Yale Course

Problem Set #1

Please feel free to work in groups on the problem set.

1. (a) Write a program to calculate the kernel density estimator for one variable with a fixed bandwidth. You may use the biweight kernel equal to

$$\begin{aligned} K(s) &= (15/16)(s^2-1)^2 \text{ for } |s| \leq 1 \\ & \quad 0 \text{ else} \end{aligned}$$

or a different kernel function (such as the normal density kernel).

(b). Generate 500 random numbers from a one dimensional uniform distribution (over range 0 to 1) evaluate the performance of the density estimator using alternative bandwidth fixed choices (0.05,0.1,0.2) at a range of points (for example, {0,0.1,0.2,...,0.8,0.9,1}). Can you tell from the estimates that the estimator is less accurate near the boundaries?

2. Prove mean-square consistency of the estimator

$$f'(x_0) = \frac{1}{nh_n^2} \sum_{i=1}^{n} K'\left(\frac{x_0 - x_i}{h_n}\right)$$

at a point x_0, under random iid sampling. State any assumptions that you need about the kernel function, the bandwidth, the sample size, and about the existence of derivatives of the function f.

3. Suppose you are interested in estimating the Average Treatment Effect (ATE(X)) parameter that was defined in class for the model:

$$Y_1 = X\beta_1 + U_1 \tag{1}$$

$$Y_0 = X\beta_0 + U_0 \tag{2}$$

where we observe

$$Y = DY_1 + (1 - D)Y_0 \tag{3}$$

with $D = 1$ if a person participates in a program, $D = 0$ otherwise.

(a) What are the identifying assumptions on the unobservables that are needed to estimate $ATE(X)$ by a cross-section regression estimator, a before-after regression, and a difference-in-difference regression?

(b) How do the assumptions change if instead you are interested in the average effect of treatment on the treated $TT(X)$?

4. The Rosenbaum and Rubin (1983) Theorem discussed in class shows that if

$$Y_0 \perp\!\!\!\perp D \mid X \tag{RR}$$

Then,

$$Y_0 \perp\!\!\!\perp D \mid P(X)$$

where $P(X)$=$\Pr(D = 1|X)$ is the "propensity score". Following the argument of the proof given in class, show that (RR) implies

$$Y_0 \perp\!\!\!\perp D \mid Z(X),$$

where $Z(X) = \ln(\frac{P(X)}{1-P(X)})$. Thus, matching could proceed on the log-odds ratio instead of on the propensity score.

5. The JTPA data (available on the internet and described below) include two groups - a randomized-out experimental control group and a nonexperimental comparison group. Neither of these groups received the program, so when an estimator is applied, it should give an estimated treatment effect of zero. Any deviation from zero is interpretable as bias associated with the estimator. Implement the following estimators (without controlling for X variables):

- Cross-section Estimator (in quarters 4,5,6)
- Before-After Estimator (symmetric differences, using quarters (4,-4),(5,-5) and (6,-6)
- Difference-in-Differences Estimator (again, take differences symmetrically)
- Nearest Neighbor (one nearest neighbor matching estimator) without imposing common support condition
- Nearest Neighbor estimator imposing the common support condition

What is the bias associated with each of the estimators? What difference does imposing common support make to the matching estimator ? (Note: for a,b, and c you can try experimenting with different sets of regressor control variables.)

Information on Accessing the JTPA dataset

The dataset is available from the web page located at
http://athena.sas.upenn.edu/˜petra/class222/jtpa.asc.

1. download the file called jtpa.asc
2. It contains the following variables

id - person id number

qtr - quarter relative to random assignment. The data cover 18 months (6 quarters) prior to and 18 months (6 quarters) after the date of random assignment. For example, quarter = -1 refers to earnings in the quarter prior to random assignment. Quarter = 6 refers to earnings 18 months after the date of random assignment. (The date of random assignment would be = 0, but the dataset does not contain earnings for the month of random assignment.)

d - indicator for whether in the experimental group (1=in experimental control group, 0 =in nonexperimental group (the ENPs))

earn - monthly earnings (in dollars)

aged - age category indicator

edlt10 - education less than 10 years

ed1011 - education 10 to 11 years

ed12 - education 12 years

ed1315 - education 13-15 years

edgt15 - education 15 or more years

totexp - total labor market experience

p - propensity score (=Pr(D=1|X))