10) Nonparametric and Kernel Regression

Vitor Kamada

February 2019

Reference

Tables, Graphics, and Figures from:

1) Hansen (2019). **Econometrics**. Ch 19

2) Hayeld & Racine (2008). **Nonparametric Econometrics: The np Package**. Journal of Statistical Software. Vol 27(5): 1-32

3) Hastie et al. (2017). **The Elements of Statistical Learning**. Ch 6.1 to 6.5

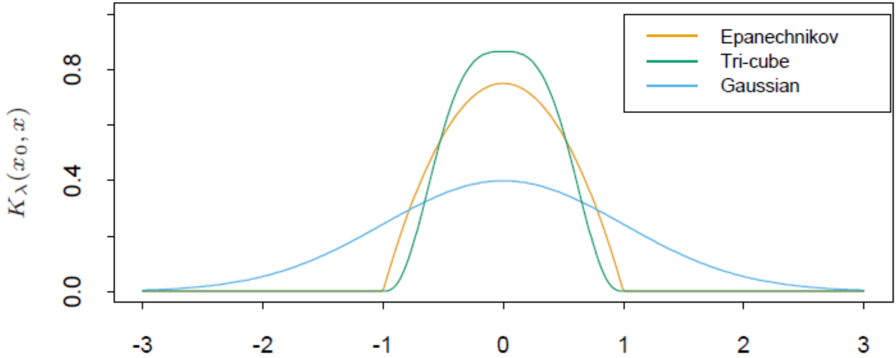
Binned Means Estimator (BME)

$$y = m(x) + \epsilon$$

$$\hat{m}(x) = \frac{\sum_{i=1}^{N} 1(|x_i - x| < h) y_i}{\sum_{i=1}^{N} 1(|x_i - x| < h)}$$

$$w_i(x) = \frac{K(|x_i - x| < h)}{\sum_{i=1}^{N} K(|x_i - x| < h)}$$

Kernel



Hastie et al. (2017)

Kernel Regression, Nadaraya-Watson, and Local Constant Estimator

Uniform Density Function on $[-1, 1]$

$$K_0(u) = \tfrac{1}{2}1(|u| \leq 1)$$

$$1(|\tfrac{x_i - x}{h}| \leq 1) = 2K_0|\tfrac{x_i - x}{h}|$$

$$w_i(x) = \frac{K_0(\frac{x_i - x}{h})}{\sum\limits_{i=1}^{N} K_0(\frac{x_i - x}{h})}$$

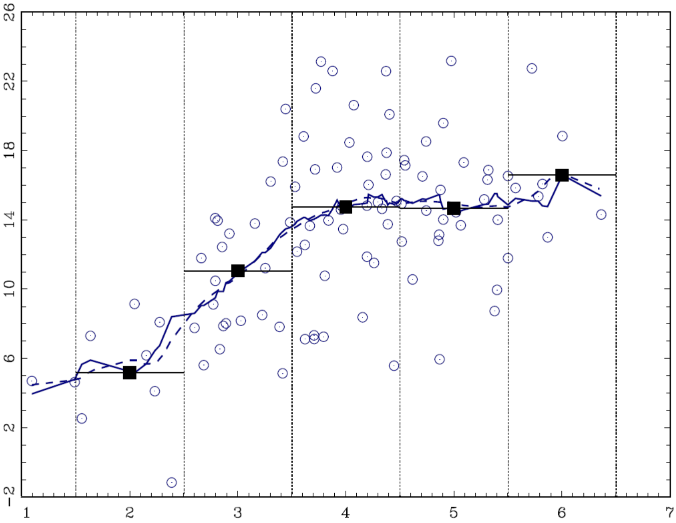
Epanechnikov Kernel

$$K_1(u) = \frac{3}{4}(1 - u^2)1(|u| \leq 1)$$

Gaussian Kernel

$$K_\phi(u) = \frac{1}{\sqrt{2\pi}} exp(-\frac{u^2}{2})$$

Blue Line: Nadaraya-Watson Regression (h=0.5)



Hansen (2018)

Dashed Line: Epanechnikov Kernel

$$\hat{m}(x) = \min_{\alpha} \sum_{i=1}^{N} K\left(\frac{x_i - x}{h}\right)(y_i - \alpha)^2$$

$m(x)$: Close to Flat Line

Perform worse for values of x near the boundary

Local Linear Estimator

$$\min_{\alpha,\beta} \sum_{i=1}^{N} K\left(\frac{x_i - x}{h}\right)[y_i - \alpha - \beta(x_i - x)]^2$$

$$z_i(x) = \begin{pmatrix} 1 \\ x_i - x \end{pmatrix}, \quad K_i(x) = K\left(\frac{x_i - x}{h}\right)$$

$$\begin{pmatrix} \hat{\alpha}(x) \\ \hat{\beta}(x) \end{pmatrix} = (Z'KZ)^{-1} Z'Ky$$

$$h \to \infty, \; \hat{m}(x) \to \hat{\alpha} + \hat{\beta} x$$

Leave-One-Out Estimator

$$\hat{e}_i = y_i - \hat{m}(x_i)$$

$h \to 0$, then $\hat{m}(x_i) \to y_i$ and $\hat{e}_i \to 0$

$$\tilde{m}_{-i}(x_i) = \frac{\sum\limits_{j \neq i} K(\frac{x_j - x}{h}) y_j}{\sum\limits_{j \neq i} K(\frac{x_j - x}{h})}$$

$$\tilde{e}_i = y_i - \tilde{y}_i$$

Cross-Validation Bandwidth Selection

$$MSE(x,h) = E\{[\hat{m}(x,h) - m(x)]^2\}$$

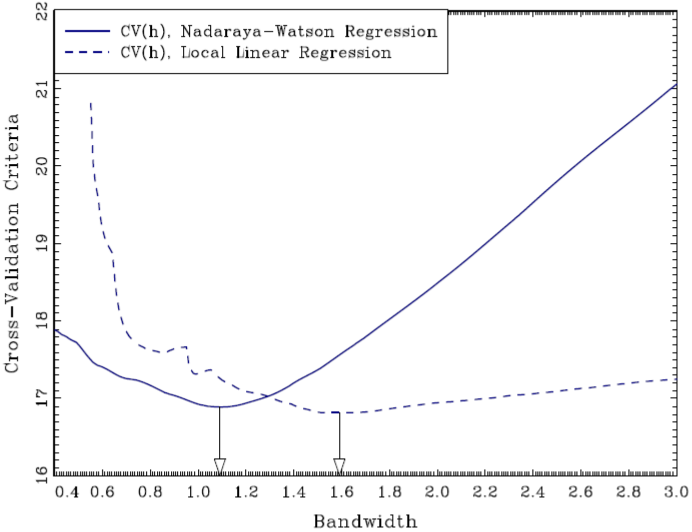
$$IMSE(h) = \int MSE(x,h) f_x(x) dx$$

$$\tilde{e}_i(h) = y_i - \tilde{m}_{-i}(x_i, h)$$

$$CV(h) = \frac{1}{N} \sum_{i=1}^{N} \tilde{e}_i(h)^2 w(x_i)$$

$$\hat{h} = \min_{h \geq h_l} CV(h)$$

Cross-Validation Criteria



Hansen (2017)

- Canadian cross-section wage data
- Random sample taken from the 1971 Canadian Census Public Use Tapes
- Male with education (Grade 13)

library("np"); data("cps71")

Local-Constant Estimator (Nadaraya-Watson)

```
NP1 <- npreg(logwage ~ age,
  regtype = "lc",
  bwmethod = "cv.aic",
  bwtype = "fixed",
  gradients = TRUE,
  ckertype = "gaussian",
  data = cps71)
```

summary(NP1); npsigtest(NP1)

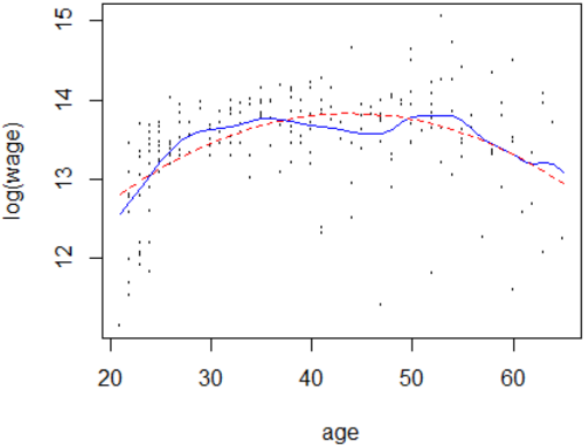
```
Regression Data: 205 training points, in 1 variable(s)
                   age
Bandwidth(s): 1.551214

Kernel Regression Estimator: Local-Constant
Bandwidth Type: Fixed
Residual standard error: 0.5244934
R-squared: 0.3261301
```

```
Kernel Regression Significance Test
Type I Test with IID Bootstrap (399 replications, Pivot = TRUE,
joint = FALSE)
Explanatory variables tested for significance:
age (1)

                   age
Bandwidth(s): 1.551214

Individual Significance Tests
P Value:
age < 2.22e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

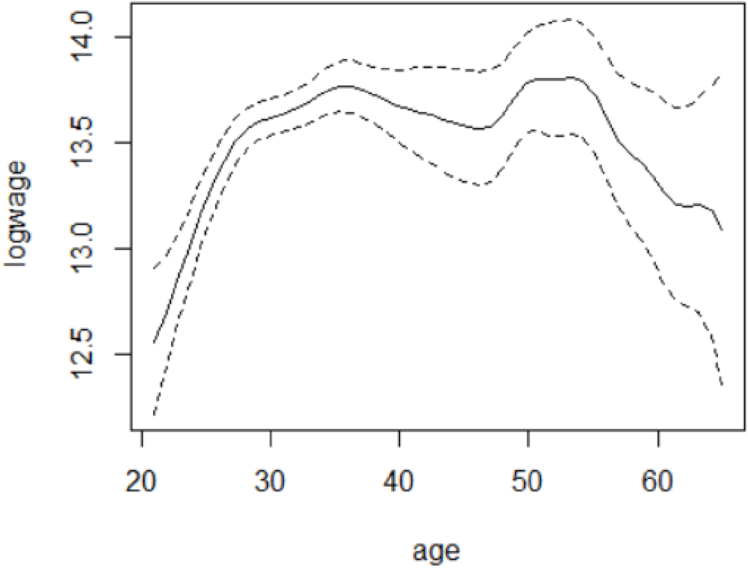
plot(cps71$age, cps71$logwage, xlab = "age", ylab = "log(wage)", cex=.1)

lines(cps71$age, fitted(NP1), lty = 1, col = "blue")

lines(cps71$age, fitted(OLS), lty = 2, col = " red")

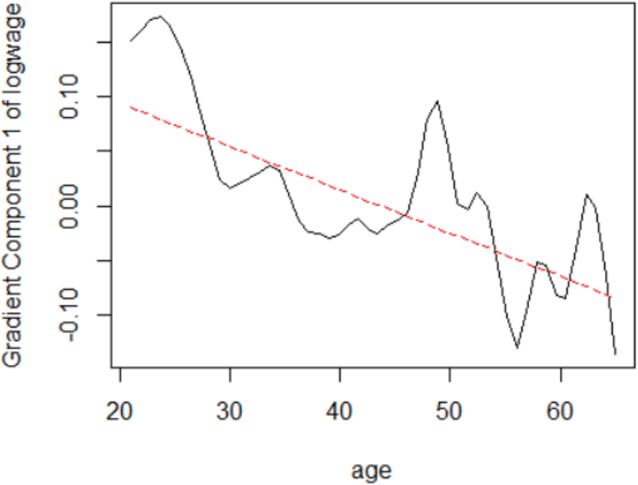


plot(NP1, plot.errors.method = "asymptotic")

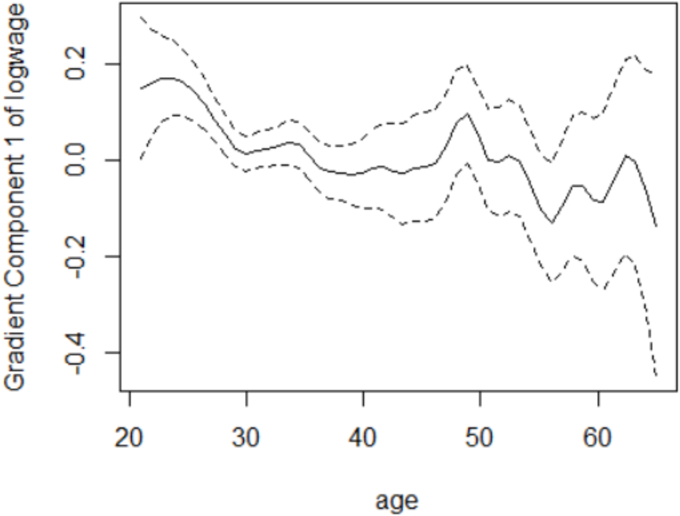


plot(NP1, gradients = TRUE)

lines(cps71$age, coef(OLS)[2]+2*cps71$age*coef(OLS)[3],
lty = 2, col = "red")



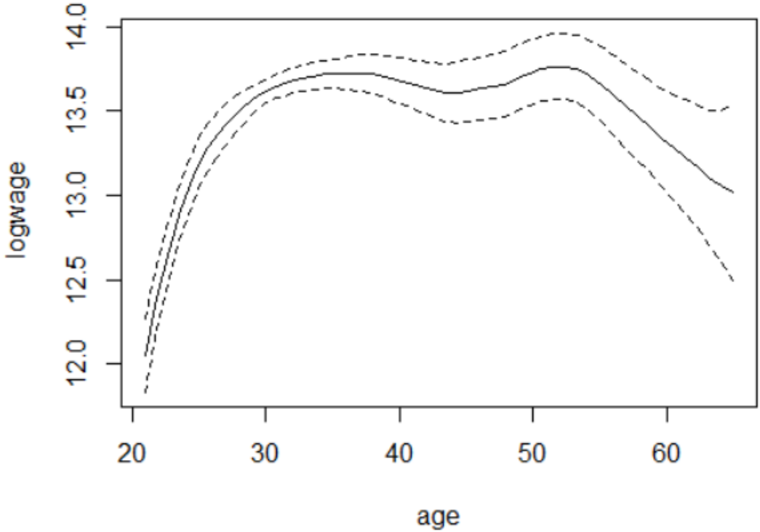
plot(NP1, gradients = TRUE, plot.errors.method = "asymptotic")



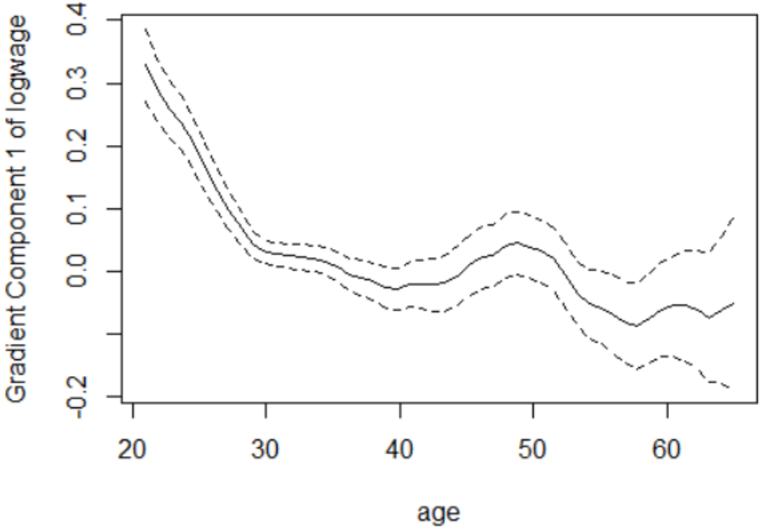
Local-Linear Estimator

```
NP2 <- npreg(logwage ~ age,
  regtype = "ll",
  bwmethod = "cv.aic",
  bwtype = "fixed",
  gradients = TRUE,
  ckertype = "epanechnikov",
  data = cps71)
```

plot(NP2, plot.errors.method = "asymptotic")



plot(NP2, gradients = TRUE, plot.errors.method = "asymptotic")



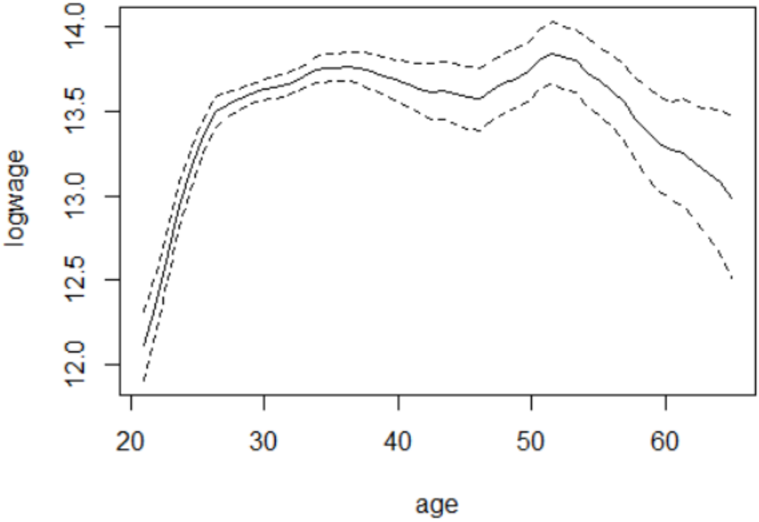
Nearest Neighbors Estimator (NNE)

$$\hat{m}_{KNN}(x_0) = \frac{1}{k} \sum_{i=1}^{N} 1[x_i \in N_k(x_0)] y_i$$

- NNE is a kernel estimator with uniform weights, except that the bandwidth is variable
- Computationally Faster

```
NP3 <- npreg(logwage ~ age,
  regtype = "ll",
  bwmethod = "cv.aic",
  bwtype = "generalized_nn",
  gradients = TRUE,
  ckertype = "epanechnikov",
  data = cps71)
```

plot(NP3, plot.errors.method = "asymptotic")



plot(NP3, gradients = TRUE, plot.errors.method = "asymptotic")