Analysis on the Difference in the Initial Examination Results of Postgraduates Based on Regression Discontinuity Design

Yuhao Wang, Yongmei Ding* and Qing Liu
College of Science, Wuhan University of Science and Technology, Wuhan 430065, Hubei, China
Email: dingyongmei@wust.edu.cn

Abstract. The high grades in postgraduates entrance examination are the stepping stone to improve the professional competition in China, which makes the critical influence factors to the initial examination of postgraduates as a hot topic. In order to measure the correlation between the academic performance of undergraduate students and the probability of successful postgraduate entrance examination, and find some effective ways to improve the success rate in the postgraduate entrance examination, we design regression discontinuity model to search the causal correlation between the preliminary entrance exam and CET-6 scores of candidates. Global parameter estimation and local non-parameter estimation were estimated respectively to improve the accuracy of model fitting. In addition, the validity and robustness of the experiment were tested by sensitivity analysis. The results show that the students who get better grades in college are more likely to have a positive effect in the preliminary entrance examination, and make the success rate of postgraduate entrance examination higher.

Keywords. Regression discontinuity design; CET-6; postgraduate entrance examination results.

1. Introduction

Currently, it is evident that an explosive growth in graduate enrolment has taken place in China, 3.41 million people have participated the national postgraduate entrance exam in 2020, with an increase of 510,000 or 17.6 percent over 2019, according to data released by the Ministry of Education. The examination results of undergraduate students are one of the indicators that directly reflect their learning ability.

Regression Discontinuity Design is one of the most widely used methods in microeconometrics. In 1960, Thistlethwaite and Campbell [1] set out to study the impact of scholarships on students’ future academic performance Taking academic performance as the index of scholarship and the score of award standard as the breakpoint, and using the samples near the breakpoint to estimate the processing effect, the idea of breakpoint regression was first proposed. The design was gradually paid attention to in the 1990s, and scholars in various fields continued to make beneficial exploration on the basis of the breakpoint regression thought. Trigonometric kernel estimation was used by McCrary [2], in local linear regression species, and the density function was calculated. Imbens and Kalyanaraman [3] began to explore how to determine the optimal bandwidth in non-parametric estimation. In order to overcome the shortcomings of the above method, such as too large bandwidth and biased confidence interval, Calonico, Cattaneo and Titunik [4] were proposed a better method called CCT to determine the optimal bandwidth.

In this paper, based on the Regression Discontinuity Design and CCT method, the results of the Unified National Graduate Entrance Examination (UNGEE) for graduate students, CET-6 and their
related influencing factors are analyzed. The results show that the score of CET-6 has a certain positive effect on the score of the UNGEE, and the higher score of CET-6 means that they have better learning habits and learning ability, and increase the success rate of postgraduate entrance examination.

2. Model Design
Regression discontinuity design is a kind of quasi-random experiment [5], which defines the probability characteristics of state changes and considers the probability to be a discontinuous function of one or several variables [6]. First, assume that \( d \) represents the effect caused by the state change, and \( X \) represents the key variable determining the state. Then equations (1) and (2) must exist, and \( d^+ \neq d^- \).

\[
\begin{align*}
    d^+ & \equiv \lim_{x \to x_0^+} E[d_i | x_i = x] \\
    d^- & \equiv \lim_{x \to x_0^-} E[d_i | x_i = x]
\end{align*}
\]

In this case, there is a critical value. If the variable sequence is greater than this critical value, the individual is disposed and the state changes. However, when the variable sequence is less than this critical value, the individual does not accept disposal and the state does not change. At the same time, the sample data on both sides of the critical value were used for regression and robustness test. First, check whether other control variables jump at the critical value. If other control variables change significantly at the critical value, then the change of explained variable at the critical value is not only caused by the changed state, and the breakpoint regression is invalid at this time. In addition, it is necessary to test whether the conditional density of the key variables that determine the disposal is continuous. If the conditional density is not continuous, it indicates that the key variable may be manipulated by an individual, thus affecting the overall result.

The application of breakpoint regression design first requires the use of configuration variables and predetermined variables to test the applicability of breakpoint regression design, then carries out global parameter estimation, and finally carries out local nonparametric estimation. In this paper, the CCT method is used to determine the optimal bandwidth for local nonparametric estimation.

3. The Empirical Analysis

3.1. Data Preprocessing and Descriptive Statistics
The College English Test-6 is a national, large-scale, standardized test administered by the Ministry of Education, with universal data. Therefore, this paper uses the score of CET-6 as an indicator to measure the score of college students. In this paper, data were collected in the form of questionnaires. 180 questionnaires were sent out and 161 valid answers were collected, with an effective rate of 89.4%.

As a non-native English-speaking country, the English level of college students in China represents the learning level and learning ability to a certain extent. Considering the main influence factors of college students’ entrance results, here, we take the examinee in school period of CET-6 results as the cause, in order to take an UNGEE results as the fruit, try to analyse the two correlations. According to the empirical data, the former is the total score of all subjects, with a full score of 500, while the latter is 710. The conventional definition of the pass line for certificates is 425. The basic statistical results of valid samples are shown in table 1.

|                | Average | Maximum | Minimum | Standard deviation | Variance  |
|----------------|---------|---------|---------|--------------------|-----------|
| CET-6          | 438.22  | 611     | 336     | 60.19              | 3623.32   |
| UNGEE          | 356.00  | 423     | 205     | 43.92              | 1929.36   |

Table 1. Examination score description statistics.
3.2. Applicability Test and Model Recognition of Regression Discontinuity Design

Firstly, the applicability of the model was tested for the data. In this paper, the pass line of CET-6 was taken as the threshold point, and the probability density of CET-6 scores was obtained by combining the sample data (as shown in figure 1). It can be observed that CET-6 scores have an obvious breakpoint effect at the pass line. Therefore, the test results with CET-6 scores as the driving variable are shown in table 2.

![Figure 1. Probability density of CET-6 scores.](image1)

**Table 2.** Examination score description statistics.

| Sample size | p  v  q | Kernel function                  | Scaling factor |
|-------------|--------|----------------------------------|---------------|
| Left        | 74     | 2 1 3 Trigonometric functions    | 0.45625       |
| Right       | 91     | 2 1 3 Trigonometric functions    | 0.45625       |

In table 2, p is the polynomial order of the point estimate, v is the polynomial order of the reciprocal estimate, and q is the polynomial order of the confidence interval. The left and right orders of the samples are the same, and the scale factor is greater than 0.05. Therefore, it can be considered that the CET-6 score is not manipulated as the driving variable.

According to the relevant theories proposed by Trochim, breakpoint regression is divided into two categories: one is the precise breakpoint regression model [7], that is, the probability of an individual receiving intervention on one side of the breakpoint is 1, and the probability on the other side is 0; The other is fuzzy breakpoint regression, that is, near the breakpoint, the probability of receiving the sensation is monotonically changing. The scatter diagram of the relationship between the driving variables and the intervention was drawn (as shown in figure 2). On the left side of the breakpoint, that is, the probability of the intervention with a CET-6 scores lower than 425 points was 0, and on the other side, the probability of the intervention was 1. It can be seen that the intervention had a clear boundary before and after the breakpoint.

![Figure 2. Scatter diagram of CET-6 and intervention.](image2)
3.3. Regression Discontinuity Design

3.3.1. Relationship Between Driving Variables and Result Variables. The horizontal axis is taken as the driving variable and the vertical axis as the result variable to draw a scatter plot (as shown in figure 3). It can be seen that there is a certain positive correlation between the scores of the preliminary test for postgraduate entrance examination and the scores of CET-6, but the noise is too large, which is not conducive to finding the jump point. Used in this article, based on the R language rdrobust package division determine rdplot function in the enclosure, and smooth after the variable relationship graph (figure 4), as a result, on the left side of the breakpoint divide 20 boxes, the average length of 5, breakpoints on the right side of a divided into 40 boxes, box the average length of 4.65.

![Figure 3](image-url) Scatter diagram of independent and dependent variables.

![Figure 4](image-url) Smoothed curves of independent and dependent variables.

3.3.2. Global Parameter Estimation. Using the characteristics of “discontinuity at breakpoint: in RD design, the following empirical model was established:

\[ Y = \alpha + D \beta_0 + \sum_{k=1}^{K} \beta_k (x - x_0)^k + \sum_{k=1}^{K} D \gamma_k (x - x_0)^k + \varepsilon \]  

(3)

Set \( x \) as the independent variable in the breakpoint regression, \( x_i \) is student i’s CET-6 score, \( x_0 \) is the passing score of CET-6; The dummy variable \( D \) is the student who has a higher score in CET-6. The student with a higher score equal 1, otherwise equal 0; \( Y \) is the dependent variable in the breakpoint regression, and \( \alpha, \beta_0, \beta_k, \gamma_k \) is the parameter to be estimated, and \( \varepsilon \) is the random disturbance term.

First to determine the number of polynomial \( K \), in a polynomial model to join the virtual variables [8], and constantly add configuration variables of multiple items for multiple order model, and compare the fitting results of AIC or regression residuals, the smaller is better, the optimal model selected in this paper according to the AIC criterion is a linear model with interaction terms added:

\[ Y = \alpha + D \beta_0 + \beta_1 (x - x_0) + \beta_2 D (x - x_0) \]  

(4)

The estimated results are shown in table 3.

| variable | Estimate | Std | T values | P values |
|----------|----------|-----|----------|----------|
| \( \alpha \) | 354.989 | 6.670 | 53.221 | 0.0000 |
| \( \beta_0 \) | 0.223 | 8.754 | -0.025 | 0.9800 |
| \( \beta_1 \) | 0.838 | 0.135 | 6.232 | 0.0000 |
| \( \beta_2 \) | 0.561 | 0.157 | -3.569 | 0.0005 |
Goodness of fit is 0.453, the standard residual error is 33.755, F value is 43.418, the reason for the abnormal P value of $\beta_0$ is interpreted as the sample size is too small, resulting in an insignificant single parameter result, but the overall model is significant, parameter estimation for the effect of global estimation precision is very high, however, the larger the interval, the more difficult it is to accurately identify the relationship function between the driving variable and the result variable. Nonparametric estimation can reduce this error by locally fitting a weighted linear or polynomial regression model, so local nonparametric estimation is also required [9].

3.3.3. Local Nonparametric Estimation. In the parameter estimation of local part, the optimal bandwidth needs to be determined first. In this paper, CCT method is used to determine the optimal bandwidth. As shown in table 4, in total, the paper puts forward to use the results of 2 kinds of choosing the way of the optimal bandwidth information, including msetwo and mserd, Calonicon, Cattaneo and Titunik CCT method put forward by the relevant information in 2014, this paper using R language developed by Calonicon, Cattaneo and Titunik rdrobust in function to calculate the optimal bandwidth, and using rdrobust function to manually specify the window width parameters of breakpoints left and right sides of the bandwidth.

The optimal bandwidth of the mserd method on both sides of the breakpoint is equal and 24.992. The bandwidth that should be considered in the sensitivity analysis is 34.153, and using the triangle kernel density function, the optimal bandwidth of the msetwo method on both sides of the breakpoint is not equal, the optimal bandwidth on the left is 18.331, and the optimal bandwidth on the right is 61.509, the bandwidth that should be considered in the sensitivity analysis is 29.917 on the left and 72.709 on the right. Combined with the detailed information of the optimal bandwidth shown in table 4, the processing effect of the intervention using CCT method, and the confidence interval of the intervention is $[-17.713, 38.090]$ at the 95% confidence level, as shown in table 5.

| Method | Bandwidth estimation | Bandwidth deviation |
|--------|----------------------|---------------------|
|        | Left | Right | Left | Right |
| mserd  | 24.992 | 24.992 | 34.153 | 34.153 |
| msetwo | 18.331 | 61.509 | 29.917 | 72.709 |

Table 5. Global parameter estimation results.

| Method   | Coefficient | std  | z       | 95% confidence interval |
|----------|-------------|------|---------|-------------------------|
| mserd    | 10.188      | 14.236 | 0.716  | [17.713, 38.090]         |
| Robustness | 0.597    |       |         | [23.601, 44.299]         |

3.3.4. Sensitivity Test. This section conducts sensitivity analysis on regression and uses the method of replacing virtual breakpoints to test whether the treatment effect still exists around the new breakpoints [10-12]. If the treatment effect exists after the replacement of breakpoints, it cannot prove that the intervention measures studied are effective. This article selects a breakpoint or so two points of 400 and 450, and the two points will be as a virtual breakpoint regression analysis, according to table 6, the replacement of regression coefficient changed little after the breakpoint, can think of at this moment the breakpoint there is no treatment effect, therefore, it is robust to take the score line of CET-6 as the breakpoint.

| Breakpoint | Coefficient | std  | z       | 95% confidence interval |
|------------|-------------|------|---------|-------------------------|
| 400        | 20.663      | 19.355 | 1.069 | [17.252, 58.619]         |
| 450        | 22.125      | 10.841 | 2.041 | [0.877, 43.372]          |
4. Conclusion
This paper collects the scores of 180 students who will take part in the UNGEE in 2020 and the scores of CET-6 during their college years through a questionnaire survey, and explores the relationship between the scores of UNGEE and the scores of CET-6 by using sharp regression discontinuity design. The results show that students with better academic performance are more likely to have a positive effect in the preliminary entrance examination. The robustness of the model estimation results is verified by sensitivity analysis. It shows that the higher score of CET-6 during the undergraduate period means better study habits and learning ability, and the success rate of postgraduate entrance examination increases.

References
[1] Thistlethwaite D L and Campbell D T 1960 Regression-discontinuity analysis: An alternative to the ex post facto experiment Journal of Educational Psychology 51 (6) 309-317.
[2] McCrary J 2008 Manipulation of the running variable in the regression discontinuity design: A density test Journal of Econometrics (142) 698-714.
[3] Imbens G and Kalyanaraman K 2009 Optimal bandwidth choice for the regression discontinuity estimator IZA Discussion Papers 79 (14726) 933-959.
[4] Calonico S, Cattaneo M and Titiunik R 2014 Robust nonparametric confidence intervals for regression discontinuity designs Econometrica 82 (6) 2295-2326.
[5] Lei G, Huang X and Xi P 2015 The impact of transfer payments on urban-rural income gap——Based on fuzzy RD analysis of China’s midwestern counties data Finance & Trade Economics (12) 35-48.
[6] Hausman C and Rapson D 2018 Regression discontinuity in time: Considerations for empirical applications Social Science Electronic Publishing 10 (1) 533-552.
[7] David W, Stefan W and Johannes S 2020 Is small family farming more environmentally sustainable? Evidence from a spatial regression discontinuity design in Germany Land Use Policy 90 (1) 1-7.
[8] Gelman A and Imbens G 2014 Why high-order polynomials should not be used in regression discontinuity designs? Journal of Business & Economic Statistics 37 (3) 255-276.
[9] Liu X 2019 Does the score line effect exist? Regression-discontinuity design based on college entrance examination scores and college graduates’ salary Research in Educational Development 39 (05) 46-53.
[10] Jia N 2020 The impact of retirement on urban household financial assets selection in the aging context——Based on fuzzy regression discontinuity design Statistical Research 37 (4) 46-58.
[11] Yang C and Lin Q 2020 Has the US’ tariffs rise policy caused trade diversion effect on China’s export?——Based on regression discontinuity design Contemporary Economic Management 42 (4) 33-39.
[12] Tao L and Han Q 2020 The causal effects of internal control compulsory disclosure policy: Empirical evidence from regression discontinuity analysis Audit & Economy Research 35 (1) 61-68.