Economics 142
Fall 2016
Problem Set 7

The data set rd.csv contains student level data for 112,008 students who finished high school and were eligible to enter college. In the specific country where the data orginate, students write a standardized test at the end of high school, called the "PSU" test. Their scores on this test, plus high school GPA, determine which colleges they can get into. Students who score at least 475 points on the PSU test are also eligible for a loan from the government for college costs, while students who score less than 475 points cannot receive the loan. In this problem set we will use regression discontinuity methods to analyze the effect of the loan program on the probability of college entry.

The variables on the data set are:

- psu = PSU test score (ranges from 300 to 700; the scores are numbers like 300.0, 300.5, 301.0, 301.5, 302.0....)
- over475 = 1 if PSU score is 475 or higher
- entercollege = 1 if student entered college
- hsgpa = high school GPA (scored from 0 to 70, 70 is "perfect")
- privatehs = 1 if student went to a private high school
- hidad = 1 if father has more than a high school education
- himom = 1 if mother has more than a high school education

1. Contruct the mean values of entercollege, hsgpa, privatehs, hidad, and himom for each integer range of PSU (e.g., for scores from 300 to 300.99), from 301 to 301.99, etc). (In other words, group scores of 300 and 300.5 together, 301 and 301.5, etc). This is sometimes called "collapsing" the data. One way to do this in R is by using the "aggregate" function.

Show plots of these mean values as a function of PSU. You should see a jump in entercollege at 475 points, but relatively smooth values of the other variables.

2. Next you will fit "local linear" regressions using different "bandwidths". To do this, you will regress one of the outcome variables, y on the following $X's$:

-$constant$
- psu
- $over475$
- a 4th variable = $X_4 = (psu - 475) \times over475$

If you fit the model:

$$y = \beta_1 + \beta_2 psu + \beta_3 over475 + \beta_4 X_4 + \varepsilon \tag{1}$$

the coefficient $\beta_3$will measure the "jump" in y at 475 points, the slope of the line to the left of 475 will be β_2,and the slope to the right will be $\beta_2 + \beta_4$.

a) using the "collapsed" data from part 1, which has 1 observation per point of psu, and a bandwidth of 10 points on each side of the 475 cutoff, fit model (1) for $y \in \{entercollege, hsgpa, hidad, himom\}$.

(HINT: what this means is that you fit the regression model (1) to the collapsed data for the subset of data with $465 \leq psu \leq 484$. This data set will have 10 observations on scores less than 475, and 10 observations on scores of 475 or higher)

b) repeat part (a) using a bandwidth of 20 points. Do you find that the estimated jumps are similar for all four variables?

c) for every bandwidth from 5 to 50, develop a graph to show the estimate of β_3 when the outcome (y) is "entercollege".

d) Now fit the same kind of model, but using as the "y" variable the parental education variables "hidad" and "himom". Use bandwidths of 10 and 20. Do you find any evidence that the family background variables "jump" at the 475 point cutoff?