12 years or completed high school  Fema…   702 2.93e7 2104999. 0.564   0.0278
##  5 12 years or completed high school  Dont…     8 3.48e5  180940. 0.00670 0.00353
##  6 12 years or completed high school  Male    408 2.23e7 1981591. 0.430   0.0279
##  7 College graduate or higher         Fema…  1879 3.81e7 1631985. 0.462   0.0153
##  8 College graduate or higher         Dont…    16 2.36e5   88136. 0.00286 0.00107
##  9 College graduate or higher         Male   1285 4.41e7 1662324. 0.535   0.0152
## 10 Some college                       Fema…  1145 4.40e7 2092723. 0.468   0.0194
## 11 Some college                       Dont…    10 8.69e5  479397. 0.00924 0.00510
## 12 Some college                       Male    778 4.91e7 2808989. 0.523   0.0195
```

```
# Chi-square test
svy_obj_linear |>
  svychisq(formula = ~ sex + edu,
           statistic = "F")
```

```
##
##  Pearson's X^2: Rao & Scott adjustment
##
## data:  NextMethod()
## F = 3.3152, ndf = 5.0531, ddf = 990.4167, p-value = 0.005459
```

The row percentages above show that a higher weighted proportion of college graduates in the sample are males (53.5%) than females (46.2%).Respondents with less than a high school diploma include fewer females (43.6%) than males (54.4%). The remaining percentage of college graduates (0.3%) and those with less than a high school diploma (2%) represent those who reported their birth sex as “don’t know”. The Chi-squared test of independence statistic and associated p value suggest that one may reject the null hypothesis that the two variables are not associated, which indicates that there is a significant difference between the distributions of educational attainment for these two groups.

Logistic Regression

This example demonstrates a multivariable logistic regression model using **svyglm** and the svy_obj_linear survey object; recall that the response should be a dichotomous 0-1 variable. The

response variable should be on the left-hand side (LHS) of the tilde in the formula command, while all covariates should be listed on the right-hand side (RHS). The “Female” is the reference group for sex effect, while “Less than high school” is the reference group for education level effect.

Computing a logistic regression:

```
logistic_model = svy_obj_linear |>
  svyglm(formula = SeekCancerInfo ~ edu + sex,
         family = quasibinomial())

# For displaying general summary statistics
summary(logistic_model)
```

```
##
## Call:
## svyglm(formula = SeekCancerInfo ~ edu + sex, design = svy_obj_linear,
##     family = quasibinomial())
##
## Survey design:
## Called via srvyr
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        -0.95760    0.24185  -3.960 0.000106 ***
## edu12 years or completed high school  0.77897    0.26574   2.931 0.003787 **
## eduCollege graduate or higher         1.71234    0.24988   6.853 9.71e-11 ***
## eduSome college                       1.06182    0.25264   4.203 4.04e-05 ***
## sexDont Know                       -0.90625    0.63101  -1.436 0.152584
## sexMale                            -0.28258    0.08786  -3.216 0.001525 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 1.012324)
##
## Number of Fisher Scoring iterations: 4
```

```
# For displaying odds ratios and 95% confidence intervals
tidy(logistic_model,
     conf.int = T,
     conf.level = 0.95,
     exponentiate = T)
```

```
## # A tibble: 6 × 7
##   term                  estimate std.error statistic  p.value conf.low conf.high
##   <chr>                    <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)              0.384    0.242      -3.96 1.06e- 4    0.238     0.618
## 2 edu12 years or compl…    2.18     0.266       2.93 3.79e- 3    1.29      3.68
## 3 eduCollege graduate …    5.54     0.250       6.85 9.71e-11    3.39      9.07
## 4 eduSome college          2.89     0.253       4.20 4.04e- 5    1.76      4.76
## 5 sexDont Know             0.404    0.631      -1.44 1.53e- 1    0.116     1.40
## 6 sexMale                  0.754    0.0879     -3.22 1.53e- 3    0.634     0.896
```

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. 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, females appear to be statistically more likely than males to have searched for cancer information.

Linear Regression

This example demonstrates a multivariable linear regression model using **svyglm** and the svy_obj_linear survey object; 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.

Computing a linear regression:

```
linear_model = svy_obj_linear |>
  svyglm(formula = GeneralHealth ~ edu + sex,
         family = gaussian())

summary(linear_model)
```

```
##
## Call:
## svyglm(formula = GeneralHealth ~ edu + sex, design = svy_obj_linear,
##     family = gaussian())
##
## Survey design:
## Called via srvyr
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          3.13412    0.10234  30.625  < 2e-16 ***
## edu12 years or completed high school -0.34828    0.12125  -2.872 0.004534 **
## eduCollege graduate or higher        -0.74569    0.11893  -6.270 2.35e-09 ***
## eduSome college                      -0.39689    0.11670  -3.401 0.000818 ***
## sexDont Know                          0.01793    0.27250   0.066 0.947602
## sexMale                              -0.10295    0.05058  -2.036 0.043175 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.268115)
##
## Number of Fisher Scoring iterations: 2
```

Compared to those respondents with less than a high school education, those who have a high school education, completed some college, and are college graduates on average reported significantly better general health (i.e., the negative beta coefficient indicates that the average health score is lower among those with a high school education, some college, and college graduates because the health variable is coded such that lower scores correspond to better health), controlling for all variables in the model. We do not interpret the estimates for males and respondents who reported their birth sex as “don’t know” because the corresponding p-value is greater than 0.05.

Merging HINTS Survey Iterations

This section provides SAS, SPSS, Stata, and R code to combine HINTS 7 and HINTS 6 data. The provided code will generate one final sample weight for population point estimates and 100 replicate weights to compute standard errors when using the replicate method for variance estimation.

Merging HINTS 7 and HINTS 6 using SAS

This section provides SAS (Version 9.4 and higher) code for merging the HINTS 7 and HINTS 6 data. 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 =13,580) that contains one new final sample weight (for population point estimates, Merged_NWGT0) and 100 new replicate weights (Merged_NWGT1 TO Merged_NWGT100; to compute standard errors); these weights are set up using the Rizzo et al. [2008] method).

One assumption when using the SAS code below is that the analyst has already formatted each file using the formats and format assignment files provided in the downloads.

```
/*FIRST CREATE THE FORMAT FOR THE SURVEY VARIABLE*/
proc format;
      value survey
      1="HINTS 6"
      2="HINTS 7"
      ;
run;
/*************************************************************************/

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

/*PUT NAME OF LIBRARY WHERE HINTS 6 FORMATS ARE STORED*/
options fmtsearch=(Lib6);

data tempHINTS6;
      /*PUT NAME OF LIBRARY AND NAME OF EXISTING HINTS 6 DATA FILE*/
      set LibH6.hints6 _public;

      survey=1;


      format survey survey.;
run;


/* PUT NAME OF LIBRARY WHERE HINTS 7 FORMATS ARE STORED*/
options fmtsearch=(LibH7);

data tempHINTS7;
      /*PUT NAME OF LIBRARY AND NAME OF EXISTING HINTS 6 DATA FILE*/
      set
      LibH7.hints7_public
      ; 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
(Merged_NWGT0) AND 100 REPLICATE WEIGHTS (Merged_NWGT1 THRU
Merged_NWGT100)*/

data mergeHINTS6_HINTS7;
      set tempHINTS6 tempHINTS7;
      /*Create Replicate Weights for trend tests*/
      **Replicate Weights;
      array hints6wgts [50] person_finwt1-
      person_finwt50; array hints7wgts [50]
      person_finwt1-person_finwt50; array
      Merged_NWgt [100] Merged_NWGT1-Merged_NWGT100;

      **Adjust Final And Replicate Weights;
      if survey eq 1 then do i=1 to 50;   *HINTS 6;
            Merged_NWGT0=person_finwt0;
            Merged_NWgt[i]=hints6wgts[i];
            Merged_NWgt[50+i]=person_finwt0;
      end;

      else if survey eq 2 then do i=1 to 50; *HINTS
            7; Merged_NWGT0= person_finwt0;
            Merged_NWGT0=person_finwt0;
            Merged_NWgt[i]=person_finwt0;
            Merged_NWgt[50+i]=hints7wgts[i];
      end;
run;


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

proc surveyfreq data = mergeHINTS6_HINTS7 varmethod = jackknife;
      weight Merged_NWGT0;
      repweights Merged_NWGT1-Merged_NWGT100 / df = 98 jkcoefs = 0.98;
      tables seekcancerinfo chanceaskquestions;
run;
```

SAS Code to Merge HINTS 7 and HINTS 6 for the Taylor Series Linearization Method

```
/*THIS CODE MERGES TWO TEMPORARY HINTS DATA SETS CREATED 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*/
```

```
/*FIRST CREATE THE FORMAT FOR THE SURVEY VARIABLE*/
proc format;
      value survey
      1="HINTS 6"
      2="HINTS 7"
      ;
run;

/*************************************************************************/
/*CREATE TWO SEPARATE TEMPORARY DATA FILES THAT CONTAIN THE NEW
'SURVEY' VARIABLE AND BOTH CONTAIN THE SAME WEIGHT VARIABLES.*/
/* NOTE THAT IN THIS EXAMPLE WE USE THE PERSON_FINWT0 VARIABLE AS OUR
WEIGHTING VARIABLE FROM HINTS 6.
*/

/*PUT NAME OF LIBRARY WHERE HINTS 6 FORMATS ARE STORED*/
options fmtsearch=(LibH6);

data tempHINTS6;
      /*PUT NAME OF LIBRARY AND NAME OF EXISTING HINTS 6 DATA FILE*/
      set Lib6.hints6_public;
      RENAME
      PERSON_FINWT0=MERGED_FINWT0;

      survey=1;
      format survey survey.;
run;

/* PUT NAME OF LIBRARY WHERE HINTS 6 FORMATS ARE STORED*/
options fmtsearch=(LibH7);

data tempHINTS7;
      /*PUT NAME OF LIBRARY AND NAME OF EXISTING HINTS 7 DATA FILE*/
      set LibH7.hints7_public;
      RENAME PERSON_FINWT0=MERGED_FINWT0;

      survey=2;
      format survey survey.;
run;

data mergeHINTS6_HINTS7;
      set tempHINTS6 tempHINTS7;
run;

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

proc surveyfreq data = MergeHints6_Hints7 varmethod =
      TAYLOR; strata VAR_STRATUM;
      cluster
      VAR_CLUSTER;
      weight
      MERGED_FINWT0;
      tables seekcancerinfo chanceaskquestions / row col;
run;
```

Merging HINTS 7 and HINTS 6 using SPSS

This section provides SPSS (Version 22) syntax for merging the HINTS 7, and HINTS 6 data and uses Taylor linearization for variance estimates. Note that the below sample syntax is created with the assumption that there were no group differences found within HINTS 7.

Within the below example SPSS syntax, a new "survey" variable is created in both datasets that will distinguish between the two iterations once the datasets are merged. 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 = 13,580).

First, you will need to have **HINTS 7** data open. The below syntax will first save a copy of HINTS 7 and rename it as a new file called '. We highly suggest this step for several reasons, mainly being that when SPSS merges datasets the old file may be overwritten. By saving your original datafile, you can always have this available to refer to. Next, the syntax will rename the dataset to help with making sure the correct dataset is active and being edited in later syntax.
Next, the below syntax copies HINTS 6's weighting variable PERSON_FINWT0 so that both cycles' weighting variable names match (MERGED_FINWT0). Finally, the syntax creates a new variable called 'Survey' and gives each participant in HINTS 7 a "2" so that analysts can easily identify cases from this iteration.

```
**below, you should insert the filepath for your HINTS 7 data**.
SAVE OUTFILE='INSERT YOUR FILE PATH HERE\MERGED_H7.sav'
/COMPRESSED.
DATASET NAME MERGED_DATA.

DATASET ACTIVATE MERGED_DATA.
COMPUTE
MERGED_FINWT0=PERSON_FINWT0.
COMPUTE
Survey=2.
EXECUTE.
```

Next, we need to open our HINTS 6 data and rename our datafile, again to help with keeping files aligned for the merging process below. The following code will open your HINTS 6 data and rename the dataset as H6. The syntax will then create the 'Survey' variable in the HINTS 6 dataset and give each participant from HINTS 6 a value of "1". Again, this is so that once the datasets are merged, analysts can easily identify which cases were from the HINTS 6 dataset. Finally, the syntax creates copies the weighting variable PERSON_FINWT0 and names it MERGED_FINWT0 so that the key weighting variable matches the key weighting variable from our HINTS 7 dataset
Note, the analyst will need to insert the file path for where HINTS 6 is saved.

```
**below, you should insert the file path for your HINTS 6 data**.
GET FILE='INSERT YOUR FILE PATH HERE\hints6_public.sav'.
DATASET NAME H6 WINDOW=FRONT.
COMPUTE MERGED_FINWT0=PERSON_FINWT0.
COMPUTE
Survey=1.
EXECUTE.
```

Next, a plan file is required to conduct analyses in SPSS. To create a plan file and subsequently conduct analyses, paste the following syntax in the SPSS Syntax Editor:

```
* Analysis Preparation Wizard.
*INSERT DATH OF PATH TO SAMPLE DESIGN FILE IN /PLAN
FILE=. CSPLAN ANALYSIS
/PLAN FILE='INSERT YOUR FILE PATH HERE\MergePlan.csaplan'
/PLANVARS ANALYSISWEIGHT=MERGED_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 should, by this point, have two datasets open: “MERGED_H7” (which currently contains only HINTS 7 data) and “hints6_public”. Within your “MERGED_H7” dataset 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 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 are in the “Variables in New Active Dataset” box. Once you have verified all your desired variables are in that box, click “OK”.

```
DATASET ACTIVATE
MERGED_DATA. ADD FILES
/FILE=*
/FILE='H6'.
EXECUTE.

*YOU CAN USE THE CODE BELOW TO RUN SIMPLE FREQUENCIES ON TWO COMMON
VARIABLES, ‘seekcancerinfo’ AND ‘chanceaskquestions’.

*INSERT PATH OF TO ANALYSIS PLAN UNDER
/PLAN FILE. CSTABULATE
/PLAN FILE='INSERT YOUR FILE PATH HERE\MergePlan.csaplan'
/TABLES VARIABLES=seekcancerinfo chanceaskquestions
/CELLS POPSIZE TABLEPCT
/STATISTICS SE COUNT
/MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

Merging HINTS 7 and HINTS 6 using Stata

This section provides Stata (Version 10.0 and higher) code for merging the HINTS 7 and HINTS 6 data. The analyst will need to use the Rizzo, et al., (2008) method to create one new final sample weight (MERGED_NWGT0) and 100 new replicate weights (MERGED_NWGT1 thru MERGED_NWGT100) when using the replicate method for variance estimation.

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

In order to combine HINTS 7 with HINTS 6, the below sample code creates two temporary data files and generates the appropriate final sample weight (for population point estimates; MERGED_NWGT0) and 100 replicate weights (MERGED_NWGT1 through MERGED_NWGT100; to compute standard errors) on each, using the Rizzo, et al., (2008) method. Next, the two files are merged into one and the new “survey” variable is generated to distinguish between the two iterations. This survey variable can later be used to easily differentiate the cases that came from each HINTS iteration. During the merge, Stata will match up

variables that have the same name and format, creating a final merged data file (n =13,530). Note that variable names are case sensitive in Stata.

```
*Put path and name to your HINTS 6 data
use "INSERT YOUR PATH HERE\hints6_public.dta", clear

*Create final and replicate weights (merged_nwt*) for multi-cycle
datasets
gen merged_nwgt0=PERSONFINWT0

forvalues n1=1/50 {

local x1=`n1'+50
gen merged_nwgt`n1'=PERSON_FINWT`n1'
gen merged_nwgt`x1'=PERSON_FINWT0

}

save h6.dta, replace

*Put path and name to your HINTS 7 data
use "INSERT YOUR PATH HERE\hints7_public.dta", clear

*Create final and replicate weights (merged_nwt*) for multi-cycle
datasets

gen merged_nwgt0=PERSON_FINWT0

forvalues n2=1/50 {

local x2=`n2'+50
gen merged_nwgt`n2'=PERSON_FINWT0
gen merged_nwgt`x2'=PERSON_FINWT`n2'

}

save h7.dta, replace

set trace off

*Combine the 2 cycles of data & generate survey variable flagging HINTS
iteration
use h6.dta, clear

append using h7.dta, generate(survey)

label define survey 0 "HINTS 6" 1 "HINTS 7"

label values survey survey

save combined.dta, replace

* Use the code below to run simple one-way frequencies for 2 common
variables
** First, declare survey design

svyset [pw=merged_nwgt0], jkrw(merged_nwgt1-merged_nwgt100,
```

```
multiplier(0.98)) vce(jack) dof(98) mse

svy: tabulate seekcancerinfo, obs percent se

svy: tabulate chanceaskquestions, obs percent se
```

Stata Code to Merge HINTS 7 and HINTS 6 for the Taylor Series Linearization Method

In order to combine HINTS 7 with HINTS 6, the below sample code creates two temporary data files and generates the appropriate final sample weight (for population point estimates; MERGED_NWGT0) on each. No transformations are needed to the VAR_CLUSTER and VAR_STRATUM variables to support computation of standard errors. Next, the two files are merged into one and the new "survey" variable is generated to distinguish between the two iterations. This survey variable can later be used to easily differentiate the cases that came from each HINTS iteration. During the merge, Stata will match up variables that have the same name and format, creating a final merged data file (n = 13,530).

```
*Put path and name to your HINTS 6 data
use "INSERT YOUR PATH HERE\hints6_public.dta", clear

*Create final weight (merged_nwt0) for multi-cycle
datasets gen merged_nwgt0=PERSONFINWT0
save h6.dta, replace

*Put path and name to your HINTS 7 data
use "INSERT YOUR PATH HERE\hints7_public.dta", clear

*Create final weight (merged_nwt0) for multi-cycle
datasets

gen merged_nwgt0=PERSONFINWT0
save h7.dta, replace

*Combine the 2 cycles of data & generate survey variable flagging HINTS
iteration
use h6.dta, clear
append using h7.dta, generate(survey)
label define survey 0 "HINTS 6" 1 "HINTS 7"
label values survey survey
save combined.dta, replace

* Use the code below to run simple one-way frequencies for 2 common
variables
** First, declare survey design

svyset var_cluster [pw=merged_nwgt0], strata(var_stratum)

svy: tabulate seekcancerinfo, obs percent se
svy: tabulate chanceaskquestions, obs percent se
```

Merging HINTS 7 and HINTS 6 using R

This section provides R syntax for merging the HINTS 7 and HINTS 6 iterations. The code below loads HINTS 7 and HINTS 6 SAS files into R as separate data objects (make sure both files are in the same working directory).

Within the below example R syntax, appropriate final sample weight (for population point estimates; ngwt0) and 100 replicate weights (nwgt1 through nwgt100; to compute standard errors) are generated, using the Rizzo, et al., (2008) method. Next, a new "hints_edition" variable is created in both datasets that will distinguish between the two iterations once the datasets are merged. Once the two files are merged into one, variables that have the same name and format will be matched up to create a merged data file (n = 13,580).

```
# Load Required Packages

library(haven) # For loading data from SAS, SPSS, or STATA into R
library(dplyr) # For data manipulation
library(survey) # For analyzing complex survey data
library(srvyr) # For manipulating survey objects with dplyr


# Setting the working directory to file location
setwd('[WORKING DIRECTORY HERE]')
```

```r
# HINTS 7 file
df_H7    =    haven::read_sas("hints7_public.sas7bdat")

# HINTS 6 file
df_H6     =      haven::read_sas("hints6_public.sas7bdat")

# Create variable names
nwgt_var_names = c(paste0('nwgt', 1:100))
var_names = c(paste0('PERSON_FINWT', 1:50))

# Create Hints 6 group weights
df_H6 = df_H6 |>
  dplyr::mutate(hints_edition = 'Hints 6') |> dplyr::mutate(nwgt0 =
  PERSON_FINWT0)

for(i in 1:100){
  if(i <= 50){
       df_H6[nwgt_var_names[i]] = df_H6[var_names[i]]
  }

  if(i > 50){
    df_H6[nwgt_var_names[i]]   =   df_H6$PERSON_FINWT0
  }
}

# Create Hints 7 group weights
df_H7 = df_H7 |>
  dplyr::mutate(hints_edition = 'HINTS 7') |>
  dplyr::mutate(nwgt0 = PERSON_FINWT0)

for(i in 1:100){
  if(i <= 50){
   df_H7[nwgt_var_names[i]]   =   df_H7$PERSON_FINWT0
  }

  if(i > 50){
    df_H7[nwgt_var_names[i]] = df_H7[var_names[i-50]]
  }
}
```

The below syntax will merge the HINTS 7 and HINTS 6 datasets into a new file called ‘df_multi’. We highly suggest this step for several reasons, mainly being that when R merges datasets the old file may be overwritten. By saving your original datafile, you can always have this available to refer to.

```r
# Merge the data sets
df_multi = plyr::rbind.fill(df_H6, df_H7)

# Display number of respondents from both survey editions
table(df_multi$hints_edition)
```

```
##
## Hints 6                    HINTS 7
##              6252                    7278
```

The example code below can be used to run simple frequencies on two common variables ("SeekCancerInfo" and "ChanceAskQuestions") in the HINTS 7 and HINTS 6 merged data set using a replicate weights approach:

```
# Create the replicate weights survey design object
svy_obj_rep_merged = as_survey_rep(.data = df_multi,
                                   weights = nwgt0,
                                   repweights = num_range(prefix = "nwgt",
                                                          range = 1:100),
                                   type = "JKn",
                                   scale = 0.98,
                                   rscales = rep(1, times = 100))

# Crosstab
svy_obj_rep_merged |>
  dplyr::filter(ChanceAskQuestions > 0,
                SeekCancerInfo > 0) |>
  dplyr::group_by(ChanceAskQuestions, SeekCancerInfo) |>
  dplyr::summarize(n = n(),
                   total = survey_total(),
                   pct = survey_prop())
```

The example code below can be used to run simple frequencies on two common variables ("SeekCancerInfo" and "ChanceAskQuestions") in the HINTS 7 and HINTS 6 merged data set using a Taylor Series linearization approach. No transformations are needed to the VAR_CLUSTER and VAR_STRATUM variables to support computation of standard errors.

```
# Create the Taylor Series linearization survey design object
svy_obj_linear_merged = as_survey_design(.data = df_multi,
                                         ids = VAR_CLUSTER,
                                         strata = VAR_STRATUM,
                                         weights = PERSON_FINWT0,
                                         nest = T)

# Crosstab
svy_obj_linear_merged |>
  dplyr::filter(ChanceAskQuestions > 0,
                SeekCancerInfo > 0) |>
  dplyr::group_by(ChanceAskQuestions, SeekCancerInfo) |>
  dplyr::summarize(n = n(),
                   total = survey_total(),
                   pct = survey_prop())
```

References

Conrad, F. G., Couper, M. P., Tourangeau, R., & Zhang, C. (2017). Reducing speeding in web surveys by providing immediate feedback. Survey research methods, 11(1), 45–61. https://doi.org/10.18148/srm/2017.v11i1.6304

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

DeBell, M. (2022). The Visible Cash Effect with Prepaid Incentives: Evidence for Data Quality, Response Rates, Generalizability, and Cost. Journal of Survey Statistics and Methodology, 11(5), 991–1010. https://doi.org/10.1093/jssam/smac032

Dillman, D.A., Smyth, J.D., and Christian, L.M. (2009). Internet, mail, and mixed-mode surveys: The tailored design method. Hoboken, NJ: John Wiley and Sons.

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.

Hibben, K. C., Felderer, B., & Conrad, F. G. (2020). Respondent commitment: applying techniques from face-to-face interviewing to online collection of employment data. International Journal of Social Research Methodology, 25(1), 15–27. https://doi.org/10.1080/13645579.2020.1826647

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.

Rizzo, L., Moser, R. P., Waldron, W., Wang, Z., Davis, W.W. (2008). Analytic Methods to Examine Changes Across Years Using HINTS 2003 & 2005 Data. Retrieved from: https://hints.cancer.gov/docs/HINTS_Data_Users_Handbook-2008.pdf

Seidenberg, A.B., Moser, R.P., & West, B. (2023). Preferred Reporting Items for Complex Sample Survey Analysis (PRICSSA), Journal of Survey Statistics and Methodology, 11(4), 743-757.

Sherr, S., Wells, B.M. (2021, May 11-14). What You See Is What You Get: Evaluating the Use of Visible Incentives in the California Health Interview Survey [Conference presentation]. AAPOR 76th Annual Conference, Virtual Conference.

Vanette, D.L. (2016, May 12-15). Assessing the Effects and Effectiveness of Attention-check Questions in Web Surveys: Evidence from a 14 Country Cross-national Survey Experiment [Conference presentation]. AAPOR 71st Annual Conference, Austin, TX, United States.

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

Zhang, S., West, B. T., Wagner, J., Couper, M. P., Gatward, R., & Axinn, W. G. (2023). Visible cash, a second incentive, and priority Mail? An experimental evaluation of mailing strategies for a screening questionnaire in a national Push-to-Web/Mail survey. Journal of Survey Statistics and Methodology, 11(5), 1011–1031. https://doi.org/10.1093/jssam/smac041

Derived Variables List

DERIVED VARIABLE	VARIABLE DESCRIPTION
AgeGrpA	Respondent Age Recode-4 Levels
AgeGrpB	Respondent Age Recode-5 Levels
EducA	What is the highest level of school you completed? (Education recoded-4 levels)
EducB	What is the highest level of school you completed? (Education recoded-5 levels)
RaceEthn	Race/Ethnicity recode (Hisp_Cat and Race_Cat2--7 Levels)
RaceEthn5	Race/Ethnicity recode (Hisp_Cat and Race_Cat2--5 Levels)
HHInc	What is your (combined) annual household income? (IncomeRanges Recode-6 levels)
BMI	Body Mass Index (BMI)
TimeSinceDX	How long ago were you diagnosed with cancer?
smokeStat	Smoking status
PHQ-4	PHQ-4 total score
WeeklyMinutesModerateExercise	Minutes per week of at least moderate intensity exercise
ECIGUSE	Electronic Cigarette Use
IncomeRanges_IMP	Imputed Income Ranges Variable
AvgDrinksPerWeek	Average Number of Drinks Per Week
PCCScale	Patient Centered Communication Scale
ISEE_Scale	Information Seeking Experience Scale
PROMIS_Isolation_t	PROMIS Social Isolation Scale T Scores
SexualOrientation_Recode	Sexual Orientation Recode