Is MORE LESS? The Role of Data Augmentation in Testing for Structural Breaks

Yao Rao and Brendan McCabe*
Management School, The University of Liverpool, L69 7ZH, UK

March 23, 2017

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.

Keywords: Structural change, CUSUM test, Regression
JEL Classification: C01;C15;C44[1]

[1] Correspondence to: Brendan McCabe, Management School, Chatham Street, University of Liverpool, L69 7ZH, UK. E-mail address: Brendan.McCabe@liverpool.ac.uk

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 to 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

$$
\begin{aligned}
y &= X\beta + \omega_q \delta + \varepsilon \qquad (1)\\
\varepsilon &\sim (0, \sigma^2 I)
\end{aligned}
$$

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 $\omega_q = \mathbf{i}_q = (0, ..., 0, 1..., 1)'$ 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 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)[2]. The two-sided detection procedure based on the W-$CUSUM$ statistic is to accept the hypothesis of no break if

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

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

(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[3].

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

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

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

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 Figure 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 $2,000$ replications and $\delta = 0.01$. The results are given in the Table below. 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,

[3] 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).

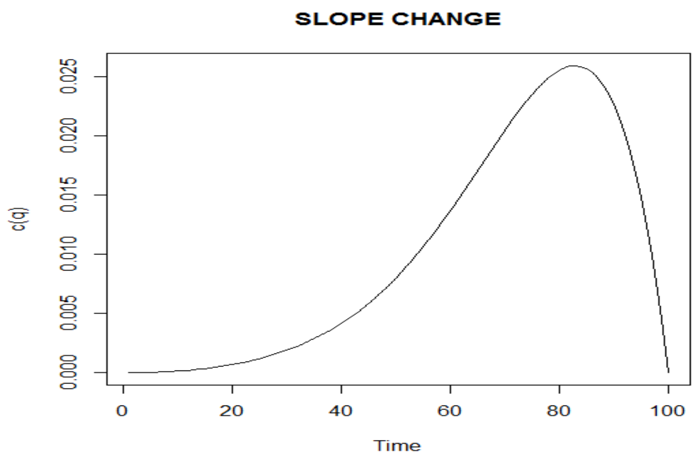

Figure 1: Cq values in trend model

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 reposition the break point to a more favourable position in the span.

4 Interest Rate and Bond Data

This section looks at the specific relationship between the 3-month US Treasury bill rate (T-$Bill$) and the 10-year bond yield (B10Y). The data are taken from the Federal Reserve data base $FRED$[4] and are monthly from January 1972

[4]FRED database at the Federal Reserve Bank of Saint Louis, from the interest rates category: http://research.stlouisfed.org/fred2/categories/22

until December 2010, giving 468 observations. More recent observations are not included due to the effect of quantitative easing in keeping short term rates very near zero. The data are plotted in Figure 2. Economic theory suggests that these

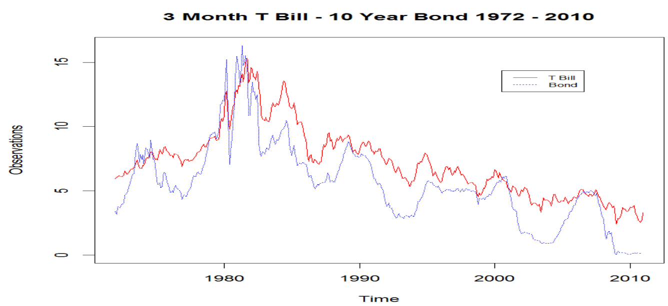

Figure 2: 3-month T-Bill and 10 Year Bond

data should individually be integrated, $I(1)$, and also be cointegrated, i.e., there exists a linear combination of the B10Y and $T\text{-}Bill$ rates which is stationary. Further, the short term $T\text{-}Bill$ rate should drive the long term bond yield with a regression slope coefficient of 1, i.e. the spread (difference between the rates) should be stationary. Since the regression slope should remain at 1, we might expect to see exogenous breaks show themselves in the intercept term of the model

$$B10Y_t = \alpha + 1 * T\text{-}Bill_t + \varepsilon_t,$$

ε_t being some stationary process.

To check for structural change, we performed an intercept stability test on the first 20 years of data from 1972:1 till 1992:12 looking for a possible break date which would match the recession of 1980-1982 and then a test on the last 20 years using 1990:1 till 2010:12 whose span covers the period of the Asian currency crisis. We used fixed-X bootstrap critical values, extracting an $AR(1)$ term from the estimated residuals before resampling to deal with possible autocorrelation in the disturbances. The $T\text{-}Bill$ variable is conditioned on as it is ancillary under the model.

For the first 20 years of data, the p-values for the $CUSUM$ and $W\text{-}CUSUM$ tests were 0.244 and 0.392 respectively and, despite the obvious increase in volatility evident in Figure 2, no break was identified in the regression intercept. For the last 20 years of data, the p-value of the $CUSUM$ test was 0.047 and subsequently June 1997 was identified as a break date. This position corresponds closely to the accepted date of the Asian crisis i.e. July 1997. The $W\text{-}CUSUM$ strictly did not find a break at the 5% level but the p-value was close at 0.064

Since no break was identified in the earlier data set we tested the complete set from 1972:1 to 2010:12 hoping that the additional observations would lead to a more emphatic conclusion with respect to the currency crisis by means of lower p-values. However, the p-values for the full set of observations increased

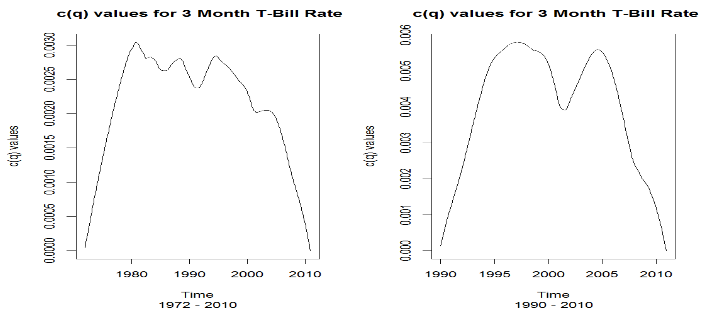

Figure 3: Cq values

to 0.176 and 0.267 for the $CUSUM$ and W-$CUSUM$ respectively and no break was identified. To shed some light on the bill rate/bond yield puzzle, we looked at (normalised) plots of the c_q values for both the most recent 20 years and complete sets of observations and these are shown in Figure 3. The c_q for the recent observations (right panel) have a clear peak in the mid to late 90's indicating that the confounding effect of the regression is not strong there, making it easier to find a break. On the other hand, the c_q for the full data (left panel) peak around 1980 with a smaller sub-peak in the early 90's and are in further decline by the late 90's, the time of the Asian crisis, making it difficult to overcome the impact of the regression design.

5 Conclusion

It appears that the position of the break location in the sample span is an important determinant of the power of break tests. As a rule of thumb, data augmentation that changes the position of the true break point from a relatively favourable to an unfavourable one may be unhelpful and may make finding breaks *more* difficult. These suggestions offer a plausible explanation for the perplexing finding sometimes encountered in empirical investigations as exemplified by the Treasury bill relationship with the bond yield as analysed here.

References

Bai, J. (1997). Estimation of a change point in multiple regressions. *Review of Economics and Statistics* 79, 551–563.

McCabe, B. P. M. and Harrison, M. J. (1980). Testing the constancy of regression relationships over time using least squares residuals. *Applied Statistics* **29**, 142-148.

McCabe, B. P. M and Rao, Y (2017). The effect of regression design on optimal tests for finding break positions. Available at SSRN http://ssrn.com/abstract=2867141.

Ploberger, W. and Krämer, W. (1990). The local power of the CUSUM and CUSUM of squares tests. *Econometric Theory*, **6**, 335-347.

Ploberger,W. and Krämer, W. (1992). The CUSUM test with OLS residuals. *Econometrica* **60**, 271-285.