

Common breaks in panel data

Jushan Bai

December 9, 2006

To highlight the main idea, we consider a simple mean shift model:

$$\begin{aligned} Y_{it} &= \mu_{i1} + e_{it}, & t = 1, 2, \dots, k_0 \\ Y_{it} &= \mu_{i2} + e_{it}, & t = k_0 + 1, \dots, T \\ & & i = 1, 2, \dots, N \end{aligned} \tag{1}$$

k_0 is the common break point.

Two cases for T : (i) T is fixed, (ii) $T \rightarrow \infty$.

Assume

$$k_0 = [T\tau_0]$$

$$E(e_{it}) = 0, \quad \text{Var}(e_{it}) = \sigma_i^2, \quad E(e_{it}e_{is}) = 0, \quad t \neq s$$

and $\sigma_i^2 \leq \bar{\sigma}^2$ for some $\bar{\sigma}^2 > 0$, for all i . In addition, we shall assume e_{it} are cross-sectionally independent for simplicity. Finally the following moment condition on e_{it} is assumed,

$$E|e_{it}|^{4+\delta} \leq M$$

On the magnitude of breaks, we consider three cases:

case 1

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N (\mu_{i2} - \mu_{i1})^2 > 0. \quad (2)$$

This would be the case if the magnitude of breaks $\mu_{i2} - \mu_{i1}$ are iid random variables with positive variance.

case 2 (a weaker condition),

$$\lim_{N \rightarrow \infty} N^{-1/2} \sum_{i=1}^N (\mu_{i2} - \mu_{i1})^2 = \infty. \quad (3)$$

case 3 (weakest)

$$\lim_{N \rightarrow \infty} \sum_{i=1}^N (\mu_{i2} - \mu_{i1})^2 = \infty. \quad (4)$$

These conditions do not require every series to have a break.

simple least squares estimation

Define the pre-break and post-break means

$$\bar{Y}_{i1} = \frac{1}{k} \sum_{t=1}^k Y_{it}$$

$$\bar{Y}_{i2} = \frac{1}{T-k} \sum_{t=k+1}^T Y_{it}$$

The sum of squared residuals for the i th equation is

$$S_{iT}(k) = \sum_{t=1}^k (Y_{it} - \bar{Y}_{i1})^2 + \sum_{t=k+1}^T (Y_{it} - \bar{Y}_{i2})^2$$

The total sum of squared residuals across all equations is given by

$$SSR(k) = \sum_{i=1}^N S_{iT}(k)$$

The least squares estimator for k_0 in the panel data model is defined as

$$\hat{k} = \operatorname{argmin}_{1 \leq k \leq T} SSR(k)$$

This estimator is easy to compute.

(Estimating a break equation by equation then averaging may not be consistent)

When $N = 1$ (a univariate series), it is well known that

$$\hat{k} = k_0 + O_p(1) \tag{5}$$

so that the difference between \hat{k} and true break point k_0 is stochastically bounded. For a fixed T , such a statement is not helpful because T is bounded and $\hat{k} - k_0$ is always bounded.

When $T \rightarrow \infty$, the statement of (5) is quite strong. It implies that

$$\hat{\tau} = \tau_0 + O_p(T^{-1})$$

where $\hat{\tau} = \hat{k}/T$. So in terms of the fraction of the sample size, $\hat{\tau}$ is T -consistent for τ_0 . Nevertheless, \hat{k} itself is not consistent for k_0 in univariate framework.

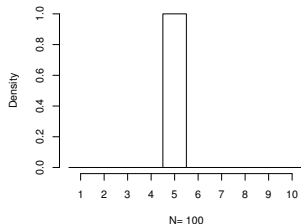
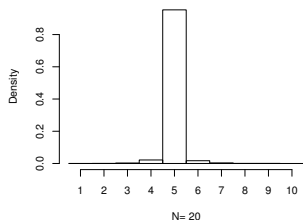
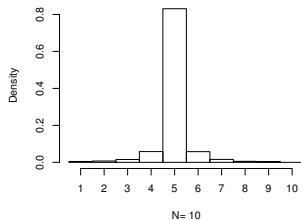
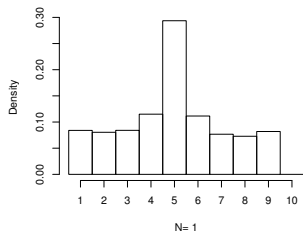
For panel data, however, much stronger statements can be made. We prove the following result:

Some simulation results

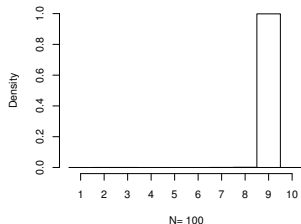
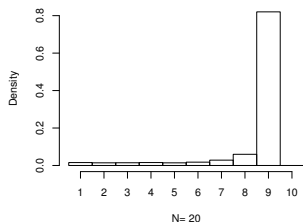
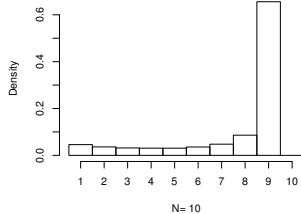
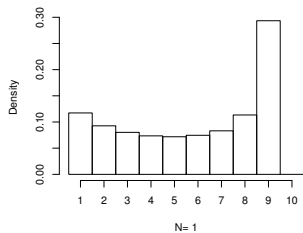
$$\mu_{i2} - \mu_{i1} \sim U(-2, 2), \quad i = 1, 2, \dots, N.$$

$$e_{it} \sim iid N(0, 1)$$

Histogram of the estimated break points ($T = 10, k_0 = 5$)



Histogram of the estimated break points ($T = 10, k_0 = 9$)



Second consistency theorem

Recall condition 3

$$\lim_{N \rightarrow \infty} \sum_{i=1}^N (\mu_{i2} - \mu_{i1})^2 = \infty.$$

As a cost to the weaker condition, we have to assume T to be large such that

$$\frac{\log(\log(T))}{T} N \rightarrow 0 \tag{6}$$

as T and N going to infinity. This restricts the relative rate at which T and N diverge. Intuitively, this means that with more observations, one can detect faint signals.

Single series limiting distribution

$$\begin{aligned} Y_t &= \mu_1 + e_t, & t = 1, 2, \dots, k_0 \\ Y_t &= \mu_2 + e_t, & t = k_0 + 1, \dots, T. \end{aligned}$$

$$\hat{k} - k_0 \xrightarrow{d} \operatorname{argmin}_\ell V(\ell) \quad (7)$$

where $V(0) = 0$ and

$$V(\ell) = (\mu_2 - \mu_1)^2 |\ell| - 2(\mu_2 - \mu_1) \sum_{s=-\ell+1}^0 e_s, \quad \ell = -1, -2, \dots$$

$$V(\ell) = (\mu_2 - \mu_1)^2 \ell - 2(\mu_2 - \mu_1) \sum_{s=1}^{\ell} e_s, \quad \ell = 1, 2, \dots$$

Panel data limiting distribution

To obtain a non-degenerate distribution, we assume

$$\lim_{N \rightarrow \infty} \sum_{i=1}^N (\mu_{i2} - \mu_{i1})^2 = \lambda > 0$$

or

$$\mu_{i2} - \mu_{i1} = N^{-1/2} \Delta_i, \text{ with } \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \Delta_i^2 = \lambda \quad (8)$$

In practice, N is fixed, this assumption provides a reasonable approximation.

assumptions (6), (8), as $N, T \rightarrow \infty$,

$$A_N(\hat{k} - k_0) \xrightarrow{d} \operatorname{argmin}_\ell \left[|\ell| + 2W(\ell) \right] \quad (10)$$

where

$$A_N = \frac{[\sum_{i=1}^N (\mu_{i2} - \mu_{i1})^2]^2}{\sum_{i=1}^N (\mu_{i2} - \mu_{i1})^2 \sigma_i^2}$$

the limiting distribution can be easily simulated.

$$P(|\ell^*| \leq 7) \simeq 0.90$$

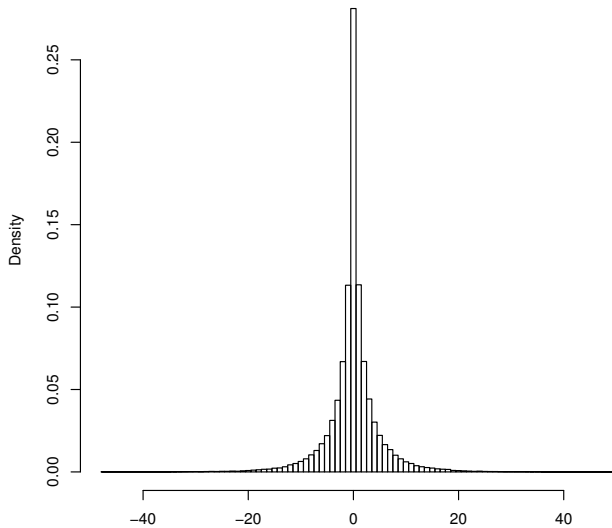
$$P(|\ell^*| \leq 11) \simeq 0.95$$

$$P(|\ell^*| \leq 20) \simeq 0.99$$

Using (10), the 90% confidence interval for k_0 is constructed as

$$[\hat{k} - \text{floor}(7/\hat{A}_N), \hat{k} + \text{ceiling}(7/\hat{A}_N)] \quad (11)$$

limiting distribution



simulation

$$\mu_{i2} - \mu_{i1} \sim \sigma \cdot U(-1, 1), \quad i = 1, 2, \dots, N$$

where σ is the standard deviation of the disturbances, i.e., $\sigma^2 = E(e_{it}^2)$. The magnitude of break is small relative to the error variance. For a single series, it would be very difficult to accurately estimate the break point.

Table 1. Coverage rate and the length of confidence intervals
($T = 100$)

Distribution of e_{it}	N	Coverage rate			Median length of CI		
		90%	95%	99%	90%	95%	99%
$N(0, 1)$	1	0.635	0.719	0.812	25	39	70
	5	0.829	0.886	0.954	9	13	23
	10	0.900	0.932	0.979	5	7	13
	15	0.937	0.968	0.989	5	7	9
	20	0.949	0.983	0.994	3	5	7
$\chi^2_{(5)}$	1	0.647	0.722	0.815	25	39	69
	5	0.824	0.885	0.944	9	13	23
	10	0.905	0.941	0.979	5	9	13
	15	0.933	0.963	0.989	5	5	9
	20	0.942	0.959	0.986	3	5	7
$t_{(5)}$	1	0.646	0.713	0.800	25	39	69
	5	0.830	0.883	0.927	9	13	23
	10	0.904	0.947	0.978	5	7	13
	15	0.939	0.967	0.989	5	5	9
	20	0.935	0.968	0.988	3	5	7

An application

Breaks in per capita GDP growth rates across 18 industrialized countries (most of European countries plus Canada, Japan, and US)

Data period: 1951-1990 (Ben-David and Papell, 1998)

With this data set, we have $N = 18$, and $T = 40$. Applying the method introduced earlier, the estimated break point is $\hat{k} = 23$, which corresponds to year 1973. The confidence intervals are

90% : (1972, 1974)

95% : (1972, 1974)

99% : (1971, 1975)

▶ Summary:

1. Consistency is possible with fixed T or even with a regime having a single period.
2. Derived the limiting distribution, useful for construction of confidence intervals.

▶ Possible extensions:

1. general regression model, AR
2. linear processes for e_{it}
3. Multiple breaks.
4. mean and variance breaks.