Testing for a structural change of gradient in regression modelling

Paul White[1], Don J. Webber[2] and Angela Helvin[3]

[1] *Department of Mathematics and Statistics, University of the West of England, Bristol, UK; Email: paul.white@uwe.ac.uk*

[2] *Department of Economics, Auckland University of Technology, Auckland, NZ; Email: don.webber@aut.ac.nz*

[3] *Department of Mathematics and Statistics, University of the West of England, Bristol, UK; Email: angela2.helvin@uwe.ac.uk*

We present a "broken stick" method to model structural changes of gradient in a regression model. The assessment of the statistical significance of the nodes in the model is assessed using bootstrap techniques. The method is illustrated using output data across the EU but more generally applies when modelling changes of gradient in a regression model whilst fixing the change point and preserving continuity.

Corresponding Author: Dr Don J Webber, Department of Economics, Auckland University of Technology, Auckland 1142, New Zealand. E-mail: Don.Webber@aut.ac.nz.

1.   Introduction

In regression analyses the concept of lack-of-fit may be paraphrased as the "need to have different regression equations in different parts of the regression space to adequately model a response variable, Y". The Chow test (Chow, 1960) is often used to determine whether a regression model can be improved upon by incorporating a structural break; if such a break is warranted then the analyst may choose to proceed with separate regression equations or incorporate a break in the model by a judicious use of dummy variables. Andrews (1993), Hansen (1992), Inder and Hao (1996) and Greene (1999) have suggested alternative methods to identify structural changes.

This paper presents an alternative approach to modelling structural breaks through the use of what may be termed a "broken stick" model and is applied to EU GDP per capita data. Broken stick regression is the modelling of two or more intersecting straight lines with the break points forming the piecewise linear regression. We consider the statistical significance of the parameters of a broken stick model both when the nodes (break points) are pre-specified based on prior reasoning and also when they identified from an examination of the data.

2.   Broken Stick Models

Let X denote an explanatory variable of a response Y. Define

$$X_1 = \begin{cases} X & X \leq k \\ \\ k & X > k \end{cases}$$

$$X_2 = \begin{cases} 0 & X \leq k \\ \\ (X - k) & X > k \end{cases}$$

and consider

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon .$$

where ε denotes a random error component. When $X \leq k$ the structural part of the above model reduces to $Y = \beta_0 + \beta_1 X$. When $X > k$ the structural part of the model reduces to $Y = \beta_0 + (\beta_1 - \beta_2)k + \beta_2 X$. If $X = k + \delta x$, $\delta x > 0$ then in the limit as $\delta x \to 0$, $Y \to \beta_0 + \beta_1 X$. This model specification is piecewise continuous and β_1 indicates the rate of change of X with Y for $X \leq k$ and β_2 indicates the rate of change of X with Y for $X > k$. If $\beta_1 = \beta_2$ then the model reduces to a simple linear relationship between X and Y.

A model formulation allowing a double "break" at $X = k_1$ and at $X = k_2$ (with $k_2 > k_1$) is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

where

$$X_1 = \begin{cases} X & \text{if } X \leq k_1 \\ \\ k_1 & \text{if } X > k_1 \end{cases}$$

$$X_2 = \begin{cases} 0 & \text{if } X \leq k_1 \\ X - k_1 & \text{if } k_1 \leq X \leq k_2 \\ k_2 - k_1 & \text{if } X > k_2 \end{cases}$$

$$X_3 = \begin{cases} 0 & \text{if } X \le k_2 \\ \\ X - k_2 & \text{if } X > k_2 \end{cases}$$

where β_1 denotes the rate of change of X with Y for $X \le k_1$; β_2 indicates the rate of change of X with Y for $k_1 < X \le k_2$ and β_3 indicates the rate of change X with Y for $X > k_2$. More generally a model comprising B broken sticks with break points k_b ($b=1, \cdots, \cdots, B-1$), $k_{b-1} < k_b$, is defined by

$$Y = \beta_0 + \sum_{b=1}^{B} \beta_b X_b + \varepsilon$$

where the general terms are defined by

$$X_b = \begin{cases} 0 & \text{if } k_{b-1} > X \\ X \quad - k_{b-1} & \text{if } k_{b-1} \le X \le k_b \\ k_b - k_{b-1} & \text{if } X > k_b \end{cases}$$

3.   Data

We applied the above models to EUROSTAT-sourced GDP per capita (in 1995 prices) data for the 1980–2008 period. The observations are raw averages for all countries included in the sample for each year. The data includes Cyprus, Czech Republic, Estonia, Hungary, Lithuania, Latvia, Malta, Poland, Slovenia and Slovak Republic from 1990 onwards only. The early 1990s was plagued by severe structural changes including declines in output and high inflation rates, the advent of The Maastricht Treaty and EMU. These events may be associated with a structural change in GDP per capita suggesting the possibility of at least one important break point in this dataset; see Figure 1.

Figure 1 GDP per capita in Euros for EU
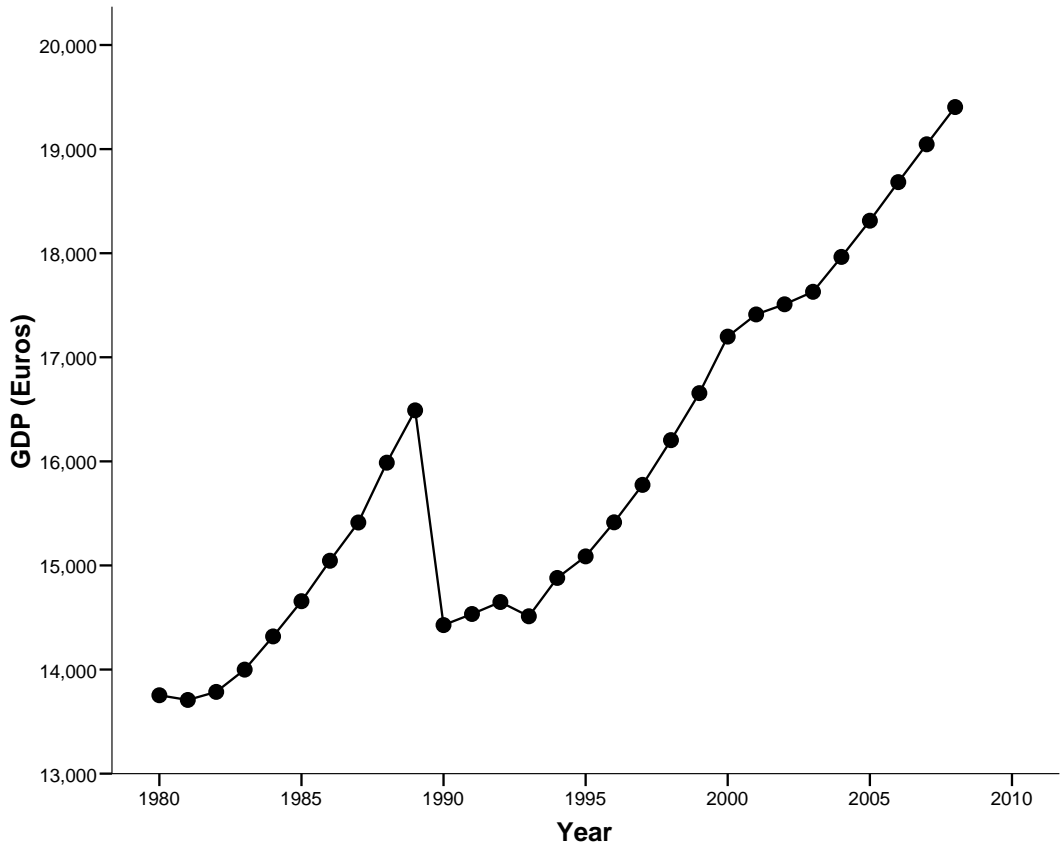
4. Applying the Single Break Model

Consider a broken stick model with a single break at *X=k* and structural specification $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. Consider GDP per capita to be the dependent variable (*Y*) and code the years (1980, 1981, …, 2008) with the values *X*=1,2,…,29. Table 1 summarises the coefficient of determination

(R^2) for each OLS estimation for each possible single break point for the sample data. Under this approach the value of *k* that minimises the within sample error sum of squares is *k*=16, corresponding to the year 1996. Estimated parameters ($\hat{\beta}_0$ =14,026 ; $\hat{\beta}_1$ =79.4; $\hat{\beta}_2$ =312.0) and a summary of standard tests of significance ($H_0 : \beta_j = 0$) is given in Table 2.

Table 1. R²s

Break	R^2	$R^2(adj)$		Break	R^2	$R^2(adj)$		Break	R^2	$R^2(adj)$
1	0.829	0.823		10	0.850	0.838		19	0.900	0.892
2	0.831	0.818		11	0.865	0.855		20	0.893	0.884
3	0.832	0.819		12	0.880	0.871		21	0.885	0.876
4	0.832	0.819		13	0.893	0.884		22	0.878	0.869
5	0.832	0.819		14	0.903	0.896		23	0.868	0.858
6	0.833	0.820		15	0.909	0.902		24	0.863	0.852
7	0.834	0.821		16	0.911	0.905		25	0.856	0.845
8	0.837	0.824		17	0.910	0.904		26	0.849	0.837
9	0.841	0.829		18	0.906	0.899		27	0.840	0.828

Table 2: Summary of standard tests of significance

	No Break			One Break (k = 16)			Two Break (k1 = 9, k2 = 14)		
Effect	Coef	t	p	Coef	t	p	Coef	t	p
Intercept (b0)	13192	47.91	<0.001	14026	53.16	<0.001	13050	57.16	<0.001
Gradient b1	183.58	11.45	<0.001	79.4	3.27	0.003	296.9	7.91	<0.001
Gradient b2				312.0	10.89	<0.001	-250.1	-5.27	<0.001
Gradient b3							330.8	20.14	<0.001
R^2	82.9%			91.1%			96.8%		

5. Assessing Significance of the Single Break Model

Standard tests of significance, such as those reported in Table 2, do not state whether the inclusion of a single breakpoint provides a real improvement in overall fit compared with a simple linear specification. Such an improvement would only be apparent if $\beta_1 \neq \beta_2$. Bootstrap procedures (Davidson and Hinkley, 1997) provide a means to validly answer this question.

A first stage in the bootstrap assessment of significance of a single break model is to fit a simple linear model and to obtain the predicted values ($\hat{y}_i$) and sample residuals (e_i). A new bootstrap data set adhering to a simple linear specification is obtained by sampling the residuals with replacement to create bootstrap residuals r_i and to form a new data set (x_i, $\hat{y}_i + r_i$), $i=1,\ldots,n$. This newly created sample is used to fit the single break model and a measure of overall model fit (R^2) is recorded. Repeating this process with B bootstrap samples produces an empiric distribution for which the corresponding observed sample statistic (R^2) may be judged against.

To assess the statistical significance of the one-break model, two versions of the bootstrap algorithm have been implemented. In the first case (Case A1) we consider the breakpoint to be fixed at $k=16$ and in the second case we consider the value of k to be determined by the data (Case A2). Algorithmically, for Case A1,
(1a)   generate a bootstrap sample using the simple linear (no break) model
(2a)   for the given bootstrap sample fit the one break model with $k=16$
(3a)   record the R^2

Steps (1a), (2a), (3a) are repeated B times. The proportion of times that the R^2 values from bootstrap samples exceed the corresponding value of R^2 for the sample data is recorded. This proportion gives a bootstrap estimate of the p-value for testing $H_0 : \beta_1 = \beta_2$ against $H_1 : \beta_1 \neq \beta_2$ assuming $k=16$.

In the second case, Case A2, the algorithm implemented has the form

(1b)   generate a bootstrap sample using the simple linear (no break) model

(2b)   fit the single break model using the given bootstrap sample; let k be bootstrap sample dependent and choose the value of k to be that value which maximises the goodness-of-fit, R^2

(3b)   record the R^2 value

Steps (1b), (2b), (3b) are repeated B times. The proportion of times that the R^2 values from bootstrap samples exceed the corresponding value of R^2 for the sample data is recorded. This proportion gives a bootstrap estimate for testing $H_0 : \beta_1 = \beta_2$ against $H_1 : \beta_1 \neq \beta_2$ without a pre-specification of k.

Table 2 summarises the fitted simple linear regression model used to generate the bootstrap samples and also summarises the single break model using k=16 (corresponding to 1996). All effects are statistically significant; the positive sign for the coefficients indicates that under this formulation the rate of change up to 1996 is positive (79.39 euros per capita per year) and this increases to 311.83 euros per capita per year. Application of Case A1, using B=5,000 bootstrap samples, provides an estimated p-value of 0.019 (97 of the 5,000 bootstrap samples had R^2 values that exceeded the observed sample R^2 (91.1%) statistic). Application of Case A2, using B=5,000 bootstrap samples, provides an estimated p-value of 0.146 (approximately fifteen percent of the bootstrap samples produced an R^2 value that exceeded the observed sample value of R^2=91.1%, when k was determined by maximizing the bootstrap sample of R^2).

6.   Applying the Two Break Model

Consider a broken stick model with a breaks at $X = k_1$ and at $X = k_2$ and with structural specification $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$. All possible pairs of values for k_1 and k_2 are considered and R^2 obtained. Figure 2 plots the values of R^2. The "top" five models, in terms of within sample goodness-

of-fit, are (k_1, k_2, R^2)=(10,11,98.1), (k_1, k_2, R^2)=(10,12,97.9), (k_1, k_2, R^2)=(10,13,97.4), (k_1, k_2, R^2)=(9,13,97.0) and (k_1, k_2, R^2)=(9,14,96.7). Taken at face value the differences in the R^2 values between these models seem quite minimal. From the possibilities listed a potential criticism of the first four models is the relatively small gap between breakpoints which may reflect local minima arising as chance random patterns in the data and on this basis the fifth model will tentatively be taken as a good two-break model.

Figure 2: Contour plot
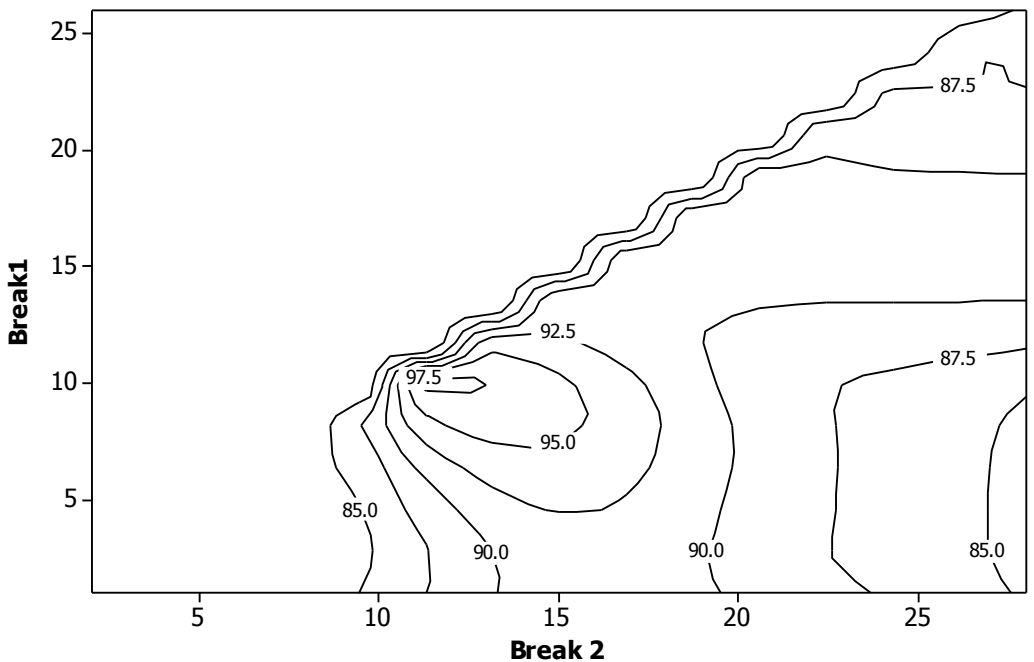
7.   Assessing Significance of the Double break model

The double break model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ reduces to the simple linear model when $\beta_1 = \beta_2 = \beta_3$. Bootstrapping may be used to test the null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3$ against H_1:"At least two values of β_j differ".

If break points can be considered to be pre-specified then the bootstrap algorithm for these hypotheses would be:

(1c)   generate a bootstrap sample using the simple linear (no break) model

(2c)   for the bootstrap sample fit the double break model with $k_1 = 9$ and $k_2 = 14$

(3c)   record the R^2 value

Steps (1c), (2c), (3c) are repeated B times. The proportion of times that the R^2 values from bootstrap samples exceed the corresponding value of R^2 is recorded. This proportion gives a bootstrap estimate for testing $H_0 : \beta_1 = \beta_2 = \beta_3$ against H_1:"At least two values of β_j differ".

The above bootstrap algorithms have been applied (B=5000 in both instances). The derived values of R^2 for the bootstrapped two break models never exceeded the sample R^2 value (96.7%) for the given two break model. Accordingly, the two-break model provides a statistically significant improvement over the simple linear (no break) specification irrespective of whether k_1 and k_2 are considered fixed with value 9 and 14 respectively or whether they are considered as values uncovered from the data under a position of ignorance.

Is the double break model an improvement on the single break model? The double break model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ reduces to the single break model when $\beta_1 = \beta_2$ or when $\beta_2 = \beta_3$. Bootstrapping may be used to test the null hypothesis $H_0 : \beta_1 = \beta_2$ or $\beta_2 = \beta_3$ against $H_1 : \beta_1 \neq \beta_2$ and

$\beta_2 \neq \beta_3$. In this set of bootstrap model the best fitting single break model will be used to generate the predicted values and residuals. A comparison of the given two break model (k_1=9; k_2=14) against similar two break bootstrap models with the null hypothesis specified by the best one break model (k_1=16) indicates that the given two break model is a statistically significant improvement over the one break model ($p<0.001$). Similarly a comparison of the given two break model (k_1=9; k_2=14) against the best possible two break model for each bootstrap sample with the null hypothesis specified by the best one break model (k_1=16) has been undertaken. In these cases 30 bootstrap samples out of 5000 bootstrap samples give a higher R^2 than the given sample R^2 value (96.7%), suggesting that even after allowing for chance effects from "over analysing the data" the given two break model may be deemed to be a significant improvement over the one break model (p=0.0006).

8.   Conclusion

The broken stick method has the property of retaining continuity and this may be advantageous when modelling a continuous process. Standard regression tests of significance do not directly assess whether gradient coefficients differ from one another, however these potential differences can be assessed using the bootstrap.

In some instances, prior reasoning or a theoretical position may suggest that a different rate of change in the dependent will occur at a pre-specified value of a predictor. In other cases, exploratory data mining may indicate a point for a different rate of change of a dependent variable with a predictor variable. The two possibilities (i.e. a pre-specified node, or a data discovered node) reflect different research positions but the bootstrap is sufficiently flexible to adjust for these two positions and permit the construction of valid tests of significance.

The method under investigation has been illustrated using output data across the EU. Prior to data collection there were no pre-conceived ideas concerning the number of breaks and location of breaks to model the data. Inspection of Figure 1 is highly suggestive of two structural breaks. The

bootstrap procedure provides evidence that a two break model is appropriate in this case even after allowing for the data discovery of these breaks.

References

Andrews, D.W.K. 1993. Tests for parameter instability and structural change with unknown change point. Econometrica 61(4), 821-856

Chow, G.C. 1960. Tests of equality between sets of coefficients in two linear regressions. Econometrica 28(3), 591-605.

Davidson, A.C. and D.V. Hinckley, 1997. Bootstrap methods and their applications. Cambridge University Press, Cambridge

Greene, C.A. 1999. Testing for a break at an unknown change-point: A test with known size in small samples. Economics Letters 64, 13-16

Hansen, B.E. 1992. Tests for parameter instability in regressions with I(1) processes. Journal of Business and Economic Statistics 10(3), 321-336

Inder, B. and Hao, K. 1996. A new test for structural change. Empirical Economics 21, 475-482