Paper 3320-2015

Using PROC SURVEYREG and PROC SURVEYLOGISTIC to Assess Potential Bias

Lucy D’Agostino McGowan, MS, Alice Toll, BS, Department of Biostatistics, Vanderbilt University

ABSTRACT

The Behavioral Risk Factor Surveillance System (BRFSS) collects data on health practices and risk behaviors via telephone survey. This study focuses on the question, “On average, how many hours of sleep do you get in a 24-hour period?” Recall bias is a potential concern in interviews and questionnaires, such as BRFSS. The 2013 BRFSS data is used to illustrate the proper methods for implementing PROC SURVEYREG and PROC SURVEYLOGISTIC, using the complex weighting scheme that BRFSS provides.

INTRODUCTION

This paper will go step by step through the process of analyzing survey data using SAS® software. The driving example will be using Behavioral Risk Factor Surveillance System (BRFSS) data, examining potential recall bias. BRFSS collects data on health practices and risk behaviors via telephone survey. Recall bias is a potential concern in interviews and questionnaires, such as BRFSS (Coughlin, 1990). This example focuses on the following question, “On average, how many hours of sleep do you get in a 24-hour period?” PROC SURVEYREG and PROC SURVEYLOGISTIC are used to establish whether the month the participant was interviewed is significantly associated with the number of hours they reported sleeping, controlling for demographics. Theoretically, since BRFSS employs random-digit dialing, the month a participant is interviewed should be random, and should not affect their responses. If there is differential recall in the hours of sleep based on the month the participant is interviewed, there is evidence for recall bias, in that the participants’ ability to recall the number of hours they sleep per night on average is influenced by the time of year that they are asked the question. Recall bias has potential to occur when participants are asked to report average behavior, because they often report their current behavior, even if it is different from their average. Previous studies have suggested a relationship between seasonal variation and sleep. A study conducted in the United States of 1571 individuals surveyed from the general population at random suggests that winter sleep increases of at least 2 hours per day relative to summer sleep was reported by nearly half of the population (Anderson et al., 1994). Another study conducted in the Netherlands found that sleep duration was 20 minutes longer during winter than during summer (Kantermann et al., 2007). We hypothesize that participants interviewed in December or January will report sleeping more on average than those interviewed in the remaining 10 months. If recall bias were not present, we would expect that the month that the participant was interviewed would not be associated with the average number of hours they sleep. If our hypothesis were shown to be true, it would indicate evidence of recall bias.

METHODS

BRFSS DATA

The Behavioral Risk Factor Surveillance System (BRFSS) is a telephone survey that collects data on preventative health practices and risk behaviors for chronic diseases in adults. BRFSS provides the sampling design and weighting scheme to allow for the calculation of valid direct estimates of prevalences. The estimates for hours of sleep were obtained from the following BRFSS question: “On average, how many hours of sleep do you get in a 24-hour period?” (Centers for Disease Control and Prevention. “Annual Survey Data.”, 2013).

COMPLEX SURVEY SAMPLING DESIGN

Complex survey sampling design requires specific analysis techniques in order to make proper inferences. Ideally, a sample should fully represent the population from which it is drawn (Figure 1a). Due to sampling limitations, however, this does not always happen. Proper weights must be applied so that the sample is representative of the population (Figure 1b). 2013 BRFSS data are used to illustrate the proper methods for implementing PROC SURVEYREG and PROC SURVEYLOGISTIC, utilizing the complex weighting scheme BRFSS provides.

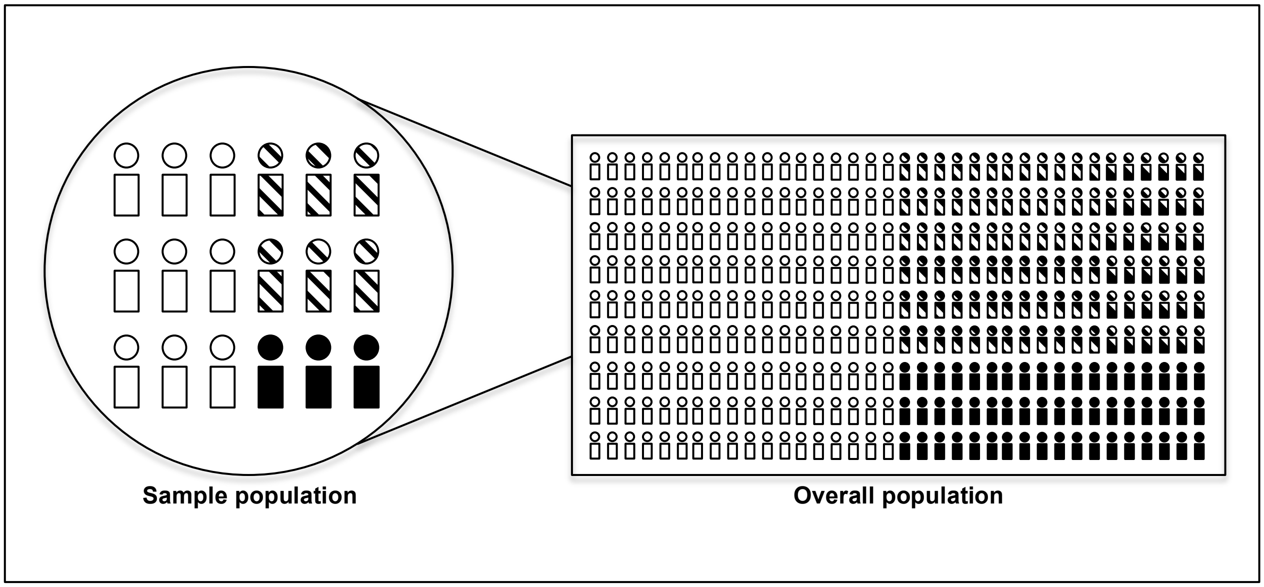

Figure 1a. Here the sample is proportionally representative of the overall population.

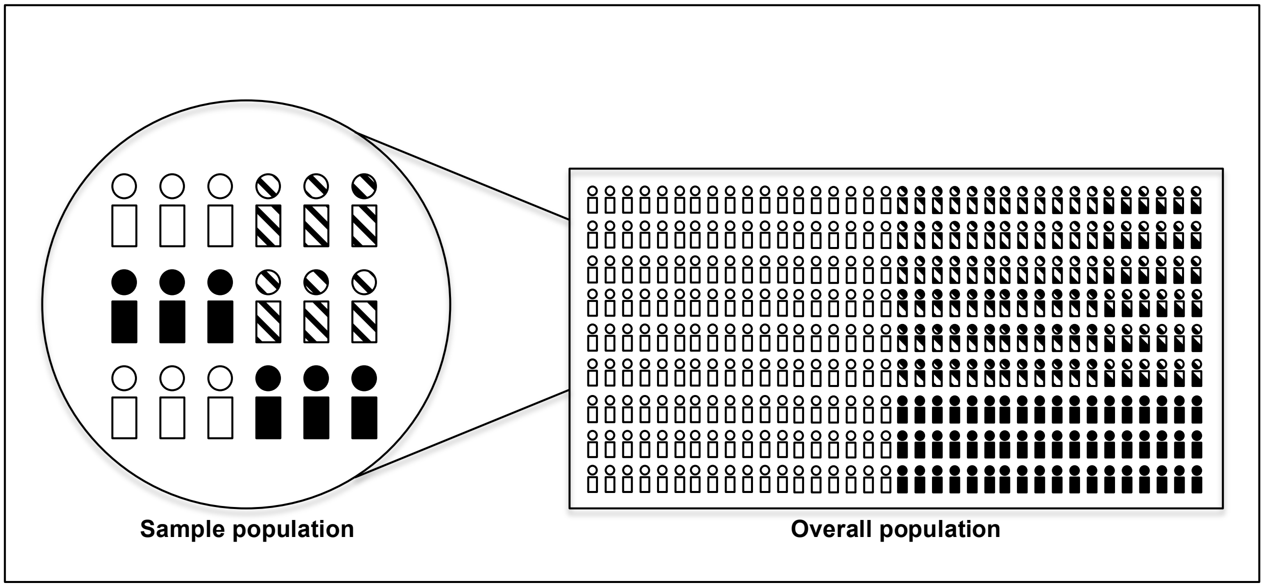

Figure 1b. Here the sample is not proportionally representative of the overall population. To correct for this, we could apply weights to the clear and filled portions of the sample to better represent the population.

WEIGHTING METHOD

BRFSS utilizes the raking weighting methodology ("Weighting the Data", 2013). This methodology involves two parts: the design weights and raking. The design weight is the following:

$$D = S \times \frac{1}{P} \times A$$

Where S is the stratum weight, which accounts for differences in the probability of the respondent's telephone number selection. We will incorporate this weight in our model in the `STRATA` statement. P is the number of residential telephone numbers in the respondent's household, and A is the number of adults in the respondent's household. The design weights are then raked into 8 margins in order to calculate the final weight. These margins are age group by gender, race/ethnicity, education, marital status, tenure, gender by race/ethnicity, age group by race/ethnicity, and phone ownership. In addition, if geographic regions are included, there are four additional margins, region, region by age group, region by gender, and region by race/ethnicity. BRFSS also performs weight trimming to temper the value of extreme weights. The main weight we will be utilizing is this trimmed weight, in the BRFSS dataset as `_LLCPWT`. Essentially, each participant is assigned a stratum weight, final weight, and a primary sampling unit that must be incorporated in the analysis process. The survey design team, BRFSS in this case, will provide these weights and their derivations.

MISSING DATA

Participants missing our outcome of interest, average hours of sleep per night, were removed from the analysis. Missing values were imputed using multiple imputation. PROC MI was used for this purpose, with PROC MIANALYZE used to combine the results of the five complete datasets (Yeats, 2009, Berglund 2010). For BRFSS data, complex imputation is often necessary, due to the branching logic and skip-patterns. For this simple example, these issues are not explored, since the variables chosen do not fall into this complex category, however it is useful to consider this. A few references that explore these complexities are listed in the "further reading" section below.

SAS SURVEY PROCEDURES

When modeling a continuous outcome, linear regression is often used. The common procedure in SAS is PROC REG. This, however, may not be a valid approach if the data arose from a complex design rather than a simple random sampling design. In such a case, PROC SURVEYREG is appropriate because it takes this design into account. Logistic regression is used when examining the relationship between a categorical outcome and a set of predictors. To perform logistic regression on a random sample, PROC LOGISTIC can be used. Again, this is not an appropriate approach if the data come from particular sampling designs, such as the BRFSS complex survey designs. In these cases, PROC SURVEYLOGISTIC is appropriate because it takes this design into account (An, 2002). Both PROC SURVEYREG and PROC SURVEYLOGISTIC use Taylor series or resampling methods to estimate sampling errors of the estimators based on the complex survey design, depending on the options chosen. The SAS Documentation, referenced in the further reading section, provides more detailed information.

DATA ANALYSIS

Analysis was conducted using SAS/STAT® version 9.4 (SAS Institute, Inc, Cary, North Carolina). Our outcome of interest is the average number of hours the participant is reporting to sleep each night. It is defined in two manners. To demonstrate PROC SURVEYREG, it is modeled continuously. To demonstrate PROC SURVEYLOGISTIC, we dichotomized sleep time. The categories are short sleep (6 or fewer hours) and adequate or long sleep (7 or greater hours). We adjusted for age group, sex (male/female), and race (non-Hispanic white, non-Hispanic black, non-Hispanic other, Hispanic). We are testing whether participants who responded in December or January differentially recalled the average number of hours they sleep per night.

MISSING DATA

To take a first look at the missing data before performing imputation, the following was implemented:

```
proc MI data=sas.sleep out=imputed seed=12345 nimpute=0;
 mcmc;
 var sleeptime decjan _llcpwt sex age race;
 run;
```

Notice above, NIMPUTE is set equal to 0, since we are only interested in our missingness pattern (Table 1).

Group	Sleep Time	Month	Weight	Sex	Age Group	Race	Frequency	Percent
1	X	X	X	X	X	X	476263	98.32
2	X	X	X	X	X	.	8138	1.68

Table 1. Missingness Pattern

From Table 1, the only variable with missing data is race. We observe a monotone missingness pattern, therefore in order to impute missing data, the following code was implemented prior to analysis:

```
proc MI data=sas.sleep out=imputed seed=12345 nimpute=5;
 class decjan sex age race;
 var sleeptime decjan _llcpwt sex age race;
 monotone logistic (race=sleeptime decjan _llcpwt sex age);
 run;
```

Here, the dependent variable, `sleeptime`, as well as all predictors and a sampling weight were included in the procedure (Yeats, 2009). In the code above, NIMPUTE is set to 5, which is the default. After the analysis is conducted, with PROC SURVEYREG or PROC SURVEYLOGISTIC, PROC MIANALYZE is used to combine the results.

LINEAR REGRESSION

Our outcome of interest is the average number of hours the participant is reporting to sleep each night. We adjusted for age group, sex, and race. Our primary hypothesis focused on whether the indicator variable for whether the participant was interviewed in December or January versus different months was associated with their self-reported average sleep time. The following is the SAS code used to implement the proper weights when performing a linear regression:

```
proc surveyreg data=sas.sleep order=formatted;
 strata _ststr;
 weight _llcpwt;
 cluster _psu;
 domain _imputation_;
 class decjan race age sex ;
 model sleeptime = decjan race age sex/solution ;
 format decjan yesno.;
 ods output ParameterEstimates=modelimpute (where=(_imputation_ ne . ));
run;
```

In the code above, `_ststr` is the stratum weight S from the design weight above, `_llcpwt` is the final trimmed weight, and `_psu` is the primary sampling unit, provided by BRFSS. To incorporate the 5 imputations, we have the `domain _imputation_` statement. The `class` statement contains all categorical variables. The `model` statement contains the model of interest. Notice that the `solution` option is also included in the model statement in order to receive the parameter estimtes.

The following will combine the 5 imputations after the regression is performed (Berglund, 2010 and Yeats, 2009). In the code above, the `ods output ParameterEstimates=modelimpute` will create a dataset with the parameters output. Each categorical variable will be saved in this dataset as the variable name and the numeric category. For example, race will be 3 variables “race 1” “race 2” and “race 3”. The spaces here are problematic for calling the variable into the PROC MIANALYZE, therefore the following code will compress the variables to avoid this:

```
data modelimpute;
 set modelimpute;
 parameter=compress(parameter);
run;
```

In the PROC MIANALYZE procedure, for the class variables, each of the dummy variable names needs to be included, leaving out the reference:

```
proc mianalyze parms=modelimpute;
 modeleffects intercept decjanAYes race1 race2 race3 age1 age2
 age3 age4 age5 sex1;
run;
```

Notice in the code above, the variable decjan is included as decjanAYes. This is because this variable was formatted as “A Yes”.

LOGISTIC REGRESSION

For the purposes of this example, we have dichotomized our outcome of interest, hours of sleep, into two categories in order to demonstrate the functionality of PROC SURVEYLOGISTIC. The categories are short sleep (6 or fewer hours) and adequate or long sleep (7 or greater hours). Again we adjusted for age group, sex, and race and are testing whether participants who responded in December or January differentially recalled the average number of hours they sleep per night. The following is the SAS code used to implement the proper weights when performing logistic regression:

```
proc surveylogistic data=sas.sleep order=formatted;
 strata _ststr;
 cluster _psu;
 weight _llcpwt;
 domain _imputation_;
 class decjan race age sex;
 model sleep2cat(event='1')= decjan race age sex;
 format decjan yesno.;
 ods output ParameterEstimates=modelimpute (where=(_imputation_ ne . ));
run;
```

In the code above, `_ststr` is the stratum weight S from the design weight above, `_llcpwt` is the final trimmed weight, and `_psu` is the primary sampling unit, provided by BRFSS. To incorporate the 5 imputations, we have the `domain _imputation_` statement. The `class` statement contains all categorical variables. The `model` statement contains the model of interest.

The following will combine the 5 imputations after the logistic regression. If you print the output dataset, here `modelimpute`, you will notice that an additional variable is included, ClassVal0. This variable allows us to skip a step when using PROC MIANALYZE as compared to combining the PROC SURVEYREG output. To account for the levels of the class variables, rather than compressing the parameters and including each level, we can use the `classvar=classval` option.

```
proc MIANALYZE parms(classvar=classval)=modelimpute;
 class decjan race age sex;
 modeleffects intercept decjan race age sex;
run;
```

RESULTS

There were 491,773 individuals included in the 2013 BRFSS survey. We excluded participants missing our outcome variable, sleep time. There were 451,388 individuals included in our analysis. 1.68 percent had missing race values (n=8138), which were imputed using multiple imputation. Participants interviewed in December or January are significantly more likely to report greater hours of sleep per night when modeled continuously, using PROC SURVEYREG, after adjusting for age group, sex (male/female), and race (non-Hispanic white/non-Hispanic black, non-Hispanic other, Hispanic) (p=0.0002, Table 2). Similarly, participants interviewed in December or January are significantly more likely to report adequate or long nightly sleep when the sleep variable was dichotomized in PROC

SURVEYLOGISTIC, adjusting for the same covariates as the linear model (p=0.002, Table 3). While these results are significant, and do suggest recall bias, the clinical significance may only be modest, since the effect size seems to be only a several minute difference between the winter and other months.

Parameter	Estimate	Std Error	95% Confidence Limits		p-value
Interviewed in December or January	0.045415	0.012313	0.02128	0.06955	0.0002

Table 2. The association between interview month and sleep time modeled continuously

Parameter	Estimate	Std Error	95% Confidence Limits		p-value
Interviewed in December or January	0.055540	0.017653	0.02094	0.09014	0.0017

Table 3. The association between interview month and sleep time modeled dichotomously

CONCLUSION

This paper outlines the statistical methods needed to analyze complex survey data, allowing the utilization of publically available data, such as BRFSS, as well as looking into biases commonly seen in survey studies. The results indicate that participants interviewed in December or January have differential recall of the average hours they sleep, as compared to participants interviewed in the remaining 10 months. Because BRFSS employs a random digit dialing survey technique, the month the participant is called should be random, and therefore should not affect any outcome of interest. Because we see a significant association here, it seems that participants asked to recall their average hours of sleep during December and January report sleeping more than those who are asked the same question in other months. It should be noted, however, that while these results are statistically significant, they are likely not clinically meaningful, since the magnitude of difference represents only a modest difference in time. This paper is meant to both show this interesting relationship between sleep recall and month interviewed, as well as provide a framework for future survey analyses in SAS.

FURTHER READING

REGARDING MULTIPLE IMPUTATION IN BRFSS DATA:

Fahimi, M. 2007. Imputation of Missing Data for the Behavioral Risk Factor Surveillance System (BRFSS): A Comprehensive Approach. Presented at American Association for Public Opinion Research Conference, Anaheim, CA, May 2007.

Moll, Philip Andrew, 2014. "A Comparison of Imputation Methods in the 2012 Behavioral Risk Factor Surveillance Survey" Scholar Archive. Paper 3503. Available at http://digitalcommons.ohsu.edu/etd/350

Rubin, D.B. 1987. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons, Inc.

SAS DOCUMENTATION:

SAS Institute Inc. 2011. SAS® 9.3 System Options: Reference, Second Edition. Cary, NC: SAS Institute Inc.

REFERENCES

An, Anthony B. 2002. “Performing Logistic Regression on Survey Data with the new SURVEYLOGISTIC Procedure.” *Proceedings of the Statistics and Data Analysis Section*, Orlando, FL: SUGI 27. Available at http://www2.sas.com/proceedings/sugi27/p258-27.pdf.

Anderson JL, Rosen LN, Mendelson WB, Jacobsen FM, Skwerer RG, Joseph-Vanderpool JR, Duncan CC, Wehr TA, Rosenthal NE. 1994. “Sleep in fall/winter seasonal affective disorder: effects of light and changing seasons*.” J Psychosom Res.* 38(4): 323-37.

Behavioral Risk Factor Surveillance System. “Weighting the Data.” Accessed October 15, 2014. http://www.cdc.gov/brfss/annual_data/2011/2011_weighting.htm

Berglund, Patricia A. 2010. “An Introduction to Multiple Imputation of Complex Sample Data using SAS® v9.4.” *Proceedings of Statistics and Data Analysis Section,* Seattle, WA: SAS GF 2010. Available at http://support.sas.com/resources/papers/proceedings10/265-2010.pdf.

Berglund, Patricia A. 2009. “Getting the Most out of the SAS® Survey Procedures: Repeated Replication Methods, Subpopulation Analysis, and Missing Data Options in SAS® v9.2” *Proceedings of Statistics and Data Analysis Section,* Washington D.C. SAS GF 2009. Available at: http://support.sas.com/resources/papers/proceedings09/246-2009.pdf

Brumback, Babette A; Dailey, Amy B; Zheng, Hao W. 2012. “Adjusting for Confounding by Neighborhood Using a Proportional Odds Model and Complex Survey Data.” *American Journal of Epidemiology*.175(11): 1133-1141.

Centers for Disease Control and Prevention. “Sleep and Sleep Disorders.” Accessed October 15, 2014. http://www.cdc.gov/sleep/.

Centers for Disease Control and Prevention. “Annual Survey Data.” Accessed October 2014. Behavioral Risk Factor Surveillance System. Annual Survey Data. Available online: http://www.cdc.gov/brfss/annual_data/annual_data.htm#2013.

Chen, Xiuhuan; Gorrell, Paul. 2008. “An Introduction to the SAS® Survey Analysis PROCs.” *Proceedings of the Statistics and Analysis Section*, Pittsburg, PA: NESUG 2008. Available at http://www.nesug.org/proceedings/nesug08/sa/sa06.pdf.

Coughlin, Steven S. 1990. “Recall bias in epidemiologic studies.” *Journal of Clinical Epidemiology,* 43(1): 87-91.

Kantermann, Thomas, Juda, Myriam, Merrow Martha, Roenneberg, Till. “The Human Circadian Clock’s Season Adjustment Is Disrupted by Daylight Saving Time.” *Current Biology*. 17(22):1996-2000.

Lohr, Sharon L. 2012. “Using SAS® for the Design, Analysis, and Visualization of Complex Surveys.” *Proceedings of the Statistics and Analysis Section,* Orlando, FL: SAS GF 2012. Available at http://support.sas.com/resources/papers/proceedings12/343-2012.pdf.

Smith, Tyler C; Smith, Besa. 2014. “Leveraging Publicly Available Data in the Classroom Using SAS® PROC SURVEYLOGISTIC.” *Proceedings of the SAS Global Forum Conference*, Washington D.C.: SAS GF 2014. Available at http://support.sas.com/resources/papers/proceedings14/1426-2014.pdf.

Thomas, Annette; Schubler, Marc N; Fischer, Joachim E; Terris Darcey D. 2009. “Employees’ sleep duration and body mass index: Potential confounders.” *Preventative Medicine,* 48(5): 467-470.

Ye, Yeats. 2009. “Multiple Imputation for Survey Data Analysis.” *Proceedings of the SouthEast SAS Users Group Conference, Birmingham*, AL: SESUG 2009. Available at http://analytics.ncsu.edu/sesug/2009/CC016.Ye.pdf.

ACKNOWLEDGEMENTS

We would like to thank our professor, Bob Johnson, for his advice and support.

CONTACT INFORMATION

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

Lucy D’Agostino McGowan
Department of Biostatistics, Vanderbilt University
ld.mcgowan@vanderbilt.edu

Alice Toll
Department of Biostatistics, Vanderbilt University
alice.e.toll@vanderbilt.edu