14.385 Nonlinear Econometric Analysis

Kernel regression

Maximilian Kasy

Department of Economics, MIT

Fall 2022

Outline

- Kernel regression: Local weighted average of outcomes.
- Tuning parameter: Bandwidth.
- Uniform confidence bands.
- Boundary bias.
- Series regression.
- Linear smoothers.

Takeaways for this part of class

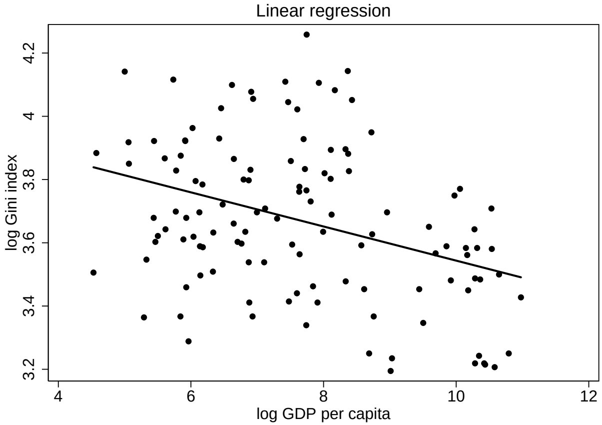
- Bandwidth governs variance-bias tradeoff:
    - Larger bandwidth $\Rightarrow$ Smaller variance.
    - Smaller bandwidth $\Rightarrow$ Smaller bias.
- Cross-validation
    - can be used to choose optimal bandwidth,
    - is easy to compute for linear smoothers.
- Bias is larger on the boundary.
  This can be reduced using local linear regression.
- A number of alternative non-parametric estimators
  can be thought of as linear smoothers.

Nonparametric Regression Estimation

- We use nonparametric regression when:
  - We are interested in the shape of the regression function.
  - We do not want to make functional form assumptions.
  - We are not directly interested in the regression function but we need an estimate to plug it in a second step estimator.
- Three classic methods:
  - Kernel regression.
  - Series regression.
  - Local linear regression.

Linear Regression

OLS: Assume linearity and minimize sum of square residuals.



But true regression (conditional expectation) may not be linear.

Kernel Regression (Nadaraya-Watson)

- Suppose we want to estimate the regression:

$$m(x_0) = E[Y|X = x_0].$$

- A kernel regression is a weighted average:

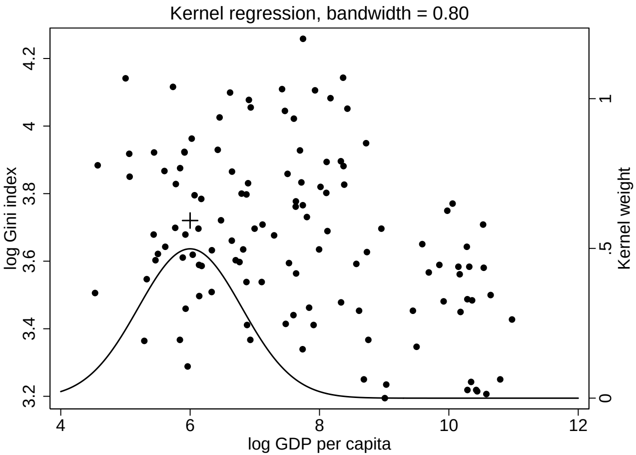
$$\widehat{m}(x_0) = \sum_{i=1}^{N} w_i Y_i,$$

  where

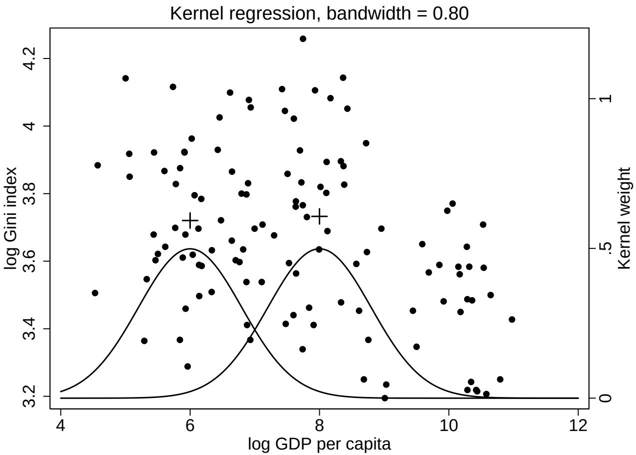
$$w_i = \frac{K\left(\frac{X_i - x_0}{h}\right)}{\sum_{j=1}^{N} K\left(\frac{X_j - x_0}{h}\right)}.$$

- Observations close to x_0 get large weights and observations distant from x_0 get small weights.

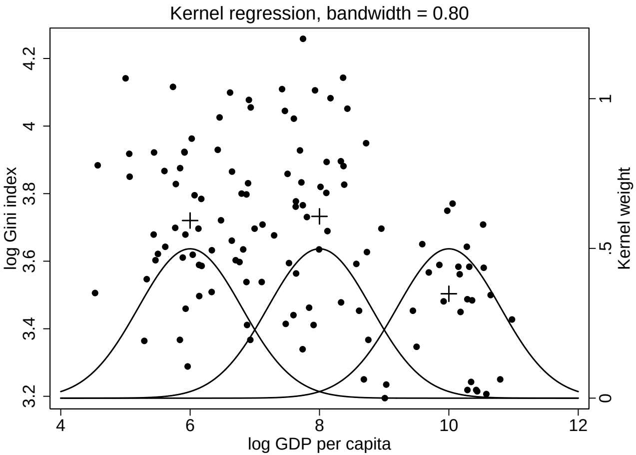
Kernel Regression



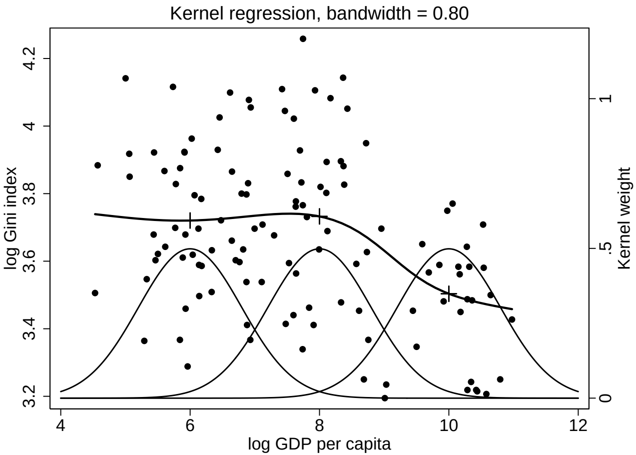
Kernel Regression



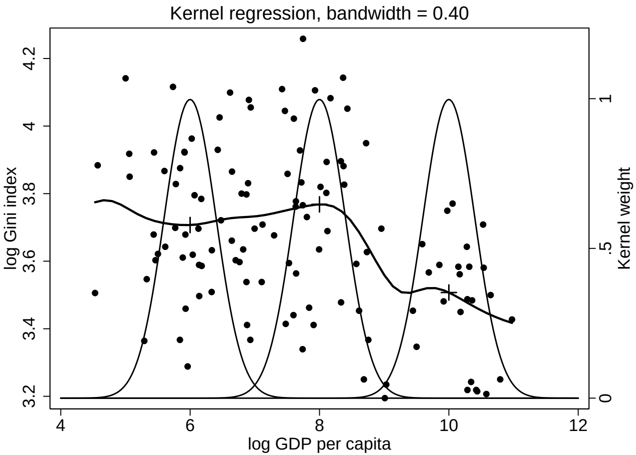
Kernel Regression



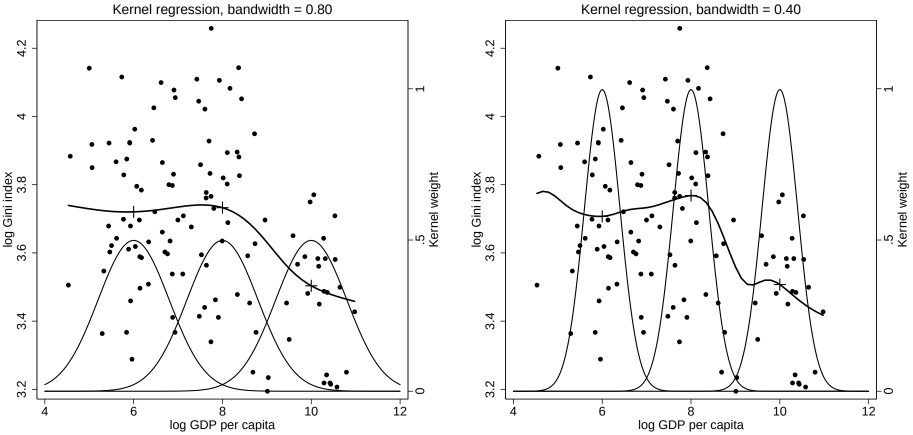
Kernel Regression



Kernel Regression



Kernel Regression



The bandwidth h is a **smoothing parameter**:

$\Rightarrow$ Large h makes regression smooth

$\Rightarrow$ Small h makes regression wiggly

Properties of Kernel Regression Estimators

- Assume that $X \in \mathbb{R}^k$.
- If $N \to \infty$, $h \to 0$, and $Nh^k \to \infty$, (and other regularity conditions hold) then

$$\widehat{m}(x_0) \xrightarrow{p} m(x_0).$$

- If, in addition, $Nh^{k+4} \to 0$, then:

$$\sqrt{Nh^k}(\widehat{m}(x_0) - m(x_0)) \xrightarrow{d} N\left(0, \frac{\sigma^2(x_0)}{f(x_0)} \int K(z)^2 dz\right),$$

  where $\sigma^2(x_0) = \text{var}(Y|X = x_0)$.

- The standard error of $\widehat{m}(x)$ can be estimated using sample analogs,

$$\widehat{s}(x_0) = \left(\frac{1}{Nh} \frac{\widehat{\sigma}^2(x_0)}{\widehat{f}(x_0)} \int K(z)^2 dz\right)^{1/2},$$

  or the bootstrap.

Choosing the Smoothing Parameter

- **Eyeballing**.
- **Plug-in**:
  - Define the **mean square error** as:

$$MSE(h) = \int \left(\widehat{m}(x) - m(x)\right)^2 f(x)dx.$$

  - A MSE-minimizing bandwidth sequence is given by:

$$h^* = cN^{-1/(k+4)},$$

    where the constant c depends on $K(z)$, $m(x)$, and $f(x)$.

  - Estimation of c by plug-in is possible but cumbersome.

Kernel regression

Tuning

Inference

Boundary bias

Alternative non-parametric regression methods

References

Choosing the Smoothing Parameter

- **Cross validation**:
  - Let
    $$CV(h) = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \widehat{m}_{-i}(X_i)\right)^2,$$
    where $\widehat{m}_{-i}(X_i)$ is the leave-i-out kernel regression estimator of $m(X_i)$ (with bandwidth h).
  - Let h_{CV} be the bandwidth sequence that minimizes $CV(h)$.
  - It can be shown that:
    $$\frac{MSE(h_{CV})}{\min_h MSE(h)} \xrightarrow{p} 1.$$

Uniform Confidence Bands

- Consider the univariate case ($k = 1$).
- Let $\mathcal{X}$ be a compact subset of the support of X. Make $\mathcal{X} = [0, 1]$.
- The goal is to obtain a band $\widehat{I}(x) = [\widehat{c}_l(x), \widehat{c}_u(x)]$, such that

$$\Pr(m(x) \in \widehat{I}(x), \forall x \in \mathcal{X}) \to 1 - \alpha. \tag{1}$$

- This is done through an approximation to the large sample behavior of

$$\sup_{x \in \mathcal{X}} \sqrt{Nh} \left(\frac{f(x)}{\sigma^2(x)} \right)^{1/2} |\widehat{m}(x) - m(x)|.$$

Kernel regression

Tuning

Inference

Boundary bias

Alternative non-parametric regression methods

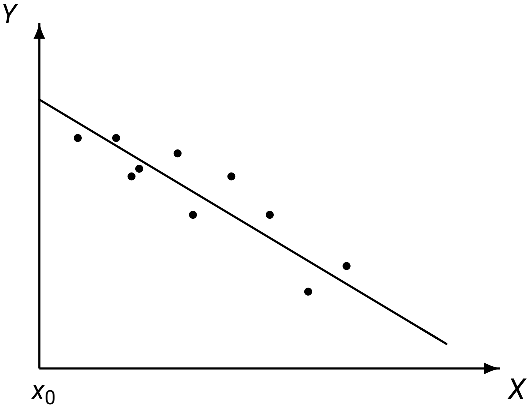
References

Uniform Confidence Bands

- Let
$$\widehat{l}(x) = \widehat{m}(x) \pm \left\{ \frac{c_\alpha}{\delta} + \delta + \frac{1}{2\delta} \ln \left(\frac{\int (K'(z))^2 dz}{4\pi^2 \int K^2(z) dz} \right) \right\} \widehat{s}(x)$$
where $\delta = \sqrt{2\ln(1/h)}$, and $\exp(-2\exp(-c_\alpha)) = 1 - \alpha$.
- Then, under regularity conditions, in particular $Nh^5 \to 0$, equation (1) holds.
- Bootstrap confidence interval are also possible.
- See Härdle and Linton (1994) for additional detail and references.

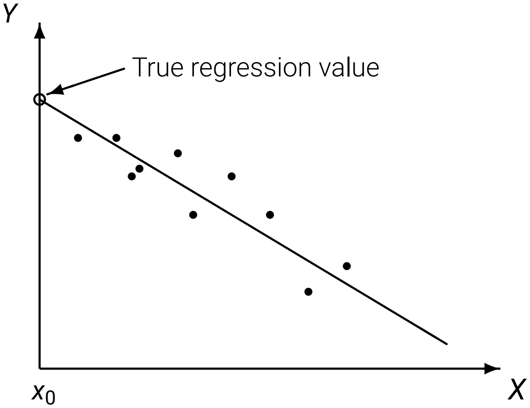
Kernel Regression: Boundary Bias

Consider x_0 at the boundary of the support of X.



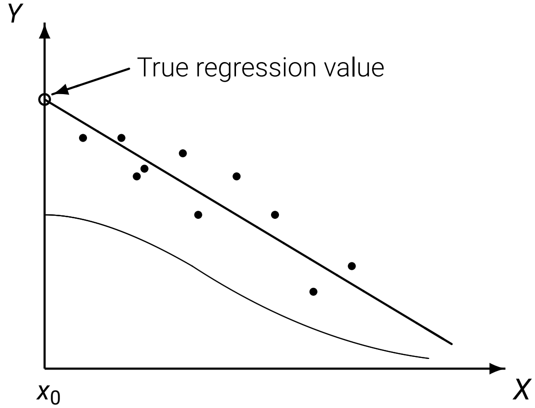
Kernel Regression: Boundary Bias

Consider x_0 at the boundary of the support of X.



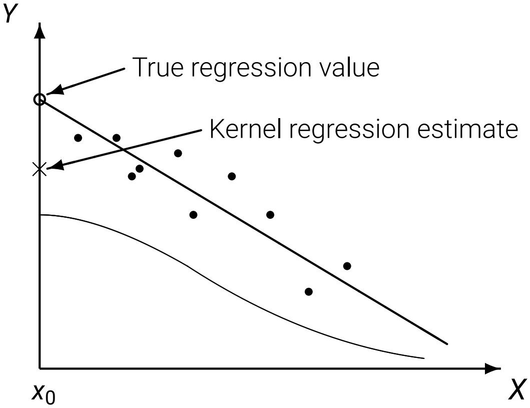
Kernel Regression: Boundary Bias

Consider x_0 at the boundary of the support of X.



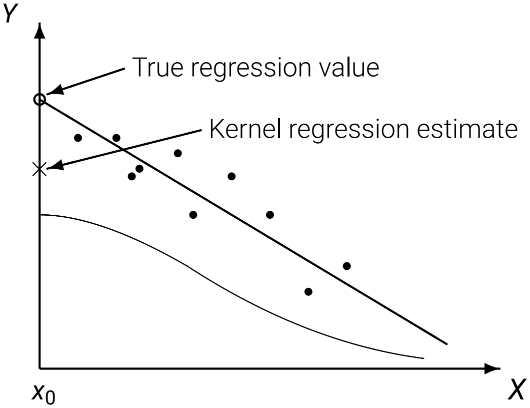
Kernel Regression: Boundary Bias

Consider x_0 at the boundary of the support of X.



Kernel Regression: Boundary Bias

Consider x_0 at the boundary of the support of X.

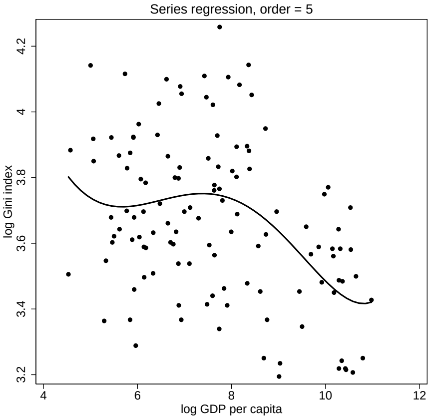


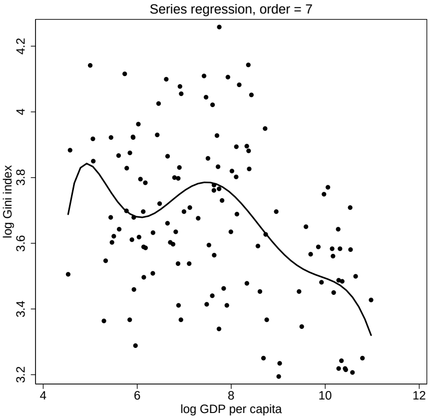
$\Rightarrow$ There is a bias because all observations that are close to the boundary have regression values smaller than the regression value of x_0.

Series Regression

Fit a polynomial of order p:

$$\sum_{i=1}^{N}(Y_i - b_0 - b_1 X_i - b_2 X_i^2 - \cdots - b_k X_i^p)^2.$$

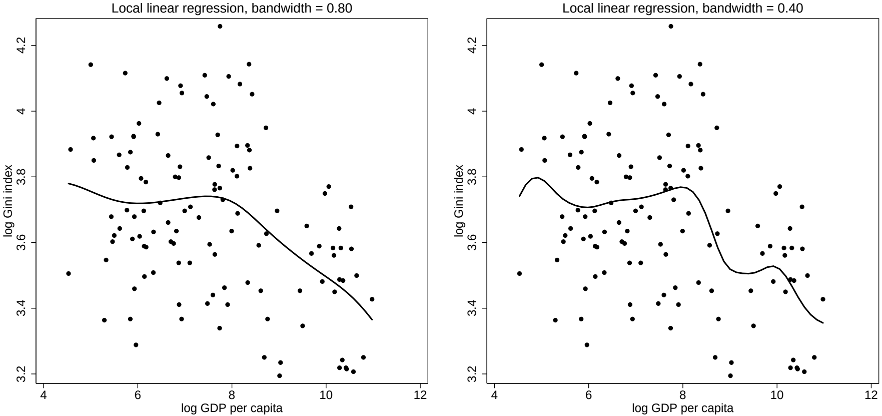




Local Linear Regression

For each $X = x_0$, minimize:

$$\sum_{i=1}^{N} K\left(\frac{X_i - x_0}{h}\right)(Y_i - b_0 - b_1 X_i)^2,$$



Local Linear and Polynomial Regression

- Local polynomial regression extends this estimator to polynomials of order p:

$$\sum_{i=1}^{N} K\left(\frac{X_i - x_0}{h}\right)(Y_i - b_0 - b_1 X_i - \cdots - b_p X_i^p)^2.$$

- Kernel regression: Special case, using $p = 0$.
- Series regression: Special case, using a constant kernel.
- Local polynomial regression can easily be estimated by the intercept value $\widehat{\beta}_0$ obtained from minimizing:

$$\sum_{i=1}^{N} K\left(\frac{X_i - x_0}{h}\right)(Y_i - b_0 - b_1(X_i - x_0) - \cdots - b_p(X_i - x_0)^p)^2.$$

- The ν-th derivative of the regression function at x_0 can be estimated by $\nu!\widehat{\beta}_\nu$.

Other Nonparametric Regression Methods

- **k-nearest neighbors**: $\hat{m}(x_0)$ = average Y_i for the k observations X_i that are closest to x_0.

- **Smoothing splines**: Let $\hat{m}(x)$ be the twice differentiable function defined on $[a,b]$ that minimizes

$$\sum_{i=1}^{N}(Y_i - \hat{m}(X_i))^2 + \lambda \int_a^b (\hat{m}''(x))^2 dx.$$

  The second term is a roughness penalty, and $\lambda \geq 0$ is a scalar smoothing parameter. Remarkably, the minimization can be solved in closed form and leads to an easily computable linear smoother.

Linear Smoothers

- An estimator $\widehat{m}(x)$ is a **linear smoother** if for each x there exists a vector $w(x) = (w_1(x), \ldots, w_N(x))$, such that:

$$\widehat{m}(x) = \sum_{i=1}^{N} w_i(x) Y_i.$$

- Examples:
  - Kernel regression:

$$w_i(x) = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^{N} K\left(\frac{X_j - x}{h}\right)}.$$

Linear Smoothers

- Series regression:

$$w_i(x) = z' \left(\sum_{j=1}^{N} Z_j Z_j' \right)^{-1} Z_i,$$

where

$$Z_i = \begin{pmatrix} 1 \\ X_i \\ \vdots \\ X_i^p \end{pmatrix} \quad \text{and} \quad z = \begin{pmatrix} 1 \\ x \\ \vdots \\ x^p \end{pmatrix}$$

- Local Polynomial Regression:

$$w_i(x) = z' \left(\sum_{j=1}^{N} K\left(\frac{X_j - x}{h}\right) Z_j Z_j' \right)^{-1} K\left(\frac{X_i - x}{h}\right) Z_i.$$

Cross-Validation of Linear Smoothers

- Cross-validation may look computationally expensive, as minimizing the cross-validation function

$$CV = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \widehat{m}_{-i}(X_i)\right)^2$$

seems to require computing the leave-one-out estimator N times for each value of h (or p for series).

- Fortunately, the cross-validation function simplifies considerable for linear smoothers:

$$CV = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_i - \widehat{m}(X_i)}{1 - w_i(X_i)}\right)^2.$$

- Using this formulation, the estimator only needs to be computed once for each value of h (or p for series).

Two-Step Estimation with a Nonparametric First Step

- There are many instances (e.g., generated regressors, propensity score weighting) where the parameter of interest θ_0 solves:

$$E[m(Z, \theta, g)] = 0$$

  in the population, and where g is an unknown functions (e.g., regression function, density function).

- In these instances, if the functional form of g is left unspecified, θ_0 is typically estimated in two-steps:
  1. Estimate g nonparametrically.
  2. Estimate θ_0 by solving:

$$\frac{1}{N}\sum_{i=1}^{N} m(Z_i, \theta, \widehat{g}) = 0.$$

References

Härdle, W. and Linton, O. (1994). Applied nonparametric methods. Handbook of econometrics, *4:2295–2339*

These slides are based on the slides by **Alberto Abadie** for previous iterations of 14.385.