Tests for multiple changes in linear regression models

Miriam Marušiaková, Marie Hušková

Department of Statistics
Charles University in Prague

ROBUST 2010, Králíky, 2.2. 2010
1. Introduction

2. Regression models

3. Test statistics

4. Simulation results

5. Applications

6. Conclusion
1. Introduction
2. Regression models
3. Test statistics
4. Simulation results
5. Applications
6. Conclusion
Introduction

Change point analysis
- Important topic in statistical and econometric research
  - Interesting theoretical problems
  - Applications in many fields
- Data sample - the model might change during the observational period

Problems
1. Test whether change(s) occurred or not
2. Estimate the number of changes and their locations
## Hypothesis testing

### Test procedures
- Max-type test statistics (CUSUM procedures)
- MOSUM type tests
- Sum-type or Bayesian type statistics
- $F$-type tests (Bai and Perron, 1998), $M$-type tests

### Approximations to critical values
- Limit null distribution
- Residual bootstrap (Antoch and Hušková, 2001)
Permutation principle

- Test statistic - expressible through the partial sums of residuals; under $H_0$ - partial sums of errors
- If $e_i$ i.i.d. $\Rightarrow (e_1, \ldots, e_n)'$ and $(e_{R_1}, \ldots, e_{R_n})'$ have the same distribution, where
  $\mathbf{R} = (R_1, \ldots, R_n)'$ is a random permutation of $(1, \ldots, n)'$ such that
  $$P(\mathbf{R} = \mathbf{r}) = \frac{1}{n!}$$
- Permutation version of the test statistic:
  errors $e_i$ replaced by permuted residuals $\hat{e}_{R_i}$
- We study conditional distribution of the permutation test statistic, given the observations $\mathbf{Y}$
How do permutation tests work?

1. From the data sample we calculate the test statistic $F_n$.
2. We generate a random permutation $R = r$ of $(1, \ldots, n)'$.
3. We calculate a permutational version $F_n(r)$ of our test statistic for $r$.
4. We repeat the last two steps $N$ times.
5. We obtain the empirical distribution of $F_n(R)$ and calculate its empirical $\alpha$-quantile $c_\alpha$.
6. We reject the null hypothesis if $F_n > c_\alpha$. 
Outline

1. Introduction
2. Regression models
3. Test statistics
4. Simulation results
5. Applications
6. Conclusion

Tests for multiple changes in linear regression models
Linear regression model

\[ Y_i = x_i'\beta_1 + e_i \quad i = 1, \ldots, t_1 \]
\[ Y_i = x_i'\beta_2 + e_i \quad i = t_1 + 1, \ldots, t_2 \]
\[ \vdots \]
\[ Y_i = x_i'\beta_{m+1} + e_i \quad i = t_m + 1, \ldots, n, \]

- unknown change points \( t_1, \ldots, t_m \)
- unknown parameters \( \beta_j = (\beta_{j1}, \ldots, \beta_{jq})' \)
- regressors \( x_i = (x_{i1}, \ldots, x_{iq})' \)
- errors \( e_1, \ldots, e_n \)
Non-trending regression and independent errors

(A1) \( e_i \) are i.i.d. with \( E e_1 = 0, E e_1^2 = \sigma^2, E |e_1|^{2+\Delta} < \infty, \Delta > 0. \)

(A2) \( e_i \) are independent of \( x_j \) for all \( i \) and \( j \).

(A3)

\[
\frac{1}{n} \sum_{i=1}^{\lfloor ns \rfloor} x_i x_i' \overset{\text{def}}{=} \frac{1}{n} C_{\lfloor ns \rfloor} \xrightarrow{P} s C > 0 \text{ uniformly in } s, 0 < s \leq 1.
\]

(A4) For all \( \eta > 0 \), there exists a \( D > 0 \), such that for all large \( n \),

\[
P \left( \frac{1}{n} \sum_{i=1}^{n} \|x_i\|^{2+\Delta} > D \right) < \eta.
\]

(A5) The true change points satisfy

\[
t_j^0 = \lfloor n \lambda_j^0 \rfloor, \quad 0 = \lambda_0^0 < \lambda_1^0 < \ldots < \lambda_{m+1}^0 = 1.
\]
Autoregressive model of order $p$

\[ x_i' = (1, Y_{i-1}, Y_{i-2}, \ldots, Y_{i-p}) \]

(B1) $e_i$ are i.i.d. with $Ee_1 = 0$, $Ee_1^2 = \sigma^2$ and $E|e_i|^4 < \infty$. 

(B2) $Y_1, \ldots, Y_p$ are independent of $e_{p+1}, \ldots, e_n$ and the roots of $t^p - \beta_{j1}t^{p-1} - \cdots - \beta_{jp}$ are $|t| < 1, \quad j = 1, \ldots, m + 1$. 

(B3) The initial values $Y_1, \ldots, Y_p$ satisfy:

\[
(Y_p - \mu_1, \ldots, Y_1 - \mu_1)' = \sum_{j=0}^{\infty} B^j e_{p-j},
\]

\[
B = \begin{pmatrix}
\beta_{11} & \cdots & \beta_{1p} \\
I_{p-1} & \cdots & 0
\end{pmatrix} ;
\]

$e_k = (e_k, 0, \ldots, 0)'$; $\mu_1 = EY_i, i \leq t_1$. 

Miriam Marušiaková

Tests for multiple changes in linear regression models
Limit behavior of regressors

- under $H_0$ ($m = 0$)

\[
\frac{C_{|ns|}}{n - p} = \frac{1}{n - p} \sum_{i=p+1}^{\lfloor ns \rfloor} x_i x'_i \xrightarrow{P} sC > 0,
\]

- under $H_A$ ($m = k$)

\[
\frac{C_{|ns|}}{n - p} \xrightarrow{P} Q_s = \sum_{j=1}^{k+1} \left( \min\{s, \lambda_j^0\} - \lambda_{j-1}^0 \right) C^{(j)} I\{s \geq \lambda_{j-1}^0\},
\]

where $C^{(j)} > 0$, $j = 1, \ldots, k + 1$ and $s \in (0, 1]$. 
Trending regression

(C1) \( e_i \) are i.i.d. with common symmetric pdf \( F \).

(C6) The regressors \( x_i = h(i/n) \) satisfy:

\[
\begin{align*}
  h_1(x) &= 1, \quad 0 \leq x \leq 1, \\
  h_2(\cdot), \ldots, h_q(\cdot) &\text{ are continuously diff. functions on } [0, 1] \\
\end{align*}
\]
and as \( n \to \infty \),

\[
\frac{1}{n} C_{[ns]} = \frac{1}{n} \sum_{i=1}^{[ns]} h(i/n)h'(i/n) \to C(s) = \int_0^s h(x)h'(x)dx > 0
\]

uniformly in \( s \in (0, 1] \), \( C(s) \) is strictly increasing in \( s \).
Trending regression (cont.)

The score function $\psi$ and $\lambda(t) = -\int \psi(e - t)dF(e), \ t \in \mathbb{R}$:

(C2) $\psi$ is non-decreasing, antisymmetric: $\psi(x) = -\psi(-x)$.

(C3) $\sigma^2(\psi) = \int \psi^2(x)dF(x) \in (0, \infty)$.

$a > 0, D_1 > 0, D_2 > 0$

$$\int (\psi(x - s_2) - \psi(x - s_1))^2 dF(x) \leq D_1 |s_2 - s_1|^a$$

$$|s_j| \leq D_2, \ j = 1, 2.$$  

(C4) $\lambda'(\cdot)$ exists and is Lipschitz in a neighbourhood of zero, $\lambda(0) = 0$ and $\lambda'(0) > 0$.

(C5) Either

(i) $\psi$ is bounded, or

(ii) $\psi(x) = x$ and $\int |x|^{2+\Delta}dF(x) < \infty$. 

Miriam Marušiaková

Tests for multiple changes in linear regression models
Outline

1. Introduction
2. Regression models
3. Test statistics
4. Simulation results
5. Applications
6. Conclusion

Miriam Marušiaková
Tests for multiple changes in linear regression models
**F-type test statistic**

Test for no change versus \( k \) (maximum \( M \)) changes

\[
F_n^\varepsilon(k, q) = \max_{(t_1, \ldots, t_k) \in T_{\varepsilon,k}} F_n(t_1, \ldots, t_k; q),
\]

\[
DF_n^\varepsilon(M; q) = \max_{k=1,\ldots,M} \max_{(t_1, \ldots, t_k) \in T_{\varepsilon,k}} F_n(t_1, \ldots, t_k; q),
\]

where

\[
F_n(t_1, \ldots, t_k; q) = \frac{1}{kq} \frac{SSR_0 - SSR_k(t)}{\hat{\sigma}_{n,k}^2(t)}
\]

\[
\hat{\sigma}_{n,k}^2(t) = \frac{SSR_k(t)}{(n - (k + 1) q)} \xrightarrow{P} \sigma^2 \quad \text{under } H_0
\]

\[
t = (t_1, \ldots, t_k)'
\]

\[
T_{\varepsilon,k} = \{(t_1, \ldots, t_k) : t_{j+1} - t_j \geq \lfloor n\varepsilon \rfloor, \forall j = 0, \ldots, k\}
\]
F-type test - equivalent expression

\[ SSR_0 - SSR_k(t) = \sum_{j=1}^{k+1} \left( \sum_{i=t_{j-1}+1}^{t_j} x_i \hat{e}_i \right)' C_{t_{j-1},t_j}^{-1} \left( \sum_{i=t_{j-1}+1}^{t_j} x_i \hat{e}_i \right) \]

\[ H_0 \equiv - \left( \sum_{i=1}^{n} x_i e_i \right)' C_n^{-1} \left( \sum_{i=1}^{n} x_i e_i \right) \]

\[ + \sum_{j=1}^{k+1} \left( \sum_{i=t_{j-1}+1}^{t_j} x_i e_i \right)' C_{t_{j-1},t_j}^{-1} \left( \sum_{i=t_{j-1}+1}^{t_j} x_i e_i \right) \]

\[ \hat{e}_i = Y_i - x_i' C_n^{-1} \sum_{i=1}^{n} x_i Y_i \overset{H_0}{=} e_i - x_i' C_n^{-1} \sum_{i=1}^{n} x_i e_i \]
\( \sum_{i=1}^{n} \rho \left( Y_i - x'_i/\beta \right) \leq \min! \), which usually leads to

\[ \sum_{i=1}^{n} \psi \left( Y_i - x'_i/\beta \right) x_{ij} = 0, \quad j = 1, \ldots, q, \quad \psi = \rho' \]

**Jurečková and Sen (1996)**

\[ \hat{\beta}_n(\psi) - \beta = C_n^{-1} \frac{1}{\lambda'(0)} \sum_{i=1}^{n} x_i \psi(e_i) + o_p(n^{-1/2}). \]

\[ \hat{\beta}_n(\psi) \sim AN \left( \beta, \frac{\sigma^2(\psi)}{(\lambda'(0))^2} C_n^{-1} \right) \]
$M$-type test statistic

\[
F_n^\varepsilon(\psi, k, q) = \max_{(t_1, \ldots, t_k) \in T_{\varepsilon, k}} F_n(\psi, t_1, \ldots, t_k; q)
\]

\[
F_n(\psi, t_1, \ldots, t_k; q) = \frac{1}{kq \tilde{\sigma}_{n,k}^2(\psi, t)}
\]

\[
\sum_{j=1}^{k+1} \left( \sum_{i=t_{j-1}+1}^{t_j} x_i \hat{e}_i(\psi) \right)' C_{t_{j-1}, t_j}^{-1} \left( \sum_{i=t_{j-1}+1}^{t_j} x_i \hat{e}_i(\psi) \right)
\]

\[
\tilde{\sigma}_{n,k}^2(\psi, t) - \sigma^2(\psi) = o_p(1), \quad \hat{e}_i(\psi) = \psi \left( Y_i - x_i' \hat{\beta}_n(\psi) \right), \quad i = 1, \ldots, n
\]
Non-trending regression and AR models

Let $W(\cdot)$ be a vector of $q$ independent standard Wiener processes on $[0, 1]$. Under $H_0 (m = 0)$ and given assumptions, as $n \to \infty$,

$$DF_n^\varepsilon(M; q) \xrightarrow{d} \max_{1 \leq k \leq M} \sup_{(\lambda_1, \ldots, \lambda_k) \in \Lambda_{\varepsilon,k}} F(\lambda_1, \ldots, \lambda_k; q)$$

with

$$F(\lambda_1, \ldots, \lambda_k; q) \overset{\text{def}}{=} \frac{1}{kq} \sum_{j=1}^{k} \frac{\|\lambda_j W(\lambda_{j+1}) - \lambda_{j+1} W(\lambda_j)\|^2}{\lambda_j \lambda_{j+1} (\lambda_{j+1} - \lambda_j)}$$

and the supremum is taken over the set

$$\Lambda_{\varepsilon,k} = \{ (\lambda_1, \ldots, \lambda_k) : \lambda_{j+1} - \lambda_j \geq \varepsilon, \forall j = 0, \ldots, k \}.$$
Trending regression

\[
DF_n^{\varepsilon}(\psi, M, q) \xrightarrow{d} \max_{1 \leq k \leq M} \left\{ \sup_{(\lambda_1, \ldots, \lambda_k) \in \Lambda_{\varepsilon, k}} F_h(\lambda_1, \ldots, \lambda_k; q) \right\},
\]

where

\[
F_h(\lambda_1, \ldots, \lambda_k; q) = \frac{1}{kq} \left\{ -W_h(1)'C(1)^{-1}W_h(1) + \sum_{j=1}^{k+1} (W_h(\lambda_j) - W_h(\lambda_{j-1}))' (C(\lambda_j) - C(\lambda_{j-1}))^{-1} (W_h(\lambda_j) - W_h(\lambda_{j-1})) \right\},
\]

\[
W_h(t) = \int_0^t h(x)dW(x),
\]

\[
\Lambda_{\varepsilon, k} = \{(\lambda_1, \ldots, \lambda_k) : \lambda_{j+1} - \lambda_j \geq \varepsilon, \forall j = 0, \ldots, k\}.
\]
Permutation version of $F$-type test statistic

$$DF_n^\varepsilon(M, q, R) = \max_{1 \leq k \leq M} \max_{(t_1, \ldots, t_k) \in T_{\varepsilon, k}} F_n(t_1, \ldots, t_k; q, R)$$

with

$$F_n(t_1, \ldots, t_k; q, R) = \frac{1}{kq \, \hat{\sigma}_n^2} \left[ - \left( \sum_{i=1}^{n} x_i \hat{e}_{R_i} \right)' C_n^{-1} \left( \sum_{i=1}^{n} x_i \hat{e}_{R_i} \right) + \sum_{j=1}^{k+1} \left( \sum_{i=t_{j-1}+1}^{t_j} x_i \hat{e}_{R_i} \right)' C_{t_{j-1}, t_j}^{-1} \left( \sum_{i=t_{j-1}+1}^{t_j} x_i \hat{e}_{R_i} \right) \right],$$

where $\hat{e}_{R_1}, \ldots, \hat{e}_{R_n}$ are permuted $L_2$-residuals.
Permutation version of $M$-type test statistic

\[
DF_{n}^{\varepsilon}(\psi, M, q, R) = \max_{1 \leq k \leq M} \max_{(t_1, \ldots, t_k) \in T_{\varepsilon, k}} F_{n}(\psi, t_1, \ldots, t_k; q, R)
\]

with

\[
F_{n}(\psi, t_1, \ldots, t_k; q, R) = \frac{1}{kq \hat{\sigma}_{n}^{2}(\psi)} \left[ - \left( \sum_{i=1}^{n} h(i/n) \hat{e}_{R_{i}}(\psi) \right)' C_{n}^{-1} \left( \sum_{i=1}^{n} h(i/n) \hat{e}_{R_{i}}(\psi) \right) \right. \\
+ \sum_{j=1}^{k+1} \left( \sum_{i=t_{j-1}+1}^{t_{j}} h(i/n) \hat{e}_{R_{i}}(\psi) \right)' C_{t_{j-1}, t_{j}}^{-1} \left( \sum_{i=t_{j-1}+1}^{t_{j}} h(i/n) \hat{e}_{R_{i}}(\psi) \right) \right],
\]

where $\hat{e}_{R_{1}}(\psi), \ldots, \hat{e}_{R_{n}}(\psi)$ are permuted $M$-residuals.
Theorem

Let \( Y = (Y_1, \ldots, Y_n)' \) follow the model with \( m \geq 0 \) changes. Then under the considered assumptions, for arbitrary \( x \in \mathbb{R} \), as \( n \to \infty \),

\[
P\left( \max_{1 \leq k \leq M} \max_{(t_1, \ldots, t_k) \in T_{\epsilon,k}} F_n(t_1, \ldots, t_k, q, R) \leq x \mid Y \right) \xrightarrow{P} P\left( \max_{1 \leq k \leq M} \sup_{(\lambda_1, \ldots, \lambda_k) \in \Lambda_{\epsilon,k}} F(\lambda_1, \ldots, \lambda_k) \leq x \right),
\]

where \( F(\lambda_1, \ldots, \lambda_k) \) is a random variable such that

\[
F_n(t_1, \ldots, t_k, q) \xrightarrow{d} F(\lambda_1, \ldots, \lambda_k) \quad \text{under } H_0, \quad n \to \infty.
\]
Outline

1. Introduction
2. Regression models
3. Test statistics
4. Simulation results
5. Applications
6. Conclusion

Miriam Marušiaková  Tests for multiple changes in linear regression models
QQ plot for the null distr. of $DF_n^\varepsilon(q, M)$ vs. $DF_n^\varepsilon(q, M, R)$

$n = 180, M = 2, \varepsilon = 0.15, q = 1$
SPC plots for $DF_n^\varepsilon(q, M)$ with respect to perm. distr.

Size-power-curves plots; $n = 180, M = 2, \varepsilon = 0.15, q = 1$
Outline

1 Introduction
2 Regression models
3 Test statistics
4 Simulation results
5 Applications
6 Conclusion

Tests for multiple changes in linear regression models
Global temperature data

http://www.metoffice.gov.uk/research/hadleycentre/obsdata
Fitted models

Segmented model with linear / quadratic trend with \( k \) changes

\[
Y_i = \beta_{j0} + \beta_{j1} \left( \frac{i}{n} \right) + e_i
\]

\[
Y_i = \beta_{j0} + \beta_{j1} \left( \frac{i}{n} \right) + \beta_{j2} \left( \frac{i}{n} \right)^2 + e_i
\]

\[i = t_{j-1} + 1, \ldots, t_j, \quad j = 1, \ldots, k + 1, \quad k = 1, 2, 3, 4\]

\[n = 159 \text{ (years 1850, \ldots, 2008)}\]

\[n \varepsilon = \lfloor 159 \ast 0.05 \rfloor = 7 \text{ years}\]
### F-type tests

<table>
<thead>
<tr>
<th>q</th>
<th>$F_n^\varepsilon(1, q)$</th>
<th>$F_n^\varepsilon(2, q)$</th>
<th>$F_n^\varepsilon(3, q)$</th>
<th>$F_n^\varepsilon(4, q)$</th>
<th>$F_n^\varepsilon(5, q)$</th>
<th>$DF_n^\varepsilon(5, q)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>49.6</td>
<td>47.7</td>
<td>40.7</td>
<td>37.8</td>
<td>33.6</td>
<td>49.6</td>
</tr>
<tr>
<td>3</td>
<td>14.5</td>
<td>20.6</td>
<td>17.4</td>
<td>16.0</td>
<td>14.3</td>
<td>20.6</td>
</tr>
</tbody>
</table>

### Permutation critical values

<table>
<thead>
<tr>
<th></th>
<th>90%</th>
<th>95%</th>
<th>99%</th>
<th>90%</th>
<th>95%</th>
<th>99%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_n^\varepsilon(1, q)$</td>
<td>5.893</td>
<td>6.921</td>
<td>8.937</td>
<td>4.938</td>
<td>5.563</td>
<td>6.768</td>
</tr>
<tr>
<td>$F_n^\varepsilon(2, q)$</td>
<td>5.363</td>
<td>5.965</td>
<td>6.876</td>
<td>4.589</td>
<td>5.037</td>
<td>5.895</td>
</tr>
<tr>
<td>$F_n^\varepsilon(3, q)$</td>
<td>5.014</td>
<td>5.497</td>
<td>6.649</td>
<td>4.313</td>
<td>4.765</td>
<td>5.378</td>
</tr>
<tr>
<td>$F_n^\varepsilon(4, q)$</td>
<td>4.813</td>
<td>5.274</td>
<td>6.340</td>
<td>4.193</td>
<td>4.542</td>
<td>5.182</td>
</tr>
<tr>
<td>$F_n^\varepsilon(5, q)$</td>
<td>4.664</td>
<td>5.068</td>
<td>6.036</td>
<td>4.101</td>
<td>4.380</td>
<td>5.028</td>
</tr>
<tr>
<td>$DF_n^\varepsilon(5, q)$</td>
<td>6.080</td>
<td>6.956</td>
<td>8.937</td>
<td>5.093</td>
<td>5.767</td>
<td>6.768</td>
</tr>
</tbody>
</table>
Segmented models with linear / quadratic trend
<table>
<thead>
<tr>
<th>Outline</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
</tr>
<tr>
<td>2</td>
<td>Regression models</td>
</tr>
<tr>
<td>3</td>
<td>Test statistics</td>
</tr>
<tr>
<td>4</td>
<td>Simulation results</td>
</tr>
<tr>
<td>5</td>
<td>Applications</td>
</tr>
<tr>
<td>6</td>
<td>Conclusion</td>
</tr>
</tbody>
</table>
Conclusion

Tests for detection of multiple changes

- $F$-type tests (Bai and Perron, 1998)
- Generalized $M$-type tests

Models

- Linear regression models with non-trending or trending regressors and independent errors
- Autoregressive models

Approximations to critical values

- Bootstrap with or without replacement
Some important publications

**Bai and Perron (1998).**
*Estimating and testing linear models with multiple structural changes.*  
Econometrica, 66, 47–78.

**Hušková and Picek (2005)**
*Bootstrap in detection of changes in linear regression*  

**Kirch (2006)**
*Resampling Methods for the Change Analysis of Dependent Data*  
PhD thesis, University of Cologne
R DEVELOPMENT CORE TEAM (2008)
R: A language and environment for statistical computing.
R Foundation for Statistical Computing, Vienna, Austria

ZEILEIS, LEISCH, HORNIK AND KLEIBER (2002)
strucchange: An R Package for Testing for Structural Change in Linear Regression Models
Journal of Statistical Software 7, 1 – 38