Exercise 1: RDD - paper discussion

In the lecture you discussed the theoretical framework of the regression discontinuity design (RDD). With this in mind, read the following paper:

Carpenter, C., Dobkin, C. (2009). The Effect of Alcohol Consumption on Mortality: Regression Discontinuity Evidence from the Minimum Drinking Age, *American Economic Journal: Applied Economics*, 1, 164-82.

Please consider the following components of the paper, discuss them critically and be prepared to present them in class:

- research question
- contribution
- methodology (type of RDD and underlying details)
- key identifying assumption
- data
- results / findings
- interpretation of the results / implications / policy relevance
- limitations of the paper

Exercise 2: RDD - practical application

In the following exercise you will estimate the effects of the MLDA on alcohol consumption. The dataset NHIS.data contains a sample drawn from the *National Health Interview Sample* adult-files 1997-2007 which Carpenter et al. (2009) used in their study. The data set includes several variables such as:

- $days\_21$: days to the 21^{st} birthday.
- age_yrs: age in years.
- $drinks\_alcohol$: whether someone drinks alcohol.
- $perc\_days\_drink$: percent of days someone drinks.
- and further self-explaining control variables.

(a) Create a carefully labeled graph showing the relationship between the running variable ($days\_21$) and whether or not someone drinks alcohol ($drinks\_alcohol$). Also highlight the MLDA threshold. You will need to make age cells (similar to Fig.1 of the paper). Experiment with different bin sizes. Which jump do you observe at the MLDA? Given the pre-MLDA drinking level and the potential jump that you observe at the MLDA, calculate the precentage increase.

(b) You are given the following model:

$$y_{ai} = \alpha + \pi D_{ai} + g(age_i) + v_{ai}$$

with y_{ai} indicating the binary outcome whether someone drinks alcohol or not. $g(age_i)$ is a flexible polynomial of one's age in days which takes different forms on either side of the age cutoff. D_{ai} is a binary variable indicating whether someone's age is above or below the MLDA. Estimate the previous model using different order polynomials going from least to most flexible (i.e. form linear to quadratic, cubic, and quartic). Your sample should include individuals aged three years before and three years after the MLDA. Use robust standard errors and compare the different specifications. Do the results prove robust? How do they compare to the graph that you previously plotted? Which coefficient contains the pre-MLDA drinking level? Given the pre-MLDA drinking level and the potential jump at the MLDA, calculate the precentage increase for your most preferred specification.

(c) Check if the regression estimates are robust to different bandwidths. In the previous question the bandwidth was defined to range from three years before and three years after the MLDA. In this exercise your are asked to reduce the bandwidth stepwise by 60 days (i.e. 30 days on each side of the cut-off). Using a quadratic polynomial run separate regressions using the different bandwidths. Then, plot the π coefficients that you obtain from each regression, that is, plot the bandwidth on the x axis and the estimates of the coefficients including the confidence interval on the right. Do your findings prove robust?

(d) Check if there is any evidence of sharp changes in the characteristics at the MLDA. Make a regression table that presents these estimates. You can use a quadratic polynomial.