Freie Universität Bozen
Libera Università di Bolzano
Università Liedia de Bulsan

xtbreak: Estimation of and testing for structural breaks in Stata

US Stata Conference 2021

Jan Ditzen[1], Yiannis Karavias[2], Joakim Westerlund[3]

[1]Free University of Bozen-Bolzano, Bozen, Italy
www.jan.ditzen.net, jan.ditzen@unibz.it

[2]University of Birmingham, UK
https://sites.google.com/site/yianniskaravias/
i.karavias@bham.ac.uk

[3]Lund University, Lund, Sweden

August 05, 2021

Motivation

- In time series or panel time series structural breaks (or change points) in the relationships between key variables can occur.
- Estimations and forecasts depend on knowledge about structural breaks.
- Structural breaks might influence interpretations and policy recommendations.
- Break can be unknown or known and single and multiple breaks can occur.
- Examples: Financial Crisis, oil price shock, Brexit Referendum, COVID19,...
- Question: Can we estimate when the breaks occur and test them?

Literature

- Time Series:
  - Andrews (1993) test for parameter instability and structure change with unknown change point.
  - Bai and Perron (1998) propose three tests for and estimation of multiple change points.
- Panel (Time) Series:
  - Wachter and Tzavalis (2012) single structural break in dynamic independent panels.
  - Antoch et al. (2019); Hidalgo and Schafgans (2017) single structural break in dependent panel data.
  - Ditzen et al. (2021); Karavias et al. (2021) single and multiple breaks in panel data with cross-section dependence.
- `xtbreak` introduces estimation of and tests for multiple structural breaks in time series and panel data based on Bai and Perron (1998) and Ditzen et al. (2021); Karavias et al. (2021).

Econometric Model I

- Static linear panel regression model with s breaks:

$$
\begin{aligned}
y_{i,t} &= x'_{i,t}\beta + w'_{i,t}\delta_1 + u_{i,t}, & t &= 1, ..., T_1, \quad i = 1, ..., N \\
y_{i,t} &= x'_{i,t}\beta + w'_{i,t}\delta_2 + u_{i,t}, & t &= T_1 + 1, ..., T_2 \\
&\dots \\
y_{i,t} &= x'_{i,t}\beta + w'_{i,t}\delta_{s+1} + u_{i,t}, & t &= T_s, ..., T
\end{aligned}
$$

- $\tau_s = (T_1, T_2, ..., T_s)$ are break points of the s breaks.
- x_t is a $(1 \times p)$ vector of variables without structural breaks.
- w_t is a $(1 \times q)$ vector of variables with structural breaks.
- Fixed effects can be included in $x_{i,t}$, pooled constant can be included in $x_{i,t}$ or $w_{i,t}$
- Error $u_{i,t}$ contains unobserved heterogeneity $(u_{i,t} = f'_t\gamma_i + \epsilon_{i,t})$.

Econometric Model II

- The model can be expressed in matrix form:

$$Y_i = X_i\beta + W_i(\tau_s)\delta + U_i \tag{1}$$

- where $Y_i = (y_{i,1}, .., y_{i,T})'$, $W_i = (w_{i,1}, ..., w_{i,T})'$, $\delta = (\delta_1', ..., \delta_{s+1}')'$ and:

$$W_i(\tau_s) = \begin{pmatrix} w_{1,i} & 0 & \cdots & 0 \\ 0 & w_{2,i} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & w_{s+1,i} \end{pmatrix}$$

- $w_{s,i}$ is $(T_s \times q)$.
- Aim: Estimation and testing of breaks $\tau_s = (T_1, T_2, ..., T_s)$.

Estimation of breaks

Unknown Breakpoints

- Main idea: if the model has the true number of breaks and the true point in time, then the SSR should be smaller than for a model with a larger or smaller number of breaks.
- `xtbreak` implements the dynamic programming algorithm from Bai and Perron (2003). Idea is to calculate the SSR for all *necessary* subsamples.

- For example: Break in period 2 ($T_1 = 2$), then $SSR = SSR(1,2) + SSR(3,T)$.

Start \ End	1	2	3	...	T
1	$\ddots$	$SSR(1,2)$	$SSR(1,3)$	$\cdots$	$SSR(1,T)$
2		$\ddots$	$SSR(2,3)$	$\cdots$	$SSR(2,T)$
3			$\ddots$		$SSR(3,T)$
$\vdots$				$\ddots$	
T					$\ddots$

Estimation of breaks

- Point of break is determined by minimum of the SSR for a given number of breaks $\hat{b}$.
- Confidence intervals can be constructed around the estimated following Bai (1997); Bai and Perron (1998); Karavias et al. (2021):

$$\left[\hat{b} \pm \left\lfloor c_\alpha \frac{\hat{\delta}(\hat{b})' R' \hat{\Phi}_X R \hat{\delta}(\hat{b})}{N\left(\hat{\delta}(\hat{b})' R' \hat{\Omega}_X R \hat{\delta}(\hat{b})\right)} \right\rfloor \pm 1\right]$$

- where $\hat{\Omega}_X = \frac{1}{NT}\sum_{i=1}^{N} X_i' X_i$, $\hat{\Phi}_X = \frac{1}{NT}\sum_{i=1}^{N} \hat{\sigma}^2_{\epsilon,i} X_i' X_i$.

Three tests for breaks

- Three hypotheses (Bai and Perron, 1998):
  1. No break vs. s breaks ‣ Details Hypothesis 1
     $H_0 : \delta_1 = \delta_2 = ... = \delta_{s+1}$ vs $H_1 : \delta_k \neq \delta_j$ for some $j \neq k$.
  2. No break vs $1 \leq s \leq s^*$ breaks ‣ Details Hypothesis 2
     $H_0 : \delta_1 = \delta_2 = ... = \delta_{s+1}$ vs $H_1 : \delta_k \neq \delta_j$ for some $j \neq k$ and $s = 1, ..., s^*$
  3. s breaks vs $s + 1$ breaks ‣ Details Hypothesis 3
     $H_0 : \delta_j = \delta_{j+1}$ for one $j = 1, .., s$ vs. $H_1 : \delta_j \neq \delta_{j+1}$ for all $j = 1, ..., s$.

xtbreak[1]

For the estimation of breakpoints:

xtbreak estimate *depvar* [*indepvars*] [*if*] [, *general_options* showindex]

Testing for breaks:

xtbreak test *depvar* [*indepvars*] [*if*] [, *general_options*]

general_options are:

break_point_options *panel_options* nobreakvariables(varlists) noconstant breakconstant vce(ssr|hac|nw)

[1] This command is work in progress. Options, functions and results might change.

xtbreak I

If the break is estimated, then `break_point_options` are:

```
breaks(real) minlength(real) error(real)
```

- `breaks(real)` number of breaks.
- `showindex` display index of confidence interval rather than dates.
- `minlength(real)` minimal length of segments in %.
- `error(real)` minimal difference between SSRs for partial break model.

If an unknown break point is tested, then `break_point_options` are:

```
hypothesis(1|2|3) breaks(real) minlength(real) level(real)
error(real) wdmax
```

- hypothesis() which hypothesis to test.
- `breaks(real)` number of breaks.
- `level` which level the weighted (only hypothesis 2) test is evaluated at.
- `wdmax` weighted max test (only hypothesis 2).

xtbreak II

If the breakpoint is known then `break_point_options` are:

`breakpoints(numlist [,index fmt(string)])`

panel_options are specific for panel data sets:

`nofixedeffects csd csa(varlist, deterministic[(varlist)])`

`csanobreak(varlist, deterministic[(varlist)])`

- `nofixedeffects` omits fixed effects model. If `noconstant` not used, assume pooled OLS model.
- `csa` and `csanobreak` define variables added as cross-section averages. Suboption `deterministic` treats variables as deterministic cross-section averages.
- `csd` automatically select cross-section averages.

`xtbreak update`

- Updates `xtbreak` from GitHub.

Excess Mortality and number of COVID cases in the US I

- Question: can we identify structural breaks in the relationship between excess mortality and number of COVID19 cases in the US in 2020 and 2021?
- Excess mortality, em_t is defined as the difference between the actual deaths and the average over 2015 to 2019.
- Time between positive covid test, nc_t and death between 1 to 2 weeks.
- $em_t = \beta_0 + \beta_1 nc_{t-1} + \epsilon_t$, with em_t excess mortality and nc_t new cases.
- Three potential regimes:
  1. high death rates, but possible under reporting of cases
  2. lower death rates and more precise reporting of cases
  3. Effect of vaccines
- Weekly data from 2020 week 5 to 2021 week 24 ($T = 72$).

Excess Mortality and number of COVID cases in the US II

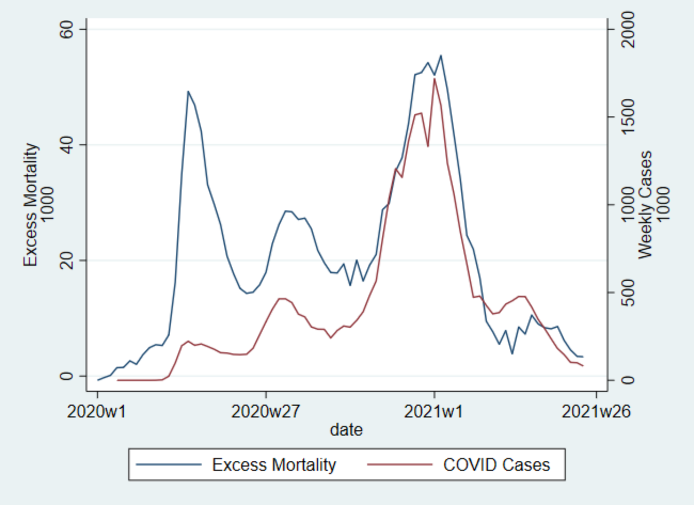

Figure: Excess Mortality and COVID cases in the US. Data from CDC and World In Data.

Excess Mortality and number of COVID cases in the US III

- Excess mortality in the first wave highest, despite relatively "small" number of infections.
- In the second wave less excess mortality.
- Third wave worst in terms of excess mortality and number of cases, but given the cases, mortality could be much higher.
- Can we identify breaks in the relationship between COVID cases and excess mortality?
- Disclaimer: This is an **example** for the use of `xtbreak` and should be treated purely as such!

Unknown Breakdates

Test of 0 vs up to 5 breaks

- Unknown number and dates of breaks.
- Use hypothesis 2 to test for up to 5 breaks: H_0 : no breaks vs $H_1 : 1 \leq s \leq 5$
- xtbreak estimates the breakpoints and then performs the test.

```
. xtbreak test ExcessMortality L1.new_cases, hypothesis(2) breaks(5)

Test for multiple breaks at unknown breakdates
(Bai & Perron. 1998. Econometrica)
H0: no break(s) vs. H1: 1 <= s <= 5 break(s)
```

		Bai & Perron Critical Values		
	Test Statistic	1% Critical Value	5% Critical Value	10% Critical Value
UDmax(tau)	130.10	12.37	8.88	7.46

```
Estimated break points:  2020w20  2021w8
* evaluated at a level of 0.95.
```

- Reject hypothesis of no breaks, 2 breaks identified.

Unknown Breakdates

Test for no vs 2 breaks

- We can now test for no vs. 2 breaks.

```
. xtbreak test ExcessMortality L1.new_cases, hypothesis(1) breaks(2)
Test for multiple breaks at unknown breakdates
(Bai & Perron. 1998. Econometrica)
H0: no break(s) vs. H1: 2 break(s)
```

		Bai & Perron Critical Values		
	Test Statistic	1% Critical Value	5% Critical Value	10% Critical Value
supW(tau)	130.10	9.36	7.22	6.28

```
Estimated break points:  2020w20  2021w8
```

- Test statistic and estimated break dates are (as expected) the same.

Estimation of breakdates

- So far we tested if there are breaks.
- Estimating the breakpoints allows to construct confidence intervals.

```
. xtbreak estimate ExcessMortality L1.new_cases, breaks(2)
 Estimation of break points
                                                    T     =      72
                                                    SSR   =   1519.53
----------------------------------------------------------------------
  #      Index    Date                      [95% Conf. Interval]
----------------------------------------------------------------------
  1        16    2020w20                    2020w19       2020w21
  2        56     2021w8                     2021w7        2021w9
----------------------------------------------------------------------

. xtbreak estimate ExcessMortality L1.new_cases, breaks(2) showindex
 Estimation of break points
                                                    T     =      72
                                                    SSR   =   1519.53
----------------------------------------------------------------------
  #      Index    Date             [95% Conf. Interval]
----------------------------------------------------------------------
  1        16    2020w20           15            17
  2        56     2021w8           55            57
----------------------------------------------------------------------
```

Confidence Intervals

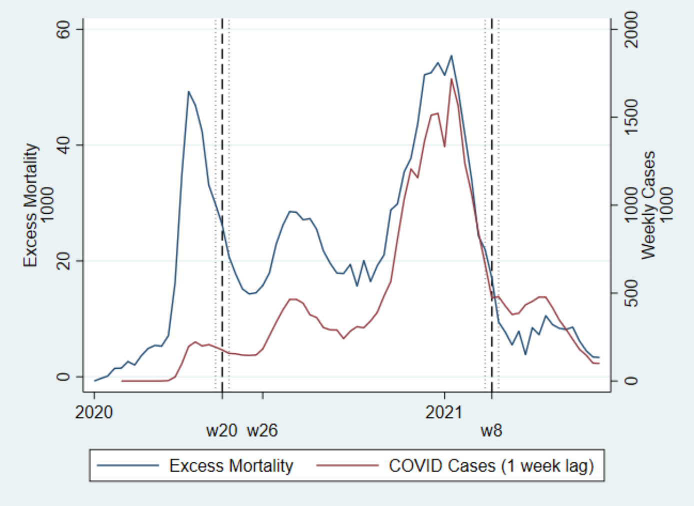

Figure: Excess Mortality and COVID cases in the US. Data from CDC and World In Data. 95% confidence interval marked by dotted lines.

Postestimation

- `xtbreak estimate` has several post estimation features:
  - `estat indicator` creates indicator variable with $1, .., \hat{s}+1$ for each segment.
  - `estat split` *varlist* creates a new variable for each segment (breaks). List of new variable names saved in r(varlist).
- To see how β_1 changes we can run a simple OLS regression after using `estat split`.

Postestimation

```
. qui xtbreak estimate ExcessMortality L1.new_cases, breaks(2)

. estat split
New variables created: L_new_cases1 L_new_cases2 L_new_cases3

. reg ExcessMortality `r(varlist)´

      Source |       SS           df       MS      Number of obs   =        72
-------------+----------------------------------   F(3, 68)        =    218.95
       Model |  14678.1401         3  4892.71336   Prob > F        =    0.0000
    Residual |  1519.52511        68  22.3459576   R-squared       =    0.9062
-------------+----------------------------------   Adj R-squared   =    0.9020
       Total |  16197.6652        71   228.13613   Root MSE        =    4.7272

------------------------------------------------------------------------------
ExcessMort~y | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
L_new_cases1 |   .1517681   .0106782    14.21   0.000     .1304601    .1730761
L_new_cases2 |   .0284604   .0013397    21.24   0.000     .0257872    .0311337
L_new_cases3 |  -.0034063   .0040829    -0.83   0.407    -.0115537    .0047411
       _cons |    8.91028   .9357773     9.52   0.000     7.042966    10.77759
------------------------------------------------------------------------------
```

Disclaimer: This is an **example** for the use of `xtbreak` and should be treated purely as such!

Conclusion

- Introduced new community contributed package called `xtbreak`
- Estimation and test for breaks at known and unknown points in time.
- Three tests for time series and panel data included, following Bai and Perron (1998); Ditzen et al. (2021); Karavias et al. (2021).
- Estimation and tests can be applied to time series and panel models, including models with cross-section dependence.
- For the ado files, further details and examples see our GitHub page or

```
net install xtbreak, from(https://janditzen.github.io/xtbreak/)
```

References I

Andrews, D. W. K. 1993. Tests for Parameter Instability and Structural Change With Unknown Change Point. Econometrica 61(4): 821–856.

Antoch, J., J. Hanousek, L. Horvath, M. Huskova, and S. Wang. 2019. Structural breaks in panel data: Large number of panels and short length time series. Econometric Reviews 38(7).

Bai, B. Y. J., and P. Perron. 1998. Estimating and Testing Linear Models with Multiple Structural Changes. Econometrica, 66(1): 47–78.

Bai, J. 1997. Estimation of a change point in multiple regression models. Review of Economics and Statistics 79(4): 551–560.

Bai, J., and P. Perron. 2003. Computation and analysis of multiple structural change models. Journal of Applied Econometrics 18(1): 1–22.

Ditzen, J., Y. Karavias, and J. Westerlund. 2021. Testing for Multiple Structural Breaks in Panel Data .

References II

Hidalgo, J., and M. Schafgans. 2017. Inference and testing breaks in large dynamic panels with strong cross sectional dependence. Journal of Econometrics 96(2).

Karavias, Y., J. Westerlund, and P. Narayan. 2021. Structural Breaks in Interactive Effects Panels and the Stock Market Reaction to COVID-19 .

Wachter, S. D., and E. Tzavalis. 2012. Detection of structural breaks in linear dynamic panel data models. Computational Statistics & Data Analysis .

Test Hypothesis 1 back

No break vs. s breaks

$$H_0 : \delta_1 = \delta_2 = ... = \delta_{s+1} \text{ vs } H_1 : \delta_k \neq \delta_j \text{ for some } j \neq k$$

- Wald test with test statistic:

$$F_T(\tau_s^0) = \frac{N(T-p-(s+1)q)-p-(s+1)q}{sq} \hat{\delta}' R' \left(R \hat{V}(\hat{\delta}) R' \right)^{-1} R\hat{\delta}$$

- R imposes the restrictions such that $R\delta' = (\delta_1' - \delta_2', ..., \delta_s' - \delta_{s+1})'$.
- $\hat{V}(\hat{\delta})$ is an estimate of the variance.

Test Hypothesis 1 ▸ back

No break vs. s breaks

- If the break dates are known, then (Andrews, 1993)

$$F_T(\tau) \sim \chi^2(sq).$$

- If the break dates are unknown, then *supF* test statistic is used:

$$\sup F_T(s,q) = \sup_{\tau \in \tau_\eta} F_T(\tau, q)$$

- τ_ϵ is a subset of $[0, T]^s$ and represent all possible combination of break points with a minimal length of each set of η.
- Asymptotic critical values depending on the number of breaks s and regressors q are given in Bai and Perron (1998, Table 1).

Test Hypothesis 2 back

No break vs. $1 \leq s \leq s^*$ breaks

- Test if a maximum of s^* breaks occurs.
- "Double Maximum" test, where the maximum of the test using hypothesis 1 for the number of breaks between 1 and s^* is taken.

$$\text{WDmax}F_T(s,q) = \max_{1 \leq s \leq s^*} \left\{ \frac{c_{\alpha,1,q}}{c_{\alpha,s,q}} \sup_{\tau \in \tau_\eta} F_T(\tau, q) \right\}$$

- $c_{\alpha,s,q}$ is the critical value at a level of α for s breaks and q regressors.
- Asymptotic critical values depending on the number of breaks s and regressors q are given in Bai and Perron (1998, Table 1).

Test Hypothesis 3 ▸ back

s breaks vs. $s+1$ breaks

- Idea: test each s segments for an additional break within the segment.

$$F(s+1|s) = \frac{SSR(\hat{T}_1, ..., \hat{T}_s) - \min_{1 \leq j \leq s+1} \left\{ \inf_{\tau \in \Lambda_{j,\eta}} SSR(\hat{T}_1, ..., \hat{T}_{j-1}, \tau, \hat{T}_j, ..., \hat{T}_s) \right\}}{\hat{\sigma}_s^2}$$

$$\Lambda_{j,\eta} = \left\{ \tau; \hat{T}_{j-1} + \left(\hat{T}_j - \hat{T}_{j-1} \right) \eta \leq \tau \leq \hat{T}_j - \left(\hat{T}_j - \hat{T}_{j-1} \right) \eta \right\}$$

$$\hat{\sigma}_s^2 = \frac{SSR(\hat{T}_1, ..., \hat{T}_s)}{N(T-1) - sq - p}$$

$$SSR(\hat{T}_1, ..., \hat{T}_{s+1}) = \min_{\tau \in \tau_\eta} SSR(\tau)$$

- Looks complicated.... but it is essentially the difference of the minimum of combinations of the SSR with s and $s+1$ breaks.
- Asymptotic critical values depending on the number of breaks s and regressors q are given in Bai and Perron (1998, Table 2).