

Statistics for Applications

Chapter 7: Further questions in regression

Linear regression and lack of identifiability (1)

Consider the following model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

with:

1. $\mathbf{Y} \in \mathbb{R}^n$ (dependent variables), $\mathbf{X} \in \mathbb{R}^{n \times p}$ (deterministic design) ;
2. $\boldsymbol{\beta} \in \mathbb{R}^p$, unknown;
3. $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 I_n)$.
 - ▶ Previously, we assumed that X had rank p , so we could invert $X'X$.
 - ▶ What if X is not of rank p ? E.g., if $p > n$?
 - ▶ $\boldsymbol{\beta}$ would no longer be identified: estimation of $\boldsymbol{\beta}$ is vain (unless we add more structure).

Linear regression and lack of identifiability (2)

- ▶ What about prediction ? $\mathbf{X}\beta$ is still identified.
- ▶ $\hat{\mathbf{Y}}$: orthogonal projection of \mathbf{Y} onto the linear span of the columns of \mathbf{X} .
- ▶ $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^\dagger\mathbf{X}\mathbf{Y}$, where A^\dagger stands for the (Moore-Penrose) pseudo inverse of a matrix A .
- ▶ Similarly as before, if $k = \text{rank}(\mathbf{X})$:
 - ▶ $\frac{\|\hat{\mathbf{Y}} - \mathbf{Y}\|_2^2}{\sigma^2} \sim \chi_{n-k}^2$,
 - ▶ $\|\hat{\mathbf{Y}} - \mathbf{Y}\|_2^2 \perp\!\!\!\perp \hat{\mathbf{Y}}$.

Linear regression and lack of identifiability (3)

- ▶ In particular:

$$\mathbb{E}[\|\hat{\mathbf{Y}} - \mathbf{Y}\|_2^2] = (n - k)\sigma^2.$$

- ▶ Unbiased estimator of the variance:

$$\hat{\sigma}^2 = \frac{1}{n - k} \|\hat{\mathbf{Y}} - \mathbf{Y}\|_2^2.$$

Linear regression in high dimension (1)

Consider again the following model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

with:

1. $\mathbf{Y} \in \mathbb{R}^n$ (dependent variables), $\mathbf{X} \in \mathbb{R}^{n \times p}$ (deterministic design) ;
2. $\boldsymbol{\beta} \in \mathbb{R}^p$, unknown: to be estimated;
3. $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 I_n)$.
 - ▶ For each i , $X_i \in \mathbb{R}^p$ is the vector of covariates of the i -th individual.
 - ▶ If p is too large ($p > n$), there are too many parameters to be estimated (overfitting model), although some covariates may be irrelevant.
 - ▶ Solution: Reduction of the dimension.

Linear regression in high dimension (2)

- ▶ **Idea:** Assume that only a few coordinates of β are nonzero (but we do not know which ones).
- ▶ Based on the sample, select a subset of covariates and estimate the corresponding coordinates of β .
- ▶ For $S \subseteq \{1, \dots, p\}$, let

$$\hat{\beta}_S \in \operatorname{argmin}_{\mathbf{t} \in \mathbb{R}^S} \|\mathbf{Y} - \mathbf{X}_S \mathbf{t}\|^2,$$

where \mathbf{X}_S is the submatrix of \mathbf{X} obtained by keeping only the covariates indexed in S .

Linear regression in high dimension (3)

- ▶ Select a subset S that minimizes the prediction error penalized by the complexity (or size) of the model:

$$\|\mathbf{Y} - \mathbf{X}_S \hat{\beta}_S\|^2 + \lambda |S|,$$

where $\lambda > 0$ is a tuning parameter.

- ▶ If $\lambda = 2\hat{\sigma}^2$, this is the *Mallow's C_p* or *AIC* criterion.
- ▶ If $\lambda = \hat{\sigma}^2 \ln n$, this is the *BIC* criterion.

Linear regression in high dimension (4)

- ▶ Each of these criteria is equivalent to finding $\beta \in \mathbb{R}^p$ that minimizes:

$$\|\mathbf{Y} - \mathbf{X}\mathbf{b}\|_2^2 + \lambda\|\mathbf{b}\|_0,$$

where $\|\mathbf{b}\|_0$ is the number of nonzero coefficients of \mathbf{b} .

- ▶ This is a computationally hard problem: nonconvex and requires to compute 2^n estimators (all the $\hat{\beta}_S$, for $S \subseteq \{1, \dots, p\}$).
- ▶ *Lasso* estimator: replace $\|\mathbf{b}\|_0$ with $\|\mathbf{b}\|_1$ and the problem becomes convex.

$$\hat{\beta}^L \in \operatorname{argmin}_{\mathbf{b} \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\mathbf{b}\|_2^2 + \lambda\|\mathbf{b}\|_1,$$

where $\lambda > 0$ is a tuning parameter.

Linear regression in high dimension (5)

- ▶ How to choose λ ?
- ▶ This is a difficult question (see grad courses in statistics).
- ▶ A *good choice* of λ will lead to an estimator $\hat{\beta}$ that is very close to β and will allow to recover the subset S^* of all $j \in \{1, \dots, p\}$ for which $\beta_j \neq 0$, with high probability.

Nonparametric regression (1)

- ▶ In the linear setup, we assumed that $Y_i = \mathbf{X}_i' \boldsymbol{\beta} + \varepsilon_i$, where \mathbf{X}_i are deterministic.
- ▶ This has to be understood as working conditionally on the design.
- ▶ This is to assume that $\mathbb{E}[Y_i | \mathbf{X}_i]$ is a linear function of \mathbf{X}_i , which is not true in general.
- ▶ Let $f(x) = \mathbb{E}[Y_i | \mathbf{X}_i = x]$, $x \in \mathbb{R}^p$: How to estimate the function f ?

Nonparametric regression (2)

Let $\rho = 1$ in the sequel.

- ▶ One can make a parametric assumption on f .
- ▶ E.g., $f(x) = a + bx$, $f(x) = a + bx + cx^2$, $f(x) = e^{a+bx}$, ...
- ▶ The problem reduces to the estimation of a finite number of parameters.
- ▶ LSE, MLE, all the previous theory for the linear case could be adapted.
- ▶ What if we do not make any such parametric assumption on f ?

Nonparametric regression (3)

- ▶ Assume f is smooth enough: f can be well approximated by a piecewise constant function.
- ▶ Idea: Local averages.
- ▶ For $x \in \mathbb{R}$: $f(t) \approx f(x)$ for t close to x .
- ▶ For all i such that X_i is close enough to x ,

$$Y_i \approx f(x) + \varepsilon_i.$$

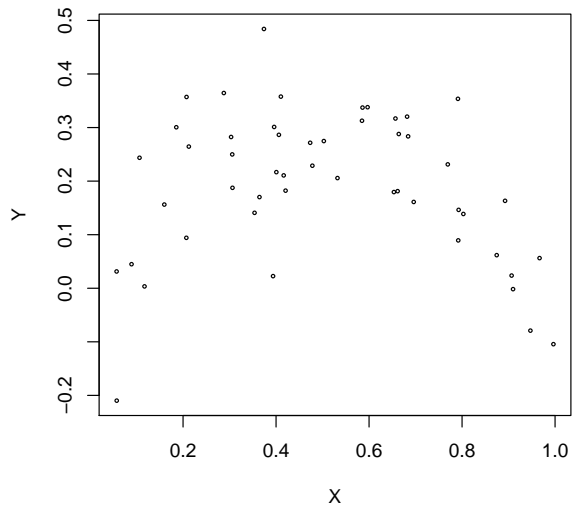
- ▶ Estimate $f(x)$ by the average of all Y_i 's for which X_i is close enough to x .

Nonparametric regression (4)

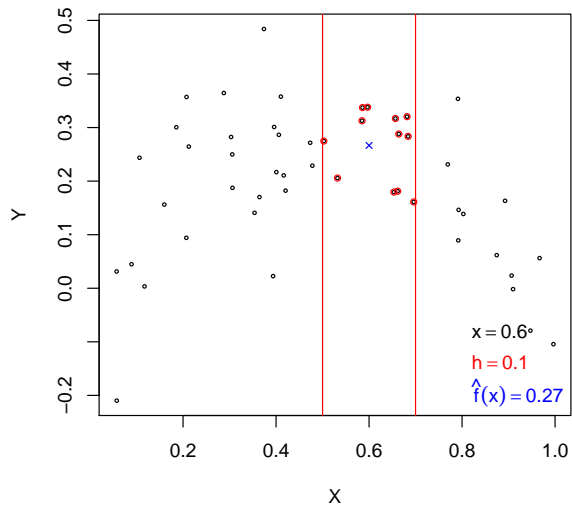
- ▶ Let $h > 0$: the window's size (or bandwidth).
- ▶ Let $I_x = \{i = 1, \dots, n : |X_i - x| < h\}$.
- ▶ Let $\hat{f}_{n,h}(x)$ be the average of $\{Y_i : i \in I_x\}$.

$$\text{▶ } \hat{f}_{n,h}(x) = \begin{cases} \frac{1}{|I_x|} \sum_{i \in I_x} Y_i & \text{if } I_x \neq \emptyset \\ 0 & \text{otherwise.} \end{cases}$$

Nonparametric regression (5)



Nonparametric regression (6)



Nonparametric regression (7)

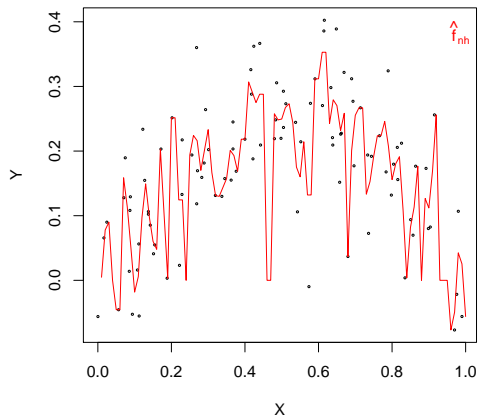
How to choose h ?

- ▶ If $h \rightarrow 0$: overfitting the data;
- ▶ If $h \rightarrow \infty$: underfitting, $\hat{f}_{n,h}(x) = \bar{Y}_n$.

Nonparametric regression (8)

Example:

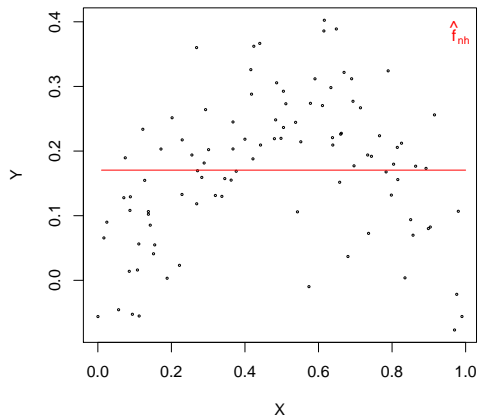
- ▶ $n = 100$, $f(x) = x(1 - x)$,
- ▶ $h = .005$.



Nonparametric regression (9)

Example:

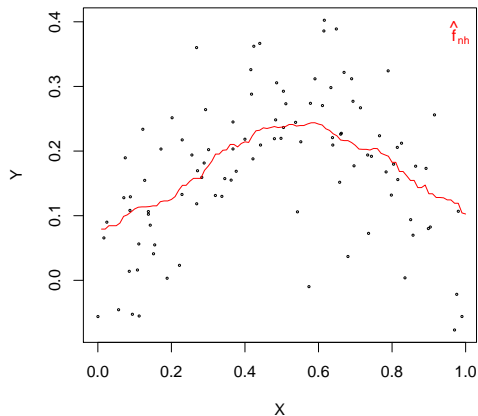
- ▶ $n = 100$, $f(x) = x(1 - x)$,
- ▶ $h = 1$.



Nonparametric regression (10)

Example:

- ▶ $n = 100$, $f(x) = x(1 - x)$,
- ▶ $h = .2$.



Nonparametric regression (11)

Choice of h ?

- ▶ If the smoothness of f is known (i.e., quality of local approximation of f by piecewise constant functions): There is a *good* choice of h depending on that smoothness
- ▶ If the smoothness of f is unknown: Other techniques, e.g. *cross validation*.