16) Survey Data - Weighting, Clustering, and Stratification

Vitor Kamada

August 2018

2004 Medical Expenditure Panel Survey (MEPS)

Sample design includes: stratification, clustering, multiple stages of selection, and disproportionate sampling

Sampling weights reflect adjustments:

1) for survey nonresponse

2) to population control totals from the Current Population Survey

Clusters and Stratification

Primary Sampling Units (PSUs): clusters at the first level of sampling

Stratification: partitions the population into distinct groups, often by demographic variables

Weights

Sampling Weight: Inverse of the probability of being included in the sample

Hispanics and blacks are oversampled at rates of 2 and 1.5 times

Oversampled observations will have lower weights than undersampled

Survey Design Setup

use http://www.stata-press.com/data/heus/heus_mepssample

egen clusterid=group(varpsu varstr)

list wtdper varstr varpsu clusterid race_bl eth_hisp in 4/11

	wtdper	varstr	varpsu	cluste~d	race_bl	eth_hisp
4.	4760.71	43	2	236	Black race	Not Hispanic
5.	15117.74	83	1	81	Not black race	Not Hispanic
6.	6243.832	8	2	203	Not black race	Not Hispanic
7.	11900.83	8	2	203	Not black race	Not Hispanic
8.	8889.356	109	3	422	Not black race	Hispanic
9.	6122.333	151	1	146	Not black race	Not Hispanic
10.	6830.828	151	1	146	Not black race	Not Hispanic
11.	14023.06	9	2	204	Not black race	Not Hispanic

svyset [pweight=wtdper], strata(varstr) psu(varpsu)

Alternative Cluster and Weight Options

```
generate race_bl_pct = race_bl*100
quietly mean exp_tot race_bl_pct
estimates store noadjust
quietly mean exp_tot race_bl_pct, vce(cluster clusterid)
estimates store cluster
quietly mean exp_tot race_bl_pct [pw=wtdper]
estimates store weights
quietly mean exp_tot race_bl_pct [pw=wtdper], vce(cluster clusterid)
estimates store clust_wgt
quietly svy: mean exp_tot race_bl_pct
estimates store survey
estimates table *, b(%7.1f) modelwidth(9)
```

Variable	noadjust	cluster	weights	clust_wgt	survey
exp_tot	3685.2	3685.2	3838.9	3838.9	3838.9
race_bl_pct	13.8	13.8	10.8	10.8	10.8

mean exp_tot, over(race_bl)

test [exp_tot]_subpop_1 = [exp_tot]_subpop_2

```
Mean estimation                         Number of obs    =      19,386
    _subpop_1: race_bl = Not black race
    _subpop_2: race_bl = Black race
```

Over	Mean	Std. Err.	[95% Conf.	Interval]
exp_tot				
_subpop_1	3730.723	77.56493	3578.689	3882.757
_subpop_2	3401.788	154.1159	3099.707	3703.868

```
. test [exp_tot]_subpop_1 = [exp_tot]_subpop_2
 ( 1)  [exp_tot]_subpop_1 - [exp_tot]_subpop_2 = 0
       F(  1, 19385) =    3.63
            Prob > F =    0.0566
```

svy: mean exp_tot, over(race_bl)

test [exp_tot]_subpop_1 = [exp_tot]_subpop_2

```
Number of strata =       203        Number of obs    =        19,386
Number of PSUs   =       448        Population size  =   187,973,715
                                    Design df        =           245

    _subpop_1: race_bl = Not black race
    _subpop_2: race_bl = Black race
```

Over	Mean	Linearized Std. Err.	[95% Conf.	Interval]
exp_tot				
_subpop_1	3926.681	110.3909	3709.244	4144.117
_subpop_2	3114.646	168.0148	2783.708	3445.583

```
. test [exp_tot]_subpop_1 = [exp_tot]_subpop_2

Adjusted Wald test

 ( 1)  [exp_tot]_subpop_1 - [exp_tot]_subpop_2 = 0

       F(  1,   245) =   15.87
            Prob > F =    0.0001
```

Regression: Alternative Cluster and Weight Options

```
quietly regress exp_tot age i.female i.race_bl i.reg_south, vce(robust)
estimates store robust
quietly regress exp_tot age i.female i.race_bl i.reg_south, ///
    vce(cluster clusterid)
estimates store cluster
quietly regress exp_tot age i.female i.race_bl i.reg_south [pw=wtdper]
estimates store weights
quietly regress exp_tot age i.female i.race_bl i.reg_south [pw=wtdper], ///
    vce(cluster clusterid)
estimates store clust_wgt
quietly svy: regress exp_tot age i.female i.race_bl i.reg_south
estimates store survey
estimates table *, b(%7.2f) se(%7.2f) p(%7.4f) modelwidth(9) drop(_cons) ///
    title(Alternative cluster and weight options: Linear regression estimates)
```

Linear Regression Estimates

Variable	robust	cluster	weights	clust_wgt	survey
age	129.13	129.13	129.91	129.91	129.91
	4.68	4.52	5.62	5.33	5.25
	0.0000	0.0000	0.0000	0.0000	0.0000
female					
Female	895.51	895.51	769.09	769.09	769.09
	138.29	134.52	188.06	184.53	188.00
	0.0000	0.0000	0.0000	0.0000	0.0001
race_bl					
Black race	-92.19	-92.19	-317.78	-317.78	-317.78
	167.13	167.38	182.56	196.65	188.59
	0.5812	0.5820	0.0818	0.1068	0.0933
reg_south					
South	-256.71	-256.71	-325.40	-325.40	-325.40
	132.97	145.14	164.93	163.80	158.92
	0.0536	0.0776	0.0485	0.0476	0.0417

legend: b/se/p

Second National Health and Nutrition Examination Survey (NHANES II)

McDowell et al (1981)

```
use http://www.stata-press.com/data/r15/nhanes2f
svyset psuid [pweight=finalwgt], strata(stratid)
sum finalwgt stratid psuid
```

Variable	Obs	Mean	Std. Dev.	Min	Max
finalwgt	10,337	11320.85	7304.457	2000	79634
stratid	10,337	16.65986	9.499389	1	32
psuid	10,337	1.482151	.4997055	1	2

logistic highbp height weight age female black

Logistic regression

Log likelihood = **-5819.4894**

Number of obs	= **10,337**
LR chi2(**5**)	= **2444.67**
Prob > chi2	= **0.0000**
Pseudo R2	= **0.1736**

highbp	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
height	.9652624	.0035406	-9.64	0.000	.9583478	.9722269
weight	1.050648	.0019344	26.83	0.000	1.046864	1.054447
age	1.048466	.0015337	32.35	0.000	1.045465	1.051477
female	.6806283	.0438292	-5.97	0.000	.5999247	.7721884
black	1.43975	.1063157	4.94	0.000	1.245753	1.663959
_cons	.8882276	.5547718	-0.19	0.849	.261143	3.021135

svy: logistic highbp height weight age female black

Number of strata	=	31	Number of obs	=	10,337
Number of PSUs	=	62	Population size	=	117,023,659
			Design df	=	31
			F(5, 27)	=	278.72
			Prob > F	=	0.0000

highbp	Odds Ratio	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
height	.9661196	.0051213	-6.50	0.000	.955731	.9766212
weight	1.052379	.0027142	19.79	0.000	1.046857	1.057929
age	1.05057	.002046	25.33	0.000	1.046405	1.054751
female	.6227382	.0353865	-8.34	0.000	.554592	.6992578
black	1.399584	.1495559	3.15	0.004	1.125512	1.740395
_cons	.6635987	.5628155	-0.48	0.632	.1176734	3.742251

svy, subpop(female): logistic highbp height weight age black

Number of strata	=	31	
Number of PSUs	=	62	
Number of obs	=	10,337	
Population size	=	117,023,659	
Subpop. no. obs	=	5,428	
Subpop. size	=	60,901,624	
Design df	=	31	
F(4, 28)	=	166.32	
Prob > F	=	0.0000	

highbp	Odds Ratio	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
height	.9639434	.0073091	-4.84	0.000	.9491511	.9789662
weight	1.051365	.0036191	14.55	0.000	1.04401	1.058773
age	1.067257	.0034789	19.97	0.000	1.060185	1.074376
black	1.63308	.2377472	3.37	0.002	1.213551	2.197641
_cons	.2958584	.3449688	-1.04	0.304	.0274351	3.190519