SAS® GLOBAL FORUM 2020

MARCH 29 – APRIL 1
WASHINGTON, DC

USERS PROGRAM

A SAS® Macro for Calibration of Survey Weights

Tony An, PhD

SAS Institute Inc.

A SAS® Macro for Calibration of Survey Weights

Tony An, PhD

Tony An is a Principal Research Statistician Developer at SAS Institute. He developed several survey analysis procedures in SAS/STAT such as SURVEYMEANS and SURVEYREG, and SURVEYLOGISTIC. He received his PhD in statistics from Iowa State University. His areas of expertise include survey data analysis, nonresponse in survey sampling, regression analysis.

USERS PROGRAM

SAS® GLOBAL FORUM 2020

OUTLINE

- Introduction
- Calibration methods
- SurveyCalibrate macro
- Example
- Discussion

WHAT IS SURVEY SAMPLING?

- Study a finite population
- Collect data from probability samples
- Estimate finite population parameters
- Make statistically valid inferences

SAMPLING WEIGHTS

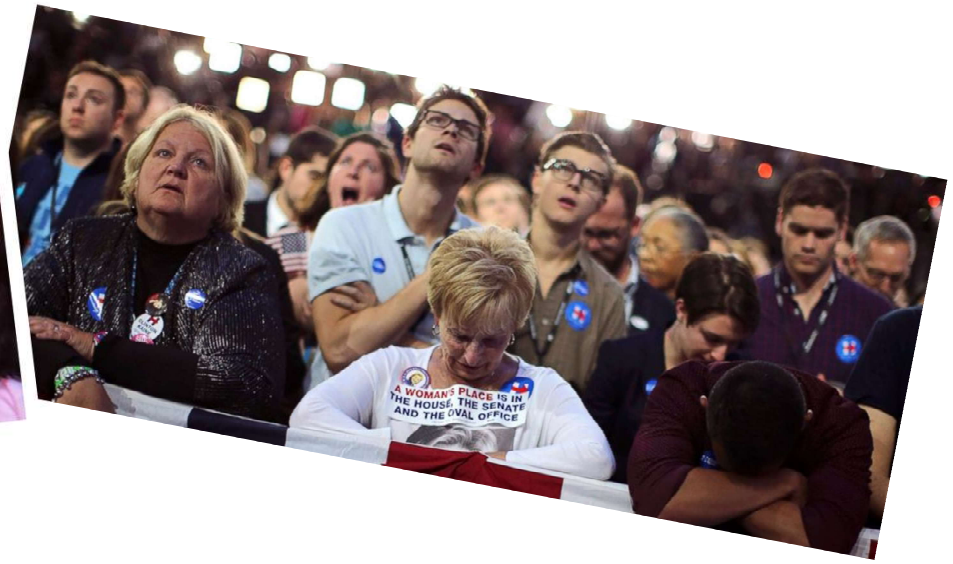
- Reduce bias
- Reflect sample design
- Adjust for nonresponses
- Estimate the variance

INCORRECT WEIGHTING

Why 2016 election polls missed their mark

BY ANDREW MERCER, CLAUDIA DEANE AND KYLEY MCGEENEY

Supporters of presidential candidate Hillary Clinton watch televised coverage of the U.S. presidential election at Comet Tavern in the Capitol Hill neighborhood of Seattle on Nov. 8. (Photo by Jason Redmond/AFP/Getty Images)



WEIGHT AJUSTMENTS

- Nonresponse adjustment to reduce bias
- Adjustment to match external sources
- Weight trimming and smoothing

AN EXAMPLE

- Restaurant utility usage
- Franchise or independent
- Known number of restaurants in each category

WEIGHTED SUM OF RESTRAUTNTS

Restaurant Type	Sum in Sample	Known Totals
Franchise	231	251
Independent	267	210

REMEDY - CALIBRATION

- Adjust the weights -> Calibration weights
- Calibration weights are as “close” as possible to the originals weights
- Estimates over control variables with calibration weights match known quantities

CALIBRATION METHOD

- Define a distance function ***G*** to measure the "closeness" between two sets of weights
- Find a solution that minimizes ***G*** under the *constraints* – matching known population totals ***T*** for a set of controls variables ***X***

CALIBRATION METHOD

$$\sum_{i=1}^{n} w_i G(\tilde{w}_i, w_i) = \min_{\{\mathbf{v}:\sum_{i=1}^{n} v_i \mathbf{x}_i = \mathbf{T}\}} \sum_{i=1}^{n} w_i G(v_i, w_i)$$

$$\tilde{w}_i = w_i \psi(\hat{\boldsymbol{\lambda}}' \mathbf{x}_i) \qquad \sum_{i=1}^{n} w_i \psi(\hat{\boldsymbol{\lambda}}' \mathbf{x}_i)\mathbf{x}_i = \mathbf{T}$$

COMMON CALIBRATION METHODS

- Linear
- Exponential (raking)
- Truncated linear
- Logit (truncated exponential)

LINEAR METHOD

Distance function: $$G(v, w) = \frac{1}{2}(v/w - 1)^2$$

Calibration weights

$$\begin{aligned} \tilde{w}_i &= w_i(1 + \hat{\boldsymbol{\lambda}}'\mathbf{x}_i) \\ &= w_i\left(1 + \mathbf{x}_i'\left(\sum_{i=1}^{n} w_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1}\left(\mathbf{T} - \sum_{i=1}^{n} w_i \mathbf{x}_i\right)\right) \end{aligned}$$

$$\hat{\boldsymbol{\lambda}} = \left(\sum_{i=1}^{n} w_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1}\left(\mathbf{T} - \sum_{i=1}^{n} w_i \mathbf{x}_i\right)$$

LINEAR METHODS

- Solution always exists
- Might have negative calibration weights

EXPONENTIAL METHOD

Distance function: $$G(v, w) = 1 + \frac{v}{w}\left(\log\left(\frac{v}{w}\right) - 1\right)$$

Calibration weights: $$\tilde{w}_i = w_i \exp(\hat{\boldsymbol{\lambda}}' \mathbf{x}_i)$$

EXPONENTIAL METHOD

- Calibration weights always positive
- Solution might not exist
- Might produce extremely large weights
- Equivalent to the raking method when there are two categorical control variables

TRUNCATED LINEAR METHOD

Distance function:

$$G(v,w) = \begin{cases} \frac{1}{2}(\frac{v}{w} - 1)^2 & \text{if } L < \frac{v}{w} < U \\ \infty & \text{otherwise} \end{cases}$$

Calibration weights:

$$\tilde{w}_i = \begin{cases} w_i L & \text{if } \hat{\boldsymbol{\lambda}}'\mathbf{x}_i < L - 1 \\ w_i(1 + \hat{\boldsymbol{\lambda}}'\mathbf{x}_i) & \text{if } L - 1 \leq \hat{\boldsymbol{\lambda}}'\mathbf{x}_i \leq U - 1 \\ w_i U & \text{if } \hat{\boldsymbol{\lambda}}'\mathbf{x}_i > U - 1 \end{cases}$$

TRUNCATED LINEAR METHODS

- Weights are bounded
- Solution might not exist
- Computation requires more resources

LOGIT METHOD

Distance function:

$$G(v,w) = \begin{cases} \left(\left(\frac{v}{w} - L\right)\log\left(\frac{v/w - L}{1-L}\right) + \left(U - \frac{v}{w}\right)\log\left(\frac{U - v/w}{U-1}\right)\right)\frac{(U-L)w}{(1-L)(U-1)} & \text{if } L < v/w < U \\ \infty & \text{otherwise} \end{cases}$$

Calibration weights:

$$\tilde{w}_i = w_i \psi(\hat{\boldsymbol{\lambda}}' \mathbf{x}_i)$$

LOGIT METHODS

- Weights are bounded
- Solution might not exist
- Computation requires more resources

%SurveyCalibrate MACRO

```
%macro SurveyCalibrate(
DATA=,       /* Input data set name                                          */
OUT=,        /* Output data set name                                         */
/*  Calibration parameters                                                  */
METHOD=,     /* LINEAR | EXPONENTIAL | TRUNLINEAR | LOGIT                   */
WEIGHT=,     /* Original weight variable                                     */
CALWT=,      /* Calibration weight variable, default CalWt                   */
CONTROLVAR=,/* Auxiliary control variables for calibration                  */
CTRLTOTAL=, /* Marginal totals for CONTROLVAR                               */
EPS=,        /* Convergence criterion for stopping iteration, default=0.01 */
MAXITER=,    /* Maximum number of iteration, default=25                      */
LOWER=,      /* Lower bound, must be in (0,1)                                */
UPPER=,      /* Upper bound, must be bigger than 1 or .                      */
NOINT=,      /* Do not keep sum of sampling weights unchanged                */
/* Replication parameters                                                   */
NOREPWT=,    /* Request no replicate weights                                 */
VARMETHOD=, /* BRR | JK | BOOTSTRAP, default is JK                           */
REPS=,       /* Number of replicates for bootstrap or brr                    */
CLUSTER=,    /* Cluster variables                                            */
STRATA=,     /* Strata variables                                             */
SEED=,       /* Random seed                                                  */
FAY=,        /* Fay coefficient for BRR varmethod                            */
RATE=,       /* FPC for bootstrap replicate weights                          */
OUTJKCOEFS=,/* OUTJKCOEFS data set                                          */
REPWEIGHTS= /* Replicate weight variables                                   */
);
```

MACRO PARAMETERS

- Calibration Parameters
- Replication parameters
- Some parameters can be left blank

REQUIRED CALIBRATION PARAMETERS

- DATA= /* Input data set name */
- OUT= /* Output data set name */
- WEIGHT=/* Original weight variable */
- CONTROLVAR=/* Auxiliary control variables for calibration */
- CTRLTOTAL= /* Marginal totals for CONTROLVAR */

SPECIFYING CONTROLS

- Control variables can be either continuous or categorical variables
- For categorical variables, you need to create indicator variables with data step before calling the macro

CATEGORICAL CONTROL VARIABLES

```
data myDataSet; set myDataSet;
   Male = 0; Female = 0;
   if (Gender='M') then Male = 1;
   if (Gender='F') then Female = 1;
run;
%SurveyCalibrate(
   DATA=myDataSet,
   ...
   CONTROLVAR=HouseholdIncome Male Female,
   CTRLTOTAL =12345678        300  400,
   ... );
```

OPTIONAL CALIBRATION PARAMETERS

- METHOD= /* LINEAR | EXPONENTIAL | TRUNLINEAR | LOGIT */
- CALWT= /* Calibration weight variable, default CalWt */
- EPS= /* Convergence criterion for stopping iteration, default=0.01 */
- MAXITER= /* Maximum number of iteration, default=25 */
- LOWER= /* Lower bound, must be in (0,1) */
- UPPER= /* Upper bound, must be bigger than 1 or . */
- NOINT= /* Do not keep sum of sampling weights unchanged */

CALIBRATION FOR REPLICATES

- Use the same calibration method for each replicate
- Ensure the correct variance estimation after the calibration
- Skip if no variance estimation needed

REPLICATION PARAMETERS

- NOREPWT= /* Request no replicate weights */
- VARMETHOD= /* BRR | JK | BOOTSTRAP, default is JK */
- REPS= /* Number of replicates for bootstrap or brr */
- CLUSTER= /* Cluster variables */
- STRATA= /* Strata variables */
- SEED= /* Random seed */
- FAY= /* Fay coefficient for BRR varmethod */
- RATE= /* FPC for bootstrap replicate weights */
- OUTJKCOEFS=/* OUTJKCOEFS data set */
- REPWEIGHTS= /* Replicate weight variables */

TIPS ON SPECIFYING PARAMETERS

- Leave optional parameters blank if not sure
- Run any other survey procedures to generate replicate weights before calibration
- Try and rerun by specifying more optional parameters
- Examine the replicate weights with various methods

AN EXAMPLE

- National Health and Nutrition Examination Survey I (NHANES I) Epidemiologic Followup Study (NHEFS)
- 174 observations from the 1992 NHEFS vital and tracing status data set

VARIABLES IN THE DATA SET

- **ID**, unit identification
- **VarStrata**, stratum identification
- **VarPSU**, identification for primary sampling units
- **SWeight**, sampling weight associated with each unit
- **Age**, the subject's reported age at the 1992 interview
- **VitalStatus**, vital status of subject in 1992 contact
- **PovArInd**, indicator subject's household location in terms of poverty area (1 = poverty area, 2 = nonpoverty area)
- **Gender**, gender of subject (1 = male, 2 = female)

THE DATA

Obs	ID	VarStrata	VarPSU	SWeight	Age	VitalStatus	PovArInd	Gender
1	1	3	1	13312	66	1	1	1
2	2	3	1	7941	71	3	1	2
3	3	3	1	16048	.	4	1	1
4	4	3	3	9298	58	3	1	1
5	5	3	2	15336	56	3	1	2
6	6	3	1	14744	63	1	1	1
7	7	3	2	83729	70	1	2	2
8	8	3	3	106492	57	1	2	1
9	9	3	3	78083	81	3	2	2
10	10	3	3	55957	79	3	2	1

KNOWN TOTALS vs THE REALITY

PovArInd	Known Population	Sample Estimate	Gender	Known Population	Sample Estimate
Poverty	536207	1507352	Male	3503378	3018151
Nonpoverty	6554845	5583700	Female	3587674	4072901

CREATE INDICATOR VARIABLES

```
data Mortality; set Mortality;
   Poverty=0; NonPoverty=0; Male=0; Female=0;
   if (Gender=1)   then Male      =1;
   if (Gender=2)   then Female    =1;
   if (PovArInd=1) then Poverty   =1;
   if (PovArInd=2) then NonPoverty=1;
run;
```

CALIBRATION

```
%SurveyCalibrate(
   DATA           = Mortality,
   OUT            = Final,
   METHOD         = TRUNLINEAR,
   WEIGHT         = SWeight,
   CONTROLVAR     = Poverty  NonPoverty Male     Female,
   CTRLTOTAL      = 536207   6554845    3503378  3587674,
   UPPER          = 2.0,
   VARMETHOD      = bootstrap,
   SEED           = 100,
   CLUSTER        = VarPSU,
   STRATA         = VarStrata
);
```

MACRO MESSAGES

- NOTE: After 7 iterations, the lower bound is set to LOWER=0.3522109375 for the TRUNLINEAR method.
- NOTE: The calibration weights Cal_Sweight are created by using the TRUNLINEAR method with LOWER=0.3522109375 and UPPER=2 bounds.

THE RESULTS

PovArInd	Known Population	Sample Estimate	After Calibration	Gender	Known Population	Sample Estimate	After Calibration
Poverty	536207	1507352	536207	Male	3503378	3018151	3503378
Non-poverty	6554845	5583700	6554845	Female	3587674	4072901	3587674

WEIGHT CHANGES

```
data Final; set Final;
   wt_change=Cal_SWeight/SWeight;
proc surveymeans data = Final
   min max quartiles;
   var wt_change;
run;
```

WEIGHT CHANGES

Quantiles						
Variable	Percentile		Estimate	Std Error	95% Confidence Limits	
wt_change	0	Min	0.352211	.	.	.
	25	Q1	0.352211	0.002115	0.34803605	0.35638583
	50	Median	0.361594	0.170975	0.02412915	0.69905895
	75	Q3	1.036222	0.079973	0.87837322	1.19407140
	100	Max	1.351921	.	.	.

ANALYSIS

	Using Original Weights		Using Calibration Weights	
Variable	**Mean**	**Std Error**	**Mean**	**Std Error**
Age	65.073909	0.949498	65.126584	1.155297
VitalStatus=1	0.644459	0.034795	0.659089	0.036309
VitalStatus=3	0.267700	0.028865	0.270262	0.029592
VitalStatus=4	0.034766	0.011432	0.026019	0.016890
VitalStatus=5	0.016649	0.012291	0.012743	0.013272
VitalStatus=6	0.036426	0.028146	0.031887	0.028144

DISCUSSION

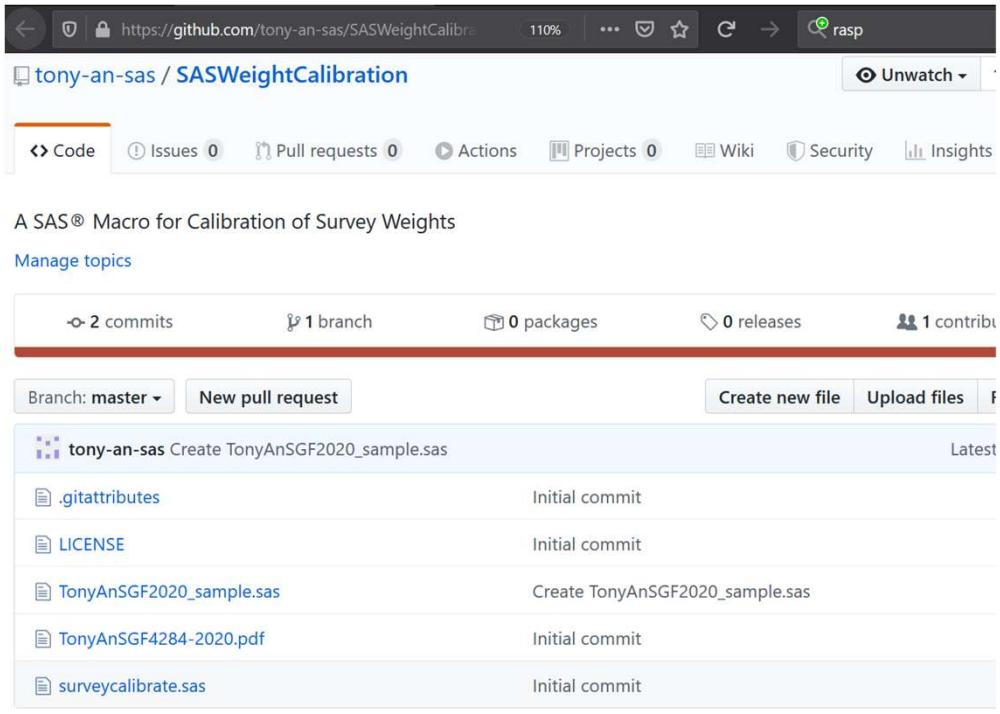
- Flexible and easy to use
- Accommodates most calibration needs
- Provides replicate weights for future analyses

DISCUSSION

- No magic rule for choosing a calibration method
- Use the linear method first
- Try the exponential method if the linear method fails
- Compromise with truncated linear or logit methods at the price of computation resources
- Experiment with different settings

HOW TO GET THE MACRO

https://github.com/tony-an-sas/SASWeightCalibration



Thank you!

Contact Information
tony.an@sas.com

Reminder:

Complete your session survey in the conference mobile app.


SAS®
GLOBAL
FORUM
2020
USERS PROGRAM
MARCH 29 – APRIL 1 | WASHINGTON, DC | #SASGF