xtbreak: Estimation of and testing for structural breaks in Stata
US Stata Conference 2021

Jan Ditzen\textsuperscript{1}, Yiannis Karavias\textsuperscript{2}, Joakim Westerlund\textsuperscript{3}

\textsuperscript{1}Free University of Bozen-Bolzano, Bozen, Italy
www.jan.ditzen.net, jan.ditzen@unibz.it

\textsuperscript{2}University of Birmingham, UK
https://sites.google.com/site/yianniskaravias/
i.karavias@bham.ac.uk

\textsuperscript{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:

$$y_{i,t} = x'_{i,t} \beta + w'_{i,t} \delta_1 + u_{i,t}, \quad t = 1, \ldots, T_1, \quad i = 1, \ldots, N$$

$$y_{i,t} = x'_{i,t} \beta + w'_{i,t} \delta_2 + u_{i,t}, \quad t = T_1 + 1, \ldots, T_2$$

$$\cdots$$

$$y_{i,t} = x'_{i,t} \beta + w'_{i,t} \delta_{s+1} + u_{i,t}, \quad t = T_s, \ldots, T$$

- $\tau_s = (T_1, T_2, \ldots, T_s)$ 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:

\[ Y_i = X_i\beta + W_i(\tau_s)\delta + U_i \]  

(1)

- where \( Y_i = (y_{i,1}, \ldots, y_{i,T})' \), \( W_i = (w_{i,1}, \ldots, w_{i,T})' \), \( \delta = (\delta'_1, \ldots, \delta'_{s+1})' \)

and:

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

- \( w_{s,i} \) is \( (T_s \times q) \).

- Aim: Estimation and testing of breaks \( \tau_s = (T_1, T_2, \ldots, T_s) \).
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 $SSR = SSR(1, 2) + SSR(3, T)$. 

<table>
<thead>
<tr>
<th>Start</th>
<th>1</th>
<th>2</th>
<th>End</th>
<th>\cdots</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>\ddots</td>
<td>$SSR(1, 2)$</td>
<td>$SSR(1, 3)$</td>
<td>\cdots</td>
<td>$SSR(1, T)$</td>
</tr>
<tr>
<td>2</td>
<td>\ddots</td>
<td>\ddots</td>
<td>$SSR(2, 3)$</td>
<td>\cdots</td>
<td>$SSR(2, T)$</td>
</tr>
<tr>
<td>3</td>
<td>\ddots</td>
<td>\ddots</td>
<td>\ddots</td>
<td>$SSR(3, T)$</td>
<td></td>
</tr>
<tr>
<td>\vdots</td>
<td>\ddots</td>
<td>\ddots</td>
<td>\ddots</td>
<td>\ddots</td>
<td>\ddots</td>
</tr>
<tr>
<td>T</td>
<td>\ddots</td>
<td>\ddots</td>
<td>\ddots</td>
<td>\ddots</td>
<td>\ddots</td>
</tr>
</tbody>
</table>
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):

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

- where $\hat{\Omega}_X = \frac{1}{N T} \sum_{i=1}^{N} X_i' X_i$, $\hat{\Phi}_X = \frac{1}{N T} \sum_{i=1}^{N} \hat{\sigma}_{\epsilon,i}^2 X_i' X_i$. 
Three tests for breaks

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

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(varlist
ts) noconstant breakconstant vce(ssr|hac|nw)
```

1This command is work in progress. Options, functions and results might change.
If the break is estimated, then break_point_options are:

\[ \text{breaks(real)} \quad \text{minlength(real)} \quad \text{error(real)} \]

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

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

\[ \text{hypothesis(1|2|3)} \quad \text{breaks(real)} \quad \text{minlength(real)} \quad \text{level(real)} \quad \text{error(real)} \quad \text{wdmax} \]

- \text{hypothesis()} which hypothesis to test.
- \text{breaks(real)} number of breaks.
- \text{level} which level the weighted (only hypothesis 2) test is evaluated at.
- \text{wdmax} weighted max test (only hypothesis 2).
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](https://github.com).
Excess Mortality and number of COVID cases in the US

- 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 xtbreset 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: $H_0 : \text{no breaks} \ vs \ H_1 : 1 \leq s \leq 5$
- `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)
$H_0: \text{no break(s)} \ vs. \ H_1: 1 \leq s \leq 5 \ \text{break(s)}$

<table>
<thead>
<tr>
<th></th>
<th>Bai &amp; Perron</th>
<th>Critical Values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1% Critical</td>
<td>5% Critical</td>
</tr>
<tr>
<td>Test Statistic</td>
<td>Value</td>
<td>Value</td>
</tr>
<tr>
<td>UDmax(tau)</td>
<td>130.10</td>
<td>12.37</td>
</tr>
</tbody>
</table>

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.

```
. xtbbreak 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)

<table>
<thead>
<tr>
<th>Test Statistic</th>
<th>Bai &amp; Perron Critical Values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1% Critical Value</td>
</tr>
<tr>
<td>supW(tau)</td>
<td>130.10</td>
</tr>
</tbody>
</table>

Estimated break points: 2020w20 2021w8

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

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

```
. xtbbreak 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
```

```
. xtbbreak 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 \(\beta_1\) changes we can run a simple OLS regression after using `estat split`.
Postestimation

\[ \text{. qui xtbreak estimate ExcessMortality L1.new_cases, breaks(2)} \]
\[ \text{. estat split} \]
New variables created: L_new_cases1 L_new_cases2 L_new_cases3

\[ \text{. reg ExcessMortality `r(varlist)'} \]

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 72</th>
<th>F(3, 68) = 218.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>14678.1401</td>
<td>3</td>
<td>4892.71336</td>
<td></td>
<td>0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>1519.52511</td>
<td>68</td>
<td>22.3459576</td>
<td>R-squared = 0.9062</td>
<td>0.9020</td>
</tr>
<tr>
<td>Total</td>
<td>16197.6652</td>
<td>71</td>
<td>228.13613</td>
<td>Root MSE = 4.7272</td>
<td></td>
</tr>
</tbody>
</table>

| ExcessMortality | 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](https://janditzen.github.io/xtbreak/)
  
  ```stata
  net install xtbreaK, from(https://janditzen.github.io/xtbreak/)
  ```
References


Test Hypothesis 1

No break vs. $s$ breaks

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

- Wald test with test statistic:

$$F_T (\tau_0^s) = \frac{N(T - p - (s + 1)q) - p - (s + 1)q}{sq} \hat{\delta}' R' \left( R \hat{V}(\hat{\delta}) R' \right)^{-1} R \hat{\delta}$$

- $R$ imposes the restrictions such that $R \delta' = (\delta'_1 - \delta'_2, \ldots, \delta'_s - \delta'_{s+1})'$.

- $\hat{V}(\hat{\delta})$ is an estimate of the variance.
Test Hypothesis 1

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 \( \sup F \) test statistic is used:
  \[ \sup F_T(s, q) = \sup_{\tau \in \tau_\eta} F_T(\tau, q) \]

- \( \tau_\epsilon \) is a subset of \([0, T]^s\) and represent all possible combination of break points with a minimal length of each set of \( \eta \).

- 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

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.

$$WDmaxF_T(s, q) = \max_{1 \leq s \leq s^*} \left\{ \frac{c_{\alpha,1,q}}{c_{\alpha,s,q}} \sup_{\tau \in \mathcal{T}_\eta} F_T(\tau, q) \right\}$$

- $c_{\alpha,s,q}$ is the critical value at a level of $\alpha$ 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

$s$ breaks vs. $s + 1$ breaks

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

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

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

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

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

- 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).