Communications for Statistical Applications and Methods
2019, Vol. 26, No. 6, 539–556

https://doi.org/10.29220/CSAM.2019.26.6.539
Print ISSN 2287-7843 / Online ISSN 2383-4757

On study for change point regression problems using a difference-based regression model

Jong Suk Park[a], Chun Gun Park[1,b], Kyeong Eun Lee[a]

[a]Department of Statistics, Kyungpook National University, Korea;
[b]Department of Mathematics, Kyonggi University, Korea

Abstract

This paper derive a method to solve change point regression problems via a process for obtaining consequential results using properties of a difference-based intercept estimator first introduced by Park and Kim (*Communications in Statistics - Theory Methods*, 2019) for outlier detection in multiple linear regression models. We describe the statistical properties of the difference-based regression model in a piecewise simple linear regression model and then propose an efficient algorithm for change point detection. We illustrate the merits of our proposed method in the light of comparison with several existing methods under simulation studies and real data analysis. This methodology is quite valuable, "no matter what regression lines" and "no matter what the number of change points".

Keywords: change point, difference-based intercept estimator, difference-based regression model, piecewise linear regression

1. Introduction

A piecewise linear regression model is a special type of relationship between a dependent variable and one (or more) explanatory variables that consists of piecewise lines. In this model, the matter of interest is the boundary points of the adjacent lines called break points, change points or jointpoints (Muggeo, 2008). Change point problems occur in fields such as molecular biology, machine learning, and econometrics. The results derived from statistical inference could be misleading if there exist change points; consequently, it is important to ascertain the threshold values where the effect of the independent variable changes (Ulm, 1991; Betts *et al.*, 2007; Muggeo, 2008).

Significant literature has been developed in many fields related to the change point regression problems since the 1950's (Quandt, 1958). Many studies have focused on testing for change points rather than estimates since then (Kim and Siegmund, 1989; Andrews, 1993; Andrews and Ploberger, 1994). However, later statisticians have focused on estimates (Loader, 1996; Bai, 1997; Julious, 2001; Muggeo, 2003; Zhou and Liang, 2008). Subsequently, related software programs on detecting change points have been developed. Some R packages related to these problems include strucchange (Zeileis *et al.*, 2002), segmented (Muggeo, 2008), and chngpt (Fong *et al.*, 2017), which have worked only under certain conditions of a continuous or discontinuous type, a single covariate or multiple covariates, and no change point, one change point or multiple change points. For example, the R package segmented (Muggeo, 2003, 2008) supports the continuous type of a mean

[1] Corresponding author: Department of Mathematics, Kyonggi University, 154-42, Gwanggyosan-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do 16227, Korea. E-mail: cgpark@kgu.ac.kr

Published 30 November 2019 / journal homepage: http://csam.or.kr

function that is a connected line at the change point, and allows multiple change points. The chngpt package proposed by Fong *et al.* (2017) also supports both continuous and discontinuous types in logistic regression models as well as linear regression models and allows two-phase regression models. Most of all R packages for change point detection support only one covariate, except for the segmented package which is available for several covariates under strong conditions (https://cran.r-project.org/web/packages/segmented/segmented.pdf).

Park and Kim (2019) first introduced a difference-based regression model (DBRM) which is useful for outlier detection in multiple linear regression models. The proposed outlier detection approach uses a difference-based intercept estimator that is influenced by anomalous data. Based on the DBRM we can apply properties of this estimator to the piecewise regression model. In particular, we propose an efficient algorithm for change point detection in a piecewise simple linear regression model (PSLR). Compared to the previously mentioned methods, our proposed method has advantages that can be applied to various circumstances: continuous or discontinuous types, single covariate or multiple covariates, and no change point, one change point or multiple change points.

The remainder of this paper is organized as follows. In Section 2, we briefly describe piecewise linear regression models. In Section 3, we utilize the process of the DBRM (Park and Kim, 2019) and derive the statistical properties of the difference-based coefficient estimator in the PSLR. An algorithm to detect change points will then be given in Section 4. In Section 5 and Section 6, we illustrate the merits of our proposed method in comparison with several existing methods by simulation studies and real data analysis. The article concludes with a discussion in Section 7.

2. Piecewise linear regression model

In this section, we consider a piecewise regression model. To do this, we use a multiple change point regression model described by Chen *et al.* (2011). Assume that the number of regimes is $r+1$ with r change points, response variable is the $\mathbf{y} = (y_1, y_2, \ldots, y_N)'$, and the number of regressors is p, where change points are defined in terms of only one of the regressors, x. Other regressors are expressed as $z_1, \ldots, z_{p-1}$. Then general piecewise linear regression model is:

$$y_i = \begin{cases} \alpha_{1,1} + \beta_{1,1} x_i + \sum_{l=2}^{p} \beta_l z_{i,l-1} + \epsilon_i, & \text{if } i \le n_1, \\ \alpha_{1,2} + \beta_{1,2} x_i + \sum_{l=2}^{p} \beta_l z_{i,l-1} + \epsilon_i, & \text{if } n_1 < i \le n_1 + n_2, \\ \quad \vdots & \quad \vdots \\ \alpha_{1,m} + \beta_{1,m} x_i + \sum_{l=2}^{p} \beta_l z_{i,l-1} + \epsilon_i, & \text{if } \sum_{l=1}^{m-1} n_l < i \le \sum_{l=1}^{m} n_l, \\ \quad \vdots & \quad \vdots \\ \alpha_{1,r+1} + \beta_{1,r+1} x_i + \sum_{l=2}^{p} \beta_l z_{i,l-1} + \epsilon_i, & \text{if } \sum_{l=1}^{r} n_l < i \le N. \end{cases} \tag{2.1}$$

Here, N is the total number of observations. As for $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_r)'$ is the $r \times 1$ vector, and its parameters are change point parameters which satisfy $\delta_1 < \delta_2 < \cdots < \delta_r$. n_1 is the number of observations in the first regime ($x_i \le \delta_1$), n_m is the number of observations in the m^{th} regime ($\delta_{m-1} < x_i \le \delta_m$) for $m = 2, \ldots, r$, and $n_{r+1} = N - \sum_{m=1}^{r} n_m$.

A piecewise linear regression model is classified by Hawkins (1980) as continuous and discontinuous. Continuous type here means that the regression function is a connected line at the change point, δ_m that satisfies the following equation: $\alpha_{1,m} + \beta_{1,m}\delta_m = \alpha_{1,m+1} + \beta_{1,m+1}\delta_m$ for $m = 1, 2, \ldots, r$. If this is not satisfied, the model is discontinuous.

Below, we use Equation (2.1) with $r = 1$ and $p = 1$, assuming that change point location, δ is unknown and this model has two regimes. Without loss of generality, assume that $n = \#(x_i \le \delta)$ (i.e.,

$x_1 \leq \cdots \leq x_n < \delta \leq x_{n+1} \leq \cdots \leq x_N$). Then the formula can be expressed as:

$$y_i = \begin{cases} \alpha_1 + \beta_1 x_i + \epsilon_i, & \text{if } i \leq n, \\ \alpha_2 + \beta_2 x_i + \epsilon_i, & \text{if } i > n, \end{cases} \tag{2.2}$$

where α_1, α_2, β_1, and β_2 are unknown parameters and $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. For these two phase linear models, Fong *et al.* (2017) distinguished them into step, hinge, segmented, and stegmented. We now address change point problems using the Model (2.2), and use the four types to illustrate.

3. Difference-based regression model for PSLR

3.1. Fitting PSLR using difference-based regression model

In this section, after introducing the DBRM (Park and Kim, 2019), we explain how to apply the process of the DBRM to the PSLR. Park and Kim (2019) proposed an outlier-detection approach using the properties of an intercept estimator in the DBRM. This method uses only the estimator of the intercept: it does not require estimating the other parameters in the DBRM. In this paper, we first use the DBRM process to detect change points and explain how to apply DBRM to the PSLR.

First, we describe the DBRM without change points. Then, the simple linear regression can be expressed as: $y_i = \alpha + \beta x_i + \epsilon_i$, $i = 1, \ldots, N$. Let $y_{(k)i}$ be the difference between y_i and y_k and $x_{(k)i}$ be the difference between x_i and x_k, where $k = u, \ldots, N$. Here, u is the minimum number of observations to estimate parameters, and we set $u = 5$. Also, we assume that $i = 1, \ldots, k-1$ due to the effect of the change point. The DBRM for the k^{th} observation is written as

$$y_{(k)i}|\epsilon_k = -\epsilon_k + \beta x_{(k)i} + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right), \tag{3.1}$$

where β is the slope of regression line.

Second, we describe the DBRM with change point using Equation (2.2). This model is divided into two parts according to the value of k affected by the location of the change point. For the first part ($k \leq n$), the DBRM for the k^{th} observation is

$$y_{(k)i}|\epsilon_k = -\epsilon_k + \beta_1 x_{(k)i} + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right), \tag{3.2}$$

where β_1 is the slope of the first regression line. For the second part ($k > n$), the DBRM for the k^{th} observation is

$$y_{(k)i}|\epsilon_k = -\epsilon_k + \alpha_1 - \alpha_2 + (\beta_1 - \beta_2)x_k + \beta_1 x_{(k)i} + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right), \tag{3.3}$$

where $\alpha_1, \alpha_2, \beta_1$, and β_2 are the coefficients of each regression line in Equation (2.2). For $k = u, \ldots, N$, Model (3.1), (3.2), and (3.3) can be expressed as a simple linear regression model:

$$y_{(k)i}|\epsilon_k = \alpha^{(k)} + \beta^{(k)} x_{(k)i} + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right), \tag{3.4}$$

where $i = 1, \ldots, k-1$, and $\alpha^{(k)}$ and $\beta^{(k)}$ are regression coefficients of these models for the k^{th} observation. In simple notation, the matrix form is as follows.

$$\begin{pmatrix} y_1 - y_k \\ y_2 - y_k \\ \vdots \\ y_{k-1} - y_k \end{pmatrix} = \begin{pmatrix} 1 & (x_1 - x_k) \\ 1 & (x_2 - x_k) \\ \vdots & \vdots \\ 1 & (x_{k-1} - x_k) \end{pmatrix} \begin{pmatrix} \alpha^{(k)} \\ \beta^{(k)} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{k-1} \end{pmatrix}.$$

Then we estimate the k^{th} coefficients, $\alpha^{(k)}$ and $\beta^{(k)}$ in the DBRM (3.4) using the least square estimators. The result of the DBRM without change point leads to the estimators as the following two parts: intercept estimator $\hat{\alpha}^{(k)}|\epsilon_k$ and slope estimator $\hat{\beta}^{(k)}|\epsilon_k$.

$$\hat{\alpha}^{(k)}|\epsilon_k = -\left(y_k - \bar{y}_{(k)}\right) + \left(x_k - \bar{x}_{(k)}\right)\left(\hat{\beta}^{(k)}|\epsilon_k\right), \tag{3.5a}$$

$$\hat{\beta}^{(k)}|\epsilon_k = \frac{\sum_{j=1}^{k-1}\left(x_j - \bar{x}_{(k)}\right)\left(y_j - \bar{y}_{(k)}\right)}{\sum_{j=1}^{k-1}\left(x_j - \bar{x}_{(k)}\right)^2}, \tag{3.5b}$$

where $\bar{x}_{(k)} = \sum_{j=1}^{k-1} x_j/(k-1)$ and $\bar{y}_{(k)} = \sum_{j=1}^{k-1} y_j/(k-1)$. The result of the DBRM with change point is divided into two parts based on where k is located: when $k = n$, n is the number of observations in the first regime. In this case, $\hat{\alpha}^{(n)}|\epsilon_n$ and $\hat{\beta}^{(n)}|\epsilon_n$ are expressed as

$$\hat{\alpha}^{(n)}|\epsilon_n = -(y_n - \bar{y}_{(n)}) + (x_n - \bar{x}_{(n)})\left(\hat{\beta}^{(n)}|\epsilon_n\right), \tag{3.6a}$$

$$\hat{\beta}^{(n)}|\epsilon_n = \frac{\sum_{j=1}^{n-1}\left(x_j - \bar{x}_{(n)}\right)\left(y_j - \bar{y}_{(n)}\right)}{\sum_{j=1}^{n-1}\left(x_j - \bar{x}_{(n)}\right)^2}, \tag{3.6b}$$

where $\bar{x}_{(n)} = \sum_{j=1}^{n-1} x_j/(n-1)$ and $\bar{y}_{(n)} = \sum_{j=1}^{n-1} y_j/(n-1)$. In the case of $k = n+1$, the expression of $\hat{\alpha}^{(n+1)}|\epsilon_{n+1}$ and $\hat{\beta}^{(n+1)}|\epsilon_{n+1}$ using Equation (3.6) is noted as

$$\hat{\alpha}^{(n+1)}|\epsilon_{n+1} = (n-1)\mathcal{R}_{n+1}\left(\hat{\alpha}^{(n)}|\epsilon_n\right) - (y_{n+1} - y_n) + (x_{n+1} - x_n)\left(\hat{\beta}^{(n)}|\epsilon_n\right), \tag{3.7a}$$

$$\hat{\beta}^{(n+1)}|\epsilon_{n+1} = \mathcal{P}_{n+1}\left(\hat{\beta}^{(n)}|\epsilon_n\right) + \mathcal{Q}_{n+1}\left(y_n - \bar{y}_{(n)}\right), \tag{3.7b}$$

where

$$\mathcal{R}_{n+1} = \frac{\sum_{j=1}^{n-1}\left(x_j - \bar{x}_{(n)}\right)^2 - (x_{n+1} - x_n)(x_n - \bar{x}_{(n)})}{n\sum_{j=1}^{n-1}\left(x_j - \bar{x}_{(n)}\right)^2 + (n-1)(x_n - \bar{x}_{(n)})^2},$$

$$\mathcal{P}_{n+1} = \frac{n\sum_{j=1}^{n-1}\left(x_j - \bar{x}_{(n)}\right)^2}{n\sum_{j=1}^{n-1}\left(x_j - \bar{x}_{(n)}\right)^2 + (n-1)\left(x_n - \bar{x}_{(n)}\right)^2},$$

$$\mathcal{Q}_{n+1} = \frac{(n-1)(x_n - \bar{x}_{(n)})}{n\sum_{j=1}^{n-1}\left(x_j - \bar{x}_{(n)}\right)^2 + (n-1)(x_n - \bar{x}_{(n)})^2}.$$

3.2. Statistical properties of coefficient estimators

For $k = u, \ldots, N$, we provide mean and variance of the k^{th} coefficient estimators, $\hat{\alpha}^{(k)}|\epsilon_k$ and $\hat{\beta}^{(k)}|\epsilon_k$ in the following two parts: no change point and one change point.

- **No change point**: In the case of Model (3.1) without change point, the mean and variance of coefficient estimators, $\hat{\alpha}^{(k)}|\epsilon_k$ and $\hat{\beta}^{(k)}|\epsilon_k$ for the k^{th} observation are expressed as

$$E\left(\hat{\alpha}^{(k)}\right) = E\left(E\left(\hat{\alpha}^{(k)}|\epsilon_k\right)\right) = E\left(-\epsilon_k\right) = 0, \tag{3.8a}$$

$$\mathrm{Var}\left(\hat{\alpha}^{(k)}\right) = E\left(V\left(\hat{\alpha}^{(k)}|\epsilon_k\right)\right) + V\left(E\left(\hat{\alpha}^{(k)}|\epsilon_k\right)\right) = \frac{\sum_{j=1}^{k-1}\left(x_j - x_k\right)^2}{(k-1)\sum_{j=1}^{k-1}\left(x_j - \bar{x}_{(k)}\right)^2}\sigma^2 + \sigma^2, \tag{3.8b}$$

$$E\left(\hat{\beta}^{(k)}\right) = E\left(E\left(\hat{\beta}^{(k)}|\epsilon_k\right)\right) = \beta, \tag{3.8c}$$

$$\mathrm{Var}\left(\hat{\beta}^{(k)}\right) = E\left(V\left(\hat{\beta}^{(k)}|\epsilon_k\right)\right) + V\left(E\left(\hat{\beta}^{(k)}|\epsilon_k\right)\right) = \frac{\sigma^2}{\sum_{j=1}^{k-1}\left(x_j - \bar{x}_{(k)}\right)^2}, \tag{3.8d}$$

where $\bar{x}_{(k)} = \sum_{j=1}^{k-1} x_j/(k-1)$ and β is the coefficient of regression line.

- **One change point**: When $k = n$ of Model (3.2) and $k = n+1$ of Model (3.3), respectively, the mean and variance are expressed as following two parts. If $k = n$, there are

$$E\left(\hat{\alpha}^{(n)}\right) = E\left(E\left(\hat{\alpha}^{(n)}|\epsilon_n\right)\right) = E\left(-\epsilon_n\right) = 0, \tag{3.9a}$$

$$\mathrm{Var}\left(\hat{\alpha}^{(n)}\right) = E\left(V\left(\hat{\alpha}^{(n)}|\epsilon_n\right)\right) + V\left(E\left(\hat{\alpha}^{(n)}|\epsilon_n\right)\right) = \frac{\sum_{j=1}^{n-1}\left(x_j - x_n\right)^2}{(n-1)\sum_{j=1}^{n-1}\left(x_j - \bar{x}_{(n)}\right)^2}\sigma^2 + \sigma^2, \tag{3.9b}$$

$$E\left(\hat{\beta}^{(n)}\right) = E\left(E\left(\hat{\beta}^{(n)}|\epsilon_n\right)\right) = \beta_1, \tag{3.9c}$$

$$\mathrm{Var}\left(\hat{\beta}^{(n)}\right) = E\left(V\left(\hat{\beta}^{(n)}|\epsilon_n\right)\right) + V\left(E\left(\hat{\beta}^{(n)}|\epsilon_n\right)\right) = \frac{\sigma^2}{\sum_{j=1}^{n-1}\left(x_j - \bar{x}_{(n)}\right)^2}, \tag{3.9d}$$

where $\bar{x}_{(n)} = \sum_{j=1}^{n-1} x_j/(n-1)$ and β_1 is the coefficient of the first regression line in Equation (3.2). If $k = n+1$, there are

$$E\left(\hat{\alpha}^{(n+1)}\right) = E\left(E\left(\hat{\alpha}^{(n+1)}|\epsilon_{n+1}\right)\right) = E\left(\lambda_{n+1} - \epsilon_{n+1}\right) = \lambda_{n+1}, \tag{3.10a}$$

$$\mathrm{Var}\left(\hat{\alpha}^{(n+1)}\right) = E\left(V\left(\hat{\beta}^{(n+1)}|\epsilon_{n+1}\right)\right) + V\left(E\left(\hat{\beta}^{(n+1)}|\epsilon_{n+1}\right)\right) = \frac{\sum_{j=1}^{n}\left(x_j - x_{n+1}\right)^2}{n\sum_{j=1}^{n}\left(x_j - \bar{x}_{(n+1)}\right)^2}\sigma^2 + \sigma^2, \tag{3.10b}$$

$$E\left(\hat{\beta}^{(n+1)}\right) = E\left(E\left(\hat{\beta}^{(n+1)}|\epsilon_{n+1}\right)\right) = \beta_1, \tag{3.10c}$$

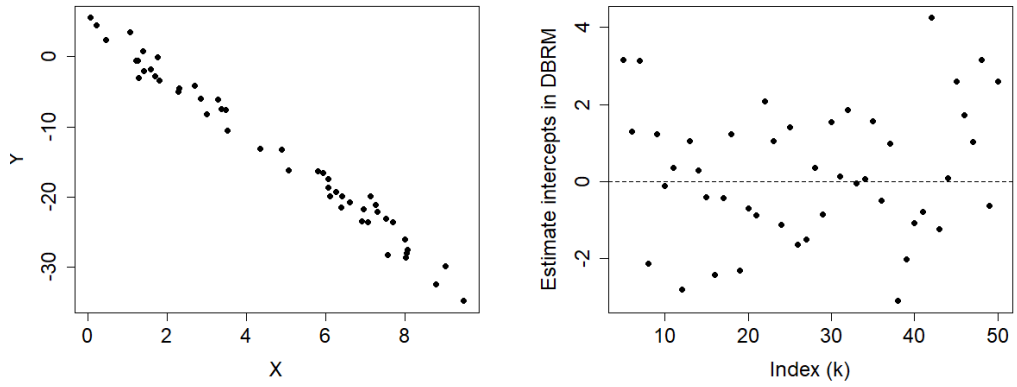
$$\mathrm{Var}\left(\hat{\beta}^{(n+1)}\right) = E\left(V\left(\hat{\beta}^{(n+1)}|\epsilon_{n+1}\right)\right) + V\left(E\left(\hat{\beta}^{(n+1)}|\epsilon_{n+1}\right)\right) = \frac{\sigma^2}{\sum_{j=1}^{n}\left(x_j - \bar{x}_{(n+1)}\right)^2}, \tag{3.10d}$$

where $\lambda_{n+1} = \alpha_1 - \alpha_2 + (\beta_1 - \beta_2)x_{n+1}$. For $k = n+2, n+3, \ldots, N$, the difference based observations aren't of the linear relationship between $x_{(k)i}$ and $y_{(k)i}$, $i = 1, 2, \ldots, k-1$ with the first regime, (x_j, y_j), $j = 1, 2, \ldots, n$, so $E(\hat{\alpha}^{(n+2)}) \neq \lambda_{n+2} - \epsilon_{n+2}$.

In accordance with Equation (3.8), the mean of the intercept estimator is zero. If $k \leq n$, results are also similar to the previous result. However, in accordance with Equation (3.10), the mean of the intercept estimator is non-zero value, λ_{n+1}. Prior to the change point, the mean of the intercept estimator is zero, but beyond the change point, the mean of the intercept estimator is non-zero. The estimated intercept is highly influenced by which observation is removed and then by whether the removed observation is the change point (or not). Hence, we use the difference-based intercept estimator for change point detection.

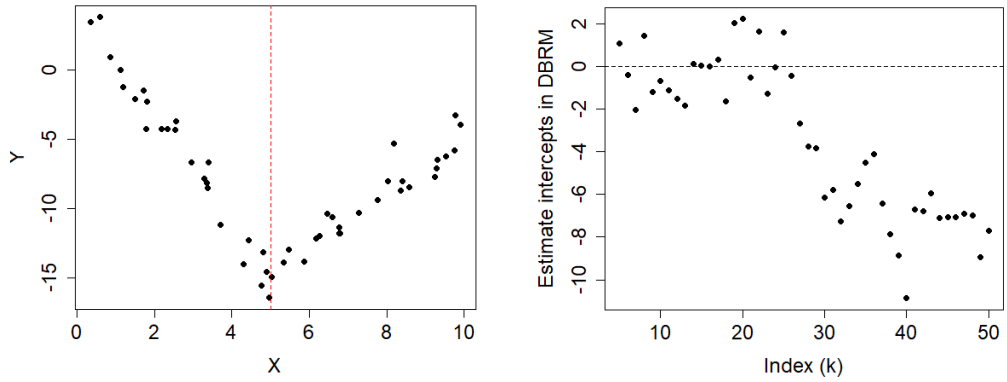
4. Hypothesis testing for change point detection

In this section, we explain characteristics of an intercept estimator, $\hat{\alpha}^{(k)}|\epsilon_k$, $k = u, \ldots, N$, and propose a testing method for change point detection using properties of the intercept estimator. In order to detect locations of change points, we also further develop algorithm using the properties of $\hat{\alpha}^{(k)}|\epsilon_k$.



(a) Scatter plot (b) Estimated intercepts

Figure 1: *Intercept estimates in the DBRM without change point: (a) scatter plot of* $\mathbf{x}$ *and* $\mathbf{y}$*; (b) estimated intercepts in the DBRM without the* k^{th} *observation,* $k = 5, \ldots, N$*. DBRM = difference-based regression model.*



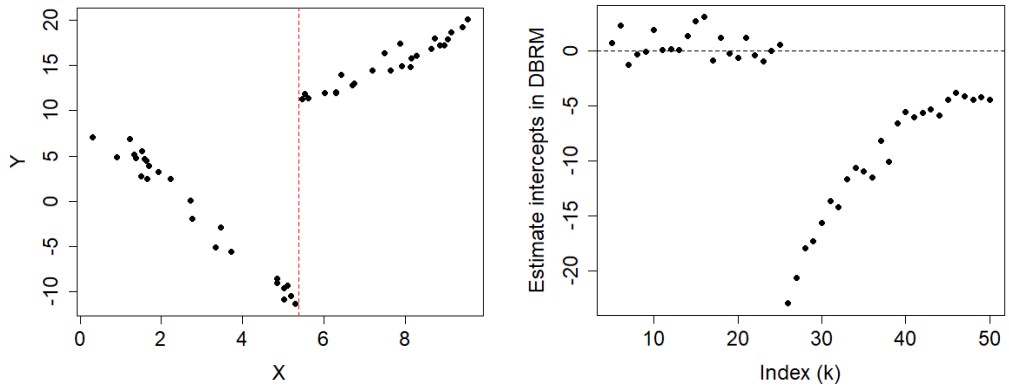
(a) Scatter plot (b) Estimated intercepts

Figure 2: *Intercept estimates in the DBRM with one change point (continuous type) with true number of observations in the first regime* $(n) = 25$*: (a) scatter plot of* $\mathbf{x}$ *and* $\mathbf{y}$*; (b) estimated intercepts in the DBRM without the* k^{th} *observation,* $k = 5, \ldots, N$*. DBRM = difference-based regression model.*

4.1. Characteristics of intercept estimator

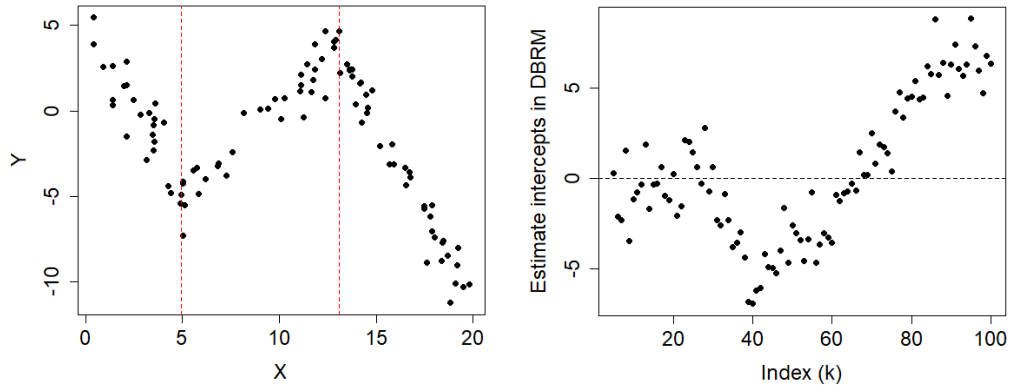
The difference-based intercept estimator is highly affected by the change point. We now explain the characteristics of the intercept estimator in three cases: no change point, one change point, and two change points.

- **No change point**: Assume a simple linear regression model ($y = -4x + 5$) without change point. In Figure 1(b), we display the scatter plot of the k^{th} observation and the intercept estimate without the k^{th} observation with $k = u, \ldots, N$. As a result, estimated $\hat{\alpha}^{(k)}|\epsilon_k$'s are random at zero.
- **One change point**: Assume a piecewise linear regression model with one change point. This model also has two types of continuous or discontinuous. If the model is a continuous type, $\hat{\alpha}^{(k)}|\epsilon_k$'s appear randomly at zero prior to the true change point. But, $\hat{\alpha}^{(k)}|\epsilon_k$'s have the same sign values and tend to increase or decrease beyond the true change point in Figure 2(b). However, if the model is



(a) Scatter plot (b) Estimated intercepts

Figure 3: *Intercept estimates in the DBRM with one change point (discontinuous type) with true number of observations in the first regime* $(n) = 25$*: (a) scatter plot of* $\mathbf{x}$ *and* $\mathbf{y}$*; (b) estimated intercepts in the DBRM without the* k^{th} *observation,* $k = 5, \ldots, N$*. DBRM = difference-based regression model.*



(a) Scatter plot (b) Estimated intercepts

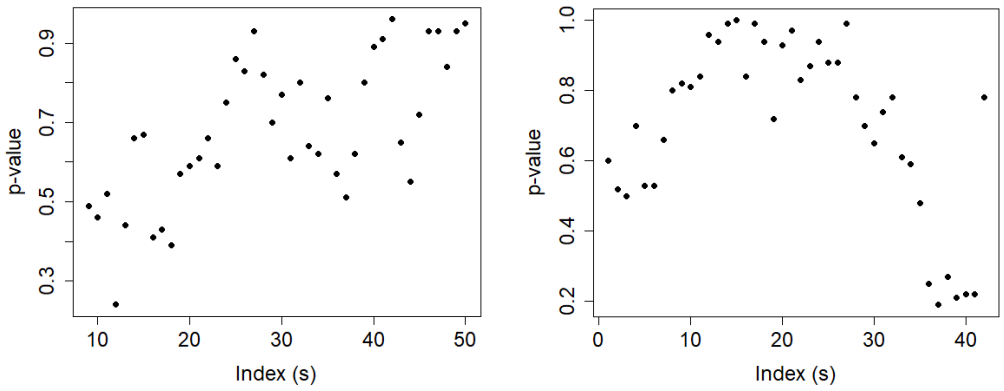
Figure 4: *Intercept estimates in the DBRM with two change points with true number of observations in the first regime* $(n) = 25$*, and in the second regime* $(n_2) = 35$*: (a) scatter plot of* $\mathbf{x}$ *and* $\mathbf{y}$*; (b) estimated intercepts in the DBRM without the* k^{th} *observation,* $k = 5, \ldots, N$*. DBRM = difference-based regression model.*

discontinuous type, the estimated intercepts have the highest or lowest point beyond the true change point in Figure 3(b).

- **Two change points**: Assume a continuous piecewise linear regression model with two change points. As shown above, $\hat{\alpha}^{(k)}|\epsilon_k$'s appear randomly at zero prior to the first change point. But passing each change point, the intercept estimates have different patterns in Figure 4(b).

4.2. A testing method for change point detection

In most cases, a simple graphical analysis is able to detect a change point, but in other cases a hypothesis test is required. In accordance with Section 3.2, if there is no change point, intercept estimators, $\hat{\alpha}^{(k)}|\epsilon_k$'s are random at zero. However, if there are change points, intercept estimators have patterns



(a) Randomness test of ascending data (b) Randomness test of descending data

Figure 5: *Bartels test of intercept estimators without change point.*

beyond the change point. Accordingly, we detect change points using these characteristics. In order to do so, we consider rank test (Bartels, 1982) which perform better than an ordinary run test in our simulation.

Bartels (1982) considered the rank version of the von Neumann's ratio statistic and obtained the critical values of this statistic under the randomness hypothesis. Suppose R_t is the rank of the t^{th} observation in T observations. The null hypothesis H_0 of randomness is rejected at large in absolute value of the test statistic

$$\text{Bar}^* = \frac{\text{Bar} - E[\text{Bar}]}{\sqrt{D[\text{Bar}]}} = \frac{\text{Bar} - 2}{2\sqrt{5/(5T+7)}}, \tag{4.1}$$

where $\text{Bar} = \sum_{t=1}^{T-1}(R_t - R_{t+1})^2 / \sum_{t=1}^{T}(R_t - \bar{R})^2$. It is known that Bar^* is asymptotically standard normal distributed under H_0. In this paper, we use the rank, the corresponding sequential number of $\{\hat{\alpha}^{(u)}, \hat{\alpha}^{(u+1)}, \ldots, \hat{\alpha}^{(s)}\}$ with $s = u + v - 1, \ldots, N$. Here, v is the minimum number of observations for the test, and we set $v = 5$. For this test, we use R package randtests (Mateus and Caeiro, 2015).

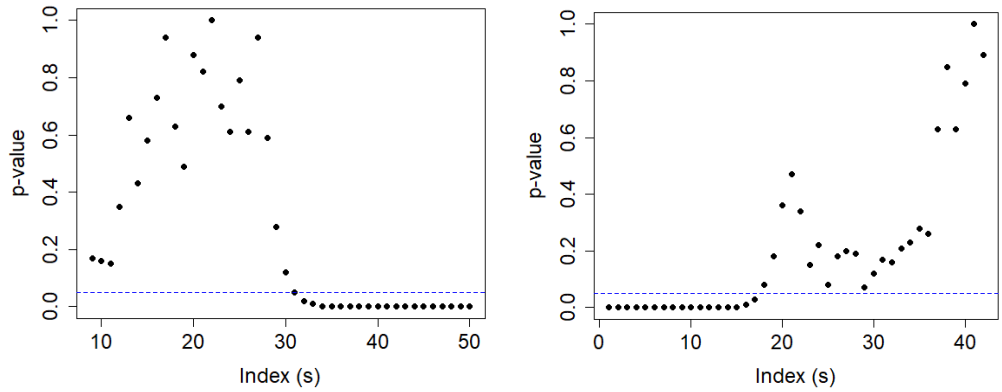
Let us revisit the simulation data described in Section 4.1 and apply the randomness test proposed by Bartels (1982). Here, the indices for ascending and descending data are same as the indices and s of algorithm in Section 4.3. As a result, Figures 5–8 show the corresponding plots for this rank test. In Figures 5(a) and (b), the test results show that all points are random if there is no change point. However, if there are change points, the test results indicate that all points are not random, regardless of the type of piecewise regression (Figures 6–8).

4.3. Algorithm

In accordance with Section 4.2, the result of randomness test is highly affected by which observation is a change point or not. Therefore, we propose a computing algorithm consisting of the following four steps in order to detect change point.

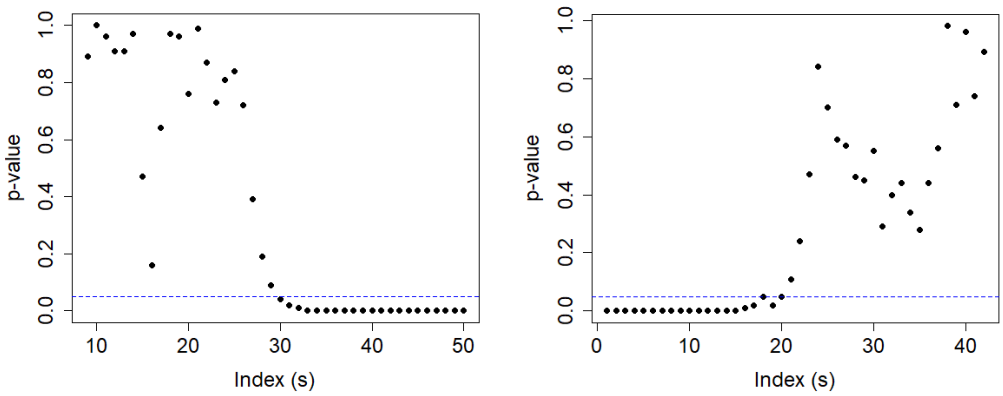
- Step 1. For $k = u, \ldots, N$, fit the Model (3.4):

$$y_{(k)i}|\epsilon_k = \alpha^{(k)} + \beta^{(k)} x_{(k)i} + \epsilon_i, \quad i = 1, \ldots, k-1;$$



(a) Randomness test of ascending data (b) Randomness test of descending data

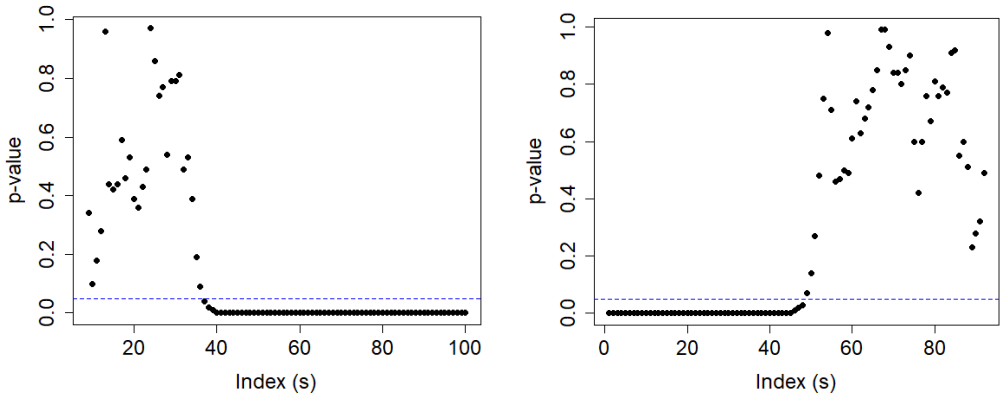
Figure 6: *Bartels test of intercept estimators with one change point (continuous type) with true number of observations in the first regime* $(n) = 25$ *(estimated* $\hat{n}$ *is* 24*).*



(a) Randomness test of ascending data (b) Randomness test of descending data

Figure 7: *Bartels test of intercept estimators with one change point (discontinuous type) with true number of observations in the first regime* $(n) = 25$ *(estimated* $\hat{n}$ *is* 25*).*

- Step 2. Estimate the difference-based intercept estimators:
  - Do $k = u : N\{$
    - Estimate intercepts, $\hat{\alpha}^{(k)}|\epsilon_k$;
    - Estimate intercepts $\hat{\alpha}^{(N-k+1)}|\epsilon_{N-k+1}$; $\}$
  - Lets $\hat{\boldsymbol{\alpha}}_{\text{asc}} = (\hat{\alpha}^{(u)}_{\text{asc}}, \hat{\alpha}^{(u+1)}_{\text{asc}}, \ldots, \hat{\alpha}^{(N)}_{\text{asc}})$ and $\hat{\boldsymbol{\alpha}}_{\text{des}} = (\hat{\alpha}^{N-u+1}_{\text{des}}, \ldots, \hat{\alpha}^{(2)}_{\text{des}}, \hat{\alpha}^{(1)}_{\text{des}})$;
- Step 3. Perform the rank test (Bartels, 1982):
  - Do $s = (u + v - 1) : N\{$
    - Put $\hat{\boldsymbol{\alpha}}_{\text{asc},s} = (\hat{\alpha}^{(u)}_{\text{asc}}, \hat{\alpha}^{(u+1)}_{\text{asc}}, \ldots, \hat{\alpha}^{(s)}_{\text{asc}})$ and $\hat{\boldsymbol{\alpha}}_{\text{des},s} = (\hat{\alpha}^{(N-u+1)}_{\text{des}}, \ldots, \hat{\alpha}^{(N-s+1)}_{\text{des}})$;
    - Perform the randomness test for sequence $\hat{\boldsymbol{\alpha}}_{\text{asc},s}$ and $\hat{\boldsymbol{\alpha}}_{\text{des},s}$;



(a) Randomness test of ascending data (b) Randomness test of descending data

Figure 8: *Bartels test of intercept estimators with two change points with true number of observations in the first regime* (n) = 25, *and in the second regime* (n_2) = 35 *(estimated* $\hat{n}$ *is* 28*).*

  * Compute the p-value of each test and put $p_{\text{asc},s}$ and $p_{\text{des},s}$; }
- Rewrite $\mathbf{p}_{\text{asc}} = (p_{\text{asc},u+v-1}, \ldots, p_{\text{asc},N})$ and $\mathbf{p}_{\text{des}} = (p_{\text{des},N-u-v+2}, \ldots, p_{\text{des},2}, p_{\text{des},1})$;

• Step 4. Decide if change points exist:

- Using the results of Step 3, the hypothesis, H_0: there is no change point, H_0 is rejected if $p_{\text{asc},N} > 0.05$ and $p_{\text{des},1} > 0.05$;
- If H_0 is rejected, i.e., there are change points, then the number of change points is determined by the following procedures:
  * Let $\max_{\text{asc}}$ be $\arg\min\{s : p_{\text{asc},s} < 0.05, s = (u+v-1), \ldots, N\}$ and $\min_{\text{des}}$ be $\arg\max\{s : p_{\text{des},s} < 0.05, s = 1, \ldots, (N-u-v+2)\}$;
  * If $\max_{\text{asc}} \geq \min_{\text{des}}$, there is one change point;
  * Otherwise, there are two or more change points;

• Step 5. Detect the location of first change points:

- If there is one change point between $\max_{\text{asc}}$ and $\min_{\text{des}}$, assume that the number of observations in the first regime, $\hat{n}$ is the median between $\max_{\text{asc}}$ and $\min_{\text{des}}$;
- If there are two or more change points, repeat the following loop until $\max_{\text{asc}} \geq \min_{\text{des}}$ {
  * Put $N = \min_{\text{des}} - 1$;
  * Perform Steps 1–4; }
- Then there is first change point between $\max_{\text{asc}}$ and $\min_{\text{des}}$. Here, assume that the number of observations in the first regime, $\hat{n}$ is the median between $\max_{\text{asc}}$ and $\min_{\text{des}}$.

The other change points can be detected with the remaining data except for the first regime corresponding to $\hat{n}$ detected using the above algorithm.

5. Simulation studies

We conduct simulations to evaluate the performance of our approach compared to other existing approaches: R packages chngpt (Fong *et al.*, 2017; Fong and He, 2018) and segmented (Muggeo, 2008, 2017). First, chngpt provides both estimation and hypothesis testing functionalities for four types of piecewise regression models. Second, segmented package offers a module to estimate the parameters in a GLM with segmented relationships. This package supports the continuous model.

5.1. Simulation setting

In order to assess and compare the performance of the proposed difference-based intercept estimators, simulations have been conducted under different conditions: sample sizes ($N = 50, 100$), the number of change points ($r = 0, 1, 2$), types of piecewise regression, and the number of independent variables. We generated 100 data sets for each model. The model is divided as follows.

- Model 1. This model for comparison is $y_i = 5 - 4x_i + \epsilon_i$, with $i = 1, \ldots, N$, where $\epsilon_i \overset{\text{iid}}{\sim} N(0, 1.5)$ and $N = 50$. The covariate x_i is generated with $U(0, 10)$.
- Model 2. Consider the Model (2.2) with one change point. We also divide this model using four types separated by Fong *et al.* (2017). The data generating processes for the covariate variable $\mathbf{x}$ is the same as in Model 1 and $\epsilon_i \overset{\text{iid}}{\sim} N(0, 1)$. Also, $n = 25$ and $N = 50$. We set four type models according to the values of the parameters α_1, α_2, β_1, and β_2 as follows.
  - hinge: This model is zero slope prior to the change point and continuous at the change point. We set parameters $\beta_1 = 0$, $\alpha_1 = 2$, $\beta_2 = 5$, and $\alpha_2 = \alpha_1 + (\beta_1 - \beta_2)\delta$ with change point location $\delta = x_n$.
  - segmented: This model generalizes the hinge model by allowing non-zero slope prior to the change point. Accordingly, put the parameter of the model be $\beta_1 = -4$, $\alpha_1 = 5$, $\beta_2 = 2$, and $\alpha_2 = \alpha_1 + (\beta_1 - \beta_2)\delta$ with change point location $\delta = x_n$.
  - step: In this model, both regression lines have a slope of 0, and are discontinuous at the change point. We set parameters $\beta_1 = 0$, $\alpha_1 = 12$, $\beta_2 = 0$, and $\alpha_2 = 3$.
  - stegmented: According to Fong *et al.* (2017), stegmented model is viewed as the fusion of the segmented and step models. Both regression liens have different slopes and intercepts. We set parameters $\beta_1 = -4$, $\alpha_1 = 10$, $\beta_2 = 2$, and $\alpha_2 = 2$.
- Model 3. Consider the model with two change points. We assume that the model is continuous at the change points:

$$y_i = \begin{cases} \alpha_1 + \beta_1 x_i + \epsilon_i, & \text{if } i \le n, \\ \alpha_2 + \beta_2 x_i + \epsilon_i, & \text{if } n < i \le n + n_2, \\ \alpha_3 + \beta_3 x_i + \epsilon_i, & \text{if } i > n + n_2, \end{cases}$$

  where $\epsilon_i \overset{\text{iid}}{\sim} N(0, 1)$ and $N = 100$. We set $\beta_1 = -2$, $\alpha_1 = 5$, $\beta_2 = 1$, $\alpha_2 = \alpha_1 + (\beta_1 - \beta_2)\delta_1$, $\beta_3 = -2$, and $\alpha_3 = \alpha_2 + (\beta_2 - \beta_3)\delta_2$ with change point locations $\delta_1 = x_n$ and $\delta_2 = x_{n+n_2}$ with $n = 25$ and $n + n_2 = 60$. The covariate x_i is generated with $U(0, 20)$.
- Model 4. We assume a multiple linear regression model with two explanatory variables and one change point:

$$y_i = \begin{cases} \alpha_1 + \beta_1 x_i + \gamma_1 z_i + \epsilon_i, & \text{if } i \le n, \\ \alpha_2 + \beta_2 x_i + \gamma_2 z_i + \epsilon_i, & \text{if } i > n, \end{cases}$$

Table 1: Comparison among three methods: DCD, SEG, and CHN

Model		r	DCD	SEG	CHN
Model 1		No	0.02	0.03	1.00
Model 2	hinge	One	0.00	0.00	0.00
	segmented		0.00	0.00	0.00
	step		0.00	0.86	0.00
	stegmented		0.00	0.00	0.00
Model 3		One	0.00	0.00	0.00
Model 4	Connected	Two	0.00	0.00	0.03
	Disconnected		0.00	0.06	0.04

DCD = our difference-based change point detection method; SEG = R package segmented proposed by Muggeo (2008, 2017); CHN = R package chngpt proposed by Fong *et al.* (2017), Fong and He (2018); r = the number of true change points.

where $\epsilon_i \overset{\text{iid}}{\sim} N(0,1)$ and $N = 50$. The change point is defined in terms of only one of the regressors, x. The location of the change point, δ is x_{25}. The covariate x_i and z_i are generated with $U(0,1)$, respectively. We set two types models according to the values of the parameters as follows.

- Continuous type: This model consists of a connected line at the change point. We set parameters $\theta_1 = (\alpha_1, \beta_1, \gamma_1 = (5, -3, 2)$ and $\theta_2 = (\alpha_2, \beta_2, \gamma_2 = (\alpha_1 + (\beta_1 - \beta_2)\delta, 2, 2)$.
- Discontinuous type: This model consists of a disconnected line at the change point. We set parameters $\theta_1 = (\alpha_1, \beta_1, \gamma_1 = (5, -3, 2)$ and $\theta_2 = (0, 2, 2)$.

Here, $u = 5$ and $v = 5$ for estimation and testing of the difference-based intercept estimator. To compare the performance of our difference-based change point detection method and R packages, segmented and chngpt, we consider the following setting of these packages. In segmented, we use "davies.test" and "pscore.test" of these functions for testing and "segmented" function for detection of change points. In chngpt, we use testing function, "chngpt.test" that tests each of the four types, respectively. The significance level for the rank test of Bartels (1982) is 0.05.

5.2. Simulation results

In this section, performances of our proposed procedure are evaluated by results of 100 replicates for each model. The results are summarized in Table 1, where we report the proportions of false detections among 100 replications.

We evaluate the ratio of falsely detecting a change point when there is no change point (Model 1). As a result, in our method (DCD), the ratio of false detections among 100 replications is 0.01, and is remarkably accurate. This implies that our approach can identify the existence of change points under various circumstances. However, in SEG, the ratio is 0.03 and in CHN, it is close to one. CHN tends to find the most appropriate change point for the model.

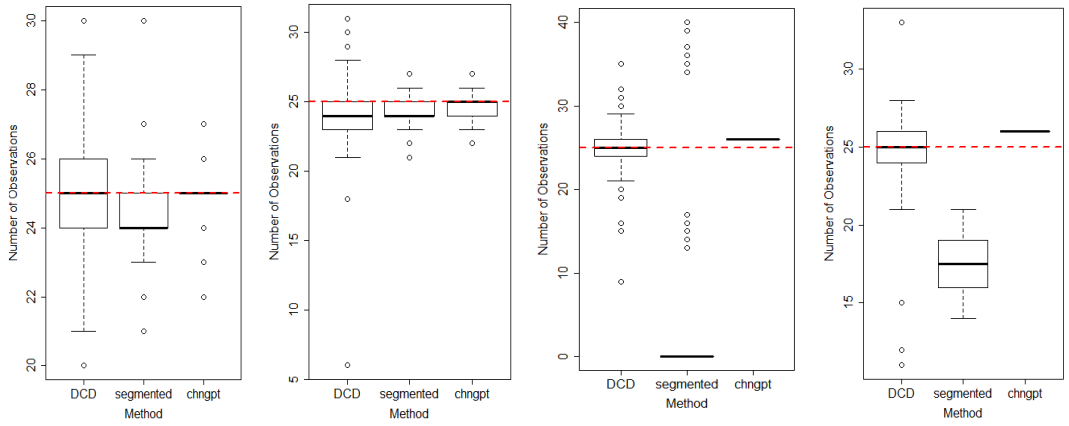
We evaluate the ratio of not detecting change points when there are change points (Models 2–4). First, our method (DCD) and CHN produce negligible percentages of false detection of change points for all types of model. However, results of SEG show a high ratio in the continuous model, but not in the discontinuous model. This is because this package is intended to be suitable for a continuous model. Second, for Model 3, all three methods show accurate results. Finally, in simulation of Model 4, our method shows accurate results regardless of the type of model, but SEG and CHN perform worse than our method.

In the following, we evaluate the accuracy of change point estimation when there are change points. The performance of the change point estimator has been evaluated, through mean and standard

Table 2: Performances of DCD: mean, SD, and ARB of estimated $\hat{n}$ when repeated 100 times; true number of observation in the first regime (n) = 25

Model		DCD			SEG			CHN		
		Mean	SD	ARB	Mean	SD	ARB	Mean	SD	ARB
Model 1		-	-	-	-	-	-	6.01	0.00	75.96
Model 2	hinge	24.78	1.84	0.88	24.59	4.64	1.64	24.99	0.10	0.04
	segmented	24.24	2.73	3.04	24.26	1.17	2.96	24.76	0.83	0.96
	step	24.87	3.37	0.52	3.63	1.09	85.48	26.00	0.92	4.00
	stegmented	24.44	2.79	2.24	17.41	9.95	30.36	26.00	0.00	4.00
Model 3		24.80	3.29	0.80	67.24	1.66	168.96	18.02	0.00	27.92
Model 4	Connected	24.91	1.49	0.36	24.39	3.34	2.44	24.77	1.52	0.92
	Disconnected	25.75	1.74	3.00	31.46	8.14	25.84	34.21	3.67	36.84

DCD = our difference-based change point detection method; SEG = R package segmented proposed by Muggeo (2008, 2017); CHN = R package chngpt proposed by Fong *et al.* (2017), Fong and He (2018); SD = standard deviation; ARB = absolute relative bias.



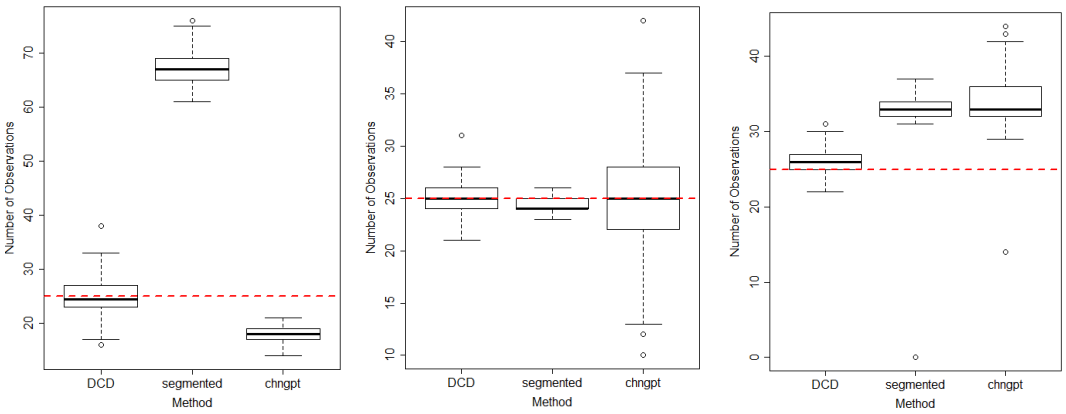
(a) Hinge (b) Segmented (c) Step (d) Stegmented

Figure 9: *Boxplots of n estimated by each method on simulated data for Model 2 with N = 50 and true number of observations in the first regime* (n) = 25.

deviation (SD). The performance of the three methods is also evaluated via the absolute relative bias (ARB) (Chen *et al.*, 2011) that represents the percentage error of the estimate $\hat{n}$ compared to the true value n ($|\ (\hat{n} - n)/n\ | \times 100$). Table 2, Figure 9, and Figure 10 show the results.

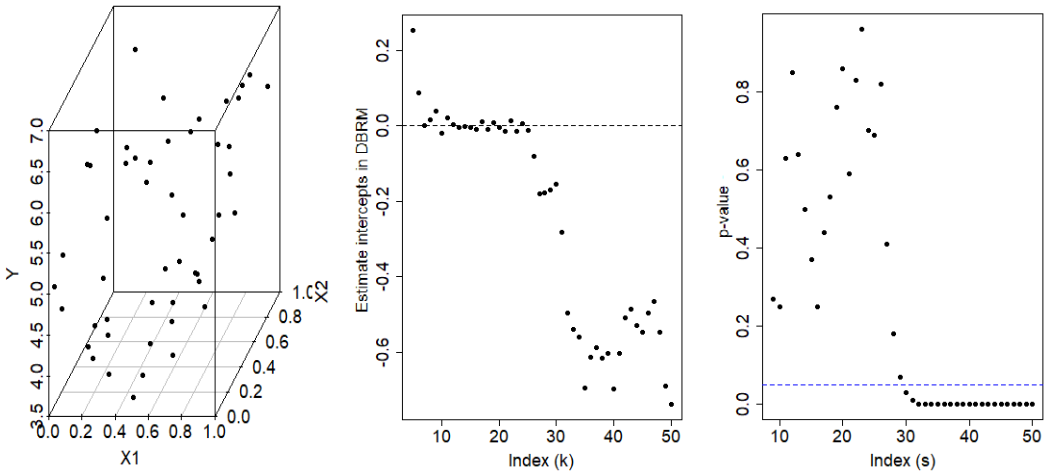
We first discuss Model 2 with only one change point. Our method and CHN estimate the location of change point with reasonable accuracy (Table 2). However, results of SEG estimate the location of change point accurately in the continuous model, but not accurately in the discontinuous model. In the simulation of Model 3 and Model 4, our method (DCD) works better than the other two methods. The results in Figure 9 and Figure 10 indicate that our method is more accurate than the other two methods. It is shown that our method is superior in the case of one change point and in the case of two change points in the simple linear regression model as well as superior in the case of one change point in the multiple regression model.

Figure 11 and Figure 12 display scatter plots of the intercept estimates and the results of the rank test for Model 4. The intercept estimate is influenced by the change point effect. This means that the



(a) Two change point (b) Model 4 connected (c) Model 4 disconnected

Figure 10: *Boxplots of the first n estimated by each method on simulated data for Model 3 and Model 4: (a) Model 3: N* = 100, *true number of observations in the first regime* (n) = 25, *and in the second regime* (n_2) = 35. *(b) and (c) Model 4: N* = 50, *true number of observations in the first regime* (n) = 25.



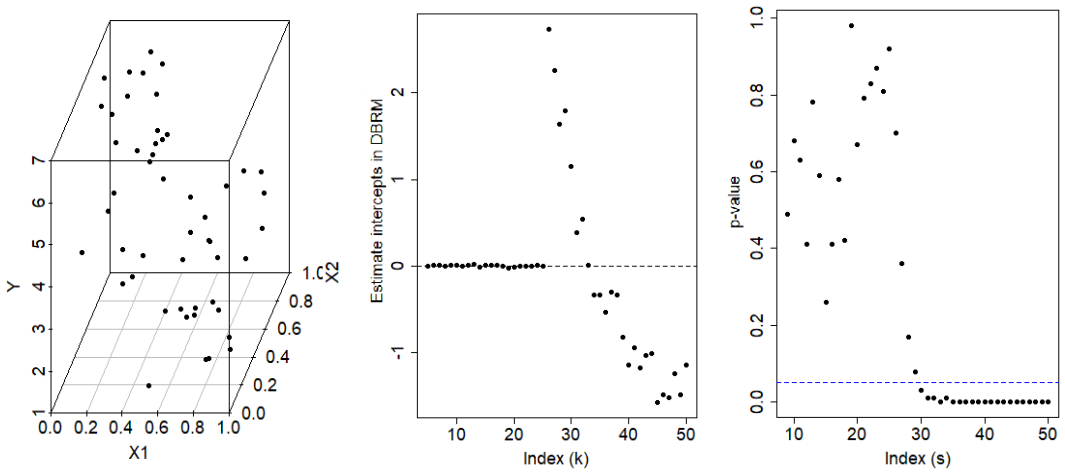
(a) Scatter plot (b) Estimated intercepts (c) Randomness test

Figure 11: *Intercept estimators and result of Bartels test on simulated data for Model 4 (continuous type), true number of observations in the first regime* (n) = 25 *(estimated* $\hat{n}$ *is* 25*).*

difference-based intercept estimators that correspond to the observations including the change point effect, are large values.

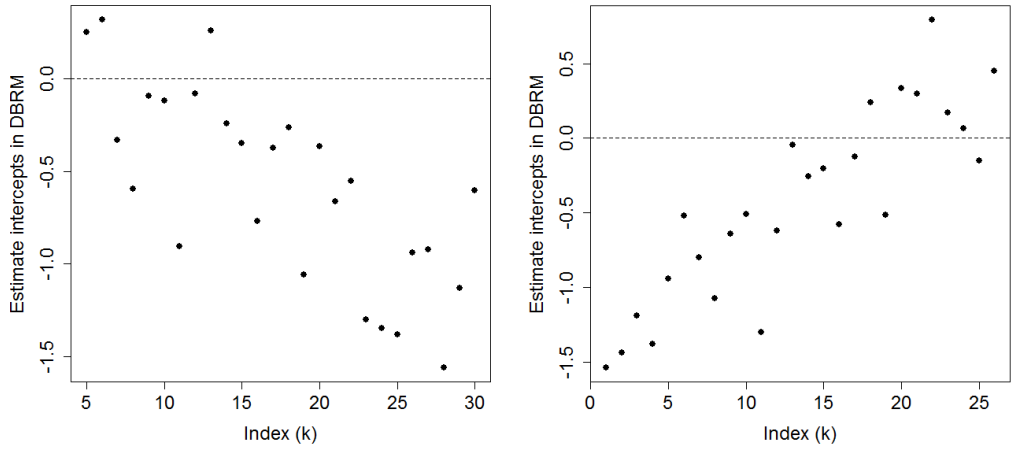
6. Real data analysis

In this section, we apply our difference-based change point detection method to Down syndrome (DS) data (Davison and Hinkley, 1997). This data set is used by Muggeo (2008) to evaluate his method. There are three explanatory variables: the number of babies with DS (cases), the number of total births (births) and the mother's mean age (age). DS risk generally increases with mother's age, but



(a) Scatter plot (b) Estimated intercepts (c) Randomness test

Figure 12: *Intercept estimators and result of Bartels test on simulated data for Model 4 (discontinuous type), true number of observations in the first regime* (n) = 25 *(estimated* $\hat{n}$ *is* 25.*).*

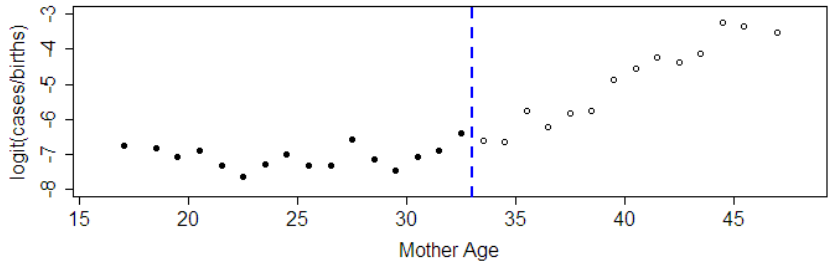


(a) Estimated intercepts of ascending data (b) Estimated intercepts of descending data

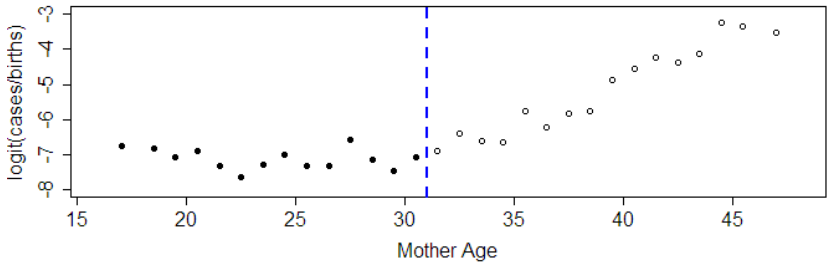
Figure 13: *Estimated intercepts in the DBRM without the* k^{th} *observation in our example data. DBRM = difference-based regression model.*

evaluation is needed to show at what point a risk change occurs.

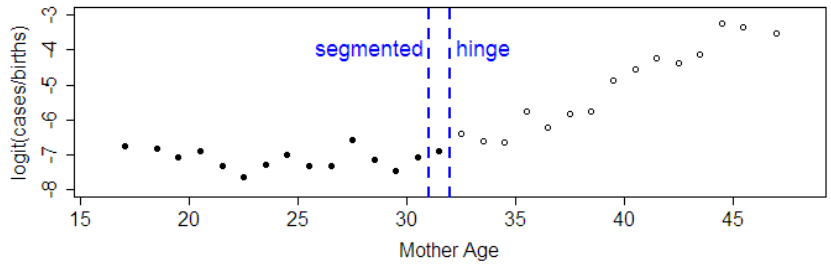
We apply our method (DCD) to this data and show the estimated intercepts in Figure 13. These estimates show a distinct trend between the 13^{th} observation and the 17^{th} observation. Even though a clear stopping rule is not provided, we can check for existence of change points using a trend of difference-based intercept estimates. Randomness tests are performed on the intercepts added one by one; in addition, we also show the scatter plot and the estimated $\hat{n}$ for each method. As a result, $\hat{n}$s of the three methods are slightly different in Figure 14. For our method, the number of observations in the first regime determined by the median of the interval between min and max, is 16. For SEG



(a) DCD



(b) SEG



(c) CHN

Figure 14: *Scatter plot and regimes obtained each method in our example data. DCD = our difference-based change point detection method; SEG = R package* segmented *proposed by Muggeo (2008, 2017); CHN = R package* chngpt *proposed by Fong et al. (2017), Fong and He (2018).*

method, under a null left slope constraint, the number of the first regime observations is 14 (Muggeo, 2008). For CHN method, under the hinge type, the number of the first regime observations is 15.

7. Discussion

In this paper, we derive a way to solve change point regression problems via a process for getting the consequential results using the properties of a difference-based intercept estimator (Park and Kim, 2019) as well as describe the statistical properties of the DBRM in a PSLR. We also propose an algorithm for change point detection.

We compare the proposed methodology with other methods available in the recent literatures, SEG and CHN. The simulation results indicate that the performance of the proposed method is good in various circumstances. First, our method can successfully identify the existence of change points under

various circumstances: continuous and discontinuous types, single covariate and multiple covariates, one change point and multiple change points. We can check for existence of change points using a trend of the difference-based intercept estimators despite not providing a clear stopping rule. We also determine the change point location as an interval and point estimation using our proposed algorithm. Our method is also more accurate than the other two methods, SEG and CHN: our method is superior in the case of one change point and in the case of two change points in the simple linear regression model, and our method is superior in the case one change point in the multiple regression model. Our method is affected by the number of samples in each regime and must have at least 15 samples. Therefore, we need to develop our method to make robust estimates regardless of the sample size in each regime as well as develop an algorithm that automatically searches for change point detection.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2018R1D1A1B070).

References

Andrews DWK (1993). Tests for parameter instability and structural change with unknown change point, *Econometrica*, **61**, 821–856.

Andrews DWK and Ploberger W (1994). Optimal tests when a nuisance parameter is present only under the alternative, *Econometrica*, **62**, 1383–1414.

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

Bartels R (1982). The rank version of von Neumann's ratio test for randomness, *Journal of the American Statistical Association*, **77**, 40–46.

Betts M, Forbes G, and Diamond A (2007). Thresholds in songbird occurrence in relation to landscape structure, *Conservation Biology*, **21**, 1046–1058.

Chen CWS, Chan JSK, Gerlach R, and Hsieh WYL (2011). A comparison of estimators for regression models with change points, *Statistics and Computing*, **21**, 395–414.

Davison A and Hinkley D (1997). *Bootstrap Methods and Their Application*, Cambridge University Press.

Fong Y and He Z (2018). Package 'chngpt': Estimation and Hypothesis Testing for Threshold Regression, R package version 2018.7-25. Available from: https://cran.r-project.org/web/ packages/chngpt/chngpt.pdf

Fong Y, Huang Y, Gilbert PB, and Permar SR (2017). chngpt: threshold regression model estimation and inference, *BMC Bioinformatics*, 18:454.

Hawkins DM (1980). A note on continuous and discontinuous segmented regressions, *Technometrics*, **22**, 443–444.

Julious SA (2001). Inference and estimation in a changepoint regression problem, *The Statistician*, **50**, 51–61.

Kim HJ and Siegmund D (1989). The likelihood ratio test for a changepoint in simple linear regression, *Biometrika*, **76**, 409–423.

Loader CR (1996). Change point estimation using nonparametric regression, *The Annals of Statistics*, **24**, 1667–1678.

Mateus A and Caeiro F (2015). Package 'randtests': Testing randomness in R, R package version 1.0. Available from: https://cran.r-project.org/web/packages/randtests/randtests.pdf

Muggeo VMR (2003). Estimating regression models with unknown break-points, *Statistics in Medicine*, **22**, 3055–3071.

Muggeo VMR (2008). Segmented: An R package to fit regression models with broken-line relationships, *R News*, **8/1**, 20–25.

Muggeo VMR (2017). Package 'segmented': Regression Models with Break-Points/Change-Points Estimation, R package version 0.5-3.0. Available from: https://cran.r-project.org/web/packages/segmented/segmented.pdf

Park CG and Kim I (2019). Robust difference-based outlier detection, *Communications in Statistics - Theory Methods*, Published online.

Quandt RE (1958). The estimation of the parameters of a linear regression system obeying two separate regimes, *Journal of the American Statistical Association*, **53**, 873–880.

Ulm K (1991). A statistical methods for assessing a threshold in epidemiological studies, *Statistics in Medicine*, **10**, 341–349.

Zeileis A, Kleiber C, Krämer W, and Hornik K (2002). Testing and Dating Structural Changes in Exchange Rate Regimes, *Computational Statistics and Data Analysis*, **54**, 1696–1706.

Zhou HL and Liang KY (2008). On estimating the change point in generalized linear models, *IMS Collections Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*, **1**, 305–320.

Received May 9, 2019; Revised September 26, 2019; Accepted October 11, 2019