Statistical Estimation in Piecewise Linear Regression Models

Abstract

The kink regression model assumes that linear regression forms are separately modelled on two sides of an unknown threshold but still continuous at the threshold. This paper considers statistical estimation for piecewise linear regression models which are widely used in various fields to capture nonlinear relationships between variables. The estimators for the kink locations and regression coefficients are obtained by using the least squares method, a detailed explanation of the estimation process is provided. Furthermore, the proposed methodology is validated through an illustrative example using Monte Carlo random simulation, demonstrating its effectiveness in accurately capturing nonlinear patterns and changes in the data.

Keywords: *piecewise linear regression; jump point; least squares estimation; parmeters estimation*

1 Introduction

In today's life, data structure is becoming more and more complex. In many cases, the simple linear model is not enough to describe the relationship between variables. More often than not, the relationship appears nonlinear. Threshold regression models are very useful tools to describe this nonlinear relationship by introducing one or more threshold parameters, also known as kink or change points. Threshold regression models offer a flexible framework to model such relationships by allowing different linear relationships in different intervals of the predictor variables. Threshold regression models can take many forms, depending on what happens at the threshold. Figure 1 displays four types of threshold effects: step, hinge, segmented and stegmented. Each type of threshold regression model has been studied by many scholars and widely used in the economy, finance, medicine, and other fields. The stegmented regression, as a special threshold regression model, specifically, there are jump points and kink points in the regression model. The kink regression model with a single kink point was first introduced by Lerman (1980). Hudson (1996) put forward the least square estimation of the model in the linear regression model with a known single kink point, but the parameter estimation is not accurate when the parameter is unknown. Fan and Li (2001) proposed a variable selection method based on penalized least squares, introducing L_1 regularization (LASSO) to address variable selection in high-dimensional data. This method not only effectively reduces model complexity but also improves prediction accuracy. Lee et al. (2011) proposed a general sup-likelihood-ratio test statistics to detect threshold effects in regression models. This paper focuses on a specific type of piecewise linear regression model that includes jumps in both the response function and its first-order derivative. Hansen (2017) considered the kink regression model with an unknown threshold, and combined the least squares estimation and grid search algorithm to estimate the regression coefficients and the kink point.

However, the kink regression models with only one threshold point are always not sufficient in practice. They may not capture multiple structural changes, which are quite common in many research fields. Motivated by this limitation, a tremendous amount of attention has been focused on the kink

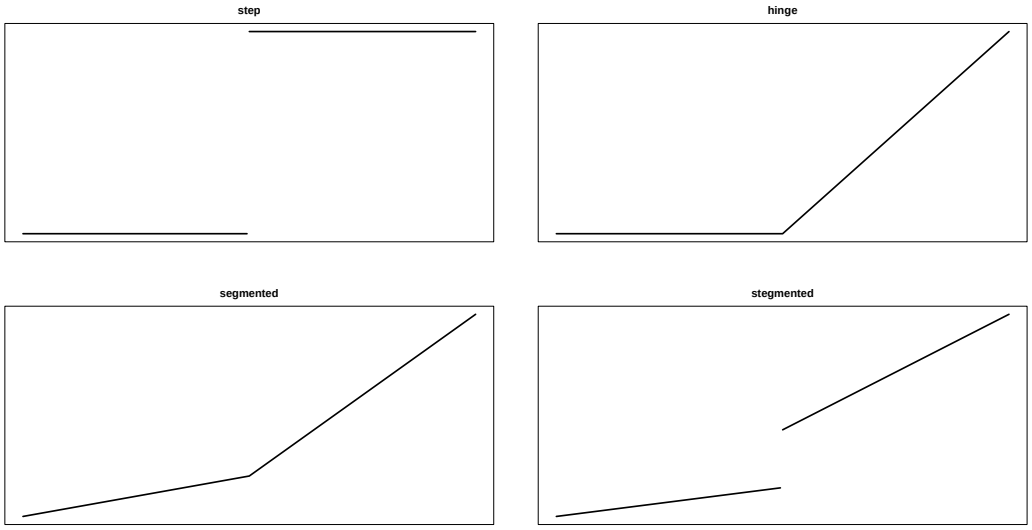

Figure 1: *Four types of Threshold regression model*

regression models with multiple change points. Chan (1993) proved that the least squares estimator has strong consistency in the discrete piecewise autoregressive model, and gave the convergence rate of the parameter estimator. Bai and Perron (1998) proposed a multiple structural change detection method based on least squares, capturing structural changes in data by introducing multiple kink points. This method has achieved significant success in economic and financial applications, particularly in studying economic cycles and financial market volatility, where structural changes are evident. Muggeo and Adelfio (2010) studied a piecewise constant model in mean regression with multiple change points and used the penalized method to select kink points. Shi et al (2020) considered the robust continuous piecewise linear regression model with multiple change points and applied it to the body mass index (BMI) and age relationship. Wan et al (2023) considered composite quantile estimation for the kink model with longitudinal data.

In this article, we consider the piecewise linear regression models. The core idea of piecewise linear regression models is to divide the interval of the predictor variable into several segments, where the relationship between the response variable and the predictor variable is linear within each segment. However, traditional piecewise linear models typically only consider jumps in the response function, ignoring changes in its first-order derivative. The model proposed in this paper not only allows jumps in the response function at certain points but also permits abrupt changes in its first-order derivative at these points. This model is particularly suitable for describing systems that exhibit abrupt changes at specific points, such as policy changes in economics or physical constraints in engineering.

The rest of this paper proceeds as follows. In Section 2, we introduce the piecewise linear regression models, and its estimation procedures for the model we have proposed under two distinct scenarios. A specific application is illustrated in Section 3. Section 4 concludes the paper.

2 Methodology and Simulation

Let Y_i be a response variable of interest, and X_i be a univariate threshold variable, and $\boldsymbol{Z}$ be a p dimensional random vector of additional covariates, $i = 1, \ldots, n$. Considered the following regression model

$$Y_i = \alpha_0 + \alpha^\top Z_i + \beta_0 X_i + \sum_{j=1}^{q_1} \beta_{1j} I(X_i > \delta_{1j}) + \sum_{k=1}^{q_2} \beta_{2k} (X_i - \delta_{2k})_+ + \varepsilon_i = \alpha_0 + \alpha^\top Z_i + g(X_i) + \varepsilon_i \quad (1)$$

where $x_+ = xI(x > 0)$ and $I(\cdot)$ is the indicator function. In model (1), $(X - \delta_k)_+ = (X - \delta_k)I(X > \delta_k)$, $g(x)$ is piecewise linear, q_1 is the number of jumps in $g(x)$, $\{\delta_{1j}, j = 1, \ldots, q_1\}$ and $\{\beta_{1j}, j = 1, \ldots, q_1\}$ are the corresponding jump positions and jump sizes. q_2 is the number of jumps in $g'(x)$ and $\{\delta_{2k}, k = 1, \ldots, q_2\}$ and $\{\beta_{2k}, k = 1, \ldots, q_2\}$ are the corresponding jump positions and jump sizes. ε_i is independent identically distribution random errors with mean 0 and unknown variance σ^2. This paper explores parameter estimation under two distinct scenarios: (i) when the number of jumps, denoted as q_1 and q_2, is known a priori, and (ii) when the number of jumps is unknown and must be inferred from the data, we will be discussed in detail.

The parameters of the model (1) can be estimated using the least squares method. Since both the response function and its first-order derivative contain jumps, the parameter estimation process is relatively complex. We will discuss parameter estimation under two scenarios: when the number of jumps q_1 and q_2 are known and when they are unknown.

2.1 When q_1 and q_2 are Known

We generate independently and identically distributed sample $\{(X_i, Z_i, Y_i), i = 1, \ldots, n\}$ from the model (1), we assuming that q_1, q_2 is known or fixed, β_{1j}, δ_{1j}, β_{2k}, δ_{2k} are assumed unknown. When the number of jumps q_1 and q_2 are known, the least squares method can be directly used to estimate the model parameters. The goal of the least squares method is to minimize the residual sum of squares:

$$S(\theta) = \sum_{i=1}^{n} (Y_i - \alpha_0 - \alpha^\top Z_i - g(X_i))^2$$

where $\theta = (\alpha_0, \alpha, \beta_0, \{\beta_{1j}\}, \{\delta_{1j}\}, \{\beta_{2k}\}, \{\delta_{2k}\})$ represents all the parameters in the model. The result of parameter estimation is

$$\hat{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

where $\mathbf{X} = \{1, Z_{i1}, \cdots, Z_{ip}, X_i, I(X_i > \delta_{11}), \cdots, I(X_i > \delta_{1q_i}), (X_i - \delta_{21})_+, \cdots, (X_i - \delta_{2q_2})_+\}$, $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_n)^\top$.

To evaluate the performance of the model, we can calculate mean squared error (MSE): MSE $= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, where $\hat{Y}_i$ are the fitted values obtained from the estimated model. The MSE provides a measure of the model's predictive accuracy, with lower values indicating a better fit to the data.

2.2 When q_1 and q_2 are Unknown

When the number of jumps q_1 and q_2 are unknown, it is necessary to first determine the number of jumps before estimating the model parameters. We often use cross-validation methods to select model.

First, we define the sum of squared residuals $S(\theta)$:

$$S_u(\theta) = \sum_{i=1}^{n} \left(Y_i - \left(\alpha_0 + \alpha^\top Z_i + \beta_0 X_i + \sum_{j=1}^{q_1} \beta_{1j} I(X_i > \delta_{1j}) + \sum_{k=1}^{q_2} \beta_{2k}(X_i - \delta_{2k})_+ \right) \right)^2 \tag{2}$$

then setting candidate ranges for q_1 and q_2, we set $q_1 \in \{0, 1, 2, \ldots, q_{1,\max}\}$, $q_2 \in \{0, 1, 2, \ldots, q_{2,\max}\}$. For each pair (q_1, q_2), fit the model and compute the model selection criterion, The specific steps are as follows:

(1) Split the data into training and validation sets.

(2) For each pair(q_1, q_2), fit the model on the training set and compute the mean squared error(MSE) on the validation set.

(3) Based on the model selection criterion, choose the (q_1, q_2)pair that minimizes the cross-validation error. For the selected q_1 and q_2, construct the design matrix $\mathbf{X}$, where is same as in section 2.1. Each row of the design matrix corresponds to an observation i, and each column corresponds to a parameter in the model, then we use the least squares method to minimize $S(\theta)$.

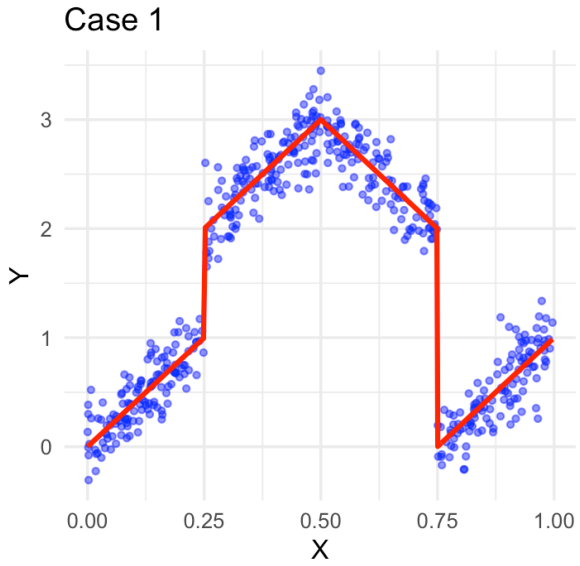

Figure 2: *Known Deltas (q1, q2 given)*

(4) Output the estimated parameters. Return the estimated parameters $\hat{\theta}$.
(5) Calculate the MSE to evaluate model prediction accuracy.

3 Simulation Studies

To illustrate the estimation process, this section investigates the proposed estimation method through Monte Carlo random simulation. To conduct the simulation, we consider the following linear regression model:

$$g(x) = 4x + I(x > 0.25) + I(x > 0.75) - 8(x - 0.5)_+ + 8(x - 0.75)_+ = \begin{cases} 4x, & 0 \le x < 0.25 \\ 4x + 1, & 0.25 \le x < 0.5 \\ -4x + 5, & 0.5 \le x < 0.75 \\ 4x, & 0.75 \le x < 1 \end{cases}$$

We generate $n = 500$ observations, and X_i is uniformly distributed over the interval [0,1], and the random errors follow a normal distribution ε_i follows from $N(0, 0.2^2)$.

When the number of jumps q_1, q_2 is given, the least squares method is employed for parameter estimation, as detailed in Section 2.1. Figure 2 displays the scatter plot along with the fitted regression curves. As anticipated, the function exhibits jumps at $x = 0.25$ and $x = 0.75$, and its first-order derivative has jumps at $x = 0.5$ and $x = 0.75$, These observations align with the theoretical expectations of the model. Furthermore, the parameter estimation results are obtained, yielding a MSE of 0.0384, which indicates a relatively high level of accuracy in the model's fit to the data.

When the number of jumps q_1, q_2 is unknown, the estimation method described in Section 2.2 is employed. Figure 3 illustrates the scatter plot along with the fitted regression curves. From the results, it is evident that the function exhibits jumps at $x = 0.25$ and $x = 0.75$, and its first-order derivative has jumps at $x = 0.5$ and $x = 0.75$. These findings are consistent with the theoretical properties of the model. Importantly, the proposed estimation method demonstrates strong performance in accurately identifying both the locations and the number of change points. The MSE of 0.0424. further confirms the robustness and precision of the estimation approach. From this example, it can be observed that the estimated parameters closely match the true parameters, and the recovered function $g(x)$ accurately

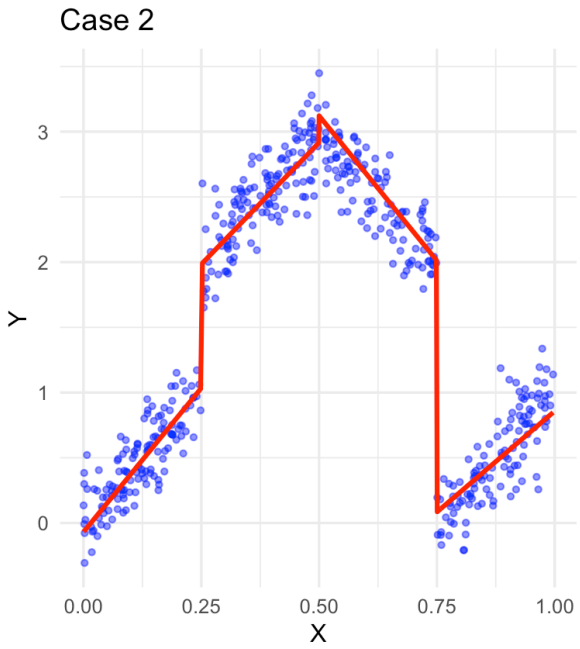

Figure 3: *Unknown Deltas (q1, q2 estimated)*

captures the jumps and kinks in the true function. The model performs well in identifying the positions and sizes of the jumps in both the response function and its derivative.

4 Discussion and Conclusion

In this article, we study the piecewise linear regression model which includes jumps in both the response function and its first-order derivative, we use the least squares method to estimate the model parameters in two scenarios: when the number of jumps q_1 and q_2 are known and when they are unknown, and the estimation process is explained. The proposed methodology offers several advantages over conventional piecewise regression models. First, by allowing for discontinuities in both the function values and their derivatives, our model can capture more complex patterns of structural change commonly observed in real-world phenomena. Second, the least squares framework provides a computationally tractable solution while maintaining desirable statistical properties. Through a simulation example, we have verified the model's ability to capture jumps and kinks.

These results suggest several promising directions for future research. The model could be fruitfully applied to various empirical domains where structural breaks are theoretically expected, such as economic time series analysis, biological growth modeling, or engineering system monitoring. From a methodological perspective, developing more computationally efficient algorithms, particularly for high-dimensional extensions of the model, would significantly enhance its practical utility. Additional theoretical work could establish formal conditions for the consistency of the jump point estimators and derive their asymptotic distributions. Furthermore, extending the current framework to accommodate other types of regression models or alternative estimation approaches would broaden the method's applicability to diverse statistical problems.

References

Bai J, Perron P. Estimating and testing linear models with multiple structural changes. *Econometrica*, 1998, 66(1): 47-78.

Chan K S. Consistency an limiting distribution of the least squares estimator of a threshold autoregressive model. *The Annals of Statistics*, 1993, 21(1): 520–533.

Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*,2001, 96(456): 1348–1360.

Hansen B E. Regression kink with an unknown threshold. *Journal of Business and Economic Statistic*, 2017, 35(2): 228–240.

Hudson D J. Fitting segmented curves whose join points have to be estimated. *Journal of the American Statistical Association*, 1996, 61(316): 1097–1129.

Lerman P M. Fitting segmented regression models by grid search. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 1980, 29(1): 77–84.

Lee S, Seo M H, Shin Y. Testing for threshold effects in regression models. *Journal of the American Statistical Association*, 2011, 106(493): 220–231.

Muggeo V M R, Adelfio G. Efficient change point detection for genomic sequences of continuous measurements. *Bioinformatics*, 2011, 27(2): 161–166.

Shi S, Li Y. Wan C. Robust continuous piecewise linear regression model with multiple change points. *Supercomput*, 2020, 76: 3623–3645.

Wan C, Zhong W, Fang Y. Composite quantile estimation for kink model with longitudinal data. *Acta Mathematica Sinica, English Series*, 2023, 39(3): 412–438.

Zhang L, Wang H J, Zhu Z. Testing for change points due to a covariate threshold in quantile regression. *Statistica Sinica*, 2014, 24(4): 1859–1877.

Zhang F, Li Q. Robust bent line regression. *Journal of Statistical Planning and Inference*, 2017, 185: 41–55.

Zhang B W, Geng J, Lai L F. Multiple change-points estimation in linear regression models via sparse group lasso. *IEEE Transactions on Signal Processing*, 2015, 63(9): 2209–2224.