A segmentation approach to regression problems with switching-points

Shao-Tung Chang
Department of Mathematics
National Taiwan Normal University
Taipei, Taiwan
schang4809@gmail.com, schang@math.ntnu.edu.tw

Kang-Ping Lu[*]
Department of Applied Statistics
National Taichung University of Science and Technology
Taichung, Taiwan
kplu@nutc.edu.tw , kplu5208@gmail.com

***Abstract*—Regression problems with switching-points arise in many fields and has been recognized as a challenging issue for modern, big data applications. Many approaches to estimating switching regression models can only deal with continuous switching- point problem successfully. However, regression problems with jump discontinuities are often encountered in reality such as those in econometrics and engineering. This article presents a segmentation method for switching regression estimations, allowing for detecting both continuous and dis-continuous switching-points. We consider using Taylor's expansion with an adjustment constant to derive the estimates of switching-points and regression parameters simultaneously. The proposed method can detect both jump-points and continuous switching-points. The proposed method is evaluated via experiments with numerical examples. The simulation results show the proposed method work well for both continuous and dis-continuous models and produce rather accurate estimates.**

Keywords—switching-point, jump, switching regression, Taylor expansion

I. Introduction

Change-point problems have arisen in many substantive fields and change-point detection has been recognized as a challenging problem for modern, big data applications [1]. In many real regression problems, it occurs frequently that the effect of the covariates on the outcome changes abruptly at some places, which are called switching-points (SPs) or change-points. Locating change-points is an important issue in switching regressions because they are indications of the time and the way of changes in data pattern. This information is usually key to decision making. However, in classical regressions, the methods utilized for modelling the non-linear effect, such as spline regressions or non-parametric smoothing, often neglected this significance of switching-points and the resulted regression parameters are not directly interpretable. Thus, switching regression models have been advocated by researchers as better alternatives for regression problems with change-points.

Switching-point problems occur in many fields including, medicine climatology, ecology, and econometrics, e.g. [2], [3] and [4]. In medical science, switching-points occur as threshold values that the exposure has no impact on the response up to an unknown threshold. In econometrics, it is of interest to access switching regressions to discover the reasons of abrupt changes in the time series data.

Significant literature has been developed for switching-regressions since Quandt [5] introduced switching-regression models. Much work has focused on testing for switching-points, e.g. [6,7]. Later researchers have focused on estimations for switching regressions [2, 8, 9]. A major difficulty in detecting switching-points for regression models is the non-smoothness of the likelihood function with respect to the switching-point regarded as a parameter. Muggeo [2] adopted Taylor- expansions to bypass non-differentiability, allowing for simultaneous maximum likelihood (ML) inference on regression parameters and switching-points. Muggeo's segmented regression (SR) method allows for multiple switching-points but restricts to continuous regression lines. Muggeo's segmentation algorithm converges fast and produces adequate estimates. Furthermore, Muggeo [9] proposed the R package segmented for switching regressions so that Muggeo's segmented method has been employed in many real applications such as [10, 11,12]. However, the continuity assumption may not be suitable in a switching regression problem such as in economics or finance [13]. Our objective in this study is to advance the segmentation method to relax the continuity restriction on the regression lines.

II. Methodology

A. Muggeo's method

We first review Muggeo's [2] segmented regression (SR) method. Assume Y and Z are the response and the single predictor respectively. For simplicity, other predictors are omitted. The segmented model with K SPs can be denoted by

$$E(Y)=\alpha_0+\beta_0 z+\sum_k \beta_k (z-\psi_k)_+ \tag{1.1}$$

$\boldsymbol{\psi}=(\psi_1,\psi_2,\ldots,\psi_K)^T$ are switching-points and

$(z-\psi_k)_+=(z-\psi_k)\times I(z>\psi_k)$, $k=1,\ldots,K$.

By the first-order Taylor's expansion around $\psi_k^{(0)}$

$(z-\psi_k)_+ \doteq (z-\psi_k^{(0)})_+ +(\psi_k-\psi_k^{(0)})(-1)I(z>\psi_k^{(0)})$, $k=1,\ldots,K$.

Hence,

$$E(Y)\doteq\alpha_0+\beta_0 z+\sum_k \beta_k (z-\psi_k^{(0)})_+ +\sum_k \beta_k(\psi_k-\psi_k^{(0)})(-1)I(z>\psi_k^{(0)}) \tag{1.2}$$

$$E(Y)=\alpha_0+\beta_0 z+\sum_k \beta_k U_k^{(0)}+\sum_k \gamma_k V_k^{(0)} \tag{1.3}$$

where

$$U_k^{(s)}=(z-\psi_k^{(s)})_+,\ V_k^{(s)}=-I(z>\psi_k^{(s)}),\ \gamma_k^{(s)}=\beta_k^{(s)}(\psi_k-\psi_k^{(s)}) \tag{1.4}$$

In the $(s+1)$ th iteration,

$$\psi_k^{(s+1)}\doteq\psi_k^{(s)}+\frac{\gamma_k^{(s)}}{\beta_k^{(s)}},\ k=1,\ldots,K\ . \tag{1.5}$$

979-8-3503-1327-7/23/$31.00 ©2023 IEEE

Muggeo's segmented regression method allows for multiple switching-points detection but restricted to continuous regression lines.

B. Proposed method

As aforementioned equation (1.1), Muggeo (2003) proposed a segmented regressions models with K switching-points but the sub-models are limited to be continuous. For real data, the continuity assumption may not be simply assumed in a switching regression problem. For example, the regression problems in finance or econometrics often occur the shift in mean from one segment to the adjacent one. In such cases, Muggeo's method is not feasible. Because of the efficacy and the popularity of the SR method, we consider advancing SR to be available for discontinuous models.

Since jump-points occur at the junction between two adjoining segments, we modify Muggeo's equation (1.1) as follows.

$$E(Y)=\alpha_0+\beta_0 z+\sum_k \alpha_k I(z>\psi_k)+\sum_k \beta_k (z-\psi_k)_+ \qquad (1.6)$$

Based on Muggeo's method, we summarize equation (1.6) as

$$E(Y)=\alpha_0+\beta_0 z+\sum_k \beta_k U_k^{(0)}+\sum_k \gamma_k V_k^{(0)} \qquad (1.7)$$

For the $(s+1)$ iteration,

$$U_k^{(s)}=(z-\psi_k^{(s)})_+,\ V_k^{(s)}=-I(z>\psi_k^{(s)}),\ \gamma_k^{(s)}=\beta_k^{(s)}(\psi_k-\psi_k^{(s)})-\alpha_k^{(s)} \qquad (1.8)$$

Thus, the updated estimates for regression parameters can be derived via Taylor's expansion similarly.

$$\psi_k^{(s+1)}=\frac{\gamma_k^{(s)}+\alpha_k^{(s)}}{\beta_k^{(s)}}+\psi_k^{(s)} \qquad (1.9)$$

Note that the $(k+1)$-th model is as follows $E(y)=(\alpha_0+\cdots+\alpha_k-\beta_1\psi_1-\cdots-\beta_k\psi_k)+(\beta_0+\cdots+\beta_k)z$.

Hence,

$$\alpha_0^{(s+1)}=\bar{y}_0-\beta_0^{(s)}\bar{z}_0,\ \alpha_1^{(s+1)}=\bar{y}_1-(\beta_0^{(s)}+\beta_1^{(s)})\bar{z}_0-\alpha_0^{(s+1)},\ \cdots,$$
$$\alpha_k^{(s+1)}=\bar{y}_k-(\beta_0^{(s)}+\cdots+\beta_k^{(s)})\bar{z}_k-(\alpha_0^{(s+1)}+\cdots+\alpha_{k-1}^{(s+1)}) \qquad (1.10)$$

where, $\bar{y}_k$ and $\bar{z}_k$ are means of response and independent variables for data in the k-th model respectively. Moreover, another intuitive method for selecting the initial value of jump magnitude α is to find the place at which the difference in the regression mean between two consecutive groups formed by several nearby data points appears atypically larger than others such as the plots shown in Fig. 1. We compute the mean of every four consecutive data points and the difference in the mean between two successive groups such as the plots shown in Fig. 1. Fig. 1(b) shows an extremely large absolute value of jump when the dependent variable x is approximately equal to 7.53. Thus, the initial values can be set $\alpha^{(0)}=2.48$ and $\psi^{(0)}=7.53$ approximately.

III. Experiments

We show the effectiveness of the proposed method through some experiments with numerical examples. We consider three cases as shown in Fig. 2-6. The true model parameters and the sample size (n) considered for generating data are shown in each figure respectively. The error terms follow a normal distribution with a constant standard deviation 0.01. We denote the proposed method by DSR. Note, to simplify, we use SP to represent either the locations or the number of switching points.

(a) Generated data (b)Selecting the initial of α and φ

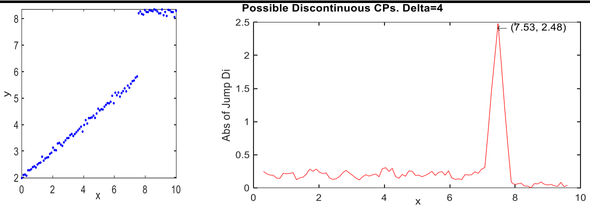
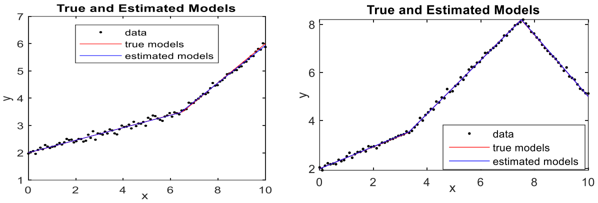


Fig. 1 Deciding the initial α and ψ

Case 1.1. n=100, SP=6.5,	**Case 1.2**. n=100, SPs:3.3, 7.5
$E(y)=\beta_{i0}+\beta_{i1}x$,	$(\beta_{10},\beta_{11})=(2,\ 0.455)$
$(\beta_{10},\beta_{11})=(2,\ 0.231)$	$(\beta_{20},\beta_{21})=(-0.193,1.119)$
$(\beta_{20},\beta_{21})=(-1.143,\ 0.714)$	$(\beta_{30},\beta_{31})=(17.8,-1.28)$
Same results by DSR, SR	Same results by DSR, SR



Estimates: $\text{SP}=6.4065$	Estimates: $\text{SP}=3.33,\ 7.51$
$(\hat{\beta}_{10},\hat{\beta}_{11})=(2.005,\ 0.229)$	$(\hat{\beta}_{10},\hat{\beta}_{11})=(1.996,0.461)$
$(\hat{\beta}_{20},\hat{\beta}_{21})=(-0.919,\ 0.686)$	$(\hat{\beta}_{20},\hat{\beta}_{21})=(-0.186,1.117)$
	$(\hat{\beta}_{30},\hat{\beta}_{31})=(17.864,-1.286)$

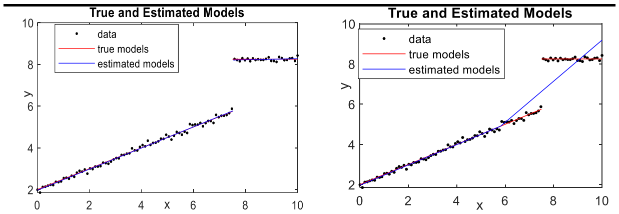
Fig. 2 Estimations for a continuous model with one or two SPs

Fig. 2 shows both methods (Muggeo's SR and the proposed DSR) perform well for continuous models with a single or multiple switching-points. Furthermore, the two methods result in the same estimates with high precision.

But, for discontinuous models such as those in Fig. 3-6, Muggeo's SR cannot work well in all discontinuous cases. On the contrary, the proposed DSR perform well and produce quite accurate estimates in all discontinuous models of a single or multiple jumps.

Case 2. n=100, SP=7.5, $E(y)=\beta_{i0}+\beta_{i1}x$, $(\beta_{10},\beta_{11})=(2,\ 0.5)$, $(\beta_{20},\beta_{21})=(8.25,\ 0)$

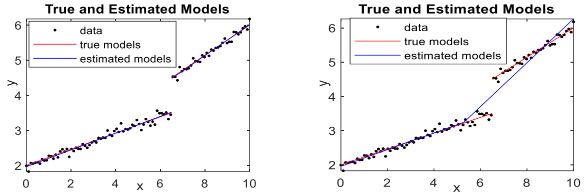
Proposed DSR	Muggeo's SR



Estimates: $SP = 7.5253$	Estimates: $SP = 5.8116$
$(\hat{\beta}_{10},\hat{\beta}_{11}) = (1.969, 0.506)$	$(\hat{\beta}_{10},\hat{\beta}_{11}) = (1.968, 0.505)$
$(\hat{\beta}_{20},\hat{\beta}_{21}) = (8.033, 0.024)$	$(\hat{\beta}_{20},\hat{\beta}_{21}) = (-1.032, 1.022)$

Fig. 3 Estimations for a model with 1 jump

Case 3. n=100, SP=6.5, $E(y)=\beta_{i0}+\beta_{i1}x$, $(\beta_{10},\beta_{11})=(2,\ 0.231)$, $(\beta_{20},\beta_{21})=(1.714,\ 0.429)$

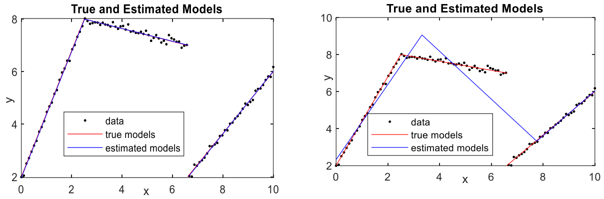


Proposed DSR	Muggeo's SR
Estimates: $SP = 6.52$	Estimates: $SP = 5.28$
$(\hat{\beta}_{10},\hat{\beta}_{11}) = (1.963, 0.239)$	$(\hat{\beta}_{10},\hat{\beta}_{11}) = (1.955, 0.244)$
$(\hat{\beta}_{20},\hat{\beta}_{21}) = (1.612, 0.440)$	$(\hat{\beta}_{20},\hat{\beta}_{21}) = (-0.142, 0.640)$

Fig. 4 Estimations for 1jump-point and changed beta

Case 4. n=100; SP: 2.5, 6.6; $E(y)=\beta_{i0}+\beta_{i1}x$

$(\beta_{10},\beta_{11})=(2,\ 2.4)$, $(\beta_{20},\beta_{21})=(8.6098,-0.2439)$
$(\beta_{30},\beta_{31})=(-5.7647, 1.1765)$

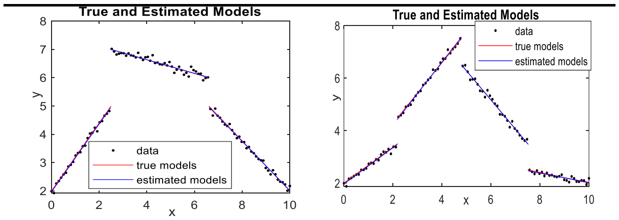


Proposed DSR	Muggeo's SR
Estimates: SP=2.51, 6.63	Estimates: SP=3.23, 7.79
$(\hat{\beta}_{10},\hat{\beta}_{11}) = (1.969, 2.399)$	$(\hat{\beta}_{10},\hat{\beta}_{11}) = (2.321, 2.031)$
$(\hat{\beta}_{20},\hat{\beta}_{21}) = (8.599, -0.241)$	$(\hat{\beta}_{20},\hat{\beta}_{21}) = (13.353, -1.294)$
$(\hat{\beta}_{30},\hat{\beta}_{31}) = (-5.882, 1.190)$	$(\hat{\beta}_{30},\hat{\beta}_{31}) = (-6.031, 1.206)$

Fig. 5 Estimations for a segmented model with 2 SPs

Fig. 3-5 shows that Muggero's SR can provide an estimated model but the estimations are unsatisfied. Furthermore, Muggeo's method cannot work totally for models with multiple jumps such as the cases in Fig. 6. On the contrary, the proposed method perform well and produce rather accurate estimates for both continuous and discontinuous models with a single or multiple switching-points.

Case 5.1. n=100,	**Case 5.2**. n=100,
SP: 2.5, 6.6,	SP: 2.2, 4.8, 7.5
$(\beta_{10},\beta_{11})=(2,\ 1.2)$	$(\beta_{10},\beta_{11})=(2, 0.682)$
$(\beta_{20},\beta_{21})=(7.610, -0.244)$	$(\beta_{20},\beta_{21})=(1.962, 1.154)$
$(\beta_{30},\beta_{31})=(10.824, -0.882)$	$(\beta_{30},\beta_{31})=(11.833, -1.111)$
	$(\beta_{40},\beta_{41})=(4, -0.2)$



Estimates: $SP = 2.47,\ 6.62$	$SP = 2.17,\ 4.80,\ 7.53$
$(\hat{\beta}_{10},\hat{\beta}_{11}) = (1.969, 1.200)$	$(\hat{\beta}_{10},\hat{\beta}_{11}) = (1.961, 0.691)$
$(\hat{\beta}_{20},\hat{\beta}_{21}) = (7.599, -0.241)$	$(\hat{\beta}_{20},\hat{\beta}_{21}) = (1.811, 1.198)$
$(\hat{\beta}_{30},\hat{\beta}_{31}) = (10.706, -0.869)$	$(\hat{\beta}_{30},\hat{\beta}_{31}) = (11.827, -1.111)$
	$(\hat{\beta}_{40},\hat{\beta}_{41}) = (3.783, -0.176)$

Fig. 6 Estimations for models of 2 and 3 jumps respectively by DSR

IV. Conclusion

Muggeo (2003) introduced a segmented regression method for solving linear regression problem with continuous switching-points. Muggeo's method is efficient and work well for continuous models and it can be carried out by using the segmented package in R. Thus, Muggeo's method is popular. However, in many real applications such as finance and industrial engineering, the continuity assumption may not be reasonable. This article presents a method to advance the feasibility of the segmented method to dis-continuous models. The experiments with numerical examples shown the efficacy of the method. The proposed method can deal with regression problems with dis-continuous jumps and connected switching-points simultaneously. Furthermore, the new method can provide quite accurate estimates.

Acknowledgment

This work was supported by National Science and Technology Council, Taiwan (Grant numbers NSTC 111-2118-M-003-002 and NSTC 111-2118-M-025-001)

References

[1] P. Fearnhead and G. Rigaill, "Changepoint Detection in the Presence of Outliers," Journal of the American Statistical Association, vol. 114, 2019, pp. 169-183.

[2] Muggeo, V.M.R. Estimating regression models with unknown breakpoints. Stat. Med. 22, 3055–3071, 2003.

[3] J. D. Toms and M.-A. Villard, "Threshold detection: matching statistical methodology to ecological questions and conservation planning objectives," Avian Conserv. Ecol., vol. 10(1): 2, 2015, http://dx.doi.org/10.5751/ACE-00715-100102.

[4] R. Werner, D. Valev, D. Danov, and V. Guineva, "Study of structural break points in global and hemispheric temperature series by piecewise regression," Adv. Space Res. 56, pp. 2323–2334, 2015.

[5] R.E. Quandt, "The estimation of the parameters of a linear regression system obeying two separate regimes," J. Am. Stat. Assoc., vol. 53, pp. 873–880, 1958.

[6] Kim, H.J., Siegmund, D.: The likelihood ratio test for a changepoint in simple linear regression. Biometrika 76, 409–423 1989.

[7] Andrews, D.W.K.: Tests for parameter instability and structural change with unknown change point. Econometrica 61, 821–856, 1993.

[8] Julious, S.A.: Inference and estimation in a changepoint regression problem. J. R. Stat. Soc. Ser. D, Stat. 50, 51–61, 2001.

[9] Muggeo, V.M.R.: Segmented: an R package to fit regression models with broken-line relationships. News. R Proj. 8(1), 20–25, 2008.

[10] Shaukat M.H., Alotaibi N., Hussain I., and Shrahili M. The Analysis of the Incidence Rate of the COVID-19 Pandemic Based on Segmented Regression for Kuwait and Saudi Arabia, Hindawi Mathematical Problems in Engineering, Volume 2021, Article ID 2644506, 11 pages, 2021, https://doi.org/10.1155/2021/2644506.

[11] Muggeo V.M.R. Sottile G., Porcu M. Modelling COVID-19 outbreak: segmented regression to assess lockdown effectiveness, Technical Report April, 2020. DOI: 10.13140/RG.2.2.32798.28485.

[12] Küchenhoff H., Günther F., Höhle M., Bender A. Analysis of the early Covid-19 epidemic curve in Germany by regression models with change points, medRxiv preprint, 2020, doi: https://doi.org/10.1101/2020.10.29.20222265.

[13] Chen C.W.S., Chan J.S., Gerlach K.R., & Hsieh W.Y.L. A comparison of estimators for regression models with change points. Stat. Comput., 21, 395-414, 2011.