Paper 3958-2019

Don't Let Complex Survey Data Get the Best of You! SAS® Survey PROCs for Categorical Data Analysis

Charlotte Baker, Virginia Polytechnic Institute and State University

ABSTRACT

Data from US federal health surveys frequently use complex survey structures, rendering traditional procedures not useful for analysis. The SAS survey procedures exist but have not yet become a regularly used asset in analysis. Instead, users frequently choose to use other programs or add-ons for even the most basic of analyses. This paper demonstrates why the survey procedures such as SURVEYFREQ, SURVEYLOGISTIC, and SURVEYREG should be in everyone's toolbox when using complex survey data in research or practice.

INTRODUCTION

Categorical data analysis includes, but is not limited to, calculating basic frequencies of categorical variables, analyzing the simple relationships between two variables, and analyzing the relationships between multiple variables at the same time. No matter which of these situations you are faced with, it is extremely important to know what SAS® tools to use and how to interpret the results from those tools. An important factor to consider is whether the data was created with multi-stage probability sampling and whether it is intended to be used as such.

Multi-stage probability sampling is the combined use of various sampling methods, including cluster sampling and stratified sampling, to collect information on multiple segments of a population in order to ensure truly random but representative selection. Weights can be applied to this data to provide regional or national estimates while taking into account known distributions of age, sex, race, etc. For this paper, we will refer to this entire sampling process as complex survey sampling. Many surveys in the United States, such as the Behavioral Risk Factor Surveillance System (BRFSS) and the National Health and Nutrition Examination Survey (NHANES), utilize complex survey sampling to produce data that represents the health status of the entire United States.

Table 1 lists SAS procedures that may be very familiar to SAS users (non-survey procedures) and can often use weights. It also includes procedures (survey procedures) that can take advantage of all the components of complex survey samples including strata and clusters.

Non-Survey Procedures	Survey Procedures
PROC FREQ	PROC SURVEYFREQ
PROC REG	PROC SURVEYREG
PROC LOGISTIC	PROC SURVEYLOGISTIC
PROC MEANS	PROC SURVEYMEANS
PROC PHREG	PROC SURVEYPHREG
	PROC SURVEYSELECT
PROC MI/PROC MIANALYZE	PROC SURVEYIMPUTE

Table 1. SAS Survey and Non-Survey Procedures

Of the procedures listed in Table 1, several are useful for categorical data analysis. This paper will specifically discuss some considerations for writing and running the SURVEYFREQ procedures, the SURVEYLOGISTIC procedure, and the SURVEYREG procedure code, as well

as important output to pay attention to as you implement these procedures in your research.

DATA AND SCENARIO

To explain and demonstrate the procedures in this paper, we will use the following data and scenario. We are conducting a study using the 2017 BRFSS data from the Centers for Disease Control and Prevention (CDC). This data is freely available and the location can be found in the References section of this paper.

To obtain the best results for data that comes from complex survey data, we start with the complete original data set. The data set should include variables representing the strata, weights, replicate weights, and/or clusters that were utilized in the sampling. Some data may not include all four of these, but if they are present they should be taken into account. It is acceptable to delete unneeded variables but all observations should be retained. Doing this ensures that any statistics will appropriately take advantage of the sampling design. An appropriate sample, if needed, can be obtained by using the SURVEYSELECT procedure.

To create the data for this paper, we used the following syntax:

```
data brfss2;
set brfss;
keep _age_g sex wtkg3 persdoc2 _rfbmi5 cvdcrhd4 _race_g1
  _psu _STSTR _LLCPWT;
wtkg3 = wtkg3*.01;
if sex = 9 then sex = .;
if persdoc2 in (7,9) then persdoc2 = .;
if persdoc2 in (1,2) then persdoc2 = 1;
if _rfbmi5 = 9 then _rfbmi5 = .;
if cvdcrhd4 in (7,9) then cvdcrhd4 = .;
run;
```

We are interested in knowing a) whether there is a relationship between heart disease and access to care, b) whether that relationship is different for people who are normal or underweight compared to those that are overweight or obese, and c) how heart disease and access to care influence weight. To make sure we can use this data to answer our questions, we used the codebook from BRFSS 2017 to identify what variables might be useful. The variables _psu, _ststr, and _llcpwt are noted in the BRFSS 2017 documentation to be the cluster, strata, and weight variables respectively. The other variables being kept in this data step are our variables of interest – heart disease (cvdcrhd4), a marker of BMI level (_rfbmi5), a marker of weight (WTKG3), and a marker of access to care (persdoc2) – and other variables that could be potentially used in analysis (age, sex, race). The heart disease variable is a yes/no variable. The BMI variable is two levels – underweight/normal weight and overweight/obese. The access to care variable is an indicator of whether someone has a doctor or not and is also a yes/no variable. The modifications being made using the IF-THEN statements are only being used to create two level variables for the example analyses. Instructions in the codebook indicate that the variable WTKG3 has 2 implied decimals. To insert these decimals, we multiply WTKG3 by 0.01.

ENTER THE SURVEY PROCS!

Using the SAS survey procedures is not drastically different than using procedures that SAS users use for non-complex survey data. An overall piece of guidance is that for the survey procedures there can only be one WEIGHT statement. However, multiple STRATA, replicate weights (REPWEIGHTS), or CLUSTER statements can be used in the same procedure. If

using REPWEIGHTS, STRATA and CLUSTER statements will be ignored. Let's review basics of our three procedures through five examples.

DIFFERENCES IN THE PROC SURVEYFREQ AND PROC FREQ CODE

The basic structure of PROC SURVEYFREQ code has some similarities to PROC FREQ, but also has several key differences:

```
PROC FREQ < options > ;
BY variables ;
EXACT statistic-options < / computation-options > ;
OUTPUT <OUT=SAS-data-set > output-options ;
TABLES requests < / options > ;
TEST options ;
WEIGHT variable < / option > ;
RUN;

PROC SURVEYFREQ <options> ;
BY variables;
CLUSTER variables;
REPWEIGHTS variables </ options> ;
STRATA variables </ option> ;
TABLES requests </ options> ;
WEIGHT variable;
RUN;
```

The differences between the FREQ procedure and PROC SURVEYFREQ are highlighted in yellow above. As we have discussed, PROC SURVEYFREQ takes into account sampling clusters and strata that PROC FREQ cannot, ensuring that standard errors are accurate. This is the primary reason for using PROC SURVEYFREQ instead of PROC FREQ. The only required statements for either procedure are the PROC statements and RUN. It is good practice to specify the data set the procedures are to use. If more than one-way tables are necessary or we only need information on specific variables, the TABLES statement is also required.

EXAMPLE 1

To begin our study, we are interested in finding out how many residents of the United States have ever been diagnosed with heart disease and how many are underweight/normal weight or overweight/obese. We are also interested in finding out how many people have access to care.

If we were to use PROC FREQ and the weight variable _llcpwt our code might look like this:

```
proc freq data = brfss2;
tables _rfbmi5 cvdcrhd4 persdoc2;
weight _llcpwt;
run;
```

Output 1 contains the results from PROC FREQ:

OVERWEIGHT OR OBESE CALCULATED VARIABLE				
_RFBMI5	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	80366617	34.60	80366617	34.60
Yes	1.5191E8	65.40	2.3228E8	100.00
Frequency Missing = 23372665.985				

EVER DIAGNOSED WITH ANGINA OR CORONARY HEART DISEASE				
CVDCRHD4	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Yes	10274735	4.05	10274735	4.05
No	2.4351E8	95.95	2.5378E8	100.00
Frequency Missing = 1869088.1729				

At Least One Doctor				
PERSDOC2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
At Least One Doc	1.9729E8	77.55	1.9729E8	77.55
No Doc	57113496	22.45	2.5441E8	100.00
Frequency Missing = 1248159.1519				

Output 1. Obesity, Heart Disease, and Access to Care Weighted Frequency Output from PROC FREQ

If we were to use PROC SURVEYFREQ and the weight (_llcpwt), strata (_ststr), and cluster (_psu) variables, our code might look like this:

```
proc surveyfreq data = brfss2;
cluster _psu;
strata _ststr;
weight _llcpwt;
tables _rfbmi5 cvdcrhd4 persdoc2;
run;
```

Output 2 contains the results from PROC SURVEYFREQ:

Data Summary	
Number of Strata	1659
Number of Clusters	450016
Number of Observations	450016
Sum of Weights	255653205

OVERWEIGHT OR OBESE CALCULATED VARIABLE					
_RFBMI5	Frequency	Weighted Frequency	Std Err of Wgt Freq	Percent	Std Err of Percent
No	135781	80366617	425668	34.5989	0.1655
Yes	277789	151913922	466519	65.4011	0.1655
Total	413570	232280539	451185	100.0000	
Frequency Missing = 36446					

EVER DIAGNOSED WITH ANGINA OR CORONARY HEART DISEASE					
CVDCRHD4	Frequency	Weighted Frequency	Std Err of Wgt Freq	Percent	Std Err of Percent
Yes	25389	10274735	140612	4.0486	0.0554
No	420720	243509382	484319	95.9514	0.0554
Total	446109	253784117	472945	100.0000	
Frequency Missing = 3907					

At Least One Doctor					
PERSDOC2	Frequency	Weighted Frequency	Std Err of Wgt Freq	Percent	Std Err of Percent
At Least One Doc	374890	197291550	502049	77.5502	0.1434
No Doc	73313	57113496	386046	22.4498	0.1434
Total	448203	254405046	473182	100.0000	
Frequency Missing = 1813					

Output 2. Obesity, Heart Disease, and Access to Care Weighted Frequency Output from PROC SURVEYFREQ

While the weighted frequencies obtained from PROC FREQ (Output 1) and PROC SURVEYFREQ (Output 2) appear to be equal, we know those obtained from PROC SURVEYFREQ are more accurate for the nationally weighted estimate because they take into account the strata and clusters. The “Sum of Weights” in the first table of results from PROC

SURVEYFREQ (Output 2) gives the total weighted population size of 255,653,205 people whereas the total from PROC FREQ does not total this same amount even though it is close.

DIFFERENCES IN THE PROC SURVEYLOGISTIC AND PROC LOGISTIC CODE

The differences between the LOGISTIC procedure and PROC SURVEYLOGISTIC are highlighted in yellow below.

```
PROC LOGISTIC <options>;
BY variables;
CLASS variable <(options)> <variable <(options)> …> </ options>;
CODE <options>;
CONTRAST 'label' effect values<, effect values, …> </ options>;
EFFECT name=effect-type(variables </ options>);
EFFECTPLOT <plot-type <(plot-definition-options)>> </ options>;
ESTIMATE <'label'> estimate-specification </ options>;
EXACT <'label'> <INTERCEPT> <effects> </ options>;
EXACTOPTIONS options;
FREQ variable;
ID variables;
LSMEANS <model-effects> </ options>;
LSMESTIMATE model-effect lsmestimate-specification </ options>;
MODEL variable <(variable_options)> = <effects> </ options>;
MODEL events/trials = <effects> </ options>;
NLOPTIONS options;
ODDSRATIO <'label'> variable </ options>;
OUTPUT <OUT=SAS-data-set> <keyword=name <keyword=name …>> </ option>;
ROC <'label'> <specification> </ options>;
ROCCONTRAST <'label'> <contrast> </ options>;
SCORE <options>;
SLICE model-effect </ options>;
STORE <OUT=>item-store-name </ LABEL='label'>;
STRATA effects </ options>;
TEST equation1 <,equation2, …> </ option>;
UNITS <independent1=list1 <independent2=list2 …>> </ option>;
WEIGHT variable </ option>;
RUN;

PROC SURVEYLOGISTIC <options>;
BY variables;
CLASS variable <(v-options)> <variable <(v-options)> …> </ v-options>;
CLUSTER variables;
CONTRAST 'label'   effect values <, …effect values> </ options>;
DOMAIN variables <variable*variable variable*variable*variable …>;
EFFECT name = effect-type (variables </ options>);
ESTIMATE <'label'> estimate-specification </ options>;
FREQ variable;
LSMEANS <model-effects> </ options>;
LSMESTIMATE model-effect lsmestimate-specification </ options>;
MODEL events/trials = <effects </ options>>;
MODEL variable <(v-options)> = <effects> </ options>;
OUTPUT <OUT=SAS-data-set> <options> </ option>;
REPWEIGHTS variables </ options>;
SLICE model-effect </ options>;
STORE <OUT=>item-store-name </ LABEL='label'>;
STRATA variables </ option>;
```

```
TEST equation1 <,equation2, …> </ option>;
UNITS independent1 = list1 <…independentk = listk> </ option>;
WEIGHT variable;
RUN;
```

PROC SURVEYLOGISTIC takes into account sampling clusters and strata that PROC LOGISTIC cannot. The only required statements for either procedure are the PROC statement, MODEL statement, and RUN. It is good practice to specify the data set the procedures are to use. If using CLASS or EFFECT statements they must come before the MODEL statement. DOMAIN is a statement present in several of the survey procedures but not in PROC SURVEYFREQ. We use it instead of the BY statement to do analyses on particular subgroups of the population. The DOMAIN statement allows you to get results for each level of a specific variable the same way you would with BY. If you were only interested in results for one group (for example males), you would only pay attention to those results and ignore the results for the other groups or levels present in the DOMAIN variable.

EXAMPLE 2

We now want to look at the relationship between heart disease and access to care. We can use PROC SURVEYFREQ or PROC SURVEYLOGISTIC for this.

Our PROC SURVEYFREQ code might look like this:

```
proc surveyfreq data = brfss2;
cluster _psu;
strata _ststr;
weight _llcpwt;
tables persdoc2*cvdcrhd4 / or;
run;
```

Just as with PROC FREQ, we can use options such as OR to obtains relative risk and odds ratio results for our analyses.

Output 3 contains the results from PROC SURVEYFREQ:

Table of PERSDOC2 by CVDCRHD4						
PERSDOC2	CVDCRHD4	Frequency	Weighted Frequency	Std Err of Wgt Freq	Percent	Std Err of Percent
At Least One Doc	Yes	23898	9471978	131941	3.7503	0.0523
	No	347542	186257011	503713	73.7456	0.1486
	Total	371440	195728989	499966	77.4958	0.1440
No Doc	Yes	1404	753109	48874	0.2982	0.0193
	No	71504	56085016	382888	22.2060	0.1434
	Total	72908	56838124	385003	22.5042	0.1440
Total	Yes	25302	10225087	140367	4.0485	0.0556
	No	419046	242342027	482769	95.9515	0.0556
	Total	444348	252567113	471421	100.0000	
Frequency Missing = 5668						

Odds Ratio and Relative Risks (Row1/Row2)			
Statistic	Estimate	95% Confidence Limits	
Odds Ratio	3.7872	3.3210	4.3189
Column 1 Relative Risk	3.6523	3.2090	4.1569
Column 2 Relative Risk	0.9644	0.9623	0.9665
Sample Size = 444348			

Output 3. Simple Relationship Between Heart Disease and Access to Care Output from PROC SURVEYFREQ

Unlike the default PROC FREQ output, PROC SURVEYFREQ does not provide an N-way (2x2 in this case) table with row and column percentages for the results. However, it does include the weighted frequencies and percentages for each group so one can obtain the exact same data as needed. Just as with PROC FREQ, PROC SURVEYFREQ allows the use of options to obtain statistics such as the odds ratio for looking at the relationship between two variables. In this example, we find that people in the US who have at least one doctor are 3.79 times more likely to have been diagnosed with heart disease than people who have no doctor.

Remembering that we can also assess this relationship with a simple logistic regression, our PROC SURVEYLOGISTIC code might look like this:

```
proc surveylogistic data = brfss2;
cluster _psu;
strata _ststr;
```

```
weight _llcpwt;
class persdoc2 (ref='No Doc');
model cvdcrhd4 (event='Yes')= persdoc2;
run;
```

Output 4 contains selected results from the PROC SURVEYLOGISTIC code.

Response Profile

Ordered Value	CVDCRHD4	Total Frequency	Total Weight
1	No	419046	242342027
2	Yes	25302	10225087
	NotaskedorMissing	.	.

Probability modeled is CVDCRHD4='Yes'.

Odds Ratio Estimates

Effect	Point Estimate	95% Confidence Limits	
PERSDOC2 At Least One Doc vs No Doc	3.787	3.321	4.319

NOTE:
The degrees of freedom in computing the confidence limits is 442689.

Output 4. Simple Relationship Between Heart Disease and Access to Care Output from PROC SURVEYLOGISTIC

The fourth table in the PROC SURVEYLOGISTIC results is very important (the first table in Output 4). It specifically lets you know what outcome your code is written to model. In this case, we are modeling the probability of having been diagnosed with heart disease. The (event = 'Yes') syntax on the model statement was used to make sure this happened. Without it, SAS defaulted to 'No' as the outcome to be modeled.

Since we obtained an odds ratio from PROC SURVEYFREQ, we can compare that value to the information in the eleventh table in the PROC SURVEYLOGISTIC results (the second table in Output 4). Just as before, we find that people who have at least one doctor are 3.79 times more likely to have been diagnosed with heart disease than people who have no doctor. It is worth noting here that we ensured that our comparison would be 'At Least One Doc' vs 'No Doc' by including the CLASS statement to tell SAS that persdoc2 was a categorical variable and the (ref='No Doc') option in that same statement to tell SAS what the reference group should be.

EXAMPLE 3

The next step in this example study is to determine if the relationship found in Example 2 is different for people who are normal/underweight compared to those that are overweight/obese. As with Example 2, we can use both PROC SURVEYFREQ and PROC SURVEYLOGISTIC.

If we were using PROC FREQ instead of PROC SURVEYFREQ, we would likely use a BY statement to ask for results for our two groups of BMI from our variable _rfbmi5. Because we are using our complex survey data, we need to use a three way table instead of a two

way table because PROC SURVEYFREQ does not have a DOMAIN statement and BY won't handle the survey structure correctly. The three way table provides us with the best estimates that take into account our sampling structure. The first variable in the three way (_rfbmi5) is the one that contains the group(s) you want to see the two way relationship (between persdoc2 and cvdcrhd4) for. As you can see in Output 5 and Output 6, you can see that we get a table for each level of _rfbmi5 and the odds ratio for each table.

```
proc surveyfreq data = brfss2;
cluster _psu;
strata _ststr;
weight _llcpwt;
tables _rfbmi5*persdoc2*cvdcrhd4 / or;
run;
```

Table of PERSDOC2 by CVDCRHD4

Controlling for _RFBMI5=No

PERSDOC2	CVDCRHD4	Frequency	Weighted Frequency	Std Err of Wgt Freq	Percent	Std Err of Percent
At Least One Doc	Yes	5350	2055571	64088	2.5863	0.0803
	No	103972	57837518	373946	72.7712	0.2708
	Total	109322	59893089	377256	75.3575	0.2650
No Doc	Yes	368	146469	14815	0.1843	0.0186
	No	24509	19439043	237246	24.4582	0.2647
	Total	24877	19585512	237630	24.6425	0.2650
Total	Yes	5718	2202040	65758	2.7706	0.0824
	No	128481	77276562	420715	97.2294	0.0824
	Total	134199	79478602	423193	100.0000	

Odds Ratio and Relative Risks (Row1/Row2)

Statistic	Estimate	95% Confidence Limits	
Odds Ratio	4.7169	3.8254	5.8161
Column 1 Relative Risk	4.5893	3.7296	5.6471
Column 2 Relative Risk	0.9730	0.9704	0.9755

Sample Size = 408596

Output 5. Relationship Between Heart Disease and Access to Care for Normal/Underweight BMI Output from PROC SURVEYFREQ

Table of PERSDOC2 by CVDCRHD4						
Controlling for _RFBMI5=Yes						
PERSDOC2	CVDCRHD4	Frequency	Weighted Frequency	Std Err of Wgt Freq	Percent	Std Err of Percent
At Least One Doc	Yes	17340	6893567	111170	4.5913	0.0738
	No	216269	112437669	435253	74.8872	0.1885
	Total	233609	119331236	437847	79.4785	0.1815
No Doc	Yes	928	532956	44876	0.3550	0.0298
	No	39860	30278517	293953	20.1665	0.1803
	Total	40788	30811474	296757	20.5215	0.1815
Total	Yes	18268	7426523	119630	4.9463	0.0792
	No	256129	142716186	463674	95.0537	0.0792
	Total	274397	150142710	463505	100.0000	

Odds Ratio and Relative Risks (Row1/Row2)			
Statistic	Estimate	95% Confidence Limits	
Odds Ratio	3.4832	2.9395	4.1274
Column 1 Relative Risk	3.3397	2.8274	3.9448
Column 2 Relative Risk	0.9588	0.9555	0.9621
Sample Size = 408596			

Output 6. Relationship Between Heart Disease and Access to Care for Overweight/Obese BMI Output from PROC SURVEYFREQ

Our results show that people who are underweight or normal weight and have at least one doctor are 4.72 times more likely to have been diagnosed with heart disease compared to people that have no doctor. This result is different than for overweight and obese people in this same data set. People who are overweight or obese who have at least one doctor are 3.48 times more likely to have been diagnosed with heart disease compared to those that have no doctor. Based on this information, we have found the answer to our research question is that yes the relationship is different for people who are normal/underweight compared to those that are overweight/obese.

Using PROC SURVEYLOGISTIC, we can find similar answers.

```
proc surveylogistic data = brfss2;
domain _rfbmi5;
cluster _psu;
strata _ststr;
weight _llcpwt;
```

```
class persdoc2 (ref='No Doc');
model cvdcrhd4 (event='Yes')= persdoc2;
run;
```

We added the BMI variable to a DOMAIN statement. Whereas it was not available to use it for PROC SURVEYFREQ it is available to use here. Use of this statement continues to help us make sure to have the most accurate statistics available but, quite importantly, is how we get to see results for only the participants of interest in a study. In this example, we are interested in the results for both groups (normal/underweight and overweight/obese) but in other situations you might only be interested in the results for one of them (for example, normal/underweight). In these other situations, you only need to pay attention to the results for that level of the domain. This is best indicated in Output 7, Output 8, and Output 9.

Domain Analysis for domain OVERWEIGHT OR OBESE CALCULATED VARIABLE=No

Domain Summary	
Number of Observations	450016
Number of Observations in Domain	135781
Number of Observations not in Domain	314235
Sum of Weights in Domain	80366617

Output 7. Output from PROC SURVEYLOGISTIC DOMAIN Summary - Normal/Underweight BMI

In Output 7 we see a box from the PROC SURVEYLOGISTIC output that specifically provides information for the subjects in the brfss2 data set that have a value of NO for being overweight or obese. These subjects are the normal/underweight persons. We can see the Sum of Weights for the Domain. This is similar as the Sum of Weights we originally saw in Output 2 – it tells us the number of observations once the aspects of the complex survey sample (weights, clusters, strata, etc) are taken into account. The results in Output 8 apply to just this group of people. The results in Output 9 apply to people who are overweight or obese.

Domain Analysis for domain OVERWEIGHT OR OBESE CALCULATED VARIABLE=No

Odds Ratio Estimates			
Effect	Point Estimate	95% Confidence Limits	
PERSDOC2 At Least One Doc vs No Doc	4.717	3.825	5.816

NOTE:
The degrees of freedom in computing the confidence limits is 406937.

Output 8. Relationship Between Heart Disease and Access to Care for Normal/Underweight BMI Output from PROC SURVEYLOGISTIC

Domain Analysis for domain OVERWEIGHT OR OBESE CALCULATED VARIABLE=Yes

Odds Ratio Estimates			
Effect	Point Estimate	95% Confidence Limits	
PERSDOC2 At Least One Doc vs No Doc	3.483	2.939	4.127

NOTE:
The degrees of freedom in computing the confidence limits is 406937.

Output 9. Relationship Between Heart Disease and Access to Care for Overweight/Obese BMI Output from PROC SURVEYLOGISTIC

As we can see in our results from the PROC SURVEYLOGISTIC (Output 8 and Output 9), we have the same results as we obtained from PROC SURVEYFREQ. If normal or underweight, people who have at least one doctor are 4.72 times more likely to have been diagnosed with heart disease compared to people that have no doctor. If obese or overweight, people who have at least one doctor are 3.48 times more likely to have been diagnosed with heart disease compared to those that have no doctor. The relationship between access to care and heart disease is different based on BMI status. This leads us to look a little closer at this relationship by advancing our use of PROC SURVEYLOGISTIC.

EXAMPLE 4

Example 3 left us with a conclusion that the primary relationship of interest between access to care and heart disease was modified by BMI status. Because of this, we want to make sure we look at our regression model again and include that information in our analysis. We do this with the addition of an interaction term, persdoc2*_rfbmi5.

```
proc surveylogistic data = brfss2;
cluster _psu;
strata _ststr;
weight _llcpwt;
class persdoc2 (ref='No Doc') _rfbmi5 (ref='Yes');
model cvdcrhd4 (event='Yes')= persdoc2 _rfbmi5 persdoc2*_rfbmi5;
run;
```

In this syntax, we add BMI to the CLASS statement and to the MODEL statement. We also use persdoc2*_rfbmi5 to have SAS show us the effect of third variable on the primary relationship between heart disease and access to care. We expect that term to be statistically significant in our output since we saw the difference ourselves in Example 3.

Analysis of Maximum Likelihood Estimates						
Parameter			Estimate	Standard Error	t Value	Pr > |t|
Intercept			-3.7640	0.0344	-109.56	<.0001
PERSDOC2	At Least One Doc		0.6996	0.0344	20.35	<.0001
_RFBMI5	No		-0.3482	0.0344	-10.13	<.0001
PERSDOC2*_RFBMI5	At Least One Doc	No	0.0756	0.0344	2.20	0.0278
NOTE: The degrees of freedom for the t tests is 406937.						

Output 10. Relationship Between Heart Disease and Access to Care Taking Into Consideration the Effect of BMI Output from PROC SURVEYLOGISTIC

To see the effect of the third variable (BMI) on the primary relationship, we look at a different table from the PROC SURVEYLOGISTIC output. In our Output 10 we can see that BMI influences the relationship between access to care (persdoc2) and heart disease. We know this because the p-value associated with the t test for the line PERSDOC2*_RFBMI5 is less than 0.05 (if 0.05 is your significance cutoff). Because BMI is in the interaction term, it has to be in the model as a standalone variable as well. Even if it did not show as significant alone, it must stay in the model as long as the interaction term is significant. In this example, _RFBMI5 is significant with a p-value of <0.0001.

PROC SURVEYREG

To this point, we have discussed PROC SURVEYFREQ and PROC SURVEYLOGISTIC, two tools that are excellent for categorical data analysis with complex survey data. The REG procedure is a general all around regression tool. The benefit of it when using categorical variables is that you have the ability to look at linear (continuous) outcomes but take into consideration how groups or other qualitative information influences those outcomes. When using surveys such as the BRFSS, using the SURVEYREG procedure this can be invaluable. From the syntax below we can see that there are quite a few differences in what statements and options are available in PROC SURVEYFREQ compared to PROC REG. In PROC REG, the PROC REG statement is required. If you fit a model, as we will do, a MODEL statement is required. In PROC SURVEYREG, the PROC SURVEYREG and MODEL statements are required. The CLASS statement is required for using categorical variables. As a note, as there is no PROC SURVEYGLM, you can use PROC SURVEYREG to run models as you would in GLM. Below we can see differences in PROC REG and PROC SURVEYREG code:

```
PROC REG <options>;
<label:> MODEL dependents = <regressors> </ options>;
BY variables;
FREQ variable;
ID variables;
VAR variables;
WEIGHT variable;
ADD variables;
CODE <options>;
DELETE variables;
<label:> MTEST <equation, …, equation> </ options>;
```

```
OUTPUT <OUT=SAS-data-set> <keyword=names> <…keyword=names>;
PAINT <condition |ALLOBS> </ options> |<STATUS |UNDO>;
PLOT <yvariable*xvariable> <=symbol> <…yvariable*xvariable> <=symbol> </
options>;
PRINT <options> <ANOVA> <MODELDATA>;
REFIT ;
RESTRICT equation, …, equation;
REWEIGHT <condition |ALLOBS> </ options> |<STATUS |UNDO>;
STORE <options>;
<label:> TEST equation, <, …, equation> </ option>;
RUN;

PROC SURVEYREG <options>;
BY variables;
CLASS variables;
CLUSTER variables;
CONTRAST 'label' effect values <…effect values> </ options>;
DOMAIN variables <variable*variable variable*variable*variable …>;
EFFECT name = effect-type (variables </ options>);
ESTIMATE <'label'> estimate-specification </ options>;
LSMEANS <model-effects> </ options>;
LSMESTIMATE model-effect lsmestimate-specification </ options>;
MODEL dependent = <effects> </ options>;
OUTPUT <keyword<=variable-name> …keyword<=variable-name>> </ option>;
REPWEIGHTS variables </ options>;
SLICE model-effect </ options>;
STORE <OUT=>item-store-name </LABEL='label'>;
STRATA variables </ options>;
TEST <model-effects> </ options>;
WEIGHT variable;
RUN;
```

EXAMPLE 5

The final question we wanted to answer from our scenario was whether our categorical variables for access to care and heart disease were associated with the linear version of weight. Our PROC SURVEYREG syntax might look like this:

```
PROC SURVEYREG data = brfss2;
CLASS persdoc2 (ref='No Doc') cvdcrhd4 (ref='No');
CLUSTER _psu;
MODEL wtkg3 = persdoc2 cvdcrhd4 / solution;
STRATA _ststr;
WEIGHT _llcpwt;
RUN;
```

We include our categorical variables on the CLASS statement as well as in our MODEL. As before, we make sure to set the reference groups using the ref=' ' syntax. Because we have a CLASS statement, we also want to be sure to use the option of SOLUTION on the MODEL statement to have SAS output the parameter estimates. Without it, this part of the output is suppressed. As we can see in Output 11, people who have more than one doctor are heavier by 1 kilogram than people who have no doctor when taking into consideration heart disease status. People who have heart disease are heavier by 5 kilograms than people who do not have heart disease when adjusting for access to care.

Estimated Regression Coefficients				
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	80.2648361	0.15075497	532.42	<.0001
PERSDOC2 At Least One Doc	0.9550579	0.17227080	5.54	<.0001
PERSDOC2 No Doc	0.0000000	0.00000000	.	.
CVDCRHD4 Yes	5.2669050	0.33011941	15.95	<.0001
CVDCRHD4 No	0.0000000	0.00000000	.	.

Output 11. Influence of Heart Disease and Access to Care on Weight in Kilograms - Output from PROC SURVEYREG

COMPUTING CONSIDERATIONS

The SAS survey PROCs can command significant computing resources, specifically RAM. They work on several versions of SAS including SAS 9.4 and SAS University Edition. Depending on the size of the data, the computer may require many or most resources that are available. If repeated slow-downs become a problem, accessing a machine with more processing power and RAM or accessing a server to run analyses may be necessary. However, in our experience, more computers are becoming available at a reasonable cost to make this even more accessible to all users.

CONCLUSION

PROC SURVEYFREQ, PROC SURVEYLOGISTIC, and PROC SURVEYREG are tools that should be in every programmers toolbox if they utilize complex survey sampling data. They eliminate the need for accessing additional software to do survey analysis and are capable of incorporating aspects of complex survey data such as strata, clusters, and weights that more basic procedures cannot. Spending the time to practice and incorporate the usage of these tools into a regular workflow can result in long term proficiency and continued expansion of the user base.

REFERENCES

Centers for Disease Control and Prevention. BRFSS 2017 Survey Data and Documentation, 2018, Available at https://www.cdc.gov/brfss/annual_data/annual_2017.html

SAS Institute, Inc. SAS/STAT 15.1 User's Guide. Cary, NC: SAS Institute Inc., 2019.

CONTACT INFORMATION

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

Charlotte Baker
Virginia Polytechnic Institute and State University
sesug.ops.2018@gmail.com