(R)
Statistics/Data analysis

User: Albert Lutakome
Project: Survey Analysis

©
17.0
MP—Parallel Edition

Statistics and Data Science

Copyright 1985-2021 StataCorp LLC
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC https://www.stata.com
979-696-4600 stata@stata.com

Stata license: Unlimited-user 64-core network perpetual
Serial number: 18461036
  Licensed to: TEAM BTCR
               TEAM BTCR

Notes:
1. Unicode is supported; see help unicode_advice.
2. More than 2 billion observations are allowed; see help obs_advice.
3. Maximum number of variables is set to 100,000; see help set_maxvar.

```
 1 . doedit "C:\Users\alber\OneDrive\Documents\Github Projects\P1\Ethiopia DHS Survey Data Analysis.do"

 2 . do "C:\Users\alber\AppData\Local\Temp\STD4ddc_000000.tmp"

 3 .
 4 . *1.  ********************** Merge Women and Household memebrs datasets**********************
 5 .      *********This will introuduce household characteristics to the Women dataset************
 6 .
 7 . *a.Open secondary dataset, and sort by ID variable
 8 . use "C:\Users\alber\OneDrive\Documents\Github Projects\P1\ETPR71FL.DTA", clear

 9 . sort hhid  // sort by ID Variable

10 .
11 . *b. Save temporary file of just the variables to merge in
12 . tempfile secondary_HHD

13 . save "`secondary_HHD'", replace
   (file C:\Users\alber\AppData\Local\Temp\ST_4ddc_000001.tmp not found)
   file C:\Users\alber\AppData\Local\Temp\ST_4ddc_000001.tmp saved as .dta format

14 .
15 . *c. open primary file * i.e. Women dataset, and sort by ID Variable and merge
16 . use "C:\Users\alber\OneDrive\Documents\Github Projects\P1\ETIR71FL.DTA", clear

17 . gen hhid =substr(caseid,1,12) // changed ID variabe name to match ID variable name in second file

18 . sort hhid

19 .
20 . merge m:m hhid using "`secondary_HHD'"

    Result                      Number of obs
    -----------------------------------------
    Not matched                        12,435
        from master                         0  (_merge==1)
        from using                     12,435  (_merge==2)

    Matched                            62,789  (_merge==3)
    -----------------------------------------
```

```
21 . drop if _merge ==1 // drops unmatched from master
(0 observations deleted)

22 . drop if _merge ==2 // drops unmatched from using
(12,435 observations deleted)

23 .
24 . tab _merge //check if all is well
```

Matching result from merge	Freq.	Percent	Cum.
Matched (3)	**62,789**	**100.00**	**100.00**
Total	**62,789**	**100.00**	

```
25 .
26 .
27 . save "C:\Users\alber\OneDrive\Documents\Github Projects\P1\WM_HHM_Merged.DTA", replace
(file C:\Users\alber\OneDrive\Documents\Github Projects\P1\WM_HHM_Merged.DTA not found)
file C:\Users\alber\OneDrive\Documents\Github Projects\P1\WM_HHM_Merged.DTA saved as .dta format

28 .
29 . *2  ********** Setting survey parameters for complex survey design  and installing tabout for table production ******
30 .
31 . ssc install tabout
checking tabout consistency and verifying not already installed...
all files already exist and are up to date.

32 .
33 . gen wt=hv005/1000000  //generating weight variable

34 . egen strata=group(v024 v025)

35 . * svyset [pw=x], psu(y) strata(z), where pw stands for probability weight, x = weight variable, y = cluster variable,
> = strata variable.
36 . svyset [pw=wt], psu(v021) strata(v022) singleunit(centered)

Sampling weights: wt
             VCE: linearized
     Single unit: centered
        Strata 1: v022
 Sampling unit 1: v021
           FPC 1: <zero>

37 .
38 . *3. ****************************  Demographics CALCULATION *************
39 .
40 . * a. Total population
41 . gen pop=0.

42 . replace pop=1 if hv001 >0
(62,789 real changes made)

43 . label variable pop "idividual women found"   // we created an individual value 1 foreach identified case and thats how
> e were able to calculate total population
```

```
44 . gen totpop=sum(pop)

45 . su totpop  // summing the total population of women
```

Variable	Obs	Mean	Std. dev.	Min	Max
totpop	62,789	31395	18125.77	1	62789

```
46 .
47 . * b. Dissagregation by subgroups
48 .
49 .      *1. urban  vs rural
50 .
51 . tab hv025 [iweight=wt]
```

type of place of residence	Freq.	Percent	Cum.
urban	10,404.265	15.69	15.69
rural	55,914.406	84.31	100.00
Total	66,318.672	100.00	

```
52 .
53 .      *2. population by region
54 . tab  v024 [iweight=wt]
```

region	Freq.	Percent	Cum.
tigray	4,451.724	6.71	6.71
afar	539.182359	0.81	7.53
amhara	14,787.455	22.30	29.82
oromia	26,101.062	39.36	69.18
somali	2,185.1379	3.29	72.48
benishangul	675.584222	1.02	73.49
snnpr	14,482.031	21.84	95.33
gambela	158.529294	0.24	95.57
harari	142.270151	0.21	95.78
addis adaba	2,460.7696	3.71	99.49
dire dawa	334.926599	0.51	100.00
Total	66,318.672	100.00	

```
55 .
56 .      *3. Women's age dissagregation
57 . gen m_agewm=.
   (62,789 missing values generated)

58 . replace m_agewm=1 if hv105<5
   (10,103 real changes made)

59 . replace m_agewm=2 if hv105 >=5  & hv105<15
   (19,681 real changes made)
```

```
60 . replace m_agewm=3 if hv105 >=15 & hv105<49
   (27,558 real changes made)

61 . replace m_agewm=4 if hv105 >=49 & hv105<95
   (5,401 real changes made)

62 . replace m_agewm=5 if hv105 >=95 & hv105 != 98
   (27 real changes made)

63 . replace m_agewm=6 if hv105 == 98
   (19 real changes made)

64 .
65 . label variable m_agewm "Age group of Female Household Member"

66 . label define m_agewm 1 "<5" 2 "5-15" 3 "15-49" 4 "49-95" 5 "95+" 6 "don't know"

67 . label values m_agewm m_agewm

68 .
69 . tab m_agewm
```

Age group of Female Household Member	Freq.	Percent	Cum.
<5	**10,103**	**16.09**	**16.09**
5-15	**19,681**	**31.34**	**47.44**
15-49	**27,558**	**43.89**	**91.32**
49-95	**5,401**	**8.60**	**99.93**
95+	**27**	**0.04**	**99.97**
don't know	**19**	**0.03**	**100.00**
Total	**62,789**	**100.00**	

```
70 .
71 .      *4 Urban vs rural population distribution by age groups
72 . svy: tab m_agewm hv025, per // pop age urban vs rural
   (running tabulate on estimation sample)

   Number of strata =  25                            Number of obs   =     62,789
   Number of PSUs   = 643                            Population size = 66,318.672
                                                     Design df       =        618
```

Age group of Female Household Member	type of place of residence: urban	rural	Total
<5	**1.794**	**14.29**	**16.09**
5-15	**3.518**	**28.74**	**32.25**
15-49	**8.886**	**34.08**	**42.96**
49-95	**1.48**	**7.148**	**8.628**
95+	**.0034**	**.0452**	**.0486**
don't kn	**.006**	**.0156**	**.0216**
Total	**15.69**	**84.31**	**100**

```
   Key: Cell percentage

     Pearson:
       Uncorrected   chi2(5)         = 1028.0867
       Design-based  F(3.43, 2119.61)=   97.2083     P = 0.0000
```

```
73 .
74 . *4. ***************************Descriptive statistics **********************
75 .         *a.  mean, median age per age group
76 . table m_agewm, statistic(mean hv105) statistic(median hv105) // mean and median age per age group
```

	Mean	Median
Age group of Female Household Member		
<5	**2.01168**	**2**
5-15	**9.306082**	**9**
15-49	**28.29643**	**27**
49-95	**60.66673**	**60**
95+	**95**	**95**
don't know	**98**	**98**
Total	**20.94886**	**16**

```
77 .
78 .         *b. Min, Max age per age group
79 . table m_agewm, statistic(min hv105) statistic(max hv105)
```

	Minimum value	Maximum value
Age group of Female Household Member		
<5	**0**	**4**
5-15	**5**	**14**
15-49	**15**	**48**
49-95	**49**	**94**
95+	**95**	**95**
don't know	**98**	**98**
Total	**0**	**98**

```
80 .
81 .
82 . * 5. ******************Computing Indicators *****************
83 .
84 . *a. wealth index
85 . gen wealthwm=.
(62,789 missing values generated)

86 . replace wealthwm=0 if inlist(hv270,1,2)
(27,632 real changes made)

87 . replace wealthwm=1 if inlist(hv270,3,4,5)
(35,157 real changes made)

88 . label variable wealthwm "Women Wealth Group"

89 . label define wealthwm 0 "Poor and Poorest" 1 "Middle, rich and richest"

90 . label values wealthwm wealthwm
```

```
91 .
92 .      *1. By residence
93 . svy: tab wealthwm hv025, per  // wealth index by residence
(running tabulate on estimation sample)

Number of strata =  25                            Number of obs   =     62,789
Number of PSUs   = 643                            Population size = 66,318.672
                                                  Design df       =        618
```

Women Wealth Group	type of place of residence: urban	rural	Total
Poor and	.8285	37.91	38.74
Middle,	14.86	46.4	61.26
Total	15.69	84.31	100

Key: **Cell percentage**

```
  Pearson:
    Uncorrected   chi2(1)         = 5510.5795
    Design-based  F(1, 618)       =  128.4439     P = 0.0000

94 .      *2. By region
95 . svy: tab wealthwm hv024, per  // wealth index by region
(running tabulate on estimation sample)

Number of strata =  25                            Number of obs   =     62,789
Number of PSUs   = 643                            Population size = 66,318.672
                                                  Design df       =        618
```

Women Wealth Group	region: tigray	afar	amhara	oromia	somali	benishan	snnpr	gambela	harari	addis ad	dire daw	Total
Poor and	3.252	.6242	7.975	15.33	2.507	.5149	8.233	.0982	.0558	0	.1488	38.74
Middle,	3.46	.1888	14.32	24.03	.7879	.5038	13.6	.1409	.1587	3.711	.3563	61.26
Total	6.713	.813	22.3	39.36	3.295	1.019	21.84	.239	.2145	3.711	.505	100

Key: **Cell percentage**

```
  Pearson:
    Uncorrected   chi2(10)        = 3285.3304
    Design-based  F(3.55, 2194.81)=   15.1441     P = 0.0000

96 .
97 .      *3. By gender
98 . svy: tab  v024 hv270,per
(running tabulate on estimation sample)

Number of strata =  25                            Number of obs   =     62,789
Number of PSUs   = 643                            Population size = 66,318.672
                                                  Design df       =        618
```

region	wealth index combined: poorest	poorer	middle	richer	richest	Total
tigray	1.873	1.379	1.09	.7196	1.651	6.713
afar	.6064	.0178	.0126	.0184	.1578	.813
amhara	3.515	4.46	5.097	5.507	3.718	22.3
oromia	6.645	8.683	8.543	8.71	6.777	39.36
somali	2.221	.2859	.1941	.1695	.4243	3.295
benishan	.2851	.2297	.1812	.1879	.1347	1.019
snnpr	3.756	4.476	5.07	5.209	3.325	21.84
gambela	.0754	.0228	.0205	.0298	.0906	.239
harari	.0215	.0343	.0238	.0229	.1121	.2145
addis ad	0	0	0	.0019	3.709	3.711
dire daw	.0912	.0576	.0341	.0166	.3055	.505
Total	19.09	19.65	20.27	20.59	20.4	100

Key: **Cell percentage**

  Pearson:
    Uncorrected   chi2(**40**)   = **1.53e+04**
    Design-based F(**9.82, 6071.65**)=   **24.4525**   P = **0.0000**

```
99 . tab  v024 hv270 [iweight=wt]
```

region	wealth index combined: poorest	poorer	middle	richer	richest	Total
tigray	1,241.939	914.83632	722.86381	477.20691	1,094.878	4,451.724
afar	402.14263	11.801963	8.37834	12.231225	104.6282	539.18236
amhara	2,331.194	2,957.861	3,380.1668	3,652.4062	2,465.827	14,787.45
oromia	4,407.079	5,758.449	5,665.324	5,776.0794	4,494.13	26,101.06
somali	1,473.018	189.5967	128.736365	112.4195	281.36762	2,185.138
benishangul	189.09556	152.36382	120.186537	124.59802	89.340284	675.58422
snnpr	2,491.054	2,968.668	3,362.4	3,454.7503	2,205.16	14,482.03
gambela	50.013279	15.102504	13.5968051	19.741497	60.075208	158.52929
harari	14.255755	22.734849	15.7748968	15.157989	74.346662	142.27015
addis adaba	0	0	0	1.283928	2,459.486	2,460.77
dire dawa	60.485523	38.180124	22.612544	11.025937	202.62247	334.9266
Total	12,660.28	13,029.594	13,440.04	13,656.9	13,531.86	66,318.67

```
100 .
101 . * b. Access to Education
102 .
103 . tab hv025 v149
```

type of place of residence	educational attainment: no educat	incomplet	complete	incomplet	complete	higher	Total
urban	3,356	5,105	968	3,657	678	2,605	16,369
rural	28,204	13,899	894	2,929	65	429	46,420
Total	31,560	19,004	1,862	6,586	743	3,034	62,789

```
104 .
105 . gen eduwm=.
    (62,789 missing values generated)

106 . replace eduwm=0 if v149==0
    (31,560 real changes made)

107 . replace eduwm=1 if inlist(v149,1,2,3,4,5)
    (31,229 real changes made)

108 . label variable eduwm "Highest Education"

109 . label define eduwm 1 "Above Primary" 0 "Below Primary"

110 . label values eduwm eduwm

111 .
112 .       *1. By residence
113 . svy: tab eduwm hv025, per
    (running tabulate on estimation sample)

    Number of strata =  25                          Number of obs   =     62,789
    Number of PSUs   = 643                          Population size = 66,318.672
                                                    Design df       =        618

    ------------------------------------
                |  type of place of
                |      residence
    Highest     |
    Education   | urban  rural  Total
    ------------+-----------------------
     Below Pr   | 2.956  49.05     52
     Above Pr   | 12.73  35.27     48
                |
        Total   | 15.69  84.31    100
    ------------------------------------
    Key: Cell percentage

      Pearson:
        Uncorrected   chi2(1)         = 5147.2558
        Design-based  F(1, 618)       =  354.6145     P = 0.0000

114 .
115 .
116 . * c. Overcrowding conditions/ living space
117 .
118 . gen room_crowd=.
    (62,789 missing values generated)

119 . replace hv012 = hv013 if hv012 == 0  // if dejure members (HV012) is 0 then  hv013 (de facto members) = 0 as well.
    (8 real changes made)

120 .
121 . replace room_crowd = hv012 if hv216 == 0  // if the number of rooms for sleeping (HV216) is 0 then all de jure members
    > ave no sufficient living => hv216 = 0.
    (0 real changes made)
```

```
. 
. replace room_crowd = (hv012 / hv216) if hv216 != 0 // if number of rooms is not 0,then person per room = persons/numbe
> of rooms.
(62,789 real changes made)

. 
. replace room_crowd = 98 if room_crowd >= 98  // Accounting for Invalid entires and missing values, entires with 98.
(0 real changes made)

. 
. 
. * Now Calculating living space indicator
. gen living_space = 1

. *As per standards, if persons per room is greator than 3, the no sufficient living space, hence, living1=0. I will mov
> above code up to see if there is change.
. replace living_space = 0 if room_crowd > 3
(41,316 real changes made)

. label variable living_space "Overcrowding Conditions"

. label define living_space 1 "Sufficient Living Space" 0 "Over crowded"

. 
.        *1. Computing women living in overcrowded conditions by age group
. 
. tab living_space m_agewm [iweight=wt]
```

Overcrowding Conditions	Age group of Female Household Member						
	<5	5-15	15-49	49-95	95+	don't kno	Total
0	7,843.018	16,239.96	16,818.58	3,213.847	22.784547	8.21485291	44,146.4
1	2,824.64	5,150.839	11,673.06	2,508.199	9.441473	6.094197	22,172.27
Total	10,667.66	21,390.8	28,491.636	5,722.046	32.22602	14.30905	66,318.67

```
. svy: tab living_space m_agewm, per
(running tabulate on estimation sample)

Number of strata =  25                        Number of obs   =     62,789
Number of PSUs   = 643                        Population size = 66,318.672
                                              Design df       =        618
```

Overcrowding Conditions	Age group of Female Household Member						
	<5	5-15	15-49	49-95	95+	don't kn	Total
0	11.83	24.49	25.36	4.846	.0344	.0124	66.57
1	4.259	7.767	17.6	3.782	.0142	.0092	33.43
Total	16.09	32.25	42.96	8.628	.0486	.0216	100

```
Key: Cell percentage

  Pearson:
    Uncorrected   chi2(5)         = 1968.2350
    Design-based  F(4.57, 2823.85)=  178.9794     P = 0.0000
```

```
139 .
140 .        *2. Computing women living in overcrowded conditions by age group
141 .
142 . svy: tab living_space v102, per
(running tabulate on estimation sample)

Number of strata =  25                          Number of obs   =     62,789
Number of PSUs   = 643                          Population size = 66,318.672
                                                Design df       =        618

------------------------------------------------
Overcrowd |
ing       |  type of place of
Condition |      residence
s         | urban  rural  Total
----------+-------------------------------------
        0 | 6.086  60.48  66.57
        1 | 9.602  23.83  33.43

    Total | 15.69  84.31    100
------------------------------------------------
Key: Cell percentage

  Pearson:
    Uncorrected   chi2(1)         = 4049.0700
    Design-based  F(1, 618)       =  251.6530     P = 0.0000

143 .
144 . *6 ********************Regression and variable associations
145 .
146 .         *a. Studying the significance of association of residence, sex,education and wealth in relation to overcrowdin
147 .
148 . * Here we analyse the results of the ch2 test from the cross tabulation below.
149 .
150 . tabout living_space hv025 [iw=wt] using "residencech2.xls",c(col) f(1) stats(chi2) svy nwt(wt) per pop replace // disa
> egation by residence
Survey results being calculated
----------+---------- 1 ----------+---------- 2 ----------+---------- 3 ----------+---------- 4 ----------+---------- 5
.......
Table output written to: residencech2.xls

        type of place of residence
Overcrowding Conditions urban   rural   Total   N
        %       %       %
0       38.8    71.7    66.6    44,146
1       61.2    28.3    33.4    22,172
Total   100.0   100.0   100.0   66,319

Pearson: Uncorrected chi2(1) = 4049.0700
Design-based F(1.00, 618.00) = 251.6530 Pr = 0.000

151 . tabout living_space hv024 [iw=wt] using "regionch2.xls",c(col) f(1) stats(chi2) svy nwt(wt) per pop replace // disagre
> tion by region
Survey results being calculated
----------+---------- 1 ----------+---------- 2 ----------+---------- 3 ----------+---------- 4 ----------+---------- 5
.......
Table output written to: regionch2.xls

        region
Overcrowding Conditions tigray  afar    amhara  oromia  somali  benishangul     snnpr   gambela harari  addis adaba
> re dawa       Total   N
        %       %       %       %       %       %       %       %       %       %       %       %
0       60.5    73.9    62.6    72.2    83.1    68.7    64.8    60.3    66.7    35.6    66.3    66.6    44,146
1       39.5    26.1    37.4    27.8    16.9    31.3    35.2    39.7    33.3    64.4    33.7    33.4    22,172
Total   100.0   100.0   100.0   100.0   100.0   100.0   100.0   100.0   100.0   100.0   100.0   100.0   66,319
```

Pearson: Uncorrected chi2(10) = 1813.3149
>
Design-based F(3.60, 2223.65) = 19.3672 Pr = 0.000
>

```
152 . tabout living_space v149 [iw=wt] using "educationch2.xls",c(col) f(1) stats(chi2) svy nwt(wt) per pop replace // disag
    > gation by education
    Survey results being calculated
    ──────┼── 1 ──┼── 2 ──┼── 3 ──┼── 4 ──┼── 5
    .......
    Table output written to: educationch2.xls
```

Overcrowding Conditions	educational attainment: no education	incomplete primary	complete primary	incomplete secondary	complete secondary	higher	Total	N
	%	%	%	%	%	%	%	
0	74.7	64.9	59.4	45.0	32.8	24.9	66.6	44,146
1	25.3	35.1	40.6	55.0	67.2	75.1	33.4	22,172
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	66,319

Pearson: Uncorrected chi2(5) = 4039.0055
Design-based F(4.62, 2855.80) = 79.6863 Pr = 0.000

```
153 . tabout living_space hv270 [iw=wt] using "wealthch2.xls",c(col) f(1) stats(chi2) svy nwt(wt) per pop replace // disagre
    > tion by wealth
    Survey results being calculated
    ──────┼── 1 ──┼── 2 ──┼── 3 ──┼── 4 ──┼── 5
    .......
    Table output written to: wealthch2.xls
```

Overcrowding Conditions	wealth index combined: poorest	poorer	middle	richer	richest	Total	N
	%	%	%	%	%		
0	85.3	77.6	71.3	57.5	42.9	66.6	44,146
1	14.7	22.4	28.7	42.5	57.1	33.4	22,172
Total	100.0	100.0	100.0	100.0	100.0	100.0	66,319

Pearson: Uncorrected chi2(4) = 6382.5644
Design-based F(3.74, 2311.03) = 123.5880 Pr = 0.000

```
154 . tabout living_space hv219 [iw=wt] using "hheadch2.xls",c(col) f(1) stats(chi2) svy nwt(wt) per pop replace // disagreg
    > ion by hhold head
    Survey results being calculated
    ──────┼── 1 ──┼── 2 ──┼── 3 ──┼── 4 ──┼── 5
    .......
    Table output written to: hheadch2.xls
```

Overcrowding Conditions	sex of head of household: male	female	Total	N
	%	%		
0	69.4	53.5	66.6	44,146
1	30.6	46.5	33.4	22,172
Total	100.0	100.0	100.0	66,319

Pearson: Uncorrected chi2(1) = 1026.8819
Design-based F(1.00, 618.00) = 82.7481 Pr = 0.000

```
155 .
156 .         *b. Logistic regression model to study weather wealth is explained by residence, education and sex:
157 .
158 . logit wealthwm  i.hv025 i.eduwm i.hv104 i.hv025

Iteration 0:   log likelihood = -43070.013
Iteration 1:   log likelihood = -33616.429
Iteration 2:   log likelihood =  -33196.13
Iteration 3:   log likelihood = -33186.966
Iteration 4:   log likelihood = -33186.935
Iteration 5:   log likelihood = -33186.935

Logistic regression                                     Number of obs =   62,789
                                                        LR chi2(3)    = 19766.16
                                                        Prob > chi2   =   0.0000
Log likelihood = -33186.935                             Pseudo R2     =   0.2295

------------------------------------------------------------------------------
    wealthwm | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      hv025 |
       rural |  -2.947689   .0376888   -78.21   0.000    -3.021558    -2.87382
             |
      eduwm |
Above Primary |   1.086913   .0190894    56.94   0.000     1.049499    1.124328
             |
      hv104 |
      female |  -.0584805    .018851    -3.10   0.002    -.0954278   -.0215332
       _cons |   2.220634   .0393117    56.49   0.000     2.143585    2.297684
------------------------------------------------------------------------------

159 .
160 .
end of do-file

161 .
```