
# Analytics Recommendations for HINTS 4 – Cycle 4 Data

July 2021

# Table of Contents

# Overview of HINTS

The Health Information National Trends Survey (HINTS) is a nationally-representative survey which has been administered every few years by the National Cancer Institute since 2003. The HINTS target population is all adults aged 18 or older in the civilian non-institutionalized population of the United States. The HINTS program collects data on the American public's need for, access to, and use of health-related information and health-related behaviors, perceptions and knowledge. (Hesse, et al., 2006; Nelson, et al., 2004). Previous iterations include HINTS 1 (2003), HINTS 2 (2005), HINTS 3 (2007/2008), HINTS 4 Cycle 1 (2011/2012), HINTS 4 Cycle 2 (2012/2013), and HINTS 4 Cycle 3 (Late 2013).

# HINTS 4

The HINTS 4 administration includes four mail-mode data collection cycles over four years starting in 2011. The fourth of these cycles (HINTS 4 Cycle 4) was conducted from August 2014 through November 2014 and is the focus of this report. HINTS 4 draws upon the lessons learned from prior iterations of HINTS while employing some new strategies (Link, 2005). Based on the higher response rates for the mail survey (over the RDD survey) in HINTS 3, a single-mode mail survey was implemented for all HINTS 4 cycles. For more extensive background about the HINTS program and previous data collection efforts, see Finney Rutten et al. (2012).

# Methodology

Data collection for Cycle 4 of HINTS 4 was initiated in August 2014 and concluded in November of 2014. HINTS 4 Cycle 4 was a self-administered mailed questionnaire. The sampling frame of addresses, provided by Marketing Systems Group (MSG), was grouped into three strata: 1) addresses in areas with high concentrations of minority population; 2) addresses in areas with low concentrations of minority population; and 3) addresses located in counties comprising Central Appalachia regardless of minority population. All non-vacant residential addresses in the United States present on the MSG database, including post office (P.O.) boxes, throwbacks (i.e., street addresses for which mail is redirected by the United States Postal Service to a specified P.O. box), and seasonal addresses, were subject to sampling. The protocol for mailing the questionnaires involved an initial mailing of the questionnaire, followed by a reminder postcard, and up to two additional mailings of the questionnaire as needed for non-responding households. Most households received one survey per mailing (in English), while households that were potentially Spanish- speaking received two surveys per mailing (one in English and one in Spanish). The second-stage of sampling consisted of selecting one adult within each sampled household using the Next Birthday Method. In this method, the adult who would have the next birthday in the sampled household was asked to complete the questionnaire. A $2 monetary incentive was included with the survey to encourage participation. Refer to the HINTS 4 Cycle 4 Methodology Report for more extensive information about the sampling procedures.

Two methodological experiments were embedded in Cycle 4. In the first experiment, two cover design factors (picture placement and contrast) were manipulated and compared to the original HINTS cover to examine if these factors influenced household response rate. In the second experiment approximately 15 percent of households were sent experimental questionnaires with grid items formatted so that each alternating row of the grid was shaded gray or white to determine whether the experimental grid design could help reduce item missing data rates within grid items. See the Methodology Report for more information.

## Sample Size and Response Rates

The final HINTS 4 Cycle 4 sample consists of 3,677 respondents. Note that 148 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% to 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 34.44%.

# IMPORTANT - PLEASE READ!

## July 2021 Data Update

We have discovered two errors with the weights for HINTS 4 Cycle 4. The first occurred where 5-year American Community Survey (ACS) estimates were used as the source of the population totals used in the calibration step of the weighting. The correct population should have been the 1-year ACS estimates. The 5-year estimates are based on an average of the ACS for the previous 5 years, while the 1-year estimates are based on the results of the ACS for that particular year. The HINTS estimates affected most by this error are population totals or counts (e.g., the total number of adults who have searched for information about cancer from any source). These totals will be, on average, about 2 percent lower for the 5-year estimates than if the 1-year estimates were used.

Linear and logistic regressions using the incorrect weights will be affected less than population totals because the error is in both the numerator and the denominator, which will tend to cancel the error out. Several different types of analyses, which compare results using the weights with the error and a corrected set of weights, were completed to test for differences. These involved looking at percentages, regression estimates, and trends. None of these analyses were substantively different when using the corrected weights. Virtually all resulted in percentages, regression coefficients, and significance tests that did not differ at the first decimal place.

A second, separate error with the weights was also discovered that arose from the process called raking, which is used to adjust the sampling weights for HINTS 4 Cycle 4 so the marginal values of a table sum to those known totals. These adjustment variables include demographic variables like gender and education, and it was found that the final weighted distribution had a disproportionate number of males in the higher education groups. In the interest of maintaining a distribution of education for males that is closer to the national population, new weights were created that enforced tighter controls on these two variables and have been included in this updated data release. **These new weights, released in July 2021, also correct the first error above.**

The advice to users who have completed analyses using HINTS 4 Cycle 4 but not published yet is to rerun the analyses with the correct weights found in this July 2021 data package downloaded from the HINTS website. For results that have already been published, you will want to rerun your analyses under these two scenarios:

1. If the results rely on reporting population counts or totals.
2. If a small change in the statistical significance of a result would affect your conclusions. For example, if the result is based on a result that is significant close to a 5% level (if that is the criteria used in the analysis).

**Note: In both cases, it is advised to rerun the analysis and decide if the results differ enough to merit reporting an erratum to the journal.**

If you are solely interested in calculating point estimates (means, proportions etc.), either weighted or unweighted, you can use programs including SAS, SPSS, STATA and Systat. If you plan on doing inferential statistical testing using the data (i.e., anything that involves calculating a p value or confidence interval), it is important that you utilize a statistical program that can incorporate the replicate weights that are included in the HINTS database. The issue is that the standard errors in your analyses will most likely be underestimated if you don't incorporate the jackknife replicate weights; therefore, your p-values will be smaller than they "should" be, your tests will be more liberal, and you are more likely to make a type I error. Statistical programs like SUDAAN, STATA, SAS and Wesvar can incorporate the replicate weights found in the HINTS database.

With the update of HINTS 4, Cycle 4, 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

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

**PERSON_FINWT0**: Final sample weight used to calculate population estimates. Note that estimates from the 2013 American Community Survey (ACS) of the US Census Bureau were used to calibrate the HINTS 4 Cycle 4 control totals with the following variables: Age, gender, education, marital status, race, ethnicity, and census region. In addition, variables from the 2014 National Health Interview Survey (NHIS) were used to calibrate HINTS 4 Cycle 4 data control totals regarding: Percent with health insurance and percent ever had cancer.

**PERSON_FINWT1 THROUGH PERSON_FINWT50**: Fifty replicate weights that can be used to calculate accurate standard error of estimates using the jackknife replication method. More information about how these weights were created can be found in the “HINTS 4 Cycle 4 Methodology Report” included in the data download or see Korn and Graubard (1999).

## STRATUM/CLUSTER VARIABLES FOR TAYLOR LINEARIZATION METHODS

**VAR_STRATUM:** This variable identifies the first-stage sampling stratum of a HINTS sample for a given data collection cycle. It is the variable assigned to the STRATA parameter when specifying the sample design to compute variances using the Taylor Series Linearization method. It has three values: Central Appalachia regardless of minority population (CA), high minority (HM), and low minority (LM).

**VAR_CLUSTER:** This variable identifies the cluster of sampling units of a HINTS sample for a given

data collection cycle used for estimating variances. It is the variable assigned to the CLUSTER parameter when specifying the sample design to compute variances using the Taylor Series Linearization method. It has values ranging from 1 to 50.

## OTHER VARIABLES

**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.

**TREATMENT_C4**: This variable codes for the experimental treatment groups that tested for: 1) Cover style differences; 2) Layout differences regarding shading of question grids (see Methodology report for more information).

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

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

**QDISP**: This variable codes for whether the survey returned by the respondent was considered Complete or Partial Complete. A complete questionnaire was defined as any questionnaire with at least 80% of the required questions answered in Sections A and B. A partial complete was defined as when between 50% and 79% of the questions were answered in Sections A and B. There were 148 partially complete questionnaires. Fifty-one questionnaires with fewer than 50% of the required questions answered in Sections A and B were coded as incompletely-filled out and discarded.

**INCOMERANGES_IMP:** This is the income variable (INCOMERANGES) imputed for missing data. To impute for missing items, PROC HOTDECK from the SUDAAN statistical software was used. PROC HOTDECK uses the Cox-Iannacchione Weighted Sequential Hot Deck imputation method as described by Cox (1980). The following variables were used as imputation classes given their strong association with the income variable: Education (O6), Race/Ethnicity (RaceEthn), Do you currently rent or own your house? (O15), How well do you speak English? (O9), and Were you born in the United States? (O7).

# Variance Estimation Methods: Replicate vs. Taylor Linearization

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

The HINTS 4, Cycle 4 dataset includes variance codes and replicate weights so analysts can use either Taylor Series or replication methods for variance estimation. The following points may provide some guidance regarding which method will best reflect the HINTS sample design in your analysis.

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

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

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

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

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

## Denominator Degrees of Freedom (DDF)

The HINTS 4 Cycle 4 database contains a set of 50 replicate weights to compute accurate standard errors for statistical testing procedures. These replicate weights were created using a jackknife minus one replication method; when analyzing one iteration of HINTS data, the proper denominator degrees of freedom (ddf) is 49. Thus, analysts who are only using the HINTS 4 Cycle 4 data should use 49 ddf in their statistical models. HINTS statistical analyses that involve more than one iteration of data will typically utilize a set of 50*k replicate weights, where they can be viewed as being created using a stratified jackknife method with k as the number of strata, and 49*k as the appropriate ddf. Analysts who were merging two iterations of data and making comparisons should adjust the ddf to be 98 (49*2) etc.

# 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.

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

# Appendix

The following appendices provide some coding examples using SAS, SPSS, and STATA for common types of statistical analyses using HINTS 4 Cycle 4 data. These examples will incorporate both the final sample weight (to get population estimates) and the set of 50 jackknife replicate weights to get the proper standard error, using the replication variance estimation method. The appendices also provide a coding example using SPSS, which incorporates the final sample weight and the variance codes for use with Taylor Series Linearization. Although these examples specifically use HINTS 4 Cycle 4 data, the concepts used here are generally applicable to other types of analyses. We will consider an analysis that includes gender, education level (edu) and two questions that are specific to the HINTS 4 data: seekcancerinfo & generalhealth.

- **Appendix A:** Analyzing data using SAS
- **Appendix B:** Analyzing data using SPSS
- **Appendix C:** Analyzing data using STATA

# Appendix A: Analyzing data using SAS

This section gives some SAS (Version 9.3 and higher) coding examples for common types of statistical analyses using HINTS 4 Cycle 4 data. We begin by doing data management of the HINTS 4 data in a SAS DATA step. We first decided to exclude all "Missing data (Not Ascertained)" and "Multiple responses selected in error" responses from the analyses. By setting these values to missing (.), SAS will exclude these responses from procedures where these variables are specifically accessed. For logistic regression modeling within the PROC SURVEYLOGISTIC procedure, SAS expects the response variable to be dichotomous with values (0, 1), so this variable will also be recoded at this point. It is better to use dummy variables instead of categorical variables in SAS survey procedures, such as PROC SURVEYREG. We use dummy variables for gender and education level in both PROC SURVEYLOGISTIC and PROC SURVEYREG procedures. When recoding existing variables, it is generally recommended to create new variables, rather than over-writing the existing variables. Note: New variables should always be compared to original source variables in a SAS PROC FREQ procedure to verify proper coding.

```
options fmtsearch=(LIBNAME);  *This is used to call up the formats, substitute your
library name in the parentheses;
proc format;      *First create some temporary formats;
Value Genderf
1 = "Male"
2 = "Female";

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

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

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

run;

data hints4cycle4;
set LIBNAME.H4C4_public_formatted;

/*Recode negative values to missing*/
if genderc = 1 then gender = 1;
if genderc = 2 then gender = 2;
if genderc in (-9, -6) then gender = .;
/*Recode education into four levels, and negative values to missing*/
if education in (1, 2) then edu = 1;
if education = 3 then edu = 2;
if education in (4, 5) then edu = 3;
if education in (6, 7) then edu = 4;
if education = -9 then edu = .;
```

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

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

/*Create dummy variables for proc surveylogistic and proc surveyreg procedures*/
if gender = 1 then Female = 0;
else if gender = 2 then Female = 1;

if edu = 1 then do;
HighSchool = 0;
SomeCollege = 0;
CollegeorMore = 0;
end;

else if edu = 2 then do;
HighSchool = 1;
SomeCollege = 0;
CollegeorMore = 0;
end;

else if edu = 3 then do;
HighSchool = 0;
SomeCollege = 1;
CollegeorMore = 0;
end;

else if edu = 4 then do;
HighSchool = 0;
SomeCollege = 0;
CollegeorMore = 1;
end;

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

**Proc Surveyfreq procedure**

We are now ready to begin using SAS 9.3 to examine the relationships among these variables. Using **PROC SURVEYFREQ**, we will first generate a cross-frequency table of education by gender, along with a (Wald) Chi-squared test of independence. Note the syntax of the overall sample weight, PERSON_FINWT0, and those of the jackknife replicate weights, PERSON_FINWT1—PERSON_FINWT50. The jackknife adjustment factor for each replicate weight is 0.98. This syntax is consistent for all procedures. Other data sets that incorporate replicate weight jackknife designs will follow a similar syntax.

```
proc surveyfreq data = hints4cycle4 varmethod =jackknife;
      weight person_finwt0;
      repweights person_finwt1-person_finwt50 / df = 49 jkcoefs =0.98;
      tables edu*gender / row col wchisq;
run;
```

The *tables* statement defines the frequencies that should be generated. Stand-alone 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 HINTS4-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 | HINTS4CYCLE4 |
| Number of Replicates | 50 |

Table Education by Gender

| edu | gender | Frequency | Percent | Std Err of Percent | Row Percent | Std Err of Row Percent | Column Percent | Std Err of Col Percent |
|---|---|---|---|---|---|---|---|---|
| **Less than high school** | **Male** | 102 | 4.9718 | 0.6748 | 54.0317 | 3.8061 | 10.2991 | 1.4153 |
| | **Female** | 202 | 4.2299 | 0.3579 | 45.9683 | 3.8061 | 8.1775 | 0.6894 |
| | **Total** | 304 | 9.2017 | 0.7982 | 100.0000 | | | |
| **12 years or completed high school** | **Male** | 237 | 11.0432 | 0.6939 | 48.0940 | 2.0436 | 22.8759 | 1.3646 |
| | **Female** | 424 | 11.9185 | 0.4939 | 51.9060 | 2.0436 | 23.0418 | 0.9380 |
| | **Total** | 661 | 22.9617 | 0.7669 | 100.0000 | | | |

| edu | gender | Frequency | Percent | Std Err of Percent | Row Percent | Std Err of Row Percent | Column Percent | Std Err of Col Percent |
|---|---|---|---|---|---|---|---|---|
| **Some college** | **Male** | 442 | 16.0428 | 0.5750 | 48.5947 | 1.2475 | 33.2325 | 1.1765 |
| | **Female** | 632 | 16.9707 | 0.4490 | 51.4053 | 1.2475 | 32.8090 | 0.8470 |
| | **Total** | 1074 | 33.0134 | 0.6276 | 100.0000 | | | |
| **College graduate or higher** | **Male** | 595 | 16.2166 | 0.1940 | 46.5684 | 0.3332 | 33.5926 | 0.4494 |
| | **Female** | 852 | 18.6066 | 0.1375 | 53.4316 | 0.3332 | 35.9717 | 0.2383 |
| | **Total** | 1447 | 34.8232 | 0.2489 | 100.0000 | | | |
| **Total** | **Male** | 1376 | 48.2744 | 0.2965 | | | 100.0000 | |
| | **Female** | 2110 | 51.7256 | 0.2965 | | | 100.0000 | |
| | **Total** | 3486 | 100.0000 | | | | | |

Frequency Missing = 191

| Wald Chi-Square Test | |
|---|---|
| **Chi-Square** | 21.4373 |
| | |
| **F Value** | 7.1458 |
| **Num DF** | 3 |
| **Den DF** | 49 |
| **Pr > F** | 0.0004 |
| | |
| **Adj F Value** | 6.8541 |
| **Num DF** | 3 |
| **Den DF** | 47 |
| **Pr > Adj F** | 0.0006 |

Sample Size = 3,486

The weighted percentages above show that a greater proportion of women have at least a college degree compared to men, 18.6% vs. 16.2%. The Chi-squared test of independence indicates that there is a significant difference between the educational distribution in these two groups (p-value < 0.05).

**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= hints4cycle4 varmethod=jackknife;
      weight person_finwt0;
      repweights person_finwt1-person_finwt50 / df=49 jkcoefs=0.98;
      model seekcancerinfo (descending) = Female HighSchool SomeCollege
CollegeorMore / tech=newton xconv=1e-8;
```

```
        contrast 'Overall model' intercept 1,
                          Female 1,
                          HighSchool 1,
                          SomeCollege 1,
                          CollegeorMore 1;
        contrast 'Overall model minus intercept' Female 1,
                          HighSchool 1,
                          SomeCollege 1,
                          CollegeorMore 1;
        contrast 'Gender' Female 1;
        contrast 'Education overall' HighSchool 1,
                          SomeCollege 1,
                          CollegeorMore 1;
run;
```

The response variable should be on the left-hand side (LHS) of the equal sign in the model statement, while all covariates should be listed on the right-hand side (RHS). The *descending* option requests the probability of seekcancerinfo="Yes" to be modeled. The "Male" is the reference group for gender effect while "Less than high school" is the reference group for education level effect. The option *tech=newton* requests the Newton-Raphson algorithm. The option xconv=1e-8 helps to avoid early termination of the iteration.

| Variance Estimation | |
|---|---|
| Method | Jackknife |
| Replicate Weights | HINTS4CYCLE4 |
| Number of Replicates | 50 |

| Analysis of Maximum Likelihood Estimates | | | | |
|---|---|---|---|---|
| **Parameter** | **Estimate** | **Standard Error** | **t Value** | **Pr > \|t\|** |
| **Intercept** | -0.3747 | 0.3221 | -1.16 | 0.2503 |
| **Female** | 0.1615 | 0.1296 | 1.25 | 0.2184 |
| **High School** | 0.2196 | 0.3306 | 0.66 | 0.5098 |
| **Some College** | 0.3171 | 0.3177 | 1.00 | 0.3231 |
| **College Graduate or Higher** | 0.7219 | 0.3136 | 2.30 | 0.0256 |
| **NOTE: The degrees of freedom for the t tests is 49.** | | | | |

| Odds Ratio Estimates | | | |
|---|---|---|---|
| **Effect** | **Point Estimate** | **95% Confidence Limits** | |
| **Female** | 1.175 | 0.906 | 1.525 |
| **High School** | 1.246 | 0.641 | 2.420 |
| **Some College** | 1.373 | 0.725 | 2.600 |

| Odds Ratio Estimates | | | |
|---|---|---|---|
| **Effect** | **Point Estimate** | **95% Confidence Limits** | |
| **College Graduate or Higher** | 2.058 | 1.096 | 3.866 |
| **NOTE: The degrees of freedom in computing the confidence limits is 49.** | | | |

| Contrast Test Results | | | | |
|---|---|---|---|---|
| **Contrast** | **F Value** | **Num DF** | **Den DF** | **Pr > F** |
| **Overall model** | 7.26 | 5 | 49 | <.0001 |
| **Overall model minus intercept** | 3.92 | 4 | 49 | 0.0077 |
| **Gender** | 1.55 | 1 | 49 | 0.2184 |
| **Education overall** | 4.99 | 3 | 49 | 0.0042 |

To identify levels/variables that display a significant difference in response, the rule of thumb is to examine odds ratios where the confidence interval does not contain 1 (by default, SAS will use alpha=.05 to determine statistical significance; this value can be changed by the user using code). However, significance may also be garnered from the test of whether the associated beta parameter is equal to 0 (see first regression table above). According to this model, those with college degree or more appear to be statistically more inclined to search for cancer information in comparison to those with less than a high school education.

**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= hints4cycle4 varmethod=jackknife;
      weight person_finwt0;
      repweights person_finwt1-person_finwt50 / df=49 jkcoefs=0.98;
      model generalhealth = Female HighSchool SomeCollege CollegeorMore;
      contrast 'Overall model' intercept 1,
                               Female 1,
                               HighSchool 1,
                               SomeCollege 1,
                               CollegeorMore 1;
      contrast 'Overall model minus intercept' Female 1,
                               HighSchool 1,
                               SomeCollege 1,
                               CollegeorMore 1;
      contrast 'Gender' Female 1;
      contrast 'Education overall' HighSchool 1,
                               SomeCollege 1,
                               CollegeorMore 1;
run;
```

| Variance Estimation | |
|---|---|
| Method | Jackknife |
| Replicate Weights | HINTS4CYCLE4 |
| Number of Replicates | 50 |

| Analysis of Contrasts | | | |
|---|---|---|---|
| **Contrast** | **Num DF** | **F Value** | **Pr > F** |
| **Overall model** | 5 | 3109.16 | <.0001 |
| **Overall model minus intercept** | 4 | 30.70 | <.0001 |
| **Gender** | 1 | 1.82 | 0.1841 |
| **Education overall** | 3 | 40.72 | <.0001 |

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

From the above table, we can see that Gender is not associated with general health, but Education is associated, adjusting for all variables in the model.

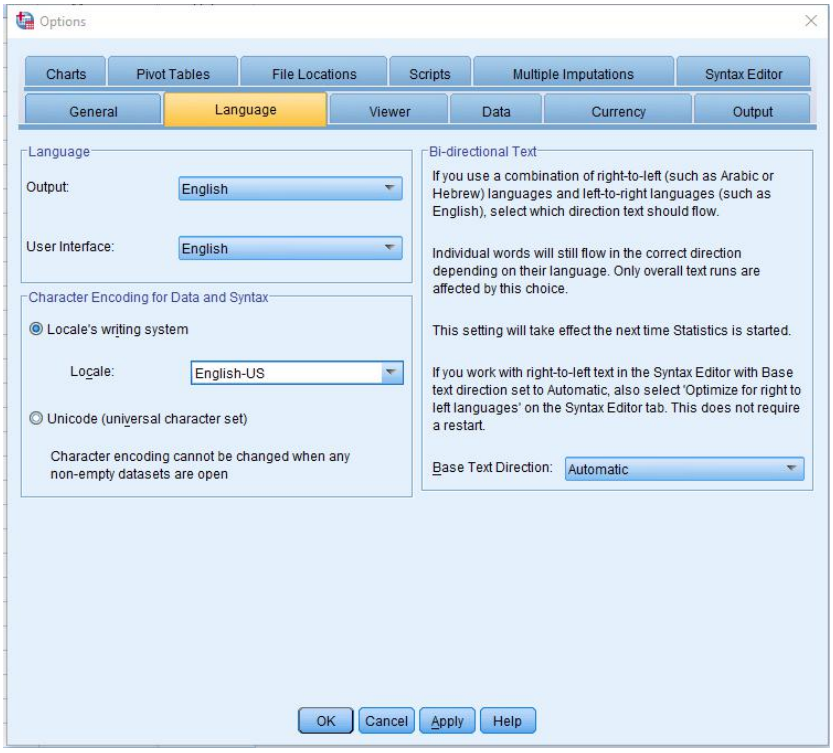
| Estimated Regression Coefficients | | | | |
|---|---|---|---|---|
| **Parameter** | **Estimate** | **Standard Error** | **t Value** | **Pr > \|t\|** |
| **Intercept** | 3.1091802 | 0.10706453 | 29.04 | <.0001 |
| **Female** | 0.0607588 | 0.04509520 | 1.35 | 0.1841 |
| **High School** | -0.4229921 | 0.11996857 | -3.53 | 0.0009 |
| **Some College** | -0.4870237 | 0.11857600 | -4.11 | 0.0002 |
| **College Graduate or Higher** | -0.8587937 | 0.11089112 | -7.74 | <.0001 |

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

From the above table, it can be seen that, compared to those respondents with Less than a High School education, those with a high school education have a significant, inverse association with general health (i.e., better reported health), controlling for all variables in the model. This association also applies to those with some college, and college graduate or higher. We don't interpret the Gender variable because it is non- significant.

# Appendix B: Analyzing data using SPSS

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




This section gives some SPSS (Version 25 and higher) coding examples for common types of statistical analyses using HINTS 4 Cycle 4 data. These examples will incorporate the stratum variable, VAR_STRATUM, and the cluster variable VAR_CLUSTER. Although these examples specifically use HINTS 4, Cycle 4 data, the concepts used here are generally applicable to other types of analyses. We will consider an analysis that includes gender, education level (edu) and two questions that are specific to the HINTS 4, Cycle 4 data: seekcancerinfo & generalhealth.

We begin by creating an analysis plan using the Complex Samples analysis procedures to specify the sample design; PERSON_FINWT0 is the sample weight variable (the final weight for the composite sample, no group differences found), VAR_STRATUM is the stratum variable, and VAR_CLUSTER is

the cluster variable. The subcommand SRSESTIMATOR specifies the variance estimator under the simple random sampling assumption. The default value is WR (with replacement), and it includes the finite population correction in the variance computation. The subcommand PRINT is used to control output from CSPLAN, and the syntax PLAN means to display a summary of plan specifications. The subcommand DESIGN with keyword STRATA identifies the sampling stratification variable, and the keyword cluster CLUSTER identifies the grouping of sampling units for variance estimation. The subcommand ESTIMATOR specifies the variance estimation method used in the analysis. The syntax TYPE=WR requires the estimation method of selection with replacement.

```
* Analysis Preparation Wizard.
*substitute your file path and library name in the parentheses of /PLAN FILE=.
CSPLAN ANALYSIS
 /PLAN FILE='(sample.csaplan)'
 /PLANVARS ANALYSISWEIGHT=PERSON_FINWT0
 /SRSESTIMATOR TYPE=WOR
 /PRINT PLAN
 /DESIGN STRATA=VAR_STRATUM CLUSTER=VAR_CLUSTER
 /ESTIMATOR TYPE=WR.
```

We completed data management of the HINTS 4 Cycle 4 data in a SPSS RECODE step. We first decided to exclude all “Missing data (Not Ascertained)” and “Multiple responses selected in error” responses from the analyses. By setting these values to missing (SYSMIS), SPSS will exclude these responses from procedures where these variables are specifically accessed. For logistic regression modeling in the CSLOGISTIC procedure, SPSS by default always uses the last (highest) level of category of the covariates as the reference, similar to SAS. Users in SPSS cannot define the reference category by themselves unless they reorder the categories to create the desired value as the reference, such as using reverse coding (see example below). To make SPSS results comparable with SAS, we reverse coded the variables in SPSS. When recoding existing variables, it is generally recommended to create new variables, rather than over-writing the existing variables. Note: New variables should always be compared to original source variables in a SPSS CROSSTABS procedure to verify proper coding.

```
RECODE GenderC (1=1) (2=2) (ELSE=SYSMIS) INTO gender.
VARIABLE LABELS gender 'gender'.
EXECUTE.

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

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

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

*Reverse coding.
RECODE gender (1=2) (2=1) (ELSE=Copy) INTO flippedgender.
```

```
VARIABLE LABELS flippedgender 'flippedgender'.
EXECUTE.

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

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

**Frequency Table and Chi-Square Test**

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

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

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

## Education by Gender

| edu | | | gender Male | Female | Total |
|---|---|---|---|---|---|
| Less than high school | Population Size | Estimate | 11548650.230 | 9825208.101 | 21373858.331 |
| | | Standard Error | 1718168.672 | 931775.393 | 1989004.289 |
| | | Unweighted Count | 102 | 202 | 304 |
| | % within edu | Estimate | 54.0% | 46.0% | 100.0% |
| | | Standard Error | 4.3% | 4.3% | 0.0% |
| | | Unweighted Count | 102 | 202 | 304 |
| | % within gender | Estimate | 10.3% | 8.2% | 9.2% |
| | | Standard Error | 1.4% | 0.7% | 0.8% |
| | | Unweighted Count | 102 | 202 | 304 |
| | % of Total | Estimate | 5.0% | 4.2% | 9.2% |
| | | Standard Error | 0.7% | 0.4% | 0.8% |
| | | Unweighted Count | 102 | 202 | 304 |
| 12 years or completed high school | Population Size | Estimate | 25651365.043 | 27684510.386 | 53335875.429 |
| | | Standard Error | 2219086.113 | 1859811.513 | 2974042.496 |
| | | Unweighted Count | 237 | 424 | 661 |
| | % within edu | Estimate | 48.1% | 51.9% | 100.0% |
| | | Standard Error | 2.7% | 2.7% | 0.0% |
| | | Unweighted Count | 237 | 424 | 661 |
| | % within gender | Estimate | 22.9% | 23.0% | 23.0% |
| | | Standard Error | 1.7% | 1.2% | 1.1% |
| | | Unweighted Count | 237 | 424 | 661 |
| | % of Total | Estimate | 11.0% | 11.9% | 23.0% |
| | | Standard Error | 0.9% | 0.7% | 1.1% |
| | | Unweighted Count | 237 | 424 | 661 |
| Some college | Population Size | Estimate | 37264486.162 | 39419820.031 | 76684306.193 |
| | | Standard Error | 3043592.389 | 2907440.113 | 4203400.666 |
| | | Unweighted Count | 442 | 632 | 1074 |
| | % within edu | Estimate | 48.6% | 51.4% | 100.0% |
| | | Standard Error | 2.8% | 2.8% | 0.0% |
| | | Unweighted Count | 442 | 632 | 1074 |
| | % within gender | Estimate | 33.2% | 32.8% | 33.0% |
| | | Standard Error | 2.1% | 1.8% | 1.5% |
| | | Unweighted Count | 442 | 632 | 1074 |
| | % of Total | Estimate | 16.0% | 17.0% | 33.0% |

| | | | | | |
|---|---|---|---|---|---|
| | | Standard Error | 1.2% | 1.2% | 1.5% |
| | | Unweighted Count | 442 | 632 | 1074 |
| College graduate or higher | Population Size | Estimate | 37668267.104 | 43219781.477 | 80888048.581 |
| | | Standard Error | 2414704.921 | 1883574.175 | 3128859.384 |
| | | Unweighted Count | 595 | 852 | 1447 |
| | % within edu | Estimate | 46.6% | 53.4% | 100.0% |
| | | Standard Error | 1.9% | 1.9% | 0.0% |
| | | Unweighted Count | 595 | 852 | 1447 |
| | % within gender | Estimate | 33.6% | 36.0% | 34.8% |
| | | Standard Error | 2.0% | 1.6% | 1.3% |
| | | Unweighted Count | 595 | 852 | 1447 |
| | % of Total | Estimate | 16.2% | 18.6% | 34.8% |
| | | Standard Error | 1.0% | 0.8% | 1.3% |
| | | Unweighted Count | 595 | 852 | 1447 |
| Total | Population Size | Estimate | 112132768.538 | 120149319.994 | 232282088.532 |
| | | Standard Error | 4778715.718 | 4332278.619 | 5929645.809 |
| | | Unweighted Count | 1376 | 2110 | 3486 |
| | % within edu | Estimate | 48.3% | 51.7% | 100.0% |
| | | Standard Error | 1.5% | 1.5% | 0.0% |
| | | Unweighted Count | 1376 | 2110 | 3486 |
| | % within gender | Estimate | 100.0% | 100.0% | 100.0% |
| | | Standard Error | 0.0% | 0.0% | 0.0% |
| | | Unweighted Count | 1376 | 2110 | 3486 |
| | % of Total | Estimate | 48.3% | 51.7% | 100.0% |
| | | Standard Error | 1.5% | 1.5% | 0.0% |
| | | Unweighted Count | 1376 | 2110 | 3486 |

**Tests of Independence**

| | | Chi-Square | Adjusted F | df1 | df2 | Sig. |
|---|---|---|---|---|---|---|
| edu * gender | Pearson | 5.731 | .848 | 2.916 | 370.334 | .466 |
| | Likelihood Ratio | 5.730 | .848 | 2.916 | 370.334 | .466 |

The adjusted F is a variant of the second-order Rao-Scott adjusted chi-square statistic. Significance is based on the adjusted F and its degrees of freedom.

The weighted percentages above show that a greater proportion of women have at least a college degree compared to men, 16.2% vs 18.6%. The Chi-squared test of independence indicates that there is not a significant difference between the educational distribution in these two groups (p-value < .05).

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

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

**Logistic Regression**

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

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

The response variable should be on the left-hand side of the BY statement, while all covariates should be listed on the right-hand side. The (LOW) option indicates that the lowest category is the reference

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

**Sample Design Information**

| | | N |
|---|---|---|
| Unweighted Cases | Valid | 2834 |
| | Invalid | 843 |
| | Total | 3677 |
| Population Size | | 187874493.094 |
| Stage 1 | Strata | 3 |
| | Units | 120 |
| Sampling Design Degrees of Freedom | | 117 |

**Parameter Estimates**

| seekcancerinfo_recode | Parameter | B | Std. Error | 95% Confidence Interval Lower | 95% Confidence Interval Upper | Hypothesis Test t | df | Sig. | Exp(B) | 95% Confidence Interval for Exp(B) Lower | 95% Confidence Interval for Exp(B) Upper |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Yes | (Intercept) | -.375 | .269 | -.908 | .158 | -1.392 | 117.000 | .166 | .687 | .403 | 1.171 |
| | Female | .162 | .114 | -.064 | .387 | 1.420 | 117.000 | .158 | 1.175 | .938 | 1.472 |
| | College Graduate or Higher | .722 | .273 | .181 | 1.262 | 2.645 | 117.000 | .009 | 2.058 | 1.199 | 3.534 |
| | Some College | .317 | .281 | -.240 | .874 | 1.128 | 117.000 | .262 | 1.373 | .787 | 2.396 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 12 Years or Completed High School | .220 | .291 | -.356 | .796 | .755 | 117.000 | .452 | 1.246 | .700 | 2.216 |

Dependent Variable: seekcancerinfo_recode (reference category = No)

Model: (Intercept), flippedgender, flippededu

a. Set to zero because this parameter is redundant.

**Odds Ratios**

| | | Odds Ratio | 95% Confidence Interval Lower | Upper |
|---|---|---|---|---|
| flippedgender | Female vs. Male | 1.175 | .938 | 1.472 |
| flippededu | College graduate or higher vs. Less than high school | 2.058 | 1.199 | 3.534 |
| | Some college vs. Less than high school | 1.373 | .787 | 2.396 |
| | 12 years or completed high school vs. Less than high school | 1.246 | .700 | 2.216 |

**Overall Model Minus Intercept**

| df | Wald Chi-Square | Sig. |
|---|---|---|
| 4.000 | 16.942 | .002 |

**Gender**

| df | Wald Chi-Square | Sig. |
|---|---|---|
| 1.000 | 2.018 | .155 |

**Education Overall**

| df | Wald Chi-Square | Sig. |
|---|---|---|
| 3.000 | 15.258 | .002 |

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, those with a college degree appear to be statistically more inclined to search for cancer information in comparison to those with less than high school education.

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 | 3394 |
| | Invalid | 283 |
| | Total | 3677 |
| Population Size | | 227789451.777 |

| | | |
|---|---|---|
| Stage 1 | Strata | 3 |
| | Units | 129 |
| Sampling Design Degrees of Freedom | | 126 |

**Parameter Estimates[a]**

| | | | 95% Confidence Interval | | Hypothesis Test | | |
|---|---|---|---|---|---|---|---|
| Parameter | Estimate | Std. Error | Lower | Upper | t | df | Sig. |
| (Intercept) | 3.109 | .098 | 2.915 | 3.303 | 31.697 | 126.000 | .000 |
| College Graduate or More | -.859 | .105 | -1.066 | -.651 | -8.192 | 126.000 | .000 |
| Some College | -.487 | .111 | -.707 | -.267 | -4.377 | 126.000 | .000 |
| 12 Years or Completed High School | -.423 | .116 | -.652 | -.194 | -3.651 | 126.000 | .000 |
| Female | .061 | .044 | -.026 | .148 | 1.382 | 126.000 | .169 |

a. Model: genhealth_recode = (Intercept) + flippededu + flippedgender

b. Set to zero because this parameter is redundant.

Compared to those respondents with less than a high school education, those who completed 12 years of school or completed high school on average reported significantly better general health (i.e., the negative beta coefficient indicates that the average health score is lower among those with some college, and the health variable is coded such that lower scores correspond to better health), controlling for all variables in the model. This association also applies to those who have completed some college and those with a college degree or higher. We do not interpret the estimates for females because the corresponding p-value is greater than .05.

**Overall Model Minus Intercept**

| df1 | df2 | Wald F | Sig. |
|---|---|---|---|
| 4.000 | 123.000 | 31.010 | .000 |

**Gender**

| df1 | df2 | Wald F | Sig. |
|---|---|---|---|
| 1.000 | 126.000 | 1.910 | .169 |

**Education**

| df1 | df2 | Wald F | Sig. |
|---|---|---|---|
| 3.000 | 124.000 | 40.661 | .000 |

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

# Appendix C: Analyzing data using STATA

This section gives some Stata (Version 10.0 and higher) coding examples for common types of statistical analyses using HINTS 4 Cycle 4 data. We begin by doing data management of the HINTS 4 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 of rather than over-writing the existing variables. Note: New variables should always be compared to original source variables in a Stata **tabulate** command to verify proper coding.

```
use "file path\hints4_cycle4_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

label define seekcancerinfo 0 "No" 1 "Yes", replace

label val seekcancerinfo seekcancerinfo



* Recode negative values to missing for svy: regress

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

**Declare survey design**

Stata requires declaring the survey design for the data set globally before any analysis. The declared survey design will be applied to all future survey commands unless another survey design is declared. Other data sets that incorporate the final sample weight and the 50 jackknife replicate weights will utilize the same code.

```
* Declare survey design for the data set

svyset [pw=person_finwt0], jkrw(person_finwt1-person_finwt50, multiplier(0.98))
vce(jack) mse
```

**Cross-tabulation**

```
* cross-tabulation

svy: tabulate edu gender, column row format(%8.5f) percent wald noadjust
```

The **svy: tabulate** command defines the frequencies that should be generated. Single variables listed in **svy: tabulate** results in one-way frequencies, while two variables will define cross-frequencies. The options **column** and **row** request column and row frequencies, respectively. The option **percent** requests the frequencies are displayed in percentage. The options **wald** and **noadjust** together request unadjusted Wald test for independence. Stata recommends default pearson test for independence. Other tests and statistics are also available; see the Stata website for more information: http://www.stata.com/

For the purposes of computing appropriate degrees of freedom for the estimator of the HINTS 4 Cycle 4 differences, we can assume as an approximation that the sample is a simple random sample of size 50 (corresponding to the 50 replicates: each replicate provides a 'pseudo sample unit') from a normal distribution. The denominator degrees of freedom (df) is equal to 49*k, where k is the number of iterations of data used in this analysis. Stata uses the number of replicates minus one as the denominator degrees of freedom and does not provide the option for user to specify the denominator degrees of freedom.

knife *: for cell counts

Number of strata = 1 Number of obs = 3486
Population size = 232282089
Replications = 50
Design df = 49

| Education | Gender | | Total |
|---|---|---|---|
| | Male | Female | |
| Less than HS | 54.03166 | 45.96834 | 1.0e+02 |
| | 10.29909 | 8.17750 | 9.20168 |

| | | | |
|---|---|---|---|
| 12 years or HS completed | 48.09402 | 51.90598 | 1.0e+02 |
| | 22.87589 | 23.04175 | 22.96168 |
| Some college | 48.59467 | 51.40533 | 1.0e+02 |
| | 33.23247 | 32.80902 | 33.01344 |
| College grad or higher | 46.56840 | 53.43160 | 1.0e+02 |
| | 33.59256 | 35.97172 | 34.82320 |
| Total | 48.27439 | 51.72561 | 1.0e+02 |
| | 1.0e+02 | 1.0e+02 | 1.0e+02 |

Key: row percentages
column percentages

Wald (Pearson):
Unadjusted chi2(3) = 21.4372
Unadjusted F(3, 49) = 7.1457 P = 0.0004
Adjusted F(3, 47) = 6.8541 P = 0.0006

The weighted percentages above show that a greater proportion of women have at least a college degree compared to men, 35.9% vs. 33.6%. The Chi-squared test of independence indicates that there is a significant difference between these the educational distribution in these two groups (p-value < 0.05).

**Logistic Regression**

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

```
* Define reference group for categorical variables for both svy: logit and
svy: regress

char gender [omit] 1

char edu [omit] 1



* Multivariable logistic regression of gender and education on seekcancerinfo

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

test _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4 _cons, nosvyadjust

test _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust

test _Igender_2, nosvyadjust
```

```
test _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust

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

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

i.gender _Igender_1-2 (naturally coded; _Igender_1 omitted)

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

(running logit on estimation sample)

Jackknife replications (50)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..... 50

Survey: Logistic regression

Number of strata = 1 Number of obs = 2834
Population size = 187874493
Replications = 50
Design df = 49
F( 4, 46) = 3.68
Prob > F = 0.0110

| seekcancer~o | Coef. | Jknife * Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| _Igender_2 | .1615276 | .1295625 | 1.25 | 0.218 | -.098838 | .4218932 |
| _Iedu_2 | .2195527 | .3306158 | 0.66 | 0.510 | -.4448447 | .8839501 |
| _Iedu_3 | .317109 | .3176806 | 1.00 | 0.323 | -.3212941 | .9555121 |
| _Iedu_4 | .7218932 | .3136296 | 2.30 | 0.026 | .0916309 | 1.352156 |
| _cons | -.374748 | .3221159 | -1.16 | 0.250 | -1.022064 | .2725681 |

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) = 7.26
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) = 3.92
Prob > F = 0.0077

Unadjusted Wald test

( 1) [seekcancerinfo]_Igender_2 = 0

F( 1, 49) = 1.55
Prob > F = 0.2184

Unadjusted Wald test

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

F( 3, 49) = 4.99
Prob > F = 0.0042

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 = 2834
Population size = 187874493
Replications = 50

Design df = 49
F( 4, 46) = 3.68
Prob > F = 0.0110

| seekcancer~o | Odds Ratio | Jknife * | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| _Igender_2 | 1.175305 | .1522754 | 1.25 | 0.218 | .9058895 | 1.524846 |
| _Iedu_2 | 1.245519 | .4117885 | 0.66 | 0.510 | .6409238 | 2.420442 |
| _Iedu_3 | 1.373152 | .4362238 | 1.00 | 0.323 | .7252099 | 2.600002 |
| _Iedu_4 | 2.058326 | .6455522 | 2.30 | 0.026 | 1.09596 | 3.865749 |

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, those with a college education or more appear to be statistically more inclined to search for cancer information compared with those who did not graduate from high school. There are no gender differences.

**Linear Regression**

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

```
* Multivariable linear regression of gender and education on

generalhealth

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

test _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4 _cons, nosvyadjust

test _Igender_2 _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust

test _Igender_2, nosvyadjust

test _Iedu_2 _Iedu_3 _Iedu_4, nosvyadjust
```

i.gender _Igender_1-2 (naturally coded; _Igender_1 omitted)

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

(running regress on estimation sample)

Jackknife replications (50)

----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5

..... 50

Survey: Linear regression

Number of strata = 1 Number of obs = 3394
Population size = 227789452
Replications = 50
Design df = 49
F( 4, 46) = 28.82
Prob > F = 0.0000
R-squared = 0.0762

| generalhea~h | Coef. | Jknife * Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| _Igender_2 | .0607588 | .0450952 | 1.35 | 0.184 | -.0298634 | .151381 |
| _Iedu_2 | -.4229921 | .1199686 | -3.53 | 0.001 | -.6640779 | -.1819062 |
| _Iedu_3 | -.4870237 | .118576 | -4.11 | 0.000 | -.7253111 | -.2487364 |
| _Iedu_4 | -.8587937 | .1108911 | -7.74 | 0.000 | -1.081638 | -.6359497 |
| _cons | 3.10918 | .1070645 | 29.04 | 0.000 | 2.894026 | 3.324334 |

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) = 3109.14
Prob > F = 0.0000

Unadjusted Wald test

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

F( 4, 49) = 30.70

Prob > F = 0.0000

Unadjusted Wald test

( 1) _Igender_2 = 0

F( 1, 49) = 1.82
Prob > F = 0.1841

Unadjusted Wald test

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

F( 3, 49) = 40.72
Prob > F = 0.0000

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