Health
Information
National
Trends
Survey
hints

OVERVIEW OF THE HINTS 5 CYCLE 3 SURVEY AND DATA ANALYSIS RECOMMENDATIONS

January 2020

CONTENTS

Overview of HINTS

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

The HINTS 5 administration includes four data collection cycles over four years, starting in 2017. The first of these cycles (HINTS 5, Cycle 1) was conducted from January through May 2017. The focus of this report is HINTS 5, Cycle 3. HINTS 5 draws upon the lessons learned from prior iterations of HINTS and incorporates an experimental design. A multi-mode survey was implemented for HINTS 5, Cycle 3, using a Web Pilot, in addition to the traditional mail survey. For more extensive background about the HINTS program and previous data collection efforts, see Finney Rutten, et al. (2012).

Methodology

Data collection for Cycle 3 of HINTS 5 began in January 2019 and concluded in April 2019. HINTS 5, Cycle 3 included a self-administered mailed questionnaire, as well as two experimental conditions in the Web Pilot. Individuals that were part of the Web Pilot were randomly assigned to either the Web Option (offering respondents a choice between responding via paper or web), or the Web Bonus (offering respondents a choice between responding via paper or web with an additional $10 incentive for those responding via web). All conditions used the same sampling frame provided by Marketing Systems Group (MSG) of addresses in the United States. All addresses were grouped into two strata: 1) addresses in areas with high concentrations of minority populations; 2) addresses in areas with low concentrations of minority populations. All non-vacant residential addresses in the United States present in the MSG database, including post office (P.O.) boxes, throwbacks (i.e., street addresses for which mail is redirected by the U.S. Postal Service to a specified P.O. box), and seasonal addresses were subject to sampling. However, P.O. box addresses were only used if they were classified as the only-way-to-get mail. The address sample was divided into three representative subsamples in order to accommodate the Web Pilot: 1) A subsample for the Paper-only treatment group also acted as the control group, where the traditional data collection procedures were used (referred to as Cycle 3); 2) A subsample for the Web Option treatment group of the Web Pilot; and 3) A subsample for the Web Bonus treatment group of the Web Pilot. The protocol for mailing the questionnaires involved an initial mailing of the questionnaire, followed by a reminder postcard, and up to two additional mailings of the questionnaire as needed for non-responding households. The protocol for the Web Pilot groups was similar; however, the language in the cover letters varied based on whether respondents were being invited to complete the survey by web and, if so, whether they were being offered a bonus incentive to do so. Those stratified into the Web Pilot groups received a link to the web survey, along with a unique PIN or access code for each respondent. Reminder postcards for both Web Pilot groups were folded and sealed so that the respondent's PIN could be included in the reminder. All households in each sample received one English survey per mailing unless someone from the household contacted Westat to request a Spanish survey, in which case the household received one Spanish survey per mailing for all subsequent mailings. The second stage of sampling consisted of selecting one adult within each sampled household using the next-birthday method. In this method, the adult who would have the next birthday in the sampled household was asked to complete the questionnaire. A $2 monetary incentive was included with the survey to encourage participation. Refer to the HINTS 5, Cycle 3 Methodology Report for more extensive information about the sampling procedures and to the Web Pilot report for more information about the methods used and the results.

Sample Size and Response Rates

The final HINTS 5, Cycle 3 sample consists of 5,438 respondents. Of these, 3,372 respondents were in the Paper only group, 986 were in the Web option group, and 1,080 were in the Web bonus group. Note that 191 of these respondents were considered partial completers who did not answer the entire survey. A questionnaire was considered to be complete if at least 80% of Sections A and B were answered. A questionnaire was considered to be partially complete if 50%–79% of the questions were answered in Sections A and B. Household response rates were calculated using the American Association for Public Opinion Research response rate 2 (RR2) formula. The overall household response rate using the next-birthday method was 30.3%. More specifically, the Cycle 3 (Paper Only) overall response rate was 30.2%, the Web Option response rate was 29.6%, and the Web Bonus group response rate was 31.5%. These response rates were not found to be significantly different.

Analyzing HINTS Data

If you are solely interested in calculating point estimates (means, proportions, etc.), either weighted or unweighted, you can use programs including SAS, SPSS, STATA, and Systat. If you plan on doing inferential statistical testing using the data (i.e., anything that involves calculating a p-value or confidence interval), it is important that you utilize a statistical program that can compute the correct variance estimates when analyzing survey data that employ a complex sampling method, such as with 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 SUDAAN, STATA, SAS, and Wesvar can incorporate the replicate weights found in the HINTS database.

With the release of HINTS 5, Cycle 3, 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 particular attention to the standard error and coefficient of variation (relative standard error) for estimates of means, proportions, and totals, and the analyst should report these when writing up results. It is important that the analyst realizes that small sample sizes for particular analyses will tend to result in unstable estimates.

Important Analytic Variables in the Database

Refer to the HINTS 5, Cycle 3 Methodology Report for more information regarding the weighting and stratification variables listed below.

Note that estimates from the 2017 American Community Survey (ACS) of the U.S. Census Bureau were used to calibrate the HINTS 5, Cycle 3 control totals with the following variables: age, gender, education, marital status, race, ethnicity, and census region. In addition, variables from the 2017 National Health Interview Survey (NHIS) were used to calibrate HINTS 5, Cycle 3 data control totals regarding: percent with health insurance and percent ever had cancer.

Final Sample and Replicate Weights for Jackknife Replication

Included with the data are several groups of 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). Recommendations and information for determining which weights to use are found in the “Recommendations for Statistical Analyses using HINTS 5 Cycle 3” section.

TG_all_FINWT0: Final sample weight used to calculate population estimates for the **combined** sample.
TG_all_FINWT1 through TG_all_FINTW50: Fifty replicate weights that can be used to calculate accurate standard error of estimates using the jackknife replication method for the **combined** sample.

TG1_FINWT0: Final sample weight used to calculate population estimates for the **paper-only** sample.
TG1_FINWT1 through TG1_FINTW50: Fifty replicate weights that can be used to calculate accurate standard error of estimates using the jackknife replication method for the **paper-only** sample.

TG2_FINWT0: Final sample weight used to calculate population estimates for the **web-option** sample.
TG2_FINWT1 through TG2_FINTW50: Fifty replicate weights that can be used to calculate accurate standard error of estimates using the jackknife replication method for the **web-option** sample.

TG3_FINWT0: Final sample weight used to calculate population estimates for the **web-bonus** sample.
TG3_FINWT1 through TG3_FINTW50: Fifty replicate weights that can be used to calculate accurate standard error of estimates using the jackknife replication method for the **web-bonus** sample.

Nwgt0: Final sample weight used to calculate population estimates for the **combined sample, controlling for group differences.** For more information on how this variable was calculated, please see Appendix A.
Nwgt1 through Nwgt150: 150 replicate weights that can be used to calculate an accurate standard error of estimates using the jackknife replication method for the **combined sample, controlling for group differences**. For more information on how these variables were calculated, please see Appendix A.

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 5 Cycle 3, this variable also incorporates the three subsamples used to accommodate the Web Pilot. 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 six values: high minority (HM) and low minority (LM) among the paper-only sample, high minority and low minority among the web-option sample (W1-HM and W1-LM, respectively), and high minority and low minority among the web-bonus sample (W2-HM and W2-LM, respectively).

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.

TG_all_FINWT0: Final sample weight used to calculate population estimates for the **combined** sample.

TG1_FINWT0: Final sample weight used to calculate population estimates for the **paper-only** sample.

TG2_FINWT0: Final sample weight used to calculate population estimates for the **web-option** sample.

TG3_FINWT0: Final sample weight used to calculate population estimates for the **web-bonus** sample.

Nwgt0: Final sample weight used to calculate population estimates for the **combined sample, controlling for group differences.** For more information on how this variable was calculated, please see Appendix A.

Other Variables

TREATMENT_H5C3: This variable codes for which group the respondent was assigned: 1) Paper only; 2) Web Option; 3) Web Bonus;

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.

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 191 partially complete questionnaires. 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).

Recommendations for Statistical Analyses Using HINTS 5 Cycle 3

Given the sampling method, each group of respondents: (1) Paper only; (2) Web Option; (3) Web Bonus; can be analyzed as a separate independent sample using the respective weights (for example, an analyst may choose to use the paper only group to be consistent with previous iterations) or the samples can be combined to analyze as a whole using the composite weights. Where possible, it is recommended that analysts use the entire sample to increase statistical power in their analyses. See below for more details for steps needed to incorporate the entire sample into your analysis.

A flow chart is provided in Appendix A to summarize and supplement the steps outlined in the subsequent subsections.

Analyzing HINTS 5 Cycle 3 Data Using the Composite Sample

Assessing for Differences Across Groups

It is strongly recommended that analysts first assess for possible group differences between their target variables. The Web Pilot Report provides some initial analyses testing for group differences and users can refer to this report to see if their variables of interest have already been assessed. If analysts want to do their own assessment for group differences with the jackknife replication variance estimation method, use the final sample weight (**nwgt0**) and 150 replicate weights (**nwgt1** through **nwgt150**) that have been provided with the data. Appendix A provides the code in SAS and STATA that created these sample weights that allow for assessing and controlling for group differences. This code was created using the Rizzo method, similar to how replicate weights are created when combining different HINTS iterations together or assessing trends over time (Rizzo, Moser, Waldron, Wang, & Davis, 2008). The SPSS data file will contain the same **nwgt0** variable that can also be used to assess for group differences using the Taylor series linearization method, along with the **var_stratum** and **var_cluster** variables (see Appendix B).

The **SAS** code below provides examples for testing for group differences using the NWGT weights.

Assessing for Group Differences with *Binary Outcomes* (with SEEKCANCERINFO as example):

```
data DATAFILENAME;
     set DATAFILENAME;
     *Set negative values to missing;
     if SeekCancerInfo < 0 then SeekCancerInfo=.;
run;
proc surveylogistic data=DATAFILENAME varmethod=jackknife;
     weight nwgt0;
     repweights nwgt1-nwgt150 /df=147 jkcoefs=.98;
     class TREATMENT_H5C3;
     model SeekCancerInfo = TREATMENT_H5C3;
run;
```

Assessing for Group Differences with *Continuous Data* (with GENERALHEALTH as example):

```
data DATAFILE;
     set DATAFILE;
     *Set negative values to missing;
     if GeneralHealth < 0 then GeneralHealth=.;
run;
proc surveyreg data=DATAFILENAME varmethod=jackknife;
     weight nwgt0;
     repweights nwgt1-nwgt150 /df=147 jkcoefs=.98;
     class TREATMENT_H5C3;
     model GeneralHealth = TREATMENT_H5C3 /solution;
run;
```

The **SPSS** code below provides examples for testing for group differences using the NWGT0 weight. Note that an analysis plan must be first created and applied; more information about setting up your analysis plan can be found in the “Analyzing Data Using SPSS” section later in the document.

```
* Analysis Preparation Wizard.
* Substitute your path and file name inside the quotes of /PLAN FILE=.
CSPLAN ANALYSIS
  /PLAN FILE="YOUR-PATH\YOUR-PLAN-NAME.csaplan"
  /PLANVARS ANALYSISWEIGHT=NWGT0
  /SRSESTIMATOR TYPE=WOR
  /PRINT PLAN
  /DESIGN STRATA=VAR_STRATUM CLUSTER=VAR_CLUSTER
  /ESTIMATOR TYPE=WR.
```

Assessing for Group Differences with *Binary Outcomes* (with SEEKCANCERINFO as example):

```
DATASET ACTIVATE DataSet1.
* Complex Samples Logistic Regression.
** First set negative values to missing.
IF SeekCancerInfo<0 SeekCancerInfo=$SYSMIS.
CSLOGISTIC SeekCancerInfo (HIGH) BY Treatment_H5C3
  /PLAN FILE="YOUR-PATH\YOUR-PLAN-NAME.csaplan"
  /MODEL Treatment_H5C3
  /INTERCEPT INCLUDE=YES SHOW=YES
  /TEST TYPE=F PADJUST=LSD
  /MISSING CLASSMISSING=EXCLUDE
  /CRITERIA MXITER=100 MXSTEP=5 PCONVERGE=[1e-006 RELATIVE] LCONVERGE=[0]
CHKSEP=20 CILEVEL=95
  /PRINT SUMMARY VARIABLEINFO SAMPLEINFO.
```

Assessing for Group Differences with *Continuous Data* (with GENERALHEALTH as example):

```
* Complex Samples General Linear Model.
** First set negative values to missing.
IF GeneralHealth<0 GeneralHealth=$SYSMIS.
CSGLM GeneralHealth BY Treatment_H5C3
  /PLAN FILE="YOUR-PATH\YOUR-PLAN-NAME.csaplan"
  /MODEL Treatment_H5C3
  /INTERCEPT INCLUDE=YES SHOW=YES
  /PRINT SUMMARY VARIABLEINFO SAMPLEINFO
  /TEST TYPE=F PADJUST=LSD
  /MISSING CLASSMISSING=EXCLUDE
  /CRITERIA CILEVEL=95.
```

The **Stata** code below provides examples for testing for group differences using the NWGT weights. 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.

```
      svyset [pw=nwgt0], jkrw(nwgt1-nwgt150, multiplier(0.98))
vce(jack) dof(147) mse
```

Assessing for Group Differences with *Binary Outcomes* (with SEEKCANCERINFO as example):

```
*Recode to 0,1 dichotomous and set negative to missing
replace seekcancerinfo = 0 if seekcancerinfo == 2
replace seekcancerinfo = . if seekcancerinfo < 0
char treatment_h5c3 [omit] 1
xi: svy: logit seekcancerinfo i.treatment_h5c3
test _Itreatment_2 _Itreatment_3 _cons, nosvyadjust
test _Itreatment_2 _Itreatment_3, nosvyadjust
xi: svy, or: logit seekcancerinfo i.treatment_h5c3
```

Assessing for Group Differences with *Continuous Data* (with GENERALHEALTH as example):

```
*Set negative to missing
replace generalhealth = . if generalhealth < 0
```

```
char treatment_h5c3 [omit] 1
xi: svy: regress generalhealth i.treatment_h5c3
test _Itreatment_2 _Itreatment_3 _cons, nosvyadjust
test _Itreatment_2 _Itreatment_3, nosvyadjust
```

Determining Statistical Weights to Use for Analyses

If an analyst finds that group differences exist, the analyst may decide to use the entire sample and control for group assignment using the TREATMENT_H5C3 variable and using the NWGT weighting variables (see example code below under Option A). Alternatively, the analyst may determine that the best course of action is to analyze only one group (such as the paper only group) and avoid any concerns about group differences (example code under Option B). In this case an analyst can use the respective group weights to analyze each group independently (e.g. Paper only sample using TG1_FINWT weight variables). One important note is that the Web Pilot Results Report found significant differences by group on several key demographic measures. It is possible that an analyst may want to consider these differences in their analyses.

If one does not find any differences by group assignment, it is recommended that analysts use the combined sample and the respective weights --the TG_all weight variables-- to increase statistical power. Example of this code in SAS is provided below.

*If Group Differences **Are** Found…*

Option A: Use the Combined Sample and Control for Group Assignment

```
proc surveylogistic data=DATAFILENAME varmethod=jackknife;
     weight nwgt0;
     repweights nwgt1-nwgt150 /df=147 jkcoefs=.98;
     *Predictor# variables in model statement are placeholders
     to substitute with your desired predictors;
     model SeekCancerInfo = TREATMENT_H5C3 predictor1 predictor2
        predictor3 predictor4…;
run;

proc surveyreg data=DATAFILENAME varmethod=jackknife;
     weight nwgt0;
     repweights nwgt1-nwgt150 /df=147 jkcoefs=.98;
     *Predictor# variables in model statement are placeholders
     to substitute with your desired predictors;
     model GeneralHealth = TREATMENT_H5C3 predictor1 predictor2
        predictor3 predictor4… /solution;
run;
```

Option B: Use One Group, only, without Accounting for Group Differences

```
proc surveylogistic data=DATAFILENAME varmethod=jackknife;
     weight tg1_finwt0;
     repweights tg1_finwt1-tg1_finwt50 /df=49 jkcoefs=.98;
     *Predictor# variables in model statement are placeholders
     to substitute with your desired predictors;
     model SeekCancerInfo = predictor1 predictor2 predictor3
       predictor4…;
run;
```

Note: example code above is for the paper-only sample ("TG1"). Weight and repweight statements may be replaced with the "TG2" weights for the web option sample or "TG3" weights for the web bonus sample.

*If Group Differences **Are NOT** Found…*

Option C: Use Combined Sample without Accounting for Group Differences

```
proc surveylogistic data=DATAFILENAME varmethod=jackknife;
     weight tg_all_finwt0;
     repweights tg_all_finwt1-tg_all_finwt50 /df=49 jkcoefs=.98;
     *Predictor# variables in model statement are placeholders
     to substitute with your desired predictors;
     model SeekCancerInfo = predictor1 predictor2 predictor3
       predictor4…;
run;
```

Note: additional sample code and output for using the combined sample without controlling for group assignment can be found in the Statistical Software Example Code section of this document.

Variance Estimation Methods: Replicate vs. Taylor Linearization

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

The HINTS 5, Cycle 3 datasets include variance codes and replicate weights so analysts can use either Taylor Series or replication methods for variance estimation. The following points may provide some guidance regarding which method will best reflect the HINTS sample design in your analysis.

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

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

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

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

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

Denominator Degrees of Freedom (DDF)

Replicate Weights: The HINTS 5, Cycle 3 database contains several sets of 50 replicate weights to compute accurate standard errors for statistical testing procedures, depending on which sample you wish to analyze (i.e., Combined; Paper only; Web Option; Web Bonus). 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. Similarly, analysts who find that there is no difference between groups on their target variables, or who are ignoring group differences and only analyzing one group, should also use 49 ddf in their statistical models. HINTS statistical analyses that involve more than one iteration of data—or a comparison for group differences in HINTS 5 Cycle 3-- 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., and analysts who are assessing for differences across the three treatment groups will use 147 (49*3) ddf.

Taylor Series: The HINTS 5, Cycle 3 database contains two variables that can be used to calculate standard errors using the Taylor series, namely VAR_STRATUM and VAR_CLUSTER (see VAR_STRATUM and VAR_CLUSTER variables in the previous section for strata definitions.). The degrees of freedom for the Taylor series, 98, is based on 50 PSUs in each of the two sampling strata (#psus - #strata = 50*2 – 2 = 98).

Statistical Software Example Code

This section provides some coding examples using SAS, SPSS, and STATA for common types of statistical analyses using HINTS 5, Cycle 3 data.

Note that these examples use the combined data and associated weights (TG_ALL_FINWT0 and TG_ALL_FINTW1 through TG_ALL_FINWT50) and do NOT account for group differences.

To instead use combined data and control for group assignment, users would need to use the NWGT0 final weight and the associated 150 replicate weights. Examples of this code are available in Appendix B. A user could also choose to only assess one group of participants within HINTS 5, Cycle 3 data, and could do so using the respective weights (e.g. TG1_FINWT for the "paper only" group and its respective final sample and 50 replicate weights).

For SAS and STATA, you'll see two sets of code: one when using replicate methods for variance estimation, and one for Taylor Series linearization. For replicate methods, these examples will incorporate both the final sample weight (to get population-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 5, Cycle 3 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 5, Cycle 3 SAS data, it is important to apply the SAS formats. To do this, follow the steps below.

1. Download all HINTS 5, Cycle 3 documents to a folder on your computer. This should be the same folder where you create the SAS library in step 3.
2. Create a permanent library using the dropdown menu that references the folder from Step 1.
3. Open the SAS program "HINTS 5 Cycle 3 Public Formats.sas."
4. Change the file location specification in the "library" statement at the top of the program to the location where you want the format library to be stored before you run this program.
5. Run the program "HINTS 5 Cycle 3 Public Formats.sas" to create a permanent SAS format library that is used to analyze the HINTS dataset.
6. Open the SAS program "HINTS 5 Cycle 3 Public 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 "HINTS 5 Cycle 3 Public Format Assignments.sas" to create the formatted SAS dataset.

Note the following:

a. Make sure to run the program "HINTS 5 Cycle 3 Public Formats.sas" BEFORE you run "HINTS 5 Cycle 3 Public Format Assignments.sas" to create the formatted HINTS dataset.

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

This section gives some SAS (Version 9.3 and higher) coding examples for common types of statistical analyses using HINTS 5, Cycle 3 data. Subsection 1 shows how to complete common analyses using replicate weights, and subsection 2 shows analyses using the Taylor series linearization approach. For either approach, we begin by doing data management of the HINTS 5 Cycle 3 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=(hints5c3);

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 hints5cycle3;
      set hints5c3.hints5cycle3_formatted;

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

     /*Recode negative values to missing for proc regress procedure*/
     if generalhealth in (-5, -9, -7) then generalhealth = .;


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

Replicate Weights Variance Estimation Method

Frequency Table and Chi-Square Test

We are now ready to begin using SAS 9.3 to examine the relationships among these variables. Using **PROC SURVEYFREQ**, we will first generate a cross-frequency table of education by gender, along with a (Wald) Chi-squared test of independence. Note the syntax of the overall sample weight, TG_all_FINWT0, and those of the jackknife replicate weights, TG_all_FINWT1—TG_all_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. Example code for using replicate weights if you find that your target variable(s) do differ by group condition can be found in Appendix B.

```
proc surveyfreq data = hints5cycle3 varmethod = jackknife;
     weight TG_all_FINWT0;
     repweights TG_all_FINWT1-TG_all_FINWT50 / df = 49 jkcoefs = 0.98;
     tables edu*gender / row col wchisq;
run;
```

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

For the purposes of computing appropriate degrees of freedom for the estimator of the HINTS5-Cycle 3 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	hints5cycle3
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	119	3.2314	0.4776	46.376	4.3306	6.6368	0.9827
	Female	209	3.7395	0.3633	53.624	4.3306	7.2937	0.7105
	Total	328	6.9736	0.6102	100			
12 years or completed high school	Male	381	11.8559	0.5994	51.0393	1.5532	24.33	1.2075
	Female	552	11.3731	0.4348	48.9607	1.5532	22.1825	0.8422
	Total	933	933	23.229	0.7549	100		
Some college	Male	673	19.7161	0.5783	49.0054	1.0365	40.4603	1.1621
	Female	901	20.5164	0.4636	50.9946	1.0365	40.0159	0.8544
	Total	1574	40.2325	0.6394	100			
College graduate or higher	Male	1030	13.9234	0.1046	47.0943	0.2503	28.5728	0.2293
	Female	1369	15.6415	0.1058	52.9057	0.2503	30.5079	0.1953
	Total	2399	29.5649	0.1496	100			
Total	Male	2203	48.7295	0.2519			100	
	Female	3031	51.2705	0.2519			100	
	Total	5234	100					

Frequency Missing =204

Wald Chi-Square Test	
Chi-Square	41.4204
F Value	13.8068
Num DF	3
Den DF	49
Pr > F	<.0001
Adj F Value	13.2433
Num DF	3
Den DF	47
Pr > Adj F	0.0001
Sample Size = 5234	

The row percentages above show that a higher weighted proportion of college graduates in the sample are women (53%) than men (47%). Respondents with less than a high school diploma include more women (54%) than men (46%). 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= hints5cycle3 varmethod=jackknife;
      weight TG_all_FINWT0;
      repweights TG_all_FINWT1-TG_all_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	hints5cycle3
Number of Replicates	50

Type 3 Analysis of Effects				
Effect	**F Value**	**Num DF**	**Den DF**	**Pf > F**
Gender	18.89	1	49	<.0001
Education	27.27	3	49	<.0001

Analysis of Maximum Likelihood Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**T value**	**Pr > |t|**	**95% confidence limits**	
Intercept	49	-0.7935	0.2041	-3.89	0.0003	-1.2036	-0.3833
Gender	49	0.4290	0.0987	4.35	<.0001	0.2306	0.6273
12 years or completed high school	49	0.1255	0.2086	0.60	0.5503	-0.2938	0.5447
Some College	49	0.9413	0.2043	4.61	<.0001	0.7661	1.6066
College graduate or higher	49	1.1864	0.2091	5.67	<.0001	0.5308	1.3519

(continued on the next page)

Odds Ratio Estimates

Effect	Point Estimate	95% Confidence Limits	
Female vs Male	1.536	1.259	1.873
12 years or completed high school vs Less than high school	1.134	0.745	1.724
Some College vs Less than high school	2.563	1.700	3.865
College graduate or higher vs Less than high school	3.275	2.151	4.986

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, women appear to be 1.54 times as likely as men to have searched for cancer information.

Linear Regression

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

```
/*Multivariable  linear  regression  of  gender  and  education  on
GeneralHealth*/
proc surveyreg data= hints5cycle3 varmethod=jackknife;
      weight TG_all_FINWT0;
      repweights TG_all_FINWT1-TG_all_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	hints5cycle3
Number of Replicates	50

(continued on the next page)

Estimated Regression of Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	2.8413999	0.11428047	24.86	<.0001
12 years or completed high school	-0.2104414	0.12700482	-1.66	0.1039
Some College	-0.2336853	0.11338544	-2.06	0.0446
College graduate or higher	-0.6492976	0.11729768	-5.54	<.0001
Female	0.0821506	0.04613621	1.78	0.0812

The table labeled Estimated Regression of Coefficients shows that respondents with some college reported better general health than those with less than a high school education (p=0.0446) when controlling for all other variables in the model. Keep in mind that the outcome, general health, is coded such that lower scores correspond to better health. This table also shows that this association applies to those with a college degree or higher (coefficient = -0.65, p<.0001) when comparing to respondents with less than a high school education. However, there's no significant difference in health score between males and females (p=0.0812) and those with only a high school diploma and those without a high school diploma (p=0.1039).

Tests of Model Effects

Contrast	Num DF	F Value	Pr > F
Model	4	31.82	<.0001
Intercept	1	6688.02	<.0001
Education	3	41.07	<.0001
Gender	1	3.17	0.0812

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.

Taylor Series Linearization Variance Estimation Method

Frequency Table and Chi-Square Test

We are now ready to begin using SAS 9.3 to examine the relationships among these variables.
Using **PROC SURVEYFREQ**, we will first generate a cross-frequency table of education by gender, along with a (Wald) Chi-squared test of independence. Note the syntax of the strata VAR_STRATUM, cluster VAR_CLUSTER, and overall sample weight TG_all_FINWT0 (no group differences). This syntax is consistent for all procedures. Other analyses that use Taylor Series approximation will follow a similar syntax.

```
proc surveyfreq data = hints5cycle3
      varmethod = TAYLOR;
      strata VAR_STRATUM;
      cluster VAR_CLUSTER;
      weight TG_all_finwt0;
```

```
    tables edu*gender / row col wchisq;
run;
```

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

Data Summary	
Number of Strata	6
Number of Clusters	300
Number of Observations	5438
Sum of Weights	252070495

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	119	3.2341	0.4579	46.3760	4.2618	6.6368	0.9574
	Female	209	3.7395	0.3697	53.6240	4.2618	7.2937	0.7033
	Total	328	6.9736	0.5931	100			
12 years or completed high school	Male	381	11.8559	0.8646	51.0393	2.3046	24.3300	1.5937
	Female	552	11.3731	0.5919	48.9607	2.3046	22.1825	1.1217
	Total	933	23.2290	1.0165	100			
Some college	Male	673	19.7161	1.0803	49.0054	2.1151	40.4603	1.9076
	Female	901	20.5164	1.0620	50.9946	2.1151	40.0159	1.5629
	Total	1574	40.2325	1.3021	100			
College graduate or higher	Male	1030	13.9234	0.6641	47.0943	1.5188	28.5728	1.2878
	Female	1369	15.6415	0.6500	52.9057	1.5188	30.5079	1.2706
	Total	2399	29.5649	0.9622	100			
Total	Male	2203	48.7295	1.1779			100	
	Female	3031	51.2705	1.1779			100	
	Total	5234	100					

Frequency Missing =204

(continued on the next page)

Wald Chi-Square Test	
Chi-Square	2.4138
F Value	0.8046
Num DF	3
Den DF	294
Pr > F	0.4921
Adj F Value	0.7991
Num DF	3
Den DF	292
Pr > Adj F	0.4952
Sample Size = 5234	

The row percentages above show that a higher weighted proportion of college graduates in the sample are women (53%) than men (47%). Respondents with less than a high school diploma include more women (54%) than women (46%). The Chi-squared test of independence statistic and associated p value suggest that one should accept the null hypothesis that the two variables are not associated, which indicates that there is not a significant difference between the distributions of educational attainment for these two groups.

Logistic Regression

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

```
/*Multivariable logistic regression of gender and education on
SeekCancerInfo*/
proc surveylogistic data= hints5cycle3 varmethod=TAYLOR;
      strata VAR_STRATUM;
      cluster VAR_CLUSTER;
      weight TG_all_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)

(continued on the next page)

Type 3 Analysis of Effects				
Effect	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Gender	21.54	1	294	<.0001
Education	29.12	3	292	<.0001

Analysis of Maximum Likelihood Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**	**Pf > |t|**	**95% Confidence Limits**	
Intercept	294	-0.7935	0.2203	-3.60	0.0004	-1.2270	-0.3599
Gender	294	0.4290	0.0924	4.64	<.0001	0.2471	0.6108
12 years or completed high school	294	0.1255	0.2392	0.52	0.6003	-0.3453	0.5962
Some College	294	0.9413	0.2254	4.18	<.0001	0.4977	1.3849
College graduate or higher	294	1.1864	0.2255	5.26	<.0001	0.7427	1.6301

Odds Ratio Estimates

Effect	Point Estimate	95% Confidence Limits	
Female vs Male	1.536	1.280	1.842
12 years or completed high school vs Less than High School	1.134	0.708	1.815
Some College vs Less than High School	2.563	1.645	3.994
College graduate or higher vs Less than High School	3.275	2.102	5.104

To identify levels/variables that display a significant difference in response, the rule of thumb is to examine odds ratios where the confidence interval does not contain 1 (by default, SAS will use alpha=.05 to determine statistical significance; this value can be changed by the user using code). However, significance may also be garnered from the test of whether the associated beta parameter is equal to 0 (see “Analysis of Maximum Likelihood Estimates” table above). According to this model, women appear to be statistically more likely than men to have searched for cancer information.

Linear Regression

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

```
/*Multivariable linear regression of gender and education on
GeneralHealth*/
proc surveyreg data= hints5cycle3 varmethod=TAYLOR;
      strata VAR_STRATUM;
      cluster VAR_CLUSTER;
      weight TG_all_FINWT0;
      class edu (ref="Less than high school") gender (ref="Male");
      model generalhealth = edu gender/solution;
run;
```

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

Estimated Regression of Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	2.8413999	0.11840930	24.00	<.0001
Female	0.0821506	0.04600196	1.79	0.0752
12 years or completed high school	-0.2104414	0.12823868	-1.64	0.1019
Some College	-0.2336853	0.11678898	-2.00	0.0463
College graduate or higher	-0.6492976	0.11626607	-5.58	<.0001

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

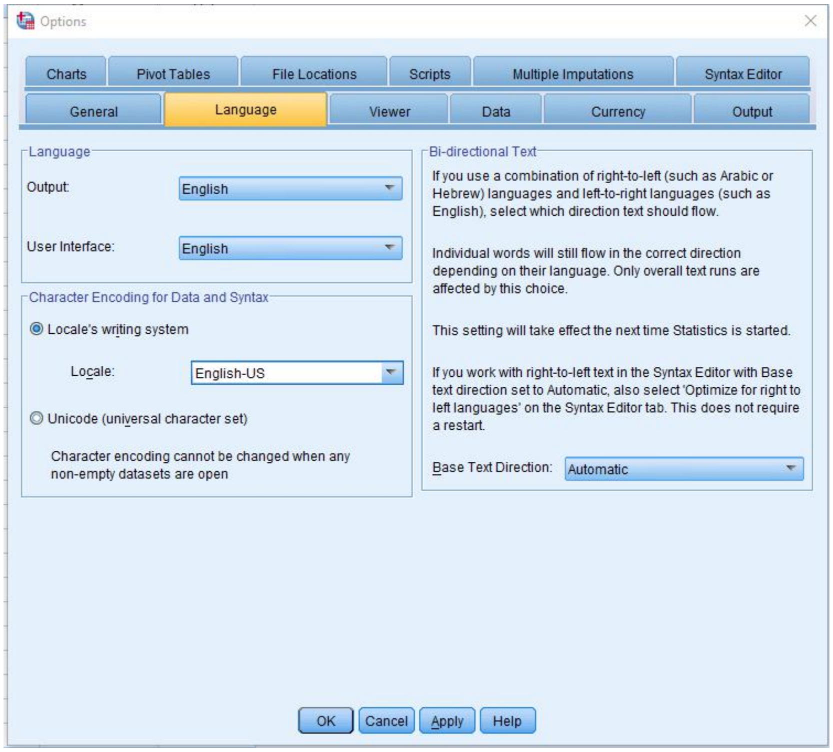
Tests of Model Effects

Contrast	Num DF	F Value	Pr > F
Overall model	4	35.58	<.0001
Intercept	1	6463.66	<.0001
Gender	1	3.19	0.0752
Education	3	41.89	<.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 5, Cycle 3 SPSS data, it is important to ensure that your SPSS environment is set up to be compatible with the dataset. Specifically, the language encoding (i.e., the way that character data are stored and accessed) must match between your environment and the dataset. We recommend locale encoding in U.S. English over Unicode encoding. To ensure compatibility, you must update the language encoding manually through the graphic user interface (GUI). In a new SPSS session, from the empty dataset window, select “Edit” > “Options…” from the menu bar. In the pop-up box, select the “Language” tab. In this tab, look for the “Character Encoding for Data and Syntax” section. Select the “Locale’s writing system” option and English-US or en-US from the “Locale:” dropdown list. “English-US” and “en-US” from the drop down are the common aliases used by SPSS to describe U.S. English encoding; if you do not see these specific aliases verbatim, choose the English alias that is most similar. Click “OK” to save your changes. You may now open the HINTS SPSS data without compatibility issues.



Language

Output: English

User Interface: English

Character Encoding for Data and Syntax

● Locale’s writing system

Locale: English-US

○ Unicode (universal character set)

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

Bi-directional Text

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

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

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

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

Base Text Direction: Automatic

This section gives some SPSS (Version 25 and higher) coding examples for common types of statistical analyses using HINTS 5, Cycle 3 data. We begin by creating an analysis plan using the Complex Samples analysis procedures to specify the sample design; TG_ALL_FINWT0 is the sample weight variable (the final weight for the composite sample, no group differences found), VAR_STRATUM is the stratum variable, and VAR_CLUSTER is the cluster variable. The subcommand SRSESTIMATOR specifies the variance estimator under the simple random sampling assumption. The default value is WR (with replacement), and it includes the finite population correction in the variance computation. The subcommand PRINT is used to control output from CSPLAN, and the syntax PLAN means to display a summary of plan specifications. The subcommand DESIGN with keyword STRATA identifies the sampling stratification variable, and the keyword cluster CLUSTER identifies the grouping of sampling units for variance estimation. The subcommand ESTIMATOR specifies the variance estimation method used in the analysis. The syntax TYPE=WR requires the estimation method of selection with replacement.

Note: if group differences are found, the final weight (NWGT0) will need to be substituted in your analysis plan.

```
* Analysis Preparation Wizard.
*substitute your library name in the parentheses of /PLAN FILE=.
CSPLAN ANALYSIS
 /PLAN FILE=’(sample.csaplan)'
 /PLANVARS ANALYSISWEIGHT=TG_all_FINWT0
 /SRSESTIMATOR TYPE=WOR
 /PRINT PLAN
 /DESIGN STRATA=VAR_STRATUM CLUSTER=VAR_CLUSTER
 /ESTIMATOR TYPE=WR.
```

We completed data management of the HINTS 5 Cycle 3 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 GenderC (1=1) (2=2) (ELSE=SYSMIS) INTO gender.
VARIABLE LABELS gender 'gender'.
EXECUTE.

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

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

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

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

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

*Add value labels to recoded variables.
```

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

Frequency Table and Chi-Square Test

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

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

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

			Gender		
Edu			Male	Female	Total
Less than high school	**Population Size**	Estimate	7879153.669	9110566.126	16989719.794
		Standard Error	1132052.432	907520.389	1480229.607
		Unweighted Count	119	209	328
	% within edu	Estimate	46.4%	53.6%	100.0%
		Standard Error	4.3%	4.3%	0.0%
		Unweighted Count	119	209	328
	% within gender	Estimate	6.6%	7.3%	7.0%
		Standard Error	1.0%	0.7%	0.6%
		Unweighted Count	119	209	328
	% of Total	Estimate	3.2%	3.7%	7.0%
		Standard Error	0.5%	0.4%	0.6%
		Unweighted Count	119	209	328
12 years or completed high school	**Population Size**	Estimate	28884258.000	27707952.346	56592210.345
		Standard Error	2349787.567	1592739.435	3031935.225
		Unweighted Count	381	552	933
	% within edu	Estimate	51.0%	49.0%	100.0%
		Standard Error	2.3%	2.3%	0.0%
		Unweighted Count	381	552	933
	% within gender	Estimate	24.3%	22.2%	23.2%
		Standard Error	1.6%	1.1%	1.0%
		Unweighted Count	381	552	933
	% of Total	Estimate	11.9%	11.4%	23.2%
		Standard Error	0.9%	0.6%	1.0%
		Unweighted Count	381	552	933
Some college	**Population Size**	Estimate	48033856.544	49983566.070	98017422.614
		Standard Error	3064908.430	2955566.477	4370348.127
		Unweighted Count	673	901	1574
	% within edu	Estimate	49.0%	51.0%	100.0%

		Standard Error	2.1%	2.1%	0.0%
		Unweighted Count	673	901	1574
	% within gender	Estimate	40.5%	40.0%	40.2%
		Standard Error	1.9%	1.6%	1.3%
		Unweighted Count	673	901	1574
	% of Total	Estimate	19.7%	20.5%	40.2%
		Standard Error	1.1%	1.1%	1.3%
		Unweighted Count	673	901	1574
College graduate or higher	**Population Size**	Estimate	33921235.072	38107150.495	72028385.567
		Standard Error	1547899.457	1453695.112	2068519.469
		Unweighted Count	1030	1369	2399
	% within edu	Estimate	47.1%	52.9%	100.0%
		Standard Error	1.5%	1.5%	0.0%
		Unweighted Count	1030	1369	2399
	% within gender	Estimate	28.6%	30.5%	29.6%
		Standard Error	1.3%	1.3%	1.0%
		Unweighted Count	1030	1369	2399
	% of Total	Estimate	13.9%	15.6%	29.6%
		Standard Error	0.7%	0.6%	1.0%
		Unweighted Count	1030	1369	2399
Total	**Population Size**	Estimate	118718503.285	124909235.035	243627738.320
		Standard Error	4309393.547	3778118.101	5743779.804
		Unweighted Count	2203	3031	5234
	% within edu	Estimate	48.7%	51.3%	100.0%
		Standard Error	1.2%	1.2%	0.0%
		Unweighted Count	2203	3031	5234
	% within gender	Estimate	100.0%	100.0%	100.0%
		Standard Error	0.0%	0.0%	0.0%
		Unweighted Count	2203	3031	5234
	% of Total	Estimate	48.7%	51.3%	100.0%

		Standard Error	1.2%	1.2%	0.0%
		Unweighted Count	2203	3031	5234

The row percentages above show that a higher weighted proportion of college graduates in the sample are women (53%) than men (47%). Respondents with less than a high school diploma include more women (54%) than men (46%).

		Chi-Square	Adjusted F	df1	df2	Significance
education * gender	Pearson	5.126	0.677	2.834	833.155	0.558
	Likelihood Ratio	5.127	0.677	2.834	833.155	0.558

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

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

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

Logistic Regression

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

```
*Multivariable logistic regression of gender and education on SeekCancerInfo.
CSLOGISTIC  seekcancerinfo_recode (LOW) BY flippedgender flippededu
 /PLAN FILE=’(sample.csaplan)’
 /MODEL flippedgender flippededu
 /CUSTOM  Label = 'Overall model minus intercept'
  LMATRIX = flippedgender 1/2 -1/2;
           flippededu 1/3 1/3 1/3 -1;
          flippededu 1/3 1/3 -1 1/3 ;
          flippededu 1/3 -1 1/3 1/3;
          flippededu -1 1/3 1/3 1/3
  /CUSTOM  Label = 'Gender'
 LMATRIX =  flippedgender 1/2 -1/2
  /CUSTOM  Label = 'Education overall'
  LMATRIX = flippededu 1/3 1/3 1/3 -1;
          flippededu 1/3 1/3 -1 1/3 ;
          flippededu 1/3 -1 1/3 1/3;
          flippededu -1 1/3 1/3 1/3
 /INTERCEPT INCLUDE=YES SHOW=YES
 /STATISTICS PARAMETER SE CINTERVAL TTEST EXP
 /TEST TYPE=CHISQUARE PADJUST=LSD
 /ODDSRATIOS FACTOR=[flippedgender(HIGH)]
 /ODDSRATIOS FACTOR=[flippededu(HIGH)]
 /MISSING CLASSMISSING=EXCLUDE
 /CRITERIA MXITER=100 MXSTEP=50 PCONVERGE=[1e-008 RELATIVE] LCONVERGE=[0]
CHKSEP=20 CILEVEL=95
 /PRINT SUMMARY COVB CORB VARIABLEINFO SAMPLEINFO.
```

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

Sample Design Information

		N
Unweighted Cases	**Valid**	5155
	Invalid	283
	Total	5438
Population Size		240838999.531
Stage 1	**Strata**	6
	Units	300
Sampling Design Degrees of Freedom		294

(continued on the next page)

Parameter Estimates

				95% Confidence Interval		Hypothesis Test				95% Confidence Interval for Exp(B)	
seekcancerinfo_recode		B	Std. Error	Lower	Upper	t	df	Sig.	Exp(B)	Lower	Upper
Yes	**(Intercept)**	-0.793	0.220	-1.227	-0.360	-3.603	294.000	0.000	0.452	0.293	0.698
	Female	0.429	0.092	0.247	0.611	4.643	294.000	0.000	1.536	1.280	1.842
	Male	.000[a]							1.000		
	College graduate or higher	1.186	0.225	0.743	1.630	5.264	294.000	0.000	3.275	2.102	5.103
	Some college	0.941	0.225	0.498	1.385	4.178	294.000	0.000	2.563	1.645	3.994
	12 years or completed high school	0.125	0.239	-0.345	0.596	0.525	294.000	0.600	1.134	0.708	1.815
	Less than high school	.000[a]							1.000		

Odds Ratios

				95% Confidence Interval	
seekcancerinfo_recode			Odds Ratio	Lower	Upper
Gender	Female vs. Male	Yes	1.536	1.280	1.842
Education	College graduate or higher vs. Less than high school	Yes	3.275	2.102	5.103
	Some college vs. Less than high school	Yes	2.563	1.645	3.994
	12 years or completed high school vs. Less than high school	Yes	1.134	0.708	1.815

Overall Model Minus Intercept

df	Wald Chi-Square	Sig.
4.000	99.144	0.000

Gender

df	Wald Chi-Square	Sig.
1.000	21.561	0.000

Education Overall

df	Wald Chi-Square	Sig.
3.000	88.040	0.000

To identify levels/variables that display a significant difference in response, the rule of thumb is to examine odds ratios where the confidence interval does not contain 1 (by default, SPSS will use alpha=.05 to determine statistical significance; this value can be changed by the user using code). However, significance may also be garnered from the test of whether the associated beta parameter is equal to 0 (see “Parameter Estimates” table above). According to this model, women appear to be statistically more likely than men to have searched for cancer information.

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

Linear Regression

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

```
* Multivariable linear regression of gender and education on GeneralHealth.
CSGLM genhealth_recode BY flippedgender 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**	5165
	Invalid	273
	Total	5438
Population Size		241189900.791
Stage 1	**Strata**	6
	Units	300
Sampling Design Degrees of Freedom		294

(continued on the next page)

Parameter Estimates

Parameter	Estimate	Std. Error	95% Confidence Interval		Hypothesis Test		
			Lower	Upper	t	df	Sig.
(Intercept)	2.841	0.118	2.608	3.074	24.006	294.000	0.000
College graduate or higher	-0.649	0.116	-0.878	-0.421	-5.587	294.000	0.000
Some college	-0.234	0.117	-0.463	-0.004	-2.002	294.000	0.046
12 years or completed high school	-0.210	0.128	-0.463	0.042	-1.642	294.000	0.102
Less than high school	.000[b]						
Female	0.082	0.046	-0.008	0.173	1.786	294.000	0.075
Male	.000[b]						

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

Overall Model Minus Intercept

df1	df2	Wald F	Sig.
4.000	291.000	35.241	0.000

Gender

df1	df2	Wald F	Sig.
1.000	294.000	3.192	0.075

Education Overall

df1	df2	Wald F	Sig.
3.000	292.000	41.634	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 5, Cycle 3 data. Subsection 1 shows how to complete common analyses using replicate weights, and subsection 2 shows analyses using the Taylor Series linearization approach. For either approach, we begin by doing data management of the HINTS 5 data. We first decided to exclude all “Missing data (Not Ascertained)”, “Multiple responses selected in error”, “Question answered in error (Commission Error)”, and “Inapplicable, coded 2 in SeekHealthInfo” responses from the analyses. By setting these values to missing (.), Stata will exclude these responses from analysis commands where these variables are specifically accessed. For logistic regression modeling within the svy: logit command, Stata expects the response variable to be dichotomous with values (0, 1), so this variable will also be recoded at this point. When recoding existing variables, it is generally recommended to create new variables rather than over-writing the existing variables. Note: New variables should always be compared to original source variables in a Stata **tabulate** command to verify proper coding.

```
use “file path\hints5_cycle3_public.dta”
* Recode negative values to missing
recode genderc (1=1 "Male") (2=2 "Female") (nonmissing=.), generate(gender)

label variable gender "Gender"

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

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


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

replace seekcancerinfo = 0 if seekcancerinfo == 2

replace seekcancerinfo = . if seekcancerinfo == -1 | seekcancerinfo == -
2 | seekcancerinfo == -6 | seekcancerinfo == -9 label define
seekcancerinfo 0 "No" 1 "Yes"

label val seekcancerinfo seekcancerinfo


* Recode negative values to missing for svy: regress

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

Replicate Weights Variance Estimation Method

Declare survey design

Stata requires that the survey design be declared for the dataset globally before any analysis. The declared survey design will be applied to all future survey commands unless another survey design is declared. In this example and declared design we are using TG_ALL_FINWT0 and its associated replicate weights (TG_ALL_FINTWT1 through TG_ALL_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.

Note: if group differences are found, the final weight (NWGT0) and its associated replicate weights (NWGT1 through NWGT150) will need to be substituted in your survey design.

```
* Declare survey design for the data set

svyset [pw=tg_all_finwt0], jkrw(tg_all_finwt1-
tg_all_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     =        5,234
                                              Population size   =  243,627,738
                                              Replications      =           50
                                              Design df         =           49

----------+--------------------------------
          |             Gender
Education |      Male     Female      Total
----------+--------------------------------
 Less tha |   3.23410    3.73954    6.97364
          | (0.47757)  (0.36328)  (0.61024)
          |
 12 years |  11.85590   11.37307   23.22897
          | (0.59936)  (0.43484)  (0.75492)
          |
 Some col |  19.71609   20.51637   40.23246
          | (0.57834)  (0.46364)  (0.63942)
          |
  College |  13.92339   15.64155   29.56494
          | (0.10464)  (0.10577)  (0.14964)
          |
    Total |  48.72947   51.27053    1.0e+02
          | (0.25192)  (0.25192)
----------+--------------------------------
  Key:  cell percentage
        (jackknife standard error of cell percentage)

  Wald (Pearson):
    Unadjusted    chi2(3)         =   41.4201
    Unadjusted    F(3, 49)        =   13.8067     P = 0.0000
    Adjusted      F(3, 47)        =   13.2432     P = 0.0000
```

```
Jknife *: for rows

Number of strata   =          1              Number of obs     =        5,234
                                             Population size   =  243,627,738
                                             Replications      =           50
                                             Design df         =           49

-------------------------------------------
          |             Gender
Education |      Male     Female      Total
----------+--------------------------------
 Less tha |  46.37601   53.62399    1.0e+02
          | (4.33057)  (4.33057)
          |
 12 years |  51.03928   48.96072    1.0e+02
          | (1.55321)  (1.55321)
          |
 Some col |  49.00543   50.99457    1.0e+02
          | (1.03651)  (1.03651)
          |
  College |  47.09426   52.90574    1.0e+02
          | (0.25026)  (0.25025)
          |
    Total |  48.72947   51.27053    1.0e+02
          | (0.25192)  (0.25192)
-------------------------------------------
  Key:  row percentage
        (jackknife standard error of row percentage)

  Wald (Pearson):
    Unadjusted    chi2(3)         =   41.4201
    Unadjusted    F(3, 49)        =   13.8067     P = 0.0000
    Adjusted      F(3, 47)        =   13.2432     P = 0.0000
```

```
Jknife *: for columns

Number of strata   =         1              Number of obs     =        5,234
                                            Population size   =  243,627,738
                                            Replications      =           50
                                            Design df         =           49

──────────┬──────────────────────────────────
          │              Gender
Education │       Male     Female      Total
──────────┼──────────────────────────────────
 Less tha │    6.63684    7.29375    6.97364
          │  (0.98272)  (0.71052)  (0.61024)
          │
 12 years │   24.33004   22.18247   23.22897
          │  (1.20752)  (0.84221)  (0.75492)
          │
 Some col │   40.46029   40.01591   40.23246
          │  (1.16209)  (0.85438)  (0.63942)
          │
  College │   28.57283   30.50787   29.56494
          │  (0.22926)  (0.19527)  (0.14964)
          │
    Total │    1.0e+02    1.0e+02    1.0e+02
          │
──────────┴──────────────────────────────────
  Key:  column percentage
        (jackknife standard error of column percentage)

  Wald (Pearson):
    Unadjusted    chi2(3)         =   41.4201
    Unadjusted    F(3, 49)        =   13.8067     P = 0.0000
    Adjusted      F(3, 47)        =   13.2432     P = 0.0000
```

For the purposes of computing appropriate degrees of freedom for the estimator of the HINTS 5, Cycle 3 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)

Jackknife replications (50)
────┼─── 1 ───┼─── 2 ───┼─── 3 ───┼─── 4 ───┼─── 5
..................................................    50

Survey: Logistic regression

Number of strata   =         1               Number of obs     =       5,155
                                             Population size   =  240,839,000
                                             Replications      =          50
                                             Design df         =          49
                                             F(   4,     46)   =       21.54
                                             Prob > F          =      0.0000

────────────────┬───────────────────────────────────────────────────────────────
                │             Jknife *
seekcancerinfo  │      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
────────────────┼───────────────────────────────────────────────────────────────
      _Igender_2 │   .4289629   .0986868     4.35   0.000     .2306442    .6272815
         _Iedu_2 │   .1254791   .2086313     0.60   0.550    -.2937811    .5447393
         _Iedu_3 │   .9413229      .2043     4.61   0.000     .5307667    1.351879
         _Iedu_4 │   1.186381   .2091227     5.67   0.000     .7661328    1.606628
           _cons │  -.7934547    .204117    -3.89   0.000    -1.203643   -.3832662
────────────────┴───────────────────────────────────────────────────────────────
```

```
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) =   29.21
            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) =   22.95
            Prob > F =    0.0000
```

```
Unadjusted Wald test

 ( 1)  [seekcancerinfo]_Igender_2 = 0

       F(  1,    49) =   18.89
            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) =   27.27
            Prob > F =    0.0000
```

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

Jackknife replications (50)
────┼─── 1 ───┼─── 2 ───┼─── 3 ───┼─── 4 ───┼─── 5
..................................................    50

Survey: Logistic regression

Number of strata   =         1                Number of obs     =       5,155
                                              Population size   =  240,839,000
                                              Replications      =          50
                                              Design df         =          49
                                              F(   4,     46)   =       21.54
                                              Prob > F          =      0.0000

───────────────┬────────────────────────────────────────────────────────────────
               │              Jknife *
seekcancerinfo │ Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
───────────────┼────────────────────────────────────────────────────────────────
    _Igender_2 │   1.535664   .1515498     4.35   0.000     1.259411    1.872513
       _Iedu_2 │   1.133691   .2365235     0.60   0.550     .7454396    1.724159
       _Iedu_3 │    2.56337   .5236965     4.61   0.000     1.700235    3.864681
       _Iedu_4 │   3.275205   .6849197     5.67   0.000      2.15143    4.985972
         _cons │   .4522796    .092318    -3.89   0.000     .3000989    .6816314
───────────────┴────────────────────────────────────────────────────────────────
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 1.54 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 Igender2 Iedu2 Iedu3 Iedu4 _cons, nosvyadjust
test Igender2 Iedu2 Iedu3 Iedu4, nosvyadjust
test Igender2, nosvyadjust
test Iedu2 Iedu3 Iedu4, nosvyadjust
```

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

Jackknife replications (50)
────┼─── 1 ───┼─── 2 ───┼─── 3 ───┼─── 4 ───┼─── 5
..................................................    50

Survey: Linear regression

Number of strata   =         1                Number of obs     =       5,165
                                              Population size   =  241,189,901
                                              Replications      =          50
                                              Design df         =          49
                                              F(   4,     46)   =       29.87
                                              Prob > F          =      0.0000
                                              R-squared         =      0.0511

───────────────┬────────────────────────────────────────────────────────────────
               │             Jknife *
generalhea~h   │      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
───────────────┼────────────────────────────────────────────────────────────────
   _Igender_2  │   .0821506   .0461362     1.78   0.081    -.0105636    .1748647
      _Iedu_2  │  -.2104414   .1270048    -1.66   0.104    -.4656671    .0447843
      _Iedu_3  │  -.2336853   .1133854    -2.06   0.045    -.4615419   -.0058287
      _Iedu_4  │  -.6492976   .1172977    -5.54   0.000    -.8850161    -.413579
        _cons  │     2.8414   .1142805    24.86   0.000     2.611745    3.071055
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) = 3600.23
            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) =   31.82
            Prob > F =    0.0000
```

```
Unadjusted Wald test

 ( 1)  _Igender_2 = 0

       F(  1,    49) =    3.17
            Prob > F =    0.0812



Unadjusted Wald test

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

       F(  3,    49) =   41.07
            Prob > F =    0.0000
```

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

Taylor Series Linearization Variance Estimation Method

Declare survey design

Stata requires that the survey design be declared for the dataset globally before any analysis. The declared survey design will be applied to all future survey commands unless another survey design is declared. In this example and declared design we are using TG_ALL_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.
Note: if group differences are found, the final weight (NWGT0) will need to be substituted in your survey design.

```
* Declare survey design for the data set (Taylor series)
svyset var_cluster [pw=tg_all_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   =          6                Number of obs     =        5,234
Number of PSUs     =        300                Population size   =  243,627,738
                                               Design df         =          294

-------------------------------------------
          |             Gender
Education |      Male     Female      Total
----------+--------------------------------
 Less tha |   3.23410    3.73954    6.97364
          | (0.45790)  (0.36970)  (0.59314)
          |
 12 years |  11.85590   11.37307   23.22897
          | (0.86456)  (0.59194)  (1.01652)
          |
 Some col |  19.71609   20.51637   40.23246
          | (1.08032)  (1.06197)  (1.30212)
          |
  College |  13.92339   15.64155   29.56494
          | (0.66414)  (0.64997)  (0.96222)
          |
    Total |  48.72947   51.27053    1.0e+02
          | (1.17787)  (1.17787)
-------------------------------------------
  Key:  cell percentage
        (linearized standard error of cell percentage)

  Wald (Pearson):
    Unadjusted    chi2(3)         =    2.4138
    Unadjusted    F(3, 294)       =    0.8046     P = 0.4921
    Adjusted      F(3, 292)       =    0.7991     P = 0.4952
```

```
(running tabulate on estimation sample)

Number of strata   =         6             Number of obs     =       5,234
Number of PSUs     =       300             Population size   =  243,627,738
                                           Design df         =         294

──────────┬────────────────────────────────
          │            Gender
Education │      Male     Female      Total
──────────┼────────────────────────────────
 Less tha │  46.37601   53.62399    1.0e+02
          │ (4.26179)  (4.26179)
          │
 12 years │  51.03928   48.96072    1.0e+02
          │ (2.30458)  (2.30458)
          │
 Some col │  49.00543   50.99457    1.0e+02
          │ (2.11509)  (2.11509)
          │
  College │  47.09426   52.90574    1.0e+02
          │ (1.51879)  (1.51879)
          │
    Total │  48.72947   51.27053    1.0e+02
          │ (1.17787)  (1.17787)
──────────┴────────────────────────────────
  Key:  row percentage
        (linearized standard error of row percentage)

  Wald (Pearson):
    Unadjusted    chi2(3)         =    2.4138
    Unadjusted    F(3, 294)       =    0.8046     P = 0.4921
    Adjusted      F(3, 292)       =    0.7991     P = 0.4952
```

```
(running tabulate on estimation sample)

Number of strata   =         6                Number of obs     =       5,234
Number of PSUs     =       300                Population size   =  243,627,738
                                              Design df         =         294

──────────┬──────────────────────────────
          │             Gender
Education │      Male     Female      Total
──────────┼──────────────────────────────
 Less tha │   6.63684    7.29375    6.97364
          │ (0.95742)  (0.70329)  (0.59314)
          │
 12 years │  24.33004   22.18247   23.22897
          │ (1.59365)  (1.12166)  (1.01652)
          │
 Some col │  40.46029   40.01591   40.23246
          │ (1.90763)  (1.56291)  (1.30212)
          │
  College │  28.57283   30.50787   29.56494
          │ (1.28779)  (1.27061)  (0.96222)
          │
    Total │   1.0e+02    1.0e+02    1.0e+02
          │
──────────┴──────────────────────────────
  Key:  column percentage
        (linearized standard error of column percentage)

  Wald (Pearson):
    Unadjusted    chi2(3)         =    2.4138
    Unadjusted    F(3, 294)       =    0.8046     P = 0.4921
    Adjusted      F(3, 292)       =    0.7991     P = 0.4952
```

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   =         6                Number of obs     =       5,155
Number of PSUs     =       300                Population size   =  240,839,000
                                              Design df         =         294
                                              F(   4,    291)   =       24.53
                                              Prob > F          =      0.0000

------------------------------------------------------------------------------
               |             Linearized
seekcancerinfo |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------+--------------------------------------------------------------
    _Igender_2 |   .4289629   .0923815     4.64   0.000       .24715    .6107757
       _Iedu_2 |   .1254791    .239105     0.52   0.600    -.3450952    .5960534
       _Iedu_3 |   .9413229   .2253038     4.18   0.000     .4979103    1.384735
       _Iedu_4 |   1.186381   .2253628     5.26   0.000     .7428517    1.629909
         _cons |  -.7934547    .220213    -3.60   0.000    -1.226848   -.3600611
------------------------------------------------------------------------------


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,   294) =   27.53
            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,   294) =   24.79
            Prob > F =    0.0000
```

```
Unadjusted Wald test

 ( 1)  [seekcancerinfo]_Igender_2 = 0

       F(  1,   294) =   21.56
            Prob > F =    0.0000
```

```
Unadjusted Wald test

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

       F(  3,   294) =   29.35
            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   =         6                  Number of obs     =      5,155
Number of PSUs     =       300                  Population size   =  240,839,000
                                                Design df         =        294
                                                F(   4,    291)   =      24.53
                                                Prob > F          =     0.0000

--------------------------------------------------------------------------------
               |             Linearized
seekcancerinfo | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------+----------------------------------------------------------------
    _Igender_2 |   1.535664   .1418669     4.64   0.000     1.280371     1.84186
       _Iedu_2 |   1.133691   .2710713     0.52   0.600      .708153    1.814942
       _Iedu_3 |    2.56337   .5775369     4.18   0.000     1.645279    3.993769
       _Iedu_4 |   3.275205   .7381095     5.26   0.000     2.101921    5.103412
         _cons |   .4522796   .0995978    -3.60   0.000     .2932152    .6976337
--------------------------------------------------------------------------------
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 1.54 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   =         6                Number of obs     =        5,165
Number of PSUs     =       300                Population size   =  241,189,901
                                              Design df         =          294
                                              F(   4,    291)   =        35.24
                                              Prob > F          =       0.0000
                                              R-squared         =       0.0511

------------------------------------------------------------------------------
             |             Linearized
generalhea~h |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Igender_2 |   .0821506   .0459841     1.79   0.075    -.0083493    .1726504
     _Iedu_2 |  -.2104414    .128189    -1.64   0.102    -.4627258     .041843
     _Iedu_3 |  -.2336853   .1167437    -2.00   0.046    -.4634447    -.003926
     _Iedu_4 |  -.6492976    .116221    -5.59   0.000    -.8780282   -.4205669
       _cons |     2.8414   .1183634    24.01   0.000     2.608453    3.074347
------------------------------------------------------------------------------
```

```
Unadjusted Wald test

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

       F(  5,   294) = 3718.31
            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,   294) =   35.60
            Prob > F =    0.0000
```

```
Unadjusted Wald test

 ( 1)  _Igender_2 = 0

       F(  1,   294) =    3.19
            Prob > F =    0.0750
```

```
Unadjusted Wald test

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

       F(  3,   294) =   41.92
            Prob > F =    0.0000
```

From the above table, it can be seen that compared to those respondents with less than a high school education, those with some college 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.

Merging HINTS Survey Iterations

This section provides SAS, SPSS, and Stata codes to combine HINTS 5, Cycle 3 and HINTS 5, Cycle 2 survey iterations. The provided code will generate one final sample weight for population point estimates and 100 replicate weights to compute standard errors.

Merging HINTS 5, Cycle 3 and HINTS 5, Cycle 2 using SAS

This section provides SAS (Version 9.3 and higher) code for merging the HINTS 5, Cycle 2 and HINTS 5, Cycle 3 iterations. It is suggested that analysts first assess for group differences in the HINTS 5, Cycle 3 data on the variables of interest (See "Recommendations for Statistical Analyses Using HINTS 5 Cycle 3: Assessing for Differences Across Groups" for more information).

The code below is created assuming there are **no differences** between the different modalities in HINTS 5 Cycle 3 and uses the total weights (TG_all weights). An analyst might find that there are group differences on the variables of interest in HINTS 5 Cycle 3. In this case, the analyst would need to create a new variable in both data files that would allow the analyst to differentiate between the 4 groups (HINTS 5 Cycle 2, HINTS 5 Cycle 3 [H5C3] Paper Only, H5C3 Web Option, and H5C3 Web Bonus groups). Next, the analyst will need to use the Rizzo, et al., (2008) method to create 200 replicate weights (an example of the Rizzo code is found in Appendix A).

Alternatively, the analyst could use Taylor linearization to control for group differences by using the NWGT0 weight, VAR_STRATUM, and VAR_CLUSTER variables. In both the Rizzo method and the Taylor linearization method, analysts would need to control for group differences by using the created variable that allows the analyst to differentiate between the 4 groups.

An analyst may also choose to use only one group from Cycle 3 to combine with Cycle 2. One scenario is if an analyst decides they would like to keep modality consistent across both HINTS iterations, and thus it would be appropriate to only use the "paper only" condition from Cycle 3 to combine with Cycle 2 data. This would be achieved by instead using the TG1 (instead of TG_all) variable weights from Cycle 3 in the below sample code.

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.

In order to combine HINTS 5, Cycle 2 with HINTS 5, Cycle 3, the below sample code 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. This survey variable can later be used to easily differentiate the cases that came from each HINTS iteration. 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 = 8,942) that contains one final sample weight (for population point estimates; Merged_WGT0) and 100 replicate weights (Merged_WGT1 through Merged_WGT100; to compute standard errors).

```
/*FIRST CREATE THE FORMAT FOR THE SURVEY VARIABLE*/
proc format;
      value survey
      1="HINTS 5 CYCLE 2"
      2="HINTS 5 CYCLE 3"
      ;
run;

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

```
/*CREATE TWO SEPARATE TEMPORARY DATA FILES THAT CONTAIN THE NEW ‘SURVEY’
VARIABLE.*/

/*PUT NAME OF LIBRARY WHERE HINTS 5 CYCLE 2 FORMATS ARE STORED*/;
options fmtsearch=(LibH5C2);

data tempHINTS5CYCLE2;
      length tmpstrat $10;
      /*PUT NAME OF LIBRARY AND NAME OF EXISTING HINTS 5 CYCLE 3 DATA FILE*/
      set LibH5C2.DataH5C2;

      survey=1;
      format survey survey.;

      *Recreate VAR_STRATUM variable with length of 10 to match H5C3;
      tmpstrat=var_stratum;
      drop var_stratum;
      rename tmpstrat=var_stratum;
run;

/* PUT NAME OF LIBRARY WHERE HINTS 5 CYCLE 3 FORMATS ARE STORED*/
options fmtsearch=(hints5c3);

data tempHINTS5CYCLE3;
      /*PUT NAME OF LIBRARY AND NAME OF EXISTING HINTS 5 CYCLE 3 DATA FILE*/
      set hints5c3.hints5cycle3_formatted;
      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 mergeHINTS5C2_HINTS5C3;
      set tempHINTS5CYCLE2 tempHINTS5CYCLE3;

      /*Create Replicate Weights for trend tests*/
      **Replicate Weights;
      array hints52wgts [50]  person_finwt1-person_finwt50;
      array hints53wgts [50]  TG_all_finwt1-TG_all_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 2;
            Merged_NWGT0=person_finwt0;
            Merged_NWgt[i]=hints52wgts[i];
            Merged_NWgt[50+i]=person_finwt0;
      end;

      else if survey eq 2 then do i=1 to 50; *HINTS 5 CYCLE 3;
            Merged_NWGT0=TG_all_finwt0;
            Merged_NWgt[i]=TG_all_finwt0;
            Merged_NWgt[50+i]=hints53wgts[i];
      end;
run;
/**************************************************************/
/*YOU CAN USE THE CODE BELOW TO RUN SIMPLE FREQUENCIES ON TWO COMMON
VARIABLES, ‘SEEKHEALTHINFO’ AND ‘CHANCEASKQUESTIONS’*/

/*SAS CODE*/
proc surveyfreq data = mergeHINTS5C2_HINTS5C3 varmethod = jackknife;
      weight Merged_NWGT0;
      repweights Merged_NWGT1-Merged_NWGT100 / df = 98 jkcoefs = 0.98;
      tables seekhealthinfo chanceaskquestions;
run;
```

SAS Code to Merge HINTS 5, Cycle 3 and HINTS 5, Cycle 2 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 2"
      2="HINTS 5 CYCLE 3"
      ;
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 TG_ALL_FINWT0 VARIABLE AS OUR
WEIGHTING VARIABLE FROM HINTS 5 CYCLE 3.
/*IF YOU WOULD LIKE TO USE A DIFFERENT WEIGHTING VARIABLE (E.G. YOU ARE
INTERESTED IN THE PAPER ONLY SAMPLE ONLY, TG1_FINWT0), YOU WILL NEED TO
UPDATE THE CODE IN THE HINTS5CYCLE3 RENAME LINE ACCORDINGLY*/

/*PUT NAME OF LIBRARY WHERE HINTS 5 CYCLE 2 FORMATS ARE STORED*/
options fmtsearch=(LibH5C2);

data tempHINTS5CYCLE2;
      length tmpstrat $10;
      /*PUT NAME OF LIBRARY AND NAME OF EXISTING HINTS 5 CYCLE 2 DATA FILE*/
      set LibH5C2.DataH5C2;
      RENAME PERSON_FINWT0=MERGED_FINWT0;

      survey=1;
      format survey survey.;
      *Recreate VAR_STRATUM variable with length of 10 to match H5C3;
      tmpstrat=var_stratum;
      drop var_stratum;
      rename tmpstrat=var_stratum;
```

```
run;

/* PUT NAME OF LIBRARY WHERE HINTS 5 CYCLE 3 FORMATS ARE STORED*/
options fmtsearch=(hints5c3);

data tempHINTS5CYCLE3;
      /*PUT NAME OF LIBRARY AND NAME OF EXISTING HINTS 5 CYCLE 3 DATA FILE*/
      set hints5c3.hints5cycle3_formatted;
      RENAME TG_ALL_FINWT0=MERGED_FINWT0;

      survey=2;
      format survey survey.;
run;

data mergeHINTS5C2_HINTS5C3;
      set tempHINTS5CYCLE2 tempHINTS5CYCLE3;
run;

/**************************************************************/
/*YOU CAN USE THE CODE BELOW TO RUN SIMPLE FREQUENCIES ON TWO
COMMON VARIABLES, ‘SEEKHEALTHINFO’ AND ‘CHANCEASKQUESTIONS’*/

/*SAS CODE*/
proc surveyfreq data = MergeHints5C2_Hints5c3 varmethod = TAYLOR;
      strata VAR_STRATUM;
      cluster VAR_CLUSTER;
      weight MERGED_FINWT0;
      tables seekhealthinfo chanceaskquestions / row col;
run;
```

Merging HINTS 5, Cycle 3 and HINTS 5, Cycle 2 using SPSS

This section provides SPSS (Version 25) syntax for merging the HINTS 5, Cycle 2 and HINTS 5, Cycle 3 iterations. Note that the below sample syntax is created with the assumption that there were no group differences found within HINTS 5, Cycle 3. If an analyst does find group differences in HINTS 5, Cycle 3, the NWGT0 weighting variable will need to be used instead of TG_all_FINWT0. Additionally, the analyst would need to create a new variable in both data files that would allow the analyst to differentiate between the 4 groups (HINTS 5 Cycle 2, HINTS 5 Cycle 3 Paper Only, H5C3 Web Option, and H5C3 Web Bonus groups).

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 = 8,942).

First, you will need to have **HINTS 5, Cycle 3** data open. The below syntax will first save a copy of HINTS 5, Cycle 3 and rename it as a new file called ‘MERGED_H5C3_H5C2.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 Cycle 3’s weighting variable TG_all_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 Cycle 3 a “2” so that analysts can easily identify cases from HINTS 5, Cycle 3.

```
SAVE OUTFILE='H:\HINTS\5 Cycle 3\SPSS\MERGED_H5C3_H5C2.sav'
 /COMPRESSED.
DATASET NAME MERGED_DATA.

DATASET ACTIVATE MERGED_DATA.
ALTER TYPE VAR_STRATUM (A10).
COMPUTE MERGED_FINWT0=TG_all_FINWT0.
COMPUTE Survey=2.
EXECUTE.
```

Next, we need to open our HINTS 5 CYCLE 2 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 2 data and rename the dataset as H5C2. The syntax will then create the ‘Survey’ variable in the HINTS 5, Cycle 2 dataset and give each participant from Cycle 2 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 2 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 5 Cycle 3 dataset
Note, the analyst will need to insert the file path for where HINTS 5 Cycle 2 is saved.

```
**below, you should insert the filepath for your HINTS 5 Cycle 2 data**.
GET FILE='H:\HINTS\5 Cycle 2\HINTS-5_Cycle2_SPSS\hints5_cycle2_public.sav'.
ALTER TYPE VAR_STRATUM (A10).
DATASET NAME H5C2 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='H:\HINTS\5 Cycle 3\SPSS\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_H5C3_H5C2” (which currently contains only HINTS 5 Cycle 3’s data) and “hints5_cycle2_public”. Within your “MERGED_H5C3_H5C2” 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_cycle2_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=*
 /RENAME (AccessUsingHealthApp AlcoholConditions_Cancer AlcoholConditions_Diabetes
   AlcoholConditions_HeartDisease AlcoholConditions_LiverDisease AverageSleepNight AverageSleepQuality
   AvoidDoc CancerSign_BowelBladderChange CancerSign_UnexpBleeding CancerSign_UnexpWeightLoss
   ChanceGetCancerNoDX ChangeThinking ConfidentGetHealthInf ConsiderFuture ConsiderQuit DRA
   Electronic_ECigHarms Electronic_MadeAppts EnjoyExercise EnjoyTimeInSun EverHadPSATest ExRec_Cat
   ExRec_ChangedEx ExRec_DecreasedEx ExRec_IncreasedEx ExRec_LookedInfo ExRec_NoChange
ExRec_NotHeard
   FreqGoUrgentCare FreqWearDevTrackHealth Fruit Frustrated GovPARec_HCP GovPARec_Internet
   GovPARec_Magazine GovPARec_TV HCPAlcoholConsequences HeardHepC HHAdultAge6 HHAdultGender6
   HHAdultMOB6 IncomeFeelings InfluenceCancer_EatingFiber InfluenceCancer_EatingFruitVeg
   InfluenceCancer_ProcessedMeat LotOfEffort LowNicotineAddictive LowNicotineHarmful MAILNUM
   MorningNightPerson NicotineAddictionConcern NicotineCauseCancer NicotineWantSmoke
   NotAccessed_LogInProb NotAccessed_MultipleRec NotAccessed_Uncomfortable nwgt0 nwgt1 nwgt10 nwgt100
   nwgt101 nwgt102 nwgt103 nwgt104 nwgt105 nwgt106 nwgt107 nwgt108 nwgt109 nwgt11 nwgt110 nwgt111
   nwgt112 nwgt113 nwgt114 nwgt115 nwgt116 nwgt117 nwgt118 nwgt119 nwgt12 nwgt120 nwgt121 nwgt122
   nwgt123 nwgt124 nwgt125 nwgt126 nwgt127 nwgt128 nwgt129 nwgt13 nwgt130 nwgt131 nwgt132 nwgt133
   nwgt134 nwgt135 nwgt136 nwgt137 nwgt138 nwgt139 nwgt14 nwgt140 nwgt141 nwgt142 nwgt143 nwgt144
   nwgt145 nwgt146 nwgt147 nwgt148 nwgt149 nwgt15 nwgt150 nwgt16 nwgt17 nwgt18 nwgt19 nwgt2 nwgt20
   nwgt21 nwgt22 nwgt23 nwgt24 nwgt25 nwgt26 nwgt27 nwgt28 nwgt29 nwgt3 nwgt30 nwgt31 nwgt32 nwgt33
   nwgt34 nwgt35 nwgt36 nwgt37 nwgt38 nwgt39 nwgt4 nwgt40 nwgt41 nwgt42 nwgt43 nwgt44 nwgt45 nwgt46
   nwgt47 nwgt48 nwgt49 nwgt5 nwgt50 nwgt51 nwgt52 nwgt53 nwgt54 nwgt55 nwgt56 nwgt57 nwgt58 nwgt59
   nwgt6 nwgt60 nwgt61 nwgt62 nwgt63 nwgt64 nwgt65 nwgt66 nwgt67 nwgt68 nwgt69 nwgt7 nwgt70 nwgt71
   nwgt72 nwgt73 nwgt74 nwgt75 nwgt76 nwgt77 nwgt78 nwgt79 nwgt8 nwgt80 nwgt81 nwgt82 nwgt83 nwgt84
   nwgt85 nwgt86 nwgt87 nwgt88 nwgt89 nwgt9 nwgt90 nwgt91 nwgt92 nwgt93 nwgt94 nwgt95 nwgt96 nwgt97
   nwgt98 nwgt99 OfferedAccessHCP2 OfferedAccessInsurer2 OnlineRecClinNotes OtherDevTrackHealth2
   PhysAct_HelpSleep PhysAct_ReduceAnxiety PhysAct_ReducePain Prompt QualityCareUrgentCare
   RecordsOnline_ViewResults RegExercise_Appearance RegExercise_Enjoyment RegExercise_Guilt
   RegExercise_Pressure SeenFederalCourtTobaccoMessages2 SmokeDayECig StrongNeedHealthInfo
   StrongNeedHealthInfo_OS Sunburned_Alcohol Sunburned_DayToDay Sunburned_DK
Sunburned_DontRemember
   Sunburned_Exercise Sunburned_HomeOutside Sunburned_JobOutside Sunburned_None Sunburned_Other
   Sunburned_OutdoorEvent Sunburned_ProtClothing Sunburned_Shade Sunburned_SPF15
   Sunburned_SportingEvent Sunburned_Sunbathing Sunburned_Swimming SunburnedAct_Cat SunburnedProt_Cat
   TalkHealthFriends TG1_FINWT0 TG1_FINWT1 TG1_FINWT10 TG1_FINWT11 TG1_FINWT12
TG1_FINWT13 TG1_FINWT14
   TG1_FINWT15 TG1_FINWT16 TG1_FINWT17 TG1_FINWT18 TG1_FINWT19 TG1_FINWT2
TG1_FINWT20 TG1_FINWT21
   TG1_FINWT22 TG1_FINWT23 TG1_FINWT24 TG1_FINWT25 TG1_FINWT26 TG1_FINWT27
TG1_FINWT28 TG1_FINWT29
   TG1_FINWT3 TG1_FINWT30 TG1_FINWT31 TG1_FINWT32 TG1_FINWT33 TG1_FINWT34
TG1_FINWT35 TG1_FINWT36
   TG1_FINWT37 TG1_FINWT38 TG1_FINWT39 TG1_FINWT4 TG1_FINWT40 TG1_FINWT41
TG1_FINWT42 TG1_FINWT43
   TG1_FINWT44 TG1_FINWT45 TG1_FINWT46 TG1_FINWT47 TG1_FINWT48 TG1_FINWT49
TG1_FINWT5 TG1_FINWT50
   TG1_FINWT6 TG1_FINWT7 TG1_FINWT8 TG1_FINWT9 TG2_FINWT0 TG2_FINWT1 TG2_FINWT10
TG2_FINWT11
   TG2_FINWT12 TG2_FINWT13 TG2_FINWT14 TG2_FINWT15 TG2_FINWT16 TG2_FINWT17
TG2_FINWT18 TG2_FINWT19
```

```
   TG2_FINWT2 TG2_FINWT20 TG2_FINWT21 TG2_FINWT22 TG2_FINWT23 TG2_FINWT24
TG2_FINWT25 TG2_FINWT26
   TG2_FINWT27 TG2_FINWT28 TG2_FINWT29 TG2_FINWT3 TG2_FINWT30 TG2_FINWT31
TG2_FINWT32 TG2_FINWT33
   TG2_FINWT34 TG2_FINWT35 TG2_FINWT36 TG2_FINWT37 TG2_FINWT38 TG2_FINWT39
TG2_FINWT4 TG2_FINWT40
   TG2_FINWT41 TG2_FINWT42 TG2_FINWT43 TG2_FINWT44 TG2_FINWT45 TG2_FINWT46
TG2_FINWT47 TG2_FINWT48
   TG2_FINWT49 TG2_FINWT5 TG2_FINWT50 TG2_FINWT6 TG2_FINWT7 TG2_FINWT8 TG2_FINWT9
TG3_FINWT0
   TG3_FINWT1 TG3_FINWT10 TG3_FINWT11 TG3_FINWT12 TG3_FINWT13 TG3_FINWT14
TG3_FINWT15 TG3_FINWT16
   TG3_FINWT17 TG3_FINWT18 TG3_FINWT19 TG3_FINWT2 TG3_FINWT20 TG3_FINWT21
TG3_FINWT22 TG3_FINWT23
   TG3_FINWT24 TG3_FINWT25 TG3_FINWT26 TG3_FINWT27 TG3_FINWT28 TG3_FINWT29
TG3_FINWT3 TG3_FINWT30
   TG3_FINWT31 TG3_FINWT32 TG3_FINWT33 TG3_FINWT34 TG3_FINWT35 TG3_FINWT36
TG3_FINWT37 TG3_FINWT38
   TG3_FINWT39 TG3_FINWT4 TG3_FINWT40 TG3_FINWT41 TG3_FINWT42 TG3_FINWT43
TG3_FINWT44 TG3_FINWT45
   TG3_FINWT46 TG3_FINWT47 TG3_FINWT48 TG3_FINWT49 TG3_FINWT5 TG3_FINWT50
TG3_FINWT6 TG3_FINWT7
   TG3_FINWT8 TG3_FINWT9 TG_all_FINWT0 TG_all_FINWT1 TG_all_FINWT10 TG_all_FINWT11
TG_all_FINWT12
   TG_all_FINWT13 TG_all_FINWT14 TG_all_FINWT15 TG_all_FINWT16 TG_all_FINWT17
TG_all_FINWT18
   TG_all_FINWT19 TG_all_FINWT2 TG_all_FINWT20 TG_all_FINWT21 TG_all_FINWT22 TG_all_FINWT23
   TG_all_FINWT24 TG_all_FINWT25 TG_all_FINWT26 TG_all_FINWT27 TG_all_FINWT28
TG_all_FINWT29
   TG_all_FINWT3 TG_all_FINWT30 TG_all_FINWT31 TG_all_FINWT32 TG_all_FINWT33 TG_all_FINWT34
   TG_all_FINWT35 TG_all_FINWT36 TG_all_FINWT37 TG_all_FINWT38 TG_all_FINWT39 TG_all_FINWT4
   TG_all_FINWT40 TG_all_FINWT41 TG_all_FINWT42 TG_all_FINWT43 TG_all_FINWT44
TG_all_FINWT45
   TG_all_FINWT46 TG_all_FINWT47 TG_all_FINWT48 TG_all_FINWT49 TG_all_FINWT5 TG_all_FINWT50
   TG_all_FINWT6 TG_all_FINWT7 TG_all_FINWT8 TG_all_FINWT9 TimesSunburned Treatment_H5C3
TriedQuit
   TrustCharities TrustDoctor TrustFamily TrustGov TrustReligiousOrgs UnderstandOnlineMedRec
   Vegetables WearableDevTrackHealth WeightIntention WeightPerception WillingShareData_Fam
   WillingShareData_HCP=d0 d1 d2 d3 d4 d5 d6 d7 d8 d9 d10 d11 d12 d13 d14 d15 d16 d17 d18 d19 d20 d21
   d22 d23 d24 d25 d26 d27 d28 d29 d30 d31 d32 d33 d34 d35 d36 d37 d38 d39 d40 d41 d42 d43 d44 d45 d46
   d47 d48 d49 d50 d51 d52 d53 d54 d55 d56 d57 d58 d59 d60 d61 d62 d63 d64 d65 d66 d67 d68 d69 d70 d71
   d72 d73 d74 d75 d76 d77 d78 d79 d80 d81 d82 d83 d84 d85 d86 d87 d88 d89 d90 d91 d92 d93 d94 d95 d96
   d97 d98 d99 d100 d101 d102 d103 d104 d105 d106 d107 d108 d109 d110 d111 d112 d113 d114 d115 d116
   d117 d118 d119 d120 d121 d122 d123 d124 d125 d126 d127 d128 d129 d130 d131 d132 d133 d134 d135 d136
   d137 d138 d139 d140 d141 d142 d143 d144 d145 d146 d147 d148 d149 d150 d151 d152 d153 d154 d155 d156
   d157 d158 d159 d160 d161 d162 d163 d164 d165 d166 d167 d168 d169 d170 d171 d172 d173 d174 d175 d176
   d177 d178 d179 d180 d181 d182 d183 d184 d185 d186 d187 d188 d189 d190 d191 d192 d193 d194 d195 d196
   d197 d198 d199 d200 d201 d202 d203 d204 d205 d206 d207 d208 d209 d210 d211 d212 d213 d214 d215 d216
   d217 d218 d219 d220 d221 d222 d223 d224 d225 d226 d227 d228 d229 d230 d231 d232 d233 d234 d235 d236
   d237 d238 d239 d240 d241 d242 d243 d244 d245 d246 d247 d248 d249 d250 d251 d252 d253 d254 d255 d256
   d257 d258 d259 d260 d261 d262 d263 d264 d265 d266 d267 d268 d269 d270 d271 d272 d273 d274 d275 d276
   d277 d278 d279 d280 d281 d282 d283 d284 d285 d286 d287 d288 d289 d290 d291 d292 d293 d294 d295 d296
   d297 d298 d299 d300 d301 d302 d303 d304 d305 d306 d307 d308 d309 d310 d311 d312 d313 d314 d315 d316
   d317 d318 d319 d320 d321 d322 d323 d324 d325 d326 d327 d328 d329 d330 d331 d332 d333 d334 d335 d336
   d337 d338 d339 d340 d341 d342 d343 d344 d345 d346 d347 d348 d349 d350 d351 d352 d353 d354 d355 d356
   d357 d358 d359 d360 d361 d362 d363 d364 d365 d366 d367 d368 d369 d370 d371 d372 d373 d374 d375 d376
```

```
    d377 d378 d379 d380 d381 d382 d383 d384 d385 d386 d387 d388 d389 d390 d391 d392 d393 d394 d395 d396
    d397 d398 d399 d400 d401 d402 d403 d404 d405 d406 d407 d408 d409 d410 d411 d412 d413 d414 d415 d416
    d417 d418 d419 d420 d421 d422 d423 d424 d425 d426 d427 d428 d429 d430 d431 d432 d433 d434 d435 d436
    d437 d438 d439 d440 d441 d442 d443 d444 d445 d446 d447 d448 d449 d450 d451 d452 d453 d454 d455 d456
    d457 d458 d459 d460 d461 d462 d463)
 /FILE='H5C2'
 /RENAME (ActiveDutyArmedForces BornInUSA CancerConcernedQuality CancerConfidentGetHealthInf
    CancerFrustrated CancerLotOfEffort CancerTooHardUnderstand CancerTrustCharities CancerTrustDoctor
    CancerTrustFamily CancerTrustGov CancerTrustInternet CancerTrustNewsMag CancerTrustRadio
    CancerTrustReligiousOrgs CancerTrustTelevision Caregiver_AccessHelp Caregiver_Counseling
    Caregiver_MedTrain Caregiver_RespiteCare Caregiver_SupportGroup CaregiverTraining_Cat
    CaregiverTraining_Hotline CaregiverTraining_InPerson CaregiverTraining_OnlineVideo
    CaregiverTraining_ReadingMat CaregiverTraining_Virtual Caregiving_ArrangeSvcs Caregiving_Bathing
    Caregiving_BedsChairs Caregiving_CommunicateHCP Caregiving_Dressing Caregiving_Feeding
    Caregiving_Finances Caregiving_Housework Caregiving_HowLong Caregiving_Incontinence
    Caregiving_MealPrep Caregiving_MedTasks Caregiving_Reside Caregiving_Shopping Caregiving_SpendTime
    Caregiving_Toilet Caregiving_Transportation CaregivingActivities_Cat CaregivingMedAct_Cat
    ConfidentFamilyHistory ConfidentInfoSafe Disabled Electronic_HealthInfoSE
    Electronic_LookedAssistance EmotionalSupport2 Employed EverOfferedAccessRec FamBetween9and27
    FamiliarFamilyCancer FamilyCancer_Brother FamilyCancer_Cat FamilyCancer_Children
    FamilyCancer_Father FamilyCancer_HCP FamilyCancer_Mother FamilyCancer_None FamilyCancer_OthFam
    FamilyCancer_Sister FORM_NAME FreqWorryCancerAgain HCPAdvisedLimitingSun HelpDailyChores2
    HelpPreparingMeals HelpRunErrands HelpTransportDoctor Homemaker ImagineCancer ImagineCancerAgain
    InfluenceCancer_EatingHealthy InfluenceCancer_RegExercise KnowledgePalliativeCare MailSurveyTimeHrs
    MailSurveyTimeMin MedConditions_Arthritis MostRecentCheckup2 MultiOcc NotAccessed_Other
    NotAccessed_Other_OS OccupationStatus OccupationStatus_OS OtherDevTrackHealth OtherOcc
    PCGoal_HelpFamCope PCGoal_ManageSymptoms PCGoal_MoreTime PCGoal_SocEmotSupport
PCHospiceCare
    PCMeansGivingUp PCObligatedToInform PCStopTreatments PCStrongNeedInfo PCThinkDeath PCTrustInfo
    PERSON_FINWT0 PERSON_FINWT1 PERSON_FINWT10 PERSON_FINWT11 PERSON_FINWT12
PERSON_FINWT13
    PERSON_FINWT14 PERSON_FINWT15 PERSON_FINWT16 PERSON_FINWT17 PERSON_FINWT18
PERSON_FINWT19
    PERSON_FINWT2 PERSON_FINWT20 PERSON_FINWT21 PERSON_FINWT22 PERSON_FINWT23
PERSON_FINWT24
    PERSON_FINWT25 PERSON_FINWT26 PERSON_FINWT27 PERSON_FINWT28 PERSON_FINWT29
PERSON_FINWT3
    PERSON_FINWT30 PERSON_FINWT31 PERSON_FINWT32 PERSON_FINWT33 PERSON_FINWT34
PERSON_FINWT35
    PERSON_FINWT36 PERSON_FINWT37 PERSON_FINWT38 PERSON_FINWT39 PERSON_FINWT4
PERSON_FINWT40
    PERSON_FINWT41 PERSON_FINWT42 PERSON_FINWT43 PERSON_FINWT44 PERSON_FINWT45
PERSON_FINWT46
    PERSON_FINWT47 PERSON_FINWT48 PERSON_FINWT49 PERSON_FINWT5 PERSON_FINWT50
PERSON_FINWT6
    PERSON_FINWT7 PERSON_FINWT8 PERSON_FINWT9 ProbCare_BringTest ProbCare_ProvideHist
ProbCare_RedoTest
    ProbCare_WaitLong ReceivedCareVA RecommendHPVShot RecordsOnline_Allergies
RecordsOnline_ClinNotes
    RecordsOnline_HealthProbs RecordsOnline_Immunizations RecordsOnline_Paperwork
    RecordsOnline_VisitSummary Retired SeenFederalCourtTobaccoMessages StrongNeedCancerInfo
    StrongNeedCancerInfo_OS Student SunEffectAfter1Hour TalkHealthFriends2 TimesUsedTanningBed
    TypeOfAddressA TypeOfAddressB TypeOfAddressC TypeOfAddressD Unemployed WhoOffered_Cat
    WhoOffered_HCP WhoOffered_Insurer WhoOffered_Other WhoOffered_Other_OS WithheldInfoPrivacy
    YearCameToUSA=d464 d465 d466 d467 d468 d469 d470 d471 d472 d473 d474 d475 d476 d477 d478 d479
d480
```

```
   d481 d482 d483 d484 d485 d486 d487 d488 d489 d490 d491 d492 d493 d494 d495 d496 d497 d498 d499 d500
   d501 d502 d503 d504 d505 d506 d507 d508 d509 d510 d511 d512 d513 d514 d515 d516 d517 d518 d519 d520
   d521 d522 d523 d524 d525 d526 d527 d528 d529 d530 d531 d532 d533 d534 d535 d536 d537 d538 d539 d540
   d541 d542 d543 d544 d545 d546 d547 d548 d549 d550 d551 d552 d553 d554 d555 d556 d557 d558 d559 d560
   d561 d562 d563 d564 d565 d566 d567 d568 d569 d570 d571 d572 d573 d574 d575 d576 d577 d578 d579 d580
   d581 d582 d583 d584 d585 d586 d587 d588 d589 d590 d591 d592 d593 d594 d595 d596 d597 d598 d599 d600
   d601 d602 d603 d604 d605 d606 d607 d608 d609 d610 d611 d612 d613 d614 d615 d616 d617 d618 d619 d620
   d621 d622 d623 d624 d625 d626 d627 d628 d629 d630 d631 d632 d633 d634 d635 d636 d637 d638 d639 d640
   d641 d642 d643 d644 d645 d646)
 /DROP=d0 d1 d2 d3 d4 d5 d6 d7 d8 d9 d10 d11 d12 d13 d14 d15 d16 d17 d18 d19 d20 d21 d22 d23 d24
   d25 d26 d27 d28 d29 d30 d31 d32 d33 d34 d35 d36 d37 d38 d39 d40 d41 d42 d43 d44 d45 d46 d47 d48 d49
   d50 d51 d52 d53 d54 d55 d56 d57 d58 d59 d60 d61 d62 d63 d64 d65 d66 d67 d68 d69 d70 d71 d72 d73 d74
   d75 d76 d77 d78 d79 d80 d81 d82 d83 d84 d85 d86 d87 d88 d89 d90 d91 d92 d93 d94 d95 d96 d97 d98 d99
   d100 d101 d102 d103 d104 d105 d106 d107 d108 d109 d110 d111 d112 d113 d114 d115 d116 d117 d118 d119
   d120 d121 d122 d123 d124 d125 d126 d127 d128 d129 d130 d131 d132 d133 d134 d135 d136 d137 d138 d139
   d140 d141 d142 d143 d144 d145 d146 d147 d148 d149 d150 d151 d152 d153 d154 d155 d156 d157 d158 d159
   d160 d161 d162 d163 d164 d165 d166 d167 d168 d169 d170 d171 d172 d173 d174 d175 d176 d177 d178 d179
   d180 d181 d182 d183 d184 d185 d186 d187 d188 d189 d190 d191 d192 d193 d194 d195 d196 d197 d198 d199
   d200 d201 d202 d203 d204 d205 d206 d207 d208 d209 d210 d211 d212 d213 d214 d215 d216 d217 d218 d219
   d220 d221 d222 d223 d224 d225 d226 d227 d228 d229 d230 d231 d232 d233 d234 d235 d236 d237 d238 d239
   d240 d241 d242 d243 d244 d245 d246 d247 d248 d249 d250 d251 d252 d253 d254 d255 d256 d257 d258 d259
   d260 d261 d262 d263 d264 d265 d266 d267 d268 d269 d270 d271 d272 d273 d274 d275 d276 d277 d278 d279
   d280 d281 d282 d283 d284 d285 d286 d287 d288 d289 d290 d291 d292 d293 d294 d295 d296 d297 d298 d299
   d300 d301 d302 d303 d304 d305 d306 d307 d308 d309 d310 d311 d312 d313 d314 d315 d316 d317 d318 d319
   d320 d321 d322 d323 d324 d325 d326 d327 d328 d329 d330 d331 d332 d333 d334 d335 d336 d337 d338 d339
   d340 d341 d342 d343 d344 d345 d346 d347 d348 d349 d350 d351 d352 d353 d354 d355 d356 d357 d358 d359
   d360 d361 d362 d363 d364 d365 d366 d367 d368 d369 d370 d371 d372 d373 d374 d375 d376 d377 d378 d379
   d380 d381 d382 d383 d384 d385 d386 d387 d388 d389 d390 d391 d392 d393 d394 d395 d396 d397 d398 d399
   d400 d401 d402 d403 d404 d405 d406 d407 d408 d409 d410 d411 d412 d413 d414 d415 d416 d417 d418 d419
   d420 d421 d422 d423 d424 d425 d426 d427 d428 d429 d430 d431 d432 d433 d434 d435 d436 d437 d438 d439
   d440 d441 d442 d443 d444 d445 d446 d447 d448 d449 d450 d451 d452 d453 d454 d455 d456 d457 d458 d459
   d460 d461 d462 d463 d464 d465 d466 d467 d468 d469 d470 d471 d472 d473 d474 d475 d476 d477 d478 d479
   d480 d481 d482 d483 d484 d485 d486 d487 d488 d489 d490 d491 d492 d493 d494 d495 d496 d497 d498 d499
   d500 d501 d502 d503 d504 d505 d506 d507 d508 d509 d510 d511 d512 d513 d514 d515 d516 d517 d518 d519
   d520 d521 d522 d523 d524 d525 d526 d527 d528 d529 d530 d531 d532 d533 d534 d535 d536 d537 d538 d539
   d540 d541 d542 d543 d544 d545 d546 d547 d548 d549 d550 d551 d552 d553 d554 d555 d556 d557 d558 d559
   d560 d561 d562 d563 d564 d565 d566 d567 d568 d569 d570 d571 d572 d573 d574 d575 d576 d577 d578 d579
   d580 d581 d582 d583 d584 d585 d586 d587 d588 d589 d590 d591 d592 d593 d594 d595 d596 d597 d598 d599
   d600 d601 d602 d603 d604 d605 d606 d607 d608 d609 d610 d611 d612 d613 d614 d615 d616 d617 d618 d619
   d620 d621 d622 d623 d624 d625 d626 d627 d628 d629 d630 d631 d632 d633 d634 d635 d636 d637 d638 d639
   d640 d641 d642 d643 d644 d645 d646.
EXECUTE.
/*********************************************************/

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

*INSERT PATH OF TO ANALYSIS PLAN UNDER /PLAN FILE.
CSTABULATE
/PLAN FILE='H:\HINTS\5 Cycle 3\SPSS\MergePlan.csaplan'
/TABLES VARIABLES=seekhealthinfo chanceaskquestions
/CELLS POPSIZE TABLEPCT
/STATISTICS SE COUNT
/MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

Merging HINTS 5, Cycle 3 and HINTS 5, Cycle 2 using STATA

This section provides Stata (Version 10.0 and higher) code for merging the HINTS 5, Cycle 2 and HINTS 5, Cycle 3 iterations. It is suggested that analysts first assess for group differences in the HINTS 5, Cycle 3 data on the variables of interest (See “Recommendations for Statistical Analyses Using HINTS 5 Cycle 3: Assessing for Differences Across Groups” for more information).

The code below is created assuming there are **no differences** between the different modalities in HINTS 5 Cycle 3 and uses the total weights (TG_all weights). An analyst might find that there are group differences on the variables of interest in HINTS 5 Cycle 3. In this case, the analyst would need to create a new variable in both data files that would allow the analyst to differentiate between the 4 groups (HINTS 5 Cycle 2, HINTS 5 Cycle 3 [H5C3] Paper Only, H5C3 Web Option, and H5C3 Web Bonus groups). Next, the analyst will need to use the Rizzo, et al., (2008) method to create 200 replicate weights (an example of the Rizzo code is found in Appendix A.

Alternatively, the analyst could use Taylor linearization to control for group differences by using the NWGT0 weight, VAR_STRATUM, and VAR_CLUSTER variables. In both the Rizzo method and the Taylor linearization method, analysts would need to control for group differences by using the created variable that allows the analyst to differentiate between the 4 groups.

An analyst may also choose to use only one group from Cycle 3 to combine with Cycle 2. One scenario is if an analyst decides they would like to keep modality consistent across both HINTS iterations, and thus it would be appropriate to only use the “paper only” condition from Cycle 3 to combine with Cycle 2 data. This would be achieved by instead using the TG1 (instead of TG_all) variable weights from Cycle 3 in the below sample code.

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

In order to combine HINTS 5, Cycle 2 with HINTS 5, Cycle 3, 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 = 8,942).

```
*Put path and name to your HINTS 5 Cycle 2 data
use "H:\HINTS\HINTS5-Cycle2\hints5_cycle2_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 h5c2.dta, replace

*Put path and name to your HINTS 5 Cycle 3 data
use "H:\HINTS\HINTS5-Cycle3\hints5_cycle3_public.dta", clear
*Create final and replicate weights (merged_nwt*) for multi-cycle datasets
gen merged_nwgt0=tg_all_finwt0
```

```
forvalues n2=1/50 {
local x2=`n2'+50
gen merged_nwgt`n2'=tg_all_finwt0
gen merged_nwgt`x2'=tg_all_finwt`n2'
}
save h5c3.dta, replace

set trace off

*Combine the 2 cycles of data & generate survey variable flagging HINTS
iteration
use h5c2.dta, clear
append using h5c3.dta, generate(survey)
label define survey 0 "HINTS 5 CYCLE 2" 1 "HINTS 5 CYCLE 3"
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 seekhealthinfo, obs percent se
svy: tabulate chanceaskquestions, obs percent se
```

STATA Code to Merge HINTS 5, Cycle 3 and HINTS 5, Cycle 2 for the Taylor Series Linearization Method

In order to combine HINTS 5, Cycle 2 with HINTS 5, Cycle 3, 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 = 8,942).

```
*Put path and name to your HINTS 5 Cycle 2 data
use "H:\HINTS\HINTS5-Cycle2\hints5_cycle2_public.dta", clear
*Create final weight (merged_nwt0) for multi-cycle datasets
gen merged_nwgt0=person_finwt0
save h5c2.dta, replace

*Put path and name to your HINTS 5 Cycle 3 data
use "H:\HINTS\HINTS5-Cycle3\hints5_cycle3_public.dta", clear
*Create final weight (merged_nwt0) for multi-cycle datasets
gen merged_nwgt0=tg_all_finwt0
save h5c3.dta, replace

*Combine the 2 cycles of data & generate survey variable flagging HINTS
iteration
use h5c2.dta, clear
append using h5c3.dta, generate(survey)
label define survey 0 "HINTS 5 CYCLE 2" 1 "HINTS 5 CYCLE 3"
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 seekhealthinfo, obs percent se
svy: tabulate chanceaskquestions, obs percent se
```

References

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

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

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

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

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

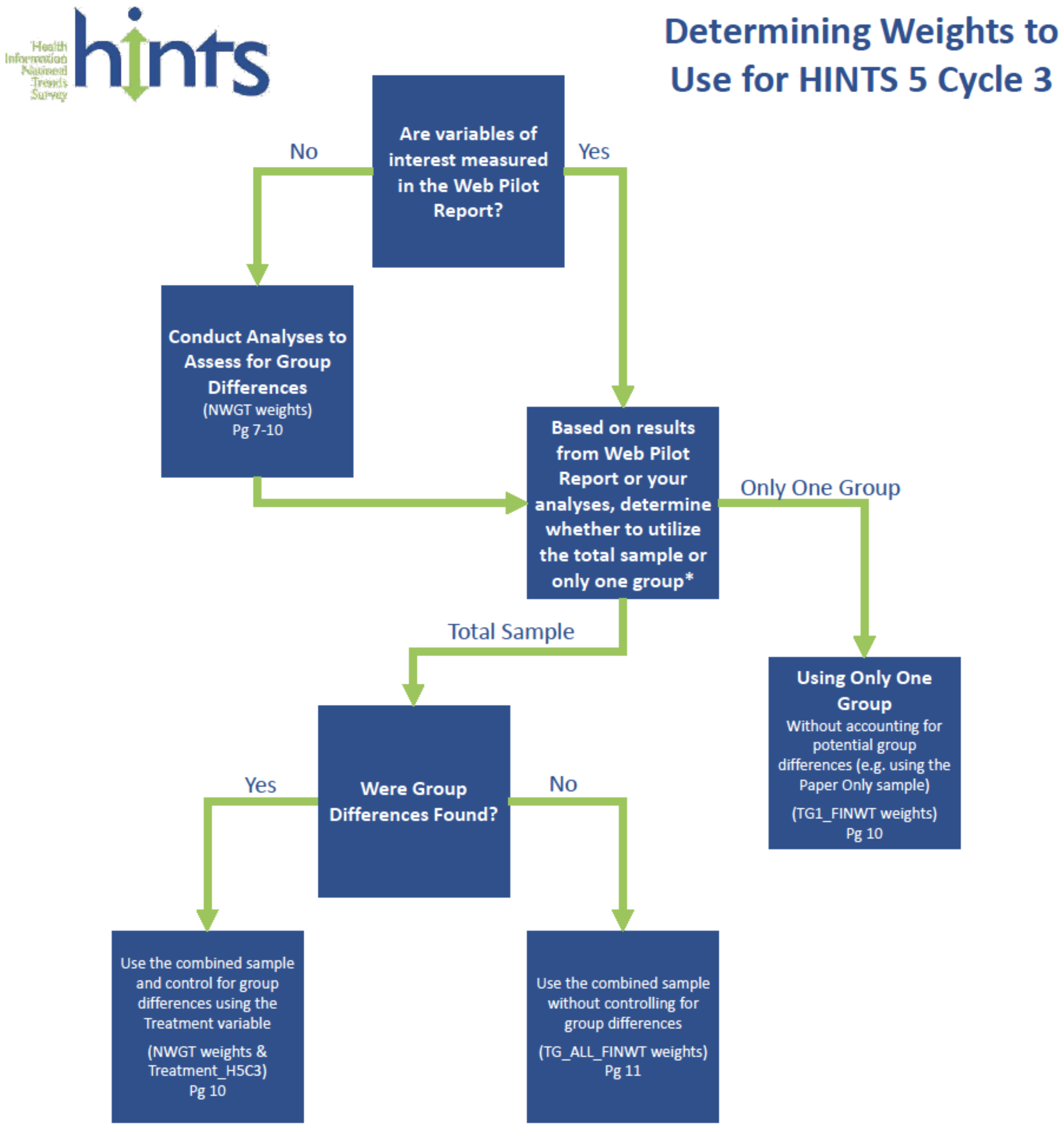
Nelson, D. E., Kreps, G. L., Hesse, B. W., Croyle, R. T., Willis, G., Arora, N. K., et al. (2004). The Health Information National Trends Survey (HINTS): development, design, and dissemination. *J Health Commun, 9*(5), 443-460; discussion 481-444.

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

Appendix A: Supplement to Determining Statistical Weights to Use for HINTS 5 Cycle 3 Analyses

Assessing for Group Differences Flow Chart


Health Information National Trends Survey
hints
Determining Weights to Use for HINTS 5 Cycle 3
Are variables of interest measured in the Web Pilot Report?
No
Yes
Conduct Analyses to Assess for Group Differences
(NWGT weights)
Pg 7-10
Based on results from Web Pilot Report or your analyses, determine whether to utilize the total sample or only one group*
Only One Group
Total Sample
Using Only One Group
Without accounting for potential group differences (e.g. using the Paper Only sample)
(TG1_FINWT weights)
Pg 10
Were Group Differences Found?
Yes
No
Use the combined sample and control for group differences using the Treatment variable
(NWGT weights & Treatment_H5C3)
Pg 10
Use the combined sample without controlling for group differences
(TG_ALL_FINWT weights)
Pg 11

*If one does find group differences then the analyst can decide whether to utilize only one group or the entire sample (recommended). If one does not find any differences by group assignment, it is recommended that analysts use the combined sample and the respective weights – the TG_ALL weight variables – to increase statistical power. Note that while it is suggested analysts use the full sample when possible, an analyst may make the decision to conduct analyses on only one group. (pg 10)

Creation of Sample Weights for the Combined Sample, Controlling for Group Differences

It is strongly recommended that analysts first assess for possible group differences between their target variables. The Web Pilot Report provides some initial analyses testing for group differences and users can refer to this report to see if their variables of interest have already been assessed. If analysts want to do their own assessment for group differences with the jackknife replication variance estimation method, use the final sample weight (**nwgt0**) and 150 replicate weights (**nwgt1** through **nwgt150**) that have been provided with the data. The SPSS data file will contain the same **nwgt0** variable that can also be used to assess for group differences using the Taylor series linearization method, along with the **var_stratum** and **var_cluster** variables

This section provides the code in SAS, SPSS, and Stata that created these sample weights that allow for assessing and controlling for group differences. For SAS and Stata, the codes were created using the Rizzo method, similar to how replicate weights are created when combining different HINTS iterations together or assessing trends over time (Rizzo, Moser, Waldron, Wang, & Davis, 2008). For SPSS, the codes only compute the nwgt0 final weight variable, as SPSS cannot incorporate replicate weights in analyses.

SAS

```
options fmtsearch=(hints5c3);

data hints5_cycle3_public (drop=i);
      set hints5_cycle3_public;

      /*Create Replicate Weights for mode tests involving Treatment*/
      **Replicate weights;
      array paperWgts [50]TG1_FINWT1-TG1_FINWT50;
      array webWgts [50]TG2_FINWT1-TG2_FINWT50;
      array webbonusWgts [50]TG3_FINWT1-TG3_FINWT50;
      array nwgt [150]nwgt1-nwgt150;

      **Adjust Final and Replicate Weights;
      If Treatment_H5C3 eq 1 then do i = 1 to 50; **paper;
            nwgt0=TG1_FINWT0;
            nwgt [i]=paperWgts [i];
            nwgt [50+i]=TG1_FINWT0;
            nwgt [100+i]=TG1_FINWT0;
      end;
      If Treatment_H5C3 eq 2 then do i = 1 to 50; **web;
            nwgt0=TG2_FINWT0;
            nwgt [i]=TG2_FINWT0;
            nwgt [50+i]=webWgts [i];
            nwgt [100+i]=TG2_FINWT0;
      end;
      If Treatment_H5C3 eq 3 then do i = 1 to 50; **WebBonus;
            nwgt0=TG3_FINWT0;
            nwgt [i]=TG3_FINWT0;
            nwgt [50+i]=TG3_FINWT0;
            nwgt [100+i]=webbonusWgts [i];
      end;
run;
```

SPSS

```
COMPUTE NWGT0=0.
```

```
IF (Treatment_H5C3=1) NWGT0=TG1_FINWT0.
IF (Treatment_H5C3=2) NWGT0=TG2_FINWT0.
IF (Treatment_H5C3=3) NWGT0=TG3_FINWT0.
EXECUTE.
```

Stata

```
set trace on

use "H:\HINTS\HINTS5-Cycle3\hints5_cycle3.dta", clear
tab treatment_h5c3

*Create final and replicate weights (nwt*) to control for group differences
*Treatment group 1 (paper-only)
keep if treatment_h5c3 == 1
generate nwgt0=tg1_finwt0
forvalues a1=1/50 {
      local b1=`a1'+50
      local c1=`a1'+100
      gen nwgt`a1'=tg1_finwt`a1'
      gen nwgt`b1'=tg1_finwt0
      gen nwgt`c1'=tg1_finwt0
}
save tg1.dta, replace

use "H:\HINTS\HINTS5-Cycle3\hints5_cycle3.dta", clear

*Create final and replicate weights (nwt*) to control for group differences
*Treatment group 2 (Web option)
keep if treatment_h5c3 == 2
generate nwgt0=tg2_finwt0
forvalues i1=1/50 {
      local j1=`i1'+50
      local k1=`i1'+100
      gen nwgt`i1'=tg2_finwt0
      gen nwgt`j1'=tg2_finwt`i1'
      gen nwgt`k1'=tg2_finwt0

}
save tg2.dta, replace

use "H:\HINTS\HINTS5-Cycle3\hints5_cycle3.dta", clear

*Create final and replicate weights (nwt*) to control for group differences
*Treatment group 3 (Web bonus)
keep if treatment_h5c3 == 3
generate nwgt0=tg3_finwt0
forvalues x1=1/50 {
      local y1=`x1'+50
      local z1=`x1'+100
      gen nwgt`x1'=tg3_finwt0
      gen nwgt`y1'=tg3_finwt0
      gen nwgt`z1'=tg3_finwt`x1'
}
save tg3.dta, replace

set trace off
```

```
*Combine treatment groups
use tg1.dta, clear
append using tg2.dta tg3.dta
```

Appendix B: Statistical Software Example Code if Group Differences are Found

If an analyst finds that group differences exist, the analyst may decide to use the entire sample and control for group assignment using the TREATMENT_H5C3 variable and using the NWGT weighting variables. This appendix provides the sample analytical code in SAS, SPSS, and Stata for this approach.

SAS

Replicate Weights Variance Estimation Method

Frequency Table and Chi-Square Test

```
proc surveyfreq data=hints5cycle3 varmethod=jackknife;
     weight NWGT0;
     repweights NWGT1-NWGT150 / df=147 jkcoefs=0.98;
     tables treatment_h5c3*edu*gender / row col wchisq;
run;
```

Logistic Regression

```
/*Multivariable logistic regression of gender and education on
SeekCancerInfo*/
proc surveylogistic data=hints5cycle3 varmethod=jackknife;
     weight NWGT0;
     repweights NWGT1-NWGT150 / df=147 jkcoefs=0.98;
     class edu (ref="Less than high school")
           gender (ref="Male")
           treatment_h5c3 (ref=first) /param=REF;
     model seekcancerinfo (descending) = treatment_h5c3 gender edu
     /tech=newton xconv=1e-8 CLPARM EXPB;
run;
```

Linear Regression

```
/*Multivariable linear regression of gender and education on
GeneralHealth*/
proc surveyreg data= hints5cycle3 varmethod=jackknife;
     weight NWGT0;
     repweights NWGT1-NWGT150 / df=147 jkcoefs=0.98;
     class edu (ref="Less than high school") gender (ref="Male")
           treatment_h5c3 (ref=first);
     model generalhealth = treatment_h5c3 edu gender/solution;
run;
```

Taylor Series Linearization Variance Estimation Method

Frequency Table and Chi-Square Test

```
proc surveyfreq data = hints5cycle3 varmethod = TAYLOR;
      strata VAR_STRATUM;
      cluster VAR_CLUSTER;
      weight NWGT0;
      tables treatment_h5c3*edu*gender / row col wchisq;
run;
```

Logistic Regression

```
/*Multivariable logistic regression of gender and education on
SeekCancerInfo*/
proc surveylogistic data=hints5cycle3 varmethod=TAYLOR;
      strata VAR_STRATUM;
      cluster VAR_CLUSTER;
      weight NWGT0;
      class edu (ref="Less than high school")
            gender (ref="Male")
           treatment_h5c3 (ref=first) /param=REF;
      model seekcancerinfo (descending) = treatment_h5c3 gender edu
      /tech=newton xconv=1e-8 CLPARM EXPB;
run;
```

Linear Regression

```
/*Multivariable linear regression of gender and education on
GeneralHealth*/
proc surveyreg data=hints5cycle3 varmethod=TAYLOR;
      strata VAR_STRATUM;
      cluster VAR_CLUSTER;
      weight NWGT0;
      class edu (ref="Less than high school") gender (ref="Male")
            treatment_h5c3 (ref=first);
      model generalhealth = treatment_h5c3 edu gender/solution;
run;
```

SPSS

Taylor Series Linearization Variance Estimation Method

Analysis Plan

```
* Analysis Preparation Wizard.
*substitute your library name in the parentheses of /PLAN FILE=.
CSPLAN ANALYSIS
 /PLAN FILE=’(sample.csaplan)'
 /PLANVARS ANALYSISWEIGHT=nwgt0
 /SRSESTIMATOR TYPE=WOR
 /PRINT PLAN
 /DESIGN STRATA=VAR_STRATUM CLUSTER=VAR_CLUSTER
 /ESTIMATOR TYPE=WR.
```

Frequency Table and Chi-Square Test

Note: The code below will produce an expected warning message of “The Test of Independence is not available because the subpopulation variable: Treatment_H5C3 is not specified as a stratification variable of the first stage.” This is an acceptable warning. Cross-tabulation code, when controlling for group difference, should only be used to determine population estimates and not for determining significant associations. Analysts should use regression models to determine significant associations while controlling for group differences.

```
* Complex Samples Crosstabs.
CSTABULATE
/PLAN FILE=”(plan filename)”
/TABLES VARIABLES= edu BY gender
/SUBPOP TABLE=Treatment_H5C3 DISPLAY=LAYERED
/CELLS POPSIZE ROWPCT COLPCT TABLEPCT
/STATISTICS SE COUNT
/TEST INDEPENDENCE
/MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

Logistic Regression

```
*Multivariable logistic regression of gender and education on SeekCancerInfo.
CSLOGISTIC  seekcancerinfo_recode (LOW) BY treatment_h5c3 flippedgender flippededu
 /PLAN FILE=’(sample.csaplan)’
 /MODEL treatment_h5c3 flippedgender flippededu
/CUSTOM  Label = 'Overall model minus intercept'
  LMATRIX =
        treatment_h5c3 -1 1/2 1/2;
 treatment_h5c3 1/2 -1  1/2;
 treatment_h5c3 1/2 1/2 -1;
        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
/CUSTOM  Label = 'Treatment group overall'
  LMATRIX =
        treatment_h5c3 -1 1/2 1/2;
 treatment_h5c3 1/2 -1  1/2;
 treatment_h5c3 1/2 1/2 -1 /INTERCEPT INCLUDE=YES SHOW=YES
 /STATISTICS PARAMETER SE CINTERVAL TTEST EXP
 /TEST TYPE=CHISQUARE PADJUST=LSD
 /ODDSRATIOS FACTOR=[flippedgender(HIGH)]
 /ODDSRATIOS FACTOR=[flippededu(HIGH)]
/ODDSRATIOS FACTOR=[treatment_h5c3(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.
```

Linear Regression

```
* Multivariable linear regression of gender and education on GeneralHealth.
CSGLM genhealth_recode BY treatment_h5c3 flippedgender flippededu
 /PLAN FILE=’(sample.csaplan)’
 /MODEL treatment_h5c3 flippededu flippedgender
 /CUSTOM  Label = 'Overall model minus intercept'
  LMATRIX =
        treatment_h5c3 -1 1/2 1/2;
 treatment_h5c3 1/2 -1  1/2;
 treatment_h5c3 1/2 1/2 -1;
        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
/CUSTOM  Label = 'Treatment group overall'
  LMATRIX =
        treatment_h5c3 -1 1/2 1/2;
 treatment_h5c3 1/2 -1  1/2;
 treatment_h5c3 1/2 1/2 -1
 /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.
```

Stata

Replicate Weights Variance Estimation Method

Declare Survey Design

```
svyset [pw=nwgt0], jkrw(nwgt1-nwgt150, multiplier(0.98)) vce(jack) dof(147)
mse
```

Cross-tabulation

```
* cross-tabulation: to obtain standard errors for total, row, and column you
must separately request each under different tabulate statements
* must also separately specify cross-tab within each treatment group
svy: tabulate edu gender if treatment_h5c3==1, cell format(%8.5f) percent se
wald noadjust
svy: tabulate edu gender if treatment_h5c3==1, row format(%8.5f) percent se
wald noadjust
```

```
svy: tabulate edu gender if treatment_h5c3==1, column format(%8.5f) percent
se wald noadjust
svy: tabulate edu gender if treatment_h5c3==2, cell format(%8.5f) percent se
wald noadjust
svy: tabulate edu gender if treatment_h5c3==2, row format(%8.5f) percent se
wald noadjust
svy: tabulate edu gender if treatment_h5c3==2, column format(%8.5f) percent
se wald noadjust
svy: tabulate edu gender if treatment_h5c3==3, cell format(%8.5f) percent se
wald noadjust
svy: tabulate edu gender if treatment_h5c3==3, row format(%8.5f) percent se
wald noadjust
svy: tabulate edu gender if treatment_h5c3==3, column format(%8.5f) percent
se wald noadjust
```

Logistic Regression

```
*   Define reference group for categorical variables for both svy: logit
and svy: regress
char gender [omit] 1
char edu [omit] 1
char treatment_h5c3 [omit] 1

*   Multivariable logistic regression of gender and education on
seekcancerinfo
xi: svy: logit seekcancerinfo i.treatment_h5c3 i.gender i.edu
test _Itreatment_2 _Itreatment_3 _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4 _cons,
nosvyadjust
test _Itreatment_2 _Itreatment_3 _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust
test _Igender_2, nosvyadjust
test _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust
test _Itreatment_2 _Itreatment_3, nosvyadjust
xi: svy, or: logit seekcancerinfo i.treatment_h5c3 i.gender i.edu
```

Linear Regression

```
*   Multivariable linear regression of gender and education on generalhealth
xi: svy: regress generalhealth i.treatment_h5c3 i.gender i.edu
test _Itreatment_2 _Itreatment_3 _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4 _cons,
nosvyadjust
test _Itreatment_2 _Itreatment_3 _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust
test _Igender_2, nosvyadjust
test _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust
test _Itreatment_2 _Itreatment_3, nosvyadjust
xi: svy, or: logit seekcancerinfo i.treatment_h5c3 i.gender i.edu
```

Taylor Series Linearization Variance Estimation Method

Declare Survey Design

```
svyset var_cluster [pw=nwgt0], strata(var_stratum)
```

Cross-tabulation

```
* cross-tabulation: to obtain standard errors for total, row, and column you
must separately request each under different tabulate statements
* must also separately specify cross-tab within each treatment group
svy: tabulate edu gender if treatment_h5c3==1, cell format(%8.5f) percent se
wald noadjust
svy: tabulate edu gender if treatment_h5c3==1, row format(%8.5f) percent se
wald noadjust
svy: tabulate edu gender if treatment_h5c3==1, column format(%8.5f) percent
se wald noadjust
svy: tabulate edu gender if treatment_h5c3==2, cell format(%8.5f) percent se
wald noadjust
svy: tabulate edu gender if treatment_h5c3==2, row format(%8.5f) percent se
wald noadjust
svy: tabulate edu gender if treatment_h5c3==2, column format(%8.5f) percent
se wald noadjust
svy: tabulate edu gender if treatment_h5c3==3, cell format(%8.5f) percent se
wald noadjust
svy: tabulate edu gender if treatment_h5c3==3, row format(%8.5f) percent se
wald noadjust
svy: tabulate edu gender if treatment_h5c3==3, column format(%8.5f) percent
se wald noadjust
```

Logistic Regression

```
*   Define reference group for categorical variables for both svy: logit
and svy: regress
char gender [omit] 1
char edu [omit] 1
char treatment_h5c3 [omit] 1

*   Multivariable logistic regression of gender and education on
seekcancerinfo
xi: svy: logit seekcancerinfo i.treatment_h5c3 i.gender i.edu
test  _Itreatment_2 _Itreatment_3 _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4 _cons,
nosvyadjust
test _Itreatment_2 _Itreatment_3 _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust
test _Igender_2, nosvyadjust
test _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust
test  _Itreatment_2 _Itreatment_3, nosvyadjust
xi: svy, or: logit seekcancerinfo i.treatment_h5c3 i.gender i.edu
```

Linear Regression

```
*   Multivariable linear regression of gender and education on generalhealth
xi: svy: regress generalhealth i.treatment_h5c3 i.gender i.edu
test _Itreatment_2 _Itreatment_3 _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4 _cons,
nosvyadjust
test _Itreatment_2 _Itreatment_3 _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust
test _Igender_2, nosvyadjust
test _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust
test _Itreatment_2 _Itreatment_3, nosvyadjust
```