Fixed effect threshold model for unbalanced panel data

Qunyong Wang[1], Yujun Lian[2]

1. Institute of Statistics and Econometrics, Nankai University
2. Lingnan (University) College, Sun Yat-sen University

August 22nd, 2019

Contents

Fixed effect panel threshold model

- Threshold model has been widely used in economics and finance, monetary policy, economic growth etc.
- Hansen (1999) proposed the threshold model for panel data. The model is only used for balanced panel.
- For unbalanced panel, the user need to transform it to balanced one.
  1. potential sample selection bias.
  2. subjectivity: longer period and less individual, or shorter period and more individual.

Contents

- single threshold model

$$y_{it} = \mu + X_{it}(q_{it} < \gamma)\beta_1 + X_{it}(q_{it} \geq \gamma)\beta_2 + u_i + e_{it}. \qquad (1)$$

with $i = 1, 2, ..., G$, $t = 1, 2, ..., T_i$. The independent variable z is regime independent and x is regime dependent.

- written as

$$y_{it} = \mu + X_{it}(q_{it}, \gamma)\beta + u_i + e_{it}. \qquad (2)$$

- grid search to minimize $S_1(\gamma)$: $[\underline{\gamma}, \overline{\gamma}]$.

$$\hat{\gamma} = \underset{\gamma}{\operatorname{argmin}} S_1(\gamma) \tag{3}$$

- As proved by Hansen (1999), $\hat{\gamma}$ is consistent, and has non-standard distribution.

- threshold effect test:

$$H_0 : \beta_1 = \beta_2; \quad H_a : \beta_1 \neq \beta_2$$

- F-statistic

$$F_0 = \left(S_0 - S_1\right)/\hat{\sigma}_1^2.$$

- bootstrap p-value:
  (1) estimate linear model, get $\hat{e}_{it}$; estimate single threshold model, get F_o.
  (2) cluster resampling of $\hat{e}_{it}$, get bootstrap residual v_{it}^*.
  (3) bootstrap dependent variable

$$y_{it}^* = X_{it}\hat{\beta}.$$

(4) use y^*, X, and q to compute F-statistic.
(5) （2）-（4）B times , get F-statistic $(F_1, F_2, ..., F_B)$.

$$Prob = P(F_j > F_0).$$

Contents

Contents

- DGP (data generating process)

$$y_{it} = \mu + \alpha z_{it} + \beta_1 x_{it} 1(q_{it} < \tau) + \beta_2 x_{it} 1(q_{it} \geq \tau) + u_i + e_{it}, \quad (4)$$

- $z_{it}, x_{it} \sim \chi^2(1) - 1$,
  individual effect: $u_i \sim \chi^2(1) - 1$,
  the idiosyncratic error $e_{it} \sim N(0, 1)$.
- parameter: $\mu = 1, \alpha = 1, \beta_1 = 1, \beta_2 = 2, \tau = 1$.

- allow cluster sizes to vary systematically, we allocate N observations among G clusters using the formula

$$N_i = \left[\frac{N \exp(\gamma i/G)}{\sum_{j=1}^{G} \exp(\gamma j/G)} \right], \quad i = 1, 2, ..., G-1. \qquad (5)$$

- $[\cdot]$ denotes the integer part of the argument, and $N_G = N - \sum N_{j=1}^{G-1} N_g$. When $\gamma = 0$ and N/G is an integer, the panel is balanced and $N_i = N/G$ for all i. As γ increases, cluster size becomes more unequal.
  For example, when $G = 50$ and $N = 500$, if $\gamma = 1$, the largest cluster size is 24, and the smallest cluster size is 6.
  If $\gamma = 2$, the largest cluster size is 24, and the smallest cluster size is 3.

Table: Simulation result for balanced and unbalanced panel

			$\gamma = 0$		$\gamma = 1$		$\gamma = 2$	
G	$\bar{T}$	Obs	RMSE(10^3)	Rate	RMSE(10^3)	Rate	RMSE(10^3)	Rate
50	5	250	8.3	0.537	34.4	0.481	24.3	0.463
50	10	500	1.7	0.633	2.9	0.579	3.0	0.574
50	20	1,000	0.4	0.815	0.8	0.763	0.8	0.755
200	5	1,000	0.5	0.775	0.9	0.734	1.2	0.709
200	10	2,000	0.1	0.921	0.2	0.890	0.2	0.864
200	20	4,000	0.05	0.977	0.077	0.962	0.071	0.958
500	5	2,500	0.104	0.935	0.195	0.888	0.196	0.889
500	10	5,000	0.041	0.994	0.049	0.981	0.059	0.967
500	20	10,000	0.026	1.00	0.0.032	0.995	0.031	0.996

- Given the number of observations and the number of clusters, the more unbalanced the panel is, the higher the RMSE is.
- The RMSE decreases with a larger number of groups or a larger number of observations. This is true for both balanced and unbalanced panels. That implies the estimator is consistent in unbalanced panel data, so is the regression coefficient.
- The coverage rate increases with sample size. But the coverage rate gets higher than the nominal level when the sample size gets larger after some point. Too small sample tends to get a tighter interval and too large sample tends to get a wider interval.
- The RMSE increases and coverage rate decreases as the panel data become more unbalanced.

Contents

- DGP:

$$y_{it} = 1 + z_{it} + x_{it} + u_i + e_{it}. \tag{6}$$

where $z_{it}, x_{it} \sim \chi^2(1) - 1$. $u_i \sim \chi^2(1) - 1$, $e_{it} \sim N(0, 1)$.

- (1) Draw a random sample. First estimate a fixed effect linear model under the null hypothesis, and get the estimate $\hat{\beta}$, the fitted value
$$\hat{y}_{it} = \hat{\mu} + x_{it}\hat{\beta} + \hat{u}_i,$$
the residual $\hat{e}_{it} = y_{it} - \hat{y}_{it}$, and the sum of squared residuals S_0.
(2) Estimate the fixed effect single threshold model under the alternative hypothesis, and get the sum of squared residuals S_1. The F-statistic is computed as
$$F = \frac{S_0 - S_1}{\hat{\sigma}_1^2}. \tag{7}$$
(3) Repeat step (1) - (2) S times, and get a series of probability values $(P_1, P_2, ..., P_S)$. The test size is
$$Pr = \frac{\sum_{i=1}^{S} 1(P_i < \alpha)}{S}, \quad \alpha = 0.1, 0.05, 0.01. \tag{8}$$

- The probability of F-statistic: wild cluster bootstrap method.
  (a) draw a random wild weight v_i^b from Rademacher distribution

$$v_i^b = \begin{cases} 1, & \text{with prob. } 1/2 \\ -1, & \text{with prob. } 1/2 \end{cases}.$$

  The bootstrap error is

$$e_i^b = \hat{e}_i v_i^b$$

  and the bootstrap dependent variable is

$$y_i^b = \hat{y}_i + \hat{e}_i v_i^b. \tag{9}$$

  (b) Repeat step (a) R times, get a series of F-statistic $F_j (j = 1, 2, ..., R)$.

$$P = \frac{\sum_{i=1}^{R} 1(F_i > F)}{R}, \tag{10}$$

Several notes

- We can also use $\tilde{y}_{it} = x_{it}\hat{\beta}$ in the bootstrap.
- The constant and individual effect will dropped off in the within-deviation transformation.

- v_i^b can also be drawn from other distributions with mean 0 and variance 1. Some options for v_i^b include:
(a) Mammen (1993) two-point distribution

$$v_i^b = \begin{cases} 1-\phi, & \text{with prob. } \phi/\sqrt{5} \\ \phi, & \text{with prob. } (1+\sqrt{5})/2 \end{cases}.$$

The golden ratio is $\phi = (1+\sqrt{5})/2$.
(b) Webb (2014) six-point distribution which assigns probability 1/6 to each of 6 points, namely, $\pm\sqrt{1/2}$, ± 1, and $\pm\sqrt{3/2}$.
Rademacher and Mammen distribution can yield only 2^G distinct bootstrap samples. Webb distribution reduces, but not eliminate this problem.
(c) Standard normal distribution.
(d) Gamma distribution with shape parameter 4 and scale parameter 1/2 as suggested by Liu(1988).

- As described in Mackinnon (2018), simulation studies suggest that wild bootstrap tests based on the Rademacher distribution perform better than ones based on other auxiliary distributions; see, among others, Davidson, Monticini, and Peel (2007); Davidson and Flachaire (2008); Finlay and Magnusson (2016). However, the Webb six-point distribution is preferred to the Rademacher when G is less than 10 or perhaps 12.

Several notes

- which residual to use: restricted model (null hypothesis) or unrestricted model (alternative)?
- It is generally better use the wild cluster bootstrap under restrictions, MacKinnon et al.(2018).
- Intuitively, since inference involves estimating the probabilities of obtaining certain results under the assumption that the null is true, inference is improved by using bootstrap datasets in which the null in fact holds. Simulation evidence on this issue is presented in, among many others, Davidson and MacKinnon (1999) and Djogbenou, MacKinnon, and Nielsen (2018).

Table: Test size of balanced and unbalanced panel

G	$\bar{T}$	Obs	α	$\gamma = 0$	$\gamma = 0.5$	$\gamma = 1$	$\gamma = 1.5$	$\gamma = 2$
50	10	500	0.10	0.112	0.118	0.108	0.096	0.13
			0.05	0.062	0.05	0.04	0.048	0.058
			0.01	0.014	0.026	0.008	0.014	0.026
200	10	2,000	0.1	0.138	0.124	0.104	0.126	0.118
			0.05	0.078	0.004	0.054	0.062	0.066
			0.01	0.03	0.008	0.014	0.016	0.016

Contents

- DGP:

$$y_{it} = \mu + \alpha z_{it} + \beta_1 x_{it} 1(q_{it} < \tau) + \beta_2 x_{it} 1(q_{it} \geq \tau) + u_i + e_{it}, \quad (11)$$

If $\beta_1 = \beta_2$, then the threshold effect disappears. The bigger difference between β_1 and β_2, the higher the test power.

- Set $\beta_1 = 1, \beta_2 = (1.1, 1.2)$. $\gamma = (0, 0.5, 1, 1.5, 2)$

Table: Test power of balanced and unbalanced panel ($\alpha = 0.05$)

G	$\bar{T}$	Obs	β_2	$\gamma = 0$	$\gamma = 0.5$	$\gamma = 1$	$\gamma = 1.5$	$\gamma = 2$
50	10	500	1.1	0.248	0.18	0.132	0.164	0.178
			1.2	0.722	0.49	0.492	0.486	0.492
200	10	2,000	1.1	0.748	0.498	0.466	0.454	0.446
			1.2	1.000	0.966	0.976	0.970	0.946

Several notes

- The test power is relatively low for small sample with $G = 50$ and $\bar{T} = 10$. An G gets bigger, the test power is improved quite a lot.
- With the same number of observations, the test power for unbalanced panel is lower than that for balanced panel. But, the test power doesn't seem to be affected much by the extent of the unbalancedness.

Contents

- Basically, `xthreg2` inherits the syntax from `xthreg`.

  `xthreg2` *depvar* [*indepvars*][*if*] [*in*] , `rx(`*varlist*`)` `qx(`*varname*`)` [`thnum(`*integer*`)` `grid(`*integer*`)` `trim(`*numlist*`)` `bs(`*numlist*`)` `thlevel(`*#*`)` `gen(`*newvarname*`)` `noreg` `nobslog` `wc(`*string*`)` *options*]

- `xthreg2` will give the same point and confidence interval estimation with `xthreg`, but the threshold effect significance test may give different critical values and different probability value because of the different bootstrap design.

- Hansen (1999)

```
. use hansen1999, clear
. set seed 123
. xthreg2 i q1 q2 q3 d1 qd1, rx(c1) qx(d1) thnum(2) trim(0.01) bs(300)
There exist time-invariant individual(s) (maybe only one obs):  d1 qd1
Estimating  the  threshold  parameters:   1st ...  2nd ...  Done
Bootstrap for single threshold
......
.................................................. +  300
Bootstrap for double threshold model:
......
.................................................. +  300

Threshold estimator (level = 95):
-----------------------------------------------------
     model |    Threshold         Lower         Upper
-----------+-----------------------------------------
      Th-1 |       0.0158        0.0139        0.0177
     Th-21 |       0.0158        0.0139        0.0177
     Th-22 |       0.5394        0.5321        0.5470
-----------------------------------------------------

Threshold effect test (bootstrap = 300):
-------------------------------------------------------------------------------------
 Threshold |       RSS        MSE      Fstat      Prob     Crit10     Crit5     Crit1
-----------+-------------------------------------------------------------------------
    Single |   16.7953     0.0024      30.25     0.047     26.942    29.982    37.916
    Double |   16.7457     0.0023      21.14     0.207     25.339    28.949    36.070
-------------------------------------------------------------------------------------
```

- unbalanced example

```
. use hansen1999,clear
. set seed 123
. drop if (year<=1977 | year>=1985) & inrange(runiform(0, 566), 1, 200) & rnormal()<0.1
(740 observations deleted)
. xtdes
      id:  1, 2, ..., 565                                        n =        565
    year:  1974, 1975, ..., 1987                                 T =         14
           Delta(year) = 1 year
           Span(year)  = 14 periods
           (id*year uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                        10      11      12        13        13      14      14

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+----------------
      138     24.42   24.42 |  11111111111111
       39      6.90   31.33 |  11.11111111111
       32      5.66   36.99 |  .1111111111111
       30      5.31   42.30 |  1.111111111111
       29      5.13   47.43 |  11111111111.11
       27      4.78   52.21 |  111.1111111111
       26      4.60   56.81 |  1111111111111.
       23      4.07   60.88 |  111111111111.1
       10      1.77   62.65 |  ..111111111111
      211     37.35  100.00 | (other patterns)
 ---------------------------+----------------
      565    100.00         |  XXXXXXXXXXXXXX
```

- triple threshold model

```
. set seed 123
. xthreg2 i q1 q2 q3 d1 qd1, rx(c1) qx(d1) thnum(3) trim(0.01) bs(300)
There exist time-invariant individual(s) (maybe only one obs):  d1 qd1
Estimating  the  threshold  parameters:   1st ...  2nd ...  3rd ...  Done
Bootstrap for single threshold
..................................................+   300
Bootstrap for double threshold model:
..................................................+   300
Bootstrap for triple threshold model:
..................................................+   300

Threshold estimator (level = 95):
-----------------------------------------------------
     model |    Threshold        Lower        Upper
-----------+-----------------------------------------
      Th-1 |       0.0154       0.0121       0.0173
     Th-21 |       0.0154       0.0129       0.0173
     Th-22 |       0.5421       0.5252       0.5487
      Th-3 |       0.9230            .            .
-----------------------------------------------------

Threshold effect test (bootstrap = 300):
-------------------------------------------------------------------------------------
 Threshold |       RSS        MSE      Fstat     Prob     Crit10     Crit5     Crit1
-----------+-------------------------------------------------------------------------
    Single |   17.3886     0.0027      36.65    0.010     24.212    27.861    36.592
    Double |   17.3201     0.0027      25.41    0.077     24.423    28.217    31.632
    Triple |   17.2933     0.0027       9.99    0.470     21.504    24.319    32.615
-------------------------------------------------------------------------------------
```

- double threshold reg. result

```
. xthreg2 i q1 q2 q3 d1 qd1, rx(c1) qx(d1) thnum(2) trim(0.01)
Fixed-effects (within) regression               Number of obs      =      6,992
Group variable: id                              Number of groups   =        550
R-sq:                                           Obs per group:
     within  = 0.1099                                         min =         10
     between = 0.1166                                         avg =       12.7
     overall = 0.1064                                         max =         14

                                                F(8,6434)         =      99.29
corr(u_i, Xb)  = -0.1886                        Prob > F          =     0.0000

------------------------------------------------------------------------------
           i |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          q1 |   .0115655   .0009494    12.18   0.000     .0097043    .0134267
          q2 |  -.0248182   .0028863    -8.60   0.000    -.0304763   -.0191602
          q3 |   .0013585   .0002162     6.28   0.000     .0009348    .0017823
          d1 |  -.0249041   .0041553    -5.99   0.000    -.0330498   -.0167583
         qd1 |   .0010209   .0013583     0.75   0.452    -.0016418    .0036837
    _cat#c.c1 |
          0  |   .0580387   .0066704     8.70   0.000     .0449624     .071115
          1  |   .1353776   .0176681     7.66   0.000     .1007422    .1700131
          2  |   .0861913   .0054647    15.77   0.000     .0754787    .0969039
       _cons |   .0636492   .0017055    37.32   0.000     .0603059    .0669925
-------------+----------------------------------------------------------------
     sigma_u |  .03530003
     sigma_e |  .04558565
         rho |  .37486107   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(549, 6434) = 6.88                  Prob > F = 0.0000
```

Contents

conclusions

- A more unbalanced panel will result in more volatile threshold estimate and tighter coverage rate
- The test size is not affected by the unbalancedness.
- The test power is affected by the unbalancedness, but it does't deteriorate when the panel becomes more unbalanced.

So?

- Our conclusions don't mean that it is be always better to use the unbalanced panel than to transform it into balanced one.
- Two choices for a unbalanced panel:
  First, transform the panel data into balanced one by dropping some observations, then estimate the threshold model using the balanced panel.
  Second, directly estimate the threshold model using the unbalanced panel.
- Both choices have pros and cons. The balanced panel has a less volatile estimate for the threshold, but less efficient estimate for regression coefficient due to less observations. The unbalanced panel has a more volatile estimate for the threshold, but more efficient estimate for regression coefficient beneficial from more observations.

Further questions

- the effect of heteroscedasticity
- sample selection bias

Thank you