Scientific Research and Essays Vol. 7(27), pp. 2251-2254, 12 July, 2012
Available online at http://www.academicjournals.org/SRE
DOI: 10.5897/SRE11.828
ISSN 1992-2248 ©2012 Academic Journals

Full Length Research Paper

Evaluation of rate trend patterns via segmented regression: Case study of crime rates of Pakistan

Atif Akbar* and Aamna Khan

Department of Statistics, Bahauddin Zakariya University, Multan, Pakistan.

Accepted 7 June, 2012

In many practical regression-type problems, one uniform regression function to the data was found inappropriate, since the functional relationship between the response and the regressor(s) changes at certain points of regressors. These points are usually called break points or change points. In such situations, regression models are applied that are obtained by a piecewise definition of the regression function. Simply, by parts regression is used to detect an abrupt change of the response function at an increase or decrease of an influential factor, and named as segmented regression. Theoretical efforts have been made in the subject when parameters of a linear regression system obeying separate regimes. A case study concerning the crime rates in Pakistan is carried out by segmented regression approach as a comparative overview of linear regression approach and the segmented regression approach. Segmented regression approach was found to be more opposite than that of linear regression analysis on the basis of Bayesian information criteria and permutation test criteria.

Key words: Change point, permutation test, rate trends, segmented regression.

INTRODUCTION

A model is a judicious tradeoff between realism and simplicity. It is important to note that a valuable model is not one that is true or realistic but one that is parsimonious, plausible and informative. Researchers in many instances delineate and examine the models in detail that originate from entirely elaborated views. In view of the fact, that a model can act to endow with a helpful guess towards some of the significant core characteristics of the population which bring about the truth concerning the properties of real world phenomena but not the mathematical structures. If the class of model is incorrectly chosen, there may possibly not subsist as such a set of parameters on the way to even fit the data very well, as a typical data may not exactly fit the model that is being used, even when that model is correct. Thus, simulation practitioners recommend increasing the complexity of a model iteratively. But data modeling does not stop with finding the best-fit parameters.

Measured data is by and large subject to measurement errors. To fit a set of data to a straight line, linear regression could be used, which guarantees the best-fit model parameters based on the least square criterion. It is accustomed to perceive rigorously precise relations between several events of carefully considered requisites that have erratic values; particularly when the relation is speckled due to unsystematic disparity where the relationship is defined with the help of an exact mock-up. Thus, theorist possibly will use regression models in support of studying how changes in one or more explanatory variables will change the value of response variable. The choice of using linear or nonlinear regression that crop up as the two types of regression analysis depends upon the problem that is to be solved.

Linear regression analysis can be done either by means of the entire minutes or by carving up (segments) of data. The segments are introduced to see if there are abrupt changes in the relation under investigation. As in many practical regression-type problems, we cannot fit one uniform regression function to the data, since the functional relationship between the response and the regressor(s), changes at certain points of regressors. These points are usually called break points or change

*Corresponding author. E-mail: atifakber@yahoo.co.uk.

points. In such situations, regression models that are obtained by a piecewise definition of the regression function are applied. Simply, by parts regression is used to detect sudden change of the response function at an increase or decrease of an influential factor and that may be named as segmented regression, broken-line regression, piecewise regression or the multiphase regression with the continuity constraint (Bellman and Roth, 1969).

Thus, such a regression is a method in statistical regression analysis whereby the independent variable(s) are segmented (divided into sequential groups according to their value) and the regression analysis is performed separately for the segments: that is the regression within the segments is linear. The boundaries between the segments are perhaps called breakpoints, join points, knots or the transition points. The breakpoint can be taken as a critical or safe value beyond or below which undesired/desired effects occur. The resulting regression equations may show a discontinuity at the breakpoints. These breakpoints or the boundaries may be considered as a turning point, thus changing the linear phase of regression analysis into nonlinear. Seeing that, segments are introduced to witness the unexpected changes that possibly will turn up in our daily life and at the same time apt as a part of our personal experiences, such as the death of a loved one, or collective episodes, such as war or societal economic crisis. Breaks or jumps in the parameters that relate security returns to position variables possibly will come to pass due to a number of factors such as major changes in market sentiments, burst or creation of speculative bubbles, regime switches in a monetary and debt management policies (Pesaran and Timmermann, 2002). For example, to study; is the consumption pattern of the people of northern areas of Pakistan people today the same as it was before 2005 earthquake? Do the firms in the steel industry and the firms in the chemical industry have similar dividend policies, etc. Statistically these questions can be answered by testing whether two sets of observations can be regarded as belonging to the same regression model. Thus to deal with such situations one can easily go with segmented regression. Furthermore in different fields of sciences such as biology, economics, demography, etc., rate trend patterns usually occur and one may use segmented regression to describe rate trend patterns. Thus to study the mortality transition or the average annual rate of population growth, one make use of segmented regression and not only this one may take advantage of segmented regression for comparison purposes (Kim et al., 2004). Accordingly, for evaluation of these patterns, one may use different approaches to handle such situations that usually turn up in a scientific manner, in order to avoid complications and obstacles that are a hindrance just around the corner of finding solutions for different problems. Thus the segmented regression technique can be used to handle some of such situations in a fairly unusual style. Thus an assortment of efforts have been made for the exposure of structural breaks in different fields considering different conditions such as working under the assumption of known and unknown time points at which such turning points come into contact. The main consideration while proceeding with segmented regression is to consider the number of change points and their effects on the inferences (Stinson and Lubov, 1982).

Analysis of the source materials has been done that takes into account both the theoretical and empirical issues and provided details about the spur for the present work by describing what others have done. Quandt (1958) estimated the parameters of a linear regression system obeying at most two separate regimes. Hudson (1966) reported a method to find the overall least squares solution when a complete curve to be fit consists of two or more sub models. McGee and Carleton (1970) outlined an approach that uses hierarchical clustering to cluster the points into segments that represent the individual regimes of the piecewise function and perform standard linear regression on them. They combined the theory of regression analysis and cluster analysis to define a suitable method for finding the regression parameter estimates when the sampled data is supposed to be generated by more than a simple regression model.

Lerman (1980) proposed the grid search method to fit segmented line regression of rate trend patterns. Kim et al. (2000) worked out to describe cancer rates using the join point regression and proposed a permutation test to determine the number of significant join points. Hussain (2005) insinuated the breakpoints after which inflation is harmful to economic growth and estimated threshold level of inflation in Pakistan. Yu et al. (2007) extended the Hudson's continuous fitting method to multiple join point model and compared the computational efficiencies of the Lerman's grid search and Hudson's continuous fitting. Nagata et al. (2008) used segmented regression approach to estimate the effect of a new road traffic law against alcohol impaired driving in Japan. The result of their study gives an initiative that all traffic injuries declined significantly after the new traffic law.

MATERIALS AND METHODS

To evaluate the change point status, segmented regression approach was adopted in this article where, attention was disbursed to the three types of crime rates, that is, murder rate, dacoity rate and cattle theft rates in Pakistan for the year 1996 to 2006, by using the permutation test criteria and BIC along with the approach of McGee and Carleton (1970) for comparison purpose and also test for autocorrelation assuming constant variance.

As a phenomenological approach that we used in our study to commence, give good reasons for, and keep an eye on as well as make unequivocal, the assumptions underpinning our research design consisted of permutation tests that are the special kind of randomization test and is known under various names as approximate permutation test, Monte Carlo permutation test, and random permutation test. However, the all permutation tests are

theoretically same; and exist for any test statistic, regardless of whether or not its distribution is known. Permutation tests are, computationally intensive and are primarily used to provide a p-value. For testing between two different join point models, the permutation test is used repeatedly. A simpler model with fewer join points is called the null model, and a more complicated model called the alternative model. The alternative model fits better because it is more complicated (Kim et al., 2000).

The second method used is Bayesian information criterion (BIC), that is, a statistical criterion for model selection and is sometimes also named as Schwarz information criterion (Schwarz, 1978). The formulation for the BIC is:

$$BIC = -2\,ln\,L + k\,ln(n) \quad (1)$$

Under the assumption that the model errors or disturbances are normally distributed, it becomes:

$$BIC = n\,ln\,(RSS/n) + k\,ln(n) \quad (2)$$

where RSS is the residual sum of squares, from the estimated model.

We worked with BIC under the assumption that the model errors are normally distributed and while using BIC, we prefer the model out of all with lower value of BIC. In a change point analysis, a change point model allows different parts of a dataset to obey different probability laws, that is, all data received before some specified time *t* follows one model, while data received after time *t* follows the other model. A vital issue is determining the number of change points. The likelihood is not sufficient for this purpose because it will always prefer more change points. For this, we use Bayesian model selection by computing the probability of the data for each number of change points. We take up BIC to put side by side alternative models as information criteria penalize models with additional parameters.

Furthermore, employed the McGee and Carleton (1970) approach for the evaluation of rate trend patterns via segmented regression or simply to detect the discontinuities in the selected rates for study.

Thus using this approach of piecewise regression, our point of interest is to detect the P line segments in a data set, know where the P-1 breaks occurred in our single equation model and to discuss the various characteristics of intercept and slope of each line segment.

In our analysis, we used the join point regression 3.20 to analyze join point models that yields almost the same results for the three data sets as we get with the help of the above outlined method. We test whether or not there is a statistically significant change appearing in a trend. The program offers a set up in which one may select different options for the modeling of data and to specify the modeling methods. Further we also used *NLREG 6.40*, for model fitting.

RESULTS AND DISCUSSION

We contemplated the selected data sets while assuming no autocorrelation with constant variance and take 0.05 and 0.01 as the overall significance level for permutation test. The results of the different time points where the change occur in the data sets under consideration were broached, further the model statistics were made known and the estimated regression coefficients were revealed using the standard and general parameterization. We then select the model with the help of permutation test results and BIC. Further, the results of the approach suggested by McGee and Carleton (1970) are given for the same data sets for comparison.

Different models selected by three different approaches are given for each data set where $\hat{y}_p$ represent the selected model from permutation tests approach, $\hat{y}_m$ represent the selected model from McGee and Carleton (1970) approach, and $\hat{y}_b$ shows the result of BIC as favored model.

Working with the murder rate data, the two approaches, permutation tests approach with overall significance level 0.05 and 0.01 for both the uncorrelated and autocorrelated error models and McGee and Carleton (1970) approach produced the same results in order to check the existence of a change point in a selected data set. Thus their results goes in the favor of a model with no join point whereas; BIC preferred a model with 2 join points and the selected models are given by:

$$\hat{y}_p = -84636.318182+47.045455\,x \quad (3)$$

$$\hat{y}_m = -84636.318182+47.045455\,x \quad (4)$$

$$\hat{y}_b = -1068650.783146+539.886517\,x \quad (5)$$

$$\hat{y}_b = 1013085.621348-502.023596\,x \quad (6)$$

$$\hat{y}_b = -295795.919583+152.417175\,x \quad (7)$$

where for (5), *t* =1996 through 1998, for (6), *t* =1999 through 2000 and for (7), *t* =2001 through 2006.

For dacoity rates, results of two approaches goes in favor of single join point model and in case of uncorrelated error model and BIC preferred a model with 2 join points.

$$\hat{y}_p = -107.272727+0.727273\,x \text{ for } t = 1996 \text{ through } 2001 \quad (8)$$

$$\hat{y}_p = -585199.672727+293.127273\,x \text{ for } t = 2002 \text{ through } 2006 \quad (9)$$

$$\hat{y}_m = -116236.641+58.85716\,x \text{ for } t = 1996 \text{ through } 2003 \quad (10)$$

$$\hat{y}_m = -555851.91+278.501036x \text{ for } t = 2004 \text{ through } 2006 \quad (11)$$

$\hat{y}_b$ =-547359.764706+274.823529 *x* for *t* =1996 through 1997 (12)

$\hat{y}_b$ =78993.411765-38.823529 *x* for *t* =1998 through 2001 (13)

$\hat{y}_b$ =-607326.047059+304.164706 *x* for *t* =2002 through 2006 (14)

It however differs if we set out an autocorrelated error model with ρ = 0.9, 0.6, 0.8, etc. Under this environment, the selected model is the one with no join point.

$\hat{y}_p$ = -280133.605274+140.856683 *x* (15)

For cattle theft rates, the results of permutation tests approach with overall significance level 0.05 and 0.01 goes in favor of single join point model and in case of uncorrelated error model and McGee and Carleton (1970) approach, the results favor a linear model, that is, a zero join point model, whereas BIC preferred a model with 4 join points.

$\hat{y}_p$ = 137715.714286-65.714286x for *t* =1996 through 2003 (16)

$\hat{y}_p$ = -5020438.500000+2509.500000*x* for *t* =2004 through 2006 (17)

$\hat{y}_m$ = -1078862.25 + 542.972683*x* (18)

$\hat{y}_b$ =-3744115.241379+1878.551724*x* for *t* =1996 through 1997 (19)

$\hat{y}_b$ = 748466.796552-371.113793*x* for *t* =1998 through 2002 (20)

$\hat{y}_b$ = -2436563.327586+1219.810345*x* for *t* =2003 through 2004 (21)

$\hat{y}_b$ = -7902646.086207+3947.396552*x* for *t* =2004 through 2005 (22)

$\hat{y}_b$ =-2881331.000000+1443.000000*x* for *t* =2005 through 2006 (23)

Similarly, results differ if we set out an autocorrelated error model with ρ = 0.9, 0.6, 0.8, etc. Under this environment, the selected model is the one with no join point.

$\hat{y}_p$ =-955196.529170+481.092182*x* (24)

CONCLUSION AND RECOMMENDATION

In view of all discussed previously, the segmented regression provide a better fit weighing against the linear regression analysis to describe rate trend patterns. It is recommended that autocorrelation should be removed while implementing segmented regression, as the difference between linear regression and segmented regression narrows down with the increase in degree of autocorrelation for the data sets considered.

REFERENCES

Bellman R, Roth R (1969). Curve fitting by segmented straight lines. J. Am. Stat. Assoc., 64: 1079-1084.

Hudson D (1966). Fitting segmented curves whose join points have to be estimated. J. Am. Stat. Assoc., 61: 1097-1129.

Hussain M (2005). Inflation and growth: estimation of threshold point for Pakistan economic policy department. State Bank of Pakistan.

Kim HJ, Fay MP, Feuer EJ, Midthune, DN (2000). Permutation tests for joinpoint regression with applications to Cancer rates. Stat. Med., 19: 335-351.

Kim H, Fay, MP, Yu B, Barrett MJ, Feuer EJ (2004). Comparability of segmented line regression models. Biometrika, 60: 1005-1014.

Lerman PM (1980). Fitting segmented regression models by grid search. Appl. Stat., 29: 77-84.

McGee VE, Carleton WT (1970). Piecewise regression. J. Am. Stat. Assoc., 65: 1109-1124.

Nagata T, Setoguchi S, Hemenway D, Perry M J (2008). Effectiveness of a law to reduce alcohol impaired driving in Japan. BioMed. J., 14: 19-23.

Pesaran MH, Timmermann A (2002). A market timing and return prediction under model instability. J. Empir. Finan. 9: 495-510.

Quandt RE (1958). The estimation of parameters of a linear regression system obeying two separate regimes. J. Am. Stat. Assoc., 53: 873-880.

Schwarz G (1978). Estimating the dimension of a model. Ann. Stat., 6: 461-464.

Stinson TF, Lubov A (1982). Segmented regression, threshold effects, and police expenditures in small cities. Am. J. Agric. Eco., 64: 738-746.

Yu B, Barrett M, Kim HJ, Feuer EJ (2007). Estimating joinpoints in continuous time scale for multiple change-point models. Comp. Stat. Data Anal., 51: 2420-2427.