College cost and time to complete a degree

Extension from the paper by

Pietro Garibaldi, Francesco Giavazzi, Andrea Ichino, and Enrico Rettore

Seminar paper

Department of Economics

University of Bern

With

Prof. Dr. Blaise Melly

Tran Thi Bao Tram

Master in Economics

14-105-928

Thi.tran@students.unibe.ch

10.06.2015

Table of content

1. Introduction

It has become a tendency for students around the world to extend their study beyond the normal completion time in recent years. In the United States, the median time to complete PhD was 9 years (1978) and increased to 10.1 years in 2003 (Hoffer and Welch, 2006). The "Pathways to Prosperity" study by the Harvard Graduate School of Education in 2011 shows that just 56% of college students complete four-year degrees within six years and only 29% of those who start two-year degrees finish them within three years. In Europe, the survey by Brunello and Winter-Ebmer (2003) finds that just 31.2% undergraduates in Sweden complete their degrees at least one year later than the required time, and only 30.8% in Italy to 0 in UK.

The reasons for this phenomenon can be increased enrollment among students less academically prepared for college, inability to cope with competing demands of study, family and jobs; and increase in direct cost of education (Bound et al, 2010).

There has been some papers study the effect of financial incentives on completion rates and time to degree in order to find solutions to this problem. However, they found small or no effects in the desired direction. The paper by Garibaldi, Giavazzi, Ichino and Rettore is the first to provide quasi-experimental evidence on this effect. The authors use regression discontinuity design with 4^{th}-order polynominal regression to find the effect of tuition on the probability of late graduation. In order to ensure the validity of this research method, they applied McCrary test (2008) at ten cut-off points for tuition fees and found that this method is valid as there is no jump at the thresholds. However, they found that the monotonicity assumption is violated, thus they simply reported the effect of instrument (official tuition) on the late graduation probability, not of actually paid tuition. That is if the official tuition assigned to students in the last regular year increased by 1,000 euros, the probability of late graduation would decrease by 5.2 percentage points. In addition, they also question whether this decline leads to higher drop-out rate or a reduction in the quality of study and found there is no evidence of these consequences.

In this paper, I replicate the results and test the robustness of these results by using local linear estimation, checked with polynominal regression with restricted data and applying McCrary test (2008) for all threshold together. My motivation in the paper is that local linear method is often considered to be more accurate in estimating the causal effect than polynominal regression and McCrary test should be applied for all thresholds not at each threshold.

This paper is organized as follows. The introduction gives an overview of late graduation problem around the world and the importance of the paper by Garibaldi et al (2012). The second section introduces the econometrics model and identification strategy used by the authors. The next section discusses the data manipulation and

summary of main results. In the section 4, I discuss my extension in econometric methods to this paper to test the robustness of the result and the final section is the conclusion.

2. Regression discontinuity design

Regression discontinuity (RD) research design exploits the fact that some rules are quite arbitrary and therefore provide good quasi-experiments when comparing people affected and not unaffected by the rules. Suppose we want to estimate the effect of binary treatment D on outcome Y and a variable X affect both D and Y. Assume that the impact of X on Y is smooth but the relationship between D and X is discontinuous at a known x_0. Therefore, the discontinuity in the relationship of X and Y can be attributed to D. We have two cases: Sharp RD and fuzzy RD. Sharp RD is used when treatment status is a deterministic and discontinuous function of a covariate X_i.

$$D_i = 1(x_i >= x_0)$$

While fuzzy RD design exploits discontinuities in the probability or expected value of treatment conditional on a covariate (Angrist, 2008). Thus the discontinuity can be used as an instrument variable for treatment status.

$$P[Di = 1|xi] = \begin{cases} g0(xi)\ if\ xi \geq x0 \\ g1(xi)\ if\ xi < x0 \end{cases}, \text{where } g1(x0) \neq g0(x0)$$

The probability of treatment can be written as

$$E\,[D_i|x_i] = P\,[D_i=1|x_i] = g_0(x_i) + [g_1(x_i)-g_0(x_i)]T_i$$

where

$$Ti = 1(x_i >= x_0)$$

We have two approaches of estimating treatment effect: parametric and non-parametric methods. In the parametric method, fuzzy RD can be seen as 2SLS estimation strategy. In the simplest model uses only Ti as an instrument, without the interaction terms, the first stage is

$$D_i = \gamma_o + \gamma_1 x_i + \gamma_2 x_i^2 + \ldots + \gamma_p x_i^p + \pi T_i + \epsilon_{1i}$$

The second stage is then

$$Y_i = \mu + \kappa_1 x_i + \kappa_2 x_i^2 + \ldots + \kappa_p x_i^p + \rho\pi T_i + \epsilon_{2i}$$

One disadvantage of polynominal method is that it provides the global estimates of the regression function over all values of X. This use of data far away from the cut-off

point may be misleading, and this method also depends much on the correctly specified order of the model.

In the non-parametric method, the treatment effect is local "Wald" estimator

$$\mathrm{E}\,[\mathrm{Y1}-\mathrm{Y0}|\,\mathrm{x}=\mathrm{xo}]\;=\frac{\lim\limits_{x\downarrow x0}\mathrm{E}\,[\mathrm{Y}|\mathrm{X}=\mathrm{x}]-\lim\limits_{x\uparrow x0}\mathrm{E}\,[\mathrm{Y}|\mathrm{X}=\mathrm{x}]}{\lim\limits_{x\downarrow x0}\mathrm{E}\,[\mathrm{D}|\mathrm{X}=\mathrm{x}]-\lim\limits_{x\uparrow x0}\mathrm{E}\,[\mathrm{D}|\mathrm{X}=\mathrm{x}]}$$

This method depends on the choice of bandwidth, specifically the small bandwidth would help avoid misspecifying model due to nonlinearities but at the cost of not enough observations. Imbens and Kalyanaraman (2012) suggest a way to choose optimal bandwidth

$$\hat{h}_{IK,p}=\left\{\frac{\hat{V}_{IK,p}}{2(P+1)\hat{B}_{IK,p}^{2}+\hat{R}_{IK,p}}\right\}^{\frac{1}{2p+3}}n^{\frac{-1}{2p+3}}$$

where n is the size of the sample, p is the order of polynominal regression, $\hat{B}_{IK,p}$and $\hat{V}_{IK,p}$ represent nonparametric consistent estimators of asymptotic bias and asymptotic variance. $\hat{R}_{IK,p}$ is introduced to avoid small denominators in moderate-size samples.

One important feature of regression discontinuity design is that the density of the running variable X must be continuous at the cut-off points, or else this will suggest manipulation around the thresholds. This feature can be tested formally by McCrary test (2008). In addition, one may include some covariates in the model to reduce bias. However, there should be no jump in the relationship between covariates and running variable X, if not, one may cast doubt on identifying assumption since these covariates should not be affected by treatment D.

At Bocconi university, upon enrollment, students are assigned to one of twelve tuition levels based on their family income. However, the actually paid tuition is different from the official tuition assigned since Bocconi university administration reserves the right to make its own reassessment of a family's ability to pay through the income tax declaration and further inquiries. The students just below and above these thresholds are identical in terms of pre-Bocconi characteristics but have to pay different tuition fee. Garibaldi et al (2012) thus implement fuzzy RD which compares students' rate of late graduation with family income immediately above or below each discontinuity threshold by parametric method to estimate the treatment effect of the tuition fee on the probality of late graduation. The framework they use is based on Hahn, Todd, and van der Klaauw (2001).

The two stages IV regression are as followed

$$(1)\ \tau^p=\alpha\tau^t + \gamma_a + g(Y) + \delta X + \epsilon$$

$$(2)\ F=\beta\tau^t + \gamma_a + g(Y) + \delta X + \epsilon$$

where Y is the student's real income, g(Y) is the fourth-order polynominal in Y, X is a vector of pretreatment characteristics of students, γ_a is academic year-specific effects, τ^t is the official tuition that the students should pay according to assignment rule (here the instrument variable), and τ^p is the tuition paid, which is the treatment. The inclusion of characteristics X in the equation is to reduce bias and this should not affect the estimate of β if the assignment to treatment is orthogonal with respect to pretreatment characteristics at each threshold. The authors also include year effect γ_a since the composition of the pool of Bocconi students changed over time with respect to some observables relevant to the outcome. Therefore, conditional on time effect will make students just above the threshold comparable to those just below (Garibaldi et al, 2012).

3. Data manipulation

The empirical analysis is based on the administrative data from Bocconi University in Milan, Italy during the period 1992-2002. The dataset is considered to be very informative with 80% of Bocconi students complete their degree in more than four years.

Table 1: Descriptive statistics by late graduation status

LABELS	Conditional on being		
	On time	Late graduation	Of the total
% of the 10,216 fourth-year students from 1995 to 2002 who:			
Female	47.75	40.39	41.88
Family of origin outside Milan	37.32	40.22	39.63
Graduated from high school with top grade	36.50	23.33	25.99
In top high school tracks	73.15	65.89	67.36
Income before Bocconi in euros	43,881	38,966	39,958
Total	20.19	79.81	100,00

Based on statistics by Garibaldi et al (2012)

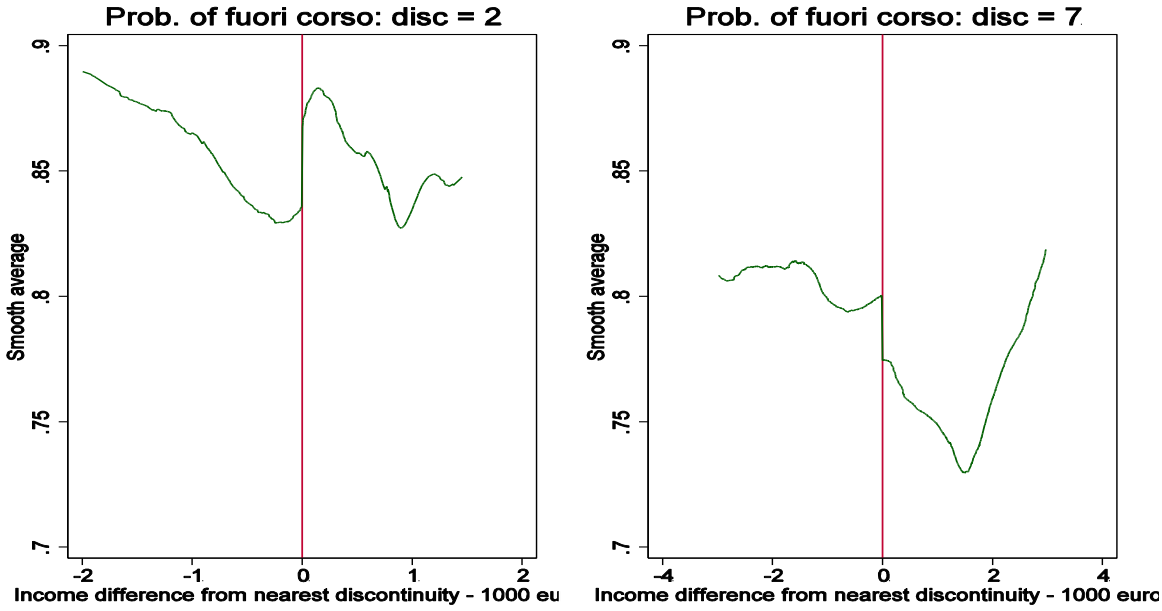
These statistics suggest that the rate of late graduation is correlated with lower ability and education performance. Particularly, the proportion of students who were graduated from high school with top grade and in top high school tracks are higher for students graduated on time than late graduation. The family origin outside Milan does not matter much, while the fractions of on-time graduation is higher for female.

The data is restricted to fourth-year students. The reason for this is because the students who enrolled in the fourth regular year of the program do not know the tuition they have to pay if they graduate late since they are uncertain about their families' income and the future possible adjustments of the tuition structure by Bocconi administration each year. Therefore, fourth-year official tuition in a given year is a good predictor of official tuition in the following year.

The analysis is then restricted to those whose family income differs by no more than ±3000 euros from the discontinuity threshold in order to get better estimation for the polynominal regression function.

The figure 1 shows that there are discontinuities in the rate of late graduation at some threshold over income, which suggests treatment effect may be used with regression discontinuity.

Figure 1: Late graduation on the running variable income



(based on Garibaldi et al, 2012)

4. Results

The fourth-order polynominal regression with and without covariates results are given in the table 2. It is shown that official tuition has strong correlation with actually paid tuition at 52.8% (model 1) and an increase of 1,000 euros of official tuition would reduce the rate of late graduation by 5.2%. The results are similar in other models with covariates. Income before studying in Bocconi is statistically insignificant at 5% level, which suggests that there might be no manipulation around the thresholds. Since if

family can alter their declared taxable income to be assigned to lower tuition level, there will be a concentration of probability mass below the threshold and the discontinuity research design would become invalid. Other covariates including female, family origin, high school profile are all significant, though they do not affect the coefficient interested in a relevant way. This means the students' characteristics are balanced around the thresholds.

Table 2: Regression discontinuity estimates of the effects of official tuition

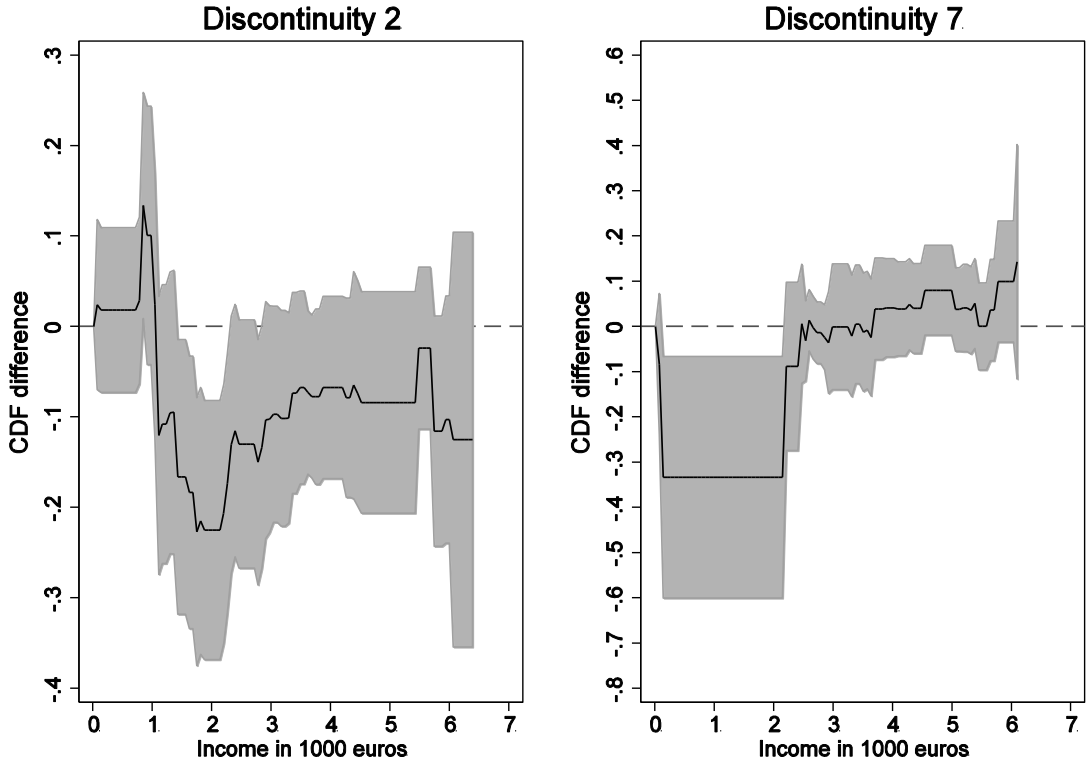
	Paid tuition 1	Late graduation 2	Paid tuition 3	Late graduation 4	Paid tuition 5	Late graduation 6
Official tuition	0.528	-0.052	0.531	-0.054	0.562	-0.047
	(0.055)	(0.023)	(0.055)	(0.023)	(0.060)	(0.025)
Female			0.010	-0.031	0.008	-0.031
			(0.029)	(0.010)	(0.029)	(0.010)
Family of origin outside Milan			-0.003	0.029	-0.002	0.029
			(0.028)	(0.010)	(0.028)	(0.010)
Highschool grade			-1.564	-0.660	-1.564	-0.662
			(0.136)	(0.045)	(0.137)	(0.045)
Highschool type			0.071	-0.054	0.071	-0.054
			(0.031)	(0.010)	(0.031)	(0.010)
Income before Bocconi			0.008	-0.001	0.008	-0.001
			(0.001)	(0.000)	(0.001)	(0.000)
Constant	2.571	0.870	3.844	1.500	14.074	12.185
	(0.528)	(0.161)	(0.539)	(0.165)	(43.724)	(11.421)
Academic year dummies	Yes	Yes	Yes	Yes	Yes	Yes
Same g(Y) for all thresholds	Yes	Yes	Yes	Yes	No	No
Different g(Y) for low, medium, and high threshold	No	No	No	No	Yes	Yes
Observations	6,985	6,985	6,790	6,790	6,790	6,790
R2	0.529	0.0371	0.545	0.0695	0.545	0.0696

(based on Garibaldi et al, 2012)

To test the validity of the research design, Garibaldi et al (2012) apply McCrary test (2008) to test the null hypothesis of continuity of the density of the running variable at the discontinuity point (here the income). If individuals can manage to manipulate the running variable, one might expect to see a discontinuity in the density at the cut-off point. By this test, the authors find that the t-statistics of the tests at each threshold are

all insignificant at 5% significant level, and thus cannot reject the hypothesis that the density of covariate is continuous at the cut-off points. The t-values at the ten cut-off points are 0.30, 0.70, 0.90, 1.10, 0.30, -0.41, 1.29, -0.38, -0.20, -0.62 respectively. However, we will see below that this test should be applied for all the thresholds together, not at each threshold as was exploited here.

Figure 2: Test of monotonicity: cdf



However, in order to interpret causal effect, we need to have the same assumptions as in IV framework (Hahn, Todd, and van der Klaauw, 2001). As can be seen from the table, the effect of official tuition on paid tuition is significant, thus ensures the first stage assumption. The monotonicity assumption implies that no one would have to pay a lower actual tuition if her official tuition shifts from low to high, and at least one student should pay a higher tuition in this event (Garibaldi et al, 2012), which means $\tau^{p}_{h} >= \tau^{p}_{l}$. The authors then test the inequality in that the cumulative distribution function (cdf) for those in a right neighborhood of the cut-off point should not be above the cdf for those in the left at any value of its support. The figure 2 illustrates that this hypothesis is rejected at these threshold, thus suggests the lack of monotonicity. The reason for this might be due to the reassessment of families' ability to pay by Bocconi university administrators, by which the actually paid tuition does not correspond exactly to the official tuition assigned to students.

Therefore, we cannot identify the causal effect, but only the effect of official tuition on the rate of late graduation.

5. Robustness tests

5.1. Local linear estimation

While replicating the results by Garibaldi et al (2012) by using IV regression, I find that the adjusted R2 is very poor, -5.5% and -1.2% for without and with covariates respectively (columns 1-2), which means the models failed to capture the variation of the observations. Moreover, as discussed in Gelman, Imbens (2014), it is argued that the estimators for causal effects based on regression discontinuity controlling high-order polynominals of forcing variables can be misleading. That is because we do not have methods for choosing the order of polynominal that is optimal for a good estimator of the causal effect while the estimates can be highly sensitive to the degree of polynominal. Gelman, Imbens (2014) also argue that the confidence intervals by high-order polynominal are too narrow, which may result in misleading inference.

Table 3: Regression discontinuity estimates of paid tuition on late graduation

	2SLS Polynominal of order 4th +/-3		+/-1.5		Local linear +/-1.5
	1	2	3	4	5
Actually paid tuition	-0.0985**	-0.101**	-0.0603	-0.0624	-0.0278
Female		-0.0298***		-0.0346***	
Family of origin outside Milan		0.0283***		0.0349***	
Highschool grade		-0.818***		-0.778***	
Highschool type		-0.0469***		-0.0436***	
Income before Bocconi		5.08e-06		-0.000151	
Constant	0.935***	1.683***	0.728**	1.428***	
Academic year dummies	Yes	Yes	Yes	Yes	
Observations	6,985	6,790	4,496	4,392	6,985
R2	-0.055	-0.012	0.003	0.043	

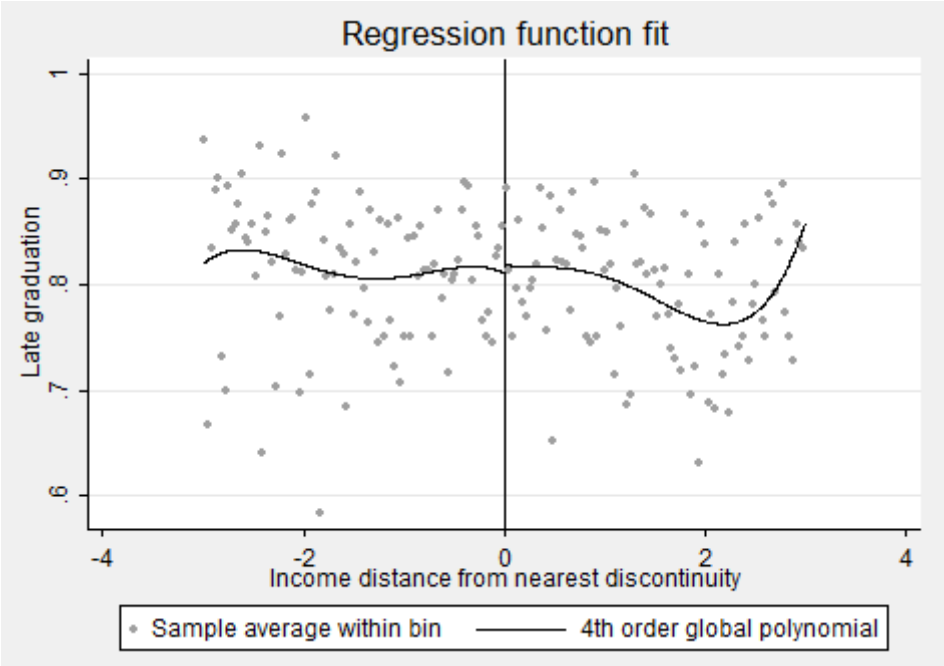
*** p<0.01, ** p<0.05, * p<0.1

Therefore, in this paper the local linear estimation using fuzzy RD is applied to test the robustness of the result. The income is normalized to the nearest threshold with the bandwidth of 1.5, which is optimally chosen by Imben and Kalyanaraman (2012). I found that the treatment effect is insignificant at 10% level, which means that the actually paid tuition has no effect on the probability of late graduation (column 5. To test this result on polynominal regression, I restrict the sample to just 1.5 thousand euros from the threshold and get the same result of no treatment effect and the goodness of fit adjust R2 are now better at 0.3% and 4.3% without and with

covariates. Thus the result found by Garibaldi et al (2012) is unconvincing and one explanation for this might be that the sample is too large for the model to be correctly specified.

The following figure illustrates the relationship between rate of late graduation and income which is normalized at the thresholds using 4^{th} order local polynominal regression. This also suggests there seems to be no discontinuity in this relationship, thus we may not find any effect of tuition on the late graduation probability.

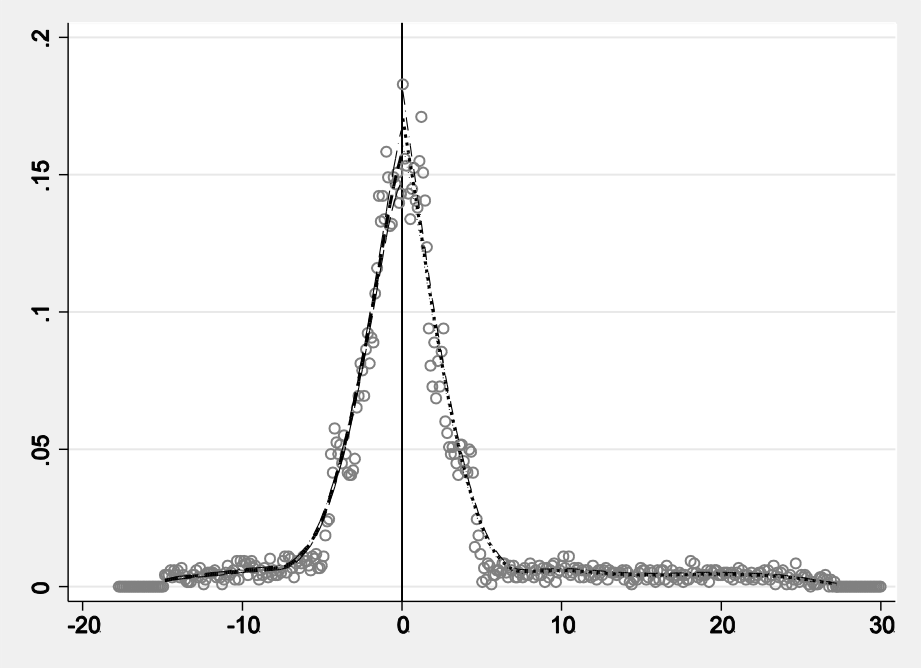
Figure 3: Late graduation rate over income



5.2. Test the continuity of the density with McCrary test for all thresholds

To test the validity of discontinuity regression design, Garibaldi et al (2012) carried out McCrary (2008) test at ten cut-off points. This test checks if there is a jump in the density of the running variable (here the income). If there is, it means that people can manipulate the running variable around the threshold, and the regression discontinuity design becomes invalid. However, McCrary test is used for all thresholds combined, not at each threshold as Garibaldi et al (2012) did in their paper. Thus McCrary (2008) test is carried out again for the whole thresholds together and t-value amounts to 1.93 so the null hypothesis that there is no discontinuity is insignificant at 5% but can be rejected at 10% significant level. Therefore, we may doubt the manipulation of income around the thresholds.

Figure 4: McCrary test for all thresholds combined



6. Conclusion

The paper by Garibaldi et al (2012) suggests that regression discontinuity design can be exploited to estimate the effect of tuition on the late graduation, which has been a widespread phenomenon around the world. By using fourth-order polynominal regression, the authors find that increasing official tuition of 1,000 euros can help reduce the probability to prolong study by 5.2%. However, they reason that the lack of monotonicity prevent us from getting the treatment effect of tuition paid on the rate of late graduation.

This paper challenges this conclusion by first looking at the regression results and find that the goodness of fit for the polynominal is very bad, which suggests incorrect specification of the model. Then the local linear regression is applied and it turns out that the treatment effect is statistically insignificant at all levels. The result is checked again with the polynominal model but restricting the income to just 1.5 thousand euros from the threshold and get the same result of no treatment effect. In addition, McCrary test is carried out for all thresholds together and the null hypothesis of continuity in the density of running variable cannot be rejected at 5% level but become significant at 10%, thus the validity of the discontinuity research design is weakly accepted.

List of tables

List of figures

References

Angrist, Joshua D. and Joern Steffen Pischke, "Mostly harmless econometrics: An empiricist's companion", *Princeton University Press* (2008).

Bound, John, Michael F.Lovenhein, and Sarah Turner, "Increasing Time to Baccalaureate Degree in the United States", NBER Working Paper No.15892 (2010)

Brunello, Giorgio, and Rudolf Winter-Ebmer, "Why do students expect to stay longer in college? Evidence from Europe", Econometrics Letters 80 (2003), 247-253

Colonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik, "Robust data-driven inference in the regression- discontinuity design", *The Stata Journal* (2014), 909-946.

Cattaneo, Matias, Luke Keele and Rocio Titiunik, and Gonzalo Vazquez Bare, "Identification in regression discontinuity designs with multiple cutoffs" (2015).

Garibaldi, Pietro, Francesco Giavazzi, Andrea Ichino and Enrico Rettore, "College cost and time to complete a degree: Evidence from tuition discontinuities", *The Review of Economics and Stattistics* (2012).

Hahn, Jinyhong, Petra Todd, and Wilbert van der Klaauw, "Identidication and estimation of treatment effects with a regression discontinuity design", *Econometrica* 69 (2001), 201-209.

Imbens, Guido, and Thosmas Limieux, "Regression discontinuity designs: A guide to practice", *Journal of Econometrics* 142 (2008), 615-635.

McCrary, Justin, "Manipulation of the running variable in the regression discontinuity design: A density test", *Journal of Econometrics* 142 (2008), 698-714.