OVERVIEW OF THE HINTS 6 SURVEY (2022) AND DATA ANALYSIS RECOMMENDATIONS

April 2023

CONTENTS

Overview of HINTS 3
HINTS 6 3
Methodology 3
Sample Size and Response Rates 4
Analyzing HINTS Data 4
Important Analytic Variables in the Database 5
Variance Estimation Methods: Replicate vs. Taylor Linearization 6
Denominator Degrees of Freedom (DDF) 7
Statistical Software Example Code 8
Analyzing Data Using SAS 8
Analyzing Data Using SPSS—Taylor Series 19
Analyzing Data Using Stata 28
Analyzing Data Using R 46
Merging HINTS Survey Iterations 53
Merging HINTS 6 and HINTS 5, Cycle 4 using SAS 53
Merging HINTS 6 and HINTS 5, Cycle 4 using SPSS 56
Merging HINTS 6 and HINTS 5, Cycle 4 using Stata 57
Merging HINTS 6 and HINTS 5, Cycle 4 using R 59
References 62

Overview of HINTS

The Health Information National Trends Survey (HINTS) is a nationally representative survey that has been administered every few years by the National Cancer Institute since 2003. The HINTS target population is all adults aged 18 or older in the civilian non-institutionalized population of the United States. The HINTS program collects data on the American public's need for, access to, and use of health-related information and health-related behaviors, perceptions, and knowledge. (Hesse, et al., 2006; Nelson, et al., 2004). Previous iterations include HINTS 1 (2003), HINTS 2 (2005), HINTS 3 (2007/2008), HINTS 4, Cycle 1 (2011); 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); and HINTS 5 Cycle 4 (2020).

HINTS 6

Starting with HINTS 6, data will be collected on a biennial basis. HINTS 6 draws upon the lessons learned from prior iterations of HINTS and incorporates an experimental design. A multi-mode survey was implemented using paper and web modes. Respondents were randomly assigned to one of two groups: 1) Concurrent: they were offered the two modes at the same time (web and paper) ; or 2) Sequential: they were offered one mode first (web) and the other mode later (paper). Details about the protocol for this mixed-mode experiment are detailed in the Methodology Report. For more extensive background about the HINTS program and previous data collection efforts, see Finney Rutten, et al. (2012).

Methodology

Data collection for HINTS 6 started on March 7, 2022 and concluded on November 8, 2022. HINTS 6 included two experimental conditions. HINTS 6 included an embedded methodological experiment comparing two mixed mode approaches: concurrent and sequential (also known as the control and treatment groups). Households in the concurrent (control) group received a cover letter with the link to the web survey and their unique access code as well as a paper survey with each mailing (including their first mailing). Households in the sequential (treatment) group received only a cover letter with the link to the web survey and their unique access code with their first mailing—they did not receive a paper survey in their first mailing. In subsequent mailings, these households received the link to the web survey and their unique access code as well as the paper survey. 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 6 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. Because of an unexpectedly low response, a subsample of non-respondents received a third follow-up mailing. 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. All households received a $2 incentive to encourage participation. Households in the concurrent group were also offered $10 complete the survey on the web. Refer to the HINTS 6 Methodology Report for more extensive information about the sampling and weighting procedures.
In addition to testing the concurrent and sequential mixed-mode approaches, HINTS 6 also included an additional embedded experiment meant to increase data quality by addressing issues around speeding and straight lining with web-based surveys. Straight lining is a term to describe when a respondent selects the same response to each question. Respondents who completed their survey on the web were randomly assigned to one of three groups: 1) Control (2 prompts at any time: 1 for speeding and 1 for

straight lining); 2) Treatment group 1 (4 prompts at any time: 2 for speeding and 2 for straight lining); and 3) Treatment group 2 (4 prompts: 2 in first half of the survey and 2 in the second half of the survey).

Sample Size and Response Rates

The final HINTS 6 sample consists of 6,252 respondents. Of these, 4,611 respondents were in the concurrent group, and 1,641 were in the sequential group. Note that 67 of these respondents were considered partial completers who did not answer the entire survey. A questionnaire was considered to be complete if at least 80% of Sections A and B were answered. A questionnaire was considered to be partially complete if 50%–79% of the questions were answered in Sections A and B. Household response rates were calculated using the American Association for Public Opinion Research response rate 4 (RR4) formula. The overall household response rate using the next-birthday method was 28.1%. More specifically, the concurrent group overall response rate was 29.1%, and the sequential group overall response rate was 25.8. These response rates were significantly different. 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 employed for HINTS. The issue is that the standard errors in your analyses will most likely be underestimated if you do not take into account the sampling procedure; therefore, your p-values will be smaller than they "should" be, your tests will be more liberal, and you are more likely to make a type I error. 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. Please see the Methodology Report for more information, including information on which variables were recoded or suppressed.

Important Analytic Variables in the Database

Refer to the HINTS 6 Methodology Report for more information regarding the weighting and stratification variables listed below.

Note that estimates from the 2021 American Community Survey (ACS) of the U.S. Census Bureau were used to calibrate the HINTS 6 control totals with the following variables: age, gender, education, marital status, race, ethnicity, and census region. In addition, the 2021 National Health Interview Survey (NHIS) was used to calibrate HINTS 6 data control totals regarding percent with health insurance and the 2021 National Center for Health Statistics (National Center for Health Statistics, Interactive Summary Health Statistics for Adults-2019-2021) was used for percent ever had 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 6, 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

TREATMENT_H6: This variable codes for which group the respondent was assigned: 1) Concurrent mixed-mode; 2) Sequential mixed-mode.

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.

HIGHSPANLI: This variable codes for whether the respondent was in the high Spanish linguistically isolated stratum (Yes or No).

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

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

APP_REGION: This variable codes for Appalachia subregion.

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

QDISP: This variable codes for whether the survey returned by the respondent was considered complete or partially complete. A complete questionnaire was defined as any questionnaire with at least 80% of the required questions answered in Sections A and B. A partial complete was defined as when between 50% and 79% of the questions were answered in Sections A and B. There were 67 partially complete questionnaires. One hundred forty-five questionnaires with fewer than 50% of the required questions answered in Sections A and B were coded as incompletely filled out and discarded.

INCOMERANGES_IMP: This is the income variable (INCOMERANGES) imputed for missing data. To impute for missing items, PROC HOTDECK from the SUDAAN statistical software was used. PROC HOTDECK uses the Cox-Iannacchione Weighted Sequential Hot Deck imputation method, as described by Cox (1980). The following variables were used as imputation classes given their strong association with the income variable: Education (O3), Race/Ethnicity (RaceEthn) (standard recode from O5 and O6), Do you currently rent or own your house? (O11), and how well do you speak English? (O4).

PROMPT: This variable has three levels that distinguish which prompt a web respondent received (if any): 1) Control (2 prompts at any time: 1 for speeding, 1 for straight lining); 2) Treatment group 1 (4 prompts at any time: 2 for speeding, 2 for straight lining); 3) Treatment group 2 (4 prompts: 2 in first half and 2 in the second half of the survey).

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 6 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 reflect the sample design. By providing weights for each subset of the sample, called replicate weights, end users can estimate the variance of a variety of estimators using standard weighted sums. The variability among the replicates is used to estimate the sampling variance of the point estimator.

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

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

Denominator Degrees of Freedom (DDF)

Replicate Weights: The HINTS 6 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 were merging two iterations of data and making comparisons should adjust the ddf to be 98 (49*2), etc.

Taylor Series: The HINTS 6 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 6 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 6 data, the concepts used here are generally applicable to other types of analyses. We will consider an analysis that includes gender, 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 6 SAS data, it is important to apply the SAS formats. To do this, see the "How to Format the HINTS 6 SAS Dataset" document included in the data download.

1. Download all HINTS 6 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 *"HINTS6_Final_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 *"HINTS6_Final_Formats.sas"* to create a permanent SAS format library that is used to analyze the HINTS dataset.

6. Open the SAS program "*HINTS6_Final_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 "*HINTS6_Final_Format_Assignments.sas"* to create the formatted SAS data set.

Note:

1) Make sure to run the program *"HINTS6_Final_Formats.sas"* BEFORE you run "*HINTS6_Final_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 6 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 6 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=(hints6);

proc format;     *First create some temporary formats;
      Value Genderf
      1 = "Male"
      2 = "Female";

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

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

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


data hints6;
      set hints6.hints6_public;

      /*Recode negative values to missing*/
      if birthgender = 1 then gender = 1;
      If birthgender = 2 then gender = 2;
      if birthgender in (-9, -7) then gender = .;

      /*Recode education into four levels, and negative values to
      missing*/
      if education in (1, 2) then edu = 1;
      if education = 3 then edu = 2;
```

```
      if education in (4, 5) then edu = 3;
      if education in (6, 7) then edu = 4;
      if education 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 gender genderf. 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 gender, along with a (Wald) Chi-squared test of independence. Note the syntax of the overall sample weight, Person_FINWT0, and those of the jackknife replicate weights, PERSON_FINWT1—PERSON_FINWT50. The jackknife adjustment factor for each replicate weight is 0.98. This syntax is consistent for all procedures. Other datasets that incorporate replicate weight jackknife designs will follow a similar syntax.

```
proc surveyfreq data = hints6 varmethod = jackknife;
      weight person_finwt0;
      repweights person_FINWT1-person_FINWT50 / df = 49 jkcoefs = 0.98;
      tables edu*gender / 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 6 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	HINTS6
Number of Replicates	50

Edu	gender	Frequency	Percent	Std Err of Percent	Row Percent	Std Err of Row Percent	Column Percent	Std Err of Col Percent
Less than high school	Male	155	4.0514	0.6071	59.5638	4.3824	8.2842	1.2234
	Female	228	2.7504	0.2577	40.4362	4.3824	5.3829	0.4932
	Total	383	6.8018	0.6534	100			
12 years or completed high school	Male	375	10.9269	0.7324	50.6565	1.8610	22.3430	1.4745
	Female	686	10.6437	0.4544	49.3435	1.8610	20.8314	0.8601
	Total	1061	21.5707	0.9097	100			
Some college	Male	642	18.3985	0.5018	47.1634	0.7608	37.6206	1.0270
	Female	1023	20.6117	0.3785	52.8366	0.7608	40.3403	0.7264
	Total	1665	39.0102	0.6720	100			
College graduate or higher	Male	1127	15.5286	0.1665	47.6083	0.3207	31.7522	0.4353
	Female	1582	17.0888	0.1772	52.3917	0.3207	33.4454	0.3182
	Total	2709	32.6173	0.2725	100			
Total	Male	2299	48.9054	0.3801			100	
	Female	3519	51.0946	0.3801			100	
	Total	5818	100					

Frequency Missing = 434

Rao-Scott Chi-Square Test	
Pearson Chi-Square	23.5647
Design Correction	1.9915
First-Order Chi-Square	11.8328
Second-Order Chi-Square	6.3850
DF	1.62
Pr > ChiSq	0.0272
F Value	3.9443
Num DF	1.62
Den DF	79.32
Pr > F	0.0312
Sample Size = 5818	

The row percentages above show that a higher weighted proportion of college graduates in the sample are female (52.4%) than male (47.6%). Respondents with less than a high school diploma include fewer females (40.4%) than males (59.6%). The statistic for the Chi-square test of independence and its associated p-value indicate that the distributions of educational attainment between men and women are significantly different.

Logistic Regression

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

```
/*Multivariable logistic regression of gender and education on
SeekCancerInfo*/
proc surveylogistic data= hints6 varmethod=jackknife;
      weight person_FINWT0;
      repweights person_FINWT1-person_FINWT50 / df=49 jkcoefs=0.98;
      class edu (ref="Less than high school")
            gender (ref="Male")/param=REF;
      model seekcancerinfo (descending) = gender 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 "Male" is the reference group for gender effect, while "Less than high school" is the reference group for education level effect. The option *tech=newton* requests the Newton- Raphson algorithm. The option *xconv=1e-8* helps to avoid early termination of the iteration.

Variance Estimation	
Method	Jackknife
Replicate Weights	HINTS6
Number of Replicates	50

Type 3 Analysis of Effects				
Effect	**F Value**	**Num DF**	**Den DF**	**Pf > F**
Gender	64.23	1	49	<.0001
Education	68.16	3	49	<.0001

Analysis of Maximum Likelihood Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**T value**	**Pr > |t|**	**95% confidence limits**	
Intercept	49	-1.5423	0.2282	-6.76	<.0001	-2.0010	-1.0837
Gender	49	0.7142	0.0891	8.01	<.0001	0.5351	0.8933
12 years or completed high school	49	0.2099	0.2509	0.84	0.4069	-0.2943	0.7140
Some College	49	0.9454	0.2455	3.85	0.0003	0.4520	1.4388
College graduate or higher	49	1.5114	0.2226	6.79	<.0001	1.0642	1.9587

(continued on the next page)

Odds Ratio Estimates

Effect	Point Estimate	95% Confidence Limits	
Female vs Male	2.043	1.708	2.443
12 years or completed high school vs Less than high school	1.234	0.745	2.042
Some College vs Less than high school	2.574	1.571	4.216
College graduate or higher vs Less than high school	4.533	2.898	7.090

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, females appear to have 2.04 times higher odds than males to have searched for cancer information.

Linear Regression

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

```
/*Multivariable linear regression of gender and education on
GeneralHealth*/
proc surveyreg data= hints6 varmethod=jackknife;
      weight PERSON_FINWT0;
      repweights PERSON_FINWT1-PERSON_FINWT50 / df=49 jkcoefs=0.98;
      class edu (ref="Less than high school")
            gender (ref="Male");
      model generalhealth = edu gender /solution;
run;
```

Variance Estimation	
Method	Jackknife
Replicate Weights	HINTS6
Number of Replicates	50

(continued on the next page)

Estimated Regression of Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	3.0815626	0.10024930	30.74	<.0001
12 years or completed high school	-0.3462532	0.12124913	-2.86	0.0063
Some College	-0.4540517	0.11396035	-3.98	0.0002
College graduate or higher	-0.7772833	0.10809175	-7.19	<.0001
Female	0.0750102	0.03756681	2.00	0.0514

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. However, there's no significant difference in health score between males and females (p=0.0514).

Tests of Model Effects

Contrast	Num DF	F Value	Pr > F
Model	4	33.80	<.0001
Intercept	1	9280.05	<.0001
Education	3	41.25	<.0001
Gender	1	3.99	0.0514

The table labeled Test of Model Effects also shows that the association between gender and general health is not significant, but the association between education and general health is 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 gender, along with a (Wald) Chi-squared test of independence. Note the syntax of the strata VAR_STRATUM, cluster VAR_CLUSTER, and overall sample weight PERSON_FINWT0. This syntax is consistent for all procedures. Other analyses that use Taylor Series approximation will follow a similar syntax.

```
proc surveyfreq data = hints6 varmethod = TAYLOR;
      strata VAR_STRATUM;
      cluster VAR_CLUSTER;
      weight person_finwt0;
      tables edu*gender / 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	6252
Sum of Weights	258418467

edu	gender	Frequency	Percent	Std Err of Percent	Row Percent	Std Err of Row Percent	Column Percent	Std Err of Col Percent
Less than high school	Male	155	4.0514	0.5744	59.5638	4.2369	8.2842	1.1613
	Female	228	2.7504	0.2737	40.4362	4.2369	5.3829	0.5163
	Total	383	6.8018	0.6280	100			
12 years or completed high school	Male	375	10.9269	0.8082	50.6565	2.4565	22.3430	1.5466
	Female	686	10.6437	0.5690	49.3435	2.4565	20.8314	0.9780
	Total	1061	21.5707	0.9067	100			
Some college	Male	642	18.3985	1.1159	47.1634	2.1571	37.6206	1.8709
	Female	1023	20.6117	0.9592	52.8366	2.1571	40.3403	1.5800
	Total	1665	39.0102	1.2409	100			
College graduate or higher	Male	1127	15.5286	0.7025	47.6083	1.4903	31.7522	1.5200
	Female	1582	17.0888	0.7097	52.3917	1.4903	33.4454	1.2778
	Total	2709	32.6173	1.0250	100			
Total	Male	2299	48.9054	1.1621			100	
	Female	3519	51.0946	1.1621			100	
	Total	5818	100					

Frequency Missing = 434

(continued on the next page)

Rao-Scott Chi-Square Test	
Pearson Chi-Square	23.5647
Design Correction	3.1565
First-Order Chi-Square	7.4655
Second-Order Chi-Square	7.0736
DF	2.84
Pr > ChiSq	0.0618
F Value	2.4885
Num DF	2.84
Den DF	557.13
Pr > F	0.0629
Sample Size = 5818	

The row percentages above show that a higher weighted proportion of college graduates in the sample are females (52.4%) than males (47.6%). Respondents with less than a high school diploma include fewer females (40.4%) than males (59.6%). The Chi-squared test of independence statistic and associated p value suggest that one may fail to reject the null hypothesis that the two variables are not associated, which indicates that there is not a significant difference between the distributions of educational attainment for these two groups.

The results of these tests based on Taylor Series linearization contradict the results using replication shown in the previous section (in SAS, the distributions of educational attainment between males and females were determined to be statistically different using the replication method). This is a good example of how the variance estimation method used can affect the outcome of a statistical test. Both education and gender are variables used in the raking process as part of the HINTS weighting procedure. As a result, the standard errors based on replication are much smaller than those based on Taylor Series linearization, which in turn results in significant differences 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 gender and education on
SeekCancerInfo*/
proc surveylogistic data= hints6 varmethod=TAYLOR;
      strata VAR_STRATUM;
      cluster VAR_CLUSTER;
      weight person_FINWT0;
      class edu (ref="Less than high school")
            gender (ref="Male")/param=REF;
      model seekcancerinfo (descending) = gender 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 gender effect, while "Less than high school" is the reference group for education level effect. The option tech=newton requests the Newton-Raphson algorithm. The option xconv=1e-8 helps to avoid early termination of the iteration.

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

Type 3 Analysis of Effects				
Effect	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Gender	62.06	1	196	<.0001
Education	62.02	3	194	<.0001

Analysis of Maximum Likelihood Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**	**Pf > |t|**	**95% Confidence Limits**	
Intercept	196	-1.5423	0.2209	-6.98	<.0001	-1.9779	-1.1068
Gender	196	0.7142	0.0907	7.88	<.0001	0.5354	0.8930
12 years or completed high school	196	0.2099	0.2377	0.88	0.3784	-0.2589	0.6787
Some College	196	0.9454	0.2279	4.15	<.0001	0.4960	1.3948
College graduate or higher	196	1.5114	0.2132	7.09	<.0001	1.0910	1.9319

Odds Ratio Estimates

Effect	Point Estimate	95% Confidence Limits	
Female vs Male	2.043	1.708	2.443
12 years or completed high school vs Less than High School	1.234	0.772	1.971
Some College vs Less than High School	2.574	1.642	4.034
College graduate or higher vs Less than High School	4.533	2.977	6.902

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, females appear to have statistically higher odds than males to have searched for cancer information.

Linear Regression

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

```
/*Multivariable linear regression of gender and education on
GeneralHealth*/
proc surveyreg data= hints6 varmethod=TAYLOR;
      strata VAR_STRATUM;
      cluster VAR_CLUSTER;
      weight person_FINWT0;
      class edu (ref="Less than high school")
      gender (ref="Male");
      model generalhealth = edu gender/solution;
run;
```

Estimated Regression of Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	3.0815626	0.10199601	30.21	<.0001
Female	0.0750102	0.04059290	1.85	0.0661
12 years or completed high school	-0.3462532	0.11270882	-3.07	0.0024
Some College	-0.4540517	0.10739644	-4.23	<.0001
College graduate or higher	-0.7772833	0.10888438	-7.14	<.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. We do not interpret the estimates for Female because the corresponding p-value is greater than .05.

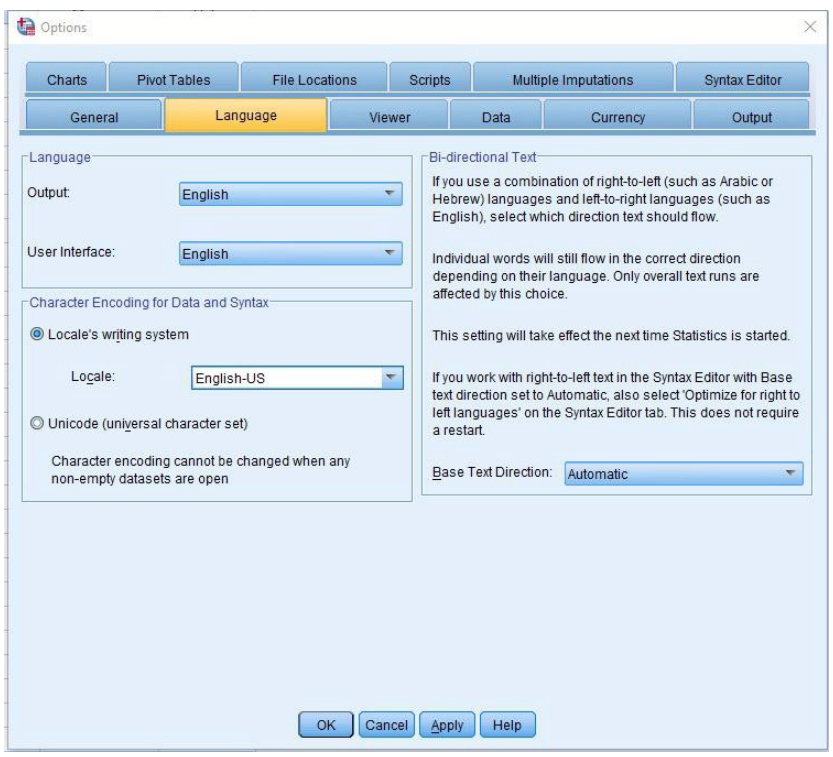
Tests of Model Effects

Contrast	Num DF	F Value	Pr > F
Overall model	4	34.18	<.0001
Intercept	1	8410.91	<.0001
Gender	1	3.41	0.0661
Education	3	38.86	<.0001

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

Analyzing Data Using SPSS—Taylor Series

Prior to opening the HINTS 6 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 6 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 6 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 BirthGender (1=1) (2=2) (ELSE=SYSMIS) INTO gender.
VARIABLE LABELS gender 'gender'.
EXECUTE.

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

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

*Recode negative values to missing for CSGLM procedure.
```

```
RECODE GeneralHealth (1 thru 5=Copy) (ELSE=SYSMIS) INTO genhealth_recode.
VARIABLE LABELS genhealth_recode 'genhealth_recode'.
EXECUTE.

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

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

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

Frequency Table and Chi-Square Test

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

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

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

Edu			Gender		
			Male	Female	Total
Less than high school	Population Size	Estimate	9673126.590	6566822.070	16239948.66
		Standard Error	1416389.370	651864.583	1566287.245
		Unweighted Count	155	228	383
	% within edu	Estimate	59.6%	40.4%	100.0%
		Standard Error	4.2%	4.2%	0.0%
		Unweighted Count	155	228	383
	% within gender	Estimate	8.3%	5.4%	6.8%
		Standard Error	1.2%	0.5%	0.6%
		Unweighted Count	155	228	383
	% of Total	Estimate	4.1%	2.8%	6.8%
		Standard Error	0.6%	0.3%	0.6%
		Unweighted Count	155	228	383
12 years or completed high school	Population Size	Estimate	26089156.55	25412934.25	51502090.80
		Standard Error	2031080.146	1382808.472	2369688.448
		Unweighted Count	375	686	1061
	% within edu	Estimate	50.7%	49.3%	100.0%
		Standard Error	2.5%	2.5%	0.0%
		Unweighted Count	375	686	1061
	% within gender	Estimate	22.3%	20.8%	21.6%
		Standard Error	1.5%	1.0%	0.9%
		Unweighted Count	375	686	1061
	% of Total	Estimate	10.9%	10.6%	21.6%
		Standard Error	0.8%	0.6%	0.9%
		Unweighted Count	375	686	1061
Some college	Population Size	Estimate	43928337.38	49212484.77	93140822.15
		Standard Error	3284813.460	2385657.541	4177434.204
		Unweighted Count	642	1023	1665
	% within edu	Estimate	47.2%	52.8%	100.0%

		Standard Error	2.2%	2.2%	0.0%
		Unweighted Count	642	1023	1665
	% within gender	Estimate	37.6%	40.3%	39.0%
		Standard Error	1.9%	1.6%	1.2%
		Unweighted Count	642	1023	1665
	% of Total	Estimate	18.4%	20.6%	39.0%
		Standard Error	1.1%	1.0%	1.2%
		Unweighted Count	642	1023	1665
College graduate or higher	Population Size	Estimate	37075981.84	40801112.37	77877094.21
		Standard Error	1630403.356	1558490.974	2195330.245
		Unweighted Count	1127	1582	2709
	% within edu	Estimate	47.6%	52.4%	100.0%
		Standard Error	1.5%	1.5%	0.0%
		Unweighted Count	1127	1582	2709
	% within gender	Estimate	31.8%	33.4%	32.6%
		Standard Error	1.5%	1.3%	1.0%
		Unweighted Count	1127	1582	2709
	% of Total	Estimate	15.5%	17.1%	32.6%
		Standard Error	0.7%	0.7%	1.0%
		Unweighted Count	1127	1582	2709
Total	Population Size	Estimate	116766602.4	121993353.5	238759955.8
		Standard Error	4509117.143	2786627.776	5095266.521
		Unweighted Count	2299	3519	5818
	% within edu	Estimate	48.9%	51.1%	100.0%
		Standard Error	1.2%	1.2%	0.0%
		Unweighted Count	2299	3519	5818
	% within gender	Estimate	100.0%	100.0%	100.0%
		Standard Error	0.0%	0.0%	0.0%
		Unweighted Count	2299	3519	5818
	% of Total	Estimate	48.9%	51.1%	100.0%

		Standard Error	1.2%	1.2%	0.0%
		Unweighted Count	2299	3519	5818

The row percentages above show that a higher weighted proportion of college graduates in the sample are females (52.4%) than males (47.6%). Respondents with less than a high school diploma include more males (59.6%) than females (40.4%).

Tests of Independence

		Chi-Square	Adjusted F	df1	df2	Significance
edu * gender	Pearson	23.565	2.696	2.871	562.786	.048
	Likelihood Ratio	23.648	2.705	2.871	562.786	.047

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 gender and education on SeekCancerInfo.
CSLOGISTIC seekcancerinfo_recode (LOW) BY flippedgender flippededu
/PLAN FILE='(sample.csaplan)'
/MODEL flippedgender flippededu
/CUSTOM Label = 'Overall model minus intercept'
LMATRIX = flippedgender 1/2 -1/2;
        flippededu 1/3 1/3 1/3 -1;
        flippededu 1/3 1/3 -1 1/3 ;
        flippededu 1/3 -1 1/3 1/3;
        flippededu -1 1/3 1/3 1/3
/CUSTOM Label = 'Gender'
LMATRIX = flippedgender 1/2 -1/2
/CUSTOM Label = 'Education overall'
LMATRIX = flippededu 1/3 1/3 1/3 -1;
        flippededu 1/3 1/3 -1 1/3 ;
        flippededu 1/3 -1 1/3 1/3;
        flippededu -1 1/3 1/3 1/3
/INTERCEPT INCLUDE=YES SHOW=YES
/STATISTICS PARAMETER SE CINTERVAL TTEST EXP
/TEST TYPE=CHISQUARE PADJUST=LSD
/ODDSRATIOS FACTOR=[flippedgender(HIGH)]
/ODDSRATIOS FACTOR=[flippededu(HIGH)]
/MISSING CLASSMISSING=EXCLUDE
/CRITERIA MXITER=100 MXSTEP=50 PCONVERGE=[1e-008 RELATIVE] LCONVERGE=[0] CHKSEP=20
```

CILEVEL=95
/PRINT SUMMARY COVB CORB VARIABLEINFO SAMPLEINFO.

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

Sample Design Information

		N
Unweighted Cases	**Valid**	5802
	Invalid	450
	Total	6252
Population Size		238177982.3
Stage 1	**Strata**	4
	Units	200
Sampling Design Degrees of Freedom		196

(continued on the next page)

Parameter Estimates

seekcancerinfo_recode		B	Std. Error	95% Confidence Interval Lower	95% Confidence Interval Upper	Hypothesis Test t	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B) Lower	95% Confidence Interval for Exp(B) Upper
Yes	**(Intercept)**	-1.542	0.221	-1.978	-1.107	-6.986	196	0.000	0.214	0.138	0.331
	Female	0.714	0.091	0.535	0.893	7.881	196	0.000	2.043	1.708	2.442
	College graduate or higher	1.511	0.213	1.091	1.932	7.092	196	0.000	4.533	2.978	6.901
	Some college	0.945	0.228	0.496	1.395	4.150	196	0.000	2.574	1.642	4.034
	12 years or completed high school	0.210	0.238	-0.259	0.679	0.883	196	0.378	1.234	0.772	1.971

Odds Ratios

		seekcancerinfo_recode	Odds Ratio	95% Confidence Interval Lower	95% Confidence Interval Upper
Gender	Female vs. Male	Yes	2.043	1.708	2.442
Education	College graduate or higher vs. Less than high school	Yes	4.533	2.978	6.901
	Some college vs. Less than high school	Yes	2.574	1.642	4.034
	12 years or completed high school vs. Less than high school	Yes	1.234	.772	1.971

Overall Model Minus Intercept

df	Wald Chi-Square	Sig.
4.000	210.977	0.000

Gender

df	Wald Chi-Square	Sig.
1.000	62.104	0.000

Education Overall

df	Wald Chi-Square	Sig.
3.000	188.109	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, females appear to be statistically more likely than males to have searched for cancer information.

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

Linear Regression

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

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

Sample Design Information

		N
Unweighted Cases	**Valid**	5776
	Invalid	476
	Total	6252
Population Size		237958448.1
Stage 1	**Strata**	4
	Units	200
Sampling Design Degrees of Freedom		196

Parameter Estimates

Parameter	Estimate	Std. Error	95% Confidence Interval		Hypothesis Test		
			Lower	Upper	t	df	Sig.
(Intercept)	3.082	0.102	2.880	3.283	30.223	196	0.000
College graduate or higher	-0.777	0.109	-0.992	-0.563	-7.141	196	0.000
Some college	-0.454	0.107	-0.666	-0.242	-4.229	196	0.000
12 years or completed high school	-0.346	0.113	-0.568	-0.124	-3.073	196	0.002
Female	0.075	0.041	-0.005	0.155	1.849	196	0.066

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. We do not interpret the estimates for gender because the corresponding p-values for female are greater than .05.

Overall Model Minus Intercept

df1	df2	Wald F	Sig.
4	193	33.679	0.000

Gender

df1	df2	Wald F	Sig.
1	196	3.417	0.066

Education Overall

df1	df2	Wald F	Sig.
3	194	38.493	0.000

From the above table, we can see that education, but not gender, is 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 6 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 6 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\hints6_public.dta"

* Recode negative values to missing
recode BirthGender (1=1 "Male") (2=2 "Female") (nonmissing=.),

generate(gender)

label variable gender "Gender"

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

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


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


replace SeekCancerInfo = 0 if SeekCancerInfo == 2

replace SeekCancerInfo = . if SeekCancerInfo == -1 | SeekCancerInfo == -2 |
SeekCancerInfo == -6 | SeekCancerInfo == -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 gender, cell format(%8.5f) percent se wald noadjust
svy: tabulate edu gender, row format(%8.5f) percent se wald noadjust
svy: tabulate edu gender, 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.

(results on subsequent pages)

Jknife *: for cell counts

Number of strata = 1

Number of obs = 5,818
Population size = 238,759,956
Replications = 50
Design df = 49

Education	Gender Male	Female	Total
Less than high school	4.05140	2.75039	6.80179
	(0.60706)	(0.25767)	(0.65339)
12 years	10.92694	10.64372	21.57066
	(0.73237)	(0.45437)	(0.90967)
Some college	18.39854	20.61170	39.01024
	(0.50176)	(0.37847)	(0.67204)
College	15.52856	17.08876	32.61732
	(0.16654)	(0.17721)	(0.27253)
Total	48.90544	51.09456	1.0e+02
	(0.38012)	(0.38012)	

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

Wald (Pearson):

Unadjusted	chi2(3)	=	17.2854	
Unadjusted	F(3, 49)	=	5.7618	P = 0.0019
Adjusted	F(3, 47)	=	5.5266	P = 0.0025

Jknife *: for rows

Number of strata	= 1	Number of obs	= 5,818
		Population size	= 238,759,956
		Replications	= 50
		Design df	= 49

Education	Gender Male	Female	Total
Less than high school	59.56378	40.43622	1.0e+02
	(4.38243)	(4.38243)	
12 years	50.65650	49.34350	1.0e+02
	(1.86104)	(1.86104)	
Some college	47.16336	52.83664	1.0e+02
	(0.76075)	(0.76076)	
College	47.60833	52.39167	1.0e+02
	(0.32066)	(0.32066)	
Total	48.90544	51.09456	1.0e+02
	(0.38012)	(0.38012)	

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

Wald (Pearson):

Unadjusted	chi2(3)	=	17.2854	
Unadjusted	F(3, 49)	=	5.7618	P = 0.0019
Adjusted	F(3, 47)	=	5.5266	P = 0.0025

```
Jknife *: for columns

Number of strata   =          1                Number of obs      =         5,818
                                               Population size    =   238,759,956
                                               Replications       =            50
                                               Design df          =            49
```

Education	Gender: Male	Female	Total
Less than high school	8.28416	5.38293	6.80179
	(1.22341)	(0.49322)	(0.65339)
12 years	22.34300	20.83141	21.57066
	(1.47446)	(0.86011)	(0.90967)
Some college	37.62064	40.34030	39.01024
	(1.02703)	(0.72638)	(0.67204)
College	31.75221	33.44536	32.61732
	(0.43529)	(0.31820)	(0.27253)
Total	1.0e+02	1.0e+02	1.0e+02

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

Wald (Pearson):
  Unadjusted     chi2(3)          =   17.2854
  Unadjusted     F(3, 49)         =   5.7618       P = 0.0019
  Adjusted       F(3, 47)         =   5.5266       P = 0.0025
```

For the purposes of computing appropriate degrees of freedom for the estimator of the HINTS 6 differences, we can assume as an approximation that the sample is a simple random sample of size 50 (corresponding to the 50 replicates: each replicate provides a “pseudo sample unit”) from a normal distribution. The denominator degrees of freedom (df) is equal to 49*k, where k is the number of iterations of data used in this analysis. Stata uses the number of replicates minus one as the denominator degrees of freedom and does not provide the option for the user to specify the denominator degrees of freedom.

Logistic Regression

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

```
* Define reference group for categorical variables for both svy: logit and
svy: regress
char gender [omit] 1
```

```
char edu [omit] 1
* Multivariable logistic regression of gender and

education on SeekCancerInfo

xi: svy: logit SeekCancerInfo i.gender i.edu
test _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4 _cons, nosvyadjust
test _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust
test _Igender_2, nosvyadjust
test _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust
xi: svy, or: logit SeekCancerInfo i.gender i.edu
```

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

```
. xi: svy: logit SeekCancerInfo i.gender  i.edu
     i.gender   _Igender_1-2     (naturally coded; _Igender_1 omitted)
     i.edu      _Iedu_1-4        (naturally coded; _Iedu_1 omitted)
     (running logit on estimation sample)

Survey: Logistic regression

Number of  strata       =      1              Number of obs            =             5,802
                                              Population size          =       238,177,982
                                              Replications             =                50
                                              Design df                =                49
                                              F(     4,       46)      =             51.43
                                              Prob > F                 =            0.0000
```

seekcancerinfo	Coef.	Jknife * Std. Err.	t	P>|t|	[95% Conf.	Interval]
_Igender_2	.7142369	.0891223	8.01	0.000	.535139	.8933348
_Iedu_2	.2098593	.2508817	0.84	0.407	-.2943063	.7140249
_Iedu_3	.9453871	.2455282	3.85	0.000	.4519798	1.438794
_Iedu_4	1.511432	.2225701	6.79	0.000	1.064161	1.958703
_cons	-1.54231	.2282339	-6.76	0.000	-2.000963	-1.083657

```
Unadjusted Wald test

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

       F( 5, 49)   =     44.01
       Prob > F    =     0.0000

Unadjusted Wald test

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

       F( 4, 49)   =     54.78
       Prob > F    =     0.0000

Unadjusted Wald test

( 1)  [seekcancerinfo]_Igender_2 = 0

       F(  1,49)   =     64.23
       Prob > F    =     0.0001


Unadjusted Wald test

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

       F(  3,49)   =     68.16
       Prob > F    =     0.0000
```

```
i.gender   _Igender_1-2  (naturally coded; _Igender_1 omitted)
i.edu      _Iedu_1-4     (naturally coded; _Iedu_1 omitted)
(running logit on estimation sample)

Survey: Logistic regression

Number of strata     =      1             Number of obs        =          5,802
                                          Population size      =    238,177,982
                                          Replications         =             50
                                          Design df            =             49
                                          F(    4,      46)    =          51.43
                                          Prob > F             =         0.0000
```

seekcancerinfo	Odds Ratio	Jknife * Std. Err.	t	P>|t|	[95% Conf.	Interval]
_Igender_2	2.042627	.1820436	8.01	0.000	1.707686	2.443264
_Iedu_2	1.233504	.3094637	0.84	0.407	.7450482	2.042194
_Iedu_3	2.57381	.6319427	3.85	0.000	1.57142	4.215611
_Iedu_4	4.533218	1.008959	6.79	0.000	2.898405	7.090128
_cons	.2138864	.0488161	-6.76	0.000	.135205	.3383559

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

To identify levels/variables that display a significant difference in response, the rule of thumb is to examine odds ratios where the confidence interval does not contain 1 (by default, Stata will use alpha=.05 to determine statistical significance; this value can be changed by the user using code). However, significance may also be garnered from the test of whether the associated beta parameter is equal to 0 (see first regression table above). According to this model, females appear to be 2.04 times as likely as males 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 gender and education

on GeneralHealth

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

test _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4 _cons, nosvyadjust

test _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust

test _Igender_2, nosvyadjust
```

```
test _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust


i.gender    _Igender_1-2      (naturally coded; _Igender_1 omitted)
 i.edu      _Iedu_1-4         (naturally coded; _Iedu_1 omitted)
(running regress on estimation sample)

Survey: Linear regression

   Number of strata     =          1                  Number of obs       =        5,776
                                                       Population size     =  237,958,448
                                                       Replications        =           50
                                                       Design df           =           49
                                                       F(    4,      46)   =        31.73
                                                       Prob > F            =       0.0000
                                                       R-squared           =       0.0546

   ------------------------------------------------------------------------------
                    |             Jknife *
   generalhealth    |       Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
   -----------------+------------------------------------------------------------
      _Igender_2    |  .0750102     .0375668    2.00   0.051    -.0004832   .1505035
         _Iedu_2    | -.3462532     .1212491   -2.86   0.006    -.5899124  -.1025939
         _Iedu_3    | -.4540517     .1139604   -3.98   0.000    -.6830636  -.2250398
         _Iedu_4    | -.7772833     .1080917   -7.19   0.000    -.9945018  -.5600649
           _cons    |  3.081563     .1002493   30.74   0.000     2.880104   3.283021
   ------------------------------------------------------------------------------


      Unadjusted Wald test

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

              F( 5,  49)    =      3704.40
              Prob > F      =      0.0000


      Unadjusted Wald test

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

              F( 4,  49)    =      33.80
              Prob > F      =      0.0000
```

```
Unadjusted Wald test
 ( 1) _Igender_2 = 0

        F(  1, 49)    =      3.99
        Prob > F      =      0.0514

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

        F(  3, 49)    =      41.25
        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. We do not interpret the gender variable because it is non-significant.

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 gender, cell format(%8.5f) percent se wald noadjust
svy: tabulate edu gender, row format(%8.5f) percent se wald noadjust
svy: tabulate edu gender, 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     =        5,818
Number of PSUs     =       200                Population size   =   238,759956
                                              Design df         =          196
```

Education	Gender Male	Female	Total
Less than high school	4.05140	2.75039	6.80179
	(0.57443)	(0.27369)	(0.62802)
12 years	10.92694	10.64372	21.57066
	(0.80818)	(0.56895)	(0.90672)
Some college	18.39854	20.61170	39.01024
	(1.11589)	(0.95916)	(1.24091)
College	15.52856	17.08876	32.61732
	(0.70248)	(0.70973)	(1.02504)
Total	48.90544	51.09456	1.0e+02
	(1.16213)	(1.16213)	

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

Wald (Pearson):

Unadjusted	chi2(3)	=	7.6720	
Unadjusted	F(3, 196)	=	2.5573	P = 0.0564
Adjusted	F(3, 194)	=	2.5312	P = 0.0584

(running tabulate on estimation sample)

Number of strata	=	4	Number of obs	=	5,818
Number of PSUs	=	200	Population size	=	238,759,956
			Design df	=	196

Education	Gender: Male	Gender: Female	Total
Less than high school	59.56378	40.43622	1.0e+02
	(4.23691)	(4.23691)	
12 years	50.65650	49.34350	1.0e+02
	(2.45648)	(2.45648)	
Some college	47.16336	52.83664	1.0e+02
	(2.15709)	(2.15709)	
College	47.60833	52.39167	1.0e+02
	(1.49026)	(1.49026)	
Total	48.90544	51.09456	1.0e+02
	(1.16213)	(1.16213)	

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

Wald (Pearson):

Unadjusted	chi2(3)	=	7.6720	
Unadjusted	F(3, 196)	=	2.5573	P = 0.0564
Adjusted	F(3, 194)	=	2.5312	P = 0.0584

```
      (running tabulate on estimation sample)

Number of strata   =         4                Number of obs     =        5,818
Number of PSUs     =       200                Population size   =  238,759,956
                                              Design df         =          196
```

Education	Gender Male	Female	Total
Less than high school	8.28416	5.38293	6.80179
	(1.16134)	(0.51634)	(0.62802)
12 years	22.34300	20.83141	21.57066
	(1.54658)	(0.97796)	(0.90672)
Some college	37.62064	40.34030	39.01024
	(1.87092)	(1.57996)	(1.24091)
College	31.75221	33.44536	32.61732
	(1.52001)	(1.27782)	(1.02504)
Total	1.0e+02	1.0e+02	1.0e+02

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

       Wald (Pearson):
         Unadjusted    chi2(3)       =       7.6720
         Unadjusted    F(3, 196)     =       2.5573   P = 0.0564
         Adjusted      F(3, 194)     =       2.5312   P = 0.0584
```

The results of these tests based on Taylor Series linearization contradict 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.) This is a good example of how the variance estimation method used can affect the outcome of a statistical test. Both education and gender are variables used in the raking process as part of the HINTS weighting procedure. As a result, the standard errors based on replication are much smaller than those based on Taylor Series linearization, which in turn results in significant differences 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 gender [omit] 1
char edu [omit] 1
```

```
* Multivariable logistic regression of gender and

education on seekcancerinfo

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

test _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4 _cons, nosvyadjust

test _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust

test _Igender_2, nosvyadjust

test _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust

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

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

```
i.gender  _Igender_1-2   (naturally coded; _Igender_1 omitted)
 i.edu     _Iedu_1-4      (naturally coded; _Iedu_1 omitted)
(running logit on estimation sample)

Survey: Logistic regression

Number of strata    =          4                  Number of obs      =          5,802
Number of PSUs      =        200                  Population size    =    238,177,982
                                                  Design df          =            196
                                                  F(   4,     193)   =          51.94
                                                  Prob > F           =         0.0000
```

seekcancerinfo	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
_Igender_2	.7142369	.0906322	7.88	0.000	.5354975	.8929763
_Iedu_2	.2098593	.2376315	0.88	0.378	-.2587836	.6785021
_Iedu_3	.9453871	.2278119	4.15	0.000	.4961098	1.394664
_Iedu_4	1.511432	.2131146	7.09	0.000	1.09114	1.931724
_cons	-1.54231	.2207776	-6.99	0.000	-1.977715	-1.106906

```
Unadjusted Wald test

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

```
       F( 5,  196)   =      42.20
       Prob > F      =      0.0000


Unadjusted Wald test

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

       F( 4,  196)   =      52.74
       Prob > F      =      0.0000


Unadjusted Wald test

( 1)  [seekcancerinfo]_Igender_2 = 0

       F( 1,  196)   =      62.10
       Prob > F      =      0.0000

Unadjusted Wald test

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

       F( 3,  196)   =      62.70
       Prob > F      =      0.0000

i.gender _Igender_1-2       (naturally coded; _Igender_1 omitted)
 i.edu   _Iedu_1-4          (naturally coded; _Iedu_1 omitted)
(running logit on estimation sample)

Survey: Logistic regression

Number of strata      =        4                  Number of  obs       =          5,802
Number of PSUs        =      200                  Population  size     =    238,177,982
                                                  Design df            =            196
                                                  F(    4,     193)    =          51.94
                                                  Prob > F             =         0.0000
```

seekcancerinfo	Odds Ratio	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
_Igender_2	2.042627	.1851277	7.88	0.000	1.708298	2.442388
_Iedu_2	1.233504	.2931195	0.88	0.378	.7719901	1.970923
_Iedu_3	2.57381	.5863445	4.15	0.000	1.64232	4.033621
_Iedu_4	4.533218	.9660947	7.09	0.000	2.977667	6.901398
_cons	.2138864	.0472213	-6.99	0.000	.1383851	.3305803

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, women appear to be 2.04 times as likely as men 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 gender and education on generalhealth

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

test _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4 _cons, nosvyadjust
test _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust
test _Igender_2, nosvyadjust
test _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust

i.gender  _Igender_1-2   (naturally coded; _Igender_1 omitted)
i.edu     _Iedu_1-4      (naturally coded; _Iedu_1 omitted)
(running regress on estimation sample)

Survey: Linear regression

Number of strata    =          4                Number of obs      =         5,776
Number of PSUs      =        200                Population size    =   237,958,448
                                                Design df          =           196
                                                F(   4,     193)   =         33.68
                                                Prob > F           =        0.0000
                                                R-squared          =        0.0546
```

generalhealth	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
_Igender_2	.0750102	.0405788	1.85	0.066	-.0050171	.155037
_Iedu_2	-.3462532	.1126698	-3.07	0.002	-.5684539	-.1240524
_Iedu_3	-.4540517	.1073592	-4.23	0.000	-.6657793	-.2423241
_Iedu_4	-.7772833	.1088467	-7.14	0.000	-.9919443	-.5626223
_cons	3.081563	.1019607	30.22	0.000	2.880482	3.282643

```
Unadjusted Wald test

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

             F( 5,  196)   =      3651.10
             Prob > F      =      0.0000
```

```
Unadjusted Wald test

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

             F( 4,  196)   =      34.20
             Prob > F      =      0.0000

Unadjusted Wald test

 ( 1) _Igender_2 = 0

             F( 1,  196)   =      3.42
             Prob > F      =      0.0660


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

             F( 3,  196)   =      38.89
             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. We don't interpret the gender variable because it is non-significant.

Analyzing Data Using R

This section gives some R (v 4.2.2) coding examples for common types of statistical analyses using HINTS 6 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 6 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.

```
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("hints6_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(gender = case_match(factor(BirthGender),
                                    '1' ~ 'Male',
                                    '2' ~ 'Female')) |>

  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$gender = relevel(factor(df$gender, ordered = F),
                    ref = 'Male')
```

```
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.

```
# Declare Survey Design

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(gender) == F) |>
  dplyr::group_by(edu, gender) |>
  dplyr::summarize(n = n(),
                   total = survey_total(),
                   pct = survey_prop())
```

```
## # A tibble: 8 × 7
## # Groups:   edu [4]
##   edu                                gender     n     total total…¹   pct  pct_se
##   <fct>                              <fct>  <int>     <dbl>   <dbl> <dbl>   <dbl>
## 1 Less than high school              Male     155  9673127.  1.47e6 0.596  0.0438
## 2 Less than high school              Female   228  6566822.  6.23e5 0.404  0.0438
## 3 12 years or completed high school Male     375 26089157.  1.81e6 0.507  0.0186
## 4 12 years or completed high school Female   686 25412934.  1.12e6 0.493  0.0186
## 5 College graduate or higher         Male    1127 37075982.  3.01e5 0.476  0.00321
## 6 College graduate or higher         Female  1582 40801112.  3.84e5 0.524  0.00321
## 7 Some college                       Male     642 43928337.  1.24e6 0.472  0.00761
## 8 Some college                       Female  1023 49212485.  8.22e5 0.528  0.00761
## # … with abbreviated variable name ¹total_se
```

```r
# Chi-square test

svy_obj_rep |>
  svychisq(formula = ~ gender + edu,
           statistic = "F")

##
##  Pearson's X^2: Rao & Scott adjustment
##
## data:  NextMethod()
## F = 4.6411, ndf = 1.7392, ddf = 85.2215, p-value = 0.01586
```

The row percentages above show that a higher weighted proportion of college graduates in the sample are female (52.4%) than male (47.6%). Respondents with less than a high school diploma include fewer females (40.4%) than males (59.6%). 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 “Male” is the reference group for gender effect, while “Less than high school” is the reference group for education level effect.

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

# For displaying general summary statistics
summary(logistic_model)

##
## Call:
## svyglm(svy_obj_rep, formula = SeekCancerInfo ~ edu + gender,
##     family = quasibinomial())
##
## Survey design:
## Called via srvyr
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         -1.54231    0.22823  -6.758 2.33e-08 ***
## edu12 years or completed high school  0.20986    0.25088   0.836  0.40730
## eduCollege graduate or higher         1.51143    0.22256   6.791 2.08e-08 ***
## eduSome college                       0.94539    0.24552   3.851  0.00037 ***
## genderFemale                          0.71424    0.08912   8.014 3.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 5778.992)
```

```
##
## 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, females appear to be 2.04 times as likely as males 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: 5 × 7
##   term                          estim…¹ std.e…² stati…³  p.value conf.…⁴ conf.…⁵
##   <chr>                           <dbl>   <dbl>   <dbl>    <dbl>   <dbl>   <dbl>
## 1 (Intercept)                     0.214  0.228   -6.76  2.33e- 8   0.135   0.339
## 2 edu12 years or completed hig…   1.23   0.251    0.836 4.07e- 1   0.744   2.04
## 3 eduCollege graduate or higher   4.53   0.223    6.79  2.08e- 8   2.90    7.10
## 4 eduSome college                 2.57   0.246    3.85  3.70e- 4   1.57    4.22
## 5 genderFemale                    2.04   0.0891   8.01  3.29e-10   1.71    2.44
## # … with abbreviated variable names ¹estimate, ²std.error, ³statistic,
## #   ⁴conf.low, ⁵conf.high
```

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.

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

summary(linear_model)

##
## Call:
## svyglm(svy_obj_rep, formula = GeneralHealth ~ edu + gender, family = gaussian())
##
## Survey design:
## Called via srvyr
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         2.96358    0.11274  26.287  < 2e-16 ***
## edu12 years or completed high school -0.27363    0.12832  -2.132   0.0385 *
## eduCollege graduate or higher        -0.67084    0.12127  -5.532 1.54e-06 ***
```

```
## eduSome college                      -0.34157    0.12737  -2.682   0.0102 *
## genderFemale                          0.05092    0.03589   1.418   0.1629
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 7112.641)
##
## 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. However, there's no significant difference in reported general health between males and females (p = 0.16).

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)
```

Crosstab and Chi-Square Test

We are now ready to create a cross-tabulation table to examine frequency of education by gender.

```
# Crosstab

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

## # A tibble: 8 × 7
## # Groups:   edu [4]
##   edu                                gender     n     total total_se   pct pct_se
##   <fct>                              <fct>  <int>     <dbl>    <dbl> <dbl>  <dbl>
## 1 Less than high school              Male     155  9673127. 1416389. 0.596 0.0424
## 2 Less than high school              Female   228  6566822.  651865. 0.404 0.0424
## 3 12 years or completed high school Male     375 26089157. 2031080. 0.507 0.0246
## 4 12 years or completed high school Female   686 25412934. 1382808. 0.493 0.0246
## 5 College graduate or higher         Male    1127 37075982. 1630403. 0.476 0.0149
## 6 College graduate or higher         Female  1582 40801112. 1558491. 0.524 0.0149
## 7 Some college                       Male     642 43928337. 3284813. 0.472 0.0216
## 8 Some college                       Female  1023 49212485. 2385658. 0.528 0.0216
```

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

##
##  Pearson's X^2: Rao & Scott adjustment
##
## data:  NextMethod()
## F = 2.6956, ndf = 2.8714, ddf = 562.7859, p-value = 0.04772
```

The row percentages above show that a higher weighted proportion of college graduates in the sample are females (52.4%) than males (47.6%). Respondents with less than a high school diploma include fewer females (40.4%) than males (59.6%). 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 “Male” is the reference group for gender effect, while “Less than high school” is the reference group for education level effect.

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

# For displaying general summary statistics
summary(logistic_model)

##
## Call:
## svyglm(formula = SeekCancerInfo ~ edu + gender, design = svy_obj_linear,
##     family = quasibinomial())
##
## Survey design:
## Called via srvyr
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         -1.54231    0.22078  -6.986 4.52e-11 ***
## edu12 years or completed high school  0.20986    0.23763   0.883    0.378
## eduCollege graduate or higher         1.51143    0.21311   7.092 2.47e-11 ***
## eduSome college                       0.94539    0.22781   4.150 5.00e-05 ***
## genderFemale                          0.71424    0.09063   7.881 2.37e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 0.996206)
##
## 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: 5 × 7
##   term                          estim…¹ std.e…² stati…³  p.value conf.…⁴ conf.…⁵
##   <chr>                           <dbl>   <dbl>   <dbl>    <dbl>   <dbl>   <dbl>
## 1 (Intercept)                     0.214  0.221   -6.99  4.52e-11   0.138   0.331
## 2 edu12 years or completed hig…   1.23   0.238    0.883 3.78e- 1   0.772   1.97
## 3 eduCollege graduate or higher   4.53   0.213    7.09  2.47e-11   2.98    6.90
## 4 eduSome college                 2.57   0.228    4.15  5.00e- 5   1.64    4.03
## 5 genderFemale                    2.04   0.0906   7.88  2.37e-13   1.71    2.44
## # … with abbreviated variable names ¹estimate, ²std.error, ³statistic,
## #   ⁴conf.low, ⁵conf.high
```

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.

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

summary(linear_model)

##
## Call:
## svyglm(formula = GeneralHealth ~ edu + gender, design = svy_obj_linear,
##     family = gaussian())
##
## Survey design:
## Called via srvyr
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          2.96358    0.11884  24.937  < 2e-16 ***
## edu12 years or completed high school -0.27363    0.13079  -2.092  0.03775 *
## eduCollege graduate or higher        -0.67084    0.12749  -5.262 3.79e-07 ***
## eduSome college                      -0.34157    0.12686  -2.692  0.00772 **
## genderFemale                          0.05092    0.04364   1.167  0.24476
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for gaussian family taken to be 1.222734)
##
## 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 Female because the corresponding p-value is greater than .05.

Merging HINTS Survey Iterations

This section provides SAS, SPSS, Stata, and R code to combine HINTS 6 and HINTS 5, Cycle 4 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 6 and HINTS 5, Cycle 4 using SAS

This section provides SAS (Version 9.4 and higher) code for merging the HINTS 6 and HINTS 5, Cycle 4 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 = 10,117) 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 5 Cycle 4"
      2="HINTS 6"
      ;
run;
/*************************************************************************/

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

/*PUT NAME OF LIBRARY WHERE HINTS 5 CYCLE 4 FORMATS ARE STORED*/
options fmtsearch=(LibH5C4);

data tempHINTS5CYCLE4;
      /*PUT NAME OF LIBRARY AND NAME OF EXISTING HINTS 5 CYCLE 4 DATA FILE*/
      set LibH5C4.hints5_cycle4_public;

      survey=1;
```

```
      format survey survey.;
run;

/* 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 LibH6.hints6_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 mergeHINTS5C4_HINTS6;
      set tempHINTS5CYCLE4 tempHINTS6;
      /*Create Replicate Weights for trend tests*/
      **Replicate Weights;
      array hints54wgts [50]  person_finwt1-person_finwt50;
      array hints6wgts [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 5 CYCLE 4;
            Merged_NWGT0=person_finwt0;
            Merged_NWgt[i]=hints54wgts[i];
            Merged_NWgt[50+i]=person_finwt0;
      end;

      else if survey eq 2 then do i=1 to 50; *HINTS 6;
            Merged_NWGT0= person_finwt0;
            Merged_NWGT0=person_finwt0;
            Merged_NWgt[i]=person_finwt0;
            Merged_NWgt[50+i]=hints6wgts[i];
      end;
run;


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

proc surveyfreq data = mergeHINTS5c4_HINTS6 varmethod = jackknife;
      weight Merged_NWGT0;
      repweights Merged_NWGT1-Merged_NWGT100 / df = 98 jkcoefs = 0.98; tables
      seekcancerinfo chanceaskquestions;
run;
```

SAS Code to Merge HINTS 6 and HINTS 5, Cycle 4 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 5 CYCLE 4"
      2="HINTS 6"
      ;
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 5 CYCLE 4 FORMATS ARE STORED*/
options fmtsearch=(LibH5C4);

data tempHINTS5CYCLE4;
      /*PUT NAME OF LIBRARY AND NAME OF EXISTING HINTS 5 CYCLE 4 DATA FILE*/
      set LibH5C4.hints5_cycle4_public;
      RENAME PERSON_FINWT0=MERGED_FINWT0;

      survey=1;
      format survey survey.;
run;

/* 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 LibH6.hints6_public; RENAME PERSON_FINWT0=MERGED_FINWT0;

      survey=2;
      format survey survey.;
run;

data mergeHINTS5C4_HINTS6;
      set tempHINTS5CYCLE4 tempHINTS6;
run;

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

proc surveyfreq data = MergeHints5C4_Hints6 varmethod = TAYLOR;
      strata VAR_STRATUM;
      cluster VAR_CLUSTER;
      weight MERGED_FINWT0;
```

```
      tables seekcancerinfo chanceaskquestions / row col;
run;
```

Merging HINTS 6 and HINTS 5, Cycle 4 using SPSS

This section provides SPSS (Version 22) syntax for merging the HINTS 5, Cycle 4 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 6.

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 = 10,117).

First, you will need to have **HINTS 6** data open. The below syntax will first save a copy of HINTS 6 and rename it as a new file called 'MERGED_H6_H5C4.sav. 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 6 a "2" so that analysts can easily identify cases from this iteration.

```
**below, you should insert the filepath for your HINTS 6 data**.
SAVE OUTFILE='INSERT YOUR FILE PATH HERE\MERGED_H6_H5C4.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 5 CYCLE 4 data and rename our datafile, again to help with keeping files aligned for the merging process below. The following code will open your HINTS 5 Cycle 4 data and rename the dataset as H5C4. The syntax will then create the 'Survey' variable in the HINTS 5, Cycle 4 dataset and give each participant from HINTS 5, Cycle 4 a value of "1". Again, this is so that once the datasets are merged, analysts can easily identify which cases were from the HINTS 5, Cycle 4 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 6 dataset
Note, the analyst will need to insert the file path for where HINTS 5 Cycle 4 is saved.

```
**below, you should insert the file path for your HINTS 5 Cycle 4 data**.
GET FILE='INSERT YOUR FILE PATH HERE\hints5_cycle4_public.sav'.
DATASET NAME H5C4 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_H6_H5C4” (which currently contains only HINTS 6 data) and “hints5_cycle4_public”. Within your “MERGED_H6_H5C4” 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 hints5_cycle4_public 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='H5C4'.
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 6 and HINTS 5, Cycle 4 using Stata

This section provides Stata (Version 10.0 and higher) code for merging the HINTS 6 and HINTS 5, Cycle 4 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 6 with HINTS 5, Cycle 4, 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 = 10,117). Note that variable names are case sensitive in Stata.

```
*Put path and name to your HINTS 5 Cycle 4 data
use "INSERT YOUR PATH HERE\hints5_cycle4_public.dta", clear

*Create final and replicate weights (merged_nwt*) for multi-cycle datasets
gen merged_nwgt0=person_finwt0
forvalues n1=1/50 {
local x1=`n1'+50
gen merged_nwgt`n1'=person_finwt`n1'
gen merged_nwgt`x1'=person_finwt0
}
save h5c4.dta, replace

*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=person_finwt0
forvalues n2=1/50 {
local x2=`n2'+50
gen merged_nwgt`n2'=person_finwt0
gen merged_nwgt`x2'=person_finwt`n2'
}
save h6.dta,replace

set trace off

*Combine the 2 cycles of data & generate survey variable flagging HINTS
iteration
use h5c4.dta, clear
append using h6.dta, generate(survey)
label define survey 0 "HINTS 5 CYCLE 4" 1 "HINTS 6"
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 6 and HINTS 5, Cycle 4 for the Taylor Series Linearization Method

In order to combine HINTS 5, Cycle 4 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 = 10,117).

```
*Put path and name to your HINTS 5 Cycle 4 data
use "INSERT YOUR PATH HERE\hints5_cycle4_public.dta", clear

*Create final weight (merged_nwt0) for multi-cycle datasets
gen merged_nwgt0=person_finwt0
save h5c4.dta, replace

*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=person_finwt0
save h6.dta, replace

*Combine the 2 cycles of data & generate survey variable flagging HINTS
iteration
use h5c4.dta, clear
append using h6.dta, generate(survey)
label define survey 0 "HINTS 5 CYCLE 4" 1 "HINTS 6"
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 6 and HINTS 5, Cycle 4 using R

This section provides R syntax for merging the HINTS 5, Cycle 4 and HINTS 6 iterations. The code below loads HINTS 6 and HINTS 5 Cycle 4 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 = 10,117).

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

# HINTS 5 Cycle 4 file
df_H5C4 = haven::read_sas("hints5_cycle4_public.sas7bdat")

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

# Create Hints 5 Cycle 4 group weights
df_H5C4 = df_H5C4 |>
  dplyr::mutate(hints_edition = 'Hints 5 Cycle 4') |>
  dplyr::mutate(nwgt0 = PERSON_FINWT0)

for(i in 1:100){
  if(i <= 50){
    df_H5C4[nwgt_var_names[i]] = df_H5C4[var_names[i]]
  }

  if(i > 50){
    df_H5C4[nwgt_var_names[i]] = df_H5C4$PERSON_FINWT0
  }
}

# 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$PERSON_FINWT0
  }

  if(i > 50){
    df_H6[nwgt_var_names[i]] = df_H6[var_names[i-50]]
  }
}
```

The below syntax will merge the HINTS 6 and HINTS 5 CYCLE 4 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_H5C4, df_H6)

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

##
## Hints 5 Cycle 4         HINTS 6
##            3865            6252
```

The example code below can be used to run simple frequencies on two common variables ("SeekCancerInfo" and "ChanceAskQuestions") in the HINTS 6 and HINTS 5 Cycle 4 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 6 and HINTS 5 Cycle 4 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

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

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.

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

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