Weighting the Daybreak Poll

Erik Meijer

July 25, 2016

1 Overview

As with nearly all survey data, we use weights to ensure that the sample matches the population of interest on a number of characteristics that are potentially related to outcomes of interest. Characteristics that are matched are sex, age, race-ethnicity, education, household size, family income, and voting in the previous presidential election (2012). The weighting method largely follows the default weighting procedure of the UAS (USC, 2016), but with a few differences. It is also similar to the method we used for the successful RAND Continuous 2012 Presidential Election Poll (Gutsche, Kapteyn, Meijer, & Weerman, 2014; Kapteyn, Meijer, & Weerman, 2012), but again, small differences in the poll design led to small differences in the weighting procedure.

The weights are constructed using *raking*, which is a common method in survey research (see, e.g., Lu & Gelman, 2003; Valliant, Dever, & Kreuter, 2013, pp. 358–361; Kolenikov, 2014). Specifically, this procedure ensures that the following distributions are exactly the same in the weighted poll data as in the reference population:

- **Race-ethnicity:** (1) Non-Hispanic white alone; (2) Non-Hispanic African American alone or in combination; (3) Other non-Hispanic; (4) Hispanic.
- **Sex by education:** The cross-classification of sex and education. Sex has categories (0) Male and (1) Female. Education has categories (1) High school or less; (2) Some college but no degree; (3) Associate's degree; (4) Bachelor's degree; (5) Advanced degree. The combination of these two has $2 \times 5 = 10$ categories.
- **Sex by age:** The cross-classification of sex and age. Age (in years) has categories (0) 18–21; (1) 22–34; (2) 35–44; (3) 45–54; (4) 55–64; (5) 65+. To keep the weights (more) stable in the course of the poll, we define age using election day (November 8, 2016) as the reference date in the sample, so that individuals do not shift categories between waves. A minor exception to this date is the bound between categories 0 and 1, which uses November 4, 2016. The reason is that individuals born after November 4, 1994 were not age-eligible to vote in the previous election (2012), whereas those who were born on November 4, 1994 or earlier were age-eligible. This cross-classification results in 12 categories.

- **Household size by income:** The cross-classification of household size and family income. The former has categories (1) Single household; (2) 2 or 3 persons; (3) 4 or more persons. The latter has categories (1) Less than $35,000; (2) $35,000–$74,999; and (3) $75,000 or more. This cross-classification would result in $3 \times 3 = 9$ categories, but for single households, we combine income groups (2) and (3) into one category ($35,000 or more), so the number of categories matched is 8.

- **Prior voting:** The best predictor of future voting behavior is past voting behavior, and any discrepancies in composition with respect to past voting behavior thus are likely to give biased predictions of voting behavior in the 2016 election. Therefore, we use voting in the 2012 presidential election as an additional reference characteristic. This has the following categories: (1) 18–21 years old, not age-eligible to vote in 2012; (2) Age eligible in 2012, but did not vote; (3) Voted for Mitt Romney; (4) Voted for Barack Obama; and (5) Voted for someone else.

Except for the prior voting variable, these characteristics are measured in the UAS through the "MyHousehold" survey, which respondents are asked to complete quarterly. Each record in the election poll survey data contains the variables from most recent (at the time the poll survey was completed by the respondent) MyHousehold survey. Panel members who are U.S. citizens were invited to complete a pre-election survey (UAS47). Those who are old enough to have been age-eligible for voting in the 2012 presidential election were asked whether they voted in 2012, and if so, for whom they voted.

In our preparation of the RAND 2012 Continuous Presidential Election Poll, we found that members of RAND's American Life Panel were very accurate in their reporting of their voting four years earlier: More than 90% of the reports in 2012 about voting in 2008 coincided with their reports immediately after the 2008 election, for those panel members that participated in both surveys (Gutsche et al., 2014; Kapteyn et al., 2012).

2 Reference distributions

Reference distributions of the socio-economic and demographic variables were obtained from the May 2016 basic monthly data of the Current Population Survey (CPS), obtained from `http://thedataweb.rm.census.gov/ftp/cps_ftp.html`, which were the most recent available when we prepared the weighting procedure. We used the script from the National Bureau of Economic Research (`http://www.nber.org/data/cps_basic_progs.html`) to read them into Stata. We restricted the sample to U.S. citizens age 18 and over and used the CPS's person weight (`pwsswgt`) to obtain the most accurate estimates. The appendix presents the frequency distributions used.

We use the following information to compute the reference distribution of the prior voting variable:

- As estimated from the CPS, 6.90% of U.S. citizens 18 and over are 18–21 years old.
- Voter turnout in 2012 was obtained from the estimates provided by the United States Election Project (McDonald, 2014). Translating this to a reference percentage is somewhat inexact, because there are slight differences in the population that is covered by the UAS versus the population that is eligible to vote. Specifically, the former includes individuals who are ineligible to vote because of their criminal history, whereas the latter includes U.S. citizens who live outside the U.S. We do not know about the criminal history of our panel members, but individuals who would have been ineligible in 2012 should say that they did not vote. Hence, we include these in the denominator. Thus, after weighting, we would like the poll to be as representative as possible for the population consisting of the voting age population minus non-U.S. citizens but including individuals living elsewhere who are eligible to vote (2.13%). Based on these considerations, we decided to use the following calculation for the denominator:

Voting age population	240,957,993
Non-citizens (8.4%)	−20,240,471
Overseas eligible	5,127,418
Total	225,844,940

  For the numerator, we use the number of valid votes in the 2012 presidential election, which was 129,067,662 (FEC, 2013). Thus, we use 129,067,662/225,844,940 = 57.15% as the fraction who voted, and 100 − 57.15 = 42.85% as the fraction who did not vote. These fractions are applied to the population who was old enough to vote in 2012, that is, to 100 − 6.90 = 93.10% of our population of interest. Consequently, the target fraction of our population who did not vote in 2012 but was old enough is 42.85% of 93.10%, or 39.90%.
- Among the valid votes for president in 2012, Mitt Romney received 47.21%, Barack Obama received 51.06%, and all other candidates jointly received 1.73% (FEC, 2013). These fractions are applied to the population who voted, which from the previous points we compute as 57.15% of 93.10%. Hence, the target fraction for Mitt Romney voters is $47.21\% \times 57.15\% \times 93.10\% = 25.12\%$, and similarly for Obama and other candidates.

Table 5 in the appendix summarizes the complete frequency distribution used.

3 Practical issues

We use the Stata user-written package `survwgt` (Winter, 2015) to compute the weights. Each day's poll sample (a combined sample of the last 7 days) is separately weighted, as are the samples for additional analyses (e.g., analyses that focus on the additional questions, which use slightly different samples). Compared to the default UAS weighting procedure (USC, 2016), we use a few simplifications:

- We do not trim weights. This may lead to larger standard errors, but less bias, and this makes the weighting program faster and more stable (less likely to run into technical problems), which is desirable for an automated program that is run every night.

- Because we exclude panel members who were recruited as part of the special Native American subpanel (as well as the Los Angeles County subpanel), there is no need to weight in two steps. Note that Native Americans and Los Angeles County residents who were recruited as part of the nationally representative recruiting batches are eligible for the poll, so excluding the special subpanels does not compromise representativeness. Panel members from these subpanels are participating in the poll surveys, so their answers can be used for analyses that focus on these populations, but they are not included in the poll results. This excludes a little less than 10% of the observations, but the recruiting strategy for the LA County subpanel does not currently allow us to compute meaningful weights (see USC, 2016 for more discussion about this) and the Native American subpanel would receive very low weights and therefore contribute very little numerically. The practical advantage of simplifying the weighting program outweighs the negligible addition in precision for the purposes of this poll.

- We do not impute missing weighting variables. If any of the variables used in the weighting process is missing, the observation is excluded from the sample. Again, this has practical advantages. It leads to a loss of about 3% of the sample, about evenly split between panel members whose household size we do not know and panel members who did not remember whether they voted in 2012 or whom they voted for. There are a few other missings, but not many. We find this 3% acceptable.

Computing proper standard errors for a complex weighted survey is nontrivial, and simplified methods (e.g., treating the weights as known inverse probabilities of selection) typically lead to biased standard errors; see, for example, Lu and Gelman (2003). To compute correct standard errors, we use the Stata package `bsweights` (Kolenikov, 2010) with the aforementioned package `survwgt` to compute a set of 80 *replication weights*, which cluster by respondent. Stata's "`svy`" facilities (StataCorp, 2011) then use these to compute clustered survey bootstrap standard errors. Specifically, we use the `svy: mean` and `svy: ratio` commands (with the `over` option for estimates by subgroup) and `lincom` for computing the standard errors of differences, specifically differences between results for Clinton and Trump.

References

FEC. (2013). *Official 2012 presidential general election results.* Washington, DC: Federal Election Commission. Retrieved from `http://www.fec.gov/pubrec/fe2012/2012presgeresults.pdf`

Gutsche, T. L., Kapteyn, A., Meijer, E., & Weerman, B. (2014). The RAND Continuous 2012 Presidential Election Poll. *Public Opinion Quarterly*, *78*, 233–254. doi: 10.1093/poq/nfu009

Kapteyn, A., Meijer, E., & Weerman, B. (2012). *Methodology of the RAND Continuous 2012 Presidential Election Poll* (Working Paper No. WR-961). Santa Monica, CA: RAND Corporation. doi: 10.2139/ssrn.2146149

Kolenikov, S. (2010). Resampling variance estimation for complex survey data. *Stata Journal*, *10*, 165–199.

Kolenikov, S. (2014). Calibrating survey data using iterative proportional fitting (raking). *Stata Journal*, *14*, 22–59.

Lu, H., & Gelman, A. (2003). A method for estimating design-based sampling variances for surveys with weighting, poststratification, and raking. *Journal of Official Statistics*, *19*, 133–151.

McDonald, M. P. (2014). *2012 november general election turnout rates.* Gainesville, FL: United States Elections Project. Retrieved from `http://www.electproject.org/2012g`

StataCorp. (2011). *Stata survey data reference manual: Release 12*. College Station, TX: StataCorp.

USC. (2016). *UnderStandingAmericaStudy: Weighting procedure.* Los Angeles, CA: USC Dornsife Center for Economic and Social Research. Retrieved from `https://uas.usc.edu/documents/uas/UAS%20Weighting%20Procedures.pdf`

Valliant, R., Dever, J. A., & Kreuter, F. (2013). *Practical tools for designing and weighting survey samples*. New York, NY: Springer.

Winter, N. (2015). *SURVWGT: Stata module to create and manipulate survey weights.* Retrieved from `http://EconPapers.repec.org/RePEc:boc:bocode:s427503`

A Appendix: Reference distributions

This appendix presents the reference distributions, as derived from the CPS, vote counts, and other sources, that are used as target distributions in the weighting procedure.[1] Table 1 shows race-ethnicity, Table 2 shows sex × education, Table 3 shows sex × age, Table 4 shows household size × income, and Table 5 shows prior voting.

Table 1: Reference distribution of race-ethnicity

Category	Percent
1.Non-hispanic white alone	68.90
2.Non-hispanic black alone or in combination	12.81
3.Other non-hispanic	6.34
4.Hispanic	11.95

Table 2: Reference distribution of sex by education

Category	Percent
01.Male; ≤High School	19.62
02.Male; Some College	9.22
03.Male; Associate	4.34
04.Male; Bachelor	9.52
05.Male; Advanced	5.37
11.Female; ≤High School	19.52
12.Female; Some College	10.36
13.Female; Associate	5.63
14.Female; Bachelor	10.48
15.Female; Advanced	5.95

[1] More precisely, we use the exact estimated population counts and not these (rounded) percentages.

Table 3: Reference distribution of sex by age

Category	Percent
00.Male; 18–21	3.52
01.Male; 22–34	11.05
02.Male; 35–44	7.43
03.Male; 45–54	8.43
04.Male; 55–64	8.40
05.Male; 65+	9.23
10.Female; 18–21	3.38
11.Female; 22–34	11.35
12.Female; 35–44	7.86
13.Female; 45–54	8.87
14.Female; 55–64	9.09
15.Female; 65+	11.39

Table 4: Reference distribution of household size by income

Category	Percent
11. 1 <$35,000	8.55
12. 1; ≥$35,000	7.10
21. 2–3; <$35,000	14.78
22. 2–3; $35,000–$74,999	18.54
23. 2–3; ≥$75,000	20.92
31. 4+; <$35,000	5.81
32. 4+; $35,000–$74,999	8.91
33. 4+; ≥$75,000	15.38

Table 5: Reference distribution of prior voting

Category	Percent
1.Too young	6.90
2.Did not vote	39.90
3.Romney	25.12
4.Obama	27.17
5.Someone else	0.92