SAS Macro *%multisurr* to perform regression calibration for logistic regression with multiple surrogates for one exposure

Ruifeng Li, Edie Weller, Donna Spiegelman

May 4, 2004

Abstract

The SAS Macro *%multisurr* described in this documentation perform regression calibration for multiple surrogates with one exposure as discussed in the paper by Weller et al (submitted to Biostatistics, 2004). This type of data is often encountered in occupational studies where the measurement of exposure can be quite complex and is characterized by numerous factors of the workplace; therefore, multiple surrogates often describe one exposure. In this paper, methodology is developed along the lines of the regression calibration method to adjust the estimates of exposure-response associations for the bias and additional uncertainty due to exposure measurement error. The health outcome is assumed to be binary and related to the quantitative measure of exposure by a logistic link function. The relationship between the conditional mean of quantitative exposure measurement and job characteristics is assumed to be linear. A simulation option is available in the macro to evaluate the performance (percent bias, MSE and coverage probability) of the estimator for the data at hand.

Keywords: regression calibration, measurement error, multiple surrogates, occupational study

Contents

1 Description

%multisurr is a SAS macro with two sub-macros: *%adjfun*, for calculating the adjusted coefficients of exposure effect-response associations from regression calibration with multiple surrogates for one exposure; and *%simdata*, to generate main and validation datasets for simulation study.

This macro is appropriate for the following situation:

- One exposure is of interest with multiple surrogates measured for this exposure.
- The measurement error model relating the surrogates (**W**) and other covariates measured without error (**Z**) is linear.
- The logit of the probability of the binary outcome (D) is linearly related to the surrogates (**W**) and other covariates measured without error (**Z**).
- The data can be divided into two parts:

  (1) Validation study. In this data set, the true exposure information (X), multiple surrogates (**W**) and covariates measured without error (**Z**) are available on all n_2 subjects .

  (2) Main study. In this data set, the multiple surrogates (**W**), covariates measured without error (**Z**), and the binary outcome (D) are available on all n_1 subjects. There is no information about the true exposure (X) on these subjects.

The inputs for the SAS macro must include 1) the binary outcome (D) and surrogates (**W**) in the main study and 2) the true exposure quantity (X) and surrogates (**W**) in validation study. If there are covariates measured without error (**Z**), which the user is interested in adjusting for in the exposure-effect association, the **Z**'s should be provided for both the main study and the validation study. There is an option to perform a simulation study to examine the performance of the estimator. The criteria evaluated are the percent bias, MSE and the coverage probability. An additional option is available to allow for the exclusion of the covariates measured without error (**Z**).

The outputs from the SAS macro include:

- The estimates from the uncorrected logistic regression model among the main study subjects.
- The estimates from the measurement error model among the validation study subjects.
- The adjusted coefficients and their corresponding odds ratio (OR), 95% CI and p-values, and the covariance matrix between the adjusted coefficients.
- Simulation results, if requested (includes percent bias, MSE and coverage probability).

2 Invocation

The SAS code should include four parts: define weights; define SAS main macro *%multisurr* and its two sub-macros *%adjfun* and *%simdata*; read data into SAS for the input parameters of *%multisurr*; and input these values into *%multisurr*. More specific, to use the SAS main macro *%multisurr*, the user should excute a four-step invocation:

Step 1

Put the following SAS statement as part of the program:

%include "multisurr.sas"

Step 2

Define two SAS temporary data sets about the weights for the macro inputs incr_or1 and incr_or2.

Step 3

Read all the main and validation datasets for the macro inputs main_dataset and valid_dataset.

Step 4

Invoke *multisurr* with the following inputs :

1. true_exposure: the name of the true exposure in the data sets.

2. surrogates: the series names of the surrogates for the true exposure. Note, no comma between each other, just one by one with space. They are either binary with values 0/1 or continuous.

3. dependent_Var: the name of the binary outcome with values 0/1.

4. confounders: the series names of the covariates measured perfectly and used for adjustment. They are either binary with values 0/1 or continuous.

5. main_dataset: SAS main dataset containing outcome, surrogates, and confounders. No missing values are allowed, and no 0 cell counts by outcome.

6. valid_dataset: SAS validation dataset containing true exposure, surrogates, and confounders. No missing values are allowed, and no singular/ill-defined design matrix is allowed.

7. incr_or1: SAS data set to provide the increment of Odds ratios for each covariate in the uncorrected logistic model with interception.
   For example, if you have 3 surrogates W1, W2, W3 for your true exposure X, with 3 perfectly measure covariates Z1, Z2 and Z3 as confounders, and the increments of unit are all 1, then the following SAS code should be included before you call multisurr:

```
data incr_oc1;
  input  name $  weight;
  cards;
  intercept  1
  W1         1
  W2         1
  W3         1
  Z1         1
  Z2         1
```

```
  z3          1
     ;
run;
```

8. incr_or2: SAS data set to provide the increment of Odds ratios for the final corrected model using true exposure as covariate. Use the above example, the following SAS code should be included before you call multiplSurr:

```
data incr_or2;
  input name $ weight;
  cards;
  intercept  1
  X          1
  Z1         1
  Z2         1
  Z3         1
  ;

run;
```

The following 9-14 options for simulation study

9. simulation: T or F. T if you want to perform a simulation study; F otherwise.

10. nsim: the total number of simulations. Default is 2000.

11. nmain: the sample size for the main datasets in the simulation study. It can be the same number of the real main data set or else.

12. nvalid: the sample size for the validation datasets in the simulation study. It can be the same number of the real validation data set or else.

13. includeConfounders: T or F. T if you do have confounders and would like to include them in the simulation study; to include them in the simulation study; F if either you do not have confounders in the dataset or you don't want to include them even they exist.

14. seed: starting seed to generate simulation datasets.

3 Illustrative Example

In this section, we use the data presented in the Weller et al. paper as an example. The main dataset consists of 1040 workers (Greaves et a l., 1997) and the validation dataset consists of 83 workers in the exposure assessment study (Woskie et al., 1994). The following files are available on the web site for this example:

- *input.dat* contains the data in ASCII format.
- *multisurr.sas* is the SAS code to perform adjustments and simulation study.
- *testme.sas* contains the SAS code to run the test program.

Suppose all the files are saved in the current directory, here is the example SAS code in the file named **example.sas**:

```
/** step 1: define sas macro multisurr  **/
%include "multisurr.sas";

/** step 2: get incr1 and incr2 SAS data sets for units of OR **/
data incr1;
   input name $ weight;
   cards;
    intercept 1
```

```
    plant2    1
    grinding  1
    str       1
    syn       1
    agecat1   1
    agecat2   1
    agecat3   1
    racec     1
    smokenow  1
    ;
run;

data incr2;
   input name $ weight;
    cards;
    intercept 1
    truex     1
    agecat1   1
    agecat2   1
    agecat3   1
    racec     1
    smokenow  1
    ;
run;


/** step 3: read ASCII datafile into current temporary data sets **/
 data example;
   infile "input.dat" firstobs=2;
   input plant1 plant2 grinding str syn agecat1 agecat2 agecat3 racec
         smokenow weezmost valid truex;

run;


data valid;
  set example;
  if valid=1;
```

```
run;


data main;
  set example;
  if valid=0;
run;


%let filedirectory=/udd/strui/edie/FINAL/SASversion;
/* provide the pass where you save the multisurr.sas,
   adjfun.sas and simdata.sas
*/

%include "/udd/strui/edie/FINAL/SASversion/multisurr.sas";
/** step 4:  call multisurr macro with the corresponding inputs. **/
%multisurr(
    true_exposure=truex,
    surrogates= plant2 grinding str syn,
    dependent_Var = weezmost,
    confounders = agecat1 agecat2 agecat3 racec smokenow,
    main_dataset = main,
    valid_dataset = valid,
    incr_or1 = incr1,
    incr_or2 = incr2,
    simulation = T,
    nsim = 2000,
    nmain = 1040,
    nvalid =  83,
    includeConfounders=T,
    seed= 1256787
);
```

Once you have done editing the above SAS file, save it as example.sas, then run in any unix window with SAS version 8 (or higher) software as the following:

```
SAS example.sas
```

Then the output is saved as a file called **example.lst**

```
Note: omit the outputs from proc reg and proc logistic.

     MEASUREMENT ERROR CORRECTION OF REGRESSION ESTIMATES
FOR LOGISTIC REGRESSION WITH MULTIPLE SURROGATES FOR ONE EXPOSURE

References: Edie A. Weller, Donna Spiegelman, Don Milton, Ellen Eisen
Regression calibration for logistic regression with multiple surrogates
for one exposure
Programmers: Ruifeng Li & Edie Weller

Main study regression coefficients: Uncorrected
```

	WT	B	SE	OR	95% CI	p
intercep	1.00	-2.539	0.266	0.079	0.047 - 0.133	0.000
plant2	1.00	0.746	0.212	2.109	1.390 - 3.198	0.000
grinding	1.00	-0.349	0.324	0.706	0.374 - 1.332	0.282
str	1.00	0.496	0.195	1.641	1.119 - 2.407	0.011
syn	1.00	0.616	0.221	1.851	1.200 - 2.855	0.005
agecat1	1.00	-0.109	0.193	0.897	0.615 - 1.308	0.570
agecat2	1.00	-0.182	0.249	0.834	0.512 - 1.358	0.465
agecat3	1.00	-0.092	0.263	0.912	0.544 - 1.528	0.726
racec	1.00	0.159	0.198	1.173	0.796 - 1.729	0.420
smokenow	1.00	1.113	0.163	3.042	2.210 - 4.188	0.000

```
NOTE: 0.000 means it is less than 0.001.

Main study regression coefficients: Corrected
```

	WT	B	SE	OR	95% CI	p
intercep	1.00	-2.712	0.291	0.066	0.038 - 0.117	0.000
truex	1.00	1.056	0.385	2.875	1.353 - 6.108	0.006
agecat1	1.00	-0.035	0.203	0.965	0.648 - 1.437	0.861
agecat2	1.00	-0.159	0.259	0.853	0.513 - 1.418	0.540
agecat3	1.00	-0.090	0.273	0.914	0.535 - 1.561	0.741
racec	1.00	0.154	0.204	1.166	0.781 - 1.741	0.452
smokenow	1.00	1.091	0.168	2.978	2.144 - 4.138	0.000

NOTE: 0.000 means it is less than 0.001.

Measurement error model

	B	SE	P-value	tau
intercep	0.151	0.071	0.037	
plant2	-0.036	0.075	0.635	0.000
grinding	0.098	0.067	0.145	0.015
str	0.501	0.048	0.000	0.857
syn	0.298	0.061	0.000	0.127
agecat1	-0.070	0.061	0.259	
agecat2	-0.022	0.072	0.764	
agecat3	-0.002	0.070	0.979	
racec	0.005	0.047	0.910	
smokenow	0.020	0.038	0.598	

NOTE: The R square for the measurement error regression model is: 0.6832,

```
      and the adjusted R square is: 0.6442.
NOTE: 0.000 means it is less than 0.001.

------------------------------------------------------------------
 Simulation results (nsim=2000,nvalid=83,nmain=1040,
        start seed=1256787,  include confounders)
------------------------------------------------------------------


     BIAS       MSE     PBIAS    MEDIAN   COVPROB

-0.060065  0.176508 -5.688094 0.8330519     96.25
```

4 Warnings

The macro requires SAS version 8 or higher because ODS is used and the long variable names (more than 8 characters).

The names for surrogates and confounders are SAME for both validation and main datasets.

5 Credits

This is a SAS version of the original Splus functions, written by Edie Weller. Ruifeng Li has programmed them to the current SAS version. Dr. Donna Spiegelman has given valuable suggestions. Questions can be directed to Ruifeng Li:

strui@channing.harvard.edu

6 See Also

There is a Splus version written by Edie Weller and Ruifeng Li, please refer the corresponding manuscript for help.

There is another SAS macro called *%blinplus* to deal with multiple true exposure, each of them can only have one surrogate, please refer the corresponding manuscript for help.

7 References

Weller E., Spiegelman D., Milton D., Eisen E, *Regression Calibration for Logistic Regression with Multiple Surrogates for One Exposure*

Greaves IA, Eisen EA, Smith TJ, Pothier LJ, Kreibel D, Woskie SR, Kennedy SM, Shalat, S and Monson, RR (1997). Respiratory health of automobile workers exposed to metal-working fluid aerosols: respiratory symptoms. *American Journal of Industrial Medicine* **32**, 450-459.

Woskie SR, Smith TJ, Hallock MF, Hammond, SK, Rosenthal F, Eisen EA, Kreibel D, Greaves IA (1994). Size-selective pulmonary dose indices for metal-working fluid aerosols in machining and grinding operations in the automobile manufacturing industry. *American Industrial Hygiene Association* **55**,20-29.