**Paper 854-2017**

# Using Proc Logistic and Weighting of Public Use Data in the Classroom

Tyler C Smith, MS, PhD and Besa Smith, MPH, PhD, San Diego, CA

## ABSTRACT

The rapidly evolving informatics capabilities of the past two decades have resulted in amazing new data-based opportunities. Large public use datasets are now available for easy download and utilization in the classroom. Days of classroom exercises based on static, clean, easily maneuverable samples of 100 or less are over. Instead we have large and messy "real-world data" at our finger tips allowing for educational opportunities not available in years past. There are now hundreds of public use data sets available for download and analysis in the classroom. Many of these sources are survey-based and require the understanding of weighting techniques. These techniques are necessary for proper variance estimation allowing for sound inferences through statistical analysis. This example uses the California Health Interview Survey to present and compare weighted and non-weighted results using SAS® PROC SURVEYLOGISTIC.

## INTRODUCTION

Data storage and accessibility advances offer unique opportunities in the classroom when utilizing data. This paper presents links and examples of easily downloadable public use datasets and describes the definition of "public use data" and advantages and disadvantages of using these types of easily accessible data. IRB implications of using such data are discussed and examples of case studies that may be developed for classroom use in an online or onsite environment are presented. Lastly, a contrast of using weighting or non-weighting in a logistic regression is presented.

## PUBLIC USE DATA FILES

Public use data files are data files prepared by investigators or data suppliers with the intent of making them available for public use. "Publicly Available" data means that the general public can obtain the data without the use of special permissions and that these data are not individually identified or maintained in a readily identifiable form. 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 do not constitute human subjects research as defined by 45 CFR 46.102. Further, they support that analyzing these data do not require IRB review unless a project merges multiple data sets and in so doing enables the identification of individuals whose data are analyzed.

## ADVANTAGES and DISADVANTAGES OF PUBLIC USE DATA

Advantages include a readily available dataset that is often large with some that have accumulated over many years of data collection. Use of such previously collected data can save researchers a tremendous amount of time and money. 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 on how they may handle public use data) is extremely efficient in compressed academic time settings.

One of the primary disadvantages of public use datasets is the methodology and manner of the original collection of the data. Data having previously been collected to meet prior specific objectives may not be relevant to the current objectives of your work. As such, target and sample populations as well as specific measurement instruments and variables collected may differ from what is needed. Further, there may be limitations due to sample size, response rate, or assessment method. 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.

## POTENTIAL DATA FILES AVAILABLE FOR DOWNLOAD

There are many repositories that are being created to house data sets as well as portals that have been created to help find data.

A National University portal: http://nu.libguides.com/c.php?g=492273&p=3367564

https://www.kaggle.com/datasets

https://aws.amazon.com/public-datasets/

https://www.data.gov/education

http://www.data-planet.com/

https://www.assetmacro.com/market-data/

http://www.datasets.co/

https://nssdc.gsfc.nasa.gov/

https://www.census.gov/

## EXAMPLES OF PUBLIC USE DATA

The following two examples will describe how to download two different public use data sets and how to read them into SAS. The first example will cover the CDC's Behavioral Risk Factor Surveillance System and the second will describe and illustrate how to download the California Health Interview Survey.

## EXAMPLE 1: 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
- Available annually (1987 to 2015)
- 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/, download the SAS .xpt file (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;
```

## EXAMPLE 2: 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 2015)
- 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 simple steps to download a SAS data set from the site.

**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 ($X_1$, $X_2$, …, $X_p$) 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

Adjusted, Non-weighted Odds Output:

**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.

## 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/](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;
```

**PROC SURVEYLOGISTIC output:**

**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.

The main differences in the weighted and non-weighted results appear in the following table. Though of limited statistical difference (significance is maintained) in number of reported chronic diseases, there is notable differences in magnitude of the measure of effect. With a difference of 5.66 non-weighted versus a 4.68 weighted for 4 chronic diseases and reporting mental health. There were more meaningful differences between weighted and non-weighted in age years with some of them changing direction and others losing significance.

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) |

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

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

## SUMMARY

Public use data offers many opportunities to enhance the learning experience with 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 ~140 peer-reviewed publications in scientific journals, ~300 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 as appointment at the University of California, San Diego where she is an associate 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. She is currently Regional Director of Medical Affairs Statistical Research at ICON clinical research organization. Dr. Smith has a BS in biology; MPH in biometry, and PhD in epidemiology. With nearly 20 years leveraging health analytics in longitudinal studies and medical health outcomes research, she has ~80 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