Bergen Course

Problem Set #2

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

1. Computer Problem #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. Computer Problem #2

   (a) Draw 500 random variables $x_i$ from a N(0,4) distribution. Then draw another 500 random variables $\varepsilon_i$ from a N(0,2) distribution. From these random variables, generate the random variable

   $$y_i = g(x_i) + \varepsilon_i,$$

   where $g(x_i) = 2x_i^2$.

   (b) Next, write a program to estimate the function $g(x_i)$ using kernel regression. (You can run a weighted regression on a constant, with the weights given by $\frac{K(\frac{x_0 - x_i}{h_n})}{\sum_{i=1}^{n} K(\frac{x_0 - x_i}{h_n})}$, or you can program up the formula $g(x_0) =$

$\frac{\sum_{i=1}^{n} y_i K(\frac{x_0 - x_i}{h_n})}{\sum_{i=1}^{n} K(\frac{x_0 - x_i}{h_n})}$ directly). Evaluate the regression function at at least 10 values between -10 and 10.

(c) Plot the true $g(x_i)$ function ($=2x_i^2$) against the one one you estimated on one plot to see how close the estimated one is to the true one.

(d) Explore how the estimated function changes as you vary the bandwidth choice and as you vary the number of data points.

3. 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 $f$.