```
------------------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------
      name:  <unnamed>
       log:  D:\jason\workshop\complex survey deisgn\complex survey2.log
  log type:  text
 opened on:  28 Jun 2021, 09:31:00

. 
. 
. ***********************************************************************
. * Example 1. Survey data collected with a complex survey design
. ***********************************************************************
. 
. webuse highschool, clear

. 
. des weight height race sex county sampwgt ncounties state school nschools

              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------------------------------------------------------
------------------------------------------------------------------
weight          float   %9.0g                 weight (lbs.)
height          float   %9.0g                 height (in.)
race            byte    %9.0g      race       1=white, 2=black, 3=other
sex             byte    %9.0g      sex        1=male, 2=female
county          byte    %9.0g                 County ID (PSU)
sampwgt         double  %9.0g                 sampling weight
ncounties       byte    %9.0g                 Stage 1 FPC
state           byte    %9.0g                 State ID (strata)
school          byte    %9.0g                 School ID (SSU)
nschools        int     %9.0g                 Stage 2 FPC


. sum weight height race sex county sampwgt ncounties state school nschools

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      weight |      4,071    159.9516    34.43119   68.53333   332.6444
      height |      4,071     430.287    25.06648   355.1282   512.8205
        race |      4,071    1.144436    .3968771          1          3
         sex |      4,071     1.52395    .4994874          1          2
      county |      4,071    1.473348    .4993505          1          2
-------------+---------------------------------------------------------
     sampwgt |      4,071    1965.119    1261.919   350.9584   10387.62
   ncounties |      4,071    10.92655    .9974217         10         12
       state |      4,071    25.57799     14.2365          1         50
      school |      4,071    5.438467     2.86066          1         10
    nschools |      4,071     89.0253    28.42123         45        130

. 
. 
. 
. tab1 sex, mis

-> tabulation of sex

    1=male, |
   2=female |      Freq.     Percent        Cum.
------------+-----------------------------------
       male |      1,938       47.61       47.61
     female |      2,133       52.39      100.00
------------+-----------------------------------
      Total |      4,071      100.00
```

```
. quietly svyset county [pw = sampwgt], fpc(ncounties) strata(state)|| school, fpc(nschools)


. ************************************************************
. * 1.1 Analyzing the data without the sampling deisgn variables
. ************************************************************
.
. reg weight height race

      Source |       SS           df       MS      Number of obs   =     4,071
-------------+----------------------------------   F(2, 4068)      =    625.17
       Model |  1134354.91         2  567177.455   Prob > F        =    0.0000
    Residual |  3690657.81     4,068  907.241349   R-squared       =    0.2351
-------------+----------------------------------   Adj R-squared   =    0.2347
       Total |  4825012.72     4,070  1185.50681   Root MSE        =     30.12

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   .6667443   .0188561    35.36   0.000     .6297761    .7037125
        race |   2.187205   1.190937     1.84   0.066    -.1476834    4.522094
       _cons |  -129.4429   8.303603   -15.59   0.000    -145.7225   -113.1633
------------------------------------------------------------------------------

.
. ************************************************************************
. * 1.2 Analyzing the data, using only the personal weights
. ************************************************************************
.
. reg weight height race [pweight=sampwgt]
(sum of wgt is   8.0000e+06)

Linear regression                               Number of obs     =      4,071
                                                F(2, 4068)        =     545.46
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2789
                                                Root MSE          =     29.148

------------------------------------------------------------------------------
             |               Robust
      weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   .7179642   .0217759    32.97   0.000     .6752715    .7606569
        race |   1.152795   1.270376     0.91   0.364    -1.337837    3.643428
       _cons |  -151.6543   9.629201   -15.75   0.000    -170.5328   -132.7758
------------------------------------------------------------------------------
.
. ************************************************************************
. * 1.3. Analyzing the data, using all the sampling design variables
. ************************************************************************
.
. svy: reg weight height race
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =        50                  Number of obs     =      4,071
Number of PSUs     =       100                  Population size   =  8,000,000
                                                Design df         =         50
                                                F(   2,     49)   =     323.01
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2789

------------------------------------------------------------------------------
             |             Linearized
      weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   .7179642   .0287941    24.93   0.000     .6601295    .7757989
        race |   1.152795   1.668818     0.69   0.493    -2.199124    4.504715
       _cons |  -151.6543   12.04756   -12.59   0.000    -175.8525   -127.4561
------------------------------------------------------------------------------
```

```
.
.
. ***********************************
. * Example 2: Analyzie the sub-sample
. ***********************************
.
.
*************************************************************************************************
. * 2.1. Analyzing the data from a subpopulation, without considering the sampling design and
the change to the variance estimate
.
*************************************************************************************************
.
. reg weight height race if sex ==2

      Source |       SS           df       MS      Number of obs   =     2,133
-------------+----------------------------------   F(2, 2130)      =     62.52
       Model |  127576.528         2  63788.2638   Prob > F        =    0.0000
    Residual |  2173084.82     2,130  1020.22761   R-squared       =    0.0555
-------------+----------------------------------   Adj R-squared   =    0.0546
       Total |  2300661.34     2,132  1079.10945   Root MSE        =    31.941

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   .4219641   .0406362    10.38   0.000     .3422733    .5016549
        race |   8.113019   1.739819     4.66   0.000     4.701098    11.52494
       _cons |  -35.92156   17.05037    -2.11   0.035    -69.35867   -2.484446
------------------------------------------------------------------------------

.
.
*************************************************************************************************
. * 2.2. Analyzing the data from a subpopulation, without considering the change to the variance
estimate
.
*************************************************************************************************
.
.
. svy: reg weight height race if sex ==2
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =        50                  Number of obs      =      2,133
Number of PSUs     =       100                  Population size    =  4,151,979
                                                Design df          =         50
                                                F(   2,     49)    =      31.60
                                                Prob > F           =     0.0000
                                                R-squared          =     0.0565

------------------------------------------------------------------------------
             |             Linearized
      weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   .4335144   .0565368     7.67   0.000     .3199569    .5470719
        race |   5.160693   2.646177     1.95   0.057    -.1543106     10.4757
       _cons |  -40.00938   23.22344    -1.72   0.091    -86.65502    6.636268
------------------------------------------------------------------------------

.
.
*************************************************************************************************
. * 2.3. Analyzing the data from a subpopulation, considering both the sampling design and the
change to the variance estimate
.
*************************************************************************************************
.
.
. svy, subpop(if sex ==2): reg weight height race
```

```
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =        50                  Number of obs     =      4,071
Number of PSUs     =       100                  Population size   =  8,000,000
                                                Subpop. no. obs   =      2,133
                                                Subpop. size      =  4,151,979
                                                Design df         =         50
                                                F(   2,     49)   =      31.69
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0565

------------------------------------------------------------------------------
             |             Linearized
      weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   .4335144   .0564673     7.68   0.000     .3200964    .5469323
        race |   5.160693   2.644868     1.95   0.057    -.1516809    10.47307
       _cons |  -40.00938   23.19292    -1.73   0.091    -86.59373    6.574983
------------------------------------------------------------------------------

.
.
. ********************************************************************************
. * Example 3: The error message caused by strata with only one sampling uint
. ********************************************************************************
.
. use http://www.stata-press.com/data/r15/nhanes2b, clear

.
. svyset psuid [pweight=finalwgt], strata(stratid)

      pweight: finalwgt
          VCE: linearized
  Single unit: missing
     Strata 1: stratid
         SU 1: psuid
        FPC 1: <zero>

. svy: mean hdresult
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =      31       Number of obs    =       8,720
Number of PSUs   =      60       Population size  =  98,725,345
                                 Design df        =          29

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
    hdresult |   49.67141          .            .           .
--------------------------------------------------------------
Note: Missing standard error because of stratum with single
      sampling unit.

.
.
.
*************************************************************************************************
. * 3.1. if there are strata with only one sampling unit, Stata does not calculate the standard
error
. *************************************************************************************************
.
. svyset psuid [pweight=finalwgt],  strata(stratid) singleunit(missing)

      pweight: finalwgt
          VCE: linearized
  Single unit: missing
```

```
     Strata 1: stratid
         SU 1: psuid
        FPC 1: <zero>

. svy: mean hdresult
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =      31       Number of obs    =       8,720
Number of PSUs   =      60       Population size =  98,725,345
                                  Design df       =          29

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
    hdresult |   49.67141          .              .           .
--------------------------------------------------------------
Note: Missing standard error because of stratum with single
      sampling unit.

.
.
*************************************************************************************************
. * 3.2.Treat the only unit of the stratum as being selected with certaintly and selecting this
sampling unit does not contribute to the standared error
.
*************************************************************************************************
.
. svyset psuid [pweight=finalwgt],  strata(stratid) singleunit(certainty)

      pweight: finalwgt
          VCE: linearized
  Single unit: certainty
     Strata 1: stratid
         SU 1: psuid
        FPC 1: <zero>

. svy: mean hdresult
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =      31       Number of obs    =       8,720
Number of PSUs   =      60       Population size =  98,725,345
                                  Design df       =          29

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
    hdresult |   49.67141   .3829811      48.88813     50.4547
--------------------------------------------------------------
Note: Strata with single sampling unit treated as certainty
      units.

.
.
*************************************************************************************************
. * 3.3. using the average of the variances from the strata with multiple sampling units for
each stratum with one sampling unit
.
*************************************************************************************************
.
. svyset psuid [pweight=finalwgt],  strata(stratid) singleunit(scaled)

      pweight: finalwgt
          VCE: linearized
  Single unit: scaled
     Strata 1: stratid
```

```
         SU 1: psuid
        FPC 1: <zero>

. svy: mean hdresult
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =      31       Number of obs    =       8,720
Number of PSUs   =      60       Population size =  98,725,345
                                  Design df       =          29

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
    hdresult |   49.67141   .3959672      48.86157    50.48126
--------------------------------------------------------------
Note: Variance scaled to handle strata with a single sampling
      unit.

.
.
*************************************************************************************************
. * 3.4. using the average of the variances from all strata with multiple sampling units for
each stratum with one sampling unit
.
*************************************************************************************************
.
. svyset psuid [pweight=finalwgt],  strata(stratid) singleunit(centered)

      pweight: finalwgt
          VCE: linearized
  Single unit: centered
     Strata 1: stratid
         SU 1: psuid
        FPC 1: <zero>

. svy: mean hdresult
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =      31       Number of obs    =       8,720
Number of PSUs   =      60       Population size =  98,725,345
                                  Design df       =          29

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
    hdresult |   49.67141   .3838218      48.88641    50.45642
--------------------------------------------------------------
Note: Strata with single sampling unit centered at overall
      mean.

.
.
. log close
      name:  <unnamed>
       log:  D:\jason\workshop\complex survey deisgn\complex survey2.log
  log type:  text
 closed on:  28 Jun 2021, 09:31:02
-------------------------------------------------------------------------------------------------
-------------------------------------------------------------------
```