

---

Subject: Weighting data after merging survey rounds with different levels of representation

Posted by [jswindle](#) on Wed, 11 Jan 2017 06:11:07 GMT

[View Forum Message](#) <> [Reply to Message](#)

---

Hello DHS experts and other forum users,

My problem relates to weighting after merging survey rounds.

I have merged the IR for the Malawi 2000, 2004, and 2010 surveys. When I weight this data, I am unsure how to best create the strata given different levels of representation across the surveys. The 2000 and 2004 surveys were representative across the 10 largest districts and across the three regions. The 2010 survey was representative across all 27 districts, as well as across the three regions.

I am combining these survey data with district level, time-varying data for foreign aid.

Currently I am doing the following, but am unsure about the commands. I am especially looking for guidance about the fifth through seventh lines below, which begin with "egen mw\_00\_strata..."

Thank you.

\* Weight the dataset

```
generate weight = v005/10000000
```

```
recode survey (2000=1) (2004=2) (2010=3)
```

```
egen clusters=group(survey v021), label
```

```
egen mw_00_strata = group(survey region urban), label
```

```
egen mw_04_strata = group(survey region urban), label
```

```
egen mw_10_strata = group(survey region district urban), label
```

```
gen strata = .
```

```
replace strata = mw_00_strata if year==2000
```

```
replace strata = mw_04_strata if year==2004
```

```
replace strata = mw_10_strata if year==2010
```

```
svyset clusters [pweight=weight], strata(strata) singleunit(centered)
```

---

Subject: Re: Weighting data after merging survey rounds with different levels of representation

Posted by [Bridgette-DHS](#) on Wed, 11 Jan 2017 23:56:27 GMT

[View Forum Message](#) <> [Reply to Message](#)

---

Following is a response from Senior DHS Stata Specialist, Tom Pullum:

Unfortunately, the stratification variable is often incorrect or labelled incorrectly in surveys conducted before about 2010. The following lines will work if you change the path.

set more off

```

set maxvar 10000

cd e:\DHS\DHS_data\IR_files

use MWIR41FL.dta, clear
gen survey=1
append using MWIR4DFL.dta
replace survey=2 if survey==.
append using MWIR61FL.dta
replace survey=3 if survey==.

* The strata in MW41 are given by s006
* The strata in MW4D are given by group(sdist v025)
* The strata in MW61 are given by v022

gen mw_00_strata = s006
egen mw_04_strata = group(sdist v025), label
gen mw_10_strata = v022

gen strata_temp=.
replace strata_temp=mw_00_strata if survey==1
replace strata_temp=mw_04_strata if survey==2
replace strata_temp=mw_10_strata if survey==3

egen strata=group(survey strata_temp)

tab strata survey, table clean

```

---

Subject: Re: Weighting data after merging survey rounds with different levels of representation

Posted by [jswindle](#) on Thu, 12 Jan 2017 02:20:24 GMT

[View Forum Message](#) <> [Reply to Message](#)

---

Hello Tom and Bridgett,

Thank you for your very helpful and prompt reply.

I ran the code you shared and got some results that surprised me. When I ran the final command of "tab strata survey, table clean" I got an error message saying that I could not use those options. When I instead ran "tab strata survey", I got these interesting results:

```
tab strata survey
```

```
group(survey)

```

strata\_tem survey

p) 1 2 3 Total

1	226	0	0	226
2	416	0	0	416
3	607	0	0	607
4	320	0	0	320
5	17	0	0	17
6	253	0	0	253
7	190	0	0	190
8	470	0	0	470
9	27	0	0	27
10	447	0	0	447
11	767	0	0	767
12	174	0	0	174
13	556	0	0	556
14	172	0	0	172
15	438	0	0	438
16	433	0	0	433
17	611	0	0	611
18	187	0	0	187
19	459	0	0	459
20	195	0	0	195
21	340	0	0	340
22	800	0	0	800
23	105	0	0	105
24	121	0	0	121
25	450	0	0	450
26	331	0	0	331
27	186	0	0	186
28	193	0	0	193
29	28	0	0	28
30	188	0	0	188
31	435	0	0	435
32	185	0	0	185
33	239	0	0	239
34	67	0	0	67
35	22	0	0	22
36	591	0	0	591
37	193	0	0	193
38	787	0	0	787
39	95	0	0	95
40	614	0	0	614
41	285	0	0	285
42	0	420	0	420
43	0	283	0	283
44	0	47	0	47
45	0	850	0	850

46 0 40 0 40  
47 0 732 0 732  
48 0 81 0 81  
49 0 693 0 693  
50 0 263 0 263  
51 0 690 0 690  
52 0 78 0 78  
53 0 625 0 625  
54 0 31 0 31  
55 0 789 0 789  
56 0 101 0 101  
57 0 705 0 705  
58 0 307 0 307  
59 0 403 0 403  
60 0 42 0 42  
61 0 735 0 735  
62 0 230 0 230  
63 0 3,553 0 3,553  
64 0 0 92 92  
65 0 0 754 754  
66 0 0 825 825  
67 0 0 318 318  
68 0 0 33 33  
69 0 0 789 789  
70 0 0 35 35  
71 0 0 786 786  
72 0 0 60 60  
73 0 0 718 718  
74 0 0 45 45  
75 0 0 821 821  
76 0 0 32 32  
77 0 0 781 781  
78 0 0 138 138  
79 0 0 650 650  
80 0 0 76 76  
81 0 0 832 832  
82 0 0 480 480  
83 0 0 646 646  
84 0 0 53 53  
85 0 0 723 723  
86 0 0 55 55  
87 0 0 746 746  
88 0 0 44 44  
89 0 0 786 786  
90 0 0 41 41  
91 0 0 823 823  
92 0 0 127 127  
93 0 0 668 668

94 0 0 197 197  
 95 0 0 755 755  
 96 0 0 29 29  
 97 0 0 706 706  
 98 0 0 42 42  
 99 0 0 778 778  
 100 0 0 70 70  
 101 0 0 747 747  
 102 0 0 81 81  
 103 0 0 737 737  
 104 0 0 63 63  
 105 0 0 831 831  
 106 0 0 37 37  
 107 0 0 782 782  
 108 0 0 35 35  
 109 0 0 767 767  
 110 0 0 90 90  
 111 0 0 761 761  
 112 0 0 66 66  
 113 0 0 723 723  
 114 0 0 85 85  
 115 0 0 778 778  
 116 0 0 137 137  
 117 0 0 746 746

Total 13,220 11,698 23,020 47,938

The part of these results that I found surprising is that the number of strata per survey vary in strange way. There are 41 categories for 2000, 22 categories for 2004, and 54 categories for 2010. The result for 2010 makes sense; there were 27 districts and when stratified by urban/rural you get 54. The result for 2004, I believe comes from 11 districts categories stratified by urban/rural; those 11 district categories are the ten largest districts that were sampled in a representative manner and then there is one big catch-all for the other 17 districts, hence the huge total of 3,553 respondents in the catch-all rural category (at least that is my guess). The 2000 results are perplexing. From what I can gather in the final report for the 2000 Malawi DHS, the sampling was done in the same manner as the 2004 survey, so I'm not sure why there are 41 categories here. Thoughts?

Once I have calculate the strata correctly, would the rest of this code (pasted below) work to appropriately survey set the data?

```

generate weight = v005/10000000
egen clusters=group(survey v021), label
svyset clusters [pweight=weight], strata(strata) singleunit(centered)
  
```

Or would you simply do:

```
generate weight = v005/10000000  
svyset [pweight=weight], psu(v021) strata(strata)
```

In case it is relevant for deciding how to svyset the data, my ultimate goal is to do a three-level mixed effects model with the higher orders being the districtyear and district variables.

A final issue I am facing if I do this sort of mixed effects model is whether the 2000 and 2004 data from the 17 districts that are not sampled sufficiently to be representative could be appropriately incorporated into such a model. I realize that is outside the purview of the DHS surveys, but I'm guessing you have faced these types of issue before in your own research. Any thoughts?

thank you kindly,  
Jeff

---

Subject: Re: Weighting data after merging survey rounds with different levels of representation

Posted by [Bridgette-DHS](#) on Fri, 13 Jan 2017 16:45:52 GMT

[View Forum Message](#) <> [Reply to Message](#)

---

Following is a response from Senior DHS Stata Specialist, Tom Pullum:

Sorry about that mistake--"table clean" is an option with "list" and not with "tab". Don't know why I added that--it definitely would not run.

You definitely need something like "egen clusters=group(survey v021)" to get unique identifiers for the clusters across the three surveys and the strata across the three surveys.

The change you observe in the number of strata from one survey to the next is not implausible (although usually the definitions are the same from one survey to the next.) In general, the number of strata is in the range of 20 to 60. I suggest you look at the report. I HOPE it will confirm what I passed on to you.

DHS estimates at the stratum level are always representative, in terms of being unbiased. It is true that some older documentation mentioned a lack of representativeness, but that actually refers just to higher standard errors when there are fewer cases. That can be an issue, but bias is NOT an issue. The sampling is designed so that small strata tend to be over-sampled (large strata correspondingly tend to be under-sampled) in order to get more stable estimates.

The generic term we use for the first national sub-division is "region". The generic term for the second level is "district". In general the strata are the combinations of region and urban/rural.

---