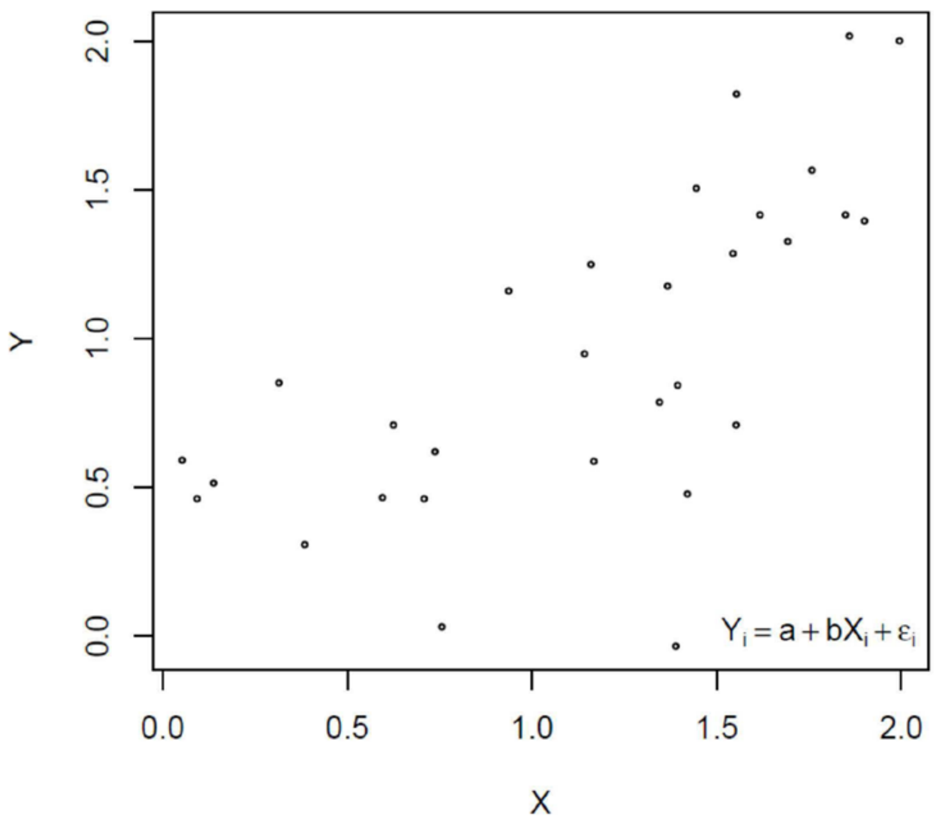
Statistics for Applications

Chapter 7: Regression

Heuristics of the linear regression (1)

Consider a cloud of i.i.d. random points $(X_i, Y_i), i = 1, \ldots, n$ **:**



Heuristics of the linear regression (2)

- **Idea:** Fit the *best* line fitting the data.
- Approximation: $Y_i \approx a + bX_i, i = 1, \ldots, n$, for some (unknown) $a, b \in \mathbb{R}$.
- Find $\hat{a}, \hat{b}$ that approach a and b.
- More generally: $Y_i \in \mathbb{R}, X_i \in \mathbb{R}^d$,

$$Y_i \approx a + X_i^\top b, \quad a \in \mathbb{R}, b \in \mathbb{R}^d.$$

- **Goal:** Write a rigorous model and estimate a and b.

Heuristics of the linear regression (3)

Examples:

Economics: Demand and price,

$$D_i \approx a + bp_i, \quad i = 1, \ldots, n.$$

Ideal gas law: $PV = nRT$,

$$\log P_i \approx a + b \log V_i + c \log T_i, \quad i = 1, \ldots, n.$$

Linear regression of a r.v. Y on a r.v. X (1)

Let X and Y be two real r.v. (non necessarily independent) with two moments and such that $Var(X) \neq 0$.

The *theoretical linear regression* of Y on X is the *best approximation in quadratic means* of Y by a linear function of X, i.e. the r.v. $a + bX$, where a and b are the two real numbers minimizing $\mathbb{E}\left[(Y - a - bX)^2\right]$.

By some simple algebra:

- $b = \dfrac{cov(X,Y)}{Var(X)}$,
- $a = \mathbb{E}[Y] - b\mathbb{E}[X] = \mathbb{E}[Y] - \dfrac{cov(X,Y)}{Var(X)}\mathbb{E}[X]$.

Linear regression of a r.v. Y on a r.v. X (2)

If $\varepsilon = Y - (a + bX)$, then

$$Y = a + bX + \varepsilon,$$

with $\mathbb{E}[\varepsilon] = 0$ and $cov(X, \varepsilon) = 0$.

Conversely: Assume that $Y = a + bX + \varepsilon$ for some $a, b \in \mathbb{R}$ and some centered r.v. ε that satisfies $cov(X, \varepsilon) = 0$.

E.g., if $X \perp\!\!\!\perp \varepsilon$ or if $\mathbb{E}[\varepsilon | X] = 0$, then $cov(X, \varepsilon) = 0$.

Then, $a + bX$ is the theoretical linear regression of Y on X.

Linear regression of a r.v. Y on a r.v. X (3)

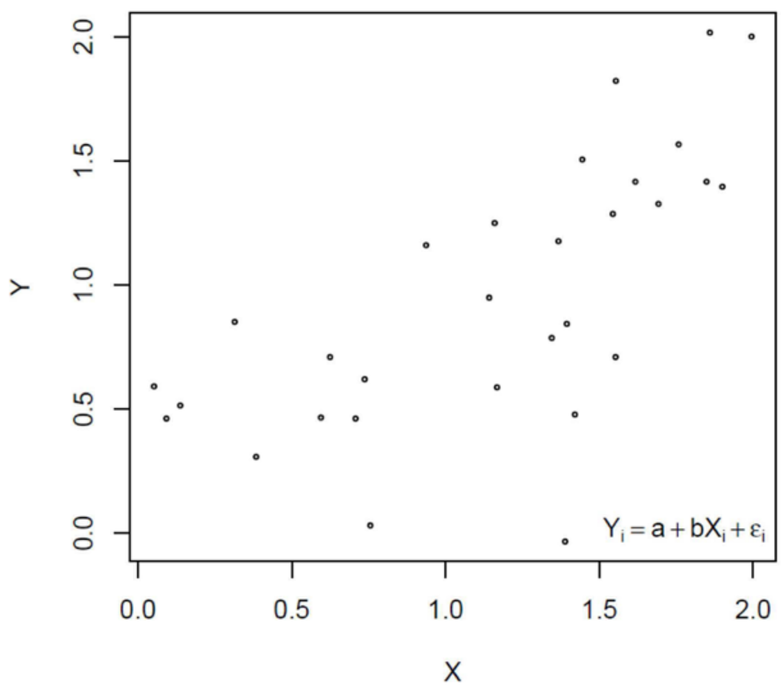
A sample of n i.i.d. random pairs $(X_1, \ldots, X_n)$ with same distribution as (X, Y) is available.

We want to estimate a and b.

Linear regression of a r.v. Y on a r.v. X (3)

A sample of n i.i.d. random pairs $(X_1, \ldots, X_n)$ with same distribution as (X, Y) is available.

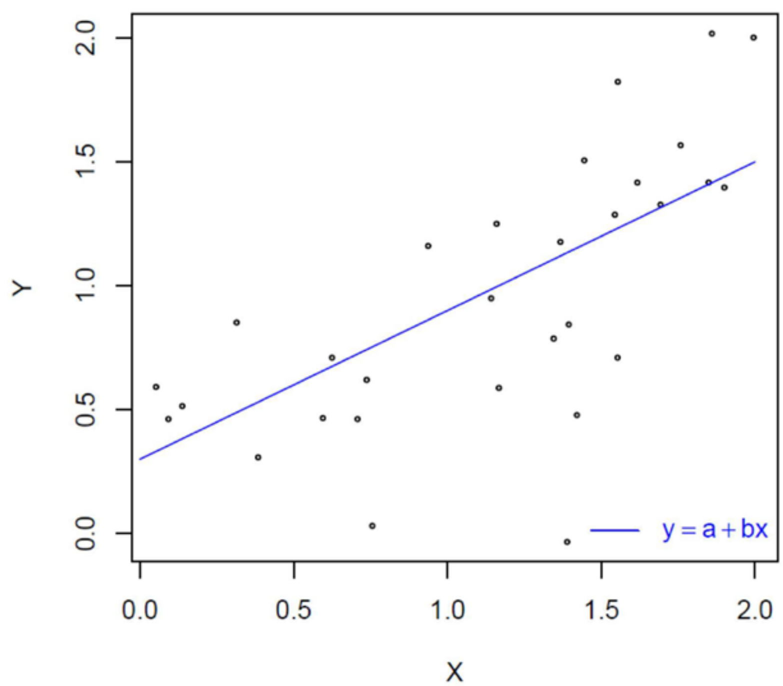
We want to estimate a and b.



Linear regression of a r.v. Y on a r.v. X (3)

A sample of n i.i.d. random pairs $(X_1, \ldots, X_n)$ with same distribution as (X, Y) is available.

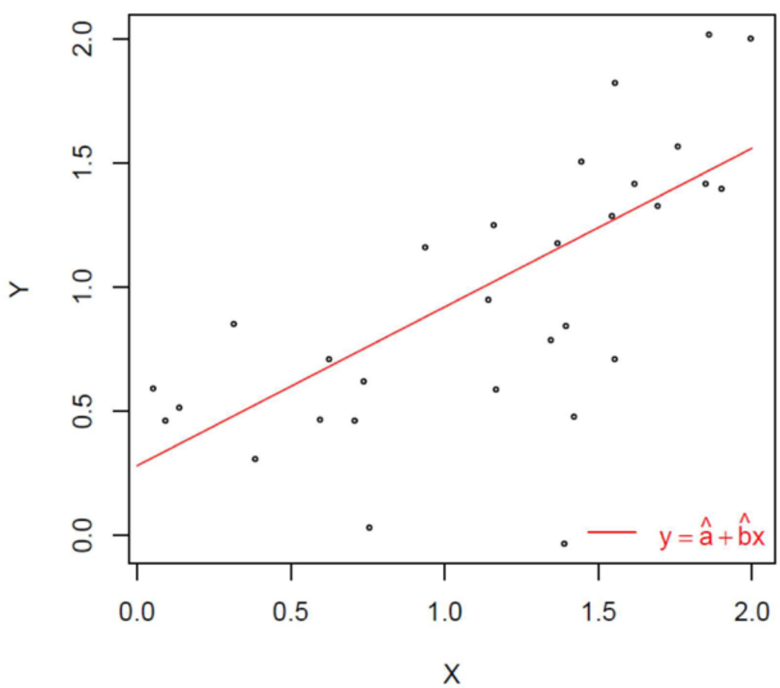
We want to estimate a and b.



Linear regression of a r.v. Y on a r.v. X (3)

A sample of n i.i.d. random pairs $(X_1, \ldots, X_n)$ with same distribution as (X, Y) is available.

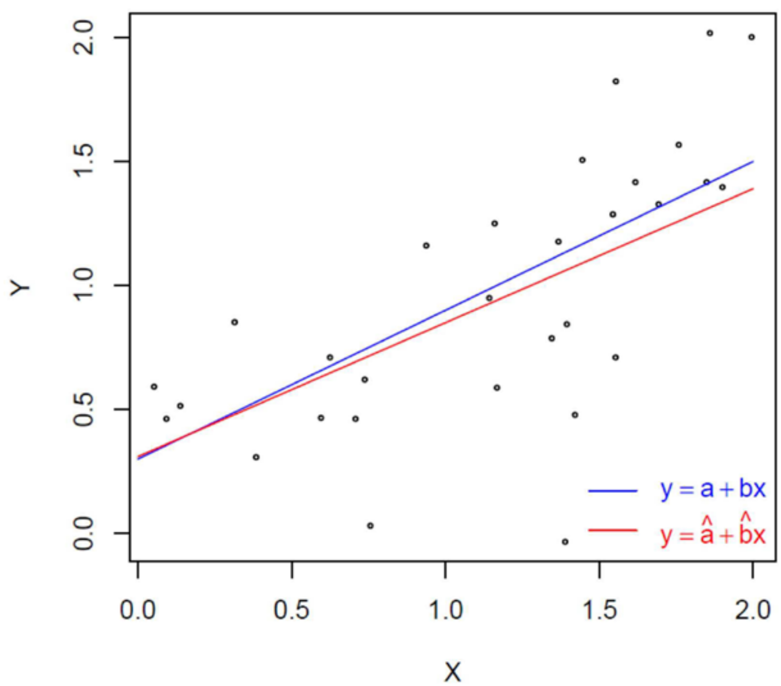
We want to estimate a and b.



Linear regression of a r.v. Y on a r.v. X (3)

A sample of n i.i.d. random pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ with same distribution as (X, Y) is available.

We want to estimate a and b.



Linear regression of a r.v. Y on a r.v. X (4)

Definition

The *least squared error (LSE)* estimator of (a, b) is the minimizer of the sum of squared errors:

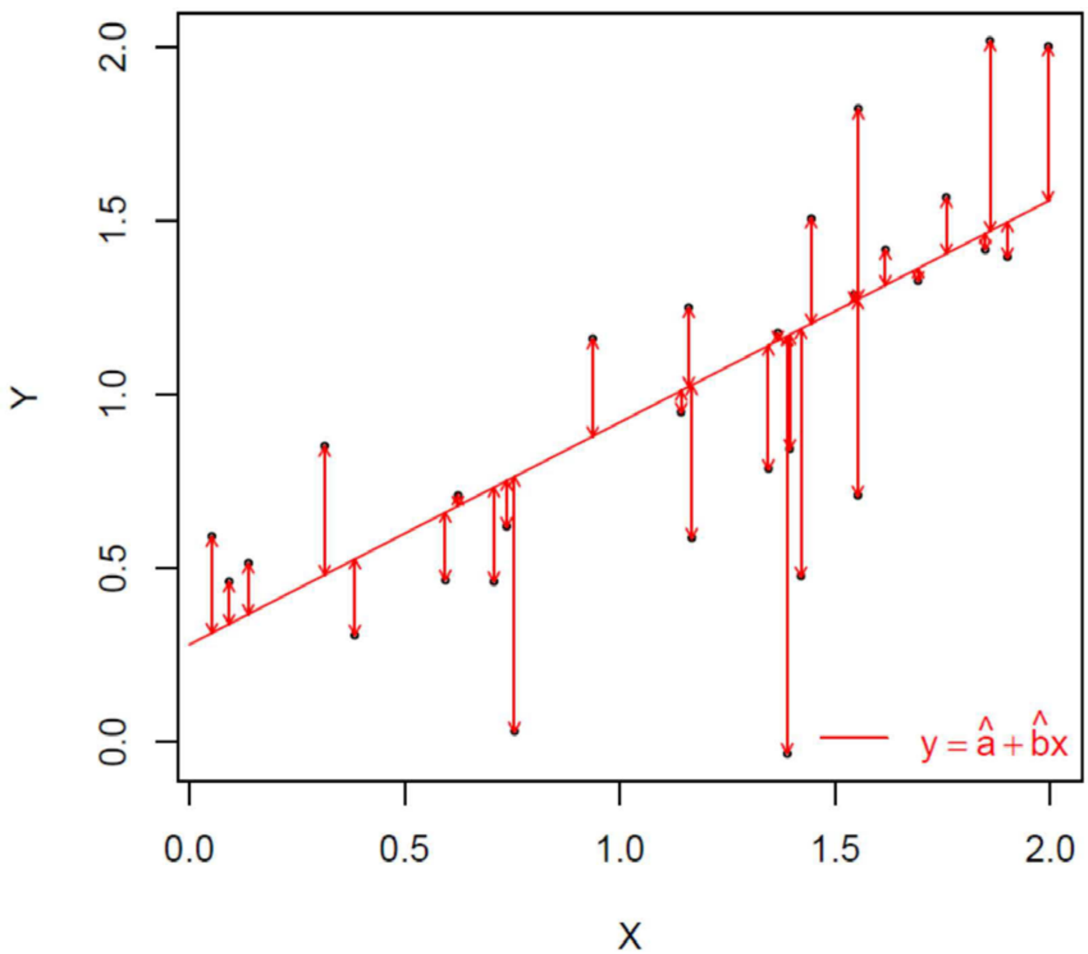
$$\sum_{i=1}^{n}(Y_i - a - bX_i)^2.$$

$(\hat{a}, \hat{b})$ is given by

$$\hat{b} = \frac{\overline{XY} - \bar{X}\bar{Y}}{\overline{X^2} - \bar{X}^2},$$

$$\hat{a} = \bar{Y} - \hat{b}\bar{X}.$$

Linear regression of a r.v. Y on a r.v. X (5)



Multivariate case (1)

$$Y_i = \mathbf{X}_i \ \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n.$$

Vector of *explanatory variables* or *covariates*: $\mathbf{X}_i \in \mathbb{R}^p$ (wlog, assume its first coordinate is 1).

Dependent variable: Y_i.

$\boldsymbol{\beta} = (a, \mathbf{b} \ \ )$; $\beta_1 (= a)$ is called the *intercept*.

$\{\varepsilon_i\}_{i=1,\ldots,n}$: noise terms satisfying $cov(\mathbf{X}_i, \varepsilon_i) = \mathbf{0}$.

Definition

The *least squared error (LSE)* estimator of $\boldsymbol{\beta}$ is the minimizer of the sum of square errors:

$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{t} \in \mathbb{R}^p}{\operatorname{argmin}} \ \sum_{i=1}^{n} (Y_i - \mathbf{X}_i \ \mathbf{t})^2$$

Multivariate case (2)

LSE in matrix form

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$.

Let $\mathbf{X}$ be the $n \times p$ matrix whose rows are $\mathbf{X}_1^\top, \ldots, \mathbf{X}_n^\top$ ($\mathbf{X}$ is called the *design*).

Let $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n$ (unobserved noise)

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

The LSE $\hat{\boldsymbol{\beta}}$ satisfies:

$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{t} \in \mathbb{R}^p}{\operatorname{argmin}} \; \|\mathbf{Y} - \mathbf{X}\mathbf{t}\|_2^2.$$

Multivariate case (3)

Assume that $rank(\mathbf{X}) = p$.

Analytic computation of the LSE:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}\quad\mathbf{X})^{-1}\mathbf{X}\quad\mathbf{Y}.$$

Geometric interpretation of the LSE

$\mathbf{X}\hat{\boldsymbol{\beta}}$ is the orthogonal projection of $\mathbf{Y}$ onto the subspace spanned by the columns of $\mathbf{X}$:

$$\mathbf{X}\hat{\boldsymbol{\beta}} = P\mathbf{Y},$$

where $P = \mathbf{X}(\mathbf{X}\quad\mathbf{X})^{-1}\mathbf{X}\quad$.

Linear regression with deterministic design and Gaussian noise (1)

Assumptions:

The design matrix $\mathbf{X}$ is deterministic and $rank(\mathbf{X}) = p$.

The model is *homoscedastic*: $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d.

The noise vector $\boldsymbol{\varepsilon}$ is Gaussian:

$$\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 I_n),$$

for some known or unknown $\sigma^2 > 0$.

Linear regression with deterministic design and Gaussian noise (2)

LSE = MLE: $\hat{\boldsymbol{\beta}} \sim \mathcal{N}_p\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X} \quad \mathbf{X})^{-1}\right)$.

Quadratic risk of $\hat{\boldsymbol{\beta}}$: $\mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2\right] = \sigma^2 \text{tr}\left((\mathbf{X} \quad \mathbf{X})^{-1}\right)$.

Prediction error: $\mathbb{E}\left[\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2\right] = \sigma^2(n-p)$.

Unbiased estimator of σ^2: $\hat{\sigma}^2 = \frac{1}{n-p}\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2$.

Theorem

$$(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}.$$

$$\hat{\boldsymbol{\beta}} \perp\!\!\!\perp \hat{\sigma}^2.$$

Significance tests (1)

Test whether the j-th explanatory variable is significant in the linear regression ($1 \leq j \leq p$).

$H_0 : \beta_j = 0$ v.s. $H_1 : \beta_j = 0$.

If γ_j is the j-th diagonal coefficient of $(\mathbf{X} \quad \mathbf{X})^{-1}$ ($\gamma_j > 0$):

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 \gamma_j}} \sim t_{n-p}.$$

Let $T_n^{(j)} = \dfrac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 \gamma_j}}$.

Test with non asymptotic level $\alpha \in (0, 1)$:

$$\delta_\alpha^{(j)} = \mathbb{1}\{|T_n^{(j)}| > q_{\frac{\alpha}{2}}(t_{n-p})\},$$

where $q_{\frac{\alpha}{2}}(t_{n-p})$ is the $(1 - \alpha/2)$-quantile of t_{n-p}.

Significance tests (2)

Test whether a **group** of explanatory variables is significant in the linear regression.

$H_0 : \beta_j = 0, \forall j \in S$ v.s. $H_1 : \exists j \in S, \beta_j = 0$, where $S \subseteq \{1, \ldots, p\}$.

Bonferroni's test: $\delta_\alpha^B = \max_{j \in S} \delta_{\alpha/k}^{(j)}$, where $k = |S|$.

δ_α has non asymptotic level at most α.

More tests (1)

Let G be a $k \times p$ matrix with $rank(G) = k$ ($k \leq p$) and $\boldsymbol{\lambda} \in \mathbb{R}^k$.

Consider the hypotheses:

$$H_0 : G\boldsymbol{\beta} = \boldsymbol{\lambda} \text{ v.s. } H_1 : G\boldsymbol{\beta} = \boldsymbol{\lambda}.$$

The setup of the previous slide is a particular case.

If H_0 is true, then:

$$G\hat{\boldsymbol{\beta}} - \boldsymbol{\lambda} \sim \mathcal{N}_k \quad 0, \sigma^2 G(\mathbf{X} \quad \mathbf{X})^{-1} G \qquad ,$$

and

$$\sigma^{-2}(G\hat{\boldsymbol{\beta}} - \boldsymbol{\lambda}) \quad G(\mathbf{X} \quad \mathbf{X})^{-1} G \quad ^{-1} (G\boldsymbol{\beta} - \boldsymbol{\lambda}) \sim \chi^2_k.$$

More tests (2)

Let $S_n = \frac{1}{\hat{\sigma}^2} \frac{(G\hat{\boldsymbol{\beta}} - \boldsymbol{\lambda}) \quad \left(G(\mathbf{X} \quad \mathbf{X})^{-1} G \quad \right)^{-1} (G\boldsymbol{\beta} - \boldsymbol{\lambda})}{k}$.

If H_0 is true, then $S_n \sim F_{k,n-p}$.

Test with non asymptotic level $\alpha \in (0, 1)$:

$$\delta_\alpha = \mathbb{1}\{S_n > q_\alpha(F_{k,n-p})\},$$

where $q_\alpha(F_{k,n-p})$ is the $(1-\alpha)$-quantile of $F_{k,n-p}$.

Definition

The *Fisher distribution with* p *and* q *degrees of freedom*, denoted by $F_{p,q}$, is the distribution of $\frac{U/p}{V/q}$, where:

$U \sim \chi^2_p$, $V \sim \chi^2_q$,

$U \perp\!\!\!\perp V$.

Concluding remarks

Linear regression exhibits correlations, **NOT** causality

Normality of the noise: One can use goodness of fit tests to test whether the residuals $\hat{\varepsilon}_i = Y_i - \mathbf{X}_i^\top \hat{\boldsymbol{\beta}}$ are Gaussian.

Deterministic design: If $\mathbf{X}$ is not deterministic, all the above can be understood conditionally on $\mathbf{X}$, if the noise is assumed to be Gaussian, conditionally on X.

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 $X \quad X$.

   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 ? $\mathbf{X}\boldsymbol{\beta}$ is still identified.

$\hat{\mathbf{Y}}$: orthogonal projection of $\mathbf{Y}$ onto the linear span of the columns of $\mathbf{X}$.

$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X} \quad \mathbf{X})^{\dagger}\mathbf{X}\mathbf{Y}$, where $A^{\dagger}$ stands for the (Moore-Penrose) pseudo inverse of a matrix A.

Similarly as before, if $k = rank(\mathbf{X})$:

$$\frac{\|\hat{\mathbf{Y}} - \mathbf{Y}\|_2^2}{\sigma^2} \sim \chi^2_{n-k},$$

$$\|\hat{\mathbf{Y}} - \mathbf{Y}\|_2^2 \perp\!\!\!\perp \hat{\mathbf{Y}}.$$

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 $\boldsymbol{\beta}$ are nonzero (but we do not know which ones).

Based on the sample, select a subset of covariates and estimate the corresponding coordinates of $\boldsymbol{\beta}$.

For $S \subseteq \{1, \ldots, p\}$, let

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

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:

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

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

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

If $\lambda = \hat{\sigma}^2 \log n$, this is the *BIC* criterion.

Linear regression in high dimension (4)

Each of these criteria is equivalent to finding $\boldsymbol{\beta} \in \mathbb{R}^p$ 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 $\hat{\boldsymbol{\beta}}_S$, for $S \subseteq \{1, \ldots, p\}$).

Lasso estimator:

$$\text{replace} \|\mathbf{b}\|_0 = \sum_{j=1}^{p} \mathbb{1}\{b_j = 0\} \qquad \text{with} \qquad \|\mathbf{b}\|_1 = \sum_{j=1}^{p} |b_j|$$

and the problem becomes convex.

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

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

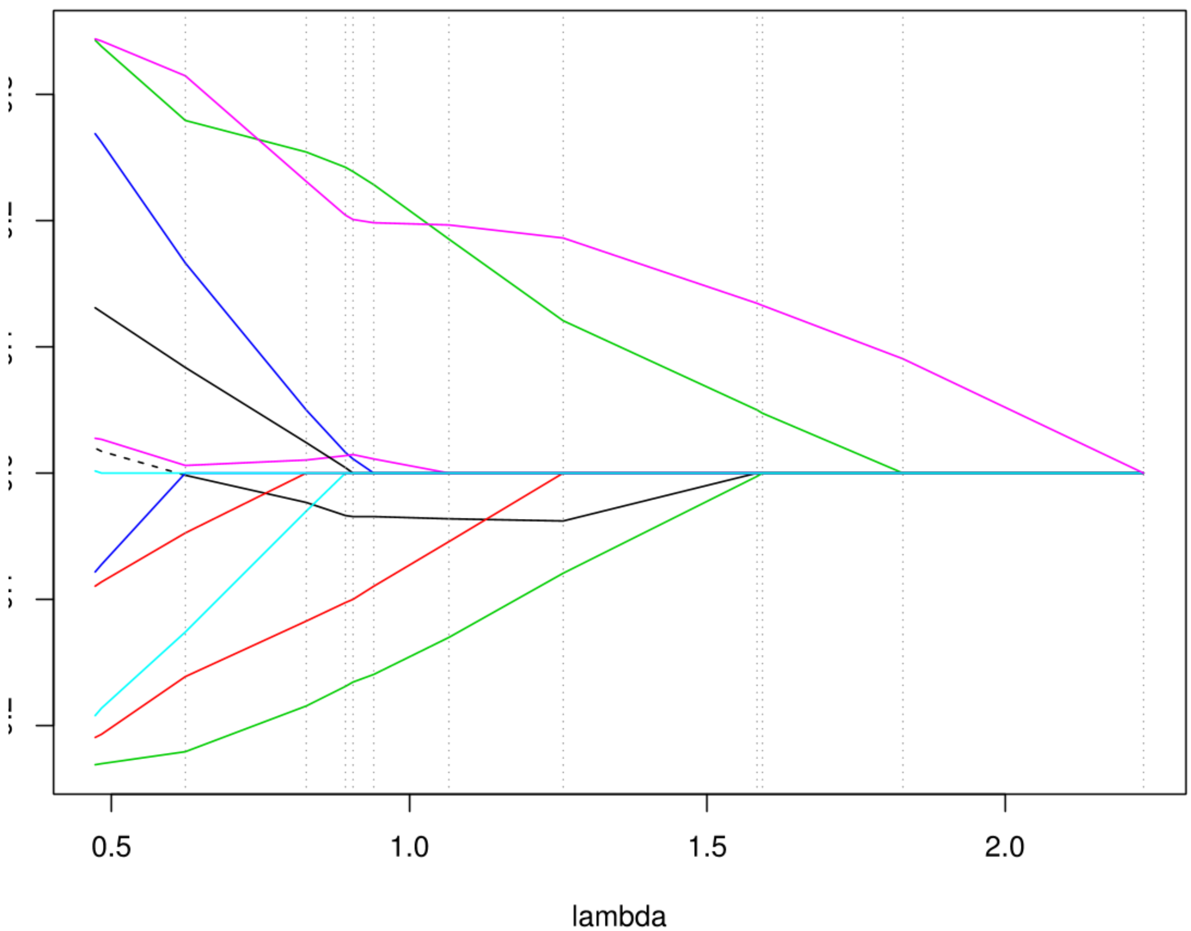
Linear regression in high dimension (5)

How to choose λ ?

This is a difficult question (see grad course 18.657: "High-dimensional statistics" in Spring 2017).

A *good choice* of λ with lead to an estimator $\hat{\boldsymbol{\beta}}$ that is very close to $\boldsymbol{\beta}$ and will allow to recover the subset S^* of all $j \in \{1, \dots, p\}$ for which $\boldsymbol{\beta}_j = 0$, with high probability.

Linear regression in high dimension (6)



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 $p = 1$ 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)

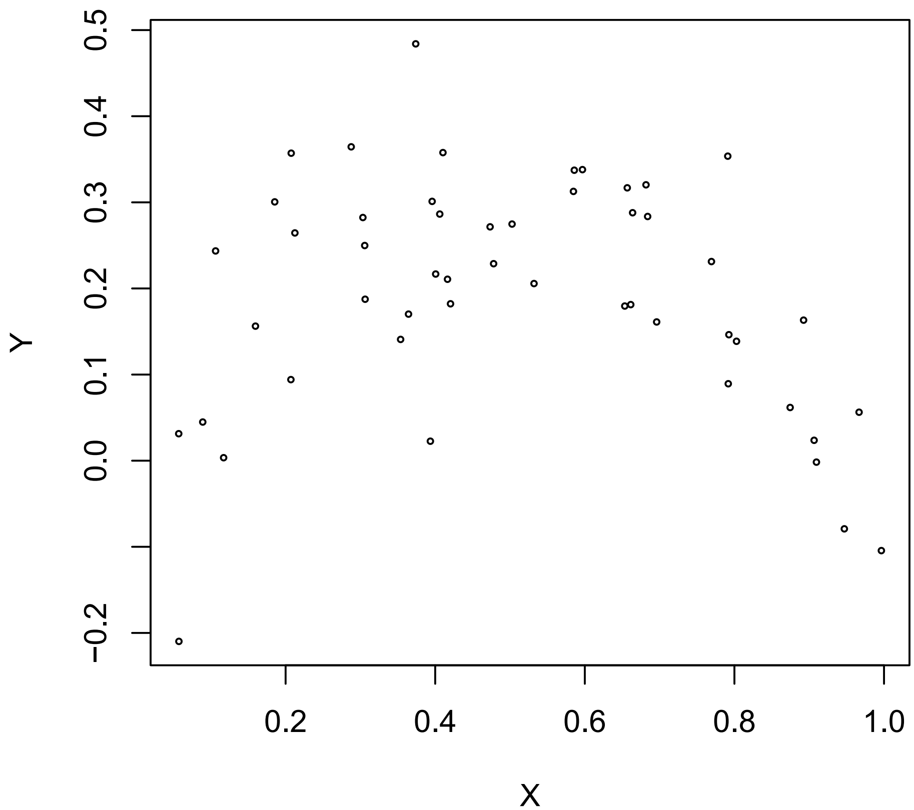
Let $h > 0$: the window's size (or bandwidth).

Let $I_x = \{i = 1, \dots, n : |X_i - x| < h\}$.

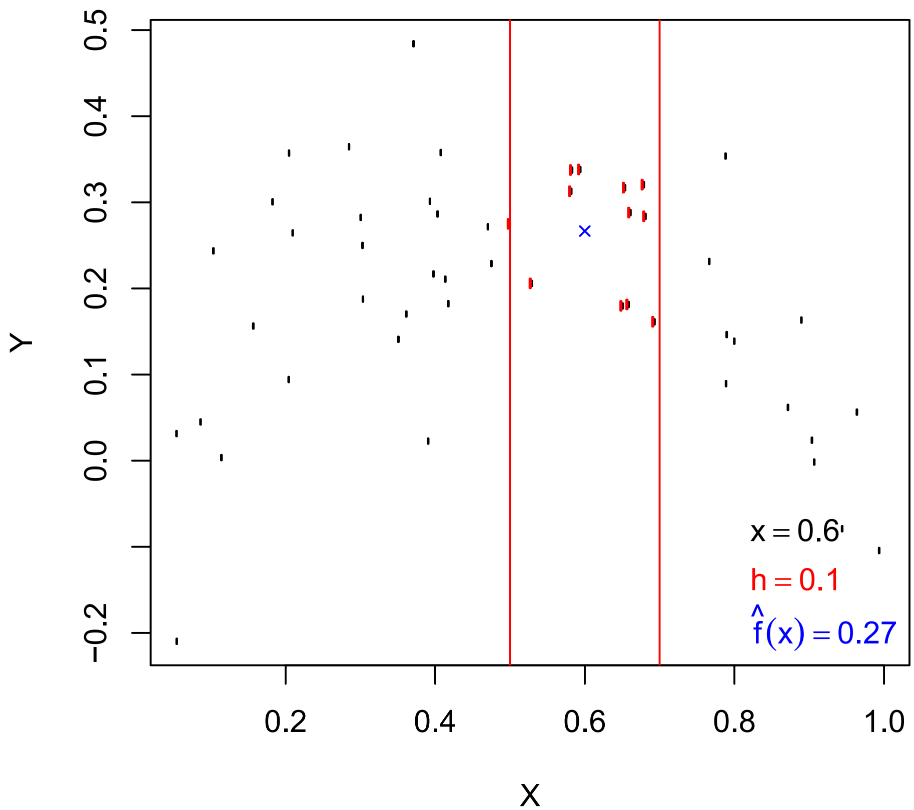
Let $\hat{f}_{n,h}(x)$ be the average of $\{Y_i : i \in I_x\}$.

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

Nonparametric regression (5)



Nonparametric regression (6)



Nonparametric regression (7)

How to choose h ?

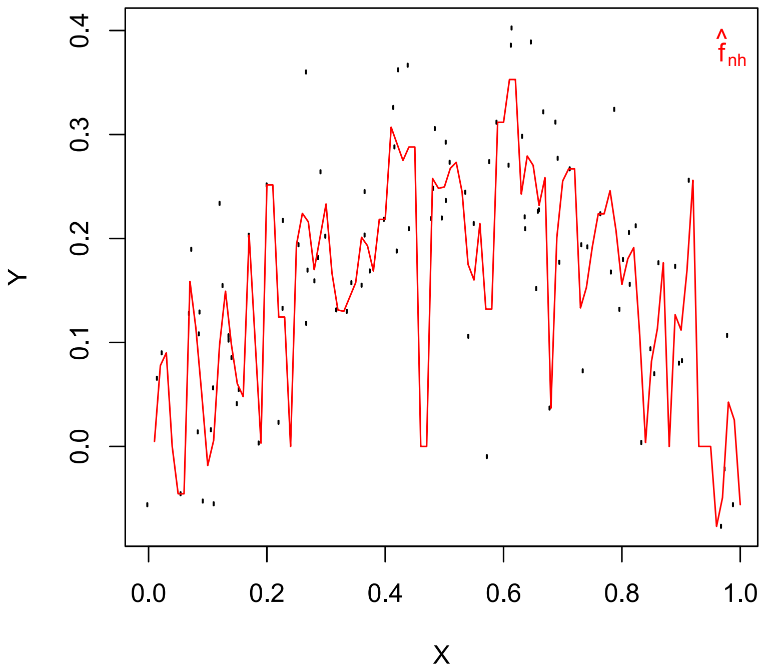
If $h \to 0$: overfitting the data;

If $h \to \infty$: underfitting, $\hat{f}_{n,h}(x) = \bar{Y}_n$.

Nonparametric regression (8)

Example:

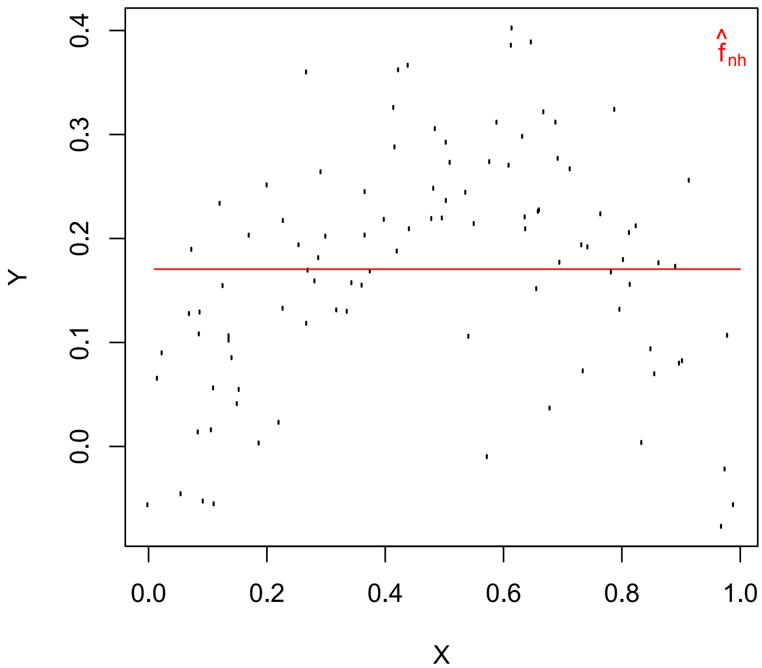
$n = 100,\ f(x) = x(1-x),$
$h = .005.$



Nonparametric regression (9)

Example:

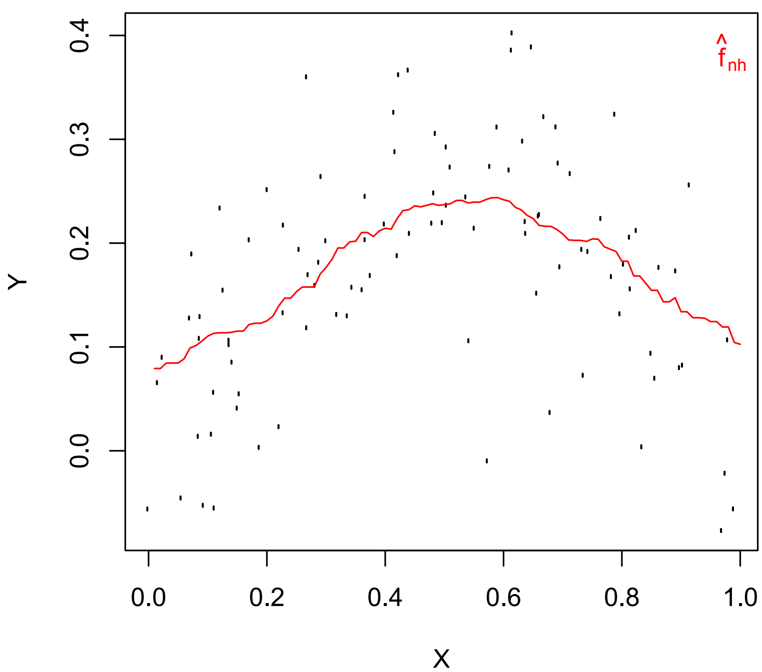
$n = 100,\ f(x) = x(1-x),$
$h = 1.$



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*.