Exercise 1: Non-Parametric Methods

In this task we focus on methods which do not make assumptions regarding the functional forms. You will use the *NHIS* dataset that you already obtained for tutorial 6.

(a) Create a histogram to plot the distribution of the variable *days_21*, which indicates the number of days above or below the 21^{st}birthday. Why might a histogram not be optimal to obtain the underlying distribution?

(b) Plot kernel density estimates for *days_21* using *kdens*. Compare various kernels as well as bandwidths and explain the differences. In a second step plot within the same figure kernel densities of the variable *days_21* but separately for the two different values of the binary variable *drinks_alcohol*. Compare them and describe what you observe.

(c) For age bins of 30 days plot the means of *drinks_alcohol*. Why might a parametric approach not be suitable in these cases?

(d) Add to your plot from exercise (c) for the area below and above of the cut-off a linear fit as well as a kernel-weighted local polynomial fit (*lpoly*).

(e) Use the *rdrobust* command to implement a local polynomial regression-discontinuity estimation. Vary the bandwidth as well as the kernel function (e.g. triangular, epanechnikov, etc.). Compare your estimates with the graphs of exercise (d) and your results from tutorial 6 where you used parametric estimation techniques.

Exercise 2: Matching

For this exercise you are given a data extract of the SOEP-student version which can be used for educational purpose. The sample has already been adjusted and restricted to employed individuals aged 20 to 60. Outcome variable y indicates the cross hourly wage. Treatment indicator D contains the information whether someone attended college or not. There are further control variables included which are summarized by matrix X.

(a) Regress the treatment indicator D on all control variables (i.e. X) using a logit model. In a second step predict for each individual the respective propensity score. Plot the distribution of the propensity score for treatment and control group. What can you say about the common support? How many observations within the treatment/control group are located on or off the common support.

(b) Use the *psmatch2* command to conduct a NN-1 (one nearest neighbor) matching with replacement. Interpret your outcomes. Which advantages and disadvantages do you face if you include more neighbors?

(c) Use the *psgraph* command to plot the distribution of the propensity score that you obtained from sub-exercise (b). Do you observe any difference compared to your graph from exercise (a).

(d) Use your results form exercise (a) to compare it with the $2 * 2$ table (showing treatment assignment vs. common support) which is displayed if you execute the $psmatch2$ command.

(e) Discuss briefly potential disadvantages of the nearest neighbor matching? Which alternative matching approaches do you suggest?

(f) Do the matching assumptions hold in the present exercise and do you identify a causal effect?