Analytics Recommendations for HINTS-FDA

June, 2016

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), HINTS 4 Cycle 3 (Late 2013), HINTS 4 Cycle 4 (2014)

HINTS-FDA

The HINTS-FDA administration was conducted from May 2015 through September 2015, and is the focus of this report. HINTS-FDA was a special round of HINTS data collection conducted by the National Cancer Institute (NCI) in partnership with the Food and Drug Administration (FDA) to combine the traditional HINTS topics of health communication, cancer knowledge, and cancer risk behaviors with an assessment of the public's knowledge of medical devices, communications related to product recalls, diet supplement labeling, risk perceptions about new tobacco products, perceptions of tobacco product harm, and tobacco product claims. HINTS-FDA draws upon the lessons learned from prior iterations of HINTS while employing some new strategies. HINTS-FDA was conducted by mail using a protocol similar to that used in HINTS 4 with a goal of obtaining 3,500 completed questionnaires. For more extensive background about the HINTS program and previous data collection efforts, see Finney Rutten et al. (2012).

Methodology

Data collection for HINTS-FDA was initiated in May 2015 and concluded in September 2015. HINTS-FDA was a self-administered mailed questionnaire. Because of the unique nature of the HINTS-FDA instrument and the specific goals of FDA, the regular sampling strategy of HINTS was altered in an effort to include more current and former smokers in the study. Using data from the Behavioral Risk Factor Surveillance System (BRFSS), county-level smoking rates were used to group addresses into sampling strata of high, medium-high, medium-low, and low smoking rates. The high and the medium-high strata were then oversampled to increase the yield of current smokers. The sampling frame consisted of a database of addresses used by Marketing Systems Group (MSG) to provide random samples of addresses. 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. A total of four mailings were sent out as part of HINTS-FDA. The mailing protocol followed a modified Dillman approach (Dillman, et. al., 2009) with a total of four mailings: an initial mailing, a reminder postcard, and two follow-up mailings. All households in the sample received the first mailing and reminder postcard, while only non-responding households received the subsequent survey mailings. Most households received one survey per mailing (in English), while households that were flagged as potentially Spanish-speaking received two surveys per mailing (one English and one Spanish). The second-stage of sampling consisted of selecting one adult within each sampled household. In keeping with HINTS 4, data collection for HINTS-FDA implemented the Next Birthday Method to select the one adult in the household. Questions were included on the survey instrument to assist the household in selecting the adult in the household having the next birthday. A $2 monetary incentive was included with the survey to encourage participation. Refer to the HINTS-FDA Methodology Report for more extensive information about the sampling procedures.

Sample Size and Response Rates

The final HINTS-FDA sample consists of 3,738 respondents. Note that 143 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 33.04%.

Analyzing HINTS Data

If you are solely interested in calculating point estimates (means, proportions etc.), either weighted or unweighted, you can use programs including SAS, SPSS, STATA and Systat. If you plan on doing inferential statistical testing using the data (i.e., anything that involves calculating a p value or confidence interval), it is important that you utilize a statistical program that can incorporate the replicate weights that are included in the HINTS database. The issue is that the standard errors in your analyses will most likely be underestimated if you 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. Currently, SPSS is not able to incorporate these replicate weights.

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-FDA 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 2014 American Community Survey (ACS) of the US Census Bureau were used to calibrate the HINTS-FDA control totals with the following variables: Age, gender, education, marital status, race, ethnicity, and census region. In addition, variables from the 2015 National Health Interview Survey (NHIS) were used to calibrate HINTS-FDA 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-FDA Methodology Report” included in the data download, or see Korn and Graubard (1999).

STRATUM: This variable codes for whether the respondent was in the Low, Medium-Low, Medium-High or High smoking rate 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 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 143 partially complete questionnaires. Sixty-three 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).

Denominator Degrees of Freedom (DDF)

The HINTS-FDA 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-FDA 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.

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

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

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

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

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, SUDAAN, and STATA for common types of statistical analyses using HINTS-FDA 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. Although these examples specifically use HINTS-FDA 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-FDA data: seekhealthinfo & friendsusetobacco.

- **Appendix A:** Analyzing data using SAS
- **Appendix B:** Analyzing data using SUDAAN
- **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-FDA data. We begin by doing data management of the HINTS-FDA 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=(hints);  *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 seekhealthinfof
1 = "Yes"
0 = "No";
run;


data hints_fda;
set hints.hints_fda_09052017_public;

/*Recode negative values to missing*/
if Selfgender = 1 then gender = 1;
if Selfgender = 2 then gender = 2;
if selfgender<0 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 <0 then edu = .;

/*Recode seekhealthinfo to 0-1 format for proc rlogist procedure, and negativ
values to missing */
if seekhealthinfo = 2 then seekhealthinfo = 0;
if seekhealthinfo<0 then seekhealthinfo = .;
```

```
/*Recode negative values to missing for proc regress procedure*/
if FriendsUseTobacco<0 then FriendsUseTobacco=.;

/*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. seekhealthinfo seekhealthinfof.;
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 = hints_fda 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	HINTS_FDA
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	101	5.1751	0.7016	47.7668	5.1824	10.5386	1.4289
	Female	125	5.659	0.6163	52.2332	5.1824	11.1192	1.2212
	Total	226	10.8341	0.7306	100			
12 years or completed high school	Male	273	10.9943	0.9144	51.7689	2.6027	22.3889	1.8765
	Female	410	10.2429	0.6818	48.2311	2.6027	20.126	1.3183
	Total	683	21.2372	1.1637	100			
Some college	Male	432	16.2467	0.9375	49.5629	1.5959	33.085	1.8515
	Female	643	16.5333	0.5169	50.4371	1.5959	32.4858	0.9951
	Total	1075	32.78	1.0987	100			
College graduate or	Male	677	16.6899	0.1884	47.4838	0.4475	33.9875	0.4163
	Female	820	18.4588	0.2745	52.5162	0.4475	36.269	0.4847

higher	Total	1497	35.1487	0.3447	100			
Total	Male	1483	49.106	0.2881			100	
	Female	1998	50.894	0.2881			100	
	Total	3481	100					

Frequency Missing = 257

Wald Chi-Square Test	
Chi-Square	16.8463
F Value	5.6154
Num DF	3
Den DF	49
Pr > F	0.0022
Adj F Value	5.3862
Num DF	3
Den DF	47
Pr > Adj F	0.0029

Sample Size = 3,481

The weighted percentages above show that a greater proportion of women have at least a college degree compared to men, 18.46% vs. 16.69%. 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 **PROC SURVEYLOGISTIC**; recall that the response should be a dichotomous 0-1 variable.

```
/*Multivariable logistic regression of gender and education on
SeekHealthInfo*/
proc surveylogistic data= hints_fda varmethod=jackknife;
weight person_finwt0;
repweights person_finwt1-person_finwt50 / df=49 jkcoefs=0.98;
model seekhealthinfo (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 seekhealthinfo="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	HINTS FDA
Number of Replicates	50

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.0896	0.238	0.1417	0.7066
Female	1	0.4313	0.1365	9.9866	0.0016
HighSchool	1	0.6944	0.229	9.1966	0.0024
SomeCollege	1	1.4254	0.2701	27.8392	<.0001
CollegeorMore	1	2.0935	0.2443	73.4401	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Female	1.539	1.178	2.011
HighSchool	2.003	1.278	3.137
SomeCollege	4.159	2.45	7.063
CollegeorMore	8.113	5.026	13.095

Contrast Test Results

Contrast	DF	Wald Chi-Square	Pr > ChiSq
Overall model	5	439.8074	<.0001
Overall model minus intercept	4	91.7227	<.0001
Gender	1	9.9866	0.0016
Education overall	3	86.4777	<.0001

To identify levels/variables that display a significant difference in response, the rule of thumb is to examine odds ratios where the confidence interval does not contain 1 (by default, SAS will use alpha=.05 to determine statistical significance; this value can be changed by the user using code). However, significance may also be garnered from the test of whether the associated beta parameter is equal to 0 (see first regression table above). According to this model, those with a high school education, some college, or a college degree or more appear to be statistically more inclined to search for health information. Females are more likely than males to search for health 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 six levels as a continuous variable (FriendsUseTobacco).Note that higher values on FriendsUseTobacco indicate the more friends of the participant using tobacco.

```
/*Multivariable linear regression of gender and education on
FriendsUseTobacco*/
proc surveyreg data= hints_fda varmethod=jackknife;
weight person_finwt0;
repweights person_finwt1-person_finwt50 / df=49 jkcoefs=0.98;
model FriendsUseTobacco = 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	HINTS_FDA
Number of Replicates	50

Analysis of Contrasts

Contrast	Num DF	F Value	Pr > F
Overall model	5	212.03	<.0001
Overall model minus intercept	4	19.57	<.0001
Gender	1	5.68	0.0211
Education overall	3	26.03	<.0001

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

From the above table, we can see that both Gender and Edu are associated with the number of friends who use tobacco, adjusting for all variables in the model.

Estimated Regression of Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1.8695126	0.21522769	8.69	<.0001
Female	-0.2300369	0.09655703	-2.38	0.0211
HighSchool	-0.3457126	0.20981474	-1.65	0.1058
SomeCollege	-0.3203905	0.21604006	-1.48	0.1445
CollegeorMore	-1.0064264	0.21229034	-4.74	<.0001

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

From the above table, it can be seen that, those with a college or higher education level have a significantly inverse linear association with number of friends who use tobacco (i.e., less friends using tobacco), controlling for all variables in the model. This association also applies to gender variable. We don't interpret those with Some College or High School education level because they are non-significant.

Appendix B: Analyzing data using SUDAAN

This section gives some SUDAAN (Version 11.0 and higher) coding examples for common types of statistical analyses using HINTS-FDA data. We begin by doing data management of the HINTS-FDA 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 RLOGIST procedure, SUDAAN 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 SAS PROC FREQ procedure to verify proper coding.

```
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 seekhealthinfof
1 = "Yes"
0 = "No";

run;


data hints_fda; /*CREATE A TEMPORARY DATA FILE FOR ANALYSIS*/
 set hints.hints_fda_09052017_public;

/*Recode negative values to missing and create new gender variable*/
if selfgender = 1 then gender = 1;
if selfgender = 2 then gender = 2;
if selfgender 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 < 0 then edu = .;

/*Recode seekhealthinfo to 0-1 format for proc rlogist procedure, and negative
values to missing */
if seekhealthinfo = 2 then seekhealthinfo = 0;
if seekhealthinfo<0 then seekhealthinfo = .;

/*Recode negative values to missing for proc regress procedure*/
if FriendsUseTobacco<0 then FriendsUseTobacco=.;
```

```
/*Apply formats to recoded variables */
format gender genderf. edu educationf. seekhealthinfo seekhealthinfof.;
run;
```

We are now ready to begin using SUDAAN to examine the relationships among these variables. Using **proc crosstab**, we will first generate a cross-frequency table of education and 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—PERSONFINWT50. 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 crosstab data= hints_fda design=jackknife ddf = 49;
weight person_finwt0;
jackwgts person_finwt1-person_finwt50 / adjjack=.98;
class gender edu;
tables edu*gender;
test chisq;
run;
```

Since this procedure is mainly for categorical variables, each variable should be specified as such by inclusion in the class statement (which is ubiquitous in all SUDAAN procedures). 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. In general, the PROC CROSSTAB procedure may be used to investigate n-way variable frequencies, along with their relationships. This is accomplished by the *test* statement, which defines various types of independence tests: here a Chi-Squared test is implemented. Other tests and statistics are also available; see the SUDAAN site link for more information.

The HINTS-FDA database for a single iteration 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. Thus, the proper denominator degrees of freedom (ddf) should be 49 when one iteration of HINTS data is being analyzed. Thus, analysts who are only using the HINTS-FDA data should use 49 ddf in their statistical models.

HINTS-FDA databases with more than one iteration of data will contain 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 these should adjust the ddf to be 98 (49*2) etc.

Variance Estimation Method: Replicate Weight Jackknife
By: EDU, GENDER

	Are you male or female?

What is the highest grade or level of schooling you completed?		Total	Male	Female
Total	**Sample Size**	3481	1483	1998
	Col Percent	100	100	100
	Row Percent	100	49.11	50.89
Less than HS	**Sample Size**	226	101	125
	Col Percent	10.83	10.54	11.12
	Row Percent	100	47.77	52.23
12 years or completed HS	**Sample Size**	683	273	410
	Col Percent	21.24	22.39	20.13
	Row Percent	100	51.77	48.23
Some college	**Sample Size**	1075	432	643
	Col Percent	32.78	33.08	32.49
	Row Percent	100	49.56	50.44
College graduate or higher	**Sample Size**	1497	677	820
	Col Percent	35.15	33.99	36.27
	Row Percent	100	47.48	52.52

Variance Estimation Method: Replicate Weight Jackknife
Chi Square Test of Independence for EDU and GENDER

ChiSq	5.6154
P-value for ChiSq	.0022
Degress of Freedom ChiSq	3

Logistic Regression

This example demonstrates a multivariable logistic regression model using **PROC RLOGIST** (*RLOGIST* is used to differentiate it from the SAS procedure, PROC LOGISTIC, and is used with SAS-callable SUDAAN); recall that the response should be a dichotomous 0-1 variable.

```
/*Multivariable logistic regression of gender and education on
Seekhealthinfo*/
proc rlogist data = hints_fda design = jackknife ddf = 49;
weight person_finwt0;
jackwgts person_finwt1-person_finwt50 / adjjack = 0.98;
class gender edu;
model seekhealthinfo = gender edu;
reflev gender=1 edu=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). Categorical variables should also be

included in the class statement. By default, the reference level of each categorical variable is that of the highest numeric level. In this example, males and those with less than a high school education were changed to be the reference values for gender and education, respectively, byusing the reflev statement to explicitly define another reference level.

Variance Estimation Method: Replicate Weight Jackknife
Working Correlations: Independent
Link Function: Logit
Response variable SEEKHEALTHINFO. Have you ever looked for information about cancer from any source?
by: Independent Variables and Effects.

Independent variables and effects	Beta Coeff.	SE Beta	T-test B=0	P-value T-Test B=0
Intercept	-.09	0.24	-.38	.7082
Gender				
Male	0.00	0.00	-	-
Female	0.43	0.14	3.16	.0027
Education Level				
Less than HS	0.00	0.00	-	-
12 years or HS completed	0.69	0.23	3.03	.0039
Some College	1.43	0.27	5.28	<.0001
College graduate or higher	2.09	0.24	8.57	<.0001

Contrast Test Results

Contrast	Degrees of Freedom	Wald F	P-value Wald Chi-Sq
Overall Model	5	87.96	<.0001
Model minus intercept	4	22.93	<.0001
Intercept	--	--	--
Gender	1	9.99	.0027
Edu	3	28.83	<.0001

Odds Ratio Estimates

Independent variables and effects	Odds Ratio	Lower 95% Limit OR	Upper 95% Limit OR
Intercept	0.91	0.57	1.48
Gender			
Male	1.00	1.00	1.00
Female	1.54	1.17	2.03
Education Level			
Less than HS	1.00	1.00	1.00

12 years or HS completed	2.00	1.26	3.17
Some College	4.16	2.42	7.16
College graduate or higher	8.11	4.97	13.26

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, SUDAAN 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 high school education, some college, or a college degree or more appear to be statistically more inclined to search for health information. Females are more inclined than males to search for health information.

Linear Regression

This example demonstrates a multivariable linear regression model using **PROC REGRESS** (REGRESS is used to differentiate it from the SAS procedure, PROC REG, and is used with SAS-callable SUDAAN); recall that the response should be a continuous variable. For the purposes of this example, we decided to use an outcome with six levels as a continuous variable (FRIENDSUSETOBACCO). Note that higher values on FRIENDSUSETOBACCO indicate the more friends of the participant using tobacco.

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). Categorical variables should also be included in the class statement. By default, the reference level of each categorical variable is that of the highest numeric level. In this example, males and those with less than a high school education were changed to be the reference values for gender and education, respectively, byusing the reflev statement to explicitly define another reference level.

```
/*Multivariable linear regression of gender and education on GeneralHealth*/
proc regress data = hints_fda design = jackknife ddf = 49;
weight person_finwt0;
jackwgts person_finwt1-person_finwt50 / adjjack = 0.98;
class gender edu;
model friendsusetobacco = gender edu ;
reflev gender=1 edu=1;
run;
```

Variance Estimation Method: Replicate Weight Jackknife
Working Correlations: Independent
Link Function: Identity
Response variable FRIENDSUSETOBACCO: C17. Of the five closest friends or acquaintances that you spend time with, how many of them use tobacco?
by: Contrast.

Contrast	Degrees of Freedom	Wald F	P-value Wald F
Overall Model	5	212.03	0.0000
Model minus intercept	4	19.57	0.0000
Intercept	--	--	--
Gender	1	5.68	.0211
Edu	3	26.03	0.0000

From the above table, we can see that both Gender and Education are associated with the outcome, adjusting for all variables in the model.

Variance Estimation Method: Replicate Weight Jackknife
Working Correlations: Independent
Link Function: Identity
Response variable FRIENDSUSETOBACCO: C17. Of the five closest friends or acquaintances that you spend time with, how many of them use tobacco?

Independent variables and effects	Beta Coeff.	SE Beta	T-test B=0	P-value T-Test B=0
Intercept	1.87	0.22	8.69	.0000
Gender				
Male	0.00	0.00	-	-
Female	-0.23	0.10	-2.38	.0211
Education Level				
Less than HS	0.00	0.00	-	-
12 years or HS completed	-0.35	0.21	-1.65	.1058
Some College	-.032	0.22	-1.48	.1445
College graduate or higher	-1.01	0.21	-4.74	.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 have a significantly inverse linear association with number of friends who use tobacco (i.e., fewer friends using tobacco), controlling for all variables in the model. This association also applies to females. We don't interpret those with Some College or High School education because they are non- significant.

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-FDA data. We begin by doing data management of the HINTS-FDA 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\hints_fda_09052017_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 seekhealthinfo to 0-1 format, and negative values to missing for
svy: logit

replace seekhealthinfo = 0 if seekhealthinfo == 2

replace seekhealthinfo = . if seekhealthinfo == -9

label define seekhealthinfo 0 "No" 1 "Yes"



* Recode negative values to missing for svy: regress

replace friendsusetobacco = . if friendsusetobacco == -5 | friendsusetobacco == -
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-FDA cycle 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.

Jknife *: for cell counts

Number of strata = 1 Number of obs = 3,481
Population size = 230,139,515
Replications = 50
Design df = 49

Education	Gender		Total
	Male	Female	
Less than HS	47.76677	52.23323	1.00E+02
	10.53862	11.11918	10.83409
12 years or HS completed	51.76894	48.23106	1.00E+02
	22.38886	20.12602	21.23721

Some college	49.56287	50.43713	1.00E+02
	33.08499	32.48576	32.78002
College graduate or higher	47.48378	52.51622	1.00E+02
	33.98753	36.26904	35.14868
Total	49.10601	50.89399	1.00E+02
	1.00E+02	1.00E+02	1.00E+02

Key: row percentage
column percentage

Wald (Pearson):
Unadjusted chi2(3) = 16.8463
Unadjusted F(3, 49) = 5.6154 P = 0.0022
Adjusted F(3, 47) = 5.3862 P = 0.0029

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 seekhealthinfo

xi: svy: logit seekhealthinfo 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 seekhealthinfo 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 = 3,460
Population size = 229,221,313
Replications = 50
Design df = 49
F(4, 46) = 21.53
Prob > F = 0.0000

seekcancer~o	Coef.	Jknife * Std. Err.	t	P>|t|	[95% Conf.	Interval]
_Igender_2	0.431311	0.136484	3.16	0.003	0.157036	0.705586
_Iedu_2	0.694412	0.2289828	3.03	0.004	0.234254	1.15457
_Iedu_3	1.425361	0.2701449	5.28	0	0.882484	1.968237
_Iedu_4	2.093467	0.2442864	8.57	0	1.602555	2.584379
_cons	-0.08959	0.2380035	-0.38	0.708	-0.56787	0.388699

Unadjusted Wald test

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

F(5, 49) = 87.96
Prob > F = 0.0000

Unadjusted Wald test

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

F(4, 49) = 22.93
Prob > F = 0.0000

Unadjusted Wald test

(1) [seekhealthinfo]_Igender_2 = 0

F(1, 49) = 9.99
Prob > F = 0.0027

Unadjusted Wald test

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

F(3, 49) = 28.83
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 = 3,460
Population size = 229,221,313
Replications = 50
Design df = 49
F(4, 46) = 21.53
Prob > F = 0.0000

seekhealthinfo	Odds Ratio	Jknife Std. Err.	t	P>|t|	[95% Conf. Interval]	
_Igender_2	1.539274	0.2100863	3.16	0.003	1.170038	2.025032
_Iedu_2	2.002531	0.4585451	3.03	0.004	1.263965	3.172658
_Iedu_3	4.159358	1.123629	5.28	0.000	2.416896	7.158047
_Iedu_4	8.112993	1.981894	8.57	0.000	4.965703	13.25505

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 high school education, some college, or a college degree or higher appear to be statistically more inclined to search for health information compared with those who did not graduate from high school, controlling for all other variables. Also, females are more statistical more inclined than makes to search for health 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 six levels as a continuous variable (friendsusetobacco). Note that higher values on FriendsUseTobacco indicate more friends (out of their 5 closest friends) of the participant use tobacco.

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

friendsusetobacco 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 = 3,432
Population size = 227,364,503
Replications = 50
Design df = 49
F(4, 46) = 18.37
Prob > F = 0.0000
R-squared = 0.0658

generalhea~h	Coef.	Jknife * Std. Err.	t	P>|t|	[95% Conf.	Interval]
_Igender_2	-0.2300369	0.096557	-2.38	0.021	-0.4240756	-0.0359983
_Iedu_2	-0.3457126	0.2098147	-1.65	0.106	-0.7673511	0.0759259
_Iedu_3	-0.3203905	0.2160401	-1.48	0.144	-0.7545392	0.1137583
_Iedu_4	-1.006426	0.2122904	-4.74	0.000	-1.43304	-0.5798129
_cons	1.869513	0.2152277	8.69	0.000	1.436996	2.302029

Unadjusted Wald test

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) = 212.03
Prob > F = 0.0000

Unadjusted Wald test

Unadjusted Wald test

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

F(4,   49) =   19.57
Prob > F =   0.0000

Unadjusted Wald test

(1) _Igender_2 = 0

F(1,   49) =   5.68
Prob > F =   0.0211

Unadjusted Wald test

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

F(3,   49) =   46.15
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 college degree or higher have a significantly negative linear association with the outcome (i.e., friends who use tobacco), controlling for all variables in the model. This association also applies to gender variable. We don't interpret those with Some College or High School education level because they are non- significant.