Xiang Yu

SAS Program Sample

```
/*************************************************************************

  Author: Xiang Yu
  Date: March 25, 2017

  Description: This program reads in person level measurement and
               Questionnaire data from the National Health and Nutrition
               Examination (NHANES) Survey provided by the Center for Disease
               Control and Prevention (CDC) to analyze association between
               thyroid disorder and selected synthetic agents.

               The program identifies important demographic covariates and
               synthetic agents associated with thyroid disease.

               Relationships between synthetic agents and thyroid disease are
               Estimated individually in a series of logistic models,
               controlling for the influence of demographic covariates.

               PROGRAM STEPS
               1. Set macro variables that define input and output libnames
                  and file names
               2. Read in NHANES SAS transport files
               3  Recode and clean data; merge data; define study sample
               4. Obtain descriptive statistics
               5. Identify demographic control variables (based on Step 4)
               6. Select synthetic agents of interest    (based on Step 4)
               7. Build base model (control variables)
               8. Estimate parameters for base model + one independent
                  variable for each agent of interest (with thyroid disorder
                  as the dependent variable)

*************************************************************************/

/******************
  STEP 1 OMMITTED
*******************/

/*************************************************************
  STEP 2: Read in SAS transport files with macro %IMPORT
*************************************************************
  macro %IMPORT
      1. read in NHANES SAS transport data sat
      2. create temporary SAS data set
      3. sort by respondent ID (SEQN) to merge all data sets

  INPUT
   path     = location of SAS transport data
   name     = name of SAS transport data
   new_name = output temporary SAS data set
*************************************************************/
```

```
%macro IMPORT(path=, name=, new_name=);

    libname in xport "&PATH/&NAME..xpt";

    data &new_name.;
        set in.&name.;
    run;

    proc sort data=&new_name.;
        by SEQN;
    run;

%mend import;

%import(path=&INPUT_DIR., name=DEMO_H,    new_name=demo);
%import(path=&INPUT_DIR., name=MCQ_H,     new_name=outcomes);
%import(path=&INPUT_DIR., name=SSPFAS_H, new_name=isomer);
%import(path=&INPUT_DIR., name=PHTHTE_H, new_name=urine);
%import(path=&INPUT_DIR., name=BMX_H,     new_name=BMI);

/*********************
  STEPS 3-5 OMMITTED
*********************/
/*****************************************************************
  STEP 6:  Select synthetic agents of interest
*****************************************************************
  The macro %SURVEYTTEST is used to identify potential important risk factors
  of disease status. The macro performs multiple t-tests to rank possible
  strong univariate relationships between the outcome (thyroid disease
  status) and risk factors (synthetic agents).

  The SAS surveymeans procedure is used to obtain descriptive statistics
  (mean, std error) that account for sampling weight and design weights in
  this NHANES study. These output statistics are used to conduct multiple 2-
  sample t-tests, using binary disease status (disease, no disease) as the
  grouping variable for each continuous indicator of synthetic agent
  concentration. P values are calculated (assuming unequal variances) for
  each variable of interest and output is ranked from the smallest P value to
  highest. The output from all of the t-tests are combined into one data set
  for comparison purposes. Small p-values indicate potential predictors of
  disease status.

  macro %SURVEYTTEST
    1.  read in a study data set
    2.  create a temporary output dataset with
            group0 = all observations
            group1 = disease status negative
            group2 = disease status positive
     3. create temporary SAS data set (ttests) ranked by p values

  INPUT
   dataname       = study data set
   dep_var        = variable of study interest
   indep_var_list = list of variables for t-tests
   strata_weight  = design weight from study data set
   cluster_weight = design weight from study data set
```

```
   person_weight  = sample weight from study data set

  OUTPUT
    ttests  -  SAS temporary data set (collection of fit statistics and p-
               value from all t-tests)

*************************************************************************/

%macro SURVEYTTEST(dataname=, dep_var =, indep_var_list=, strata_weight=,
                   cluster_weight=, person_weight=);


   * Sort by grouping variable (outcome: thyroid disease status);
   proc sort data=&dataname.;
        by &dep_var.;
   run;

   * Designate output data set names for statistics by the grouping variable;

   ods output Surveymeans.ByGroup1.Statistics =group0  /* All observations*/
              Surveymeans.ByGroup2.Statistics =group1  /* Disease*/
              Surveymeans.ByGroup3.Statistics =group2; /* No Disease*/


   * Calculate summary statistics (mean, std) that incorporate design and
     sampling weights for each group;


   proc surveymeans data=&dataname.;
      by &dep_var.;                  /* Disease status*/
      var &indep_var_list.;          /* Synthetic agents*/
          strata &strata_weight.;
          cluster &cluster_weight.;
          weight &person_weight.;
   run;


  * Merge summary statistics from each group (disease, no disease) and
    compute t statistics and p-values;

   data ttests;
     merge group1 (rename=(N=N1 mean=mean1 stderr=stderr1 lowerCLmean=
                           lowerCLmean1 upperCLmean=upperCLmean1))
           group2 (rename=(N=N2 mean=mean2 stderr=stderr2 lowerCLmean=
                           lowerCLmean2 upperCLmean=upperCLmean2));
     tstat = (mean1-mean2)/sqrt(stderr1**2 + stderr2**2 );
     df = min(N1,N2)-1;
     p=2*(1-probt(abs(tstat),df));
   run;


  * Sort results from t-test from lowest p-value to highest;

   proc sort data=ttests;
     by p;
   run;
```

```
  * Create a new variable for rank of important predictors of disease status,
    based on p-value;

   data ttests;
     set ttests;
     rank = _n_;
   run;

  * Print results;
   proc print data=ttests noobs;
     var varName;
   run;


%mend SURVEYTTEST;

%let var_list = SSNPFOA SSBPFOA SSNPFOS SSMPFOS URXCNP URXCOP URXECP
                URXMBP URXMC1 URXMEP URXMHH URXMHNC URXMHP URXMIB URXMNP
                URXMOH URXMZP;

%SURVEYTTEST(dataname=study_sample, dep_var=thyroid,
             indep_var_list=&var_list., strata_weight=SDMVSTRA,
             cluster_weight=SDMVSTRA, person_weight=WTSB2YRW);


/*****************************************************************
  STEP 7:  Build base model (control variables)
*****************************************************************

  The macro %RUN_BASEMODEL is used to establish a logistic model using the
  disease status as the binary response. This macro is aimed to identify the
  demographic covariates that would be included in Step8, where the
  candidates of synthetic agents screeend in Step 6 would be examined as
  predictors for thyroid disesase status.

  The SAS surveylogistic procedure is used to examine the model fitting and
  account for sampling weight and design weights in this NHANES study. The
  response variable is the binary status of the thyroid disease. The
  demographic preditor variables, age, gender, foreign-born status and race
  were identified in Step 5. These variables were examined individually and
  in combinations. The output fit statistics were compared to select
  the list of control variables desired for Step 8. The final list included
  age, gender and foreign-born status. The race variables was dropped because
  it was found that once the foreign-born status was included, the race
  variable failed to further improve the model fit.

  macro %run_basemodel
    1. read in a study data set
    2. create a temporary output dataset containing fit statistics

  INPUT
   dataname       = study data set
   dep_var        = binary (event=1, non-event=0) outcome variable of study
                    interest
   indep_var_list = demographic covariates of study interest
   var_label      = description of the included variable
```

```
   strata_weight  = design weight from study data set
   cluster_weight = design weight from study data set
   person_weight  = sample weight from study data set

  OUTPUT
   association_&var_label. -  Association statistics
   par_&var_labl.          -  Parameter estimates
   tests_&var_label.       -  Global test results
   fit_stats_&var_label.   -  Fit statistics


*******************************************************************/
%macro RUN_BASEMODEL(dataname=, dep_var=, ind_var_list=, var_label=,
                     strata_weight=, cluster_weight=, person_weight=);

  proc surveylogistic data=&dataname. ;
    model &dep_var. (event='1') = &ind_var_list.;
    stratum &strata_weight.;
    CLUSTER &cluster_weight.;
    WEIGHT &person_weight.;

       ods output logistic.Association=association_&var_label.
                  Surveylogistic.ParameterEstimates=par_&var_label.
                  Surveylogistic.GlobalTests=tests_&var_label.
                  Surveylogistic.FitStatistics=fit_stats_&var_label.;
  run;
  quit;

%mend RUN_BASEMODEL;

%let data=study_sample;
%let outcome=thyroid;
%let strata=SDMVSTRA;
%let cluster=SDMVSTRA;
%let p_weight=WTSB2YRW;

* Age only;
%run_basemodel(dataname=&data., dep_var=&outcome., ind_var_list=age,
               var_label=base_age, strata_weight=&strata.,
               cluster_weight=&cluster., person_weight=&p_weight.);
* Sex only;
%run_basemodel(dataname=&data., dep_var=&outcome., ind_var_list=sex,
               var_label=base_sex, strata_weight=&strata.,
               cluster_weight=&cluster., person_weight=&p_weight.);

* Race only - dummy variables;
%run_basemodel(dataname=&data., dep_var=&outcome., ind_var_list=race2 race3
               race4, var_label=base_race, strata_weight=&strata.,
               cluster_weight=&cluster., person_weight=&p_weight.);
* Nativity;
%run_basemodel(dataname=&data., dep_var=&outcome., ind_var_list=fb,
               var_label=base_fb, strata_weight=&strat.,
               cluster_weight=&cluster., person_weight=&p_weight.);
*Non-white;
%run_basemodel(dataname=&data., dep_var=&outcome., ind_var_list=non_white,
               var_label=,vbase_non_white,strata_weight=&strata.,
               cluster_weight=&cluster., person_weight=&p_weight.);
```

```
* Age, Sex, Nativity;
%run_basemodel(dataname=&data., dep_var=&outcome., ind_var_list=age sex fb,
               var_label=base, strata_weight=&strata.,
               cluster_weight=&cluster., person_weight=&p_weight.);



*****************************************************************
  STEP 8: Estimate parameters for base model + one independent
          variable for each agent of interest (with thyroid disorder
          as the dependent variable)
*****************************************************************
  macro %RUN_LOGISTIC
  surveylogistic procedure is used to account for sampling weight and design
  weight in this NHANES study. Variables identified from Step 6 is added one
  at a time to the base model.

  Codes omitted
*****************************************************************/
```