Ratio Type Statistics for Detection of Changes in Linear Regression Models

Barbora Madurkayová

Department of Probability and Statistics,
Charles University in Prague, Czech Republic
madurka@karlin.mff.cuni.cz

1. Introduction

We assume to have a set of observations $Y_1, \dots, Y_n$ obtained at time ordered points and that these data follow a linear regression model. Particularly, we are interested in studying a situation, where at some time point k^* may occur a change in regression parameters.

$$Y_k = \boldsymbol{h}^T(k/n)\boldsymbol{\beta} + \boldsymbol{h}^T(k/n)\boldsymbol{\delta} I\{k > k^*\} + e_k, \quad k = 1, \dots, n,$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$, $\boldsymbol{\delta} = \boldsymbol{\delta}_n = (\delta_1, \dots, \delta_p)^T$ and $k^* = k_n^*$ are unknown parameters. $\boldsymbol{h}^T(t) = (h_1(t), \dots, h_p(t))^T$ is such that $h_1(t) = 1$ for $t \in [0, 1]$ and $h_j(t),\ j = 2, \dots, p$ are continuously differentiable functions on $[0, 1]$. $e_1, \dots, e_n$ are generally assumed to satisfy the so called functional central limit theorem However, here we are going to focus on a simpler situation, where error terms $e_1, \dots, e_n$ are independent and identically distributed (i.i.d.) random variables, satisfying $\mathrm{E}\, e_k = 0$ and $\mathrm{Var}\, e_k = \sigma^2 > 0$ for $k = 1, \dots, n$.
The basic question, we are trying to answer, is whether a change in regression parameters occurred at some unknown time point k^* or not. Using the above introduced notation, the null hypothesis of no change an be expressed as

$$H_0 : k^* = n$$

We are going to test the null hypothesis against the alternative

$$H_1 : k^* < n,\ \boldsymbol{\delta} \neq \mathbf{0}.$$

2. General form of a ratio type test statistic

One possible form of a ratio test statistic that may be used in change-point analysis, is the following

$$\max_{n\gamma \le k \le n - n\gamma} \frac{\max\limits_{1 \le j \le k} V_{j,k}}{\max\limits_{k \le j \le n} \tilde{V}_{j,k}},$$

where $V_{j,k}$ for $j = 1, \dots, k$ denotes a statistic based on observations $Y_1, \dots, Y_k$ and $\tilde{V}_{j,k}$ for $j = k+1, \dots, n$ is an similar statistic based on observations $Y_{k+1}, \dots, Y_n$.
This type of statistic was studied for example in [Hušková, 2006] or [Horváth et al., 2007], where a ratio statistic of two CUSUM statistic was used for detection of abrupt changes in location. A report about a ratio statistic that may be used for detection of gradual changes in location may be also found in [Madurkayová, 2007].
The basic motivation for studying ratio type test statistics lies in the fact that when computing such test statistic, it is not necessary to estimate variance of the underlying model. This property makes the ratio type statistics a suitable alternative of classical (non-ratio) statistic – most of all in situations, when it is difficult to find a suitable variance estimate.

In the last section we present some applications of the proposed ratio statistic to simulated data from normal and Laplace distributions. In simulation, we took $p = 3$, $h_1(x) = 1$, $h_2(x) = x$ and $h_3(x) = x^2$, $\gamma = 0.1$ and several choices of $\boldsymbol{\delta}$.
On Figure 1 and 2, we can observe the values of ratio

$$Q_k = \frac{\max\limits_{1 \le j \le k} \boldsymbol{S}_{j,k}^T \boldsymbol{C}_{1,k}^{-1} \boldsymbol{S}_{j,k}}{\max\limits_{k \le j \le n} \tilde{\boldsymbol{S}}_{j,k}^T \boldsymbol{C}_{k+1,n}^{-1} \tilde{\boldsymbol{S}}_{j,k}},$$

computed for $k : n\gamma \le k \le n - n\gamma$ with $\gamma = 0.1$. Simulated 95% critical values for each of the two distributions are depicted by the horizontal dotted line.

$\boldsymbol{\delta}$	N(0,1) 90%	Lap(0,1) 90%	N(0,1) 95%	Lap(0,1) 95%
(0,1,0)	0.130	0.086	0.075	0.039
(0,0,1)	0.117	0.089	0.069	0.032
(0,1,1)	0.171	0.098	0.105	0.045
(1,1,1)	0.568	0.485	0.450	0.335
(0,3,0)	0.462	0.356	0.335	0.206
(0,0,3)	0.163	0.081	0.101	0.034
(0,3,3)	0.713	0.619	0.576	0.420
(3,3,3)	1.000	1.000	0.999	0.996
(0,5,0)	0.858	0.750	0.757	0.588
(0,0,5)	0.259	0.173	0.154	0.088
(0,5,5)	0.986	0.960	0.959	0.890
(5,5,5)	1.000	1.000	1.000	1.000

Table 1: *Simulated 90% and 95% rejection rates of U_{100}; $\gamma = 0.1$, $k^* = n/2 = 50$*

3. Ratio type test statistic for change in regression parameters

For situation described above, test statistics based on weighted partial sums of residuals are often used

$$\boldsymbol{S}_k = \sum_{i=1}^{k} \boldsymbol{h}(i/n) \left(Y_i - \boldsymbol{h}^T(i/n) \boldsymbol{b}_n \right), \quad k = 1, \dots, n,$$

where $\boldsymbol{b}_n$ is an L_2-estimate of parameter of the regression parameter based on observations $Y_1, \dots, Y_n$.
Let us denote

$$\boldsymbol{S}_{j,k} = \sum_{i=1}^{j} \boldsymbol{h}(i/n) \left(Y_i - \boldsymbol{h}^T(i/n) \boldsymbol{b}_k \right), \quad j, k = 1, \dots, n,\ j \le k.$$

Here, $\boldsymbol{b}_k$ denotes an L_2-estimate of of the regression parameter based on observations $Y_1, \dots, Y_k$.
And, similarly

$$\tilde{\boldsymbol{S}}_{j,k} = \sum_{i=k+1}^{j} \boldsymbol{h}(i/n) \left(Y_i - \boldsymbol{h}^T(i/n) \tilde{\boldsymbol{b}}_k \right), \quad j, k = 1, \dots, n,\ j > k$$

where $\tilde{\boldsymbol{b}}_k$ is an L_2-estimate of the regression parameter based on observations $Y_{k+1}, \dots, Y_n$.
Further we denote

$$\boldsymbol{C}_{j,k} = \sum_{i=j+1}^{k} \boldsymbol{h}(i/n) \boldsymbol{h}^T(i/n) j, k = 1, \dots, n.$$

Using this notation, we may now define the ratio type test statistic

$$U_n = \max_{n\gamma \le k \le n - n\gamma} \frac{\max\limits_{1 \le j \le k} \boldsymbol{S}_{j,k}^T \boldsymbol{C}_{1,k}^{-1} \boldsymbol{S}_{j,k}}{\max\limits_{k \le j \le n} \tilde{\boldsymbol{S}}_{j,k}^T \boldsymbol{C}_{k+1,n}^{-1} \tilde{\boldsymbol{S}}_{j,k}}.$$

4. Asymptotic properties

Null hypothesis Let us assume that the null hypothesis described above is true. Then, assuming that $e_1, \dots, e_n$ are i.i.d. random variables, satisfying $\mathrm{E}\, e_k = 0$ and $\mathrm{Var}\, e_k = \sigma^2 > 0$ for $k = 1, \dots, n$ and $\boldsymbol{h}^T(t) = (h_1(t), \dots, h_p(t))^T$ is such that $h_1(t) = 1$ for $t \in [0, 1]$ and $h_j(t),\ j = 2, \dots, p$ are continuously differentiable functions on $[0, 1]$ with $\int_0^1 h_j(t) dt = 0,\ j = 2, \dots, p$, as $n \to \infty$, U_n converges in distribution to

$$U = \sup_{\gamma \le t \le 1-\gamma} \frac{\sup\limits_{0 \le s \le t} \boldsymbol{S}^T(s,t) \boldsymbol{C}(0,t)^{-1} \boldsymbol{S}(s,t)}{\sup\limits_{t \le s \le 1} \tilde{\boldsymbol{S}}^T(s,t) \boldsymbol{C}(t,1)^{-1} \tilde{\boldsymbol{S}}(s,t)}.$$

where

$$\boldsymbol{C}(s,t) = \lim_{n \to \infty} \frac{1}{n} \boldsymbol{C}_{[ns],[nt]},$$

$$\boldsymbol{S}(s,t) = \int_0^s \boldsymbol{h}(x) dB_1(x) - \boldsymbol{C}(0,s) \boldsymbol{C}(0,t)^{-1} \int_0^t \boldsymbol{h}(x) dB_1(x),$$

for $s, t \in [0, 1]$ such that $s \le t$,

$$\tilde{\boldsymbol{S}}(s,t) = \int_t^s \boldsymbol{h}(x) dB_2(x) - \boldsymbol{C}(t,s) \boldsymbol{C}(t,1)^{-1} \int_t^1 \boldsymbol{h}(x) dB_2(x).$$

for $s, t \in [0, 1]$ such that $t \le s$ and where $\{B_1(u), 0 \le u \le 1\}$ and $\{B_2(u), 0 \le u \le 1\}$ are independent Brownian bridges.
Note This statement can be shown by the same method as was used in [Hušková and Picek, 2005].

Local alternative Now, let us now suppose that the alternative is true and that the same assumptions about random error terms and $\boldsymbol{h}(t)$ as in the previous statement hold. Further suppose that $k^* = [nt]$ for some $0 < t < 1$ and that $\boldsymbol{\delta}_n$ satisfies the conditions

$$\|\boldsymbol{\delta}_n\| \longrightarrow 0 \quad \text{and} \quad n^{1/2} \|\boldsymbol{\delta}_n\| \longrightarrow \infty.$$

Then, for $\gamma < t < 1 - \gamma$:

$$U_n \xrightarrow{P} \infty \qquad \text{as } n \to \infty.$$

This proves the consistency of the studied test statistic under the given assumptions.

Acknowledgement

The work was supported by GAČR grant 201/05/H007.

References

Horváth, L., Horváth, Z., and Hušková, M. (2008). Ratio tests for change point detection. Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor P. K. Sen, 1:293—304.
Hušková, M. (2007). Ratio type test statistics for detection of changes in time series. *Bulletin of the International Statistical Institute, Proceedings of the 56th Session, Lisboa 2007*, (976).
Hušková, M. and Picek, J. (2005). Bootstrap in detection of changes in linear regression. *The Indian Journal of Statistics Special Issue on Quantile Regression and Related Methods*, 67:200–226.
Madurkayová, B. (2007). Ratio tests for gradual changes. *Proceedings of WDS 2007, Prague.*

5. Simulation

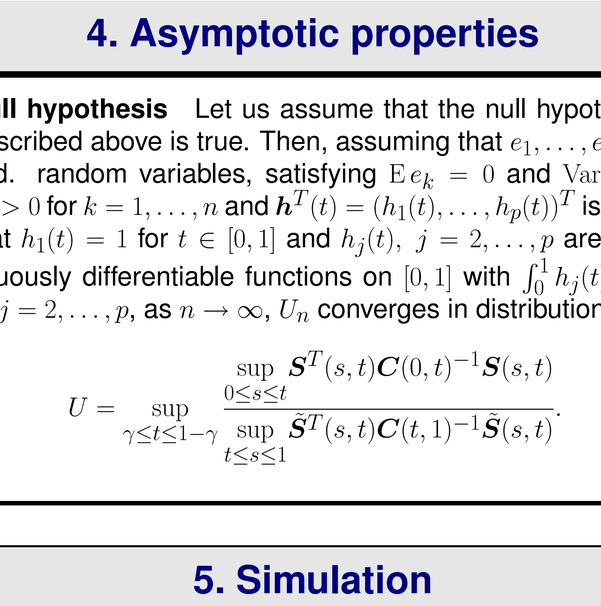

Figure 1: *The values of Q_k for simulated normal distribution samples with parameters $\mu = 0,\ \sigma = 1,\ n = 100, \gamma = 0.1$ The upper left figure refers to the null hypothesis. The other figures refer to alternatives with $k^* = n/2 = 50$ and $\boldsymbol{\delta} = (0, 0, \sqrt{27})$ (upper right), $\boldsymbol{\delta} = (0, \sqrt{27}, 0)$ (lower left), $\boldsymbol{\delta} = (3, 3, 3)$ (lower right).*

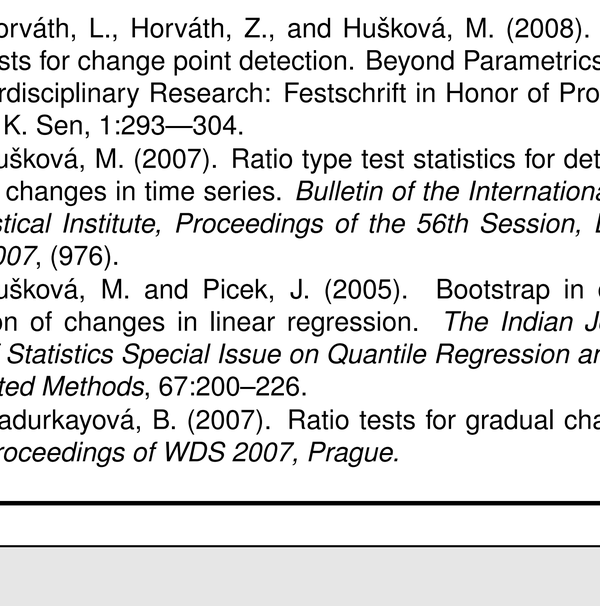

Figure 2: *The values of Q_k for simulated Laplace distribution samples with parameters $\mu = 0,\ b = 1,\ n = 100,\ \gamma = 0.1$. The upper left figure refers to the null hypothesis. The other figures refer to alternatives with $k^* = n/2 = 50$ and $\boldsymbol{\delta} = (0, 0, \sqrt{27})$ (upper right), $\boldsymbol{\delta} = (0, \sqrt{27}, 0)$ (lower left), $\boldsymbol{\delta} = (3, 3, 3)$ (lower right).*