Dean&Francis
ISSN 2959-6157

Theoretical and applied research progress of breakpoint regression method

Frank Qiu

Shanghai Experimental Foreign Language School, Shanghai, Shanghai, 201505, China

1120167094@qq.com

Abstract:

Breakpoint regression is closer to quasi-natural experiments in design, and the estimated results are more accurate, which has attracted widespread attention from the academic community in recent years. This study systematically sorts out the theoretical and applied research of breakpoint regression in the field of policy evaluation. The research results show that the theoretical research on breakpoint regression mainly focuses on model applicability, bandwidth selection, data accumulation problems, etc., and breakpoint regression is in higher education, ecological environment, fiscal taxation, scientific and technological innovation, etc. The field has been widely used. This article aims to provide new ideas for the application scenarios of breakpoint regression, and research combined with machine learning can be added in the future.

Keywords: breakpoint regression; policy evaluation; machine learning; result variables

1. Introduction

Policy evaluation is the core issue of economic analysis, but when using measurement models for research, it will inevitably be disturbed by endogeneous problems, resulting in deviations in estimated results. Therefore, it is necessary to alleviate the endogenous problems in the model. Common methods include tool variables, double difference, breakpoint regression, tendency score matching, etc. Among them, breakpoint regression is closer to quasi-natural experiments in terms of design, and the estimated results are more accurate, which has obvious advantages over other methods, and has also received widespread attention from the academic community.

Breakpoint regression design is an observational research method. Its basic logic is to use one or more critical points or turning points to divide samples and evaluate the effect of policies or interventions by comparing the average effects on both sides of breakpoints. Breakpoint regression analysis can be divided into two types according to the changing characteristics of the probability of the individual obtaining the processing effect at the breakpoint: precise breakpoint regression design and fuzzy breakpoint regression design. However, regardless of which type, the estimated steps are as follows: 1 clarify the problem and collect relevant data; 2 identify one or more breakpoints, and select the appropriate breakpoint according to the data characteristics and research problems; 3 divide the data into two parts: the control group and the experimental group. The control group is not affected by policies or intervention measures, while the experimental group Then it will be

ISSN 2959-6157

received; 4 fit the regression model in the two groups and calculate the average effect of policies or interventions; 5 Compare the average effects of the two groups to evaluate the effect of policies or interventions.

Next, this paper systematically sorts out and summarizes the research related to breakpoint regression from the two levels of theory and application, in order to provide new ideas for the application scenarios of this method.

2. Theoretical research on breakpoint regression method

2.1 Model applicability

At present, the judgment is mainly based on the diagram of the result variables. Based on the scatter diagram of the mean of the result variable, make a fitting curve and look at the shape of the potential function. If the mean of the result variable jumps at the breakpoint, it means that the processing variable has an impact. The diagram is used to show the relationship between the result variable and the processing variable. Dong Huixin and Du Liangsheng (2024) studied the impact of standard collection and simple collection on the sales of enterprises. By drawing a diagram of the relationship between enterprise sales and the probability of following the standard collection policy, it was found that there was a discontinuous change in the registration threshold value. Once the enterprise sales exceeded the threshold value, the simplified The proportion of enterprises with easy-to-levy mode has decreased significantly. If the simple levy mode continues to be adopted beyond the threshold value, enterprises will face administrative penalties, which provides an empirical basis for the use of fuzzy breakpoint regression to study the causal relationship between variables. Ma Shuzhong and He Ge (2024) provided graphic evidence that the number of products removed from the shelves and the number of uncertified products removed from the shelves appeared at breakpoints after the implementation of the fraudulent fulfillment of the order policy. As a result, it was found that the number of new products pointed out by enterprises at the time of policy implementation had an obvious downward "jump", and the number of certified products existed The significant upward "jump" means that the implementation of the fraudulent fulfillment order policy has significantly affected the decision-making of enterprises to enter the market, resulting in a significant decrease in new products, while the number of certified products has increased significantly, indicating that the products selected to enter are relatively better products. On the other hand, the number of products removed from the shelves and the number of uncertified products removed from the shelves have almost synchronously "jumped up". After the implementation of the policy, enterprises will have more underground products, especially low-quality products.

2.2 Choice of bandwidth

In the regression process, it is very important to choose the right bandwidth, and the choice of bandwidth is to weigh between accuracy and deviation. Bandwidth involves the size of the breakpoint domain, that is, to check whether there is a jump in this field. If the bandwidth selection is large, the more observed values can be estimated, which will make the estimated results more accurate, but the deviation of the average processing effect estimate will be greater. On the other hand, if the bandwidth selection is small, the accuracy will also be reduced but the deviation will be reduced. IK method and CCT method are usually used to determine bandwidth (Imbens and Kalyanaraman, 2012; Cataneo and Titiunk, 2022). Li Yongtao et al. (2023) studied the impact of individual endowment insurance incentives by setting different bandwidths. Wang He and Jia Nan (2024) reset the bandwidth to 50%, 100% and 200% of the optimal bandwidth. The results show that except for the 50% bandwidth, which is not significant, the rest of the bandwidth is significant and consistent with the positive and negative of the benchmark regression.

2.3 Data accumulation problem

Data accumulation refers to the phenomenon of excessive observation values of certain values of the configuration variable. The reasons for this phenomenon include respondents tend to approximate to a certain value when reporting certain information, and the limited reading of the measured ruler. The phenomenon of data manipulation originates from the profit-seeking motives of economic individuals and only appears at breakpoints. However, data accumulation is not a profit-seeking motive derived from economic individuals, and may appear elsewhere except the breakpoint. If the result variable is affected by the accumulation of configuration variables, then the RDD estimate may be biased. At this time, you can remove some observation values near the breakpoint and then do RDD estimation. Because the data after removing the observation value near the breakpoint is like a donut, it is called "donut RDD". As for the most appropriate observation value near the breakout, there has been no consensus in the literature. Lusnuo and Fengjin (2023) carried out "donut breakpoint regression" in the last phase before the implementation of the removal policy to exclude the impact of data accumulation before and after breakpoints. In order to exclude the estimated bias caused by the ac-

cumulation of data near the critical point of the driving variable, Wang Zhuo et al. (2024) used the donut RDD for testing. The specific practice is to remove 5% and 10% of the samples in the optimal bandwidth near the critical point of the drive variable respectively, and then carry out breakpoint regression. As a result, it was found that environmental control policies can still significantly improve the energy efficiency of enterprises, and the energy efficiency of enterprises with small tax contributions is still higher than that of enterprises with large tax contributions.

3. Applied research on breakpoint regression method

Breakpoint return has been widely used in higher education, ecological environment, fiscal taxation, scientific and technological innovation and other fields. This article will focus on these fields.

3.1 Higher education

Wang Jun and Sun Zhijun (2015) used the full sample data of two ordinary high school students in F County, and used the exogenesic impact of ordinary high school enrollment and admission on whether students can enter key high schools. According to the principle of breakpoint regression design, they studied the impact of key high schools on students' academic performance. The estimated results show that in terms of science students, the overall college entrance examination scores, mathematics and Chinese scores of key high school students are higher than those of ordinary high schools, but from a numerical point of view, this difference is not big; in terms of liberal arts students, there is no significant difference between key high schools and general high schools. This shows that key high schools only have a weak positive impact on students' academic performance. However, because students with good learning ability and learning foundation are more willing to choose science, the impact of key high schools on the academic performance of students in different subject categories may reflect the impact of key high schools on the academic performance of students with different learning abilities and academic foundations. The different dependencies of different disciplines on school resources and the different preferences within the school for the allocation of resources in different disciplines may also lead to differences in disciplines that affect this effect. In addition, key high schools have a greater impact on girls' academic performance. The impact on urban students' college entrance examination scores is obviously greater than that of rural students, but it has a greater impact on rural students' math and Chinese scores. Liu Shenglong et al. (2023), with the help of China's higher education admission rules and relevant systems, based on the breakpoint regression design framework, estimated the impact of key university education on the salary of postgraduate individuals, studying abroad and the first job found that the return of key university education is very significant, through the human capital effect, signal effect and the same The accompanying effect has an impact on personal studies and careers, making personal English scores better, study time significantly increased, and easier to get the opportunity to be sent to graduate school.

3.2 Eco-environment

Qin Zhilong and Chen Xiaoguang (2020) used the air pollution data of the Ministry of Ecology and Environment in 2011 to investigate the impact of the Beijing-Shanghai high-speed railway opened on urban air quality by the breakpoint regression method. The results show that the opening of the Beijing-Shanghai high-speed railway has significantly improved the air quality of the transit cities. After changing the time window and polynomial order of breakpoint regression, the opening of the Beijing-Shanghai high-speed railway has a steady effect on improving air quality. In addition, the Beijing-Shanghai high-speed railway has opened channels to improve the air quality of cities along the route, including replacing private cars, reducing the demand for air travel and high-energy-consuming ordinary trains. Wu Chaopeng et al. (2021) adopted the breakpoint regression method, taking the difference between the north and south of the Qinling-Huaihe dividing line in the winter heating system as a breakpoint to study the impact of air pollution on the human capital quality of the company's management. The more serious the air pollution in the city where the headquarters of the listed company is located, the lower the quality of human capital in the company's management. When the company and the city where the company is located can provide good salaries or relatively complete working and living conditions for the management, the negative impact of air pollution on the quality of human capital of management is less significant. In addition, air pollution will significantly reduce the role of the quality of human capital of management on the improvement of the future financial performance of the enterprise. The government and enterprises should actively implement the development concept of "green water and green mountains are golden mountains and silver mountains".

3.3 Fiscal taxation

Zhang Ming (2017) used the breakpoint regression method to empirically test the impact of tax collection and

ISSN 2959-6157

management on the total factor productivity of enterprises, so as to reveal the role of government tax policies in promoting the production transformation of enterprises. The study found that the strengthening of tax collection and management will have a negative impact on the development of enterprises and reduce the total factor productivity of enterprises, that is, the average total factor productivity of enterprises levied by national tax is about 3% lower than that of enterprises taxed by local taxes. Huang Baocong and Tan Guangrong (2021) evaluate the impact of financial pressure on the business performance of micro-enterprises and its mechanism based on Chinese enterprise data and breakpoint regression methods. The study found that financial pressure prompted the tax department to strengthen tax collection and management and improve tax efforts, which significantly reduced the operating performance of micro-enterprises. This effect occurs in the developed areas of the east, non-state-owned enterprises and small-scale enterprises that are more constrained by financing. The mechanism test found that financial pressure may cause enterprises to have a negative impact on their business performance by evading tax burdens, reducing the level of financing constraints and their ability to manage surpluses. This has certain reference value for the business performance of enterprises and the reform of the central government.

3.4 Scientific and technological innovation

Tang Shuxiang and Chen Qi'an (2024) discussed the impact of changes in state-owned holding on enterprise innovation in the context of deepening the reform of mixed ownership and implementing the innovation-driven development strategy. Taking Shanghai and Shenzhen A-share listed companies as a sample, the precise breakpoint regression method was used to analyze the causal effect of changes in state-owned holdings of enterprises on enterprise innovation. The results found that in general, the change of state-owned equity from non-absolute holding to absolute holding will have a significant positive stimulating effect on enterprise innovation; after distinguishing different motivational innovation behaviors, The effect is still significant, and the incentive effect on strategic innovation is obviously stronger than that of substantive innovation. Yang Haodong and Liu Li (2024) selected multiple groups of panels and time series data, and adopted synthetic control method, breakpoint regression and vector autoregression model to examine the induction and reverse effect of the COVID-19 pandemic on technological innovation from the two aspects of medical technology and digital technology. The study found that: 1 The COVID-19 pandemic not only induces the progress of technology in medical-related fields, but also forces the application innovation of digital technology; 2 In the short term, the negative impact of the increase in the intensity of the epidemic on innovation will reach its peak in a month and a half, but the significance of its effect will be rapidly reduced, and the negative impact will gradually converge to the zero horizontal line; 3 At the international level, compared with the United States, Europe and Japan and South Korea, the induction effect and reverse effect are more prominent in China, showing the resilient development of Chinese technology in the context of the epidemic.

4. Conclusion and Prospect

The typical application scenario of the breakpoint regression method is to take individuals configured below or above a certain threshold as the processing group, while individuals on the other side as the control group. The logic of breakpoint regression is that if only the individuals near the breakpoint are focused, the comparison between the processing group and the control group can be ensured to a large extent, so as to estimate the causal effect. However, in the specific application of breakpoint regression, it is generally assumed that there is a linear and polynomial distribution near the breakpoint between the result variable and the configuration variable, but these often depend on the researcher's pre-setting of the function form. Whether the variable relationship is nonlinear is not tested, and the wrong setting of the function form will lead to the estimation In a certain deviation. The first task of the use of breakpoint regression is to find suitable configuration variables and their breakpoints. In traditional breakpoint regression design, configuration variables are generally one-dimensional. Researchers can clarify the breakpoint for themselves through background analysis of the research problem. However, if multi-dimensional configuration variables are considered, the specific location of the breakpoint will change. It is not intuitive and cannot even be determined by manual observation. At this time, machine learning can be used to automatically identify the specific location of the breakpoint. Breakpoint regression has relatively strict hypothetical conditions, such as a jump around the threshold, the configuration variable must be continuous, and the individual cannot accurately control or manipulate the configuration variable to exceed the threshold. In practice, these conditions are often difficult to meet. Through a complex machine learning algorithm, features are synthesized into a tendency score as a configuration variable. The advantage is that the score obtained based on complex algorithms and a large number of behavioral characteristics is a continuous value, and the

individual allocation probability also jumps around the threshold. The score and threshold cannot be observed by the individual, so the individual cannot Manipulate your own behavior. Considering that the core of breakpoint regression is to estimate that if there is no policy impact, the distribution of the result variable should have on the right side of the breakpoint, we can use the data on the left side of the breakpoint to establish a model of the relationship between the result variable and the configuration variable, and then generalize the modeling parameters to the right side of the breakpoint, so as to estimate the right side of the breakpoint. If there is no handling policy, there should be "anti-factual results", and ensuring generalization ability is the advantage of machine learning algorithms, but this idea has not been specifically applied in social science literature.

5. Reference

[1] Dong Huixin, Du Liangsheng. Will the simple VAT collection policy affect the profit margin of enterprises? - Breakpoint regression analysis based on the registration threshold value [J]. China Public Policy Review, 2024,25(01):38-63.

[2] Ma Shuzhong, Congratulatory Song. Can consumer protection policies reduce reverse selection [J]. Economic Theory and Economic Management, 2024,44(06):77-92.

[3] Imbens G W, Kalyanaraman K. Optimal bandwidth choice for the regression discontinuity estimator[J]. The Review of Economic Studies, 2012, 79(3):933-959.

[4] Cattaneo M D, Titiunik R. Regression Discontinuity Designs[J]. Annual Review of Economics, 2022, 14:821-851.

[5] Li Yongtao, Chu Ziqiao, Liu Yutong, etc. Analysis of the impact of changes in the intensity of policy incentives on individuals' participation in endowment insurance [J]. Insurance Research, 2023, (11):92-103.

[6] Wang He, Jia Nan. The health effect of stabilizing the poverty alleviation policy: evidence based on fuzzy breakpoint regression [J]. Economic Problems, 2024,(04):84-91.

[7] Lu Sinuo, Fengjin. The impact of the direct settlement policy of off-site outpatient clinics on the expenditure of medical insurance funds [J]. Financial Research, 2023, (07):97-115.

[8] Wang Zhuo, Zhou Siyang, Yuan Ye. Environmental control, tax contribution and enterprise energy efficiency - take the "10,000 enterprises" policy as an example [J]. Economic Science, 2024,(02):115-140.

[9] Wang Jun, Sun Zhijun. Can key high schools improve students' academic performance - based on a breakpoint regression design study of ordinary high schools in F County [J]. Peking University Education Review, 2015,13(04):82-109+186.

[10] Liu Shenglong, Zhang Zhongwen, Jiang Kezhong. Education returns of key universities: empirical research based on breakpoint regression design [J]. Academic Research, 2023, (04):81-88.

[11] Qin Zhilong, Chen Xiaoguang. Does the opening of high-speed rail improve the air quality of cities along the route? - Analysis based on breakpoint regression design [J]. Environmental Economic Research, 2020,5(02):76-94.

[12] Wu Chaopeng, Li Ao, Zhang Qi. Does air pollution affect the quality of human capital in the company's management [J]. World Economy, 2021,44(02):151-178.

[13] Zhang Ming. Tax collection and management and enterprise full-factor productivity - based on empirical research on Chinese unlisted companies [J]. Journal of Central University of Finance and Economics, 2017, (01):11-20.

[14] Huang Baocong, Tan Guangguang. Is pressure the driving force? The improvement of financial pressure and the improvement of enterprise quality and efficiency [J]. Tax Economic Research, 2021,26(05):70-80.

[15] Yang Haodong, Liu Li. Research on the effect of technological innovation in the context of major public health events [J]. Scientific Research, 2024,42(03):624-636.

[16] Tang Shuxiang, Chen Qi'an. Can the change of state-owned control rights drive enterprise innovation - empirical test based on breakpoint regression [J]. Scientific and Technological Progress and Countermeasures, 2024,41(12):58-69.