Paper 1426-2014

Leveraging Publicly Available Data in the Classroom Using SAS® PROC SURVEYLOGISTIC

Tyler C Smith, MS, PhD National University, San Diego, CA
Besa Smith, MPH, PhD Analydata, San Diego, CA

ABSTRACT

The soaring number of publicly available data sets across disciplines has allowed for increased access to real-life data for use in both research and educational settings. These data often leverage cost-effective complex sampling designs including stratification and clustering, which allow for increased efficiency in survey data collection and analyses. Weighting becomes a necessary component in these survey data in order to properly calculate variance estimates and arrive at sound inferences through statistical analysis. Generally speaking, these weights are included with the variables provided in the public use data, though an explanation for how and when to use these weights is often lacking. This paper presents an analysis using the California Health Interview Survey to compare weighted and non-weighted results using SAS® PROC LOGISTIC and PROC SURVEYLOGISTIC.

INTRODUCTION

Advances in data storage, processing, and management coupled with the portability of datasets and increasing information technology solutions have created a surge of new data availabilities. To that end, the soaring number of publicly available datasets across all disciplines have allowed for increased access to real-life data for use in both research and educational settings. The days of using the same data set with 100 observations and a dozen well-built variables are over! Use of these publicly available data in the classroom allows for confirmation of reports and peer-reviewed findings as well as investigation of interesting hypotheses thought up "on the fly" in an exploratory classroom setting. This also presents students with the realization that nearly three fourths of the time spent in every analysis is often focused on the data acquisition, management, and cleaning in preparation of an analytic data set.

This paper presents links and examples of easily downloadable public use datasets and describes the definition of "public use data"; advantages and disadvantages of using these types of easily accessible data; the IRB implications of using such data; examples of case studies that may be developed for classroom use in an online or onsite environment; and a contrast of using weights or not in a logistic regression.

PUBLIC USE DATA FILES

The definition of public use data files from the National Human Subjects Protection Advisory Committee (NHRPAC) is presented as data files prepared by investigators or data suppliers with the intent of making

them available for public use. The data available to the public are not individually identified or maintained in a readily identifiable form. "Publicly Available" data means that the general public can obtain the data without the use of special permissions. Sources are not considered "publicly available" if access to the data is limited to researchers or only those with signed data use agreements. Institutional Review Boards (IRB) or other entities whose main objective is the protection of human subjects in research have acknowledged that the analysis of de-identified, publicly available data does not constitute human subjects research as defined by 45 CFR 46.102. Further, they support that analyzing these data does not require IRB review unless a project merges multiple data sets and in so doing enables the identification of individuals whose data is analyzed. The reasoning behind this is that public use data files are data files that have already been reviewed under the jurisdiction of an IRB with the intent of making them de-identifiable and thus available for public use.

Public use data files have both strong advantages as well as disadvantages. Advantages are often more relevant to the classroom and disadvantages are more relevant to research. The following area describes these advantages and disadvantages pertinent to both.

ADVANTAGES OF PUBLIC USE DATA

Advantages include a readily available dataset that is often large and accumulated over many years of data collection. Use of such previously collected data can save researchers a tremendous amount of time and money. Serial cross sections can be a good source for trend data as long as methods of data ascertainment remain consistent over time. Often these data have hundreds of variables of different types providing great examples for the classroom when discussing differences in variable types and sizes. Public use data are habitually questioned regarding their generalizability towards the general population though typically these datasets include variables to weight the study population based on the inverse of the sampling scheme and inverse of the response patterns. These weights allow for better population estimates of the sampling frame and error terms in estimation. Additionally, these datasets frequently come with detailed code books (data dictionaries) and may even include sample programming code. Using these data in a classroom setting including capstone and thesis work is acceptable and without the oversight of an IRB (check with your specific IRB for guidance in how they may handle public use data) extremely efficient in compressed academic time settings.

DISADVANTAGES OF PUBLIC USE DATA

The largest disadvantage of public use datasets are due to the nature of the methodology and manner of the original collection of the data. Having previously been collected by other researchers to meet specific objectives and hypotheses relevant to their work, these aims may or may not have anything to do with the primary objectives of your work. As such, target and sample populations as well as specific measurement instruments and variables collected may differ from what is ultimately desired. Further, there may be limitations due to sample size, response rate, or assessment. Another disadvantage is that because these data are de-identified, there is no possibility to link to other types of data in the case where you would like to investigate other hypotheses with exposures that do not exist in the original de-identified data set. These limitations are typically more relevant in research and capstone or thesis efforts though there are statistical methods that may be employed to lessen the burden of these limitations. In the classroom, these data offer a real life view of the limitations and strengths of data in general and offer additional areas for development for the student. Lastly, a major disadvantage to some researchers or professors is the lack of software capable of reading in and analyzing these data. The likelihood of the reader of this paper having access to SAS would eliminate this disadvantage!

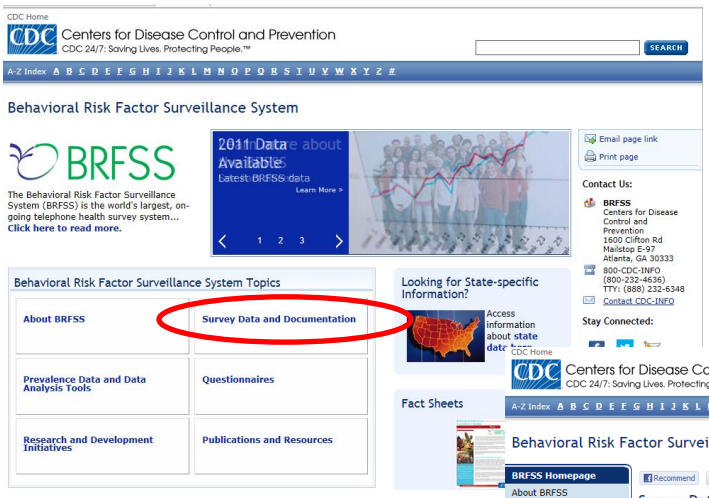
EXAMPLES OF DATA FILES AVAILABLE FOR DOWNLOAD

BEHAVIORAL RISK FACTOR SURVEILANCE SYSTEM (BRFSS)

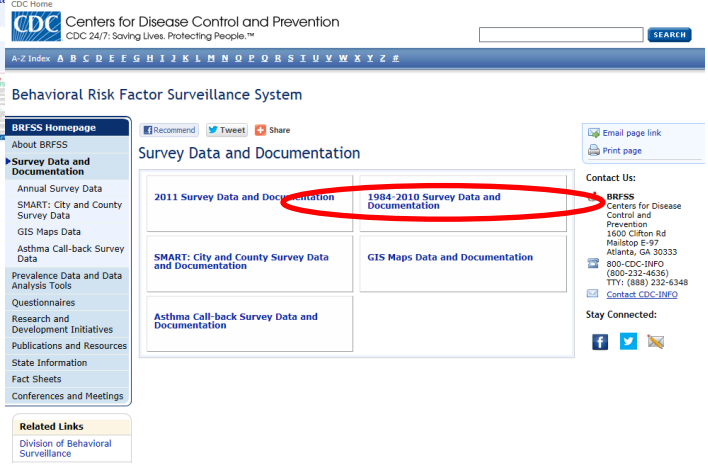
BRFSS is a health survey which evaluates behavioral risk factors and chronic diseases. It is administered by the Centers for Disease Control and Prevention and conducted by individual state health departments. The survey is the world's largest telephone survey.

- Includes computed weights
- Includes hundreds of variables
- 29 years of data available annually (1987 to 2012)
- Very large with approximately 400,000 observations per year
- Codebook and survey available
- Many peer-reviewed papers as well as reports written based on these data
- Information about limitations and strengths included

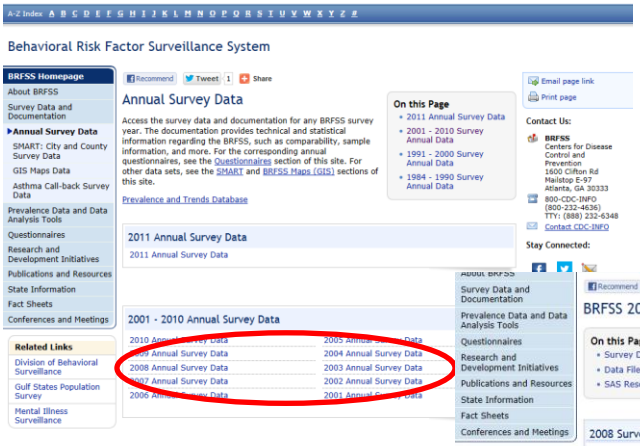
After going to the site: http://www.cdc.gov/brfss/, follow these window steps to maneuver to the data download page.



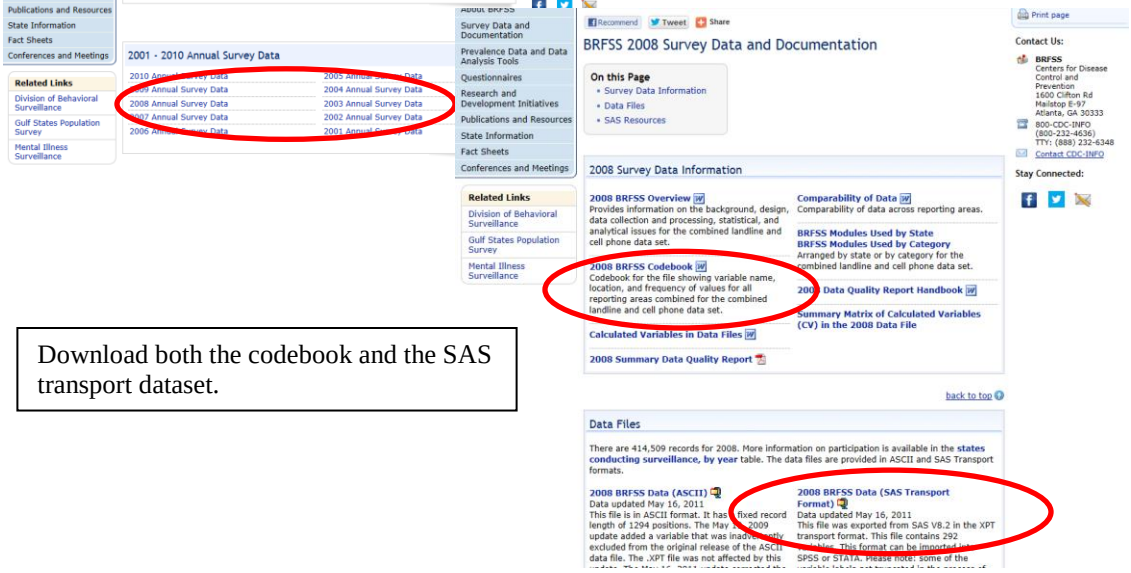
From the initial page, click "Survey Data and Documentation".



Then click on the area of survey data you are looking to maneuver to.

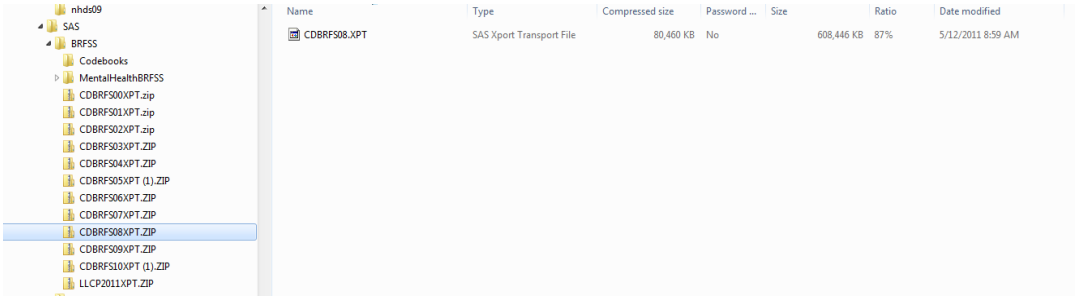


Then click on the year of survey data you are looking to maneuver to.



Download both the codebook and the SAS transport dataset.

The download will take a few minutes and you will have a “Zip” file or a .zip. Inside of the .zip file is the export file or the .XPT file



Drag and drop the .XPT file into a BRFSS directory you create on your network/computer. The transport file will be read in with the proc copy and output to the directory indicated with the libname “dataout”.

```
LIBNAME TRANSPRT XPORT 'C:\YOUR PATHWAY\BRFSS\CDBRFS08.XPT';
LIBNAME DATAOUT 'C:\ YOUR PATHWAY\BRFSS\' ;
PROC COPY IN=TRANSPRT OUT=dataout;
RUN;
```

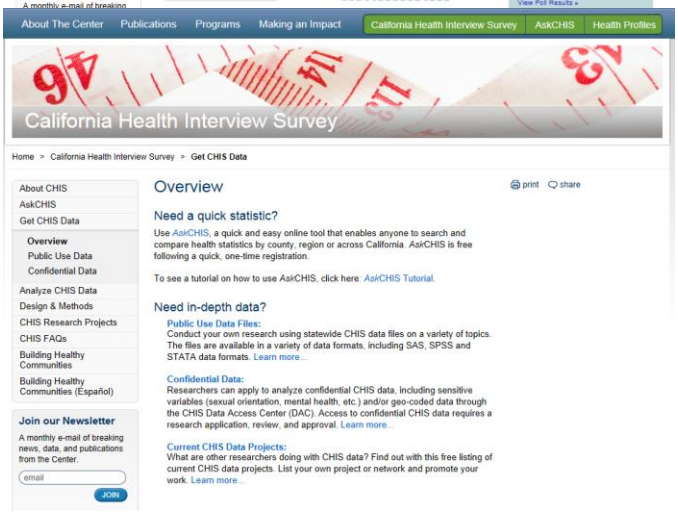
CALIFORNIA HEALTH INTERVIEW SURVEY (CHIS)

The California Health Interview Survey (CHIS) is the nation's largest state health survey with robust samples of Latinos, Asians, and American Indians.

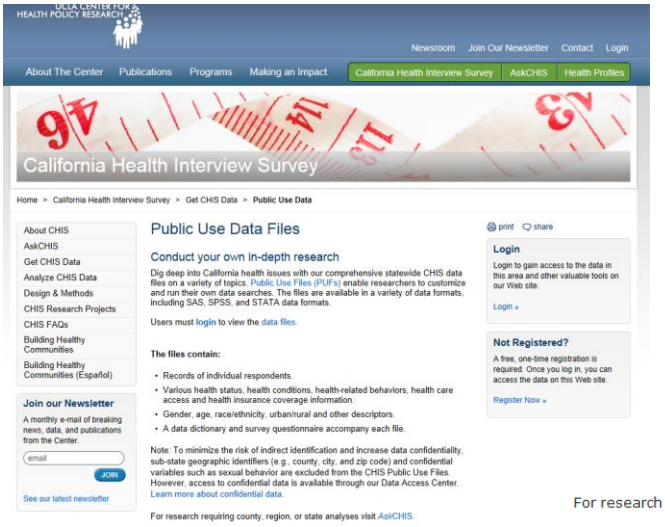
- Includes computed weights
- Includes hundreds of variables
- Serial cross-sections every two years (2001 to 2011)
- Very large with approximately 40,000 adults per year
- Also surveys adolescents and children
- Codebook and survey available
- Many peer-reviewed papers as well as reports written based on these data
- Information about limitations and strengths included

After going to the site: http://www.chis.ucla.edu/, follow these window steps to maneuver to the data download page.

From the initial page, click "Get CHIS Data".



Now click "Public Use Data".



Register and enter.

For research requiring sub-state analyses visit *Ask*CHIS.

To access the Public Use Files, select from the data years below

Click the desired year.

DOWNLOADS:

- 2009 CHIS DATA
- 2007 CHIS DATA
- 2005 CHIS DATA
- 2003 CHIS DATA
- 2001 CHIS DATA
- ORDER A CD-ROM
- CITING CHIS DATA
- METHODOLOGY REPORTS

Files

Please review the Terms of use. Select your file below.

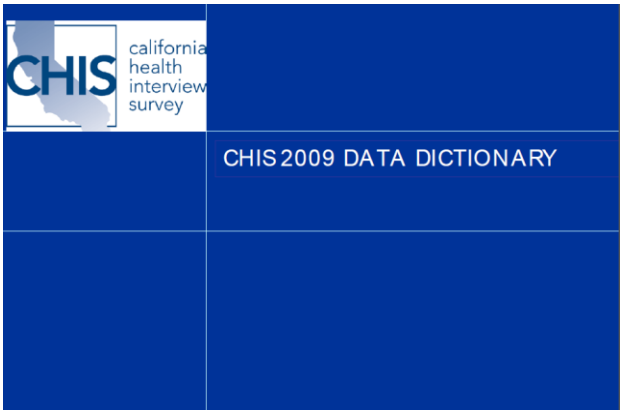
Adult, Adolescent, and Child files

File	Format	Size	
CHIS 2009 Adult *Ver. November 2012*	SAS	92 MEG	Download Now
CHIS 2009 Adult *Ver. November 2012*	SPSS	44 MEG	Download Now
CHIS 2009 Adult *Ver. November 2012*	Stata	26 MEG	Download Now
CHIS 2009 Adolescent *Ver. November 2012*	SAS	6 MEG	Download Now
CHIS 2009 Adolescent *Ver. November 2012*	SPSS	3 MEG	Download Now
CHIS 2009 Adolescent *Ver. November 2012*	Stata	3 MEG	Download Now
CHIS 2009 Child *Ver. November 2012*	SAS	14 MEG	Download Now
CHIS 2009 Child *Ver. November 2012*	SPSS	7 MEG	Download Now
CHIS 2009 Child *Ver. November 2012*	Stata	7 MEG	Download Now

rivacy Policy | Legal Notice | Data Privacy/Advisory | UCLA Center for Hea

Download the SAS data set for the desired year.

For the following example, download CHIS 2009 Adult.



Download the Data Dictionary.

CASE STUDY CLASSROOM EXAMPLE

Case Study: Gender and Flu Vaccine; CHIS 2009

You are working in a Public Health agency in California and your current efforts center around prevention of influenza among adults. You are designing a vaccine outreach program and you suspect there are differences in the acceptance of getting the flu vaccine by gender. This may impact how you target your outreach program and therefore needs to be more fully understood prior to the design phase. Using data from the California Health Interview Survey from 2009, investigate this potential.

It is expected that all students have taken and passed IRB training prior to conducting these secondary data analysis case studies. Go to this site if you have not: https://www.citiprogram.org/

From the CHIS "Public Use Data" link find 2009 data and download the SAS zipped file and the codebook.

1) What is the proportion of responders who get the flu vaccine? What is the proportion of males and proportion of females who get the flu vaccine?

   45.4% of responders reporting receiving a flu shot in the past 12 months.
   42.5% of males reported receiving a flu shot in the past 12 months.
   47.3% of females reported receiving a flu shot in the past 12 months.

2) Was there an association between flu vaccine and gender?

   An examination of the cross tab of reporting having received the flu vaccine and gender found a statistically significant difference based on a Pearson Chi-Square statistic, p-value<0.0001.

   Further analysis, including additional covariates, would be necessary to determine whether the statistically significant association was independent of other potential confounding variables.

3) Was there an association between flu vaccine and gender after adjusting for age? Please categorize age into 18-24, 25-39, 40-64, 65 or older.

After adjusting for age, women were at a statistically higher odds for reporting having received the flu vaccine in the past year when compared with men (Adjusted Odds Ratio=1.17; 95% Confidence Interval= 1.12, 1.22)

The SAS code for this short case study:

```
*************************************************************************
* FILE NAME: CHIS Adult Flu Vaccine by Gender.SAS *
*
1)    What is the proportion of responders who get the flu vaccine?
      What is the proportion of males and proportion of females who get the
      flu vaccine?

2)    Was there an association between flu vaccine and gender?

3)    Was there an association between flu vaccine and gender after adjusting
      for age? Please categorize age into 18-24, 25-39, 40-64, 65 or older.
*************************************************;

LIBNAME CHIS 'C:\YOUR PATHWAY\CHIS' ;

data tempCHIS;
  set CHIS.adult2009 (where=((ae30 in (1,2))   )); *1) Restrict your
population;

 IF SRSEX=1 THEN FEMALE=0; ELSE FEMALE=1;  *1) Recategorize or label your
variables differently;

  if 18<= srage_p <=24 then agecat=1;
  if 25<= srage_p <=39 then agecat=2;
  if 40<= srage_p <=64 then agecat=3;
  if 64<  srage_p      then agecat=4;

      flushot=0;
  if ae30=1 then flushot=1;
run;

*1)   What is the proportion of responders who get the flu vaccine?
      What is the proportion of males and proportion of females who get the
      flu vaccine?
*2)   Was there an association between flu vaccine and gender?;

proc freq data=tempCHIS;
tables  (female agecat)*flushot / chisq;
run;

*3)   Was there an association between flu vaccine and gender after adjusting
      for age? Please categorize age into 18-24, 25-39, 40-64, 65 or older.;

proc logistic data=tempCHIS;
  class   flushot (ref='0') female (ref='0') agecat (ref='1')   / param=ref;
  model flushot = female agecat  ;
title 'Multivariable Logistic Regression for Odds';
run;
```

OBJECTIVE OF THE CURRENT COMPARISON STUDY

The evolving disease burden in the US along with a growing understanding of disease comorbidities and risk factors necessitates a continuum of care that integrates all aspects of healthcare. Because psychiatric distress and impairment are likely influenced by chronic disease diagnosis and maintenance, it is important to understand the relation between these often clinically disconnected health concerns. Therefore, the objective of this analysis was to use a large cross-sectional dataset of Californians to estimate the association of psychiatric distress and impairment by evaluating the association between self-reported mental health needs and comorbid chronic diseases while controlling for known risk factors.

PROC LOGISTIC

Logistic regression is a statistical method used to evaluate many independent variables ($X_1, X_2, \ldots, X_p$) in order to predict a dichotomous outcome. Generally this outcome is denoted as Y = 1 or Y = 0 for the two possibilities.

In logistic regression the probability of an occurrence of the outcome being investigated is defined as:

$$P(Y=1) = \frac{1}{1+\exp[-\beta_0 + (\sum_{k=1}^{p} \beta_k X_k)]}$$

SAS offers several procedures to estimate the binary logit model using ML estimation which include PROC LOGISTIC, PROC GENMOD, PROC PROBIT, and PROC CATMOD. In this paper we will focus on the comparison of PROC LOGISTIC and PROC SURVEYLOGISTIC. PROC LOGISTIC is a procedure for fitting linear regression models for binary or ordinal outcomes. The following is sample code for this procedure relevant to the above described example:

```
proc logistic data=temp;
  class   Anydistorimpair (ref='0') chroniccount (ref='0') female (ref='0')
  agecat (ref='1')  currentsmoker (ref='0')  bingedrink (ref='0')
  moderatePA (ref='0') rceth (ref='1') rbmi (ref='2') / param=ref;

  model anydistorimpair = chroniccount female agecat rceth currentsmoker
  bingedrink   moderatePA rbmi / lackfit CLODDS=WALD;
  title 'Multivariable Logistic Regression CHIS Mental Health';
  run;
```

Data=temp names the input data set for the logistic regression.

Class statement allows us to establish the reference category in the categorical variables without first making "dummy" variables in a data step. In this case, we are using reference cell coding.

Param=reference requests that the parameter estimates, odds ratios, and confidence intervals be calculated using reference cell coding. The default parameter estimates would be computed using the effect coding scheme which estimates the difference in the effect of each non-reference level compared to the average effect over the other levels of the variable.

Clodds= requests for each explanatory variable, the 95% (the default alpha level because the ALPHA= option is not invoked) Wald or profile likelihood confidence intervals for the odds ratios. In this example we request the CIs based on the Wald tests

Lackfit requests the Hosmer-Lemeshow goodness of fit test for the model. The null hypothesis is that there is a good fit of the model to the observed data across the risk groups (we wish to fail to reject the null).

There are **MANY** options that are not discussed here and can be found at: https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_logistic_sect016.htm

PROC LOGISTIC output:

Number of Observations Read	39954
Number of Observations Used	39954

Response Profile

Ordered Value	Anydistorimpair	Total Frequency
1	0	36643
2	1	3311

Probability modeled is Anydistorimpair=1.

Number of observations read and number of observations used is important to check to confirm the regression is running on the numbers you expect.

Note the probability modeled is your outcome = to 1.

Class Level Information

Class	Value	Design Variables			
chroniccount	**0**	0	0	0	0
	1	1	0	0	0
	2	0	1	0	0
	3	0	0	1	0
	4	0	0	0	1
FEMALE	**0**	0			
	1	1			
agecat	**1**	0	0	0	
	2	1	0	0	
	3	0	1	0	
	4	0	0	1	
currentsmoker	**0**	0			
	1	1			

Confirm the reference category for the odds ratios are correct.

Class Level Information					
Class	**Value**	**Design Variables**			
bingedrink	**0**	0			
	1	1			
moderatePA	**0**	0			
	1	1			
rceth	**1**	0	0	0	0
	2	1	0	0	0
	3	0	1	0	0
	4	0	0	1	0
	5	0	0	0	1
RBMI	**1**	1	0	0	
	2	0	0	0	
	3	0	1	0	
	4	0	0	1	

Output not shown: Not included in this paper is the AIC (Akaikes information criterion, lower is generally better), SC (Schwartz criterion which penalizes for more parameters then the AIC, lower is generally better), and the -2 log likelihood for the model fit statistics; the likelihood ratio, score, and Wald tests for testing whether all of the parameters taken together in the fitted model are equal to 0 when compared to the model with only the intercept; significance of each variable in it's entirety (not categories of the variable) as well as the different categories.

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	**DF**	**Pr > ChiSq**
10.4038	8	0.2378

Fail to reject the null and conclude that there is a good fit of the model to the observed data across the risk groups.

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
chroniccount	4	217.6265	<.0001
FEMALE	1	88.4270	<.0001
agecat	3	500.0151	<.0001
rceth	4	8.4939	0.0751
currentsmoker	1	243.0745	<.0001
bingedrink	1	20.0002	<.0001
moderatePA	1	38.6248	<.0001
RBMI	3	12.4582	0.0060

All variables are statistically significant at the alpha=0.05 level except for race/ ethnicity. We will keep the significant variables as well as race/ethnicity to control for possible confounding.

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	**Unit**	**Estimate**	**95% Confidence Limits**	
chroniccount 1 vs 0	1.0000	1.346	1.236	1.465
chroniccount 2 vs 0	1.0000	1.962	1.742	2.210
chroniccount 3 vs 0	1.0000	2.603	2.139	3.166
chroniccount 4 vs 0	1.0000	5.659	3.680	8.704
FEMALE 1 vs 0	1.0000	1.454	1.345	1.571
agecat 2 vs 1	1.0000	0.837	0.708	0.989
agecat 3 vs 1	1.0000	0.730	0.626	0.851
agecat 4 vs 1	1.0000	0.243	0.203	0.290
rceth 2 vs 1	1.0000	0.917	0.767	1.097
rceth 3 vs 1	1.0000	0.889	0.792	0.999
rceth 4 vs 1	1.0000	0.878	0.772	1.000
rceth 5 vs 1	1.0000	1.044	0.910	1.199
currentsmoker 1 vs 0	1.0000	2.126	1.933	2.337
bingedrink 1 vs 0	1.0000	1.215	1.116	1.324
moderatePA 1 vs 0	1.0000	0.792	0.736	0.853
RBMI 1 vs 2	1.0000	1.155	0.912	1.462
RBMI 3 vs 2	1.0000	0.916	0.837	1.001
RBMI 4 vs 2	1.0000	1.077	0.977	1.189

Interpretation: After controlling for gender, age, race/ethnicity, current smoking, binge drinking, physical activity, and BMI, those in the highest category of chronic disease were at 5.66 times the odds of reporting mental health needs when compared to those without a reported chronic disease. This finding was statistically significant at the alpha=0.05 level (95% CI = 3.68, 8.70) because the confidence interval does not include 1.0. Based on the Hosmer-Lemeshow, there is a good fit of the model to the observed data across the risk groups.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	66.2	**Somers' D**	0.337
Percent Discordant	32.5	**Gamma**	0.342
Percent Tied	1.4	**Tau-a**	0.051
Pairs	121324973	**c**	0.669

Investigated the c-statistic; possible confounding. **c** is equivalent to the ROC measure. **c** ranges from 0.5 to 1, where 0.5 corresponds to the model randomly predicting the response, and a 1 corresponds to the model perfectly discriminating the response.

The C statistic is used to compare the goodness of fit of the logistic regression model (values range from 0.5 to 1.0) where a value of 0.5 indicates the model is no better than chance at making a prediction of membership in a group and a value of 1.0 indicates the model perfectly identifies those within a group and those not. Models are typically considered reasonable when the C-statistic is higher than 0.7 and strong when C exceeds 0.8 (Hosmer & Lemeshow, 2000; Hosmer & Lemeshow, 1989).

WEIGHTING

Data are often collected with complex sampling designs to ensure subgroup representation and other statistical and methodological efficiencies. There are often response differences across subgroups as well. Data should be weighted if the sample design does not give each individual an equal chance of being selected or when certain subgroups have differing probabilities of response. For example, households which have equal selection probabilities but one person is interviewed from within each household result in people from large households having a smaller chance of being interviewed. Weights are designed to lessen or eliminate the burden of sampling or response issues.

*****Sample survey data come from a finite target population and errors are not independent and identically distributed!**

This implies that: *Classical error estimation methods will give incorrect answers!*

From: http://www.chis.ucla.edu/

2.5 Weighting the Sample

To produce population estimates from the CHIS data, weights are applied to the sample data to compensate for the probability of selection and a variety of other factors, some directly resulting from the design and administration of the survey. The sample is weighted to represent the non-institutionalized population for each sampling stratum and statewide. The weighting procedures used for CHIS 2009 accomplish the following objectives:

- Compensate for differential probabilities of selection for households and persons;
- Reduce biases occurring because nonrespondents may have different characteristics than respondents;
- Adjust, to the extent possible, for undercoverage in the sampling frames and in the conduct of the survey; and
- Reduce the variance of the estimates by using auxiliary information.

PROC SURVEYLOGISTIC FOR WEIGHTED LOGISTIC REGRESSION

***Even though we are focused on the adjusted odds ratios from the logistic regression, do not forget to request cross tabs or t-tests to investigate the unadjusted associations between your independent variables and outcome of interest.

```
proc surveyfreq data = temp VARMETHOD=JACKKNIFE; *Table 2;
WEIGHT rakedw0;
REPWEIGHT rakedw1--rakedw80;
tables  (chroniccount female agecat rceth currentsmoker bingedrink
moderatePA rbmi)*Anydistorimpair / chisq;
title 'Table 2 Weighted Chisquare CHIS Mental Health';
run;
```

The following demonstrates the logistic regression model using PROC LOGISTIC followed by the logistic regression model using PROC SURVEYLOGISTIC.

```
proc logistic data=temp;
  class   Anydistorimpair (ref='0') chroniccount (ref='0') female (ref='0')
  agecat (ref='1')  currentsmoker (ref='0')  bingedrink (ref='0')
  moderatePA (ref='0') rceth (ref='1') rbmi (ref='2') / param=ref;

  model anydistorimpair = chroniccount female agecat rceth currentsmoker
  bingedrink   moderatePA rbmi / lackfit CLODDS=WALD;
  title 'Multivariable Logistic Regression CHIS Mental Health';
  run;

Proc surveylogistic data = temp VARMETHOD=JACKKNIFE;
WEIGHT rakedw0;
REPWEIGHT rakedw1--rakedw80;
  class   Anydistorimpair (ref='0') chroniccount (ref='0') female (ref='0')
  agecat (ref='1')  currentsmoker (ref='0')  bingedrink (ref='0')
  moderatePA (ref='0') rceth (ref='1') rbmi (ref='2') / param=ref;

  model anydistorimpair = chroniccount female agecat rceth currentsmoker
  bingedrink   moderatePA rbmi;
  title 'Weighted Multivariable Logistic Regression for Adjusted Odds';
run;
```

PROC SURVEYLOGISTIC output:

The SURVEYLOGISTIC Procedure

Model Information		
Data Set	WORK.TEMP	
Response Variable	Anydistorimpair	
Number of Response Levels	2	
Weight Variable	RAKEDW0	CHIS2009 RAKED WEIGHT - FULL SAMPLE
Model	Binary Logit	
Optimization Technique	Fisher's Scoring	

Number of Observations Read	39954
Number of Observations Used	39954
Sum of Weights Read	22207718
Sum of Weights Used	22207718

Number of observations read and number of observations used is the same. However, now there are weights that are being read and used also.

Note the probability modeled is the same though the total weights are now included.

Response Profile

Ordered Value	Anydistorimpair	Total Frequency	Total Weight
1	0	36643	20054897
2	1	3311	2152821

Probability modeled is Anydistorimpair=1.

The rest of the output for the procedure is presented with the same appearance though there are several differences in the results.

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
chroniccount	4	34.5029	<.0001
FEMALE	1	11.0832	0.0009
agecat	3	38.2752	<.0001
rceth	4	9.9179	0.0418
currentsmoker	1	26.2433	<.0001
bingedrink	1	2.6568	0.1031
moderatePA	1	6.0568	0.0139
RBMI	3	0.8203	0.8446

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
chroniccount 1 vs 0	1.504	1.236	1.830
chroniccount 2 vs 0	1.744	1.288	2.360
chroniccount 3 vs 0	2.514	1.374	4.599
chroniccount 4 vs 0	4.677	2.446	8.942
FEMALE 1 vs 0	1.391	1.145	1.689
agecat 2 vs 1	1.092	0.805	1.480
agecat 3 vs 1	0.837	0.627	1.117
agecat 4 vs 1	0.351	0.240	0.513
rceth 2 vs 1	1.212	0.802	1.833
rceth 3 vs 1	0.833	0.661	1.050
rceth 4 vs 1	0.836	0.573	1.220
rceth 5 vs 1	1.480	1.010	2.169
currentsmoker 1 vs 0	1.952	1.511	2.521
bingedrink 1 vs 0	1.204	0.963	1.506
moderatePA 1 vs 0	0.785	0.647	0.952
RBMI 1 vs 2	1.092	0.572	2.087
RBMI 3 vs 2	0.919	0.745	1.132
RBMI 4 vs 2	0.995	0.801	1.237

Note some of the p-values for the overall variables changed and the ORs/CIs changed.

Below are the main differences in the weighted and non-weighted results.

TABLE 3. Weighted and Non-Weighted Logistic Regression Calculated Adjusted Odds of Reporting Mental Health Needs in CHIS Adult Participants (2009).

Characteristic	Non-weighted Adjusted Odds of Participants Reporting Mental Health Needs *OR (95% CI)*		Weighted Adjusted Odds of Participants Reporting Mental Health Needs *OR (95% CI)*	
Reported Chronic Diseases				
0	1.00	--	1.00	--
1	1.35	(1.24, 1.47)	1.50	(1.24, 1.83)
2	1.96	(1.74, 2.21)	1.74	(1.29, 2.36)
3	2.60	(2.14, 3.17)	2.51	(1.37, 4.60)
4	5.66	(3.68, 8.70)	4.68	(2.45, 8.93)
Sex				
Male	1.00	--	1.00	--
Female	1.45	(1.35, 1.57)	1.39	(1.15, 1.69)
Age, years				
18 to 24	1.00	--	1.00	--
25 to 39	0.84	(0.71, 0.99)	1.09	(0.81, 1.48)
40 to 64	0.73	(0.63, 0.85)	0.84	(0.63, 1.12)
65 or older	0.24	(0.20, 0.29)	0.35	(0.24, 0.51)

TYING UP LOOSE ENDS

```
Proc surveylogistic data = temp VARMETHOD=JACKKNIFE;
WEIGHT rakedw0;
REPWEIGHT rakedw1--rakedw80;
```

CHIS included the weights in the public use data set they provided. We included them in PROC SURVEYLOGISTIC to obtain the weighted results.

From the SAS documentation: The REPWEIGHTS statement names variables that provide replicate weights for BRR or jackknife variance estimation, which you request with the VARMETHOD=BRR or VARMETHOD=JACKKNIFE option in the PROC SURVEYLOGISTIC statement. If you do not provide replicate weights for these methods by using a REPWEIGHTS statement, then the procedure constructs replicate weights for the analysis. See the sections Balanced Repeated Replication (BRR) Method and Jackknife Method for information about replicate weights.

Each REPWEIGHTS variable should contain the weights for a single replicate, and the number of replicates equals the number of REPWEIGHTS variables. The REPWEIGHTS variables must be numeric, and the variable values must be nonnegative numbers.

If you provide replicate weights with a REPWEIGHTS statement, you do not need to specify a CLUSTER or STRATA statement. If you use a REPWEIGHTS statement and do not specify the

VARMETHOD= option in the PROC SURVEYLOGISTIC statement, the procedure uses VARMETHOD=JACKKNIFE by default.

If you specify a REPWEIGHTS statement but do not include a WEIGHT statement, the procedure uses the average of each observation's replicate weights as the observation's weight.

For more information, visit the SAS Support Site:
https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_surveylogistic_sect001.htm

And

https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_surveylogistic_a0000000337.htm

SUMMARY

Public use data offers an unparalleled learning experience highlighted in real world data acquisition, cleaning, managing, and analyzing. PROC LOGISTIC has been well established in the research community for conducting regression of dichotomous or multinomial endpoints and is growing in popularity for the predictive capabilities as well. However, data sampled and presented often come with complex survey designs or response patterns that need to be addressed in the analysis. PROC SURVEYLOGISTIC offers a platform that has the same look and feel of PROC LOGISTIC though it takes into account the weights appropriately in the results.

REFERENCES

Lohr SL. Using SAS® for the Design, Analysis, and Visualization of Complex Surveys. Paper 343-2012, SAS Global Forum 2012.

Berglund PA. Enhanced Data Analysis using SAS® ODS Graphics and Statistical Graphics. Paper 343-2012, SAS Global Forum 2012

Lewis T. Considerations and Techniques for Analyzing Domains of Complex Survey Data. Paper 449-2013, SAS Global Forum 2013.

Cassell D. Wait Wait, Don't Tell Me… You're Using the Wrong Proc! Paper 193-31, SUGI 31.

ACKNOWLEDGMENTS

ABOUT THE AUTHORS AND CONTACT INFORMATION

Dr. Tyler Smith is professor of biostatistics, epidemiology, public health and health informatics; and program lead for the Health Analytics master's degree. Dr. Smith received a BS in mathematics/statistics from California State University, Chico; MS in statistics from the University of Kentucky; and PhD in epidemiology from the University of California, San Diego. With ~20 years of experience in health research leading large longitudinal studies, infant health registries, and medical health outcomes research, he has 120 peer-reviewed publications in scientific journals, >250 scientific presentations and has been PI/COI on grants totaling >$20,000,000. Currently Dr. Smith serves the SAS community through his efforts as Content Area Lead for SAS Global Forum 2014; 2015 SAS Global Forum Conference Chair; Junior Professional Award co-Chair for Western User's of SAS Software, and as part of the Executive Board for the San Diego SAS User's Group.

Tyler C Smith, MS, PhD
Associate Professor and Chair
Program Lead MS Health and Life Science Analytics
Director Health Science Research Center
Department of Community Health
School of Health and Human Services
National University
San Diego, CA 92123
tsmith@nu.edu

Dr. Besa Smith has worked in government, academic, and private industries and has served as a senior epidemiologist, senior biostatistician, and head of analytics for a 35-40 member multi-disciplinary research team. She is currently a senior scientist and founder of the health analytics consulting business, Analydata. Additionally, Dr. Smith has joint appointments with National University and the University of California, San Diego. She is an adjunct professor in the Department of Community Health in the School of Health and Human Services at NU and an assistant adjunct professor in the Department of Family and Preventive Medicine in the School of Medicine at UCSD. She teaches epidemiology and biostatistics courses to undergraduate, graduate, and medical students. Dr. Smith has a BS in biology; MPH in biometry, and PhD in epidemiology. With over 15 years leveraging health analytics in longitudinal studies and medical health outcomes research, she has >70 peer-reviewed publications in scientific journals and >100 scientific presentations.

Besa Smith, MPH, PhD
Epidemiologist and Biostatistician
Analydata
San Diego, CA 92107
besasmith@analydata.com