Paper 802-2017

The Truth Is Out There: Leveraging Census Data Using PROC SURVEYLOGISTIC

Richard Dirmyer, National Technical Institute for the Deaf, Rochester Institute of Technology

ABSTRACT

The advent of robust and thorough data collection has resulted in the term “big data.” With census data become richer, more nationally representative, and voluminous, we need methodologies that are designed to handle the manifold survey designs that Census data sets implement. The relatively nascent PROC SURVEYLOGISTIC, an experimental procedure in SAS® 9 and fully supported in SAS 9.1, addresses some of these methodologies, including clusters, strata, and replicate weights. PROC SURVEYLOGISTIC handles data that are not a straightforward random sample. Using a Census data set, this paper provides examples highlighting the appropriate use of survey weights to calculate various estimates, as well as the calculation and interpretation of odds ratios between categorical variable interactions when predicting a binary outcome.

INTRODUCTION

Survey data, both cross-sectional and longitudinal, have become widely accepted resources in the effort to estimate various characteristics of populations, both descriptively and inferentially. Categorical response, although in this paper’s case *binary*, has become of particular interest in the social sciences and policy-oriented disciplines, including the observation and prediction of general participation, say, in a program, or the labor force. Blinder (1983) began the long-standing conversation surrounding how to handle survey data and their respective underlying designs using logistic regression techniques, and the criticality of valid estimation within the technique, especially upon extrapolation and generalization to a greater population.

The SURVEYLOGISTIC procedure, fully supported in SAS 9.1, generally provides users the ability to model the relationship between a discrete response variable and set of independent variables, through maximum likelihood estimation, and includes various options for linking the mean response to the independent variables, including the default LOGIT, and CLOGLOG, GLOGIT, PROBIT link functions (An, 2002). This paper will demonstrate how to handle American Community Survey data, and its accompanying weights, for appropriate estimation, and additionally demonstrate how to prompt PROC SURVEYLOGISTIC for odds ratios of categorical variable interactions.

AMERICAN COMMUNITY SURVEY DATA

The American Community Survey (ACS) is administered by the U.S. Census Bureau, on a monthly basis to approximately 295,000 households, resulting in approximately 3.5 million records, annually. The utilities of the ACS are manifold, and perhaps most evident by the various constituents that use the data, inclusive of Federal, State, and Local agencies, non-government agencies such as small and large businesses to aid in market research, and various others to situate contextual arguments relying on population data.

Recipients of the ACS are legally obligated to respond, under Title 18 U.S.C. Section 3571 and Section 3559. ACS data includes, broadly speaking, both person-level and household-level weights, in addition to a replicate weighting structure, both of which allow appropriate estimation of the more than 320 million U.S. residents, and more than 180 million unique households.

ACS weights accompany the public-use microdata under the Successive Differences Replication (SDR) method (American Community Survey Design and Methodology, 2014; Wolter, 1984; Fay & Train, 1995; Judkins, 1990). With respect to persons, an overall person-level weight along with 80 replicate weights are provided. For any estimate *X*, 80 replicate estimates are computed based on the aforementioned replicate weights. If *X* can be considered as representing the full sample estimate, a consideration of the standard error of *X* as the sum of the squared differences, between each replicate X_r and the full sample *X,* is illustrated by Figure 1*:*

$$SE(X) = \sqrt{C_r \sum_{r=1}^{80} (X_r - X)^2}$$

Figure 1. Standard Error of full sample *X*

Where C_r is a multiplier related to the rth replicate. Under the SDR method, the value of C_r is ${}^{4}/_{R}$ where *R* represents the number of replicates within the data set (Fay & Train, 1995).

Aside from the design method of the ACS, it is important to check for any replicate weight that is negative, and whether the procedure being employed can handle negative weights. The following code can be used, if necessary, to zero-out all negative weights in the data.

```
ARRAY zeroing(*) pwgtp1-pwgtp80;

DO i = 1 TO dim(zeroing);
   IF zeroing (i) < 0 THEN zeroing (i)=0;
```

PROC SURVEYLOGISTIC IN ACTION

The following example will illustrate, in its most basic form, the use of the SURVEYLOGISTIC procedure to perform an analysis on employment, or lack thereof, using American Community Survey (ACS) data. While the variables influencing the likelihood of employment are manifold, this paper considers only a small set of categorical variables representing the highest level of education achieved (*hdegree*) and disability status (*disstat*) as the independent variables, all the while making no claims about the representativeness of this relationship.

In consideration of appropriate variance estimation, and using the formula illustrated in Figure 1, we now observe C_r = ${}^{4}/_{80} = 0.05$. While the calculation of the aforementioned formula remains trivial, we are unable to specify the weights as connected to the Successive Difference Replication (SDR) method in SAS. Keathley, Navarro, and Asiala (2010), as one example, highlight the similarities in variance estimation between the SDR and Jackknife methods using ACS data, leaving us to reason that Jackknife variance estimation is the easiest and most appropriate way to handle ACS replicate weights. The default variance estimation used by SAS when the REPWEIGHTS statement is specified, and replicate weights provided, is the Jackknife method. The use of the VARMETHOD option, then, serves only as an illustration of completeness between explicitly declaring the variance estimation method, along with the appropriate multiplier JKCOEFS. Otherwise, the VARMETHOD option does not need to be specified in the code, below.

```
PROC SURVEYLOGISTIC DATA=work.acs2015 VARMETHOD=jackknife;
   WEIGHT pwgtp;
   REPWEIGHTS pwgtp1—pwgtp80 / JKCOEFS=0.05;
   CLASS hdegree (REF='Less than High School') disstat (REF='No');
   MODEL employed (EVENT='Employed') = hdegree disstat;
RUN;
```

Since the ACS data rely on the SDR method, it is important we specify the correct multiplier of 0.05 in the JKCOEFS option, generally expressed as C_r.

OBTAINING ODDS RATIOS FOR CATEGORICAL VARIABLE INTERACTIONS

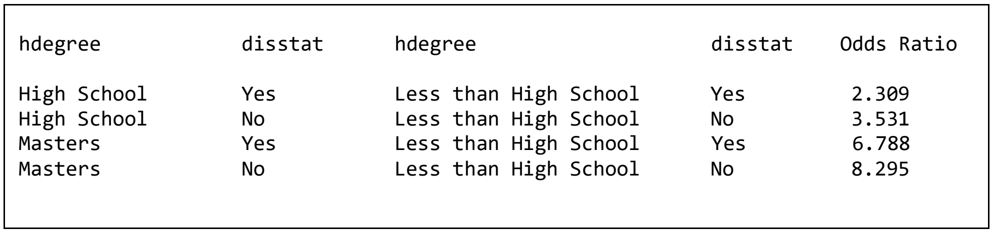
As will be demonstrated, between-group comparisons are perhaps best addressed through odds ratios within logistic regression models, the odds themselves expressed as $\frac{p_i}{1-p_i}$, where p_i represents the probability of a given event within variable *i*. As a default setting, SAS produces odds ratios within the SURVEYLOGISTIC procedure, although only for those independent variables not otherwise included as part of a linear or higher-ordered interaction term. SAS Support uses the following language in the affected PROCS: “Note that when a variable is involved in an interaction there isn’t a single odds ratio estimate for it. Rather, the odds ratio for the variable depends on the level(s) of the interacting variable(s)” to explain the nuance of odds ratios for the terms of interest, in this case an interaction between *hdegree* and *disstat*, to explore any moderating effect of educational human capital on disability status.

INSTRUCTING SAS TO PRODUCE THESE ODDS RATIOS

Specifying the PARAM option in the CLASS statement is necessary, as setting the option equal to GLM will allow the invocation of the LS MEANS statement (Agnelli, 2014). The LS MEANS statement, if including in the SURVYELOGISTIC procedure, will produce the odds ratios for linear or higher-ordered interaction terms that are specified as an interest. Additionally, the ODDSRATIO and DIFF options must also be specified within the LS MEANS statement.

```
PROC SURVEYLOGISTIC DATA=work.acs2015 VARMETHOD=jackknife;
   WEIGHT pwgtp;
   REPWEIGHTS pwgtp1--pwgtp80 / JKCOEFS=0.05;
   CLASS  hdegree (REF='Less than HS') disstat (REF='No') / PARAM=glm;
   MODEL employed (EVENT='Employed') = hdegree disstat hdegree*disstat;
   LSMEANS hdegree*disstat / ODDSRATIO DIFF;
RUN;
```

Generally, the ODDSRATIO and DIFF options will yield the odds ratios, for every possible pair of classes within each categorical variable. However, depending on the number of classes existing within a given independent variable, or explicit hypotheses targeting a specific between-group comparison, the SLICE statement is a worthwhile consideration (Agnelli, 2014), but otherwise is not discussed in this. Excerpted output from the PROC SURVEYLOGISTIC, above, is pasted below in Table 1. In full disclosure, data are representative of the 2015 American Community Survey, specifically those working-aged individuals (ages 15-64) who are participating in the labor force as either unemployed, or employed.



```
hdegree            disstat      hdegree                  disstat     Odds Ratio

High School        Yes          Less than High School    Yes          2.309
High School        No           Less than High School    No           3.531
Masters            Yes          Less than High School    Yes          6.788
Masters            No           Less than High School    No           8.295
```

Table 1. Excerpted output from LS MEANS statement.

The output, above, illustrates not only the utility of odds ratios, but additionally those obtained through the LS MEANS statement, investigating the practical significance, perhaps, of the interaction between highest degree level achieved, and disability status. While trivializing the investigation of slope-differences, we can observe through the comparison of odds ratios, between those who have earned a master’s degree and those who have earned a high school diploma, that the increase in likelihood of employment for persons without disabilities is 3.764 $(OddsRatio_{Row4} - OddsRatio_{Row2})$ fold. The same increase, for persons with disabilities, is only 3.479 fold $(OddsRatio_{Row3} - OddsRatio_{Row1})$, indicating a possible differential benefit as one obtains more educational human capital, as compared to persons with lesser educational human capital. As stated previously, I make no claims regarding the representativeness of this relationship, given its simplicity, and absence of other measures known to influence the likelihood of employment.

CONCLUSION

Handling large-scale survey data, such as the American Community Survey (ACS), must be carried out with caution. While the SURVEYLOGISTIC procedure will run without explicit declaration of the most appropriate variance estimation method, understanding the design weighting strategy of any set of survey data is critical. This paper provides a means for analyzing ACS data in the absence of the Successive Difference Replication method, as an option, instead relying on the Jackknife method for variance estimation. Furthermore, this paper provides an explanation for how to obtain the more salient output of a logistic regression model, in particular for categorical variable interactions, through the LS MEANS statement of PROC SURVEYLOGISTIC.

REFERENCES

Agnelli, R. (2014). Examples of Logistic Modeling with the SURVEYLOGISTIC Procedure.

American Community Survey Design and Methodology (2014). *Version 2.0.*

An, A. B. (2002). Performing logistic regression on survey data with the new SURVEYLOGISTIC procedure. In *Proceedings of the twenty-seventh annual SAS® users group international conference* (pp. 258-27). SAS Institute Inc. Cary, NC.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review/Revue Internationale de Statistique*, 279-292.

Fay, R. E., & Train, G. (1995). Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties. In *Proceedings of the Section on Government Statistics, American Statistical Association, Alexandria, VA* (pp. 154-159).

Judkins, D. R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, *6*(3), 223.

Keathley, D., Navarro, A., & Asiala, M. E. (2010). An analysis of alternate variance estimation methods for the American Community Survey group quarters sample. In *2010 Joint Statistical Meetings: Proceedings of the Survey Research Methods Section*.

Morel, J. G. (1989). Logistic regression under complex survey designs. *Survey Methodology*, *15*(2), 203-223.

U.S. Census Bureau. (2006). Current Population Survey: Technical Paper 66—Design and Methodology. Retrieved from U.S. Census Bureau: http://www.census.gov/prod/2006pubs/tp-66.pdf

Wolter, K. M. (1984). An investigation of some estimators of variance for systematic sampling. *Journal of the American Statistical Association*, *79*(388), 781-790.

ACKNOWLEDGMENTS

The author gratefully acknowledges the National Technical Institute for the Deaf and the Rochester Institute of Technology.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Richard Dirmyer
National Technical Institute for the Deaf
Rochester Institute of Technology
52 Lomb Memorial Dr.
Rochester, NY 14623
rcdnvd@rit.edu