postweight or calibrate? Survey post-adjustments in Stata

Pablo Cabrera Álvarez (USAL)
Modesto Escobar Mercado (USAL)

2017 Spanish Stata Users Group meeting

Madrid, 19 octubre 2017
Survey post-adjustments

- Non-response and coverage errors affect survey estimates
- Some population members are not in the sample frame (e.g. households with no internet connection in CAWI)
- Respondents may differ from non-respondents on key characteristics
- Weighting (e.g. postestratification, calibration) is a form of post-adjustment that aims to rebalance the sample using auxiliary information (Biemer & Christ, 2008)
Postestratification and calibration

- Both methods use auxiliary information in order to force the sample distribution to match the population on key auxiliary variables.
- “Postestratification could refer to any method of data analysis which involves forming units into homogeneous groups after observation of the sample” (Smith, 1991)
- Calibration doesn’t include any or some of the interactions; It is a relaxed form of postestratification (Zhang, 2000; Särndal, 2007)
Postestratification in Stata

- `svyset` is the Stata command for complex survey analysis and it allows to compute postestratification weights in the background:

```
svyset, poststrata(varname) postweight(varname)
```

- `poststrata` refers to a variable which identifies the cases in the same poststrata
- `postweight` refers to the population size of each poststratum
- The poststratification weight is computed for each estimate based on the sample of valid cases
Calibration in Stata

- *calibrate* (D’Souza, 2010) is a command to compute calibration weights:

```stata
calibrate, marginals(varlist) poptot(matrix) entrywt(varname) exitwt(varname)
```

- The *calibrate* command generates a calibration weight which forces the sample to match the population distribution on the *marginals* variables.
- The survey estimate is computed each time applying the weight *svyset, [pweight = calibwt]*
Data and methods

- Barometer 3183 (July 2017) from Centre for Sociological Research. Multistage clustered sample. Age and sex quotas to select final respondents.
- Information about past vote and voting intention available; those with no valid past vote were excluded.
- Three steps: 1) Rebalancing the sample to match the 2016 election results; 2) Estimating voting intention variables using calibration weights and svy poststratification; 3) Comparing estimates.
### Postweight vs calibrate on auxiliary variable

Past vote estimate using postweight and calibrate (%):

<table>
<thead>
<tr>
<th>Past vote</th>
<th>2016 elec.</th>
<th>Raw</th>
<th>Postwt.</th>
<th>Calib.</th>
<th>wt</th>
</tr>
</thead>
<tbody>
<tr>
<td>PP</td>
<td>22.95</td>
<td>23.54</td>
<td>22.95</td>
<td>22.95</td>
<td>0.98</td>
</tr>
<tr>
<td>PSOE</td>
<td>15.74</td>
<td>22.08</td>
<td>15.74</td>
<td>15.74</td>
<td>0.71</td>
</tr>
<tr>
<td>UP</td>
<td>14.71</td>
<td>16.97</td>
<td>14.71</td>
<td>14.71</td>
<td>0.87</td>
</tr>
<tr>
<td>Cs</td>
<td>9.08</td>
<td>9.49</td>
<td>9.08</td>
<td>9.08</td>
<td>0.96</td>
</tr>
<tr>
<td>ERC</td>
<td>1.83</td>
<td>2.92</td>
<td>1.83</td>
<td>1.83</td>
<td>0.63</td>
</tr>
<tr>
<td>Another party</td>
<td>5.53</td>
<td>8.62</td>
<td>5.53</td>
<td>5.53</td>
<td>0.64</td>
</tr>
<tr>
<td>Didn’t vote</td>
<td>30.17</td>
<td>16.38</td>
<td>30.17</td>
<td>30.17</td>
<td>1.84</td>
</tr>
</tbody>
</table>
Postweight vs calibrate on target variables (I)

<table>
<thead>
<tr>
<th></th>
<th>Raw (%)</th>
<th>Postweight (%)</th>
<th>Calibrate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Would vote</td>
<td>84.95</td>
<td>79.64</td>
<td>79.64</td>
</tr>
<tr>
<td>Wouldn’t vote</td>
<td>15.05</td>
<td>20.36</td>
<td>20.36</td>
</tr>
</tbody>
</table>
### Postweight vs calibrate on target variables (II)

<table>
<thead>
<tr>
<th></th>
<th>Raw (%)</th>
<th>Postweight (%)</th>
<th>Calibrate (%)</th>
<th>Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PP</td>
<td>25.60</td>
<td>26.64</td>
<td>28.37</td>
<td>-1.73</td>
</tr>
<tr>
<td>PSOE</td>
<td>29.69</td>
<td>27.46</td>
<td>26.90</td>
<td>0.56</td>
</tr>
<tr>
<td>UP</td>
<td>20.71</td>
<td>20.42</td>
<td>20.91</td>
<td>-0.49</td>
</tr>
<tr>
<td>Cs</td>
<td>14.16</td>
<td>16.82</td>
<td>16.12</td>
<td>0.70</td>
</tr>
<tr>
<td>ERC</td>
<td>3.87</td>
<td>2.77</td>
<td>2.94</td>
<td>-0.17</td>
</tr>
<tr>
<td>PDCat</td>
<td>0.85</td>
<td>0.83</td>
<td>0.66</td>
<td>0.18</td>
</tr>
<tr>
<td>PNV</td>
<td>1.37</td>
<td>1.28</td>
<td>1.02</td>
<td>0.26</td>
</tr>
<tr>
<td>Another party</td>
<td>3.75</td>
<td>3.78</td>
<td>3.09</td>
<td>0.69</td>
</tr>
</tbody>
</table>
## Comparing weights

<table>
<thead>
<tr>
<th>Past vote</th>
<th>Calibrate wt (full sample)</th>
<th>Postweight wt (if vote)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PP</td>
<td>0.98</td>
<td>0.87</td>
</tr>
<tr>
<td>PSOE</td>
<td>0.71</td>
<td>0.64</td>
</tr>
<tr>
<td>UP</td>
<td>0.87</td>
<td>0.76</td>
</tr>
<tr>
<td>Cs</td>
<td>0.96</td>
<td>0.90</td>
</tr>
<tr>
<td>ERC</td>
<td>0.63</td>
<td>0.53</td>
</tr>
<tr>
<td>Another party</td>
<td>0.64</td>
<td>0.84</td>
</tr>
<tr>
<td>Didn’t vote</td>
<td>1.84</td>
<td>3.12</td>
</tr>
</tbody>
</table>
Wrap-up

- Both *Postweight* and *Calibrate* are tools for rebalancing the sample.
- *Postweight* works in the background recalculating the weights for each estimate based on the valid sample but the same population totals; *Calibrate* computes a weight to force the sample to match the population, this weight is used for estimates.
- *Postweight* can only be used for general sample estimates while the weight produced by *Calibrate* can be used for general and subsample estimates.
- The estimation procedures using *Postweight* [svy poststrata and postweight options] or *Calibrate* [svy pweight option] lead to different standard errors.
** Preliminar

scalar Population = 34596892

matrix Weights = (7941236, 5443846, 5087538, 3141570, 632234, 1911558, 10438910)

matrix colnames Weights = PP PSOE UP Cs ERC Others DontVote

matrix rownames Weights = Population
Syntax (II)

** Postweight

```stata
gen poptotal = 0
foreach num of numlist 1/7 {
    replace poptotal=Weights[1,`num'] if VarPost==`num'
}
```

```
svyset _n, poststrata(VarPost) postweight(poptotal)
svy: prop Variable
```
** Calibrate (I)
quietly: sum VarPost
gen start = Population/r(N)
tab VarPost, gen(ValuesPost)
calibrate, marginals(ValuesPost1-ValuesPost7) ///
   poptot(Weights) ent(start) exit(wtcal) method(logistic)
svyset cues [pweight=wtcal]
svy: prop Variable
** Calibrate (II)**

quietly: sum VarPost if demost == 1

gen start2 = Population/\(r(N)\) if demost == 1

calibrate, `marginals`(ValuesPost1-ValuesPost7) ///

    `poptot`(Weights) `ent`(start2) `exit`(wtcal2) `method`(logistic)

quietly: sum wtcal2, d

replace wtcal2=wtcal2/\(r(\text{mean})\)

svyset cues [pweight=wtcal2]

svy: prop Variable


