Economics Letters 155 (2017) 131–134

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Is MORE LESS? The role of data augmentation in testing for structural breaks

Yao Rao, Brendan McCabe *

Management School, The University of Liverpool, L69 7ZH, United Kingdom

HIGHLIGHTS

- This paper addresses the issue of data augmentation in structural change testing.
- Theoretical and simulation analysis shows that increasing the sample size may decrease power.
- An empirical example conrms the findings.

ARTICLE INFO

Article history:
Received 14 December 2016
Received in revised form 23 March 2017
Accepted 24 March 2017
Available online 27 March 2017

JEL classification:
C01
C15
C44

Keywords:
Structural change
CUSUM test
Regression

ABSTRACT

In this paper, we examine the impact of increasing the size of a data set in detecting structural breaks. Based on an empirical application, supported by theoretical justification and a simulation experiment, we find that larger sample sizes may make it more rather than less difficult to determine the existence of a structural break.

1. Introduction

This paper investigates a curious phenomenon in structural change testing whereby a break, initially found in a given set of observations, is subsequently not detected when additional observations are obtained thereby increasing the sample size, T. The plan of the paper is as follows. Section 2 introduces the regression model used in the analysis, the *CUSUM* procedures employed to detect breaks and the numbers c_q used to assess how difficult it is to detect a break should it occur at position $q \in [2, T]$ in the sample span. Section 3 conducts some simulations to quantify the effect of increasing the sample size on the power of the *CUSUM* tests to detect the existence of a break. Section 4 attempts to detect the presence of a structural break in an interest rate-bond yield relationship based on data spanning 1972–2010 and does indeed show that a break detected with a given set of observations is not discoverable when the data set is augmented, doubling the sample size. The paper finishes with some conclusions.

2. Structural breaks in regression

The set of models we consider for the observations is the multiple regression

$$y = X\beta + \omega_q\delta + \varepsilon \qquad (1)$$
$$\varepsilon \sim (0, \sigma^2 I)$$

where y is a $N \times 1$ vector of observations, X is a $N \times k$ full rank matrix of variables that is conditioned on, β is a vector of unknown coefficients and ε is a vector of independent disturbances with zero mean and variance σ^2. The form of the structural break is captured by $\omega_q\delta$ with ω_q being a vector and δ a scalar which may be positive or negative and q is a member of a set Q. So for example, if we set $\omega_q = \mathbf{i}_q = (0, \ldots, 0, 1, \ldots, 1)'$ where the 1's start at $q = N_B + 1$, then we are considering models with a shift, in the intercept only, at the unknown position $N_B \in [1, N - 1]$. For ease of reference, we use $\tau = q/N$ to indicate the fraction of the sample span where a

* Correspondence to: Management School, Chatham Street, University of Liverpool, L69 7ZH, United Kingdom.
E-mail address: Brendan.McCabe@liverpool.ac.uk (B. McCabe).

http://dx.doi.org/10.1016/j.econlet.2017.03.033

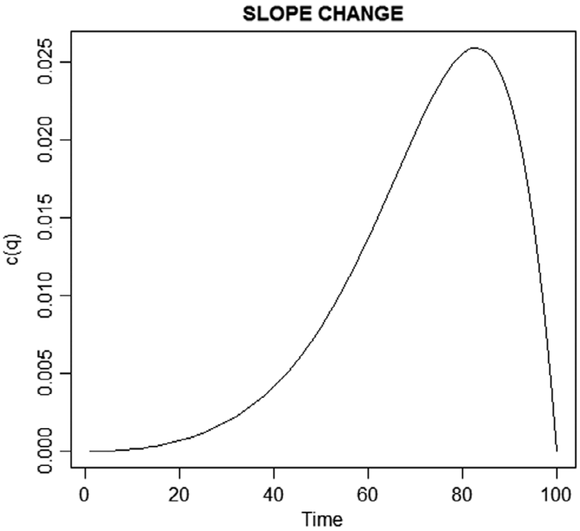

Fig. 1. Cq values in trend model.

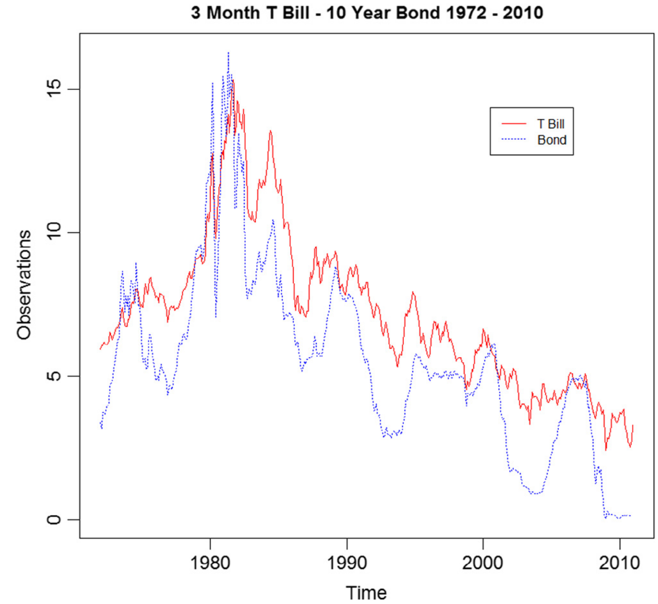

Fig. 2. 3-month T-bill and 10 year bond.

break may take place. Other notation used in the sequel includes $r = My$, $M = I - X\left(X'X\right)^{-1}X'$, the studentised r, $\widetilde{r} = r/\left(r'r\right)^{1/2}$, and $c_q = \omega_q' M\omega_q$. Two procedures for examining structural breaks were applied: a traditional residual based *CUSUM* (see McCabe and Harrison (1980) and Ploberger and Krämer (1990, 1992)) and a weighted CUSUM, *W-CUSUM*, which is equivalent to the minimum sum of squares test of Bai (1997); see McCabe and Rao (2017).[1] The two-sided detection procedure based on the *W- CUSUM* statistic is to accept the hypothesis of no break if

$$\max_{q=2,\ldots,N} c_q^{-1}\left(\xi_q'\widetilde{r}\right)^2 \leq K$$

(where K is a critical value to control the size the test) and decide there is a break at the argmax position when the test rejects. The ordinary *CUSUM* procedure is identical in structure but the test is based on $\max_{q=2,\ldots,N}\left(\xi_q'\widetilde{r}\right)^2$. Thus the procedures consist of an initial test followed by an identification step if the test rejects.[2]

A way to shed light on c_q is to note that change point detection may be thought of testing $\delta = 0$ in (1) over every possible configuration of models specified by ω_q. It is straightforward to show that the variance of $\hat{\delta}_q$, the OLS estimator of δ in $y = X\beta + \omega_q\delta + \varepsilon$, is proportional to c_q^{-1} so that accurate estimates correspond to large values of c_q . More specifically, it follows that $c_q^{-1}\left(\xi_q'\widetilde{r}\right)^2 = c_q\hat{\delta}_q^2$ and $\left(\xi_q'\widetilde{r}\right)^2 = c_q^2\hat{\delta}_q^2$. Thus, if the c_q are small in some region of the sample span, there is little chance of a break being detected should it lie therein by comparison with regions where the c_q are large. In addition, we can deduce that the *W-CUSUM* test will perform worse than the *CUSUM* test in regions with a high c_q when the true break point lies there.

3. More or less?

To assess the effect of data additions in structural break problems, it is convenient to use a stylised model as the c_q values do not then depend on the realisation of the x-variable involved. We considered linear trend model with a fixed set of 100 observations which contains the break at position 80 and subsequently these data are supplemented with additional observations from the same model, *increasing* the original sample size from T to T^*. The models are

$$\text{Model 1:} \quad \begin{cases} y_t = \alpha + \beta t + \varepsilon_t;\, t = 1, \ldots, 80 \\ y_t = \alpha + (\beta + \delta)\, t + \varepsilon_t;\, t = 81, \ldots, 100 \end{cases}$$

$$\text{Model 2:} \quad \begin{cases} (y_1, x_1)\,, \ldots, (y_{100}, x_{100}) \ \text{of Model 1} \\ \text{plus } y_t = \alpha + (\beta + \delta)\, t + \varepsilon_t;\, t = 101, \ldots, T^*. \end{cases}$$

We choose T^* to be 120, 150 and 200. Now the plot of the c_q values for the trend Model 1 are given in Fig. 1. It is unimodal, peaks roughly at $\tau = 0.8$ and gives little weight to the earlier part of the span. Thus, a break at location 80 in Model 1 would correspond $\tau = 0.8$ but in Model 2 with $T^* = 200$ a break at location 80 would correspond to $\tau = 0.4$, a position where high power is not expected.

The regression parameters were set at $\alpha = \beta = 1$ with $\varepsilon_t \sim N\,(0, 1)$. We tested for a break in the trend slope, t , with 2000 replications and $\delta = 0.01$. The results are given in the Table. When using Model 1 with $T = 100$, the *CUSUM* test suggested a break in 62% of cases when rejecting at the $\alpha = 0.05$ level, critical values being computed via the bootstrap. Then, increasing the size of the original data to $T^* = 200$ whilst keeping the break in the same position, the test was applied again to Model 2. With the additional observations, the *CUSUM* test now rejects just 5% of the time, a dramatic drop, indicating that more may sometimes be less, as the effect of shifting the relative position of the break from the advantageous $\tau = 0.8$ to $\tau = 0.4$ takes its toll. The corresponding figures for the *W-CUSUM* are from 51% rejections in Model 1 to 4% in Model 2. As expected the *W-CUSUM* has less power than the *CUSUM* in Model 1. From the Table, it is clear that additional data, that increasingly places the break location in less favourable τ positions, progressively worsens performance.

	$T = 100$	$T^* = 120$	$T = 100$	$T^* = 150$	$T = 100$	$T^* = 200$
τ	0.80	0.67	0.80	0.53	0.80	0.40
CUSUM	60%	56%	62%	20%	62%	5%
WCUSUM	49%	45%	49%	18%	51%	4%

Of course, there is no suggestion that additional data are never useful and it is easy to design experiments where additional data

[1] This working paper provides a source of background material and additional detail for the readers.

[2] Under normality, in a decision theory framework, *CUSUM* procedures can be shown to have certain optimality properties for identifying the location of the break; see McCabe and Rao (2017).