The Effect of Regression Design on Optimal Tests for Finding Break Positions

Brendan McCabe and Yao Rao

Management School, The University of Liverpool, L69 7ZH, UK

January 31, 2017

Abstract

In this paper, we derive an optimal test for determining break positions in Gaussian linear regressions. The procedure is an admissable rule in a multiple decision theory setting and the results are exact and valid in small samples. The analysis indicates that regression design can have a very significant effect on the ability of the optimal test to find the position of the break. Some regression designs make it all but impossible to successfully identify a break location in certain subsections of the sample span. Two graphical devices, the c_q and ω-plots are available to identify those subsets of the sample span where locating a break position is difficult or impossible.

Keywords: Structural change, CUSUM test, Bayes rules, Multiple decision theory, Regression

JEL Classification: C01;C12;C32;C44

1 Introduction

Structural stability has long been an important issue in statistics and econometrics. In the frequentist tradition, a lot of work has concentrated on detecting the presence of structural change and finding the corresponding location, see for example, the surveys by Perron (2006) and Aue and Horvath (2013). The corresponding Bayesian literature is also large and includes Carlin *et al.* (1992), Barry and Hartigan (1993), Stephens (1994) and Martin (2000). The present paper contributes to the structural change literature by (a) deriving a decision theory based exact small sample optimal procedure for finding the *location* of a structural change in the sample span of Gaussian linear regression models and (b) analysing of the effect that the regression design has on the performance of this procedure. We consider two specific scenarios. The first is where the observations are sequenced by the order statistics of a variable of interest as in a cross section (CS) study and the second is where the observations are ordered as a time series (TS), for example, when the model contains trends and/or seasonal components.

The modelling situation we envisage is one where there are grounds for thinking that an attribute or episode may trigger a disruption in a currently understood relationship but it is not definitive that a resultant structural change should manifest itself in the model and in the associated data. In general, the ability to determine the existence of a change depends on a) the size of the change, b) the model and c) the sample data used. The sample data employed, the span of which may vary as it is chosen by the investigator, maps the potential (unknown) fixed location of the change into a relative position (also unknown) in the sample span. Hence, the possible change location may be thought of as fixed in an absolute sense while its relative position in the sample span may vary. One has to bear in mind that large changes will overwhelm any design effect and be easily detectable. Similarly, for small changes the design effect will dominate and changes will remain undetectable. In what follows, change sizes are considered to be moderate. In the TS context, even if a disruptive episode were to consist of a one-off event and did induce a change in our model/data, its date is deemed not to be known, perhaps because the effect of the trigger event takes place with a lag or even because the event itself had been anticipated. In the CS case, we simply may not know at which level of the treatment variable an anticipated effect might take place. Determining that a change has occurred is a necessary but not a sufficient condition for finding its location.

The type of structural change we consider is often described as a structural break, in that a coefficient of the model is deemed to change at some unknown location and continues to remain at the new level for the duration of the sample. Thus, we have a model with no change, a model with the change in position 2, a model with a change in position 3 and so on until position N, the sample size. Thus there are N competing models in total and the task is to choose between them. A multiple decision theory framework is used since there are more than two decisions to be made and we follow closely the treatment of Ferguson (1961) but modify the content there to deal with structural change. Decision theory allows for the construction of an optimal break location procedure which is uniform with respect to the break parameter. It turns out that the minimum sum of squares (MSS) procedure proposed by Bai (1994,1997) in the TS context is equivalent to the decision theory approach but the latter involves the computation of just one regression as opposed to $O(N)$ such computations in MSS. The optimal procedure consists of two parts: an initial test to check if a break has occurred and, if so, an identification step to reveal the location of the break[1]. Hence, to successfully identify a break location, it is required that the test has power and that the resulting suggested location is not spurious.

We introduce the c_q plot which displays the effect that any given regression design has on the ability of the procedure to determine the existence of a break as the potential break position varies over the sample span. The c_q depends on the regression design and the types of break being investigated. These plots may be used to identify hot and blind spots in the span i.e. regions where it is, relatively speaking, easier and harder to determine if there is a break should the break be located therein; these subsets are additional to the obvious cases at the extremities of the span where it is all but impossible to determine if

[1] This sort of optimality is in contrast to studying the *power* of break tests i.e. the ability to reject the null of no break, typically against some weighted combination of the possible break points; see, for example, Andrews *et al.* (1996) and Forchini (2002).

there has been a break. So, given the possibility of a trigger situation or event, the c_q plot tells whether or not, discovering the existence of an induced break will be significantly impacted by the model and the current data. An analysis of some simple trend models and simulated data from more complicated models, quantifies the effect that certain designs may have and shows that it can be almost impossible to determine if there has been a break should it occur in certain subsets of the sample span. These subsets are recognisable using the c_q plots.

The analysis and simulations also show that it is possible for the procedure to have very high power but little ability to identify the correct break location. This phenomenon, unfortunately, leads to spurious identification of break locations. Certain design features can make it impossible to determine the correct location and we use a second graphical device, the ω-plot, to help identify such difficult cases, although the issue of spurious breaks cannot be resolved in general. As far as we are aware the analysis presented here, of the performance of an exact optimal break locating strategy in regression, is the only one currently available.

The plan for the rest of the paper is as follows. Section 2 applies the decision theory framework to regressions and a rule, based on $CUSUM$ statistics, for optimally identifying break locations is derived. In Section 3, the effects of the regression design on the $CUSUM$ are analysed. A graphical technique, the c_q plot, is suggested to identify hot spots and blind spots in the sample span. The utility of the c_q plots is explored by examining a selection of techniques for modelling trends and seasonality. Some simulations are conducted in Section 4. They illustrate how the c_q plot can identify subsections of the sample span where the optimal rules will have power and other subsections where they will not. Certain features of some designs can make it virtually impossible to identify the break location and the ω-plot is introduced to help identify such cases. An empirical analysis of is given in Section 5 while Section 6 contains conclusions.

2 Optimal Structural Break Tests in Regression

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

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

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 normal 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. The size of the breaks are calibrated against the disturbance standard deviation, σ. Much of the algebra we employ is still valid when ω_q and δ are matrices and vectors respectively, but we concentrate on the case where δ is a scalar because optimal rules that are uniform in δ^2 are available which allows procedures to be assessed independently of the value of δ (and indeed of β, σ^2). Our problem is to decide which of the hypotheses represented by the set of Q_n possibilities generated by the class $\omega_q : q \in Q$ should be accepted. So for example, if we set $\omega_q = \mathbf{i}_q = (0, ..., 0, 1..., 1)'$ where the 1's start at $q = N_B + 1$, with the convention that $\omega_0 \equiv \omega_N = (0, 0, ..., 0)'$, then we are considering models with a shift, in the intercept only, at the unknown position $N_B \in [1, N-1]$ and also the model of no shift ω_0, giving $Q_n = N$ possibilities in total. In structural break problems, ω_q is zero up to position $q-1$ and takes some non zero values thereafter. In the CS situation, q represents the observation number of the break as determined by the variable by which the observations in $[y\ X]$ were ordered while in the TS context q represents the timing or date of the break. The vector ω_q could also be specified as $\omega_q = \xi_q = (0, ..., 0, x_{q,,j}, ..., x_{N,j})'$ corresponding to a possible change in the coefficient of the jth regressor. Again in the CS case, $x_{q,j}$ is the q-th order statistic of the j-th variable according to which the observations were arranged and in TS problems q represents a date. In fact, we work with $\omega_q = \xi_q w_q$, where w_q is a sequence of *scalars* representing weights that we may wish to attach to the competing hypotheses, allowing, for example, changes of a greater magnitude to take place near the extremities of the sample span. Hence our general multiple decision problem is to decide which of the change point models defined by the set $Q = [1, 2, ..., N]$ $(Q_n = N)$, is preferred. A choice from Q identifies either the model with no breaks or a possible break point. For the purposes of this analysis, we consider that there is a single possible change but allow for the possibility that no shift has occurred. We do not consider possible simultaneous shifts in the regression variance because of confounding (See McCabe (1988) for a discussion of other multiple decision procedures as well a more general analysis of structural stability (including multiple breaks) in location and scale).

For ease of reference, we use $\tau = q/N$ to indicate the fraction of the sample span where a break may take place and when $q = N_B + 1$, τ represents the true break fraction. Other notations that are used in the following text are $r = My$, $M = I - X\left(X'X\right)^{-1}X'$, the studentised r, $\widetilde{r} = r/\left(r'r\right)^{1/2}$, and

$$c_q = \omega_q' M \omega_q.$$

We shall see later that we may interpret c_q as an index of how difficult it is to detect a change point should it occur at position q; if c_q is small it is difficult while it is relatively easier for large c_q. So, from a pragmatic point of view, it may be thought worthwhile to positively weight the hypotheses under consideration in regions where c_q is small to have any hope of success should the true break location lie there. Hence, we consider deciding between

$$
\begin{aligned}
&H_0 : \omega_q = \omega_0 \\
&H_q : \omega_q = w_q \xi_q = c_q^{-1/2} \xi_q \qquad\qquad q = 2, ..., N
\end{aligned}
\tag{2}
$$

where we have weighted the hypotheses using c_q to create larger shifts in relatively difficult regions.

Now the nuisance parameters of the problem are β, σ^2 and the sign of δ. We seek to eliminate them by invariance. The first rule is that the problem be invariant under the addition to y of vectors of the form $X\alpha$ while the second is invariance under multiplication of y by scalars $c \neq 0$.[2]

The following Proposition is a mild extension of Theorem 4.1 of Ferguson (1961, p 278) and paraphrases his formulation. Like the Neyman-Pearson Lemma, the Proposition specifies the structure of an optimal rule.

Proposition 1 *(Ferguson) The decision rule that decides to accept there is no break when*

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

and to decide that there is a break at position q^* when $c_{q^*}^{-1} \left(\xi_{q^*}' \widetilde{r}\right)^2 = \max_{q=2,\ldots,N} c_q^{-1} \left(\xi_q' \widetilde{r}\right)^2 > K$ is invariant admissible when the hypotheses are given by (2).

The proof of this proposition is the same as Ferguson's Theorem 4.1 except that b_{jj} there is replaced by the equivalent c_q here[3]. In hypothesis testing problems it is more

[2] Notice that this is not the same invariance rule that is commonly constructed when testing parameters in the *covariance* structure of the linear regression model (e.g. testing autocorrelation) where the constant $c > 0$ is used.

[3] The rule of Proposition 1 is also Bayes with respect to a prior that gives equal weight to the hypotheses and is uniform in δ^2.

usual to classify the rules by $\alpha \in (0,1)$ (rather than K) where

$$\alpha = 1 - P\left[\text{accept } H_0 | H_0 \ \text{ true}\right]$$

and by fixing $\alpha = 0.05$, say, we can find a critical value, cv, such that

$$P\left[\max_q c_q^{-1}\left(\xi_q' \widetilde{r}\right)^2 > cv\right] = \alpha$$

and implement the rule in practice. We refer to this optimal rule as a weighted cusum (W-$CUSUM$) since $\xi_q' \widetilde{r}$ cumulates the studentised residuals over q. Invariant admissibility for the W-$CUSUM$ means the probability of deciding a break at position q, given the break did occur at q, i.e. $P(\text{Dec } q | H_q)$, cannot be increased by any other invariant rule without decreasing the equivalent $P(\text{Dec } s | H_s)$ at some other point $s \neq q$.

Now the minimum least squares (MSS) procedure introduced by Bai (1997) in the TS context proceeds as follows to find a break point in the model class $y = X\beta + \omega_q \delta + \varepsilon$. Regress, for each $q \in [2, N]$, the dependent variable, y, on X and ω_q where, successively, $\omega_q = (0, ..., 0, \omega_q, ..., \omega_N)'$ and compute the sum of squares of the residuals $SS(q)$. The break date is determined to be the q that minimises $SS(q)$, i.e., the $\arg\min$. The Appendix shows that the MSS break location may be computed via a single regression and it is equivalent to the $\arg\max$ of the W-$CUSUM$. That is, it is shown that

$$\arg\min_q SS(q) = \arg\max_q \hat{\delta}_q^2 c_q = \arg\max_q c_q^{-1}\left(\xi_q' \widetilde{r}\right)^2 \tag{3}$$

where $\hat{\delta}_q$ the OLS estimator of δ. Note, the MSS and W-$CUSUM$ procedures may not be quite equivalent in practice, as applied researchers using MSS often assume *a priori* that a break does indeed exist.

3 The Effect of the Regression Design

Intuitively, because the W-$CUSUM$ procedure has to be based on an estimated model, in effect, the observations are filtered by the regression design and information is extracted from them in order to estimate the sequence of conditional means of the model. Only the remaining residual information is available for break detection and determining its location. This section looks at the effect the regression design has on the ability of the residual based rules to find breaks and their corresponding locations.

3.1 Interpreting the role of c_q

In the previous sections, the c_q play an important role in the structure of the hypotheses and correspondingly in the W-$CUSUM$ statistic. 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. From the algebra in the Appendix, it is easily seen 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, equation (7) of the Appendix shows that the numerator of the Wald statistic for testing $\delta = 0$ (for any fixed q) is given by $\hat{\delta}_q^2 c_q$ and (3) shows that using the sequence of Wald tests is essentially equivalent to using the W-$CUSUM$. Hence, if the c_q are small, say, 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. Thus a plot of the c_q, for any model, may be seen as a convenient way to identify sub-spans of the sample where the W-$CUSUM$ test will have blind spots and struggle to find possible breaks and correspondingly hot spots where, relatively speaking, discovering the existence of a break is easier. Of course, one can subsequently only identify the break *location* when the test of the null rejects and so having power in a sub-span is a necessary but not sufficient condition for having the ability to find exact break locations.

To see the effect of a regression design on the ability of detect the existence of a break, consider the simple case when testing for a possible break in means where the observations are ordered by time, i.e.,

$$\begin{aligned} y_t &= \alpha + \varepsilon_t; \quad t = 1, ..., T_B \\ y_t &= \alpha + \delta + \varepsilon_t; \quad t = T_B + 1, ..., T. \end{aligned}$$

so that X is a column of 1's and $\omega_q = \mathbf{i}_q = (0, ...0, 1, ..., 1)'$ where the 1's start in position q. The elements of the M matrix are $m_{tt} = 1 - T^{-1}$ on the diagonals and $m_{ts} = -T^{-1}$ on the off-diagonals. It is then straightforward to evaluate the c_q to get $c_q = T\tau(1 - \tau)$ which has the familiar $\cap$-shaped profile peaking at $\tau = 0.5$. Hence, if the true break point is near the middle of the sample span, the W-$CUSUM$ would have a better chance of detecting it than if the true break were to occur near the extremities. This example accords well with our intuition but, there are other models that do not and some of these are explored below. In the next sub-section, we investigate the effect that different specifications of a trend have on power.

3.2 The Effect of Trend Specification

This section looks at some of the different ways trends may be modelled and analytically derives the functional form of the corresponding c_q. Different trend specifications can have a big impact on what parts of the span are advantageous and what parts are not, from the power perspective. Some trend specifications mean that it is virtually impossible to locate the correct break date even when the corresponding test has high power. This is unfortunate as it leads to spurious identification of break dates. The ω-plot is introduced to help alleviate the spurious identification problem.

Consider the model for a break in slope coefficient of the linear trend

$$y_t = \alpha + \beta t + \varepsilon_t; \quad t = 1, ..., T_B$$
$$y_t = \alpha + (\beta + \delta)\, t + \varepsilon_t; \quad t = T_B + 1, ..., T.$$

so that X consists of a column of 1's, $\mathbf{1}$, and the trend variable $\mathbf{x} = \mathbf{t} = (1, 2, ..., T)'$ while $\omega_q = \mathbf{t}_q = (0, ...0, q, ..., T)'$. From the Appendix, it follows that

$$c_q = \mathbf{t}_q' M \mathbf{t}_q \approx T^3 \left(\frac{1}{3}\tau^3 - \tau^4 + 2\tau^5 - \frac{4}{3}\tau^6 \right) \tag{4}$$
$$= T^3 \frac{1}{3} \tau^3 (1 - \tau) \left(4\tau^2 - 2\tau + 1\right) \tag{5}$$

We use the $\approx$ to mean plus terms of a smaller order that the leading power of T, i.e., smaller than T^3 here. It is easy to see that c_q is skewed left, unimodal and maximised on $[0, 1]$ at $\tau = 0.83$ and, from (5) that that c_q is tied down at 0 at the extremities 0 and 1. Also since there are no linear or quadratic terms, (4) indicates that the test will have low power for small values of τ. The normalised c_q are plotted in the right panel of Figure 1, and it can be seen almost no weight is given to the possibility of breaks in the first half of the sample with the implication that if the true break point were to lie in that region then even those with quite a large magnitude would not be discovered. The c_q for finding a break in the intercept of the trend model is found by algebra similar to the slope case and they are given by

$$c_q = \mathbf{i}_q' M \mathbf{i}_q \approx T \left(\tau - 4\tau^2 + 6\tau^3 - 3\tau^4 \right). \tag{6}$$

There are two maxima, one at $\tau = 0.21$ and the other at $\tau = 0.79$ and the plot is as given in the left panel of Figure 1. It is seen that it is bimodal while there is a dip in the middle of the sample span around $\tau = 0.5$.

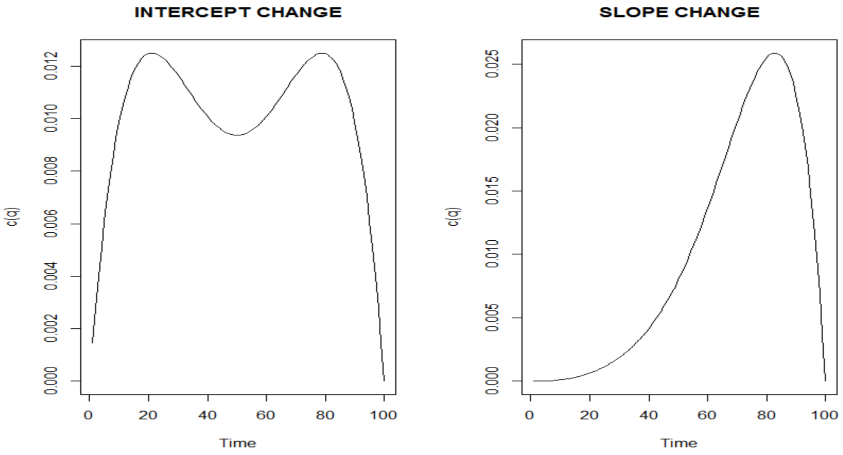

Figure 1: c_q plots for Trend Regressions

Another commonly used model is the broken (continuous) trend break given by

$$y_t = \alpha + \beta t + \varepsilon_t; \quad t = 1, ..., T_B$$

$$y_t = \alpha + \beta t + \delta\left(t - T_B\right) + \varepsilon_t; \quad t = T_B + 1, ..., T$$

which simultaneously introduces a change in slope and intercept. The $X = [\mathbf{1}\ \mathbf{t}]$ matrix is the same as before and $\omega_q = \mathbf{b}_q = (0, ...0, 1, ..., (T - q + 1))$. In this case

$$c_q = \mathbf{b}_q' M \mathbf{b}_q \approx \frac{T^3}{3}\tau^3\left(1 - \tau\right)^3.$$

These c_q have the familiar "bell" shaped plot (see the left panel in Figure 2) centered at 0.5 and are tied down at the extremities. Hence for this version of the trend model, greatest power will tend to occur should the break happen near the center of the span with $\tau = 0.5$ in contrast with $\tau = 0.8$ for the regular trend. There are also no linear or quadratic terms in either τ or $(1 - \tau)$ implying that breaks near either extremity will be hard to find. It is instructive to note that the way the trend is modelled matters. Say we symmetrise the trend around zero and use $\mathbf{x} = (\mathbf{t} - \bar{t}\mathbf{1})$ as our regression design variable.

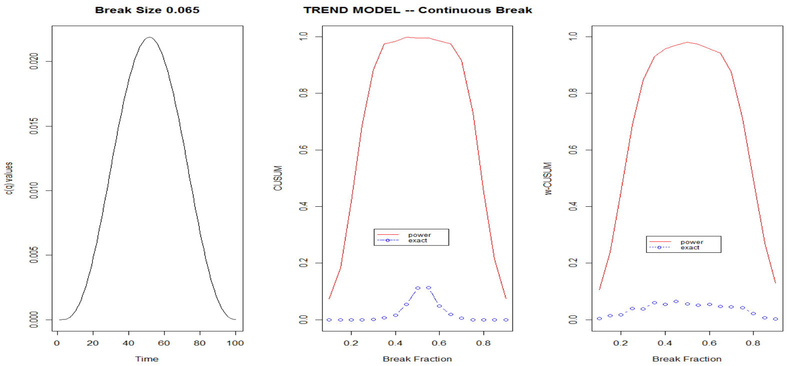

Figure 2: c_q plot along with the power and exact dating performance of the CUSUM and W-CUSUM tests for the Broken (continuous) Trend Model

Then, again using algebra similar to that above, we find in the slope case that

$$
\begin{aligned}
c_q = (\mathbf{t}_q - \bar{t}\mathbf{i}_q)' \, M \, (\mathbf{t}_q - \bar{t}\mathbf{i}_q) &\approx T^3 \left(\frac{1}{4}\tau - \frac{3}{2}\tau^2 + \frac{23}{6}\tau^3 - \frac{21}{4}\tau^4 + 4\tau^5 - \frac{4}{3}\tau^6 \right) \\
&= T^3 \frac{1}{12} \tau \left(1 - \tau\right) \left(3 - 15\tau + 31\tau^2 - 32\tau^3 + 16\tau^4\right)
\end{aligned}
$$

which is maximised at $\tau = 0.15$ and $\tau = 0.85$ and is tied down as before (compared with Figure 1). Since we have linear and quadratic terms the test will have greater power for small τ than in the corresponding ordinary trend situation but the power is lower in general. The c_q for the intercept in the demeaned model are exactly as in the ordinary case (6) and so we expect no difference in power.

As remarked earlier, even if the test has power, the regression design can have a big effect on how good the procedure is at finding break locations. For example, consider the alternative model for the trend, $t - \bar{t}$. In this case when T is odd, $t - \bar{t}$ will take the value 0 in the middle of the sample span at $t_m = (T + 1)/2$. Should a break occur at t_m, then $(t_m - \bar{t})\,\delta = 0$ for all δ, δ has no effect and y_{t_m} remains in the pre-break regime, making it impossible to find the break location. Of course, in stylised situations it is easy to identify such anomalies but, quite generally, should x_i be in the vicinity of zero for sample sub-spans and should a beak occur in one of those then finding the break location is almost impossible. For this reason we suggest that the elements of $\omega = (\omega_{2,1}, \omega_{2,1}, ...\omega_{2,N})$ be plotted to identify cases where location of the break is difficult or impossible.

Extensive simulations confirm these assessments and a selection of them is given next in Section 4.

4 Simulations

In this section, we look at some simulations to illustrate the effect of the regression design on the power of both the ordinary $CUSUM$ and W-$CUSUM$ as well as on their ability to identify the exact break location. The ordinary $CUSUM$ has a long history and, when used in conjunction with the studentised residuals, has been analysed in, for example McCabe and Harrison (1980) and Ploberger and Krämer (1990, 1992). It is unweighted and the statistic is given by

$$\max_q \left(\xi_q' \widetilde{r}\right)^2 .$$

The simulations also assess the value of the plots as a diagnostic tool to identify hot and weak spots in the sample span. The graphs in the left panels of Figure 2-5 are of the c_q plots while the two right-most panels show the performance of the procedures. The value of the break magnitude δ was calibrated to ensure that the test's maximal power, over the sample span, to reject the null of no break was close to 100%. The alternatives used were unweighted as we view the weights in this context as a technical device to discover what might be the structure of an optimal procedure for finding break locations. The tests were conducted throughout at the 5% level with critical values calculated by bootstrapping for $N = T = 100$. In all cases $\alpha = \beta = 1$, the disturbances used were $N(0,1)$ and $\tau = (N_B + 1)/N$ varies in increments of 0.05 over the span.

The relative masses that the c_q plot places on sub-spans of the sample represent predictions of the regression induced behaviour of the tests. The performance graphs on right panels show the percentage of times that the tests rejected and, following a rejection, identified the break correctly as the true break location varies across the sample span. The first sub-section simulates a cross sectional model with differing regressors. Then we look at the broken trend specification. We also consider a second TS model with a stochastically trending regressor (details in Section 4.3), looking at three different realizations of x_t. Finally, we look at a simple trigonometric model for seasonality where use of the ω-plot is highlighted.

The general picture to emerge is that the power of both tests is heavily influenced by the regression design as the c_q plots predict and the use of weights in the W-$CUSUM$ alleviates the X effect, giving a more rounded performance across the sample span while paying a penalty at the identified peaks where the $CUSUM$ has the greatest power. Neither test is uniformly superior over the entire sample span. The c_q plots are not able to predict the ability of the procedures to identify exact locations should the way the break manifests itself be zero or nearly so at the true break location. Thus, the c_q may be supplemented by the ω-plot.

4.1 Cross Section Model

We simulate a cross section model by generating a vector of explanatory variables which are then sorted to give the order statistics $x_{(i)}$. Then the model is

$$y_i = \alpha + \beta x_{(i)} + \varepsilon_i; \quad i = 1, ..., N_B$$

$$y_i = \alpha + (\beta + \delta)\, x_{(i)} + \varepsilon_i; \quad i = N_B + 1, ..., N.$$

The exogenous variables were drawn from a $\chi^2\,(3)$, a $N(5, 2)$ and a $N\,(0, 1)$, this latter choice representing an attribute measured on a standardised score, perhaps. The basic premise is that the dependent variable y_i responds differently to $x_{(i)}$ as the level of the

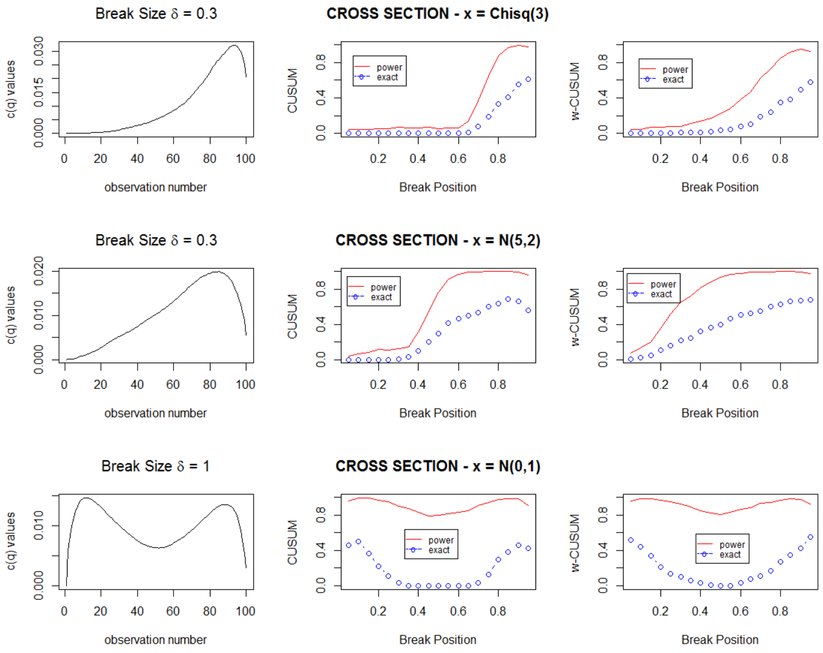

Figure 3: c_q plot along with the power and exact location performance of the CUSUM and W-CUSUM tests for the cross section Model with 3 different regressors

independent variable passes some unknown threshold which we seek to identify; perhaps having a University degree attracts a salary premium?

The X matrix consists of a column of 1's and the variable $x_{(i)}$ and these are used to calculate the usual idempotent matrix M. The $CUSUM$ and W-$CUSUM$ procedures use $r_q = \left(\sum_{i=q}^{N} x_{(i)}\widetilde{r}_i\right)^2$ and $c_q^{-1}r_q$ with weights $c_q = \mathbf{x}_{(q)}' M\mathbf{x}_{(q)}$, $\mathbf{x}_{(q)} = \left(0, ..., 0, x_{(q)}, ..., x_{(N)}\right)'$ and $\widetilde{r}_i$ is the ith element of the studentised residual vector $\widetilde{r}$. Figure 3 shows the predictions c_q in the left panel and the performance of the methods in the other two.

In the case of the $\chi^2\,(3)$ and $N(5,2)$ explanatory variables the value of δ is the same $(\delta = 0.3)$. When x is a Chi-square, there is little power and virtually no chance of exactly discovering the break location unless the break occurs at a fraction greater than $\tau = 0.6$ in the sample span and the power approaches 1 around $\tau = 0.8$. There is little to choose between the two $CUSUM$'s. When x is $N(5,2)$, the corresponding numbers are $\tau = 0.4$ and $\tau = 0.5$. This is a significant improvement in performance and the W-$CUSUM$

does smooth the regression effect. The reason for the improvement is that the Chi-square generates much greater outlier type effects than does the Normal and, given the ordering of the x variable, pushes the peak of the c_q plot towards $\tau = 1$, which makes it more difficult to find a break earlier in the span when x is Chi-square rather than Normal. This is reflected in the c_q plots. The previous pattern breaks down when x is $N(0,1)$. In this case, we used $\delta = 1$ and the performance of both tests is much better near the extremities of the span. In the middle, despite the high power, the ability to determine a break location exactly is almost zero between $\tau = 0.3$ and $\tau = 0.7$ approximately. This is doubly worrying as the presence of high power means that the procedure identifies a spurious break location. The relative information in the c_q plot about the span is reflected in the power performance but there is no indication of how badly the methods perform in the center of the span in determining the location. The reason for the dramatic change in performance is that the design now includes observations from the $N(0,1)$ that are effectively 0 or very close to it in the center of the span and hence if δ becomes non zero at one of these design points there is little, if any, impact on the corresponding observed y, making it almost impossible to identify the true break location. To save space, we do not produce the ω-plot in this obvious situation.

4.2 Trend Regression

The $CUSUM$ statistic for the continuous trend model uses

$$r_q = \left(\sum_{t=q}^{T} (t-q) \widetilde{r}_t \right)^2$$

while W-$CUSUM$ is based on $c_q^{-1} r_q$ where $c_q = \mathbf{b}_q' M \mathbf{b}_q$ where $\mathbf{b}_q = (0, ..., 0, 1, ..., (T-q+1))$. We set $\delta = 0.065$. From the panels in Figure 2 we can see that c_q plots accurately predict the power behaviour of the test but the ability to detect the correct break location is very low by virtue of the continuity of the trend which ensures that the impact on y_t immediately after the break is small. Thus for $0.3 \leq \tau \leq 0.7$, most break identifications are spurious.

4.3 Stochastic Trend

This section looks at a regression model with a stochastic trend on which we condition. The stochastic trend model is

$$y_t = \alpha + \beta x_t + \varepsilon_t$$

where $x_t = z_t / \max\{abs(z_t)\}$ and

$$z_t = 0.04 + z_{t-1} + \eta_t \quad \eta_t \backsim \; iid \; N(0,1)$$

and hence the regression variable x_t behaves like a (nonstationary) random walk with drift, i.e. has a unit root. The dependent and explanatory variables are cointegrated (differ by a stationary term) when there is no break but the x_t variable is conditioned on, given its distribution is not dependent of the parameter δ, i.e. is ancillary. The idea is to check if a known cointegrating regression relationship has coefficients that change over time.

The X matrix consists of a column of 1's and the variable x_t and is used to compute M. We look at the intercept case and $c_q = \mathbf{i}_q' M \mathbf{i}_q$ are dependent on the realisation of the data that actually occurred and while they may be computed with any data set it is not possible to know a priori what the shape will look like. The rules for a shift in the intercept are computed as usual. Figure 4 gives the c_q and performance of the two methods for three different realisations of x_t. These are displayed in three rows in the graph. Needless to say, different realisations constitute different regression designs which may or may not be similar. It appears that the sample sub-spans identified are a good

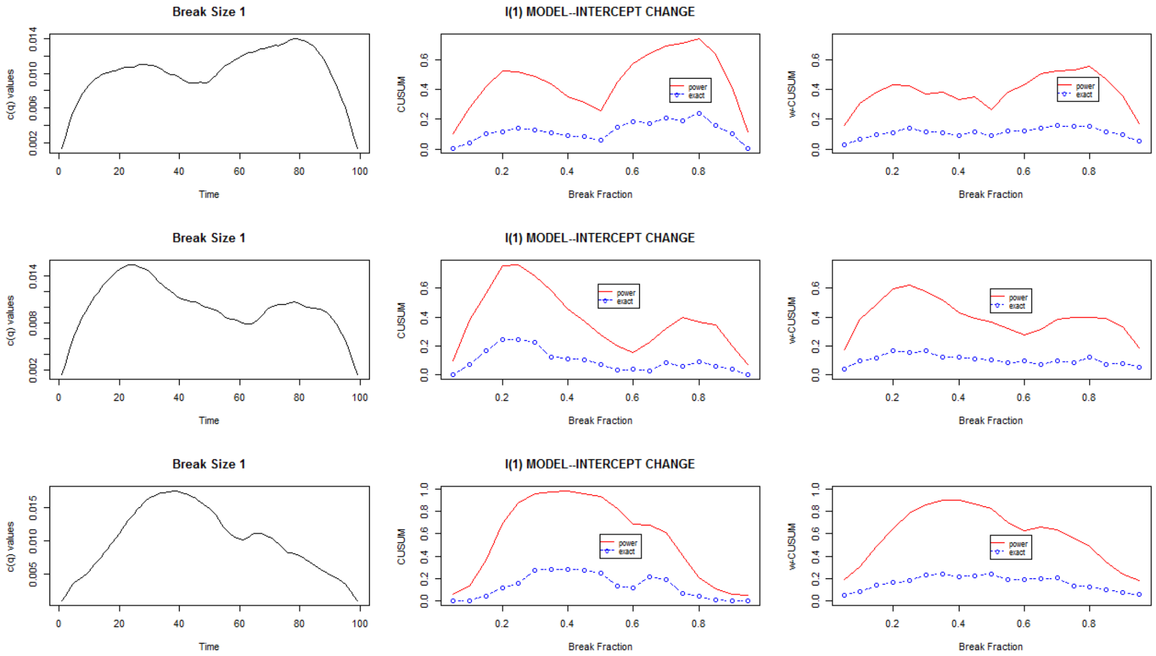

Figure 4: c_q plot along with the power and exact dating performance of the CUSUM and W-CUSUM tests for the $I(1)$ intercept Model

predictor of the overall behaviour of both the $CUSUM$ and W-$CUSUM$ procedures.

4.4 Seasonality

Consider a simple trigonometric model $x_t = \sum_{j=1}^{\frac{s}{2}} \left\{ \cos\left(\frac{2\pi j}{s} t\right) + \sin\left(\frac{2\pi j}{s} t\right) \right\}$ for an s period seasonal process and

$$y_t = \alpha + \beta x_t + \varepsilon_t; \quad t = 1, ..., T_B$$

$$y_t = \alpha + (\beta + \delta)\, x_t + \varepsilon_t; \quad t = T_B + 1, ..., T.$$

In this case, the X matrix used to construct M consists of a constant term plus the variable x_t. Hence $c_q = \mathbf{x}_q' M \mathbf{x}_q$ where $\mathbf{x}_q = (0, ..., 0, x_q, x_{q+1}, ..., x_T)'$. The $CUSUM$ uses $r_q = \left(\sum_{t=q}^{T} x_t \widetilde{r}_t \right)^2$ and the W-$CUSUM$ is based on $c_q^{-1} r_q$. We set $s = 4$ to mimic quarterly seasonality and the results are reported in Figure 5. Again, the c_q plot correctly predicts

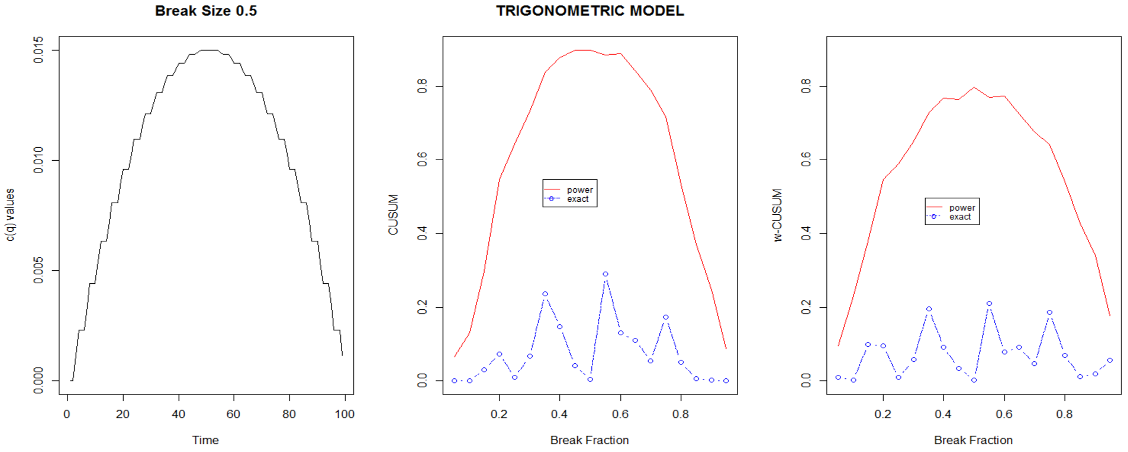

Figure 5: c_q plot along with the power and exact dating performance of the CUSUM and W-CUSUM tests for the Seasonal Model with $s = 4$

the power behaviour of both tests but fails to predict the exact dating performance. The seasonal regression has a dramatic effect on the ability to identify break date and this is due to the fact that the trigonometric functions cycle through the value zero and should a break occur at such a zero point it will be impossible to detect. The ω-plot is given in Figure 6. Our simulations introduce breaks at times $5, 10, 15, 20, \ldots$ and this sequence will match zeros of the ω-plot at times 5, 10, 25, 45 etc. which explains the dips to zero in the location finding performance in Figure 5.

5 Empirical Analysis

In this section, we analyse a cross section of banking data.

5.1 Bank Data - Cross Section

This example uses data on salary and years education to see if there is a threshold where additional years of education would yield a salary premium. The data consist of 474 observations on the logarithm of salary and years of education for employees of a US

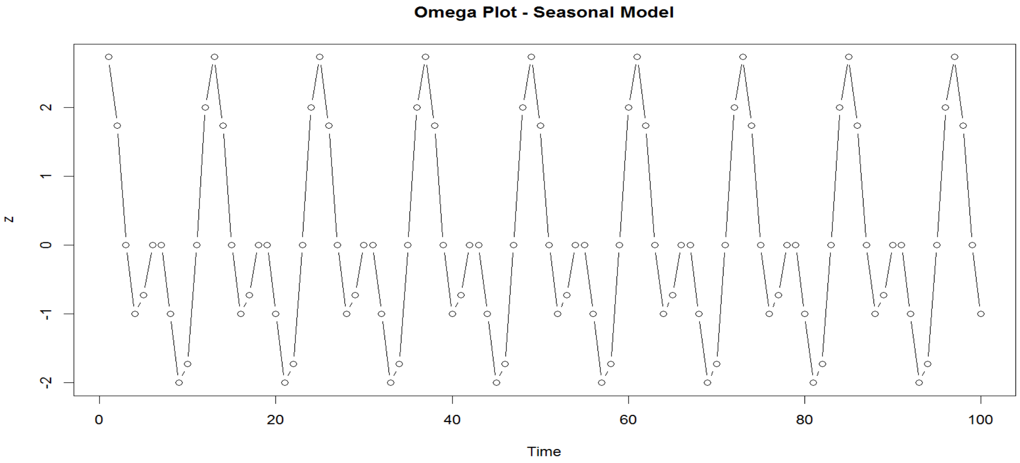

Figure 6: ω-plot for trigonometric seasonal model with $s = 4$

bank as presented in Heij *et al* (2004). Also included are two dummy variables, one for gender and the other for whether an employee was a member of a minority or not. The data are ordered by years of education and the scatter plot of log salary and education is given in Figure 7.

Critical values were found by using a fixed-X bootstrap and re-sampling from estimated residuals using the model

$$\log Salary_i = \alpha + \beta_1 * Educataion_i + \beta_2 * Dummy_g + \beta_3 * Dummy_m + \varepsilon_i.$$

The $CUSUM$ test identified a break at observation number 368 which corresponds to 16 years of education and the threshold between 15 and 16 years of education occurs at observation number 366. The W-$CUSUM$ identified observation number 370 as the break observation. The p-values for both tests were very close to 0. Hence it would appear beneficial to invest in graduate education.

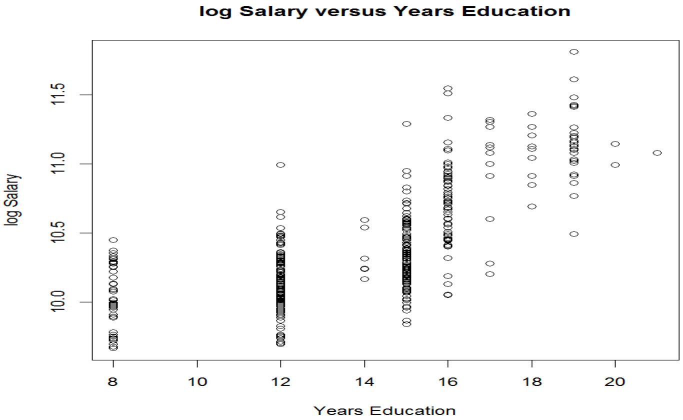

Figure 7: Scatter plot of log Salary versus Years Education

6 Conclusions

This paper introduced an optimal exact test for determining the location of a possible break point in Gaussian linear regression models. It also undertook an analysis of the effect that the regression design has on the power of the test as well as its ability to discover the exact break location. A graphical procedure, the c_q plot, is used to identify subsets of the sample span where it is difficult to detect a break should it occur there and correspondingly subsets where, relatively speaking, it is easier. It turns out that the regression design can play a major role in the ability of the optimal test to detect breaks and their locations, so much so that certain design features make it all but impossible to identify the location of moderately sized breaks. The c_q plots are good predictors of the hot spots and blind spots of the sample span as far as the power of the tests is concerned but may fail to give any indication that the tests have no ability to find the exact location of a break for certain design characteristics. The ω-plot may be used to help identify specific break locations that are difficult or impossible to detect.

References

Andrews, D.W.K., Lee, I. and Ploberger, W.(1996) Optimal change point tests for normal linear regression. *Journal of Econometrics*, **70**, 9-38.

Aue, A. and Horvath, L.(2013) Structural breaks in time series. *Journal of Time Series Analysis*, **34**, 1-16.

Bai, J. (1994) Least squares estimation of a shift in linear processes. *Journal of Time Series Analysis,* **15**, 453-472.

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

Barry, D. and Hartigan, J.A. (1993) A Bayesian Analysis for Change Point Problems, *Journal of the American Statistical Association*, **88**, 421, 309-319.

Carlin, B. P., Gelf, A. E. and Smith, A. F. M. (1992) Hierarchical Bayesian analysis of change-point problems, *Applied Statistics*, **41**, 389-405.

Ferguson, T. S. (1961) On the rejection of outliers. Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, 253-287.

Forcini, G. (2002) Optimal similar tests for structural change for the linear regression model. *Econometric Theory*, **18**, 853-867.

Heij, C., De Boer, P., Franses, P. H, Kloek, T. and Van Dijk, H. K (2004) Econometric Methods with Applications in Business and Economics. Oxford University Press.

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. (1988) A multiple decision theory analysis of structural stability in regression. *Econometric Theory* **4**, 499-508.

Martin, G. M. (2000) US Deficit Sustainability: A new approach based on multiple endogenous breaks, *Journal of Applied Econometrics*, **15**, 83-105.

Perron, P. (2006) Dealing with structural breaks, in Palgrave Handbook of Econometrics. Vol. 1: Econometrics Theory, K. Patterson and T.C. Mills (eds).

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.

Stephens, D. A. (1994) Bayesian retrospective multiple change point identification, *Applied Statistics*, **43**, 159-178.

Appendix

The Minimum Sum of Squares Procedure

In the section we specifically allow ω_q to be a matrix and δ a vector to maintain compatibility with the treatment of Bai (1997). Using the current notation, the minimum sum of squares procedure (MSS) for finding a break date is equivalent to using the argmax of a Wald statistic (see Eq (5) of Bai (1997)) i.e.

$$\arg\max_{q} \hat{\delta}_q' \omega_q' M \omega_q \hat{\delta}_q \tag{7}$$

where $\hat{\delta}_q$ is the OLS estimator of δ in the regression of y on X and ω_q with $M = \left[I - X\left(X'X\right)^{-1} X'\right]$. This equivalence simply expresses the fact that minimising residual error is equivalent to maximising fit. An obvious thing to do is to project X out of the problem since it does not depend on q. Hence, we first regress y on X to get residuals $r_y = My$ and then regress ω_q on X to get a second set of residuals $r_q = M\omega_q$, which may be a matrix. The estimate of δ is then $\hat{\delta}_q = \left(r_q' r_q\right)^{-1} r_q' r_y$ obtained by regressing r_y on r_q

for each $q \in Q$.[4] Hence

$$\hat{\delta}_q' \omega_q' M \omega_q \hat{\delta}_q = r_y' r_q' \left(\omega_q' M \omega_q\right)^{-1} r_q' r_y$$

which is just the sum of squares of the elements of the vector

$$\left(\omega_q' M \omega_q\right)^{-1/2} r_q' r_y = \left(\omega_q' M \omega_q\right)^{-1/2} \omega_q' r_y$$

where $\left(\omega_q' M \omega_q\right)^{-1/2}$ is interpreted as a square root matrix.

When δ is a scalar and $\omega_q = \xi_q w_q$, then $\left(\omega_q' M \omega_q\right)^{-1/2} r_q' r_y = w_q \xi_q' r$ using the notation $r = r_y$ of Section 2. Hence the MSS technique using $\arg\max_q \hat{\delta}_q' \omega_q' M \omega_q \hat{\delta}_q$ is equivalent to $\arg\max_q \left(w_q \xi_q' \widetilde{r}\right)^2$. Hence the MSS equals the weighted $CUSUM$ and is therefore optimal.

When $\xi_q w_q$ is a matrix product, with $w_q = \left(\xi_q' M \xi_q\right)^{-1/2}$ a square root matrix and δ is a vector, the prior is still noninformative over the possible break points but the statistic that would emerge is

$$\max_{q \in Q} \frac{r' \xi_q w_q \delta \delta' w_q' \xi_q' r}{r' r}$$

which is not in general uniform in δ as is to be expected. In the special case where $\delta = \tilde{\delta} k$ for some known vector k and scalar $\tilde{\delta}$, then the statistic is uniform and is given by

$$\max_{q \in Q} \frac{r' \xi_q w_q k k' w_q' \xi_q' r}{r' r} \tag{8}$$

and the $\arg\max_q$ gives the optimal break date estimate. For example, if the δ of the shift in the intercept and the slope of a simple regression model were considered equal then $k' = (1, 1)'$ and (8) is optimal. The continuous trend beak model is an example of this. Of course it is perfectly possible to compute $r' \xi_q w_q K w_q' \xi_q' r$ for any known weighting matrix

[4] This idea is also known as the Frisch-Waugh-Lovell Theorem.

K we choose and use the *max* and arg max as a class of break location estimating devices regardless of the dimension of the unknown δ. The Bai procedure uses the choice $K = I$. Members of such a class would no longer necessarily be Bayes rules, of course.

Trend Regression.

Consider the regression where X consists of a column on 1's and the variable x_t. Using $\Delta = \sum_{j=1}^{N} (x_j - \bar{x})^2$ and $x_2 = \sum_{j=1}^{N} x_j^2/N$ we see

$$(X'X)^{-1} = \Delta^{-1} \begin{bmatrix} x_2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}.$$

Next define $\tilde{x}_j = x_2 - \bar{x}x_j$ and $\hat{x}_j = x_j - \bar{x}$ and associated vectors $\tilde{\mathbf{x}}$ and $\hat{\mathbf{x}}$. Hence X $(X'X)^{-1} X' = \tilde{\mathbf{x}}\mathbf{1}' + \hat{\mathbf{x}}\mathbf{x}'$. After a bit of algebra in the trend notation and using orders of magnitude in line (9)

$$\begin{aligned}
c_q &= \mathbf{t}_q' M \mathbf{t}_q = \mathbf{t}_q'\mathbf{t}_q - \Delta^{-1}\mathbf{t}_q'\tilde{\mathbf{x}}\mathbf{1}'\mathbf{t}_q - \Delta^{-1}\mathbf{t}_q'\hat{\mathbf{x}}\mathbf{x}'\mathbf{t}_q \\
&= \mathbf{t}_q'\mathbf{t}_q - \Delta^{-1}\mathbf{t}_q'\left(t_2\mathbf{1} - \bar{t}\mathbf{t}\right)\mathbf{1}'\mathbf{t}_q - \Delta^{-1}\mathbf{t}_q'\left(\mathbf{t} - \bar{t}\mathbf{1}\right)\mathbf{t}'\mathbf{t}_q \\
&= \mathbf{t}_q'\mathbf{t}_q - t_2\Delta^{-1}\left(\mathbf{1}'\mathbf{t}_q\right)^2 - \Delta^{-1}\left(\mathbf{t}_q'\mathbf{t}_q\right)^2 + 2\bar{t}\Delta^{-1}\mathbf{1}'\mathbf{t}_q\mathbf{t}_q'\mathbf{t}_q \\
&\approx \left[\frac{T^3}{3} - \frac{q^3}{3}\right] - \left[\frac{T^2}{3}\frac{12}{T^3}\left(\frac{T^2}{2} - \frac{q^2}{2}\right)^2\right] - \left[\frac{12}{T^3}\left(\frac{T^3}{3} - \frac{q^3}{3}\right)^2\right] \qquad (9) \\
&+ \left[T\frac{12}{T^3}\left(\frac{T^2}{2} - \frac{q^2}{2}\right)\left(\frac{T^3}{3} - \frac{q^3}{3}\right)\right] \\
&= \frac{2}{T^2}q^5 - \frac{1}{T}q^4 - \frac{4}{3T^3}q^6 + \frac{1}{3}q^3 \\
&= T^3\left(\frac{1}{3}\tau^3 - \tau^4 + 2\tau^5 - \frac{4}{3}\tau^6\right) \\
&= T^3\frac{1}{3}\tau^3\left(1 - \tau\right)\left(4\tau^2 - 2\tau + 1\right)
\end{aligned}$$

In the case of the broken trend we get

$$c_q = \mathbf{b}_q'\mathbf{b}_q - t_2\Delta^{-1}\left(\mathbf{1}'\mathbf{b}_q\right)^2 - \Delta^{-1}\left(\mathbf{t}'\mathbf{b}_q\right)^2 + 2\bar{t}\Delta^{-1}\mathbf{1}'\mathbf{b}_q\mathbf{b}_q'\mathbf{t}$$

and $\mathbf{b}_q'\mathbf{t} = \sum_{j=1}^{T-q} j\left(q+j\right) = \frac{1}{6}\left(T-q\right)\left(T-q+1\right)\left(2T+q+1\right)$. Using orders of magnitude as before

$$\begin{aligned} c_q &\approx \frac{\left(T-q\right)^3}{3} - \frac{T^2}{3}\frac{12}{T^3}\left(\frac{\left(T-q\right)^2}{2}\right)^2 - \frac{12}{T^3}\left(\frac{1}{6}\left(T-q\right)\left(T-q\right)\left(2T+q\right)\right)^2 \\ &+ T\frac{12}{T^3}\frac{1}{6}\left(T-q\right)\left(T-q\right)\left(2T+q\right)\frac{\left(T-q\right)^2}{2} \\ &= \frac{1}{3T^3}q^3\left(T-q\right)^3 \\ &= \frac{T^3}{3}\tau^3\left(1-\tau\right)^3. \end{aligned}$$