Nonparametric Regression by Local Polynomials

Pedro Delicado
Departament d'Estadística i Investigació Operativa
Universitat Politècnica de Catalunya

Nonparametric Regression by Local Polynomials

REFERENCES:

Wand and Jones (1995)

Simonoff (1996)

Fan and Gijbels (1996)

Chapter 3 in Bowman and Azzalini (1997)

Chapters 1, 2 and 10 in Loader (1999)

Chapters 4 and 5 in Wasserman (2006)

Chapters 6 and 7 in Hastie, Tibshirani, and Friedman (2009)

The regression function

- Let (X, Y) be random variables with continuous joint distribution.
- The best prediction (in the sense of minimum mean squared prediction error) of the dependent variable Y given that the predicting variable X takes the known value x, is the conditional expectation of Y given that $X = x$,

$$m(x) = E(Y|X = x),$$

also known as regression function.

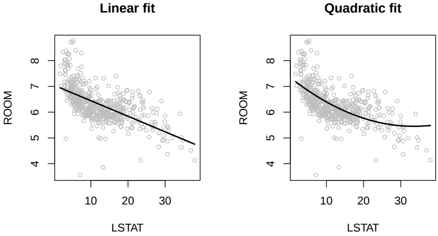
- The parametric regression models assume that the function $m(\cdot)$ is known except for a fixed finite number of unknown parameters.
- For instance, the simple linear regression model postulates that

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

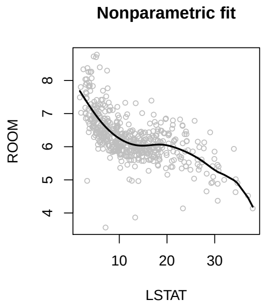
So $m(x) = \beta_0 + \beta_1 x$ is known except for two parameters: β_0, β_1.

Example. Parametric or nonparametric regression?

- Boston House-price Data, 506 neighborhoods of Boston, 1978.
- `http://lib.stat.cmu.edu/datasets/boston_corrected.txt`
- The list of variables includes:
  `ROOM` average number of rooms per dwelling,
  `LSTAT` % of the population with the lower status in a social-class classification,
  `CRIM` per capita crime rate by town,
  `AGE` proportion of owner-occupied units built prior to 1940,
  `MEDV` Median value of owner-occupied homes in $1000's
- We study `ROOM` as a function of `LSTAT`. Parametric regression.



Nonparametric fit of ROOM versus LSTAT



- The relation between variables is different when LSTAT is lower than 10%, when it is between 10% and el 20%, or when it is greater than 20%.
- In the middle range of LSTAT the values of ROOM are almost constant. In the other two sections ROOM is a decreasing function of LSTAT.
- The fall is steeper at the first section than at the third one.

The nonparametric regression model

- We observe n pairs of data (x_i, y_i) coming from the nonparametric regression model

$$y_i = m(x_i) + \varepsilon_i,\ i = 1, \ldots, n,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent r.v. with

$$E(\varepsilon_i) = 0, V(\varepsilon_i) = \sigma^2 \text{ for all } i,$$

and the predicting variable values $x_1, \ldots, x_n$ are known.

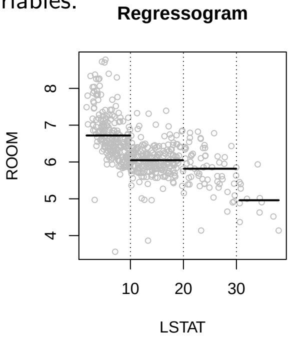
- The functional form of the regression function $m(x)$ is not specified.
- Certain regularity conditions on $m(x)$ are assumed. For instance, it is usually assumed that $m(x)$ has continuous second derivative.

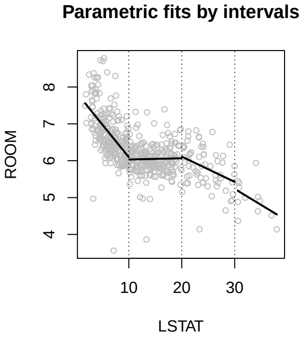
What does it mean "to fit a nonparametric regression model"?

- To provide an estimator $\hat{m}(t)$ of $m(t)$ for all $t \in \mathbb{R}$.
  - This usually implies to draw the graphic of the pairs $(t_j, \hat{m}(t_j))$, $j = 1, \ldots, J$, where t_j, $j = 1, \ldots, J$ is a regular fine grid covering the range of the observed values x_i, $i = 1, \ldots, n$.
  - An algorithm that computes $\hat{m}(t)$ for any input value t can be provided alternatively.
- To give an estimator $\hat{\sigma}^2$ of the residual variance σ^2.

Local linear regression for Boston housing data

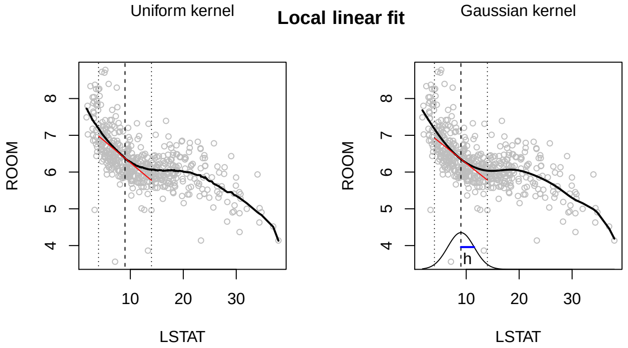
- The scatter plot of variables LSTAT and ROOM suggests that a unique linear model is not valid for the whole range of LSTAT.
- A first idea: To divide the range of LSTAT in several intervals, each of them showing an approximately linear relation between both variables.





Good results, but not entirely satisfactory. Two improvements:

- Localizing: In order to estimate the regression function at a given value t, using data (x_i, y_i) such that x_i is in an interval centered at t.
- Smoothing: Assigning to each datum (x_i, y_i) a weight $w(x_i, t)$ being a decreasing function of distance $|t - x_i|$.



Practice:

- Go to Atenea and see the html file Local linear regression: Animated graphics (html file)
- Then see the R-Markdown file that has been produced this html file.

Local linear fitting

- Weights are assigned by a kernel function K: usually, a symmetric unimodal density function centered at 0.
- The weight of (x_i, y_i) when estimating $m(t)$ is

$$w_i = w(t, x_i) = K\left(\frac{x_i - t}{h}\right) \Bigg/ \sum_{j=1}^{n} K\left(\frac{x_j - t}{h}\right),$$

- Scale parameter h: controls weight concentration around t:
  - For small values of h only the closest observations to t have a relevant weight. On the other hand, a large h allows data distant from t to be taken into account when estimating $m(t)$.
- h is called smoothing parameter or bandwidth.
- The final estimate is significantly affected by changes in the choice of the smoothing parameter, so this task is crucial in nonparametric estimation.

- Once the weights $w_i = w(t, x_i)$ have been calculated, the following weighted least squares problem is solved:

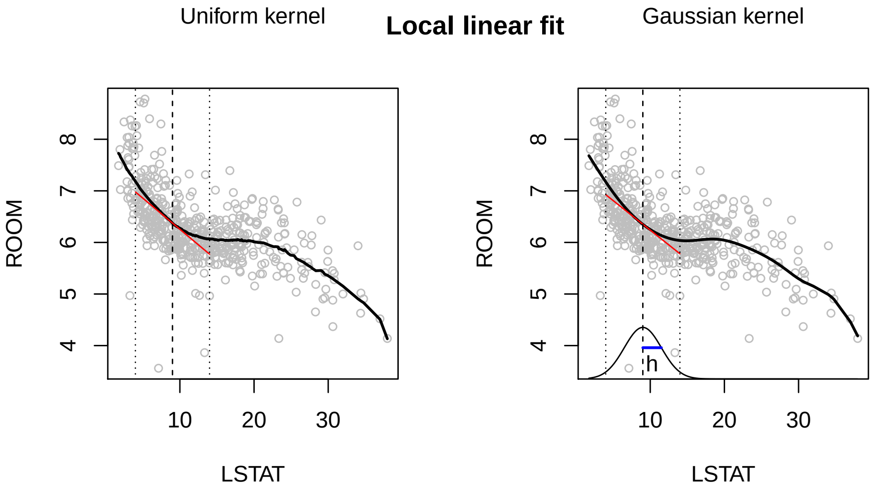
$$\min_{a,b} \sum_{i=1}^{n} w_i \left(y_i - (a + b(x_i - t))\right)^2 .$$

- The optimal parameters a and b depend on t, because the weights $w_i = w(t, x_i)$ depend on t: $a = a(t), b = b(t)$.
- The regression line fitted around t is

$$\ell_t(x) = a(t) + b(t)(x - t).$$

- Finally, the regression function estimation at point t is the value that $\ell_t(x)$ takes when $x = t$:

$$\hat{m}(t) = \ell_t(t) = a(t).$$


Local linear fit
Uniform kernel
Gaussian kernel
ROOM
LSTAT
4
5
6
7
8
10
20
30
h

Practice:

- Write your own local linear regression function.
  Use the script `02_your_llr.Rmd`
- Use it to analyze *Aircraft data* from library `sm`.
  Use the script `02_loc_pol_reg.Rmd`

Local polynomial fitting

- Consider the weighted polynomial regression problem

$$\min_{\beta_0,\ldots,\beta_q} \sum_{i=1}^{n} w_i \left(y_i - (\beta_0 + \beta_1(x_i - t) + \cdots + \beta_q(x_i - t)^q)\right)^2 .$$

- Observe that the estimated coefficients depend on t, the point for which the regression function is being estimated: $\hat{\beta}_j = \hat{\beta}_j(t)$.
- Finally, the proposed estimate for $m(t)$ is the value of the locally fitted polynomial $P_{q,t}(x) = \sum_{j=0}^{q} \hat{\beta}_j (x - t)^j$ evaluated at $x = t$:

$$\hat{m}_q(t) = P_{q,t}(t) = \hat{\beta}_0(t).$$

Moreover the estimated polynomial $P_{q,t}(x)$ allows us to estimate the first q derivatives of m at t:

$$\hat{m}_q^{(r)}(t) = \frac{d^r}{dx^r}\left(P_{q,t}(x)\right)\bigg|_{x=t} = r!\hat{\beta}_r(t).$$

Particular case: Nadaraya-Watson estimator

- When the degree of the locally fitted polynomial is $q = 0$ (that is, a constant) the resulting nonparametric estimator of $m(t)$ is known as Nadaraya-Watson estimator or, simply, kernel estimator:

$$\hat{m}_K(t) = \frac{\sum_{i=1}^{n} K\left(\frac{x_i - t}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x_i - t}{h}\right)} = \sum_{i=1}^{n} w(t, x_i) y_i.$$

- Nadaraya-Watson was proposed before local polynomial estimators.
- Observe that $\hat{m}_K(t)$ is a moving weighted mean.
- It can be proved that every local polynomial estimator is itself a weighted mean,

$$\hat{m}_q(t) = \sum_{i=1}^{n} w_q^*(t, x_i) y_i.$$

but weights $w_q^*(t, x_i)$ are not necessarily non-negative.

Local linear regression as a moving weighted average

- For weights $w_i = w(t, x_i)$, with $\sum_{i=1}^n w_i = 1$, define

$$\overline{(x-t)}_w = \sum_{i=1}^n w_i(x_i - t),\ \overline{y}_w = \sum_{i=1}^n w_i y_i,\ \overline{(x-t)y}_w = \sum_{i=1}^n w_i(x_i - t)y_i,\ \overline{(x-t)^2}_w = \sum_{i=1}^n w_i(x-t)_i^2.$$

- Solving the weighted least squares problem $\min_{a,b} \sum_{i=1}^n w_i \left(y_i - (a + b(x_i - t))\right)^2$:

$$b(t) = \frac{\overline{(x-t)y}_w - \overline{(x-t)}_w \overline{y}_w}{\overline{(x-t)^2}_w - \left(\overline{(x-t)}_w\right)^2},\ a(t) = \overline{y}_w - b(t)\overline{(x-t)}_w.$$

- Then, the local linear estimator of $m(t)$ is

$$\hat{m}_1(t) = a(t) = \bar{y}_w - \frac{\overline{(x-t)}_w}{\overline{(x-t)^2}_w - \left(\overline{(x-t)}_w\right)^2}\left(\overline{(x-t)y}_w - \overline{(x-t)}_w \overline{y}_w\right) =$$

$$\sum_{i=1}^n w_i \left(1 - \frac{\overline{(x-t)}_w}{\overline{(x-t)^2}_w - \left(\overline{(x-t)}_w\right)^2}\left((x_i - t) - \overline{(x-t)}_w\right)\right) y_i = \sum_{i=1}^n w_1^*(t, x_i) y_i.$$

Matrix formulation of the local polynomial estimator

Let

$$X_t = \begin{pmatrix} 1 & (x_1 - t) & \dots & (x_1 - t)^q \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (x_n - t) & \dots & (x_n - t)^q \end{pmatrix}$$

be the regressor matrix.

Define $Y = (y_1, \dots, y_n)^\mathsf{T}$, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\mathsf{T}$, $\beta = (\beta_0, \dots, \beta_q)^\mathsf{T}$.

Let $W_t = \mathrm{Diag}(w(x_1, t), \dots, w(x_n, t))$ be the weight matrix.

We fit the regression model $Y = X_t\beta + \varepsilon$ using weighted least squares:

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^{q+1}} \sum_{i=1}^{n} w_i \left(y_i - (\beta_0 + \beta_1(x_i - t) + \dots + \beta_q(x_i - t)^q)\right)^2 =$$

$$\arg \min_{\beta \in \mathbb{R}^{q+1}} (Y - X_t\beta)^\mathsf{T} W_t (Y - X_t\beta).$$

The solution is $\hat{\beta} = \left(X_t^\mathsf{T} W_t X_t\right)^{-1} X_t^\mathsf{T} W_t Y$.

- Solution: $\hat{\beta} = \left(X_t^{\mathsf{T}} W_t X_t\right)^{-1} X_t^{\mathsf{T}} W_t Y$.
- For $j = 0, \ldots, q$, let e_j be the $(q+1)$-dimensional vector having all its coordinates 0 except the $(j+1)$-th one, that is equal to 1.
- Then

$$\hat{m}_q(t) = \hat{\beta}_0 = e_0^{\mathsf{T}} \hat{\beta} = e_0^{\mathsf{T}} \left(X_t^{\mathsf{T}} W_t X_t\right)^{-1} X_t^{\mathsf{T}} W_t Y = S_t Y = \sum_{i=1}^{n} w_q^*(t, x_i) y_i,$$

  where $S_t = e_0^{\mathsf{T}} \left(X_t^{\mathsf{T}} W_t X_t\right)^{-1} X_t^{\mathsf{T}} W_t$ is a n-dimensional row vector.
- We say that the local polynomial regression estimator is a linear smoother because, for a fix t, $\hat{m}_q(t)$ is a linear function of $y_1, \ldots, y_n$.

The local polynomial estimator of the r-th derivative of m at point t is

$$\hat{m}_q^{(r)}(t) = r!\hat{\beta}_r(t) = r!e_r^{\mathsf{T}}\hat{\beta},$$

that is also linear in $y_1, \ldots, y_n$.

Practice:

- Local polynomial regression in R with function `locpolreg`.
  - Aircraft data.
- Local polynomial estimation in R: standard libraries and functions.

Use the script `02_loc_pol_reg.Rmd`

Linear smoothers

- A nonparametric regression estimator $\hat{m}(\cdot)$ is said to be a linear smoother when for any fix t, $\hat{m}(t)$ is a linear function of $y_1, \ldots, y_n$:

$$\hat{m}(t) = \sum_{i=1}^{n} w(t, x_i) y_i.$$

for some weight function $w(\cdot, \cdot)$.

- Linear smoothers are particular cases of linear estimators of the regression function, as OLS or Ridge regression.
- Let

$$\hat{y}_i = \hat{m}(x_i) = \sum_{j=1}^{n} w(x_i, x_j) y_j$$

be the fitted values for the n observed values x_i of the explanatory variable.

- In matrix format,

$$\hat{Y} = SY,$$

where the column vectors Y and $\hat{Y}$ have elements y_i and $\hat{y}_i$, respectively, and the matrix S has generic (i,j) element

$$s_{ij} = w(x_i, x_j).$$

- Matrix S is called the smoothing matrix, because its effect on the observed data (x_i, y_i), $i = 1, \ldots, n$, is to transform them into $(x_i, \hat{y}_i)$, $i = 1, \ldots, n$, that is a much smoother data configuration.
- For a linear smoother with smoothing matrix S ($\hat{Y} = SY$)

$$\nu = \text{Trace}(S) = \sum_{i=1}^{n} s_{ii},$$

the sum of diagonal elements of S, is the effective number of parameters.

- We have seen that local polynomial regression is a linear smoother.
- In this case, $\nu = \nu(h)$ is a decreasing function of smoothing parameter h:
    - Small values of h correspond to large numbers ν of effective parameters, that is, to highly complex and very flexible nonparametric models.
    - Large values of h correspond to small numbers ν of effective parameters, that is, to nonparametric models with low complexity and flexibility.
- The interpretation of ν as the effective number of parameters is valid for any linear nonparametric estimator.
- Then we can compare the degree of smoothing of two linear nonparametric estimators just comparing their effective numbers of parameters.

An estimator of σ^2

- The analogy with multiple linear regression suggests how the residual variance, $\sigma^2 = V(\varepsilon_i)$, can be estimated.
- In multiple linear regression with k regressors,

$$\hat{\sigma}^2 = \frac{1}{n-k}\hat{\varepsilon}^{\mathsf{T}}\hat{\varepsilon} = \frac{1}{n-k}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

  is an unbiased estimator of σ^2, the residual variance.
- A first estimator of σ^2 in nonparametric estimation:

$$\hat{\sigma}^2 = \frac{1}{n-\nu}\sum_{i=1}^{n}(y_i - \hat{m}(x_i))^2.$$

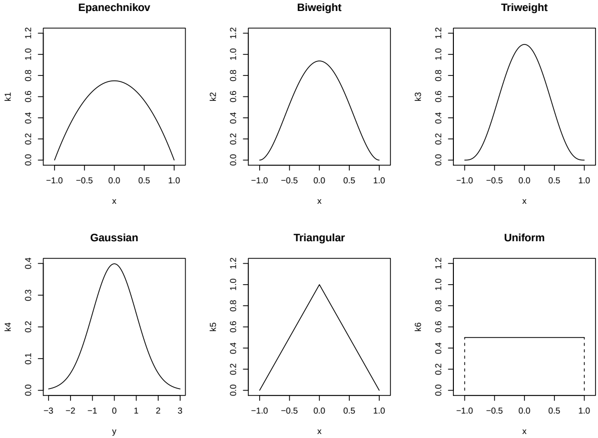
Practice:

Local polynomials as linear smoothers.

- How do the rows of smoothing matrix S look like?
- Equivalent number of parameters.
- Estimation of σ^2 when using linear smoothers.

`02_Linear_Smoothers.Rmd`

Kernel functions



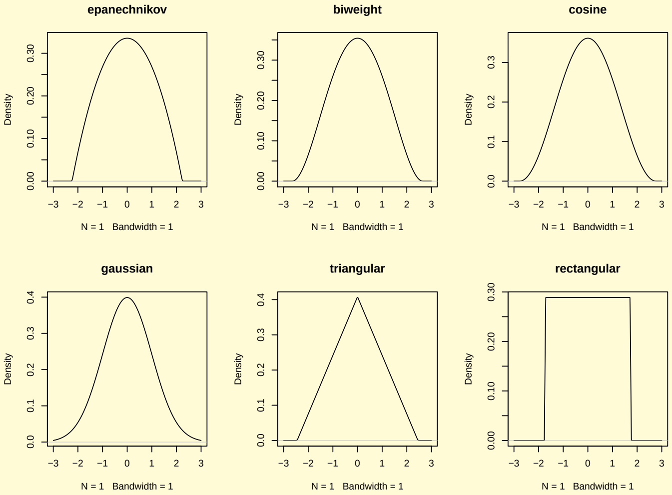
Examples of Kernel functions used in nonparametric estimation.

Kernel functions: Density functions with zero mean.

Kernel K	Expression	Variance	Efficiency
Epanechnikov (K^*)	$(3/4)(1-x^2)I_{[-1,1]}(x)$	$1/5$	1
Biweight	$(15/16)(1-x^2)^2 I_{[-1,1]}(x)$	$1/7$	0.994
Triweight	$(35/32)(1-x^2)^3 I_{[-1,1]}(x)$	$1/9$	0.987
Gaussian	$(1/\sqrt{2\pi})\exp(-x^2/2)$	1	0.951
Triangular	$(1-\lvert x\rvert)I_{[-1,1]}(x)$	$1/6$	0.986
Uniform	$(1/2)I_{[-1,1]}(x)$	$1/3$	0.930

Rescaled kernels: Density functions with zero mean and variance equal to 1.

Kernel	Original expression	Original variance	Rescaled expression
Epanechnikov	$(3/4)(1-x^2)I_{[-1,1]}(x)$	$1/5$	$(3/4\sqrt{5})(1-x^2/5)I_{[-\sqrt{5},\sqrt{5}]}(x)$
Biweight	$(15/16)(1-x^2)^2 I_{[-1,1]}(x)$	$1/7$	$(15/16\sqrt{7})(1-x^2/7)^2 I_{[-\sqrt{7},\sqrt{7}]}(x)$
Triweight	$(35/32)(1-x^2)^3 I_{[-1,1]}(x)$	$1/9$	$(35/96)(1-x^2/9)^3 I_{[-3,3]}(x)$
Gaussian	$(1/\sqrt{2\pi})\exp(-x^2/2)$	1	$(1/\sqrt{2\pi})\exp(-x^2/2)$
Triangular	$(1-\lvert x\rvert)I_{[-1,1]}(x)$	$1/6$	$(1/\sqrt{6})(1-\lvert x\rvert/\sqrt{6})I_{[-\sqrt{6},\sqrt{6}]}(x)$
Uniform	$(1/2)I_{[-1,1]}(x)$	$1/3$	$(1/2\sqrt{3})I_{[-\sqrt{3},\sqrt{3}]}(x)$



Examples of rescaled kernel functions.

Local properties of local polynomial estimator

- The term *local behavior* refers to the statistical properties of a nonparametric estimate $\hat{m}(t)$ as estimator of the unknown value $m(t)$, for a fixed value t.
- Is $\hat{m}(t)$ an unbiased estimator of $m(t)$?
  Let $\text{Bias}_{m(t)}(\hat{m}(t)) = \mathbb{E}(\hat{m}(t)) - m(t)$. Is $\text{Bias}_{m(t)}(\hat{m}(t)) = 0$?
- Is $\text{Var}(\hat{m}(t))$ going to 0 when the sample size goes to infinity?
  Is $\lim_n \text{Var}(\hat{m}(t)) = 0$?
- Is $\hat{m}(t)$ a consistent estimator of $m(t)$?
  Does $\hat{m}(t)$ converge to $m(t)$ in some sense?
- $\text{MSE}_{m(t)}(\hat{m}(t)) = \mathbb{E}((\hat{m}(t) - m(t))^2) = \text{Bias}_{m(t)}(\hat{m}(t))^2 + \text{Var}(\hat{m}(t))$.
  Is $\lim_n \text{MSE}_{m(t)}(\hat{m}(t)) = 0$?

Global properties of local polynomial estimator

- We talk about global properties when our interest is on $\hat{m}(t)$ as estimator of $m(t)$ for all $t \in [a, b]$, being $[a, b]$ the interval where the explanatory variable takes values.
- Global properties: Does the estimated function $\hat{m}$ converge to the unknown function m in some sense appropriated for functions?
- One usual way for measuring the distance between $\hat{m}$ and m is the Integrated Mean Squared Error:

$$\mathrm{IMSE}_m(\hat{m}) = \int_a^b \mathrm{MSE}_{m(t)}(\hat{m}(t))dt = \int_a^b \mathbb{E}((\hat{m}(t) - m(t))^2)dt =$$

$$\int_a^b \mathrm{Bias}_{m(t)}(\hat{m}(t))^2 dt + \int_a^b \mathrm{Var}(\hat{m}(t))dt.$$

- Is $\lim_n \mathrm{IMSE}_m(\hat{m}) = 0$?

Bias and variance of $\hat{m}_0(t)$ and $\hat{m}_1(t)$

Theorem. *Consider the nonparametric regression model*

$$Y_i = m(X_i) + \varepsilon_i,\ i = 1 \dots n$$

where $\varepsilon_1, \dots, \varepsilon_n$ *are independent r.v. with* $E(\varepsilon_i) = 0$ *and* $V(\varepsilon_i) = \sigma^2(x_i)$, $X_1, \dots, X_n$ *are independent r.v. with density* f, *with* $\Pr(a \leq X_i \leq b) = 1$, *for some* $a, b \in \mathbb{R}$. *Assume the following regularity conditions:*

1. $f(t) > 0$.
2. $f(t)$, $m''(t)$ *y* $\sigma^2(t)$ *are continuous in a neighborhood of* t.
3. K *is symmetric with support on* $[-1, 1]$, $\int_R K(u)du = 1$, $\int_{-1}^{1} uK(u)du = 0$.
4. $t \in (a, b)$.
5. $h \longrightarrow 0$ *and* $nh \longrightarrow \infty$ *when* $n \longrightarrow \infty$.

In this context, and conditioning on $X_1, \dots, X_n$, *we have the following:*

- *The Nadaraya-Watson estimator and the local linear estimator both have variance*

$$\frac{\sigma^2(t)}{nhf(t)}\int_{-1}^{1} K^2(u)du + o\left(\frac{1}{nh}\right).$$

- *The Nadaraya-Watson estimator has bias*

$$\left(\frac{m'(t)f'(t)}{f(t)} + \frac{m''(t)}{2}\right) h^2 \int_{-1}^{1} u^2 K(u)du + o(h^2).$$

- *The local linear regression estimator has bias*

$$\frac{m''(t)}{2} h^2 \int_{-1}^{1} u^2 K(u)du + o(h^2).$$

- *The Mean Squared Error (MSE) of $\hat{m}(t)$ as an estimator of $m(t)$,*

$$E[(\hat{m}(t) - m(t))^2] = Bias(\hat{m}(t))^2 + V(\hat{m}(t))$$

is $O(h^4) + O(1/(nh))$ for both estimators. Then both converge to $m(t)$ in quadratic mean and in probability.

Bias and variance of $\hat{m}_q(t)$

- Local polynomial estimators with degrees $q = 2k$ and $q = 2k + 1$ give similar asymptotic results:

$$\text{MSE}(\hat{m}_q(t)) = O(h^{4k+4}) + O\left(1/(nh)\right).$$

- The bias asymptotic expression is simpler for q odd. They do not depend on the density function of X_i.
- A general recommendation is to use the degree $q = 2k + 1$ instead of using $q = 2k$.

The bias-variance trade-off

- The Asymptotic Mean Squared Error (AMSE) is the main part of the MSE (ignoring the infinitesimal terms).
- Let us consider the AMSE for the local linear estimator:

$$\text{AMSE}(h) = \frac{(m''(t))^2}{4} h^4 \left(\int_{-1}^{1} u^2 K(u) du \right)^2 + \frac{\sigma^2(t)}{nhf(t)} \int_{-1}^{1} K^2(u) du$$

- The first term, the squared bias, increases with h.
- The second term, the variance, decreases with h.
- The optimal value h_{AMSE} represents a compromise between bias and variance.

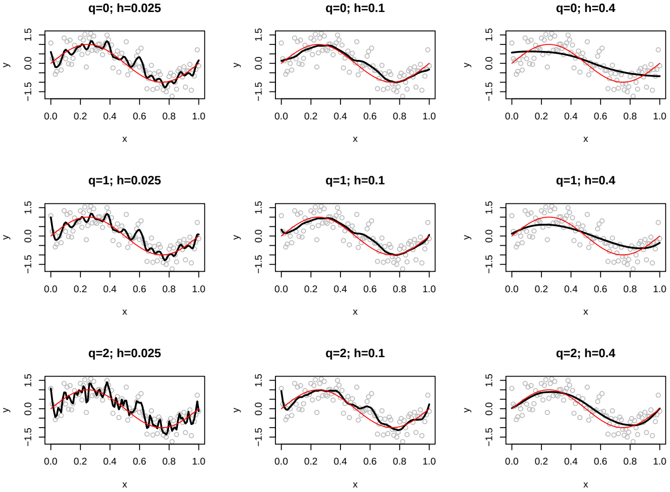
- Let $g(h) = \text{AMSE}(h)$. It has the expression $g(h) = ah^4 + b/h$. Doing $g'(h) = 0$ it follows that the minimum of g is at $h^* = (b/4a)^{1/5}$ and $g(h^*) = 5a(h^*)^4$.
- Therefore,

$$h_{\text{AMSE}} = \left(\frac{\sigma^2(t)}{nf(t)(m''(t))^2}\right)^{1/5} \left(\frac{\int_{-1}^{1} K^2(u)du}{\left(\int_{-1}^{1} u^2 K(u)du\right)^2}\right)^{1/5} n^{-1/5},$$

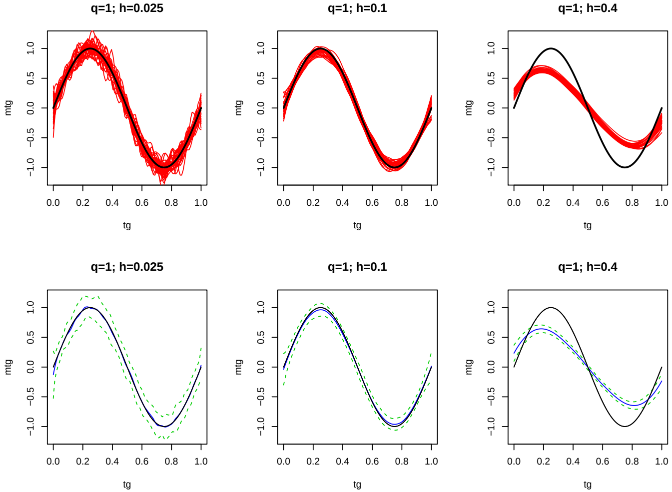
$$AMSE(h_{\text{AMSE}}) = \frac{5}{4^{4/5}} \frac{(\sigma^2(t))^{4/5}((m''(t))^2)^{1/5}}{f(t)^{4/5}}$$

$$\left(\int_{-1}^{1} K^2(u)du\right)^{4/5} \left(\int_{-1}^{1} u^2 K(u)du\right)^{2/5} n^{-4/5}.$$

Effect of bandwidth h and degree q on a single sample



Effect of bandwidth h on many samples



Practice:

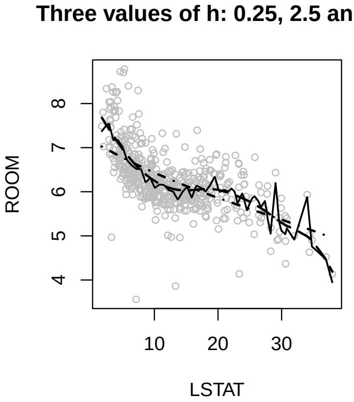
Local behaviour. Bias-variance trade-off:

`02_Bias_Var_h.Rmd`

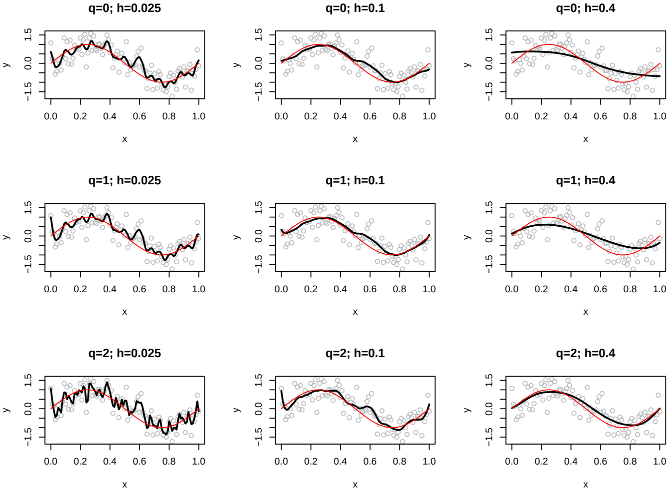
Bandwidth choice

The choice of smoothing parameter h is of crucial importance in the appearance and properties of the regression function estimator.

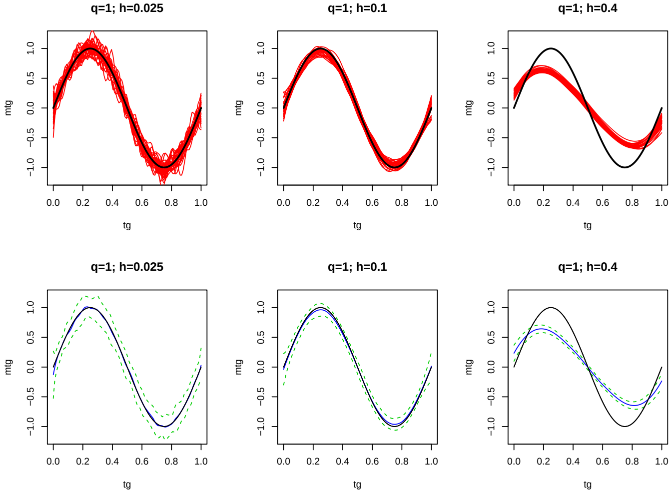
Example: Boston housing data. Local linear fit with Gaussian kernel.



Effect of bandwidth h (and degree q) on a single sample



Effect of bandwidth h on many samples



- **Estimation:** The bandwidth controls the bias-variance trade-off.
  - For h small the estimator is highly variable (applied to different samples from the same model gives very different results) and has small bias (the average of the estimators obtained for different samples is approximately the true regression function).
  - If h is large the opposite happens.
- **Prediction of new observations:** The smoothing parameter h controls the balance between fitting the observed data well and the ability to predict future observations.
  - Small values of h give great flexibility to the estimator and allow it to approach all the observed data (when h tends to 0 the estimator tends to interpolate the data), but the prediction errors will be high. There is overfitting.
  - If h is too large, there is underfitting, as it may happen with global parametric models. In this case both, the errors in the observed sample as well as the prediction errors in independent data, will be high.

Bandwidth choice: According to which criterion?

- Several criteria are sensible.
- Some of them represent global measures of estimation quality.
  - Statistics perspective.
- Other are related with prediction error for new observations.
  - Machine Learning perspective.

Bandwidth choice: Estimation based criteria

- Integrated Mean Squared Error (IMSE). A first global measure of the error made when using the nonparametric estimator $\hat{m}(t)$, $t \in [a, b]$, as an estimation of function $m(t)$, $t \in [a, b]$:

$$\text{IMSE}(\hat{m}) = \int_a^b E_{\mathbf{Z}}\left((\hat{m}(t) - m(t))^2\right) f(t)dt = \int_a^b \text{MSE}(\hat{m}(t))f(t)dt,$$

where $\mathbf{Z} = \{(x_i, Y_i) : i = 1, \ldots, n\}$ is the sample used to compute $\hat{m}$.

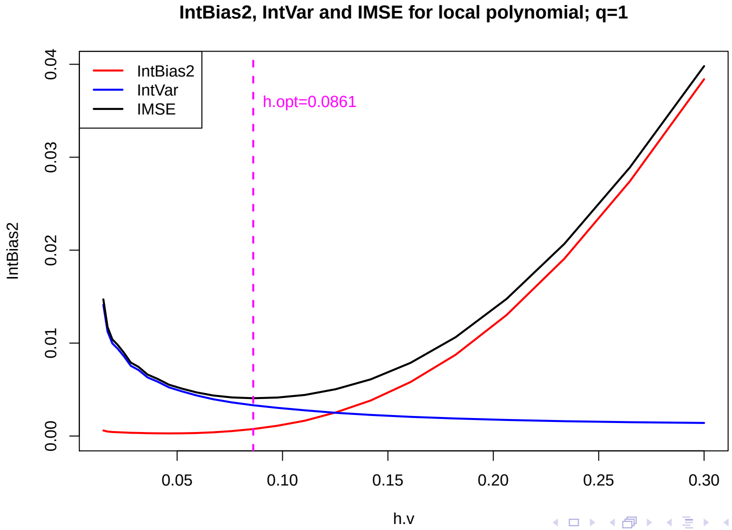
- IMSE is the sum of integrated squared bias plus integrated variance:

$$\text{IMSE}(\hat{m}) = \int_a^b \text{MSE}(\hat{m}(t))f(t)dt = \int_a^b E((\hat{m}(t) - m(t))^2)f(t)dt$$

$$= \int_a^b E(\{(\hat{m}(t) - E(\hat{m}(t))) + (E(\hat{m}(t)) - m(t))\}^2)f(t)dt$$

$$= \int_a^b \left(\text{Bias}^2(\hat{m}(t)) + V(\hat{m}(t))\right) dt.$$

Integrated Variance, integrated squared Bias and IMSE as a function of h



Practice:

Global behaviour. Global bias-variance trade-off:

02_Bias_Var_h.Rmd

- Mean Integrated Squared Error (MISE).

$$\text{MISE}(\hat{m}) = E_{\mathbf{Z}} \left(\int_a^b (\hat{m}(t) - m(t))^2 f(t) dt \right) =$$

$$E_{\mathbf{Z}} \left[E_T \{ (\hat{m}(T) - m(T))^2 \mid \mathbf{Z} \} \right] = E_{\mathbf{Z},T} \left[(\hat{m}(T) - m(T))^2 \right],$$

where $\mathbf{Z} = \{(x_i, Y_i) : i = 1, \ldots, n\}$ is the sample used to compute $\hat{m}$, and T is a random variable independent from $\mathbf{Z}$, with the same distribution that generates the independent variable values x_i, $i = 1, \ldots, n$.

- It coincides with the IMSE:

$$\text{MISE}(\hat{m}) = E_{\mathbf{Z}} \left(\int_a^b (\hat{m}(t) - m(t))^2 f(t) dt \right) \overset{\text{Fubini's Theorem}}{=}$$

$$\int_a^b E_{\mathbf{Z}} \left((\hat{m}(t) - m(t))^2 \right) f(t) dt = \text{IMSE}(\hat{m}).$$

Bandwidth choice: Prediction based criteria

- Predictive Mean Square Error (PMSE). It is the expected squared error made when predicting

$$Y = m(t) + \varepsilon$$

by $\hat{m}(t)$, where t is an observation of the random variable T, distributed as the observed explanatory variable, when T and ε are independent from the sample $\mathbf{Z}$ used to compute $\hat{m}$. Then

$$\text{PMSE}(\hat{m}) = E_{\mathbf{Z},(T,Y)}\left[(Y - \hat{m}(T))^2\right] = E_{\mathbf{Z},T,\varepsilon}\left[(\hat{m}(T) - m(T) - \varepsilon)^2\right] =$$

$$E_{\mathbf{Z},T}\left[(\hat{m}(T) - m(T))^2\right] + E_{\varepsilon}(\varepsilon^2) = \text{MISE}(\hat{m}) + \sigma^2.$$

- Observe that MISE and PMSE are equivalent criteria for evaluating a nonparametric estimator $\hat{m}(\cdot)$.
- Unfortunately both, MISE and PMSE, are **unfeasible** because they depend on the **unknown** regression function $m(\cdot)$.

- Residual Sum of Squares (RSS). An attempt to define a feasible version of
$$\text{PMSE}(\hat{m}) = E_{\mathbf{Z},(T,Y)}\left[(Y - \hat{m}(T))^2\right].$$
- Two actions in the expression of PMSE:
  - the expectation with respect to **Z** is eliminated:
$$E_{(T,Y)}\left[(Y - \hat{m}(T))^2\right].$$
  - the expectation with respect to (T, Y) is replaced by the average over the observed regressor values (x_i, y_i), $i = 1, \ldots, n$, that are distributed as (T, Y):
$$\text{RSS}(\hat{m}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{m}(x_i))^2.$$
- It is also known as error in the training sample.
- RSS is an optimistic estimation of PMSE.

- Why is the RSS an **optimistic** estimation of PMSE?
- The random quantities (T, Y) and the random sample $\mathbf{Z}$ are not independent in RSS.
  - Remember that in PMSE, the data $\mathbf{Z}$ used to compute $\hat{m}$ and those used to evaluate it, (T, Y), are independent.
- Observe that in the definition of RSS, the data are used twice: first they are used to compute $\hat{m}$, and then they are used to evaluate if $\hat{m}$ is a good estimator of m.
  - Then the estimated residuals $\varepsilon_i = y_i - \hat{m}(x_i)$ tend to be smaller than a genuine residual $\varepsilon = Y - \hat{m}(T)$, independent from $\hat{m}$.

- We have seen several global measures indicating whether a nonparametric estimator $\hat{m}$ is good or not for estimating an unknown regression function m.
- It is equivalent measuring closeness between $\hat{m}$ and m or prediction errors.
- Unfortunately these measures are unfeasible because they depend on unknown functions or quantities.
- The only exception is RSS, that is optimistically biased.
- We will see now how to obtain feasible versions of these criteria.

Bandwidth choice in practice

Several alternatives:

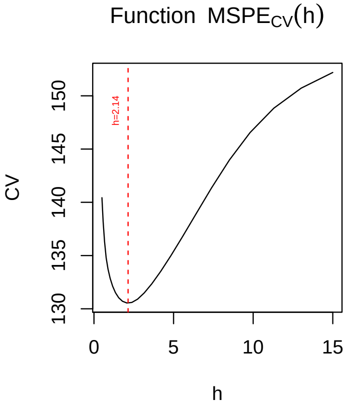
- Minimizing the average squared prediction error in a validation set.
- Leave-one-out cross-validation. It can be proved that for local polynomials

$$\text{PMSE}_{\text{CV}}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - s_{ii}}\right)^2.$$

  where s_{ii}, $i = 1, \ldots, n$, are the diagonal elements of the smoothing matrix.
- Generalized cross-validation.
- K-fold cross-validation.
- Plug-in: Specific bandwidth selector for local polynomial regression.

Example. Leave-one-out cross-validation

$\mathrm{PMSE}_{\mathrm{CV}}(h)$ as a function of h in the example of local linear regression of `ROOM` against `LSTAT`.



Plug-in bandwidth choice in the local linear estimator

- This is a bandwidth choice method specific for local linear regression.
- We have obtained before that for the local linear fit

$$\mathrm{AMISE}(\hat{m}) = \frac{h^4 \mu_2^2(K)}{4} \int_a^b (m''(x))^2 f(x) dx + (b-a) \frac{R(K)\sigma^2}{nh}.$$

- The value of h minimizing this expression is

$$h_0 = \left(\frac{R(K)\sigma^2}{\mu_2^2(K) \int_a^b (m''(x))^2 f(x) dx} \right)^{1/5} n^{-1/5}.$$

- Some quantities there are unknown: the expected value of $(m''(X))^2$ and σ^2.
- h_{PI}: Replacing the unknowns by estimations.

Estimating $E[(m''(X))^2]$ and σ^2.

- In order to estimate $\int_a^b (m''(x))^2 f(x)dx = E[(m''(X))^2]$ a local cubic polynomial regression can be fitted, using weights $w(x_i, t) = K((x_i - t)/g)$, where the bandwidth g must be chosen.
- Once $m''(t)$ has been estimated for $t = x_1, \dots, x_n$, $E[(m''(X))^2]$ is estimated as $\frac{1}{n}\sum_{i=1}^n (\hat{m}''_g(x_i))^2$.
- The optimal value of g for estimating the second derivative of $m(x)$ is

$$g_0 = C_2(K)\left(\frac{\sigma^2}{|\int_a^b m''(x)m^{(iv)}(x)f(x)dx|}\right)^{1/7} n^{-1/7}.$$

- At this point the estimation of $m''(x)$ and $m^{(iv)}(x)$ is done dividing the range of the explanatory variable in subintervals (4, for instance) and fitting a degree 4 polynomial at each subinterval.
- This last step also provides another estimation of σ^2.

Asymptotic behavior of bandwidth selectors of h

- We have seen three bandwidth selectors that do not require a validation set: h_{CV}, h_{GCV} and h_{PI}.
- The three methods provide bandwidths that converge to the value h_0 minimizing the AMISE when n goes to infinity, but their rates of convergence are different:

$$\frac{h_{\text{CV}}}{h_0} - 1 = O_p(n^{-1/10}),\ \frac{h_{\text{GCV}}}{h_0} - 1 = O_p(n^{-1/10}),\ \frac{h_{\text{PI}}}{h_0} - 1 = O_p(n^{-2/7}).$$

Practice:

Bandwidth choice: `02_Bandwidth_choice.Rmd`

Variable bandwidth

- The expression of the bandwidth h_{AMSE} minimizing the asymptotic mean square error, AMSE, of $\hat{m}(t)$ as an estimator of $m(t)$ is

$$h_{\text{AMSE}}(t) = \left(\frac{R(K)\sigma^2(t)}{\mu_2^2(K) f(t) (m''(t))^2} \right)^{1/5} n^{-1/5}.$$

- This expression suggests that sometimes it could be better to use different bandwidths at different points t.
- Variable bandwidth. The bandwidth depends on the point t where the function is being estimated: $h(t)$.

$$h_{\mathrm{AMSE}}(t) = \left(\frac{R(K)\sigma^2(t)}{\mu_2^2(K) f(t) (m''(t))^2} \right)^{1/5} n^{-1/5}.$$

When is it recommended to use a variable bandwidth?

- When the density of the explanatory variable varies considerably along the support of the explanatory variable (in areas with much data the bandwidth can be smaller than in areas where there are few observations).
- When the residual variance is a function of the explanatory variable (in areas with great residual variability it is recommended to use large values of the window).
- When the curvature of the regression function is different in different parts of the support of the explanatory variable (in areas where curvature is larger, smaller values of h should be used).

How to define a variable bandwidth in practice?

- The most common way to include a variable bandwidth is to fix the proportion s of data points to be used in the estimation of each value $m(t)$ and define $h(t)$ such that the number of data (x_i, y_i) with x_i belonging the interval $(t - h(t), t + h(t))$ is sn. The ratio s is called *span*.
- If a local polynomial of degree $q = 0$ is fitted (Nadaraya-Watson estimator) using the uniform kernel and choosing $s = k/n$, the resulting estimator is known as the *k-nearest neighbours* estimator. The choice of s (or $k = sn$) can be done by cross-validation or using a validation set.

Choosing the degree q of the local polynomial

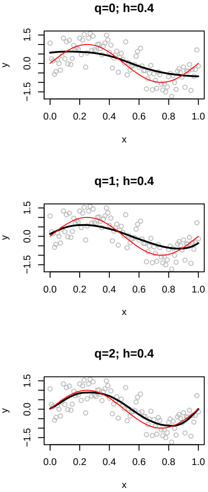
- The effect on the final estimation of the choice of the local polynomial degree, is much less important than the effect of the bandwidth choice.
- The larger is q the better are the asymptotic properties (in bias) but in practice it is recommended to use $q = r + 1$, where r is the order of the derivative of $m(t)$ that is estimated.
- When estimating $m(t)$, it is preferable to use the odd degree $q = 2k + 1$ than the preceding even degree $2k$.
- Among other advantages of local polynomials with odd degree, they are able to automatically adapt to the boundary of the explanatory variable support (when it is not the whole real line).

- To decide if it is worth fitting a local cubic model ($q = 3$) instead of just fitting a local linear model ($q = 1$), we must take into account the asymptotic expression of the local linear estimator bias:

$$\text{Bias}(\hat{m}_1(t)) = \frac{m''(t)}{2} h^2 \mu_2(K) + o(h^2).$$

- Bias is high for t in intervals where the function $m(t)$ has high curvature: large values of $|m''(t)|$.
- Therefore, if we suspect that the regression function $m(t)$ could be very bumpy it would be better to use $q = 3$ instead of $q = 1$.

Effect of degree q on a single sample



Bowman, A. W. and A. Azzalini (1997).
Applied Smoothing Techniques for Data Analysis.
Oxford: Oxford University Press.

Fan, J. and I. Gijbels (1996).
Local polynomial modelling and its applications.
London: Chapman & Hall.

Hastie, T., R. Tibshirani, and J. Friedman (2009).
The elements of statistical learning (2nd ed.).
Springer.

Loader, C. (1999).
Local regression and likelihood.
New York: Springer.

Simonoff, J. S. (1996).
Smoothing methods in statistics.
New York: Springer.

Wand, M. P. and M. C. Jones (1995).
Kernel smoothing.
London: Chapman and Hall.

Wasserman, L. (2006).
All of Nonparametric Statistics.
New York: Springer.