Forecasting when there is a Single Break

Graham Elliott

University of California, San Diego

9500 Gilman Drive

LA JOLLA, CA, 92093-0508

February 24, 2005

Abstract

Many authors attribute poor forecast performance to breaks in the model parameters. This has led to forecasting techniques that attempt to estimate the breaking process. This paper presents analytical results on why, in the context of a single break, attempts to forecast based on point estimates for the break are unlikely to improve forecasts. Instead we suggest an averaging approach that yeilds forecasts with much smaller mean squared error for many break models.

1 Introduction

Forecasting economic quantities is notoriosly difficult. Despite well developed theoretical results on regression theory, in practice models never seem to work so well. Various explanations can be made, from the size of estimation error to heterogeneity in the forecasting model relationship.

The possibility that the problem is due to breaks in the parameters of the forecasting models is suggested empirically by results such as those in Stock and Watson (1996). They showed for a vast number of relationships between macroeconomic evidence for instability of parameters linking these variables in linear regression. They examined the use of time varying parameter models as a method for improving forecast construction in this environment.

An alternative strategy that has been followed by a number of authors is to attempt to estimate the timing of the breaks and take this into account in the forecasting model. This strategy seems reasonable when it is expected that the parameter heterogeneity takes the form of a small number of breaks relative to the sample size. Methods such as least squares estimation of the breaks such as proposed by Bai and Perron (1998) have been employed to undertake the modelling of the forecasting regression.

In practice this method has often not been found to be a panacea. Pesaran and Timmerman (2002) examined a forecasting regression where they found that the method of estimating the break by least squares did not improve forecasting performance greatly. They suggested an alternate estimation strategy that hinged on reversing the order of the data, testing for a break and cutting the sample at the first sign that the test rejected the null of no breaks. They argued that this method could be seen as a data dependent rolling regression method, where the test procedure determined the sample length. This method provided gains in forecasting ability for their empirical problem.

This paper seeks to provide both an explanation for when and why least squares estimates for the break do not help with constructing forecasts, and to provide a robust approach to forecasting when there is a single break in the parameters that has good properties in situations where least squares methods succeed and better properties in cases where it fails. The model is one where there is a single break in a subset of the parameters of a linear forecasting regression. We derive analytical results that show why least squares methods fail for models that appear to be relevant empirically.

In the next section we examine the problem of forecasting when there is a single break in the parameters. We show that the standard tradeoffs between bias and estimation error from excluding or including the break parameter differ from standard regression cases. In the third section we then examine the least squares estimation of the break point and the model parameters. We then turn in a fourth section to constructing robust forecasting methods. The methods are examined in a Monte Carlo experiment in section 6. A final section draws conclusions.

2 Forecasting in the Presence of a Break

This paper considers the linear time series regression model

$$y_t = X_t'\beta + \mathbf{1}[t > \tau_0]X_t'\delta + Z_t'\gamma + u_t \quad t = 1, \cdots, T \tag{1}$$

where $\mathbf{1}[\cdot]$ is the indicator function, y_t is a scalar, X_t, β and δ are $k \times 1$ vectors, Z_t and γ are $p \times 1$, $\{y_t, X_t, Z_t\}$ are observed, τ_0, β, δ and γ are unknown and $\{u_t\}$ is a mean zero disturbance. Since we are intending to use this model as a forecasting model, we assume that $\{X_t, Z_t\}$ are available for forecasting the outcome variable y_t which is not observable at the same time. The dating convention is employed to increase the compatability of the notation with previous results available in the literature on estimation and testing for breaks. Define $Q_t = (X_t', Z_t')'$. Let '$\xrightarrow{p}$' denote convergence in probability and '$\Rightarrow$' convergence of the underlying probability measures as $T \to \infty$,and let $[\cdot]$ be the greatest lesser integer function.

The problem we are interested in is forecasting from this model when the date of the break, τ_0, is unknown to the researcher. We are assuming throughout that there is only a single break in the coefficients on the X_t covariates. In addition we are assuming that we know there are no breaks in the coefficient of some additional covariates Z_t. We assume that we are interested in the best linear projection, that is we would like to obtain the forecast procedure

$$y_{T,1} = X_{T+1}'(\beta + \delta) + Z_{T+1}'\gamma$$

so the problem reduces to finding estimators for β, δ and γ.

First consider the possibility that we ignore the break and simply run the regresssion without the break term, i.e. consider estimating the model

$$y_t = X_t'\beta^* + Z_t'\gamma^* + u_t^* \quad t = 1, \cdots, T \tag{2}$$

and using the forecast $y^*_{T,1} = X'_{T+1}\hat{\beta}^* + Z'_{T+1}\hat{\gamma}^*$ where the 'hat' indicates OLS estimation. As with the inclusion or exclusion of any covariate in a regression, when the additional variable enters with a coefficient small enough there is little effect on the bias of the forecast and a saving in terms of reducing estimation error. However when the coefficient on the omitted variable is large enough, the bias term will eventually dominate the forecast error. In the latter case we are unlikely to ignore the break in the coefficients since tests for the detection of this type of model misspecification will easily reject the null of excluding this variable, and hence this error will not be made. What is different for this problem compared to the typical decision of whether or not to include or exclude an observed regressor is that this regressor is not observed and this affects the tradeoffs between the bias and variance reduction of exclusion differ. The additional difficulties are due to the unknown value for τ_0. This affects the bias we expect to see when we exclude the variable, the impact on the sampling variation from including the covariate, and also the ease of detecting when we are sure that this variable should enter the forecasting model.

We assume the following regularity condition on model (1):

Condition 1 *(i)* $\tau_0 = [r_0 T]$ *for some* $0 < r_0 < 1$.

(ii) $T^{-1/2}\sum_{t=1}^{[sT]} Q_t u_t \Rightarrow \Omega^{1/2}\tilde{W}(s)$ *for* $s \in [0,1]$ *with* Ω *some symmetric and positive definite* $k+p \times k+p$ *matrix and* $\tilde{W}(\cdot)$ *a* $k+p \times 1$ *standard Wiener process.*

(iii) $T^{-1}\sum_{t=1}^{[sT]} Q_t Q'_t \xrightarrow{p} s\Sigma_Q = s\begin{pmatrix} \Sigma_X & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_Z \end{pmatrix}$ *uniformly in* $s \geq 0$, *where* Σ_Q *is full rank.*

In the asymptotic thought experiments considered in this paper, the data that precedes and follows the break are in the fixed proportion $r_0/(1-r_0)$. This thought experiment is standard in the breaking literature, although recently alternative asymptotics have been considered by Andrews (2003). With $\tau_0 = [r_0 T]$, the data generated by this model necessarily becomes a double-array, as τ_0 depends on T, although we do not indicate this dependence on T to enhance readability. The remaining conditions are standard high-level time series conditions, that allow for heterogeneous and autocorrelated $\{u_t\}$ and regressors $\{Q_t\}$. Condition 1 also accommodates regressions with only weakly exogenous regressors.

When this condition holds the bias in the estimated coefficients when the break is ignored

is the first term on the right hand side of

$$T^{1/2}\begin{pmatrix} \hat{\beta}^* - \beta - \delta \\ \hat{\gamma}^* - \gamma \end{pmatrix} = \left((\sum_{t=1}^{T} Q_t Q_t')^{-1} (\sum_{t=1}^{T} Q_t (X_t' - \mathbf{1}[t > \tau_0] X_t') - e_1 \right) T^{1/2}\delta \quad (3)$$
$$+ (T^{-1}\sum_{t=1}^{T} Q_t Q_t')^{-1} (T^{-1/2}\sum_{t=1}^{T} Q_t u_t)$$

where $e_1 = (I_k, 0_{kxp})'$. In the limit the coefficient on $T^{1/2}\delta$ is given by $r_0 e_1$. As we might expect when the break is close to the end of the sample this term is near one and the problem is merely that the OLS coefficients have no chance to pick up the effects of the break. The mistake in the forecast is merely this break effect on the coefficient on X_T. When the break is closer to the start of the sample then through the usual omitted variable bias effect we do pick up part of the effect, and the problem for forecasting is mitigated somewhat.

As noted and as is clear from the immediately preceeding expression the extent to which ignoring such a break in forecasting causes problems depends on the size of δ. Since the last term in (3) is $0_p(1)$ then the bias is of the same order as the sampling error when the break size is such that $T^{1/2}\delta$ is constant. Not coincidentally this is precisely the size of the break that is 'just' detectable by reasonable tests for a break. It is for breaks of this order that there is uncertainty over the break, tests have non negligible power in detecting such breaks but power is not equal to one. When breaks are of a larger magnitude, so that $T^{1/2}\delta$ is increasing in the sample size, then tests have unit power asymptotically and we know for sure that the break exists. We will refer generically to breaks that satisfy the condition that $T^{1/2}\delta$ converges to a nonzero constant as small breaks and those that diverge to be large breaks.

In many practical applications, breaks that are of interest are arguably not large in this sense. After all, formal econometric tests for the presence of breaks are employed precisely because there is uncertainty about the presence of a break. From an empirical point of view, the observed p-values are often borderline significant; in the Stock and Watson (1996) study, for instance, the QLR statistic investigated by Andrews (1993) rejects stability of 76 US postwar macroeconomic series for 23 series on the 1% level, for an additional 11 series on the 5% level, and for an additional 6 series on the 10% level.

Breaks that are small in the statistical sense of corresponding to $\delta T^{1/2}$ small are, of course, not necessarily small in an economic sense. As usual, economic and statistical significance

are two very distinct concepts. As an example, consider the possibility of a break in growth. Post-war quarterly U.S. real Gross Domestic Product growth (multiplied by 100) has a standard deviation of about unity. Even if growth is i.i.d. Gaussian, this variation will make it very difficult to detect, let alone date, a break of mean growth that is smaller than 0.25 percentage points. With 225 or so quarterly observations such a break results in $\delta T^{1/2}$ of around 4 — optimal tests for a single break in the mean of this size have power at best around 50% and if the break is not near the center of the sample far below this percentage. However breaks that lead to a change in yearly growth of the magnitude of 1 percentage point are going to be important for forecasting performance.

We now turn to the problem of including the break variable when we potentially set an incorrect break date. In this case we run the regression

$$y_t = X_t'\beta + \mathbf{1}[t > \tau]X_t'\delta + Z_t'\gamma + u(\tau)_t \quad t = 1, \cdots, T \tag{4}$$

where τ may or may not equal τ_0. Denote the OLS estimates for $\{\beta, \delta, \gamma\}$ in (1) by $\{\hat{\beta}(\tau), \hat{\delta}(\tau), \hat{\gamma}(\tau)\}$.

Theorem 1 *For the model given in (1) then when Condition (1) holds, $W(.)$ is a standard k dimensional Weiner process and $T^{1/2}\delta = \sigma d$ then for the OLS coefficients from the regression (4) when τ is incorrectly specified compared to correctly specified is*

$$T^{1/2}\left(\hat{\beta}(\tau) + \hat{\delta}(\tau) - \hat{\beta}(\tau_0) - \hat{\delta}(\tau_0)\right) \Rightarrow \sigma d\left(\frac{\min(r, r_0) - r_0}{(1-r)}\right) + \Sigma_X^{-1}\Omega_X^{1/2}\left(\frac{(1-r_0)W(r) - (1-r)W(r_0) + (r-r_0)W(1)}{(1-r)(1-r_0)}\right)$$

where.$r = [\tau T]$ and $\Omega_X^{1/2}$ the upper kxk block of $\Omega^{1/2}$ where $\Omega^{1/2}$ is chosen to be lower block triangular.

The first term on the right hand side gives the bias from using the incorrect break date. The second term shows the additional variation in the coefficient estimate relating to the stochastic estimation error.

Consider first the bias term. When the break date is chosen to be equal to or later than the correct break date then $\min(r, r_0) = r_0$ and this term is zero. Hence there is no bias term for the purposes of forecasting so long as the break date is not chosen too early. When the break date is chosen to be earlier than the true break date then there is a bias term that

depends both on the true break proportion and the difference between the one chosen and the true proportion.

The second term captures the stochastic part of the difference between the two OLS estimators. The term is normally distributed with zero mean and has variance $\Sigma_X^{-1}\Omega_X^{1/2}(r - r_0)/((1-r)(1-r_0))$. This term is a function not only of the size of the misspecification of the break date $(r - r_0)$ but also of the positions of the true and imposed break dates.

Whilst in terms of minimizing bias we would prefer to choose the break date after the true break date, the later the choice the larger is the final term in the expression. This term captures the effect on the distribution of the estimator as a function of the chosen break date τ. The later the break date the closer is r to 1 and the greater the variation in this term.

What this implies is an assymetric tradeoff between variance and bias in the choice of the break date. From the perspective of minimizing mean square error this suggests that any attempt to estimate the break date will be affected by the assymetry in the costs of incorrectly specifying the correct break date. To examine the form of the assymmetry, consider the model where $\gamma = 0$ and $X_t = 1$. In this case the contribution of estimation error from including the break term to the mean square forecast error is given by

$$v(r, r_0) = \sigma^2 \left[d^2 \left(\frac{\min(r, r_0) - r_0}{1 - r} \right)^2 + \frac{1}{1 - r} \right]$$

For $r > r_0$ this is $(1-r)^{-1}$ which increases in r, hence we want to choose r as small as possible in this range to minimize the variance. For $r < r_0$ the bias increases the smaller the choice of r, although here there is also a variance effect that depends on the value of r. Figure XX shows the tradeoffs for two values for r_0— $r_0 = 0.3$ and 0.7. For the small breaks the assymetry is present but not obviously important. There are costs to over and underestimating the true break point.

The effect of incorrectly choosing the break date when the break is large differs from the situation when the break is small. When the break is large the bias effect dominates the variance effect and so that the cost asymptotically to choosing the break date τ to be after the true break date τ_0 is small relative to the cost to estimating the break to be too early. This suggests a method for large breaks that is highly assymetric, placing most of the estimation mass on breaks subsequent to the true break.

These effects are pictured in Figure (1). Here we examine $v(r, r_0)$ for small and large d

and r_0 equal to 0.3 and 0.7. Apart from all methods doing very poorly when the choice of r is near the last observations of the sample, the effects described above hold. In all cases the tradeoffs between over and underestimating the true break point are assymmetric. However this masks a great difference between the small and large breaks case. In the first panel the break is small ($d = 5$) and whilst we see assymmetry it is relatively small. In the second panel the effect is very large — the cost to estimating a break date too early in the sample far outweighs the cost to choosing a break date too late in the sample.

3 Forecasts Based on OLS Estimates

Commonly researchers adopt the approach of 'rolling' regressions to estimate models in which they suspect there are (usually unspecified) breaks in the parameters of the forecasting regression. For any particular time period this involves dropping a prespecified number of observations from the earlier part of the sample — giving these observations zero weight — and giving even weights to the remaining observations. This approach suffers from a number of disadvantages. First, at best it would make sense only if all the parameters of the regression changed values, if there were any constant parameters this would be innefficient. Second, in updating for the next period the typical approach is to keep the forecast model data series the same length as that for the previous period by dropping the first observation of the data used for last periods forecast. This means that an observation that recieved full weight for forecasting one period before is now deemed completely useless. It would seem, at least in the context of a single break, that some more data based approach to choosing how far back to go in using observations for the forecast would be in order.

A natural approach to overcoming the problem that τ_0 is unknown is to attempt to estimate this parameter from the data. In the case where data is independent and identically distributed estimators have been proposed for the break proportion r_0. Maximum likelihood estimators have been suggested by Hinkley (1970) when $X_t = 1, Z_t = 0$, Bhattacharya (1987)). Least squares methods have been suggested in Picard (1985), Bai (1994,1997) and Antoch et. al (1996) when $X_t = 1, Z_t = 0$). The least squares methods allow for the possibility of serial dependence in the time series. The MLE has been shown to be a consistent estimator of the proportion. Yao (1987) derives an asymptotic distribution for the MLE of τ_0 for change in a parameter of the distribution for y_t, Krishnaiah and Miao

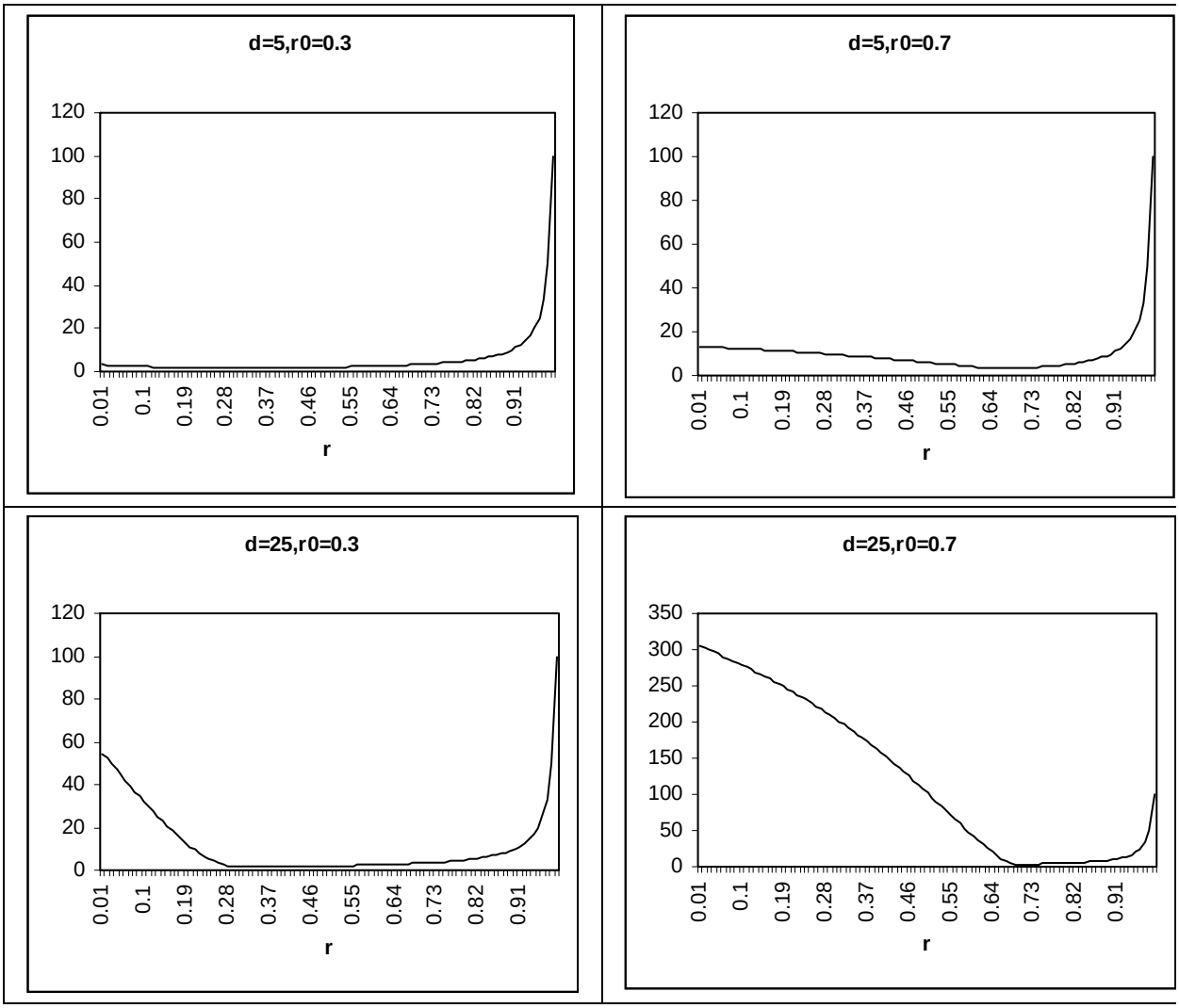

d=5,r0=0.3
120
100
80
60
40
20
0
0.01
0.1
0.19
0.28
0.37
0.46
0.55
0.64
0.73
0.82
0.91
r
d=5,r0=0.7
120
100
80
60
40
20
0
0.01
0.1
0.19
0.28
0.37
0.46
0.55
0.64
0.73
0.82
0.91
r
d=25,r0=0.3
120
100
80
60
40
20
0
0.01
0.1
0.19
0.28
0.37
0.46
0.55
0.64
0.73
0.82
0.91
r
d=25,r0=0.7
350
300
250
200
150
100
50
0
0.01
0.1
0.19
0.28
0.37
0.46
0.55
0.64
0.73
0.82
0.91
r

Figure 1:

(1988) for iid data and Nunes et al (1995) for dependent but not integrated observations when $Z_t = 0$. The rate of convergence of the estimators has also been derived (Yao (1987), Bhattacharya (1987), Bai (1994,1997)).

We will focus here on least squares estimators. Define the variance estimator $\hat{\sigma}^2(\tau) = T^{-1}\hat{u}(\tau)'\hat{u}(\tau)$ where $\hat{u}(\tau)$ is the vector of OLS residuals from the regression (1) for the stated τ. For any matrix A or row dimension T define $A(r)$ as equal to A for the first r rows and zero after (for all columns).Then the OLS estimator for the break point, denoted $\hat{\tau}$, is simply the value for τ that minimizes $\hat{\sigma}^2(\tau)$ over all the possible break dates. Minimizing the variance is identical to solving

$$\hat{\tau} = \arg \max_{\tau=[\bar{\lambda}T],\ldots,T-[\bar{\lambda}T]} \hat{\delta}(\tau)'(X(\tau)'M_Q X(\tau))\hat{\delta}(\tau)$$

where $\hat{\delta}(\tau)$ is the OLS estimator given τ, $\bar{\lambda}$ is a trimming parameter and $M_Q = I - Q(Q'Q)^{-1}Q'$. Bai (1997) shows under weaker conditions than those in Condition 1 that for this model the break proportion $\hat{r} = T^{-1}\hat{\tau}$ is a consistent estimator for the true break proportion r_0 provided that the size of the break is large. The consistency result holds so long as the break is of the order $\delta = \sigma d T^{1/2+\epsilon}$ for some $d \neq 0$ and $0 < \epsilon < \frac{1}{2}$. Hence for consistent estimation of the break date we require that the breaks be an order of magnitude larger than those that can be detected by hypothesis tests for the break. The reason for this restriction on the size of the breaks is immediately apparant from the discussion after (3) — it is for large breaks that the bias term dominates the stochastic component of the estimation error and delivers the consistency result.

Under the conditions presented here the distribution for large breaks is nonstandard but symmetric around the true break proportion. Hence from the perspective of forecasting, where the costs of errors are not symmetric around the true break proportion, the least squares approach need not be the most sensible estimator for the break point.

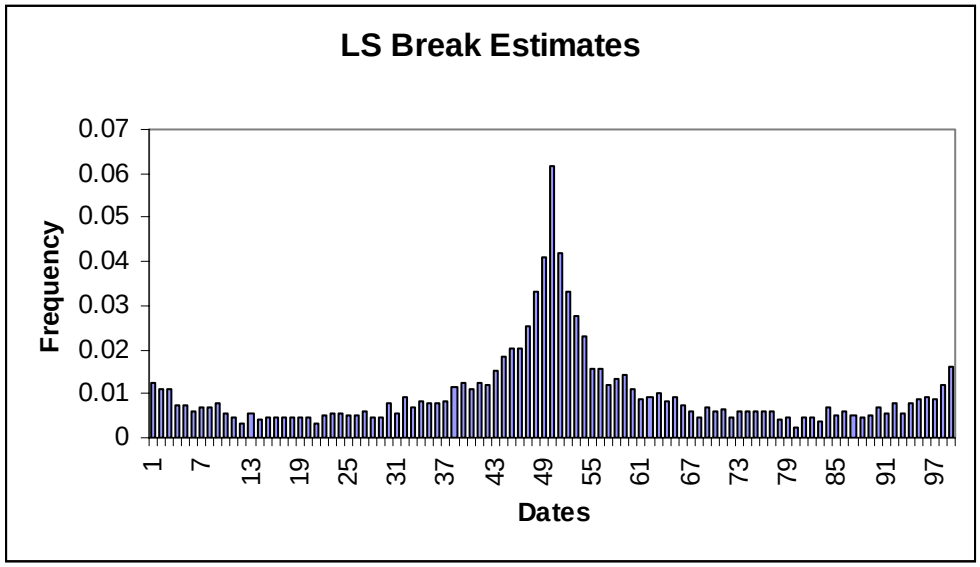
When breaks are small the least squares method does not provide a consistent estimate of the break point. Further, as shown by Elliott and Mueller (2004) the distribution of the break point estimator is a complicated functional of the underlying Brownian motions. In particular they show that when $\delta = T^{1/2}\sigma d$ that the break proportion converges to

$$\hat{r}_a = \arg \min_{\bar{\lambda}<r<1-\bar{\lambda}} \frac{M(r)'\Sigma_x^{-1}M(r)}{r(1-r)} \tag{5}$$

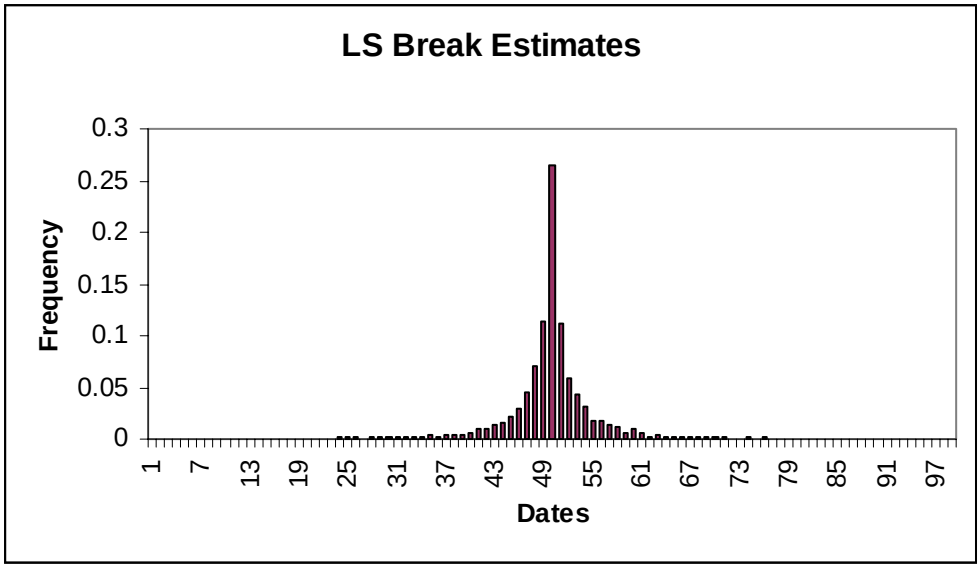
where $M(r) = \Omega_X^{1/2}(W(r) - rW(1)) + \sigma d(\min(r, r_0) - rr_0)\Sigma_x$. Whilst the shape of the distribution is not clear from the analytic results, we can examine this in a Monte Carlo experiment. Figures 2 through 5 show the distribution of break points estimated for a model where $X_t = 1, Z_t = 0$ and $T = 100$. In figures 2 and 3 the results are for a break in the center where in Figure 2 $d = 4.5$ which corresponds to a test with local power of around 50% whereas for Figure 3 $d = 10$ which corresponds to the point where local power is getting close to unity. In the latter case, the distribution has the shape somewhat closer to that expected for the large break case but in the former picture the distribution of estimated break dates is spread out across all the possible dates, quite evenly apart from a peak around the true values and lessor peaks at each end. Clearly for many of the samples the estimated break date is far from the true one. The distribution also looks symmetric in each case, which is relevant in assessing how well this estimator makes the tradeoff between the signs of the break date error.

Figures 4 and 5 repeat this experiment for a break at the 30^{th} observation. In Figure 4 $d = 5.1$ which again is roughly the point where tests have 50% power, whereas in Figure 5 the alternative is $d = 12$. For the larger break the break dates are clustered around the true break and the distribution is symmetric around this point. For the smaller break however, just as in the case of a break in the center of the sample, the estimated break points are spread all across the possible choices for a break date and there are lessor peaks at the ends of the parameter space. For the smaller break many of the estimates are very far from the true break point.

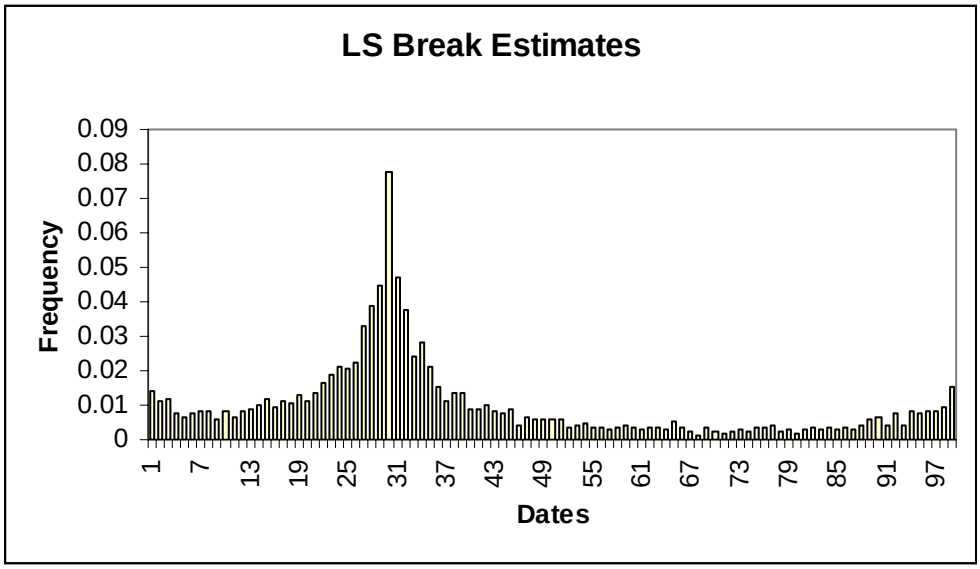
The influence of the trimming on the sampling properties of the least squares estimator for the break proportion when breaks are small has asymptotic impact, a result that follows from the dependence of the asymptotic result in (5) on $\bar{\lambda}$. Clearly from these figures it will also be empirically important. This is clear from the mass of estimates for $\hat{\tau}$ shown in the figures below, which do not trim the range of the estimator. In practice (results not reported) the influence of trimming has a very large impact on the behavior of $\hat{\tau}$ and statistics built off this parameter estimate. In the results presented later we set the trimming parameter at 2%.



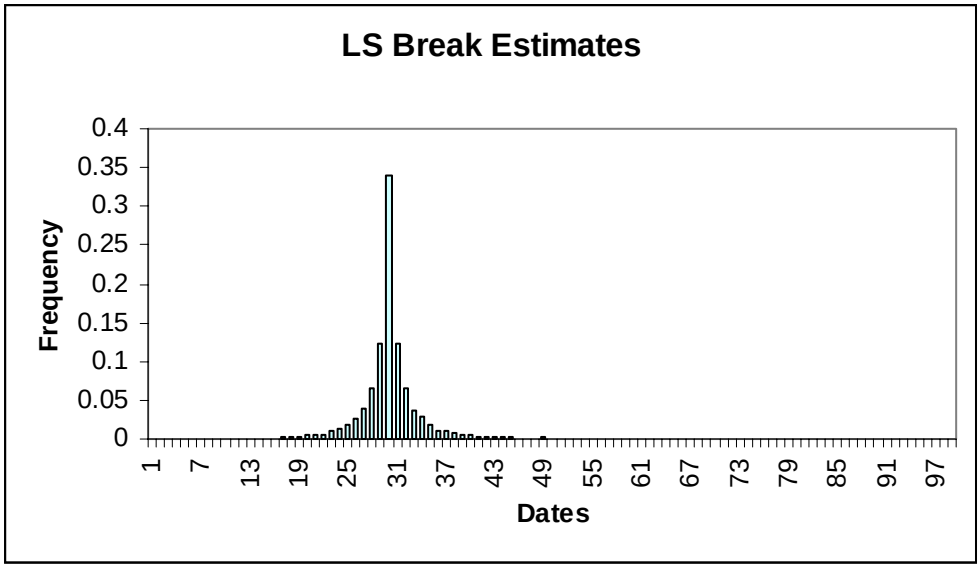
Break in constant where $d = 4.5$ and $r_0 = 0.5$, 100 observations.



Break in constant where $d = 10$ and $r_0 = 0.5$, 100 observations.



Break in constant where $d = 5.1$ and $r_0 = 0.3$, 100 observations.



Break in constant where $d = 12$ and $r_0 = 0.3$, 100 observations.

Pesaran and Timmerman (2002) provide an empirical example where these types of methods do not work well. The analytic results are suggestive of the reasons why the approach might fail. If breaks really are of the 'just' detectable range, then for many forecasting situations the break estimate may be far from the actual break point. On average we would expect this to be true. It is also clear that there is nothing in the least squares approach to estimating the break point that takes into account the assymetry in the costs to forecasting

of getting the break date wrong.

When there are large breaks then from the consistency results of Bai (1997) we have that $\left(\hat{\beta}(\hat{\tau}) + \hat{\delta}(\hat{\tau}) - \hat{\beta}(\tau_0) - \hat{\delta}(\tau_0)\right) \rightarrow^p 0$ and hence asymptotically employing the estimated break date is identical to employing the true break date in the forecasting model. When the break is small, we have the following corollary to Theorem 1 and the results of Elliott and Mueller (2004)

Corollary 1 *For the model given in (1) then when Condition (1) holds, $W(.)$ is a standard k dimensional Weiner process and $T^{1/2}\delta = \sigma d$ then for the OLS coefficients from the regression (4) when τ is incorrectly specified compared to correctly specified is*

$$T^{1/2}\left(\hat{\beta}(\hat{\tau}) + \hat{\delta}(\hat{\tau}) - \hat{\beta}(\tau_0) - \hat{\delta}(\tau_0)\right) \Rightarrow \sigma d\left(\frac{\min(\hat{r}_a, r_0) - r_0}{(1-\hat{r}_a)}\right) + \Sigma_X^{-1}\Omega_X^{1/2}\left(\frac{(1-r_0)W(\hat{r}_a) - (1-\hat{r}_a)W(r_0) + (\hat{r}_a - r_0)W(1)}{(1-\hat{r}_a)(1-r_0)}\right)$$

where $\hat{r}_a$ is defined in (5).

Thus there is an additional random variation in the coefficient estimates that will affect the MSE of the forecast error. In general this interacts in a very complicated way with the underlying estimation error so precise implications are difficult to draw out apart from noting that clearly there are extra terms in the asymptotic expression and hence the expected MSE will differ from other methods for choosing the break date to impose on the model.

For this model the least squares estimator for the break date proportion has no known optimality properties. Because of this, there is no reason to expect that other methods for estimating the break date proportion should not provide better estimates of the break date or smaller forecast MSE's. Pesaran and Timmermann (2002) examine stock prediction regressions and suggest using cusum tests on the data reversed in time. The idea is that they can use this to estimate the most recent break and then construct forecasts based only on the observations subsequent to the most recently estimated break. They found this to work better than the Bai and Perron (1998) approach, which is a multiple break extension of the least squares break estimation procedure examined here. However regardless of the break estimator employed, it will still not be consistent against small breaks.

4 Weighted Average Forecasts

From the perspective of the classical properties of forecasts generated by employing some estimate for the break date, the average performance of any method is a weighted average of the forecasts generated for each value for τ where the weights are given by the distribution of the estimator used for estimating the break point. An alternative approach to basing the forecast on a draw from this distribution is to find some weighted average of each of the models generated by each of the possible choices for the break date.

Consider the following model averageing procedure. Given a density $f(y, X, Z|\tau, \beta, \gamma, \delta)$ and some weights on each value for τ given by $\pi(\tau)$ we have a posterior density for τ given by

$$\pi(\tau|y, X, Z) = \frac{f(y, X, Z|\tau, \beta, \gamma, \delta)\pi(\tau)}{c(y, X, Z)}$$

where $c(y, X, Z) = \sum_{\tau=1}^{T-1} f(y, X, Z|\tau, \beta, \gamma, \delta)\pi(\tau)$ and plays the usual role of ensuring that the weights over τ sum to one. Given τ, we have reasonable estimators for the remaining parameters of the model in the sense that OLS provides estimators with good properties even when the regression (1) is not correctly misspecified, so we can approximate the true density $f(y, X, Z|\tau, \beta, \gamma, \delta)$ with $\hat{f}(y, X, Z|\tau)$ using these estimators.

To actually construct these weights for the models we still require a functional form assumption for the density as well as a specification of the prior weight function $\pi(\tau)$. For the first of these we will assume that the residuals of the forecasting equation are independent and normally distributed with unknown variance,which we will also replace by the least squares estimate. This gives the specification

$$\hat{f}(y, X, Z|\tau) = \left[\hat{\sigma}^2(\tau)\right]^{-T/2}$$

For the weight function $\pi(\tau)$ there is a degree of flexibility that can either relect economic reasoning on the likely positions for the breaks or alternatively can be used to 'reverse engineer' a method that has desirable properties. The family of forecast procedures now provide forecasts

$$y_{T,1} = \sum_{\tau=1}^{T-1} \left(\frac{\left[\hat{\sigma}^2(\tau)\right]^{-T/2} \pi(\tau)}{\sum_{s=1}^{T-1} \left[\hat{\sigma}^2(s)\right]^{-T/2} \pi(s)} \right) \left(X'_{T+1}(\hat{\beta}(\tau) + \hat{\delta}(\tau) + Z'_{T+1}\hat{\gamma}(\tau) \right)$$

The procedure is simple to implement, requiring the running of $T-1$ OLS regressions and then taking the weighted average of the forecast for each where the weights are solely

a function of the sums of squared residuals and $\pi(\tau)$. Each choice of $\pi(\tau)$ results in an estimator with different properties.

The comparison between the behavior of these forecast methods and estimating the break point is helped by noting that the least squares approach involves a 'weighting' where the weights are equal to $1(\tau = \hat{\tau})$, placing all of the weight where $\hat{\sigma}^2(\tau)$ is at a minimum. Notice that for even weights, i.e. $\pi(\tau) = \pi(\tau')$ for all τ, τ' results in the weighted average method giving the greatest weight to exactly the τ that corresponds with the least squares estimate. The difference is that this weight is not unity as in the estimation approach. When the likelihood is relatively flat — as it will be when the size of the break is small — the least squares estimator will tend to jump greatly in value for small changes in the data. This is why in Figures 1 and 3 the estimates are so spread out. In this case, the weights from the averaging method will be all closer together, and hence this extra degree of randomness is 'smoothed out' of the variation in the forecast. As a result, we expect better properties of the forecast. When the break is large, the likelihood is going to be much more informative as to the true break point proportion. Hence in this case both methods will work well. The likelihood is peaked, meaning that the point estimate is likely to be good. But also the weights from the averaging method are likely to be highly concentrated around this point estimate, as even relatively small variations in $\hat{\sigma}^2(\tau)$ will mean very large deviations in $\left[\hat{\sigma}^2(\tau)\right]^{-T/2}$. Thus the two approaches should give similar results when the break is large.

A second weighting approach is motivated by the likelihood that (a) for breaks at the tail end of the sample no method will work well for smaller breaks, due to the lack of observations available for estimating δ, and (b) at the very start of the sample the model picks up the effect of the break well. This weighting scheme suggests placing weights that are smallest for each end of the sample and largest in the center. A model for $\pi(\tau)$ that satisfies this is

$$\pi(\tau) = \frac{\tau}{T}\left(1 - \frac{\tau}{T}\right)$$

This weighting scheme, as in the even weighting scheme, allows the data to dominate when the break is large enough or the sample size is large enough. This follows as for an interior break the weights go to zero far enough away from τ_0 and the values for $\pi(\tau)$ ensure that no value for τ is 'zeroed' out apart from the very ends of the sample.

5 Monte Carlo Comparison

The small sample data generating processes we consider are special cases of model (1)

$$y_t = X_t'\beta + \mathbf{1}[t > \tau_0]X_t'\delta + Z_t'\gamma + u_t \quad t = 1, \cdots, T \tag{6}$$

with $T = 100$. Specifically, we consider four models: (M1) a break in the mean, such that $X_t = 1$ and there is no Z_t, and i.i.d. Gaussian disturbances $\{u_t\}$; (M2) X_t a Gaussian, stationary mean zero first-order autoregressive process with coefficient 0.5 and $Z_t = 1$ with i.i.d. Gaussian disturbances $\{u_t\}$ independent of $\{X_t\}$. The variance of the disturbances is normalized throughout such that the long-run variance Ω of $\{X_t u_t\}$ is unity.

Tables 1 and 2 present the results for the models M1 and M2 respectively. In each case we examine breaks of differing magnitudes $\delta = T^{1/2}d$ where $d = 2.5, 5, 7.5, 10$. In addition this comparison is done for break dates at each interior decile of the data, i.e. for $r_0 = \{0.1, 0.2, ..., 0.9\}$. The numbers presented are average mean square forecast errors for each method as a proportion of the mean square error that would result if the break date was known. Finally, a number of techniques are examined. In each panel the first row gives the relative MSE from ignoring completely the break. The second row gives the MSE where the break date is estimated using least squares. The third panel examines a hybrid of these two approaches where a pretest is employed to choose either omitting the break or including in with least squares estimation of the break date. The final column in each panel is the weighted average method where $\pi(\tau)$ is even over all of the break dates. In addition to these results the final row in each panel gives the empirical power (rejection rate) for the Andrews and Ploberger (1994) test for a single break.

Table 1 examines M1. When the break is relatively small, the results concur with the predictions of the asymptotic theory above. First, ignoring the break does not hurt greatly. As suggested by the results in section 2 the relative cost in ignoring the break rises with r_0, as the earlier the break the better able the omitted variable bias is to correct for this misspecification. This effect dissapears at the very end of the sample, a reflection of the asymptotic theory only applying to interior values for the break point. The tradeoff between the bias of ignoring the break and estimation error from trying to exploit sample information on the break is clearly in favour of taking the bias since the bias is small.

When we use a least squares estimator for the break point the bias can be quite large. In

linear regressions with stationary covariates we expect that the addition to estimation error from the addition of an extra covariate is on average about T^{-1}, or about 1% for a sample size of $T = 100$ and invariant to the true value of the parameter. Here however the increase when the break is set for $d = 5$ is at 12% when the break is near the center and larger when the break is towards either end of the sample. As the break gets large, this additional sampling error from estimating the break date dissapears, but not until values for d that accord with what is effectively unit power of the break test. Pre-testing helps somewhat for this problem. For $d = 5$ we have that the pretest often chooses the no break model as the forecasting model, and accordingly we have that the additional specification error results in MSE rising with r_0. The effect on MSE ranges from 2% to 13%.

For the weighted regressions, when $d = 5$ we have that the effect is to increase MSE by between a low of 3% when the true break is at the center of the sample and 5% when it is near the end of the sample. Such errors are far below that acheiveable using least squares estimates for the break. When the second version of the weighted statistic is employed — weighting the center greater than the ends — the results are even better. For $d = 5$ we have the effect of MSE between 1% and 3%. The cost is for greater d. When $d = 10$ we have that whilst for most models both weighting functions result in almost identical results, at eh very end of the sample the first weighting scheme does better.

When the break is larger, according to values for d where power of tests for a break is close to one, the results for each of the methods apart from ignoring the break are similar. Ignoring the break is obviously going to be costly when the break is large. Least squares estimation of the break date and using the weighted average method give similar results to each other and for large enough breaks to the known break case when r_0 is not too far from the end of the sample. At the end of the sample both methods have trouble forecasting relative to the known break case.

Table 2 gives the results for M2. In the case of a stochastic regressor the qualitative results remain similar although the magnitudes of the effects differ. The upward drift in the MSE as a function of r_0 and d is more apparant.

For the weighted estimators, clearly the even weights version does much more poorly than the center weighted version. The large losses, particularly when the true break is near either end of the data, is not representative but is due to a small number of outliers with a very

large mean square loss. The center weighting downweights the poorly estimated coefficients and hence removes these outliers, so the performance is much better and is directly along the lines of the results in M1. Indeed, the results for the weights2 method are very similar for the stochastic and nonstochastic regressors.

6 Conclusion

We have shown that when breaks are of the order of magnitude that we are uncertain as to their presence that least squares estimators for the break point are random. It is precisely these types of breaks that often appear to be relevant empirically, as breaks of this magnitude are precisely the size of breaks that we require tests to detect them.

The performance of the least squares estimator for these breaks carries accross to the performance of the forecasting model that conditions on the estimated break point. As a result the size of the estimation error from this forecast method can be very large. Hence the explanation of why the method fails on occassion is very likely due to the unusual feature of the method that it fails when the break is relatively small.

We proposed a model averaging method that required simply running regressions for each break date and weighting them by a function of the sums of squared residuals from these regressions. The method is simple and quick to implement and has good properties regardless of the unknown size of the break.

7 Appendix

In matrix form the model is

$$y = X\beta + (X - X(\tau))\delta + Z\gamma + u \tag{7}$$

where $y = [y_1, ...y_T]'$, $X = [X_1', ...X_T']'$, $Z = [Z_1', ..., Z_T']'$, and $u = [u_1, ..., u_T]'$. For any matrix A define $A(r)$ as equal to A for the first r rows and zero after (for all columns). Define the matrix $Q = (X, Z)$ and $E[Q'Q] = \Sigma_Q$ which is partitioned according to Q so that

$$\Sigma_Q = \begin{pmatrix} \Sigma_X & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_Z \end{pmatrix}.$$

Table 1: Monte Carlos results for Model 1, T=100										
	Method	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
d=2.5	no break	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.95
	LS	1.28	1.26	1.24	1.23	1.23	1.23	1.24	1.23	1.20
	Pre-test	1.02	1.02	1.03	1.03	1.04	1.04	1.04	1.02	0.97
	weights1	1.06	1.06	1.05	1.05	1.05	1.04	1.04	1.03	0.99
	weights2	1.02	1.02	1.02	1.02	1.01	1.01	1.01	1.00	0.96
	Power	0.07	0.12	0.15	0.18	0.19	0.18	0.16	0.12	0.08
d=5	no break	1.00	1.01	1.02	1.03	1.05	1.07	1.09	1.10	1.09
	LS	1.22	1.16	1.13	1.11	1.12	1.12	1.13	1.16	1.21
	Pre-test	1.02	1.03	1.05	1.05	1.07	1.08	1.11	1.13	1.12
	weights1	1.05	1.04	1.03	1.03	1.03	1.03	1.04	1.05	1.05
	weights2	1.02	1.02	1.01	1.01	1.01	1.01	1.02	1.03	1.03
	Power	0.14	0.31	0.46	0.55	0.58	0.55	0.47	0.33	0.16
d=7.5	no break	1.00	1.02	1.04	1.08	1.12	1.18	1.24	1.29	1.31
	LS	1.14	1.06	1.04	1.03	1.03	1.04	1.06	1.09	1.18
	Pre-test	1.02	1.03	1.03	1.03	1.04	1.05	1.08	1.16	1.30
	weights1	1.04	1.02	1.01	1.01	1.01	1.02	1.03	1.05	1.10
	weights2	1.01	1.01	1.01	1.01	1.01	1.01	1.02	1.04	1.11
	Power	0.27	0.62	0.81	0.89	0.90	0.89	0.82	0.65	0.30
d=10	no break	1.01	1.04	1.08	1.15	1.23	1.33	1.44	1.55	1.63
	LS	1.06	1.02	1.01	1.00	1.01	1.01	1.02	1.05	1.13
	Pre-test	1.02	1.02	1.01	1.01	1.01	1.01	1.03	1.10	1.40
	weights1	1.02	1.01	1.00	1.00	1.01	1.01	1.02	1.04	1.12
	weights2	1.01	1.00	1.00	1.00	1.00	1.01	1.01	1.04	1.15
	Power	0.45	0.87	0.97	0.99	0.99	0.99	0.97	0.89	0.51

Table 2: Monte Carlo results for Model 2, T=100										
	Method	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
d=2.5	no break	1.00	1.00	1.00	1.00	1.01	1.01	1.01	1.00	0.95
	LS	2.34	2.89	2.65	2.61	2.46	2.56	2.69	2.79	2.71
	Pre-test	1.02	1.02	1.03	1.04	1.04	1.04	1.03	1.03	0.97
	weights1	2.17	2.01	1.87	1.81	1.76	1.77	1.78	1.82	1.79
	weights2	1.03	1.02	1.02	1.02	1.02	1.01	1.01	1.00	0.96
	Power	0.08	0.13	0.17	0.20	0.21	0.20	0.18	0.14	0.09
d=5	no break	1.00	1.01	1.02	1.04	1.06	1.09	1.12	1.14	1.12
	LS	1.93	1.40	1.28	1.24	1.32	1.33	1.33	2.21	2.59
	Pre-test	1.03	1.04	1.04	1.07	1.07	1.09	1.10	1.14	1.12
	weights1	1.89	1.50	1.20	1.17	1.15	1.17	1.26	1.42	1.59
	weights2	1.02	1.01	1.01	1.01	1.01	1.01	1.01	1.03	1.03
	Power	0.18	0.36	0.52	0.61	0.64	0.62	0.53	0.39	0.21
d=7.5	no break	1.01	1.03	1.06	1.10	1.16	1.23	1.30	1.38	1.41
	LS	1.49	1.19	1.04	1.03	1.02	1.03	1.07	1.81	2.36
	Pre-test	1.02	1.02	1.02	1.02	1.02	1.04	1.08	1.18	1.28
	weights1	1.49	1.12	1.01	1.01	1.01	1.01	1.03	1.11	1.41
	weights2	1.01	1.01	1.00	1.00	1.01	1.01	1.01	1.04	1.10
	Power	0.34	0.66	0.83	0.91	0.93	0.91	0.84	0.68	0.36
d=10	no break	1.01	1.05	1.11	1.19	1.30	1.43	1.57	1.72	1.81
	LS	1.24	1.09	1.01	1.01	1.01	1.01	1.01	1.05	2.06
	Pre-test	1.01	1.01	1.01	1.01	1.01	1.01	1.03	1.13	1.40
	weights1	1.16	1.01	1.00	1.00	1.00	1.01	1.01	1.03	1.22
	weights2	1.01	1.00	1.00	1.00	1.00	1.01	1.01	1.03	1.15
	Power	0.49	0.85	0.96	0.99	0.99	0.99	0.97	0.86	0.53

Theorem 2 ***Proof.*** *Proof of Theorem 1*

We can write the regression as

$$y = (X - X(\tau))\delta(\tau) + Q\alpha(\tau) + \varepsilon \tag{8}$$

where $\alpha(\tau) = (\beta(\tau)', \gamma(\tau)')'$. *Now*

$$\begin{aligned}
\begin{pmatrix} \hat{\delta}(\tau) \\ \hat{\alpha}(\tau) \end{pmatrix} &= \begin{pmatrix} (X - X(\tau))'(X - X(\tau)) & (X - X(\tau))'Q \\ Q'(X - X(\tau)) & Q'Q \end{pmatrix}^{-1} \begin{pmatrix} (X - X(\tau))'y \\ Q'y \end{pmatrix} \\
&= \begin{pmatrix} (X - X(\tau))'(X - X(\tau)) & (X - X(\tau))'Q \\ Q'(X - X(\tau)) & Q'Q \end{pmatrix}^{-1} \\
&\quad x \begin{pmatrix} (X - X(\tau))'(X - X(\tau_0)) & (X - X(\tau))'Q \\ Q'(X - X(\tau_0)) & Q'Q \end{pmatrix} \begin{pmatrix} \delta \\ \alpha \end{pmatrix} \\
&\quad + \begin{pmatrix} (X - X(\tau))'(X - X(\tau)) & (X - X(\tau))'Q \\ Q'(X - X(\tau)) & Q'Q \end{pmatrix}^{-1} \begin{pmatrix} (X - X(\tau))'u \\ Q'u \end{pmatrix}
\end{aligned}$$

Now

$$\begin{aligned}
&T^{-1} \begin{pmatrix} (X - X(\tau))'(X - X(\tau_0)) & (X - X(\tau))'Q \\ Q'(X - X(\tau_0)) & Q'Q \end{pmatrix} \\
\rightarrow\ &\begin{pmatrix} (1 - r - r_0 + (r \curlyvee r_0))e_1'\Sigma_Q e_1 & (1 - r)e_1'\Sigma_Q \\ (1 - r_0)\Sigma_Q e_1 & \Sigma_Q \end{pmatrix}
\end{aligned}$$

where $(r \curlyvee r_0)$ *is the minimum of these two values. Substituting* r_0 *for* r *gives the denominator.*

By direct calculation we have that

$$\begin{pmatrix} (1 - r)e_1'\Sigma_Q e_1 & (1 - r)e_1'\Sigma_Q \\ (1 - r)\Sigma_Q e_1 & \Sigma_Q \end{pmatrix}^{-1} = \begin{pmatrix} \left(\frac{1}{r(1-r)}\right)(e_1'\Sigma_Q e_1)^{-1} & -\left(\frac{1-r}{r(1-r)}\right)(e_1'\Sigma_Q e_1)^{-1}e_1' \\ -\left(\frac{1-r}{r(1-r)}\right)e_1(e_1'\Sigma_Q e_1)^{-1} & \Sigma_Q^{-1} + \left(\frac{(1-r)^2}{r(1-r)}\right)e_1(e_1'\Sigma_Q e_1)^{-1}e_1' \end{pmatrix}$$

and multiplying these together we have

$$\begin{aligned}
&\begin{pmatrix} (X - X(\tau))'(X - X(\tau)) & (X - X(\tau))'Q \\ Q'(X - X(\tau)) & Q'Q \end{pmatrix}^{-1} \\
&\quad x \begin{pmatrix} (X - X(\tau))'(X - X(\tau_0)) & (X - X(\tau))'Q \\ Q'(X - X(\tau_0)) & Q'Q \end{pmatrix} \begin{pmatrix} \delta \\ \alpha \end{pmatrix} \\
&= \begin{pmatrix} \frac{(r \curlyvee r_0) - r r_0}{r(1-r)} I_k & 0_{kxp} \\ r^{-1}(r - (r \curlyvee r_0))e_1 & I_p \end{pmatrix} \begin{pmatrix} \delta \\ \alpha \end{pmatrix}.
\end{aligned}$$

Noting that $e_1'\alpha = \beta$ *we have that the coefficient on* γ *is zero and on* $(\beta+\delta)$ *is* $\frac{(r \Upsilon r_0)-rr_0}{r(1-r)} + r^{-1}(r-(r \Upsilon r_0)) - 1 = \frac{\min(r,r_0)-r_0}{(1-r)}$.

For the second term Condition 1(i) yeilds $T^{-1/2}Q'u \Rightarrow \Omega^{1/2}\tilde{W}(1)$ *and*

$$\begin{aligned} T^{-1/2}(X - X(\tau))'u &= T^{-1/2}X'u - T^{-1/2}X(\tau)'u \\ &\Rightarrow e_1'\Omega^{1/2}(\tilde{W}(1) - \tilde{W}(r)) \end{aligned}$$

Hence using the asymptotic result for the inverse of the denominator and these results then

$$\begin{aligned} T^{1/2}\begin{pmatrix} \hat{\beta}(\tau) + \hat{\delta}(\tau) - \beta - \delta \\ \hat{\gamma}(\tau) - \gamma \end{pmatrix} \Rightarrow\ & \sigma d \left(\frac{\min(r, r_0) - r_0}{(1-r)}\right) e_1 \\ & + \Sigma_Q^{-1}\Omega^{1/2}\tilde{W}(1) - e_1\Sigma_X^{-1}\Omega_X^{1/2}\left(\frac{W(r) - rW(1)}{1-r}\right) \end{aligned}$$

where through choosing $\Omega^{1/2}$ *to be lower block triangular with (1,1) block* Ω_X *and* $\tilde{W}(s) = [W(s)', W_z(s)']'$ *where the partition is after the* k^{th} *row.* ■

References

Andrews, D. (1993): "Tests for Parameter Instability and Structural Change with Unknown Change Point," *Econometrica*, 61, 821–856.

——— (2003): "End-of-Sample Instability Tests," *Econometrica*, 71, 1661–1694.

Bai, J. (1994): "Least Squares Esitmation of a Shift in Linear Processes," *Journal of Time Series Analysis*, pp. 453–472.

——— (1997): "Estimation of a Change Point in Multiple Regression Models," *Review of Economics and Statistics*, pp. 551–563.

Bai, J., and P. Perron (1998): "Estimating and Testing Linear Models with Multiple Structural Changes," *Econometrica*, 66, 47–78.

Bhattacharya, P. (1987): "Maximum Likelihood Estimation of a Change Point in the Distribution of Indpendent Random Variables," *Journal of Multivariate Analysis*, 23, 183–208.

Hinkley, D. (1970): "Inference About the Change-Point in a Sequence of Random Variables," *Biometrika*, pp. 1–17.

Krishnaiah, P., and B. Miao (1988): "Review About Estimation of Change Points," in *Handbook of Statistics*, ed. by P. Krishnaiah, and C. Rao. Elsevier, New York.

L. Nunes, M. K., and P. Newbold (1995): "Spurious Break," *Econometric Theory*, pp. 736–750.

Pesaran, M., and A. Timermann (2002): "Market Timing and Return Prediction Under Model Instability," *Journal of Empirical Finance*, 9, 495–510.

Picard, D. (1985): "Testing and Estimating Change-Point in Time Series," *Advances in Applied Probability*, 17, 841–867.

Yao, Y.-C. (1987): "Approximating the Distribution of the MLE Estimate of the Change-Point in a Sequence of R.V.'s," *Annals of Statistics*, pp. 1321–1328.