UnderStandingAmericaStudy

WEIGHTING PROCEDURE, V1

USC Dornsife Center for Economic and Social Research

04/08/2016

INTRODUCTION

This document provides details of the weighting procedures and benchmark distributions used to create final sample weights for data sets collected by the Center for Economic and Social Research's Understanding America Study internet panel.[1]

This version (April 2016) is superseded by the new version published on the UAS website of September 2017, which is described here.

1 SAMPLING

The UAS is a panel of US households recruited through Address Based Sampling (ABS). Eligible individuals are all adults in the contacted household aged 18 and older. In this context a household is broadly defined as anyone living together with the initial person who signed up to become an UAS member. The UAS also includes a special purpose sample of Native Americans, recruited through ABS, targeting zip-codes with a higher proportion of Native Americans. In this case, eligible individuals are all adults in the contacted household aged 18 and older, whose ethnicity is Native American. Finally, a special purpose sample of families with young children in Los Angeles County is part of the study. This sample is recruited using birth records information from the State of California in order to target addresses where children were born during the past five years.

2 WEIGHTING

In the UAS, sample weights are survey-specific. They are provided with each UAS survey and are meant to make each survey data set representative of the U.S. population aged 18 and older with respect to a pre-defined set of socio-demographic variables. Sample weights are constructed in two steps. In a first step, a *base weight* is assigned to each survey respondent in order to compensate for the disproportionate sampling of Native Americans. In a second step, *post-stratification weights* are generated to bring the final survey sample in line with the reference population as far as the distribution of key variables of interest is concerned.

[1] Mick Couper and Jon Krosnick have provided insightful and valuable comments throughout the development of the UAS weighting procedure.

The sub-sample of families with young children in Los Angeles County receives no weight as it is not possible to determine the corresponding population counts accurately.

2.1 Categorization and Imputation of Variables

For weighting purposes, we use demographic information taken from the most recent "My Household" survey, which is answered by the respondent every quarter. With the exception of age and number of household members, all other socio-demographic variables in the "My Household" survey are categorical and some, such as education and income, take values in a relatively large set. We recode all the variables used in the weighting procedure into new categorical variables with no more than 5 categories. The aim of limiting the categories is to prevent these variables from forming strata containing a very small fraction of the sample (less than 4-5%), which may cause sample weights to exhibit considerable variability. The list of recoded categorical variables used in the weighting procedure is reported in Table 1.

Table 1: List of Recoded Categorical Variables Used within the Weighting Procedure

Recoded Variable	Categories
gender	1. Male; 2. Female
age_cat	1. 18-34; 2. 35/44; 3. 45/54; 4. 55-64; 5. 65+
bornus	0. No; 1. Yes
citizenus	0. No; 1. Yes
marital_cat	1. Married; 2. Separated/Divorced/Widowed; 3. Never Married
education_cat	1. High School or Less; 2. Some College; 3. Assoc. College Degree; 4. Bachelor; 5. Master/Professional/Doctorate Degree
education_cat2	1. High School or Less; 2. Some College; 3. Bachelor or More
hisplatino	0. No; 1. Yes
race_cat	1. White; 2. Black; 3. Others; 4. Hispanic
work_cat	1. Working; 2. Unemployed; 3. Retired; 4. On leave, Disabled, Other
work_cat2	1. Working; 2. Unemployed/On Leave/Disabled; 3. Retired
work_cat3	1. Working; 2. Not Working

hhmembers_cat	1. One Member; 2. Two Members; 3. Three or Four Members; 4. Five or More Members
hhmembers_cat2	1. One Member; 2. Two or Three Members; 3. Four or More Members
hhincome_cat	1. <$30,000; 2. $30,000-$59,999; 3. $60,000-$99,999; 4. $100,000+
hhincome_cat2	1. <$35,000; 2. $35,000-74,999; 3. $75,000+

Before implementing the weighting procedure, we employ the following imputation scheme to replace missing values of recoded socio-demographic variables.

- We do not impute gender. Hence, respondents with missing gender are not assigned a sample weight. The fraction of respondents who do not report their gender in the UAS is on the order of 0.2%.

- When actual age is missing, the variable *agerange*, available in the "My Household" survey, is used to impute *age_cat*. If *agerange* is also missing, the variable *age_cat* is assigned the mode for males or females, depending on the respondent's gender.

- For binary indicators, such as *bornus*, *citizenus*, and *hisplatino*, missing values are imputed using a logistic regression.

- For ordered categorical variables, such as *education_cat*, *education_cat2*, *hhmembers_cat*, *hhmembers_cat2*, *hhincome_cat*, and *hhincome_cat2*, missing values are imputed using an ordered logistic regression.

- For non-ordered categorical variables, such as *marital_cat*, *race_cat*, and *work_cat*, missing values are imputed using a multinomial logistic regression.

Imputations are performed sequentially. That is, once *age_cat* has been imputed (if missing), the variable with the smallest number of missing values is the first one to be imputed by means of a regression featuring *gender* and *age_cat* as regressors. This newly imputed variable is then added to the set of regressors to impute the variable with the second smallest number of missing values. The procedure continues in this fashion until the variable with the most missing values (typically household income) is imputed using information on all other socio-demographic variables.

Each weighted UAS survey data set contains a binary variable, *imputation_flag*, indicating whether any of the recoded socio-economic variables listed in Table 1 and used within the weighting procedure has been imputed.

2.2 Step 1: Base Weights

In this first step, a base weight is assigned to each respondent to adjust for the disproportionate stratification of Native Americans in the UAS pool of respondents. Let $f^{s}_{non-native}$ and f^{s}_{native}, be the relative frequencies of non-Native Americans and Native Americans in the survey sample, respectively, and $f^{p}_{non-native}$ and f^{p}_{native}, the relative frequencies of non-Native Americans and Native Americans in the Census population, respectively. For a respondent i, the base weight is defined as:

$$w_{i,base} = \begin{cases} \dfrac{f^{p}_{non-native}}{f^{s}_{non-native}} & \text{if } i \text{ is not Native American} \\ \dfrac{f^{p}_{native}}{f^{s}_{native}} & \text{if } i \text{ is Native American} \end{cases}$$

Each weighted UAS survey data set includes base weights relative to their sample mean. That is:

$$relw_{i,base} = \frac{w_{i,base}}{\left(\frac{1}{N}\sum_{i=1}^{N} w_{i,base}\right)},$$

where N is the survey sample size.

These relative base weights, average to 1 and sum to the survey sample size N. They are stored in the variable *base_weight*.

2.3 Step 2: Post-stratification Weights

The execution of the sampling process for a survey is typically less than perfect. Even if the sample of panel members invited to take a survey is representative of the population along a series of dimensions, the sample of actual respondents may exhibit discrepancies because of differences in response rates across groups and/or other issues related to the fielding time and content of the survey. A second layer of weighting is therefore needed to align the final survey sample to the reference population as far as the distribution of key variables is concerned. In this second step, we perform **iterative marginal weighting** and assign survey respondents weights such that the weighted distributions of specific socio-demographic variables in the survey sample match their population counterparts (benchmark or target distributions).

The benchmark distributions against which UAS surveys are weighted are derived from the Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC) administered in March of each year. Depending on when the to-be-weighted UAS survey was collected, the following timing rule identifies the specific CPS-ASEC used by the weighting algorithm to construct target distributions:

Data Collection of UAS Survey	CPS-ASEC Used by the Weighting Algorithm
September 2013 – August 2014	2013
September 2014 – August 2015	2014
September 2015 –	2014*

* Until the 2015 CPS-ASEC becomes available.

Unless otherwise required by the aims of the survey and specified in the sample selection process, the reference population for UAS surveys is the U.S. population of those aged 18 and older, excluding institutionalized individuals and military personnel.

We adopt a **raking algorithm** to generate post-stratification weights. This procedure involves the comparison of target population relative frequencies and actually achieved sample relative frequencies on a number of socio-demographic variables independently and sequentially. More precisely, starting from the base weights as described in section 2.2, at each iteration of the algorithm weights are proportionally adjusted so that the distance between survey and population marginal distributions of each selected socio-demographic variable (or raking factor) decreases. The algorithm stops when survey and population distributions are perfectly aligned. A maximum of 50 iterations is allowed for perfect alignment of survey and population distributions to be achieved. If the process does not converge within 50 iterations, no sample weights are returned and attempts using different raking factors are made.

2.4 Trimming

Our raking algorithm trims extreme weights in order to limit variability and improve efficiency of estimators. We follow the general weight trimming and redistribution procedure described by Valliant, Dever and Kreuter (2013). Specifically, indicating with $w_{i,raking}$ the raking weight for respondent i and with $\overline{w}_{raking} = \frac{1}{N}\sum_{i=1}^{N} w_{i,raking}$ the sample average of raking weights,

I. We set the lower and upper bounds on weights equal to $L = 0.25\overline{w}_{raking}$ and $U = 4\overline{w}_{raking}$, respectively. While these values are arbitrary, they are in line with those described in the literature and followed by other surveys (Izrael, Battaglia and Frankel, 2009).

II. We reset any weights smaller than the lower bound to L and any weights greater than the upper bound to U:

$$w_{i,trim} = \begin{cases} L & w_{i,raking} \leq L \\ w_{i,raking} & L < w_{i,raking} < U \\ U & w_{i,raking} \geq U \end{cases}$$

III. We compute the amount of weight lost by trimming as $w_{lost} = \sum_{i=1}^{N} w_{i,raking} - w_{i,trim}$ and distribute it evenly among the respondents whose weights are not trimmed.

While raking weights can match population distributions of selected variables, trimmed weights typically do not. We therefore iterate the raking algorithm and the trimming procedure until a set of post-stratification weights is obtained that respect the weight bounds and align sample and population distributions of selected variables. This procedure stops after 50 iterations if an exact alignment respecting the weight bounds cannot be achieved. In this case, the trimmed weights will ensure the exact match between survey and population relative frequencies, but may take values outside the interval defined by the pre-specified lower and upper bounds.

2.5 Final Post-stratification Weights

Indicate with $w_{i,post}$ the post-stratification weight for respondent i, obtained by applying the raking algorithm to the base weights and after iterating the raking algorithm and the trimming procedure as described above in section 2.4.

Each weighted UAS survey data set includes post-stratification weights relative to their sample mean. That is:

$$relw_{i,post} = \frac{w_{i,post}}{\left(\frac{1}{N}\sum_{i=1}^{N} w_{i,post}\right)},$$

where N is the survey sample size.

These relative post-stratification weights, average to 1 and sum to the survey sample size N. They are stored in the variable *rel_weight*.

3 DEFAULT WEIGHTS

Raking can be performed on *one-way marginals*, by matching population distributions of single socio-demographic variables, such as *gender* or *education_cat*, as well as on *two-way marginals*, by matching the distributions of interaction variables, such as *gender x education_cat*. The set of raking factors may feature both single and interaction variables, such as, for instance, *gender* and *race_cat x education_cat*. The use of two-way marginals corrects for discrepancies between distributions referring to specific sub-groups that would not be accounted for by using one-way marginals alone. As an example, suppose that discrepancies in the distribution of educational attainment by gender are observed and need to be corrected. If raking is done using the single variables *gender* and *education_cat*, the resulting weights allow matching the distributions of gender and educational attainment for the entire sample, but not necessarily the distributions of educational attainment for men and women separately. In contrast, implementing the raking algorithm on the interaction variable *gender x education_cat* ensures that the distributions of educational attainment for men and women are matched to their population counterparts. Moreover, since two-way marginals subsume one-way marginals, using the interaction variable *gender x education_cat* also guarantees that the distributions of gender and education for the entire sample are matched to their population counterparts.

By default, UAS surveys are weighted using the following set of raking factors:

- *race_cat*
- *gender* x *age_cat*
- *gender* x *education_cat*
- *hhmembers_cat2* x *hhincome_cat2*

After extensive testing, raking weights produced by this combination of factors have been found to perform well across different dimensions. In particular, they exhibit moderate variability, thereby leading to better precision of weighted estimates, and allow matching the distributions of variables not used as raking factors in a satisfactory manner, thereby improving overall representativeness. Monte Carlo studies have shown that these desirable properties are robust to sample sizes ranging from 500 to 2,000 respondents, an interval that includes most of the UAS surveys.

For each UAS survey, default weights can be obtained by sending a request to uas-weights-l@mymaillists.usc.edu. For completed surveys, the data set with default weights is available for download on the UAS webpage.

4 CUSTOM WEIGHTS

Data users can customize the weighting procedure and obtain weights that better suit the goals of their research and data analysis. Custom weights can be obtained by choosing which socio-demographic variables should be used by the raking algorithm to generate post-stratification weights. Data users who want to obtain custom weights should send their request to uas-weights-l@mymaillists.usc.edu, alongside with their preferred set of raking factors. This set is **limited to a maximum of 5 variables** (single variables, interaction variables or a combination of both) selected from the pre-defined list in Table 2. Restrictions on the number and type of raking factors are imposed to ensure convergence of the algorithm and to reduce weight variability.

Table 2: List of Raking Factors for Custom Weights

Single Variables*
gender, age_cat, bornus, citizenus, marital_cat, education_cat, education_cat2, hisplatino, race_cat, work_cat, work_cat2, work_cat3, hhmembers_cat, hhmembers_cat2, hhincome_cat, hhincome_cat2
Interaction Variables
gender x age_cat, gender x bornus, gender x marital_cat, gender x education_cat, gender x education_cat2, gender x hisplatino, gender x race_cat, gender x work_cat, gender x work_cat2, gender x work_cat3, age_cat x work_cat3, gender x hhmembers_cat, gender x work_cat2 x hhincome_cat2hhmembers_cat2, gender x hhincome_cat, gender x hhincome_cat2, hhmembers_cat2 x hhincome_cat2

*Refer to Table 1 for the definition of single categorical variables.

5 WEIGHTING OUTPUT

Each weighted UAS survey data set includes the following variables:

- *imputation_flag*
  A binary variable indicating whether any of the variables used within the weighting procedure has been imputed.

- *base_weight*
  Relative base weights correcting for the over-representation of Native Americans in the survey sample. They average to one and sum to the UAS survey sample size.

- *rel_weight*
  Relative post-stratification weights which ensure representativeness of the survey sample with respect to key selected variables (raking factors). They include the correction for the over-representation of Native Americans. They average to one and sum to the UAS survey sample size.