Paper 203-2023

Survey Data Analysis in SAS

Denis B. Nyongesa, Kaiser Permanente Center for Health Research; M. Marianne Jurasic, Boston University; Gregg H. Gilbert, University of Alabama at Birmingham; Tamara Lischka, Kaiser Permanente Center for Health Research.

Abstract

While few surveys use a simple random sampling design to collect data, regular statistical software analyzes data as if the data were collected using simple random sampling. Ignoring the sampling design in analysis of survey data may lead to incorrect point estimates and their standard errors. This paper provides an overview of the SAS procedures used in analyzing survey data. The SAS procedures handled in this paper include PROC SURVEYMEANS, PROC SURVEYFREQ, and PROC SURVEYLOGISTIC. The STRATA and WEIGHT statements for the procedures mentioned above will be discussed in relation to the sampling design used during the generation of the weighted estimates. Both univariate and multivariable weighted logistic regression models will be discussed. The cross-sectional questionnaire data from the National Dental Practice-Based Research Network study entitled “Deep Caries Removal Strategies: Findings from the National Dental Practice-Based Research Network” will be used to demonstrate the above-mentioned SAS procedures in SAS 9.4.

Introduction

Data collected from a sample of respondents is called survey data. Surveys are one of the powerful tools for data collection. They enable us to collect valuable data quickly and efficiently. To gather accurate and meaningful data that can inform our decisions and lead to successful outcomes, there is need to understand the purpose and process of surveys.

There are different sampling designs, the most common being simple random sampling design. Not all data are collected using simple random design due to cost and time; thus, regular statistical software (not designed for survey data) are not suited for analysis of such data. It is important to know what kind of sampling design was used to collect the data because the point estimates and standard errors are computed differently for different sampling designs.

SAS is one of the statistical software packages that is equipped with tools designed to analyze survey data. Some common statements in most of the survey data analysis SAS procedures are:

- WEIGHT: specifies the variable that contains the sampling weights
- STRATA: specifies the stratification variables
- CLUSTER: names variables that identify the clusters in a clustered sample design

This paper focuses on the following SAS procedures:

- PROC SURVEYFREQ,
- PROC SURVEYMEANS, and
- PROC SURVEYLOGISTIC

SURVEYFREQ procedure

The SURVEYFREQ procedure produces one-way to n-way frequency and crosstabulation tables from sample survey data. The tables include estimates of population totals, population proportions (overall proportions, and row and column proportions), and corresponding standard errors.

A select set of variables will be used to demonstrate the SAS Procedures.
Table 1 below shows the distribution of strata and weight variables used in our analysis.

PRIORITY_VA	SAMP_WEIGHT	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Not VA or CHC	4.9358	333	69.67	333	69.67
VA or CHC	1.5172	145	30.33	478	100.00

Table 1: Distribution of the strata and weight variables

We have two strata for our dataset, that is, the “VA or CHC” stratum and the “Not VA or CHC” stratum. Each stratum has its own sampling weight. A census was taken from the “VA or CHC” stratum (priority) while a simple random sample was taken from the “Not VA or CHC” stratum.

Table 2 shows the basic outputs obtained from running PROC SURVEYFREQ & PROC FREQ on the same dataset. The SURVEYFREQ procedure provides both weighted and unweighted frequency distribution.

PROC SURVEYFREQ						PROC FREQ (Assumes simple random sample)			
	Freq	Weighted Freq	Std Err of Wgt Freq	Percent	Std Err of Percent	Freq	Percent	Cumulative Frequency	Cumulative Percent
Gender: What is your sex/gender?									
Missing	3	11.38880	7.13298	0.6111	0.3827	3	0.63	3	0.63
Male	272	1151	43.97872	61.7668	2.3599	272	56.90	275	57.53
Female	203	701.13060	43.81571	37.6221	2.3511	203	42.47	478	100.00
Total	478	1864	8.16046E-6	100.0000					
Hispanic: Are you of Hispanic or Latino origin?									
No	453	1771	19.79265	95.0297	1.0621	453	94.77	453	94.77
Yes	25	92.62760	19.79265	4.9703	1.0621	25	5.23	478	100.00
Total	478	1864	6.47787E-6	100.0000					
Curi_num: Approximately how many adult patients (18+ years) do you treat who have at least one deep carious lesion in a posterior tooth?									
1 - 2	66	294.99540	34.26337	15.8292	1.8385	66	13.81	66	13.81
3 - 4	100	408.11500	38.31262	21.8991	2.0558	100	20.92	166	34.73
5 - 6	88	372.81560	37.24871	20.0050	1.9987	88	18.41	254	53.14
7+	224	787.68940	44.93422	42.2667	2.4111	224	46.86	478	100.00
Total	478	1864	6.26307E-6	100.0000					

Table 2: Frequency Distribution of the gender (sex), ethnicity (Hispanic), and adult patients (18+ years) treated in a month who have at least one deep carious lesion in a posterior tooth.

There are many other analysis outputs that can be generated by the SURVEYFREQ procedure (not shown in this paper). They include but are not limited to Confidence limits, coefficients of variation, design effects, simple and weighted kappa coefficients, goodness-of-fit tests (adjusted for the sample design) such as Rao-Scott chi-square test, Rao-Scott likelihood ratio test, the Wald chi-square test, and the Wald and the log-linear chi-square test, the risk difference, the odds ratio, and relative risks. You can also test a null hypothesis of equal proportions. Carrying out these tests depends on whether you have a one-way, two-way, 2 by 2, square, or multiway frequency tables.

PROC SURVEYFREQ uses ODS Graphics to create graphs as part of its output. Available statistical graphics include weighted frequency and percent plots, which can be displayed as bar charts or dot plots in various formats. It can also be used to generate mosaic plots.

SURVEYMEANS Procedure

This procedure estimates population means, standard errors and confidence limits from the sample survey data. It can also be used to obtain T-tests by adding a T to the PROC SURVEYMEANS statement. The procedure can also be used to compute estimates of proportions for categorical variables, estimates of quantiles for continuous variables, estimates of geometric means for positive continuous variables, and ratio estimates of means and proportions (not shown in this paper).

It can be used to provide domain analysis, that is, to compute estimates for domains (subgroups) in addition to estimates for the entire study population. To accomplish this, include a DOMAIN statement.

PROC SURVEYMEANS					**PROC MEANS (Assumes simple random sample)**			
Variable	**Mean**	**Std error**	**Lower 95% CL for Mean**	**Upper 95% CL for Mean**	**Mean**	**Std error**	**Lower 95% CL for Mean**	**Upper 95% CL for Mean**
PULP1	44.08	1.76	40.62	47.55	46.48	1.65	43.23	49.72
PULP2	19.83	1.21	17.46	22.20	19.78	1.10	17.62	21.93
PULP3	36.12	1.73	32.72	39.52	33.77	1.57	30.68	36.86
SYMP1	53.72	1.84	50.09	57.34	55.89	1.70	52.55	59.23
SYMP2	22.57	1.36	19.88	25.25	21.87	1.23	19.45	24.28
SYMP3	23.71	1.53	20.71	26.72	22.24	1.37	19.55	24.94

Table 3: Mean, Standard Errors and Confidence Limits of selected variables.

From table 3, the means and standard errors obtained by the two procedures differ, and so do the confidence limits.

The standard error computed by SURVEYMEANS procedure is that of the mean under the assumption that missing values are missing completely at random (MCAR). Although the difference between the two standard errors is not great, it is important to understand that different inferences may be drawn from such results.

Several other outputs can be obtained from both PROC SURVEYMEANS & PROC MEANS. The documentation for SURVEYMEANS and MEANS procedures lists the various keywords that can be used to request additional statistical outputs, e.g., confidence limits for the mean or sum; coefficient of variation for the mean or sum; t-tests etc. These variables must also be listed on the VAR statement.

PROC SURVEYLOGISTIC

Categorical responses are common in sample surveys. They can be binary, ordinal, or nominal. We use logistic regression to investigate the relationship between such discrete responses and the explanatory/independent variables.

Suppose x is a row vector of independent variables and π is the response probability to be modelled, the linear logistic model has the form: $logit(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \alpha + x\beta$, where α is the intercept parameter and β is the vector of slope parameters.

PROC SURVEYLOGISTIC provides an opportunity to choose one of the link functions that then results into fitting a broad array of binary response models of the form, $g(\pi) = \alpha + x\beta$, where α is the intercept parameter and β is the vector of slope parameters.

The SURVEYLOGISTIC procedure uses the method of maximum likelihood estimation (MLE) to fit linear logistic regression models for discrete response survey data. This procedure incorporates complex survey sample designs, (i.e., designs with strata, clusters, and unequal weighting), for statistical inferences. The MLE is carried out with either the Fisher scoring algorithm or the Newton-Raphson algorithm. If the starting values for the parameter estimates are known, they can be specified. The link functions can also be specified, i.e., logit, probit, or the complementary log-log function.[7]

In addition to the parameter estimates, the odds ratio (OR) estimates are also displayed. You can also specify the change in the explanatory variables for which odds ratio estimates are desired.

To estimate the sampling errors of estimators based on the complex sample designs, either the Taylor series (linearization) method or replication (resampling) methods are used to compute the variances of the regression parameters and odds ratios.

With this procedure, one can specify categorical (CLASS) variables as explanatory or independent variables. The procedure also enables you to specify interaction terms, similar to the LOGISTIC procedure. The SURVEYLOGISTIC procedure provides a CONTRAST statement for specifying customized hypothesis tests concerning the model parameters. The CONTRAST statement provides estimation of individual rows of contrasts, which is particularly useful for obtaining odds ratio estimates for various levels of the CLASS variables.

The SURVEYLOGISTIC Procedure syntax

The following statements are available in the SURVEYLOGISTIC procedure:

```
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 > ;
    < label: > TEST equation1 < , . . . , equationk > < / options > ;
    UNITS independent1 = list1 < . . . independentk = listk > < / option > ;
    WEIGHT variable ;
RUN;
```

Note that the PROC SURVEYLOGISTIC and MODEL statements are required.

For detailed syntax information, see the SAS documentation (see references for the link to online version of the documentation).

In our demonstration, we will not be using all the statements in the SURVEYLOGISTIC procedure. Not all the outputs from this procedure are going to be shown.

For our example, the response is binary, that is, “greater than or equal to 50%” or “less than 50%”. This was derived from the original survey field or variable that was continuous with values ranging from 0 to 100. The derivation was informed by some previous peer-reviewed studies in this field as well the different levels of the guideline concordance testing.
When interpreting odds ratios (OR), the important points to note are: OR >1 indicates increased occurrence of an event. OR <1 indicates decreased occurrence of an event. Look at confidence Intervals (CI) and P-value for statistical significance of value. At 95% CI, a p-value of ≤ 0.05 indicates that the value is significant. If 1 is within the CI, then the OR is not significant.

For example, practitioners who responded that patients’ general health is very to extremely important were 1.6 times more likely to report higher levels of concordance for choosing selective caries removal 50% or greater of the time [OR: 1.55, 95% CI (1.02, 2.35)] than those who responded with that general health is not at all to moderately Important.

(a) Univariable model (PROC SURVEYLOGISTIC)

Table 4 below shows the odds ratios and p-values of the univariable logistic regression model (see the SAS code in appendix A).

Odds Ratio Estimates & P-values				
Effect	**OR**	**95% CL**		**Pr > F**
GENDER: Female vs Male	1.042	0.686	1.583	0.847
G_RACE2: Asian vs White/Caucasian	0.991	0.556	1.767	0.911
Black/African American vs White/Caucasian	0.787	0.333	1.861	
Other/Multi-racial/Unknown vs White/Caucasian	1.206	0.525	2.770	
HISPANIC: Hispanic or Latino vs Not Hispanic or Latino	1.436	0.525	3.933	0.480
G_PRIMARYOCCUPATION: Federal government facility vs Private practice	1.132	0.613	2.088	0.002
Other vs Private practice	2.334	1.100	4.951	
Public health clinic vs Private practice	2.718	1.548	4.772	
GENERALISTORSPECIALIST: Specialist vs General Practitioner	1.205	0.528	2.748	0.657
FP_TIME: Full-time (32+ hours per week) vs Part-time (< 32 hours per week)	0.630	0.346	1.147	0.130
PARTICIPATIONLEVEL: Full Participation vs Limited Participation	0.846	0.531	1.347	0.479
G_DENTALSCHOOLYEAR: 1986-1998 vs 1967-1985	1.316	0.754	2.299	0.009
1999-2008 vs 1967-1985	1.580	0.900	2.771	
2009-2020 vs 1967-1985	2.848	1.550	5.235	
CARI_NUM: 1 - 2 vs 7+	0.724	0.398	1.318	0.735

Odds Ratio Estimates & P-values				
Effect	**OR**	**95% CL**		**Pr > F**
3 - 4 vs 7+	0.841	0.490	1.443	
5 - 6 vs 7+	0.951	0.545	1.660	
G_TRT_DUR_IMP: Very to Extremely Important vs Not at all to Moderately Important	0.798	0.529	1.203	0.281*
G_RESTOR_IMP: Very to Extremely Important vs Not at all to Moderately Important	1.012	0.610	1.679	0.964
G_MONEY_IMP: Very to Extremely Important vs Not at all to Moderately Important	1.519	1.004	2.298	0.048
G_AGE_IMP: Very to Extremely Important vs Not at all to Moderately Important	1.238	0.795	1.927	0.343*
G_PT_PREF_IMP: Very to Extremely Important vs Not at all to Moderately Important	1.082	0.681	1.720	0.738
G_GENHLTH_IMP: Very to Extremely Important vs Not at all to Moderately Important	1.523	0.995	2.330	0.053
G_ORALHLTH_IMP: Very to Extremely Important vs Not at all to Moderately Important	2.135	1.108	4.117	0.024

Table 4: Univariate Logistic Regression model - Odds Ratio Estimates & P-values

(b) Multivariable (PROC SURVEYLOGISTIC)

The first multivariate model has all the variables from the univariate model with pvalues less than 0.20 except the two variables (*) that were specifically requested to be included in the first multivariate model. Table 5 below shows the odds ratios and p-values of the initial multivariate model.

Odds Ratio Estimates & P-values				
Effect	**OR**	**95% CL**		**Pr > F**
G_PRIMARYOCCUPATION: Federal government facility vs Private practice	1.111	0.509	2.425	0.029
Other vs Private practice	1.996	0.745	5.342	
Public health clinic vs Private practice	2.685	1.350	5.337	
FP_TIME: Full-time (32+ hours per week) vs Part-time (< 32 hours per week)	0.511	0.254	1.026	0.059
G_DENTALSCHOOLYEAR: 1986-1998 vs 1967-1985	1.548	0.844	2.838	0.006
1999-2008 vs 1967-1985	1.762	0.936	3.317	
2009-2020 vs 1967-1985	3.755	1.801	7.828	
G_TRT_DUR_IMP: Very to Extremely Important vs Not at all to Moderately Important	0.582	0.360	0.941	0.027
G_MONEY_IMP: Very to Extremely Important vs Not at all to Moderately Important	1.915	1.154	3.178	0.012
G_AGE_IMP: Very to Extremely Important vs Not at all to Moderately Important	0.872	0.511	1.487	0.613
G_GENHLTH_IMP: Very to Extremely Important vs Not at all to Moderately Important	1.965	1.131	3.413	0.017
G_ORALHLTH_IMP: Very to Extremely Important vs Not at all to Moderately Important	1.494	0.664	3.358	0.331

Table 5: Initial Multivariate Logistic Regression model - Odds Ratio Estimates & P-values

To obtain the final multivariable model, only those variables with p-value less than or equal to 0.05 (from the initial multivariate analysis) were included. Table 6 below shows the odds ratios and p-values of the final multivariate model.

Odds Ratio Estimates & P-values				
Effect	**OR**	**95% CL**		**Pr > F**
G_PRIMARYOCCUPATION: Federal government facility vs Private practice	1.306	0.630	2.710	0.028
Other vs Private practice	1.906	0.772	4.706	
Public health clinic vs Private practice	2.526	1.335	4.780	
G_DENTALSCHOOLYEAR: 1986-1998 vs 1967-1985	1.347	0.743	2.441	0.022
1999-2008 vs 1967-1985	1.682	0.920	3.074	
2009-2020 vs 1967-1985	2.902	1.464	5.752	
G_MONEY_IMP: Very to Extremely Important vs Not at all to Moderately Important	1.888	1.180	3.021	0.008
G_GENHLTH_IMP: Very to Extremely Important vs Not at all to Moderately Important	1.864	1.166	2.979	0.009
G_TRT_DUR_IMP: Very to Extremely Important vs Not at all to Moderately Important	0.622	0.394	0.983	0.042

Table 6: Final Multivariate Logistic Regression model - Odds Ratio Estimates & P values

CONCLUSIONS

Ignoring the sampling design in the analysis of survey data may lead to incorrect point estimates and their standard errors. Therefore, we should adopt usage of the right tools for the job to have accurate results and interpretation of our analyses. As demonstrated, SAS is one of the statistical analysis software packages that is well equipped with tools and procedures to analyze survey data.

REFERENCES

1. https://www.questionpro.com/blog/surveys/ (accessed on 09/15/2023)
2. https://www.questionpro.com/blog/survey-data-collection/ (accessed on 09/15/2023)
3. Rust, K., (1985), Variance Estimation for Complex Estimators in Sample Surveys, Journal of Official Statistics, 1 (4), Statistics Sweden Publishing Service. Pages 381 –397.
4. https://stats.oarc.ucla.edu/sas/seminars/sas-survey/ (accessed on 09/15/2023)
5. https://support.sas.com/documentation/onlinedoc/stat/143/introsamp.pdf (accessed on 10/10/2023)
6. https://www.lexjansen.com/nesug/nesug08/sa/sa06.pdf (Accessed on 09/15/2023)
7. https://support.sas.com/documentation/onlinedoc/stat/142/surveylogistic.pdf (Accessed on 09/15/2023)
8. Rao, J. N. K., Wu, C. F. J., and Yue, K. (1992). “Some Recent Work on Resampling Methods for Complex Surveys.” Survey Methodology 18:209–217.
9. Särndal, C.-E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.
10. Binder, D. A. (1983). “On the Variances of Asymptotically Normal Estimators from Complex Surveys.” International Statistical Review 51:279–292.
11. Wolter, K. M. (2007). Introduction to Variance Estimation. 2nd ed. New York: Springer.
12. Jurasic MM, Gillespie S, Sorbara P, et al. (2022). Deep caries removal strategies: Findings from The National Dental Practice-Based Research Network. JADA; 153 (11): 1078-1088

ACKNOWLEDGMENTS

We would like to acknowledge the following for their tremendous contribution and support towards this paper: Suzanne Gillespie, Kaiser Permanente Center for Health Research; Pina Sorbara, University of Dundee; Deborah McEdward, University of Florida; Rahma Mungia, University of Texas Health Science Center at San Antonio; Pat Ragusa, University of Rochester; Heather Weidner, HealthPartners; and William Vollmer, Kaiser Permanente Center for Health Research.

CONTACT INFORMATION

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

Denis B. Nyongesa
Kaiser Permanente Center for Health Research
Denis.B.Nyongesa@kpchr.org

APPENDIX A: SAS CODE

```
/*frequency distribution of the strata variable (priority_va) by
the sampling weights variable (samp_weight)*/

     proc freq data=wuss2023.dcrs_analysis_aim1_v3;
      table   priority_va*samp_weight / list  missing;
     run;

/*frequency distribution (proc freq)of a few selected
variables*/
     proc freq data=wuss2023.dcrs_analysis_aim1_v3;
      table age gender  hispanic g_race  g_dentalschoolyear
            g_primaryoccupation generalistorspecialist
            participationlevel fp_time  cari_num/ list missing ;
     run;

/*frequency distribution (proc surveyfreq)of a few selected
variables*/
   proc surveyfreq data=wuss2023.dcrs_analysis_aim1_v3 missing;
      table age gender  hispanic g_race  g_dentalschoolyear
            g_primaryoccupation generalistorspecialist
            participationlevel fp_time  cari_num  ;
     strata priority_va;
     weight samp_weight;
   run;
```

```
/*drop the endodontics*/
     data wuss2023.dcrs_analysis_aim1_noendo;
      set wuss2023.dcrs_analysis_aim1_v3;
      where gt_endodontics = 0;
     run;

/*frequency distribution (proc surveyfreq)of a few selected
variables*/
title "percent of time choosing between three treatment options
for each clinical scenario: by all practitioners (overall)
except endodontists";
  proc means data = wuss2023.dcrs_analysis_aim1_noendo maxdec=3
  n nmiss mean std stderr clm q1 median q3 qrange min max range;
     var pulp_1 pulp_2 pulp_3 symp_1 symp_2 symp_3;
     format pract_speciality special. gt_endodontics yn. ;
     label     pulp_1  = "option1-symptomatic"
               pulp_2  = "option2-symptomatic"
               pulp_3  = "option3-symptomatic"
               symp_1  = "option1-asymptomatic"
               symp_2  = "option2-asymptomatic"
               symp_3  = "option3-asymptomatic";

   run;

title "percent of time choosing between three treatment options
for each clinical scenario: by all practitioners (overall)
except endodontists";
  proc surveymeans data = wuss2023.dcrs_analysis_aim1_noendo
varmethod=taylor ;
     var pulp_1 pulp_2 pulp_3 symp_1 symp_2 symp_3;
     strata priority_va;
     weight samp_weight;
     format pract_speciality special. gt_endodontics yn. ;
     label   pulp_1  = "option1-symptomatic"
               pulp_2  = "option2-symptomatic"
               pulp_3  = "option3-symptomatic"
               symp_1  = "option1-asymptomatic"
               symp_2  = "option2-asymptomatic"
               symp_3  = "option3-asymptomatic";

  run;
```

```
/*univariate model*/
ods noproctitle;
%macro univariate_model (var1, var2);
title "dcrs aim 1 (asymptomatic teeth) univariate analysis:
&var2.";
proc surveylogistic data=wuss2023.dcrs_analysis_aim1_noendo
;
strata priority_va;
class   symp_1_50 &var1./param=glm;
model symp_1_50(event='greater than or equal to 50%') =
&var2. ;
weight samp_weight;
run;
%mend;

%univariate_model(gender(ref='male'), gender);
%univariate_model(g_race2(ref='white/caucasian'), g_race2);
%univariate_model(hispanic, hispanic);
%univariate_model(g_primaryoccupation (ref='private
practice '), g_primaryoccupation);
%univariate_model(generalistorspecialist (ref = 'general
practitioner'), generalistorspecialist );
%univariate_model(g_dentalschoolyear(ref='1967-1985') ,
g_dentalschoolyear);
%univariate_model(cari_num(ref='7+'), cari_num);
%univariate_model(g_trt_dur_imp(ref='not at all to
moderately important'), g_trt_dur_imp);
%univariate_model(g_restor_imp(ref='not at all to
moderately important'), g_restor_imp);
%univariate_model(g_money_imp(ref='not at all to moderately
important'), g_money_imp);
%univariate_model(g_age_imp(ref='not at all to moderately
important'), g_age_imp);
%univariate_model(g_pt_pref_imp(ref='not at all to
moderately important'), g_pt_pref_imp);
%univariate_model(g_genhlth_imp(ref='not at all to
moderately important'), g_genhlth_imp);
%univariate_model(g_oralhlth_imp(ref='not at all to
moderately important'), g_oralhlth_imp);
%univariate_model(participationlevel, participationlevel);
%univariate_model(fp_time, fp_time);
```

```
     /*multivariate model: initial*/
     ods noproctitle;
     title "dcrs aim 1 (asymptomatic teeth) multivariate
     analysis: part 1";
proc surveylogistic data=wuss2023.dcrs_analysis_aim1_noendo ;
     strata priority_va;
     class   g_primaryoccupation (ref='private practice ')
        g_trt_dur_imp(ref='not at all to moderately important')
        g_restor_imp(ref='not at all to moderately important')
        g_money_imp(ref='not at all to moderately important')
        g_age_imp(ref='not at all to moderately important')
        g_pt_pref_imp(ref='not at all to moderately important')
        g_genhlth_imp(ref='not at all to moderately important')
        g_oralhlth_imp(ref='not at all to moderately important')
        symp_1_50
        gender(ref='male')
        g_race2(ref='white/caucasian')
        hispanic
        generalistorspecialist (ref = 'general practitioner')
        participationlevel
        fp_time
        g_dentalschoolyear(ref='1967-1985')
        cari_num(ref='7+')/param=glm;
     model symp_1_50(event='greater than or equal to 50%') =
     g_primaryoccupation fp_time g_dentalschoolyear
     g_trt_dur_imp g_money_imp g_age_imp g_genhlth_imp
     g_oralhlth_imp/ rsquare;
     weight samp_weight;
     run;

     /*multivariate model: final*/
     title "dcrs aim 1 (asymptomatic teeth) multivariate
     analysis: part 2";
   proc surveylogistic data=wuss2023.dcrs_analysis_aim1_noendo ;
     strata priority_va;
     class   g_primaryoccupation (ref='private practice ')
          g_trt_dur_imp(ref='not at all to moderately important')
          g_money_imp(ref='not at all to moderately important')
          g_genhlth_imp(ref='not at all to moderately important')
          symp_1_50
          g_dentalschoolyear(ref='1967-1985') /param=glm;
     model symp_1_50(event='greater than or equal to 50%') =
     g_primaryoccupation g_dentalschoolyear g_money_imp
     g_genhlth_imp  g_trt_dur_imp   / rsquare;
     weight samp_weight;
 run;
```