PRC Stats Consultants workshop: Exploratory Data Analysis

Shih-Yi Chao

Exploratory data analysis

Created by: Shih-Yi Chao

Date created: October 18, 2018

Source data: all.dta (user created)

```
. clear

. set more off

. cd "C:\Users\sychao\Dropbox\16_2018Fall\PRC TA\Data for the Talk"
C:\Users\sychao\Dropbox\16_2018Fall\PRC TA\Data for the Talk

. use all.dta, clear
```

Explore data

1. Missing value

summarizes # of missing values

```
. mdesc

    Variable    |     Missing          Total     Percent Missing
----------------+-----------------------------------------------
        country |           0         26,282            0.00
            hp5 |         114         26,282            0.43
         female |           9         26,282            0.03
            age |          30         26,282            0.11
           age2 |          30         26,282            0.11
        partner |          41         26,282            0.16
           hedu |          86         26,282            0.33
       fulltime |       7,818         26,282           29.75
       employed |         549         26,282            2.09
        selfemp |         524         26,282            1.99
          profs |       3,659         26,282           13.92
           agri |       3,659         26,282           13.92
         finc_d |       1,449         26,282            5.51
           pt60 |         230         26,282            0.88
          urban |      14,954         26,282           56.90
       familism |          89         26,282            0.34
         extend |         290         26,282            1.10
          GDP_p |           0         26,282            0.00
            TFR |           0         26,282            0.00
         wlabor |           0         26,282            0.00
           CPI3 |           0         26,282            0.00
           GINI |           0         26,282            0.00
         mc_fam |           0         26,282            0.00
         mc_ext |           0         26,282            0.00
           wkhr |           0         26,282            0.00
```

```
             wvs |          0         26,282          0.00
             abs |          0         26,282          0.00
-----------------+---------------------------------------------
```

summarizes missing pattern

```
. misschk hp5 female age partner hedu fulltime employed profs finc_d pt60, gen(miss)

Variables examined for missing values

   #  Variable        # Missing   % Missing
-------------------------------------------
   1  hp5                 114         0.4
   2  female                9         0.0
   3  age                  30         0.1
   4  partner              41         0.2
   5  hedu                 86         0.3
   6  fulltime           7818        29.7
   7  employed            549         2.1
   8  profs              3659        13.9
   9  finc_d             1449         5.5
   10 pt60                230         0.9

Warning: this output does not differentiate among extended missing.
To generate patterns for extended missing, use extmiss option.

 Missing for |
       which |
  variables? |      Freq.     Percent        Cum.
-------------+-----------------------------------
 12__5 _____ |          1        0.00        0.00
 1_34_ ___90 |          1        0.00        0.01
 1_3__ 6789_ |          1        0.00        0.01
 1__45 _____ |          1        0.00        0.02
 1__4_ ____0 |          1        0.00        0.02
 1___5 6789_ |          1        0.00        0.02
 1___5 67___ |          1        0.00        0.03
 1___5 ___9_ |          1        0.00        0.03
 1____ 6789_ |          1        0.00        0.03
 1____ 67_9_ |          1        0.00        0.04
 1____ 67___ |          2        0.01        0.05
 1____ 6__90 |          1        0.00        0.05
 1____ 6__9_ |          1        0.00        0.05
 1____ 6____ |          7        0.03        0.08
 1____ __89_ |          6        0.02        0.10
 1____ __8__ |         14        0.05        0.16
 1____ ___9_ |         10        0.04        0.19
 1____ _____ |         63        0.24        0.43
 _2345 67890 |          1        0.00        0.44
 _2345 67_9_ |          1        0.00        0.44
 _23_5 6789_ |          1        0.00        0.45
 _23_5 __8__ |          1        0.00        0.45
 _23__ ___9_ |          1        0.00        0.45
 _2___ _____ |          3        0.01        0.46
 __3_5 _____ |          1        0.00        0.47
 __3__ 678_0 |          1        0.00        0.47
 __3__ 678__ |          1        0.00        0.48
 __3__ 67_9_ |          1        0.00        0.48
 __3__ __89_ |          1        0.00        0.48
 __3__ __8__ |          1        0.00        0.49
 __3__ ___90 |          1        0.00        0.49
 __3__ ___9_ |          3        0.01        0.50
 __3__ _____ |         13        0.05        0.55
 ___45 __8__ |          1        0.00        0.56
 ___4_ 678__ |          1        0.00        0.56
 ___4_ 6____ |          1        0.00        0.56
 ___4_ __890 |          2        0.01        0.57
```

```
  ___4_ __89_ |          3        0.01        0.58
  ___4_ __8_0 |          1        0.00        0.59
  ___4_ __8__ |          3        0.01        0.60
  ___4_ ___90 |          4        0.02        0.61
  ___4_ ____0 |          6        0.02        0.64
  ___4_ _____ |         14        0.05        0.69
  ____5 67890 |          1        0.00        0.69
  ____5 678__ |          1        0.00        0.70
  ____5 67___ |          3        0.01        0.71
  ____5 6__9_ |          1        0.00        0.71
  ____5 6____ |          4        0.02        0.73
  ____5 __890 |          1        0.00        0.73
  ____5 __89_ |          1        0.00        0.73
  ____5 __8_0 |          1        0.00        0.74
  ____5 __8__ |         12        0.05        0.78
  ____5 ___90 |          2        0.01        0.79
  ____5 ___9_ |          8        0.03        0.82
  ____5 ____0 |          5        0.02        0.84
  ____5 _____ |         35        0.13        0.97
  _____ 67890 |          1        0.00        0.98
  _____ 6789_ |         19        0.07        1.05
  _____ 678_0 |          8        0.03        1.08
  _____ 678__ |        133        0.51        1.59
  _____ 67_90 |          2        0.01        1.59
  _____ 67_9_ |         55        0.21        1.80
  _____ 67__0 |          4        0.02        1.82
  _____ 67___ |        308        1.17        2.99
  _____ 6__90 |          3        0.01        3.00
  _____ 6__9_ |        259        0.99        3.99
  _____ 6___0 |          8        0.03        4.02
  _____ 6____ |      6,984       26.57       30.59
  _____ __890 |          9        0.03       30.63
  _____ __89_ |        158        0.60       31.23
  _____ __8_0 |         46        0.18       31.40
  _____ __8__ |      3,227       12.28       43.68
  _____ ___90 |         12        0.05       43.73
  _____ ___9_ |        874        3.33       47.05
  _____ ____0 |        108        0.41       47.46
  _____ _____ |     13,808       52.54      100.00
-------------+-----------------------------------
        Total |     26,282      100.00

Missing for |
   how many |
 variables? |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |     13,808       52.54       52.54
          1 |     11,321       43.08       95.61
          2 |        866        3.30       98.91
          3 |        237        0.90       99.81
          4 |         40        0.15       99.96
          5 |          4        0.02       99.98
          6 |          3        0.01       99.99
          7 |          2        0.01      100.00
          9 |          1        0.00      100.00
------------+-----------------------------------
      Total |     26,282      100.00
```

2. Examine whether the model violates OLS assumptions

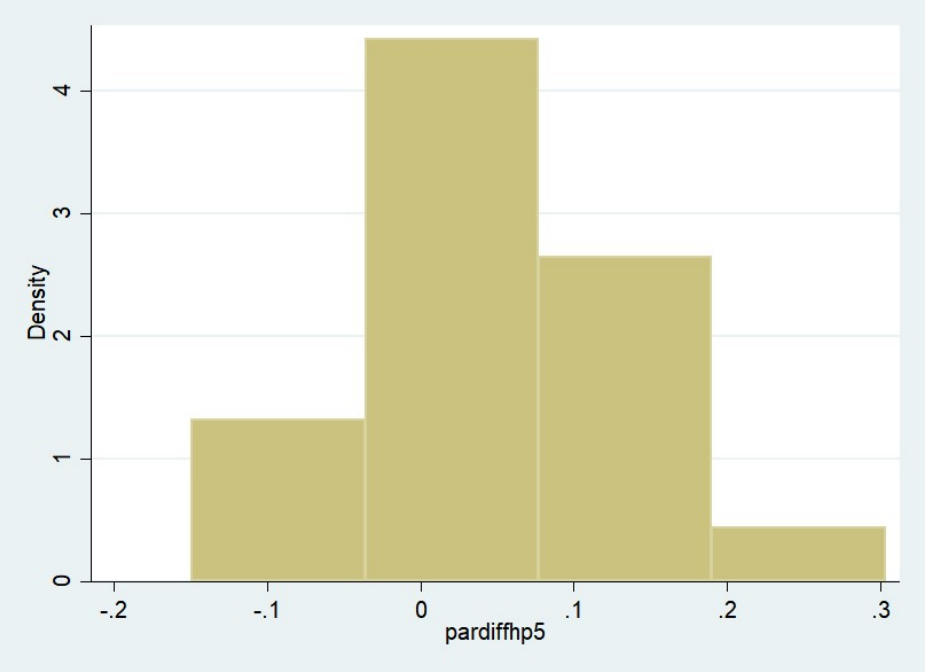
Research Question: whether family policy is associated with the gap between parents and nonparents, country-level analysis

Method: OLS

Model: happiness gap| family policy index, GDP, TFR, extend family, work hours

A. Normal Distribution of the Dependent Variable: happiness gap (countinuous)

```
. histogram pardiffhp5
(bin=4, start=-.14926076, width=.11322314)
```




```
. sum pardiffhp5, d
```

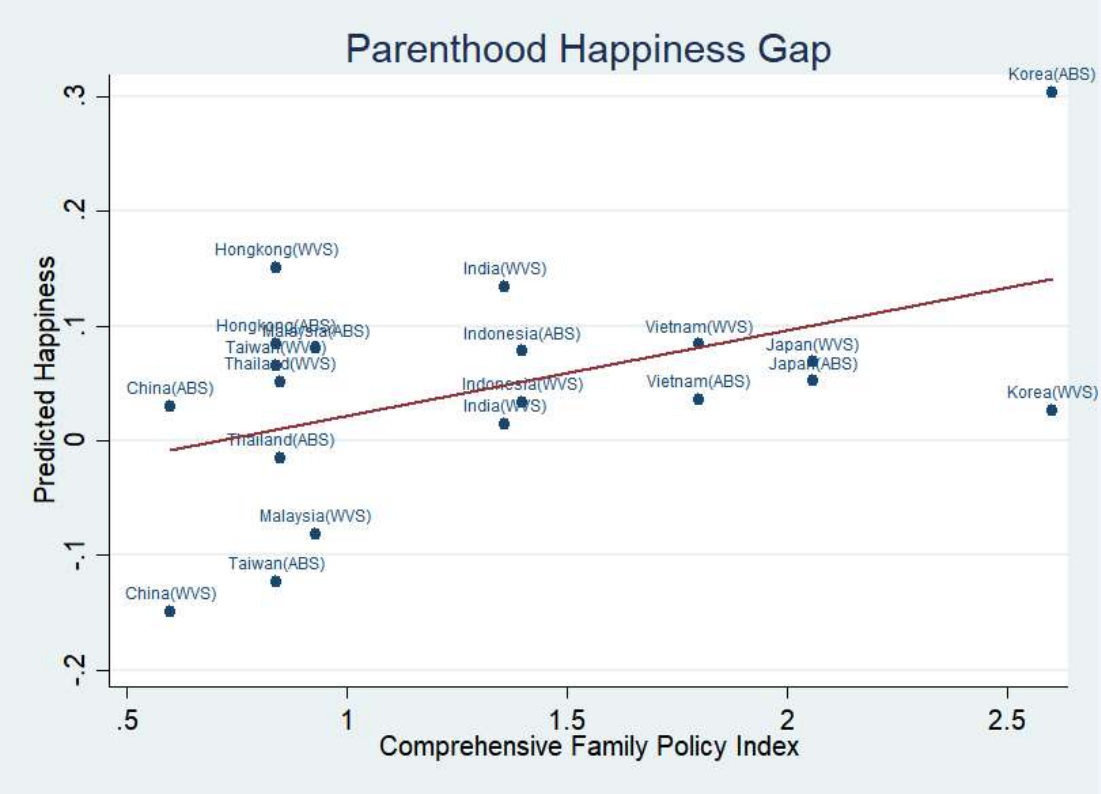
```
                        pardiffhp5
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -.1492608      -.1492608
 5%    -.1361448      -.1230288
10%    -.1022079       -.081387       Obs                  20
25%     .0202732      -.0151744       Sum of Wgt.          20

50%     .0514926                      Mean           .0461729
                        Largest       Std. Dev.      .0972646
75%     .0821371       .0846527
90%     .1423922       .1341724       Variance       .0094604
95%     .2271218       .1506119       Skewness       .2596614
99%     .3036318       .3036318       Kurtosis       4.418741
```

B. inflential cases: if you remove the cases from anaysis, the estimates have huge changes.

B-1 Scatterplots

```
. twoway (scatter pardiffhp5 CPI3, mlabel(group) mlabsize(vsmall) mlabposition(12) ti(Parenthood
Happiness Gap)) (lfit pardiffhp5 CPI3), ///
             xtitle(Comprehensive Family Policy Index) ytitle(Predicted Happiness) legend(off)
```



B-2 Two Tests for the Post estimation

Cook's D: threshold 4/n=4/20=0.2 or 1

```
. reg pardiffhp5 CPI3 GDP_p TFR mc_ext wkhr

      Source |       SS           df       MS      Number of obs   =        20
-------------+----------------------------------   F(5, 14)        =      1.10
       Model |   .05063412         5  .010126824   Prob > F        =    0.4043
    Residual |  .129113621        14  .009222402   R-squared       =    0.2817
-------------+----------------------------------   Adj R-squared   =    0.0252
       Total |  .179747741        19  .009460407   Root MSE        =    .09603

------------------------------------------------------------------------------
  pardiffhp5 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        CPI3 |     .07767   .0396367     1.96   0.070    -.0073422    .1626822
       GDP_p |   2.69e-06   4.03e-06     0.67   0.515    -5.95e-06    .0000113
         TFR |   .0316952   .0614501     0.52   0.614    -.1001022    .1634927
      mc_ext |   -.001246   .0052744    -0.24   0.817    -.0125585    .0100666
        wkhr |    .007301   .0101707     0.72   0.485     -.014513    .0291151
       _cons |  -.4426499   .5959823    -0.74   0.470    -1.720905    .8356051
------------------------------------------------------------------------------

. predict dfit,dfits

. predict d, cooksd

. list country data d if abs(d)>.2
```

```
     +---------------------------+
     |  country   data          d |
     |---------------------------|
  3. |    Korea    WVS    .297501 |
  8. | Malaysia    WVS   .2901316 |
 13. |    Korea    ABS    .369459 |
     +---------------------------+
```

DFBETAS: threshold 2/sqrt(n) or 1

```
. dfbeta
                  _dfbeta_1: dfbeta(CPI3)
                  _dfbeta_2: dfbeta(GDP_p)
                  _dfbeta_3: dfbeta(TFR)
                  _dfbeta_4: dfbeta(mc_ext)
                  _dfbeta_5: dfbeta(wkhr)

. list country data _dfbeta_1 if abs(_dfbeta_1)>.4472136

     +---------------------------+
     | country   data   _dfbeta_1 |
     |---------------------------|
  3. |   Korea    WVS   -1.156612 |
 13. |   Korea    ABS    1.333305 |
     +---------------------------+
```

C. Multicollinearity: the independent variables have no perfect correlations

vif (the variance inflation factor = 2.5)

```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
       GDP_p |      4.87    0.205394
        wkhr |      3.19    0.313701
         TFR |      3.17    0.315393
      mc_ext |      1.69    0.590375
        CPI3 |      1.29    0.777307
-------------+----------------------
    Mean VIF |      2.84
```

There are three situations in which a high VIF is not a problem

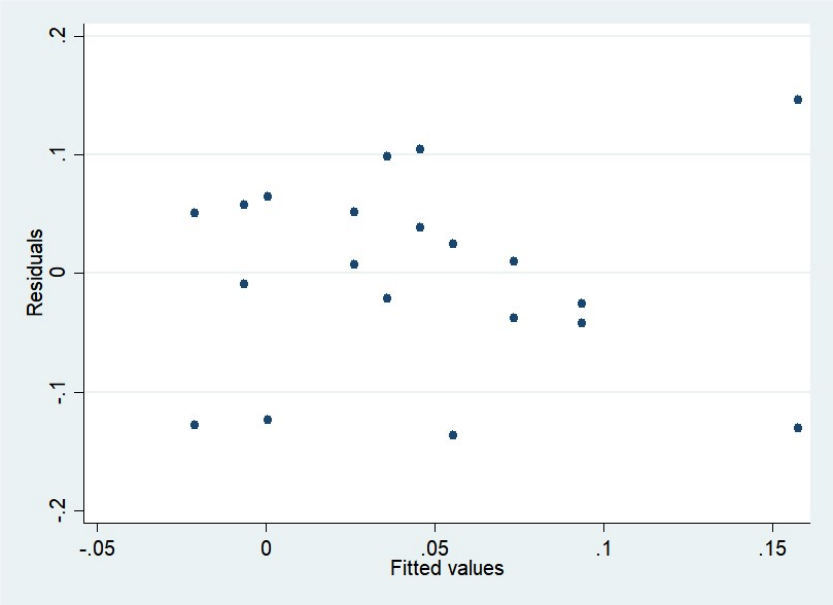
1. The variables with high VIFs are control variables and the variables of interest do not have high VIFs

2. The high VIFs are caused by the inclusion of powers or products of other variables

3. The variables with high VIFs are indicators (dummy) variables that represent a categorical variable with three or more categories

D. Heteroskedasticity: the error variance is not constant for all observations

```
. rvfplot
```




```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of pardiffhp5

         chi2(1)      =     1.07
         Prob > chi2  =   0.3016
```

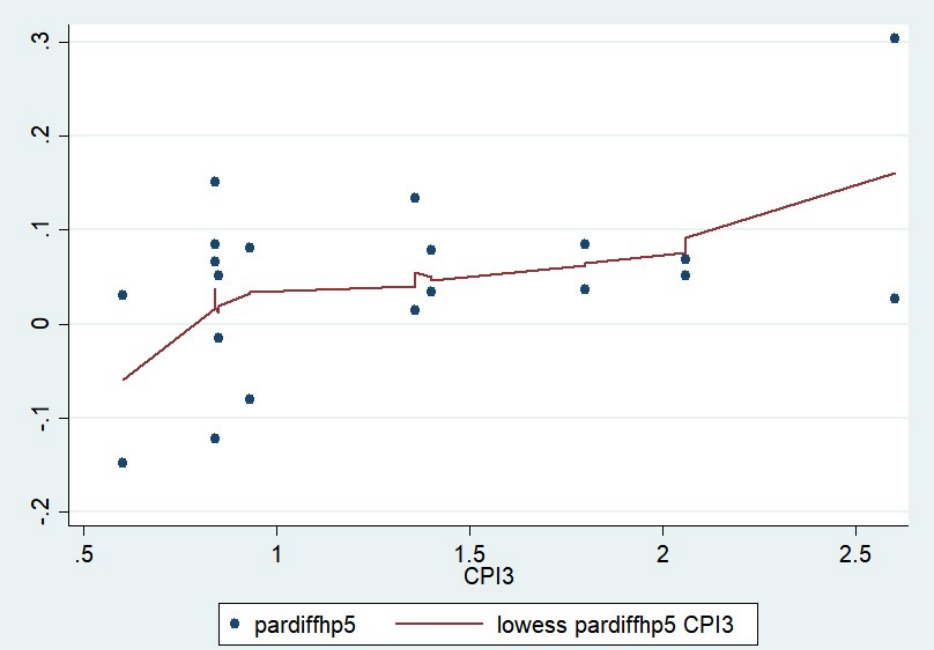
E. Omitted Variable: endogeneity

```
. ovtest

Ramsey RESET test using powers of the fitted values of pardiffhp5
       Ho:  model has no omitted variables
                  F(3, 11) =      0.86
                  Prob > F =      0.4904
```

F. Specification Error

```
. scatter pardiffhp5 CPI3 ||lowess pardiffhp5 CPI3
```




```
. acprplot CPI3, lowess
```