Problem Set I

Econometrics I - FGV EPGE

Instructor: Raul Guarini Riva

TA: Taric Latif Padovani

Problem 1

Let X be a scalar random variable with density $f(x)$. Let $K(\cdot)$ be a symmetric second-order kernel. For a given point x in the interior of the support of $f(\cdot)$, define the density estimator as

$$\hat{f}_n(x) \equiv \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right),$$

where $h > 0$ is a bandwidth parameter and $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) random variables with density $f(\cdot)$.

This exercise will show you how to ensure that $\hat{f}_n(x) \xrightarrow{p} f(x)$ as $n \to \infty$, $h \to 0$, and $nh \to \infty$. This is the same asymptotic framework as in the slides.

a) Show that $\hat{f}_n(x) \geq 0$ for all x, and that $\int_{-\infty}^{\infty} \hat{f}_n(x)dx = 1$ for all n.

b) Assume from now on that f is continuous at x. Show that $\mathbb{E}[\hat{f}_n(x)] = f(x) + o(1)$.

c) Show that $\text{Var}(\hat{f}_n(x)) = \frac{1}{nh} \cdot f(x)R(K) + o\left(\frac{1}{nh}\right)$, where $R(K) = \int_{-\infty}^{\infty} K^2(u)du$.

d) Argue that these results imply that $\hat{f}_n(x)$ is consistent for $f(x)$.

e) Now, assume that f is twice continuously differentiable at x. Show that

$$\mathbb{E}[\hat{f}_n(x)] = f(x) + \frac{h^2}{2} f''(x)R(K) + o(h^2)$$

f) Explain in words how the local convexity of f might (or might not) affect this finite-sample bias.

Hint: A very useful resource for this question is Chapter 17 from *Probability and Statistics for Economists* by Bruce Hansen.

Problem 2

Let X be a continuous random variable with density $f(\cdot)$, which is positive everywhere. Suppose the true regression function is linear, $m(X = x) = \alpha + \beta x$, and we estimate the function using the Nadaraya-Watson estimator. Assume all regularity conditions you need.

a) Calculate the bias function $B(x)$.

b) Suppose $\beta > 0$. For which regions is $B(x) > 0$ and for which regions is $B(x) < 0$?

c) Now suppose that $\beta < 0$ and re-answer the question.

d) Can you intuitively explain why the Nadaraya-Watson estimator is positively or negatively biased in these regions?

Problem 3

This is an empirical question based on Karlan and Zinman (2008, Econometrica). You will find the paper online on the class Github repo. The data used in the paper is also available there.

a) What is the main research question in the paper? What is the most striking finding? Answer in just a few sentences.

b) Your goal will be to estimate $\mathbb{P}(applied = 1|offer4 = x)$. Note that `applied` is a binary variable, while `offer4` is continuous. These are the only variables you will need. Notice that

$$\mathbb{P}(applied = 1|offer4 = x) = \mathbb{E}[applied|offer4 = x]$$

c) Use a Gaussian kernel to estimate this probability and show a plot of your estimates for a range of values of `offer4`. Do this with three different bandwidths:

- Silverman's rule of thumb: $h = 1.06 \cdot \hat{\sigma} \cdot n^{-1/5}$, where $\hat{\sigma}$ is the standard deviation of `offer4` and n is the number of observations;
- A value much *smaller* than that;
- A value much *larger* than that;

d) Do the same with the Epanechnikov kernel, using the same bandwidths as before.

e) Compare the results of the two kernels qualitatively.