PRC Stats Consultants workshop: Transforming data for analysis

Selena Caldera

October 18, 2018

Transforming data for analysis

Source data: ABS.dta (Asia Barometer Survey), WVS.dta (World Values Survey), asian countries.dta (IMF, World Bank, ILO data)

Outputs: C:18 PRC Stats Consulting_asia.dta

```
      . set linesize 80

      . set varabbrev off, perm
      (set varabbrev preference recorded)

      . global homedir  "C:\Users\Selena\OneDrive\Fa18 PRC Stats Consulting"

      . global logdir   "$homedir\log files"

      . global datadir  "$homedir\data"

      . global output   "$homedir\output"

      . cap log close

      . log using "$logdir\Oct data management presentation", replace
      ------------------------------------------------------------------------------------------------------
--------------------------
            name:  <unnamed>
             log:  C:\Users\Selena\OneDrive\Fa18 PRC Stats Consulting\log files\Oct data management
presentation.smcl
        log type:  smcl
       opened on:  17 Oct 2018, 22:02:30

      . di c(current_date)
      17 Oct 2018
```

RESEARCH QUESTION: Is family policy associated with the gap between parents and nonparents (country-level analysis)?

What datasets are we using?

1. Country-level data:

```
. use "$datadir\asian countries.dta", clear
```

add variable labels

```
. la var country  "Country/Region"

. la var GDP_p    "GDP per capita"

. la var TFR      "Total fertility rate"

. la var wlabor   "Labor force participation rate, female"

. la var CPI3     "Family policy index"

. la var GINI     "Income/wealth distribution in country"

. la var mc_fam   "Measure of familialism"

. la var mc_ext   "Country-level measure of extended family"

. la var wkhr     "Average work week (country level)"

. tempfile countries

. save `countries'
file C:\Users\Selena\AppData\Local\Temp\ST_0000000x.tmp saved

. list if _n < 16

     +-------------------------------------------------------------------------+
  1. |   country |    GDP_p |   TFR | wlabor | CPI3 | GINI | mc_fam | mc_ext |
     |     China | 1512.637 | 1.502 |     67 |   .6 | .454 |  76.58 |  15.87 |
     |-------------------------------------------------------------------------|
     |                                wkhr                                     |
     |                                  47                                     |
     +-------------------------------------------------------------------------+

     +-------------------------------------------------------------------------+
  2. |   country |    GDP_p |   TFR | wlabor | CPI3 | GINI | mc_fam | mc_ext |
     |     Japan |  36453.8 |  1.29 |     48 | 2.06 | .498 |  93.41 |  14.59 |
     |-------------------------------------------------------------------------|
     |                                wkhr                                     |
     |                                  35                                     |
     +-------------------------------------------------------------------------+

     +-------------------------------------------------------------------------+
  3. |   country |    GDP_p |   TFR | wlabor | CPI3 | GINI | mc_fam | mc_ext |
     |     Korea | 15922.18 | 1.154 |     50 |  2.6 | .312 |   90.1 |   7.86 |
     |-------------------------------------------------------------------------|
     |                                wkhr                                     |
     |                                  45                                     |
     +-------------------------------------------------------------------------+

     +-------------------------------------------------------------------------+
  4. |   country |    GDP_p |   TFR | wlabor | CPI3 | GINI | mc_fam | mc_ext |
     |    Taiwan | 15355.67 |  1.57 |     48 |  .84 | .345 |  83.79 |  15.72 |
     |-------------------------------------------------------------------------|
     |                                wkhr                                     |
     |                                  42                                     |
     +-------------------------------------------------------------------------+

     +-------------------------------------------------------------------------+
  5. |   country |    GDP_p |   TFR | wlabor | CPI3 | GINI | mc_fam | mc_ext |
     |  Hongkong | 24875.45 |  .927 |     51 |  .84 | .525 |  68.17 |   7.18 |
     |-------------------------------------------------------------------------|
     |                                wkhr                                     |
     |                                  46                                     |
     +-------------------------------------------------------------------------+

     +-------------------------------------------------------------------------+
  6. |   country |    GDP_p |   TFR | wlabor | CPI3 | GINI | mc_fam | mc_ext |
     |     India |  657.522 | 3.036 |     36 | 1.36 | .325 |  88.31 |  25.45 |
     |-------------------------------------------------------------------------|
     |                                wkhr                                     |
     |                                  42                                     |
     +-------------------------------------------------------------------------+
```

```
. tab country, missing

Country/Reg |
        ion |      Freq.     Percent        Cum.
------------+-----------------------------------
      China |          1       10.00       10.00
      Japan |          1       10.00       20.00
      Korea |          1       10.00       30.00
     Taiwan |          1       10.00       40.00
   Hongkong |          1       10.00       50.00
      India |          1       10.00       60.00
  Indonesia |          1       10.00       70.00
   Malaysia |          1       10.00       80.00
   Thailand |          1       10.00       90.00
    Vietnam |          1       10.00      100.00
------------+-----------------------------------
      Total |         10      100.00
```

2. Individual-level data: Asian Barometer Survey & World Values Survey

```
. use "$datadir\ABS.dta", clear

. la var country      "Country/Region"

. la var hp5          "Measure of individual happiness"

. la var female       "Gender"

. la var age          "Age"

. la var age2         "Age squared"

. la var partner      "Marital status"

. la var hedu         "Highest level of education completed"

. la var employed     "Occupation"

. la var fulltime     "Employed full-time?"

. la var selfemp      "Self-employed?"

. la var profs        "Status of employment"

. la var agri         "Agricultural (occupation type)"

. la var finc_d       "Family income (standardized)"

. la var pt60         "Not sure"

. la var urban        "Residence is in urban area"

. la var familism     "Family needs > indiv needs"

. la var extend       "Household member is extended family"

. la var abs          "Data source is ABS"

. tempfile abs

. save `abs'
file C:\Users\Selena\AppData\Local\Temp\ST_00000010.tmp saved

. list if _n < 3

      +------------------------------------------------------------------------+
```

```
     1. | country  | hp5  | female  | age  | age2  | partner  |          hedu  |
           |   India  |   5  |      0  |  49  | 2401  |       1  |             0  |
           |-------------------------------------------------------------------|
           |  fulltime  |  employed  |  selfemp  |                       profs   |
           |         .  |         1  |         1  |      manager or professional   |
           |-------------------------------------------------------------------|
           |                           agri |   finc_d | pt60 | urban | familism |
           | not in agruculture-related work | 1.528687 |    1 |     . |        1 |
           |-------------------------------------------------------------------|
           |               extend              |               abs               |
           |                    0              |                 1               |
           +-------------------------------------------------------------------+

           +-------------------------------------------------------------------+
     2. | country  | hp5  | female  | age  | age2  | partner  |          hedu  |
           |   India  |   5  |      1  |  42  | 1764  |       1  | higher degree  |
           |-------------------------------------------------------------------|
           |  fulltime  |  employed  |  selfemp  |                       profs   |
           |         0  |         0  |         0  |  not manager or professional   |
           |-------------------------------------------------------------------|
           |                           agri |   finc_d | pt60 | urban | familism |
           | not in agruculture-related work | 1.957147 |    1 |     . |        1 |
           |-------------------------------------------------------------------|
           |               extend              |               abs               |
           |                    0              |                 1               |
           +-------------------------------------------------------------------+

     . tab country, missing

     Country/Reg |
             ion |      Freq.     Percent        Cum.
     ------------+-----------------------------------
           China |      2,000       17.75       17.75
           Japan |      1,003        8.90       26.65
           Korea |      1,023        9.08       35.72
          Taiwan |      1,006        8.93       44.65
       Hongkong |      1,000        8.87       53.52
           India |      1,238       10.98       64.51
       Indonesia |      1,000        8.87       73.38
        Malaysia |      1,000        8.87       82.25
        Thailand |      1,000        8.87       91.13
         Vietnam |      1,000        8.87      100.00
     ------------+-----------------------------------
           Total |     11,270      100.00
```

World Values Survey

```
. use "$datadir\WVS.dta", clear
```

add variable labels

```
. la var country  "Country/Region"

. la var hp5      "Measure of individual happiness"

. la var female   "Gender"

. la var age       "Age"

. la var age2     "Age squared"

. la var partner  "Marital status"

. la var hedu     "Highest level of education completed"

. la var employed "Occupation"

. la var fulltime "Employed full-time?"
```

```
. la var selfemp  "Self-employed?"

. la var profs    "Status of employment"

. la var agri     "Agricultural (occupation type)"

. la var finc_d   "Family income (standardized)"

. la var pt60     "Not sure"

. la var urban    "Residence is in urban area"

. la var familism "Family needs > indiv needs"

. la var extend   "Household member is extended family"

. la var wvs      "Data source is WVS"

. tempfile wvs

. save `wvs'
file C:\Users\Selena\AppData\Local\Temp\ST_00000011.tmp saved

. list if _n < 4

     +-------------------------------------------------------------------------+
  1. | country  | hp5  | female  | age  | age2  | partner  |          hedu  |
     |   China  |   5  |      1  |  52  | 2704  |       1  |             0  |
     |-------------------------------------------------------------------------|
     |  fulltime  |  employed  |  selfemp  |                        profs   |
     |         1  |         1  |        0  |  not manager or professional   |
     |-------------------------------------------------------------------------|
     |                           agri |   finc_d | pt60 | urban | familism |
     |        agriculture-related work | 1.095422 |    1 |     . |        1 |
     |-------------------------------------------------------------------------|
     |               extend              |               wvs               |
     |                    0              |                 1               |
     +-------------------------------------------------------------------------+

     +-------------------------------------------------------------------------+
  2. | country  | hp5  | female  | age  | age2  | partner  |          hedu  |
     |   China  |   4  |      0  |  22  |  484  |       0  | higher degree  |
     |-------------------------------------------------------------------------|
     |  fulltime  |  employed  |  selfemp  |                        profs   |
     |         0  |         0  |        0  |  not manager or professional   |
     |-------------------------------------------------------------------------|
     |                           agri |   finc_d | pt60 | urban | familism |
     | not in agruculture-related work | 1.632074 |    0 |     . |        1 |
     |-------------------------------------------------------------------------|
     |               extend              |               wvs               |
     |                    0              |                 1               |
     +-------------------------------------------------------------------------+

     +-------------------------------------------------------------------------+
  3. | country  | hp5  | female  | age  | age2  | partner  |          hedu  |
     |   China  |   4  |      0  |  29  |  841  |       1  |             0  |
     |-------------------------------------------------------------------------|
     |  fulltime  |  employed  |  selfemp  |                        profs   |
     |         1  |         1  |        0  |  not manager or professional   |
     |-------------------------------------------------------------------------|
     |                           agri |   finc_d | pt60 | urban | familism |
     | not in agruculture-related work | 1.095422 |    0 |     . |        0 |
     |-------------------------------------------------------------------------|
     |               extend              |               wvs               |
     |                    0              |                 1               |
     +-------------------------------------------------------------------------+
```

Append individual-level datasets

```
. append using `abs', generate(source)
(label agri already defined)
(label profs already defined)
(label hedu already defined)
(label country already defined)

. assert _N == 26282

. tempfile people

. save `people'
file C:\Users\Selena\AppData\Local\Temp\ST_00000012.tmp saved
```

Merge country-level and individual-level data

```
. use `countries', clear
```

This dataset has one unique observation per country the individual-level datasets have many observations per country.

```
. merge 1:m country using `people', generate(abs_merge)
(label country already defined)

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                            26,282  (abs_merge==3)
    -----------------------------------------
```

1:m tells Stata that the key variable uniquely identifies observations in the master dataset. But the key variable identifies more than one observation in the using dataset (defines the join type).

```
. keep if abs_merge == 3
(0 observations deleted)

. drop abs_merge
```

another option to do the same thing using the keep option:

merge 1:m country using `people', keep(3) gen(abs_merge)

drop abs_merge

Do-loop to mean-center explanatory variables

```
. foreach var of varlist GDP_p TFR CPI3 {
.         qui summarize `var'
.         gen `var'_stdzd = `var' - r(mean)
.         order `var'_stdzd, a(`var')
.         }

. la var GDP_p_stdzd      "Mean-centered GDP per capita"

. la var TFR_stdzd        "Mean-centered TFR"

. la var CPI3_stdzd       "Mean-centered Family Policy Index"
```

```
. list GDP_p GDP_p_stdzd TFR TFR_stdzd CPI3 CPI3_stdzd if _n < 16

     +-------------------------------------------------------------+
     |    GDP_p   GDP_p_s~d     TFR   TFR_stdzd   CPI3   CPI3_st~d |
     |-------------------------------------------------------------|
  1. | 1512.637    -7387.69   1.502   -.3229614     .6   -.6778179 |
  2. |  36453.8    27553.47    1.29   -.5349614   2.06     .782182 |
  3. | 15922.18    7021.853   1.154   -.6709613    2.6    1.322182 |
  4. | 15355.67    6455.343    1.57   -.2549613    .84   -.4378179 |
  5. | 24875.45    15975.13    .927   -.8979614    .84   -.4378179 |
     |-------------------------------------------------------------|
  6. |  657.522   -8242.805   3.036    1.211039   1.36    .0821821 |
  7. | 1280.696   -7619.631   2.475    .6500385    1.4    .1221821 |
  8. | 5171.417    -3728.91   2.307    .4820386    .93   -.3478179 |
  9. | 2676.296   -6224.031    1.58   -.2449613    .85   -.4278179 |
 10. |  603.668   -8296.658   1.894    .0690387    1.8     .522182 |
     |-------------------------------------------------------------|
 11. | 1512.637    -7387.69   1.502   -.3229614     .6   -.6778179 |
 12. | 1512.637    -7387.69   1.502   -.3229614     .6   -.6778179 |
 13. | 1512.637    -7387.69   1.502   -.3229614     .6   -.6778179 |
 14. | 1512.637    -7387.69   1.502   -.3229614     .6   -.6778179 |
 15. | 1512.637    -7387.69   1.502   -.3229614     .6   -.6778179 |
     +-------------------------------------------------------------+

. save "$datadir\happiness_asia.dta", replace
file C:\Users\Selena\OneDrive\Fa18 PRC Stats Consulting\data\happiness_asia.dta saved
```

This is a pretty basic do-loop. You can use more complicated loops for fancier operations. Do-loops are especially helpful for cleaning longitudinal data that starts out in wide format. For example say my data has five waves of measures for each individual:

```
forval i = 1/5 {

        rename `i'age age'i'

        rename `i'employed employed`i'

        rename `i'fulltime fulltime`i'

        rename `i'finc_d finc_d`i'

        }
```

OR nest the loops for many variables:

```
   forval i = 1/5 {

        foreach var of varlist age employed fulltime finc_d {

                rename `i'`var' `var''i'

                rename `i'`var' `var'`i'

                rename `i'`var' `var'`i'

                rename `i'`var' `var'`i'

                }

        }
```

Collapse individual happiness into a country level measure

```
. use `people', clear

. collapse (mean) hp5_country = hp5, by(country)
```

now we have a single happiness variable for each country

```
. list

     +----------------------+
     |   country   hp5_co~y |
     |----------------------|
  1. |     China   3.695171 |
  2. |     Japan   3.885396 |
  3. |     Korea   3.722522 |
  4. |    Taiwan    3.74843 |
  5. |  Hongkong   3.649287 |
     |----------------------|
  6. |     India   3.873492 |
  7. | Indonesia   3.979913 |
  8. |  Malaysia   4.218736 |
  9. |  Thailand   4.148221 |
 10. |   Vietnam   4.051189 |
     +----------------------+
```

merge back to our final dataset

```
. merge 1:m country using "$datadir\happiness_asia.dta", gen(happy)
(label country already defined)
(label agri already defined)
(label profs already defined)
(label hedu already defined)

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                            26,282  (happy==3)
    -----------------------------------------

. keep if happy == 3
(0 observations deleted)

. drop happy

. order hp5_country, a(hp5)
```

alternatively, we could use egen to accomplish the same task

```
. bys country: egen hp5_country2 = mean(hp5)

. list country hp5 hp5_country hp5_country2 if _n < 16

       +-----------------------------------+
       | country   hp5   hp5_co~y   hp5_co~2 |
       |-----------------------------------|
    1. |   China     3   3.695171   3.695171 |
    2. |   China     4   3.695171   3.695171 |
    3. |   China     4   3.695171   3.695171 |
    4. |   China     2   3.695171   3.695171 |
    5. |   China     4   3.695171   3.695171 |
       |-----------------------------------|
    6. |   China     4   3.695171   3.695171 |
    7. |   China     2   3.695171   3.695171 |
    8. |   China     4   3.695171   3.695171 |
```

```
  9. |   China     5   3.695171   3.695171 |
 10. |   China     4   3.695171   3.695171 |
     |-----------------------------------|
 11. |   China     4   3.695171   3.695171 |
 12. |   China     1   3.695171   3.695171 |
 13. |   China     4   3.695171   3.695171 |
 14. |   China     4   3.695171   3.695171 |
 15. |   China     4   3.695171   3.695171 |
     +-----------------------------------+
```

collapse is more suitable when all variables in your dataset are being collapsed down. E.g. a dataset with daily measures when the analysis level is monthly.

```
. save "$datadir\happiness_asia.dta", replace
file C:\Users\Selena\OneDrive\Fa18 PRC Stats Consulting\data\happiness_asia.dta saved
```