Lecture Note 12: Non-parametric Methods

Not on final

parametric vs. non-parametric
- ① estimating the density
- ② estimating the regression function

Linear relationship: $Y = \beta_0 + \beta_1 X + U$

Conditional expectation: $E[Y|X] = \beta_0 + \beta_1 X$

Non-linear relationship? $E[Y|X] = g(X)$ ← flexible function; ← regression function

Want to estimate $Y = g(X) + U$

↳ 2 ways to estimate: ① polynomial regression
② local regression

Polynomial regression

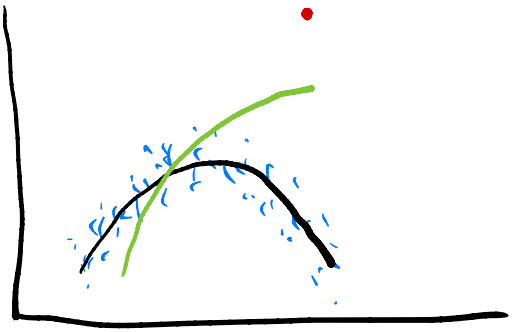
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \cdots + \beta_k X^k + U$$

→ calculus: as $k \to \infty$, polynomial converges to $g(X)$

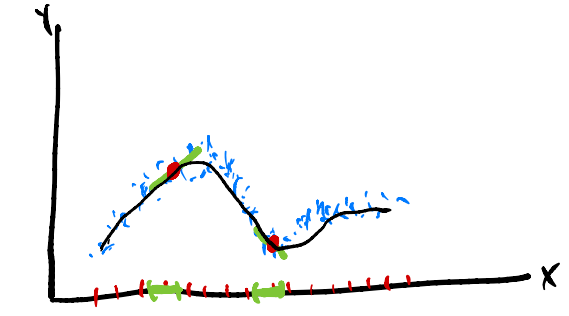
→ bias vs. variance: large k → less bias, more variance

small k → more bias, less variance

→ unappealing feature:



Local regression



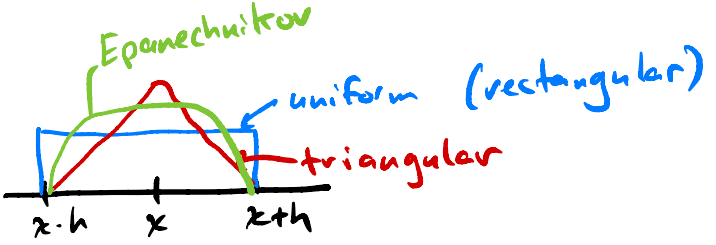
local linear regression

→ bandwidth: h (half of window)

→ bias vs. variance: large h: more bias, less var

small h: less bias, more var

→ kernel: $K\left(\frac{X_i - x}{h}\right)$



→ local linear regression:

$$\min_{b_0, b_1} \sum_i K\left(\frac{X_i - x}{h}\right)(Y_i - b_0 + b_1 X_i)^2$$

→ Stata: lpoly

→ R: kernSmooth::locpoly()

ggplot::geom_smooth()

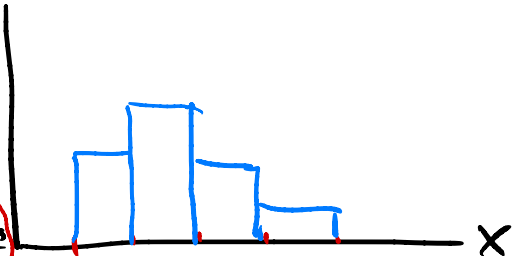
Density estimation

Histogram:

→ vertical axis: → number of obs: N_k

→ share of obs: N_k/N

→ density: $\hat{f}(x) = \frac{N_k}{N} \times \frac{\text{number bins}}{\text{width}}$



→ Stata: hist

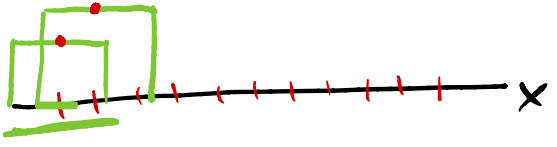
R: hist( )

ggplot::geom_histogram( )

Centered histogram

Kernel density estimator

$$\hat{f}(x) = \frac{1}{Nh} \sum_{i=1}^{N} k\left(\frac{x_i - x}{h}\right)$$



Stata: kdensity

R: ggplot::geom_density( )