Econ 883: Problem Set 2

Federico A. Bugni*

November 5, 2020
Due date: Monday, November 16, 4 p.m., as an upload in Sakai

1. Let $\{(X_i, Y_i)\}_{i=1}^n$ be an i.i.d. sample of $(X, Y) : \Omega \to \mathbb{R}^2$ with $E[Y^2] < \infty$. Also, assume that X is continuously distributed with density $f_X(x)$. Our goal is to estimate $\theta \equiv E[Y]$.

   An obvious estimator of θ would be the sample analogue estimator, i.e.,

$$\bar{Y}_n \equiv \frac{1}{n}\sum_{i=1}^n Y_i.$$

   In this problem, we want to consider an alternative estimator of θ and analyze its asymptotic properties. The idea for the new estimator comes from the law of iterated expectations, which gives:

$$\theta \equiv E[Y] = E[E[Y|X]] = \int E[Y|X = x] f_X(x) dx.$$

   Based on this, one could suggest the following estimator of θ:

$$\hat{\theta}_n \equiv \int \hat{E}[Y|X = x] \hat{f}_X(x) dx,$$

   where $\hat{E}[Y|X = x]$ and $\hat{f}_X(x)$ are the kernel regression and kernel density estimators, respectively, i.e.,

$$\begin{aligned} \hat{E}[Y|X = x] &\equiv \sum_{i=1}^n \frac{Y_i K\left(\frac{X_i - x}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{X_j - x}{h_n}\right)} \\ \hat{f}_X(x) &\equiv \frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{X_i - x}{h_n}\right), \end{aligned}$$

   K denotes the kernel function, and $\{h_n\}_{n\geq 1}$ is the bandwidth sequence.

*Email: federico.bugni@duke.edu. Please watch out for typos and mistakes.

Furthermore, we assume the following:

- $K : \mathbb{R} \to \mathbb{R}$ is a bounded univariate function that satisfies $\int K(u)du = 1$, $\int K(u)udu = 0$, $K(u) = 0$ if $|u| > 1$, $\kappa_1 = \int K(u)u^2du$, and $\kappa_2 = \int K^2(u)du$.
- For some $\delta > 0$, the functions $m_1(x) \equiv E[Y|X = x]f_X(x)$, $m_2(x) \equiv E[Y^2|X = x]f_X(x)$, and $m_3(x) \equiv E[|Y|^{2+\delta}|X = x]f_X(x)$ are twice continuously differentiable. Furthermore, their second derivatives satisfy the following Lipschitz condition; for $d = 1, 2, 3$, there is a (non-stochastic) function $M(x)$ such that $\int M(x)dx < \infty$ and

$$|m_d''(x + v) - m_d''(x)| \quad < \quad M(x) \times |v|.$$

- $\{h_n\}_{n\geq 1}$ is a bandwith sequence that satisfies $h_n \to 0$ as $n \to \infty$.

Based on this setup, answer the following questions.

(a) Show that $\hat{\theta}_n$ is consistent.

(b) Derive the asymptotic distribution of $\hat{\theta}_n$. In particular, show that

$$\sqrt{n}(\hat{\theta}_n - E[Y]) \xrightarrow{d} N(0, \Sigma)$$

under suitable (undersmoothing) conditions. Characterize these conditions and Σ.

(c) Which estimator of $\theta \equiv E[Y]$ do you prefer: $\hat{\theta}_n$ or $\bar{Y}_n$? Compare these estimators using the following criteria: assumptions required, asymptotic properties, finite sample properties (e.g., bias, efficiency, etc.), practical aspects of their implementation.

2. The body mass index (BMI) is an attempt to quantify the amount of tissue mass (muscle, fat, and bone) for individuals aged 20 or over. The value of this index is then used to categorize an individual as having underweight, normal weight, overweight, or obesity. The BMI was initially proposed in the nineteenth century by a Belgian scientist named Lambert Adolphe Jacques Quetelet, and it has been adopted by the U.S. National Institutes of Health since the 1980's.

The BMI is defined as follows:

$$BMI_i \equiv \frac{W_i}{H_i^2},$$

where W_i denotes individual i's weight measured in kilograms and H_i denotes individual i's height measured in meters.[1]

In this problem, we are interested in studying the expected BMI for males in the U.S. as a function of their age. To this end, we use Centers for Desease Control and Prevention (CDC) National Health and Nutrition Examination Survey (NHANES) for the years 2009/2010.[2] The processed data are available in Sakai in the CSV file named "CDC_data_males.csv", which has $n = 4,436$ observations of the following variables:

- age: measured in years (constructed from age measured at monthly frequency),
- weight: measured in kilograms,

[1] There is debate on how to divide the BMI scale to determine the weight categories. The U.S. National Institutes of Health proposed the following guidelines: BMIs under 18.5 are considered underweight, between 18.5 and 25 are considered normal, between 25 to 30 are considered overweight, and over 30 indicate obesity.

[2] The data are publicly available in the CDC's website: http://wwwn.cdc.gov/nchs/nhanes/search/nhanes09_10.aspx.

- height: measured metres.

(a) In this first part, we assume the expected BMI satisfies the following equation:

$$BMI_i = \beta_1 + \beta_2 Age_i + \beta_3 Age_i^2 + \varepsilon_i. \tag{1}$$

Furthermore, assume that $\{(Age_i, \varepsilon_i)\}_{i=1}^n$ is an i.i.d. sample with finite second moments and $E[\varepsilon|Age] = 0$. As a consequence,

$$E[BMI|Age] = \beta_1 + \beta_2 Age + \beta_3 Age^2. \tag{2}$$

Based on this information, complete the following tasks.

i. Estimate the parameters in Eq. (1) using OLS. Use these results to estimate the regression function $E[BMI|Age]$.

ii. Construct an asymptotically valid confidence interval for $E[BMI|Age = a]$ for every $a \in [20, 70]$. Use a confidence size of 90%.
Note: These results were covered in previous econometrics courses.

iii. Plot the parametric estimate of $E[BMI|Age]$ and its corresponding confidence interval.

(b) In this second part, we avoid making parametric assumptions. We assume that $\{(Age_i, \varepsilon_i)\}_{i=1}^n$ is an i.i.d. sample. Furthermore, for all $a \in [20, 70]$, we assume that the second order derivatives of $E[BMI|Age = a]f_{Age}(a)$ and $f_{Age}(a)$ are continuous and bounded functions, $V[BMI|Age = a]$ and $E[BMI^{2+\delta}|Age = a]$ are continuous functions, and that $f_{Age}(a) > 0$ and $E[BMI^{2+\delta}|Age = a] < \infty$ for some $\delta > 0$.

Based on this information, complete the following tasks.

i. Estimate the mean regression function nonparametrically using the Nadaraya-Watson kernel regression estimator. Implement the estimation using a second-order Gaussian kernel functions and the following bandwidths: rule-of-thumb, cross-validation, and Akaike information criterion. Briefly compare the results across bandwidths.

ii. Construct an asymptotically valid confidence interval for $E[BMI|Age = a]$ for every $a \in [20, 70]$. Use a confidence size of 90%.

iii. Plot the nonparametric estimate of $E[BMI|Age]$ and its corresponding confidence interval.

(c) Briefly compare inference results using parametric and nonparametric approaches.

3. Let $\{(X_i, Y_i)\}_{i=1}^n$ be a sample of data with $(X, Y) : \Omega \to \mathbb{R}^{q+1}$. Show that the Nadaraya Watson kernel estimator coincides with the local constant estimator, i.e., for every $x \in \mathbb{R}^q$,

$$\hat{E}[Y|X = x] = \sum_{i=1}^n \frac{Y_i K_X\left(\left\{\frac{X_{i,d} - x_d}{h_{n,d}}\right\}_{d=1}^q\right)}{\sum_{j=1}^n K_X\left(\left\{\frac{X_{j,s} - x_s}{h_{n,s}}\right\}_{s=1}^q\right)}$$

is the unique solution to

$$\arg\min_{b \in \mathbb{R}} \sum_{i=1}^n (Y_i - b)^2 K_X\left(\left\{\frac{X_{i,d} - x_d}{h_{n,d}}\right\}_{d=1}^q\right)$$

4. Exercise 2.4 in Li and Racine (2007).

5. Exercise 2.6 in Li and Racine (2007).

References

Li, Q. and J. S. Racine (2007): *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.