EC709 PS1. Nonparametric Estimation

Iván Fernández-Val

Boston University

Due on September 19

Question 1: Local Regression In this question you are going to derive the asymptotic properties of the kernel regression and locally linear regression estimators of the conditional expectation function and compare these properties with the finite sample properties of these estimators. Assume that Y and X are continuous random variables such that $g_0(x) = \mathbb{E}[Y \mid X = x]$, $f_0(x)$ is the probability density function of X, and K is a kernel function that satisfy the properties stated in the lecture notes.

1. Derive the asymptotic bias and variance of the kernel regression estimator of $g_0(x)$. Spell out the assumptions that you use in your derivations.

2. Derive the asymptotic bias and variance of the locally linear regression estimator of $g_0(x)$. Spell out the assumptions that you use in your derivations.

3. Compute numerically the bias and variance of the two estimators by simulation when $X \sim U(0,1)$, $Y = \exp(X)(1+\varepsilon)$, $\varepsilon \sim N(0,1)$ independent of X, $x = 0.5$, and $n = 1,000$. Consider the Epanechnikov kernel with 4 values for the bandwidth, $h \in \{0.05, 0.10, 0.15, 0.20\}$.

4. Compare the biases and variances that you obtain in the previous part with the approximations to the biases and variances predicted by the asymptotic theory.

[Hint: the packages `KernSmooth` and `locpol` implement local regression methods in R.]

Question 2: Gasoline Demand (Yatchew and No, 2001) Yatchew and No (2001) estimated gasoline demand curves using data from the National Private Vehicle Use Survey, conducted by Statistics Canada between October 1994 and September 1996. In this exercise you are going to use local and global nonparametric methods to estimate demand curves using the Canadian data. The file `yn.dta` contains information of $5,001$ households on the following variables: consumption of gasoline in liters per month (`gas`), price of gasoline per liter in Canadian dollars (`price`), household income before taxes with 9 categories (`income`), age of the primary driver top-coded at 66 years (`age`), number of drivers in the household (`driver`), household size (`hhsize`), year, month, province (`prov`), and indicators for urban dwellers (`urban`) and young-single (age less than 36 and household size of one).[1]

1. Estimate the regression of log consumption on log price using four methods: kernel regression, locally linear regression, series power regression, and series B-spline regression. In all the cases choose the tuning parameter (bandwidth or number of terms) using cross-validation. Plot your estimates of the demand function.

2. Report your estimates for the demand when the price is equal to 0.57, together with their standard errors.

3. Estimate the regression of log consumption on log price and log income using four methods: kernel regression, locally linear regression, series power regression, and series B-spline regression. In all the cases choose the tuning parameter (bandwidth or number of terms) using cross-validation. You can treat the income variable as continuous.

4. Obtain estimates of the demand function controlling for household characteristics using Lasso regression. Try 2 specifications: one including all the variables and another one adding the two-way interactions between all the variables.

[Hint: the package `splines` contains regression spline functions and classes and the command `poly` computes orthogonal polynomials in R. The package `hdm` implements Lasso regression in R.]

[1] See Yatchew and No (2001) for a more detailed description of the data.

Question 3: My first Lasso Regression Simulation (Belloni, Chernozhukov and Hansen, 2011) Read the paper:

> Belloni, A., Chernozhukov, V., and C. Hansen (2011), "Inference for high-dimensional sparse econometric models," arXiv preprint arXiv:1201.0220.
> `https://arxiv.org/pdf/1201.0220`

Replicate the results in Section 4.2 for the Lasso, Post-Lasso, CV Lasso, CV Post-Lasso, and Oracle estimator using 500 simulations. In particular:

1. Generate a dataset with n observations where

$$Y_i = X_i'\beta_0 + \varepsilon_i, \quad \varepsilon_i \sim i.i.d\ N(0, \sigma^2), \quad i = 1, \ldots, n,$$

   where $X_i \sim i.i.d.\ N_K(0, \Sigma)$ independently of all the ε_i's, $\Sigma_{ij} = (1/2)^{|i-j|}$, $\sigma^2 = 1$, $\beta_0 = (1, 1, 1/2, 1/3, 1/4, 1/5, 0, \ldots, 0)'$, $K = 500$, and $n = 100$.

2. Compute all estimators using the dataset obtained in the previous step.

3. Repeat 500 times steps 1 and 2.

4. Approximate the bias and prediction error using averages across simulations. For example, if $\hat{\beta}^{(1)}, \ldots, \hat{\beta}^{(500)}$ are the 500 Lasso estimates, approximate the prediction error by

$$\frac{1}{500}\sum_{s=1}^{500}\sqrt{\frac{1}{n}\sum_{i=1}^{n}[X_i'\hat{\beta}^{(s)} - X_i'\beta_0]^2}.$$

5. Tabulate the results.

[Hint: the package `hdm` implements Lasso and Post-Lasso estimators in R. The package `mvtnorm` generate draws from the multivariate normal distribution in R.]