Problem Set I

Econometrics I - FGV EPGE

Instructor: Raul Guarini Riva

TA: Taric Latif Padovani

Problem 1 – (points: 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

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

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 – (points: 1)

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 – (points: 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.

Problem 4 – (points: 2)

Let $X_i, i = 1, \ldots, n$ be an i.i.d. sample of observations with distribution $U[0, \theta]$. A natural estimator (in fact, the MLE) of θ is $X_{(n)}$, where

$$\min\{X_1, ..., X_n\} = X_{(1)} \leq \cdots \leq X_{(n)} = \max\{X_1, ..., X_n\}$$

denote the ordered values of the data. These statistics are sometimes referred to as the order statistics of the data. Consider the following root:

$$R_n = n(X_{(n)} - \theta),$$

a) Show that $R_n \leq 0$.

b) Let $J_n(x, P) = P(Rn \leq x)$ be the cdf of the root R_n above. Show that $J_n(x, P)$ converges in distribution to $J(x, P) = P(-\theta X \leq x)$, where $X \sim \exp(1)$, i.e.,

$$J(x, P) = \begin{cases} e^{x/\theta}, & x \leq 0 \\ 0 & \text{otherwise} \end{cases}$$

Hint: for every $r \in \mathbb{R}$, remember that $\left(1 + \frac{r}{n}\right)^n \xrightarrow{n\to\infty} e^r$.

c) We will now prove that the bootstrap will not work in this case. Consider a non-random sequence of probability distributions $\{P_n\}_{n=1}^{\infty}$ such that each distribution P_n puts equal mass on n distinct points in the interval $[0, \theta]$.

For each n, let $\{X_{i,n}\}_{i=1}^{n}$ be an i.i.d. sample from P_n. Denote by $X_{(n),n}$ the maximum of this n-th sample. Also, denote by $\theta(P_n)$ the maximum of the support of P_n.

Now, show that $\lim_{n\to\infty} P\left(n(X_{(n),n}) - \theta(P_n) \leq -\epsilon\right) \leq e^{-1}, \forall \epsilon > 0$.

d) Use this result to show that $J_n(x, P_n)$, which is the distribution of the bootstrapped root under the measure P_n does *not* converge to $J(x, P)$.

Hint: pick $\epsilon < \theta$.

e) Argue, using the last item ,that the bootstrap does not work in this case. Think about the relationship between the empirical distribution and the sequence of distributions $\{P_n\}_{n=1}^{\infty}$.

Problem 5 – (points: 3)

This question is about the bootstrap. Suppose that $y_i = \alpha + \beta x_i + \epsilon_i$, where $\alpha = 2$, $\beta = 5$, $\epsilon_i | x_i \sim N(0, 2 + x_i^2)$, and $x \sim N(0, 3)$.

a) Compute the conditional distribution of y_i given x_i;

b) Simulate a random sample of size $n = 100$ and create a scatter plot of your data;

c) Estimate the parameters using OLS;

d) Report the estimates and three different symmetric confidence intervals for β using different measures of standard errors:

- The usual OLS standard error that ignores heteroskedasticity;
- An heterokedasticity-robust standard error;
- A bootstrap standard error, using $B = 1000$ bootstrap samples You should also plot the histogram of $\hat{\beta}$ computed from the bootstrap samples.

It will be helpful for the next item if you structure your code so that it receives a certain sample $\{(y_1, x_1), (y_2, x_2), ..., (y_n, x_n)\}$ and spits out the three different confidence intervals.

e) Repeat the steps from items b), c), and d) for $m = 500$ times and check how frequently the three different confidence intervals cover the true value of $\beta = 5$. You can use the same code from item b) to do this, but you will need to modify it so that it runs m times and stores the results.