Apprentissage statistique et grande dimension
Chap. 1 : Introduction

M1 MA

Plan du chapitre

Qu'est-ce que l'apprentissage statistique ?

Sur et sous-apprentissage

Problématiques liées à la grande dimension

Plan du cours et informations diverses

Plan

Apprentissage statistique

- Conception, analyse et implémentation de méthodes permettant à une machine de résoudre un problème complexe.
- À l'interface entre :
  - Informatique : programmation des algorithmes, présence de jeux de données complexes, volumineux (Big Data). Quels sont les problèmes techniquement faisables/infaisables ?
  - Mathématiques (statistique) : définition des méthodes, propriétés théoriques (convergence).

Apprentissage supervisé/non supervisé (I)

- Apprentissage supervisé On s'intéresse au lien entre une quantité d'intérêt Y et des variables explicatives $X = (X^1, \ldots, X^n)$. L'objectif est de construire, à partir d'observations répétées $\{(Y_i, X_i), i = 1, \ldots, n\}$ du couple (Y, X), une fonction de prédiction $\widehat{f}$ telle que $\widehat{f}(X)$ soit proche (dans un sens à préciser) de Y.

  Exemples :

  - Régression linéaire

    $$Y_i = X_i \beta^* + \varepsilon_i,$$

    $\beta^* \in \mathbb{R}^d$, $\{\varepsilon_i\} \sim_{i.i.d.} \mathcal{N}(0, \sigma^2)$.

Apprentissage supervisé/non supervisé (II)

- Classification binaire

$$Y_i \in \{0, 1\},$$

$\widehat{f} : \mathbb{R}^d \to \{0, 1\}$ est une règle de décision minimisant (par exemple) l'erreur de classification

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\widehat{f}(X_i) \neq Y_i}$$

Exemple de problème de classification supervisée binaire

	Ciel	Temperature	Humidite	Vent	Jouer
1	Soleil	27.50	85	Faible	Non
2	Soleil	25.00	90	Fort	Non
3	Couvert	26.50	86	Faible	Oui
4	Pluie	20.00	96	Faible	Oui
5	Pluie	19.00	80	Faible	Oui
6	Pluie	17.50	70	Fort	Non
7	Couvert	17.00	65	Fort	Oui
8	Soleil	21.00	95	Faible	Non
9	Soleil	19.50	70	Faible	Oui
10	Pluie	22.50	80	Faible	Oui
11	Soleil	22.50	70	Fort	Oui
12	Couvert	21.00	90	Fort	Oui
13	Couvert	25.50	75	Faible	Oui
14	Pluie	20.50	91	Fort	Non

Exemple de problème de classification supervisée binaire

	Ciel	Temperature	Humidite	Vent	Jouer
1	Soleil	27.50	85	Faible	Non
2	Soleil	25.00	90	Fort	Non
3	Couvert	26.50	86	Faible	Oui
4	Pluie	20.00	96	Faible	Oui
5	Pluie	19.00	80	Faible	Oui
6	Pluie	17.50	70	Fort	Non
7	Couvert	17.00	65	Fort	Oui
8	Soleil	21.00	95	Faible	Non
9	Soleil	19.50	70	Faible	Oui
10	Pluie	22.50	80	Faible	Oui
11	Soleil	22.50	70	Fort	Oui
12	Couvert	21.00	90	Fort	Oui
13	Couvert	25.50	75	Faible	Oui
14	Pluie	20.50	91	Fort	Non

variables explicatives variables d'intérêt

Apprentissage supervisé/non supervisé (III)

- Apprentissage non supervisé On observe seulement $(X_1, ..., X_n)$, $X_i \in \mathbb{R}^d$ et l'objectif est de décrire la distribution jointe ou d'estimer certaines des propriétés des données. Par exemple, si $\{X_i\}_{i=1,...,n}$ i.i.d.,
  - en petite dimension ($d \leq 3$) : estimer la densité de X_1 (cf cours de statistique non-paramétrique),
  - en grande dimension : représenter les données dans un espace de dimension plus petite (analyse en composantes principales, cf cours analyse des données),
  - classer les données en sous-groupes homogènes (classification non supervisée).

Apprentissage non-supervisé : Exemple de données

Consommation de protéines en Europe : pour chacun des 25 pays de l'union européenne, relevé de la consommation moyenne journalière des 9 types de protéines.

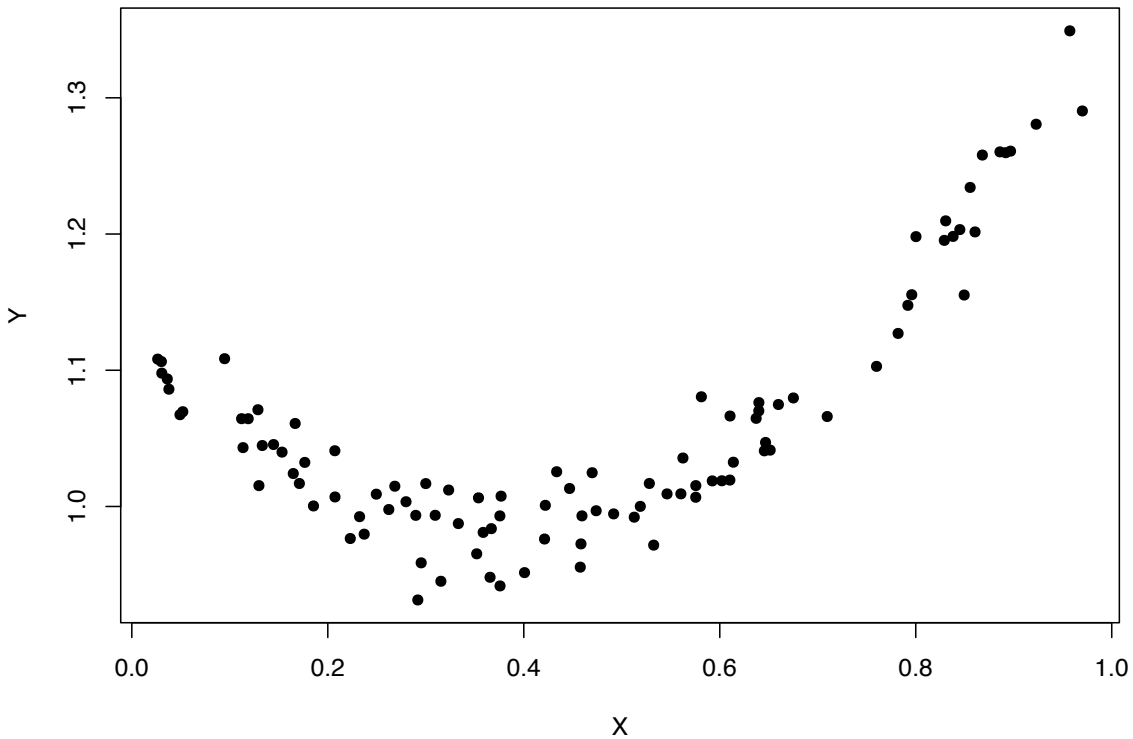
Prot. / Pays	viandr	viandb	oeuf	lait	poisson	céréale	féculent	...
Bulgaria	7,8	6,0	1,6	8,3	1,2	56,7	1,1	...
Yugoslavia	4,4	5,0	1,2	9,5	0,6	55,9	3,0	...
Romania	6,2	6,3	1,5	11,1	1,0	49,6	3,1	...
Germany	11,4	12,5	4,1	18,8	3,4	18,6	5,2	...
France	18,0	9,9	3,3	19,5	5,7	28,1	4,8	...
Norway	9,4	4,7	2,7	23,3	9,7	23,0	4,6	...
Greece	10,2	3,0	2,8	17,6	5,9	41,7	2,2	...
...	...	...	...	...	...	...	...	...

Quelle classification, et donc quel profilage des pays, en fonction de la consommation de protéines ?

Plan

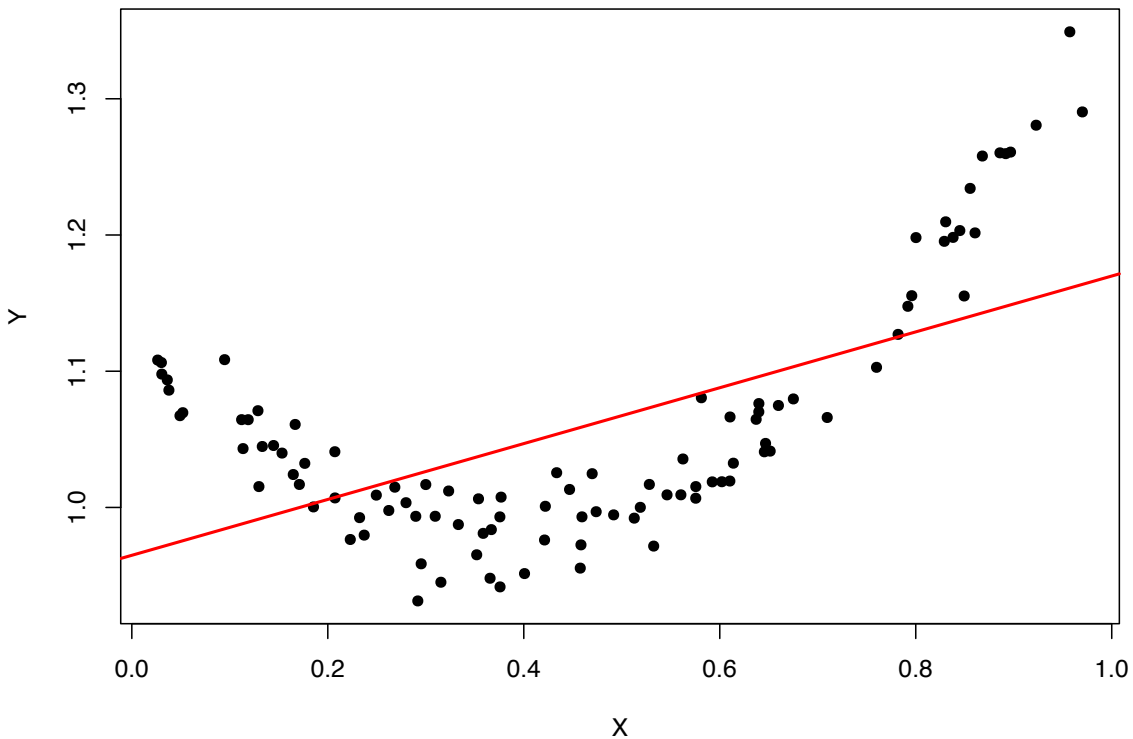
Exemple : régression linéaire (I)

Données : $\{X_i, Y_i\}$, $X_i \in \mathbb{R}$ représentées ci-dessous.



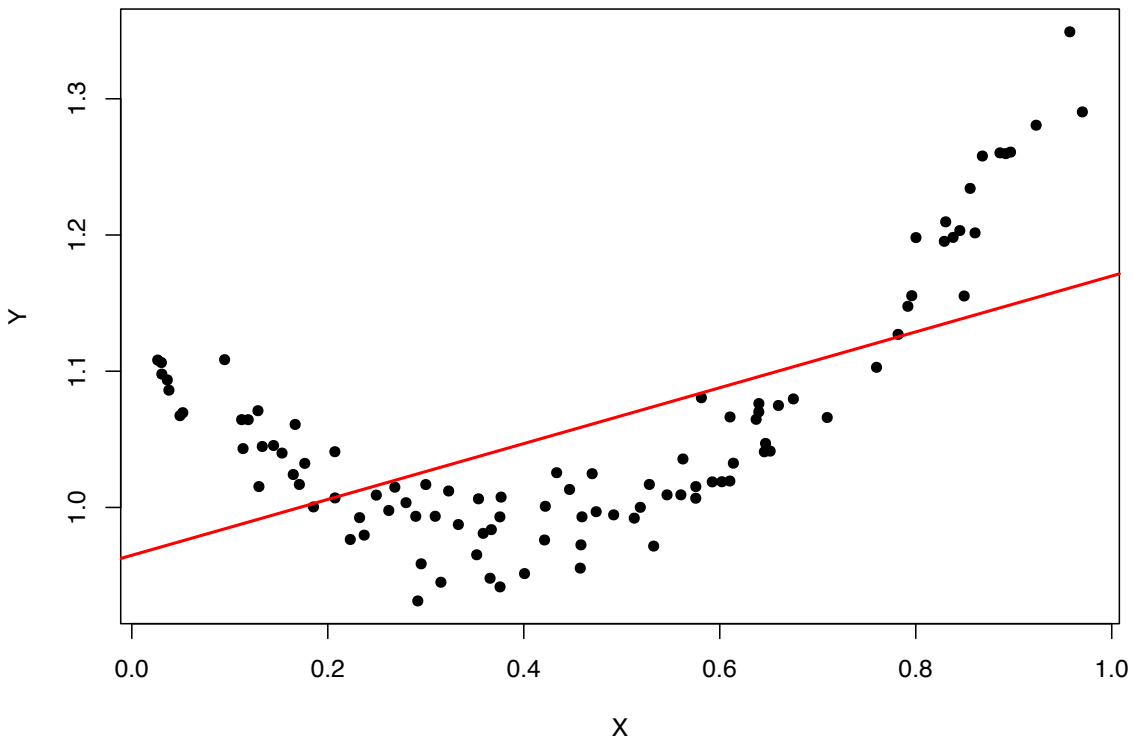
Exemple : régression linéaire (II)

On propose le modèle : $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$.



Exemple : régression linéaire (II)

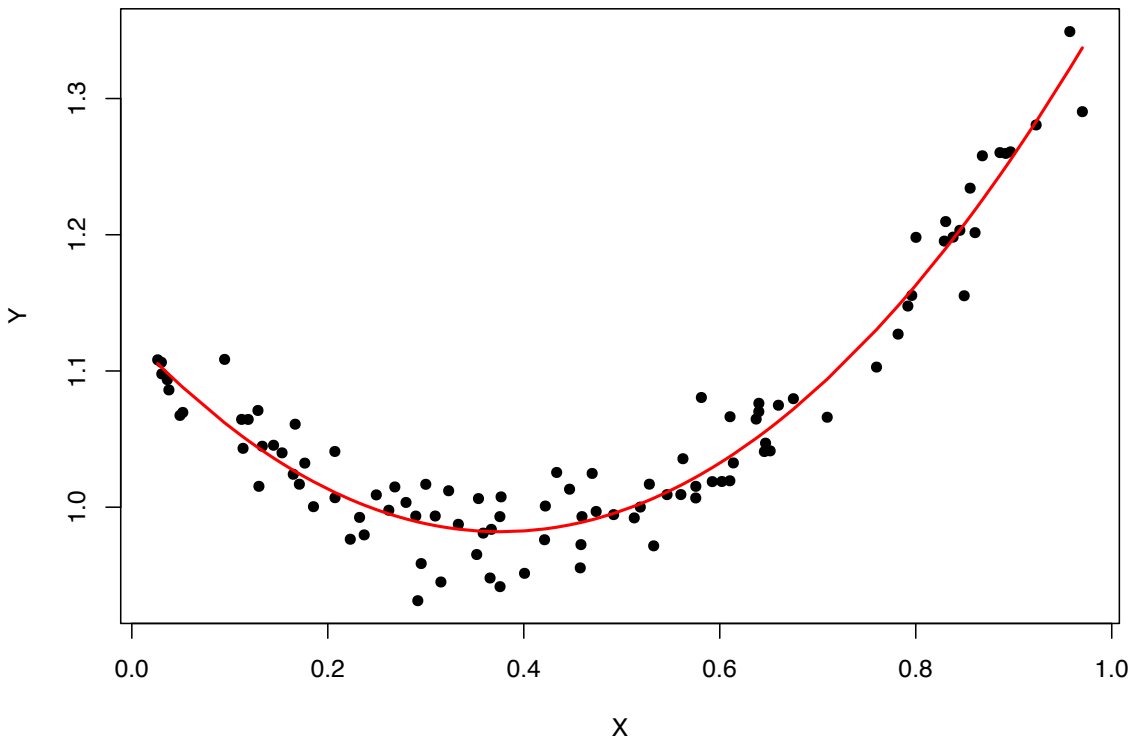
On propose le modèle : $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$.



Modèle trop simple : sous-apprentissage.

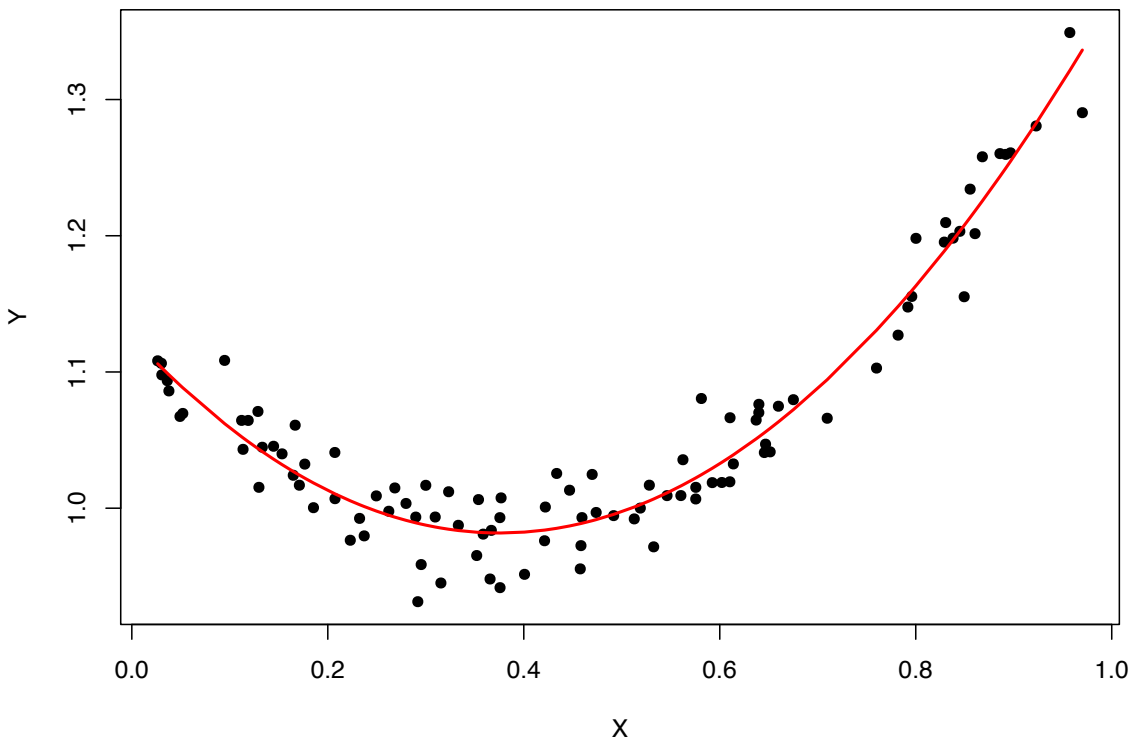
Exemple : régression linéaire (III)

Modèle linéaire quadratique : $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$.



Exemple : régression linéaire (IV)

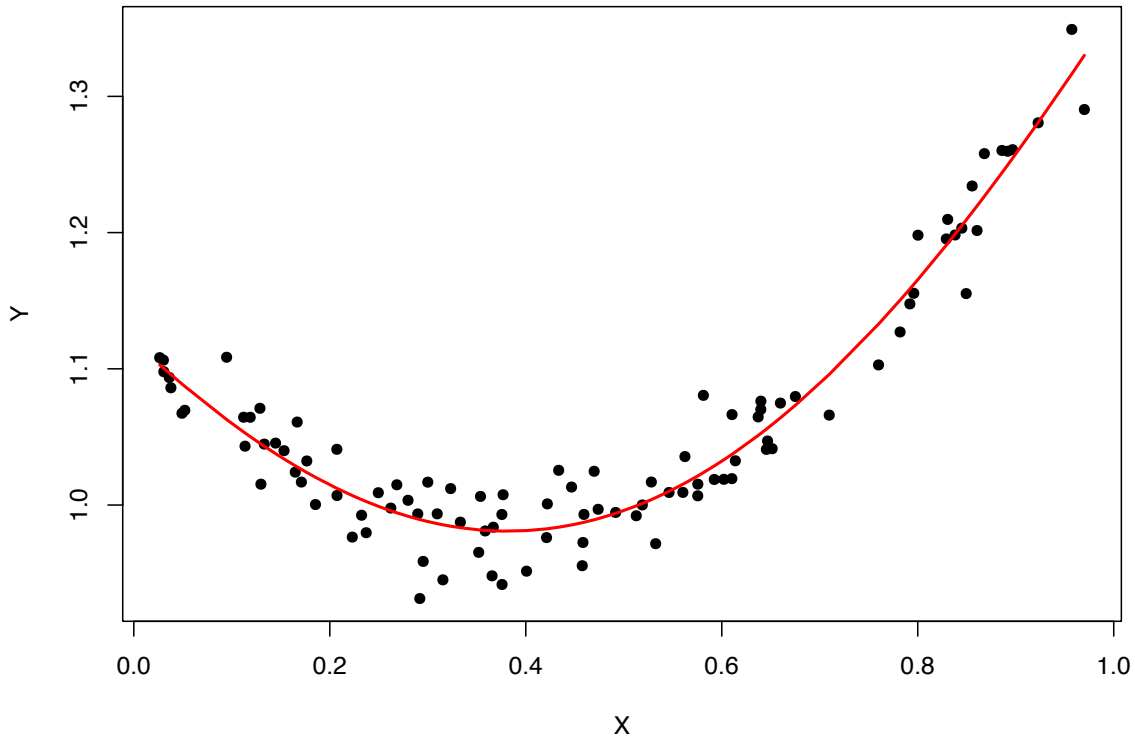
Modèle linéaire cubique : $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \varepsilon_i$.



Exemple : régression linéaire (V)

Modèle linéaire polynomial ($p = 4$) :

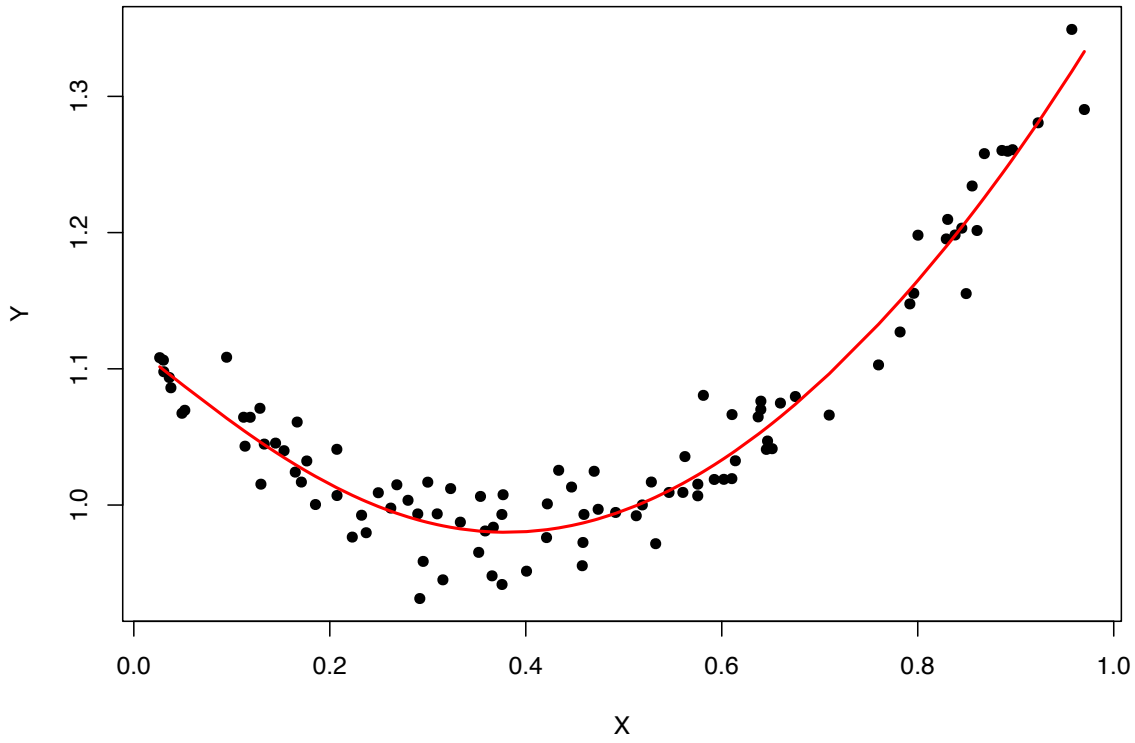
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \beta_4 X_i^4 + \varepsilon_i.$$



Exemple : régression linéaire (VI)

Modèle linéaire polynomial ($p = 5$) :

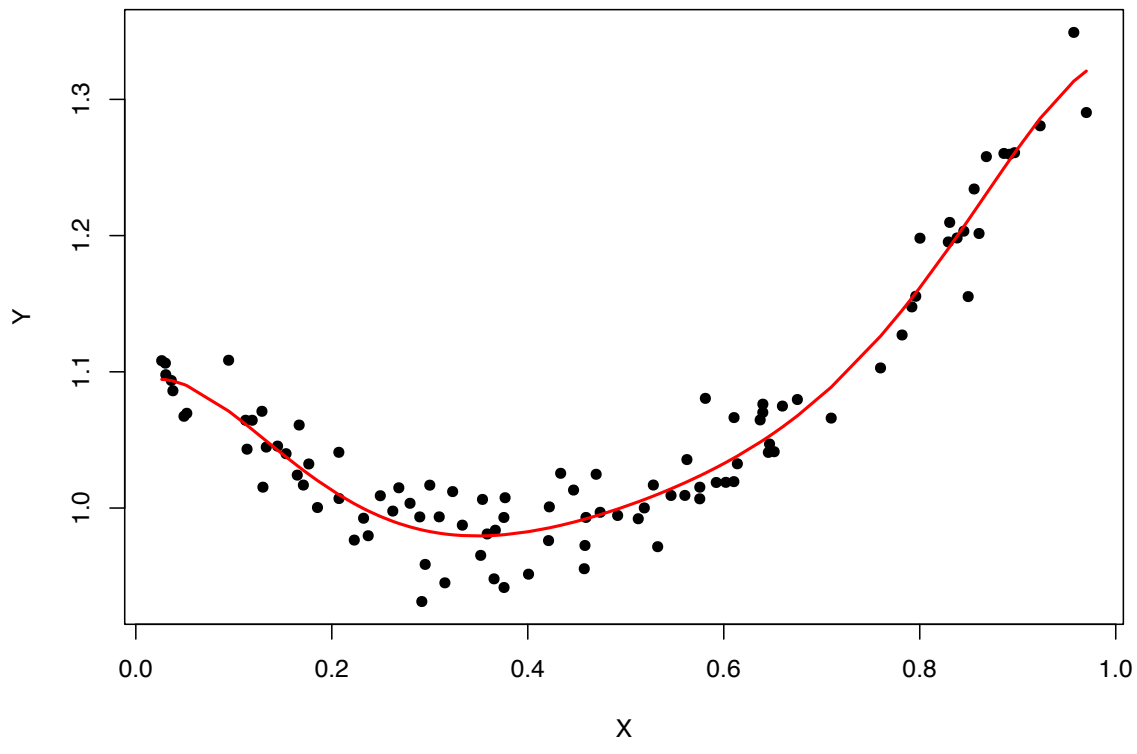
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \beta_4 X_i^4 + \beta_5 X_i^5 + \varepsilon_i.$$



Exemple : régression linéaire (VII)

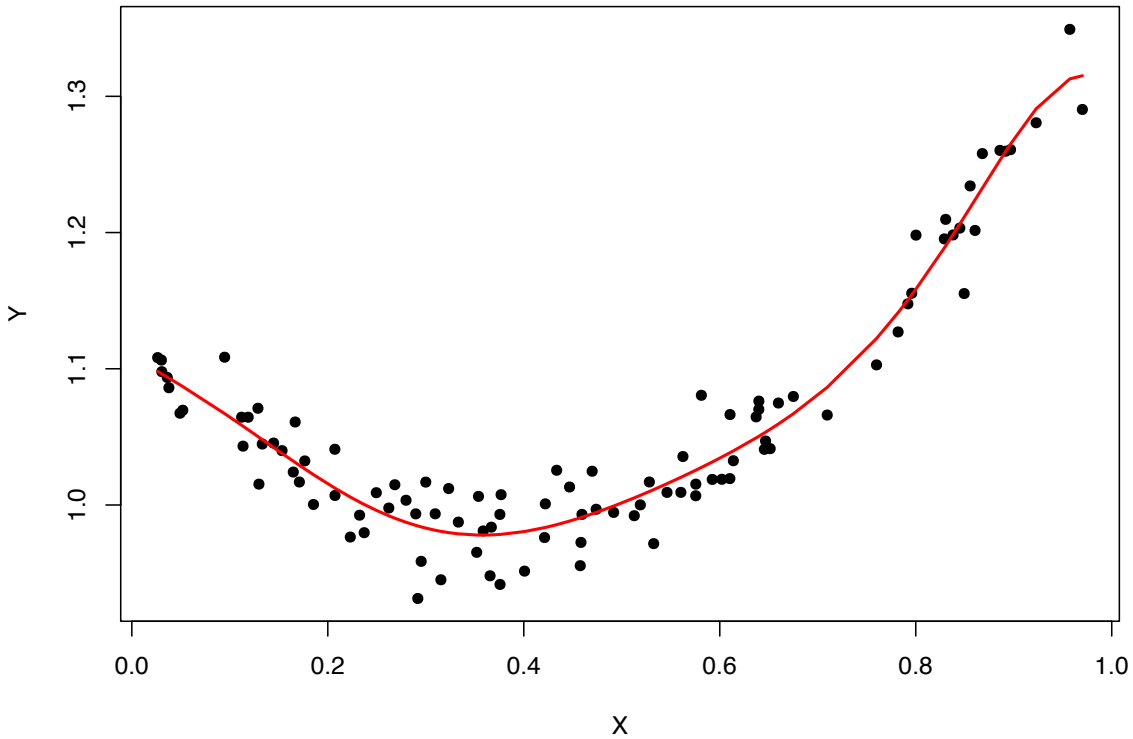
Modèle linéaire polynomial ($p = 6$) :

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \beta_4 X_i^4 + \beta_5 X_i^5 + \beta_6 X_i^6 + \varepsilon_i.$$



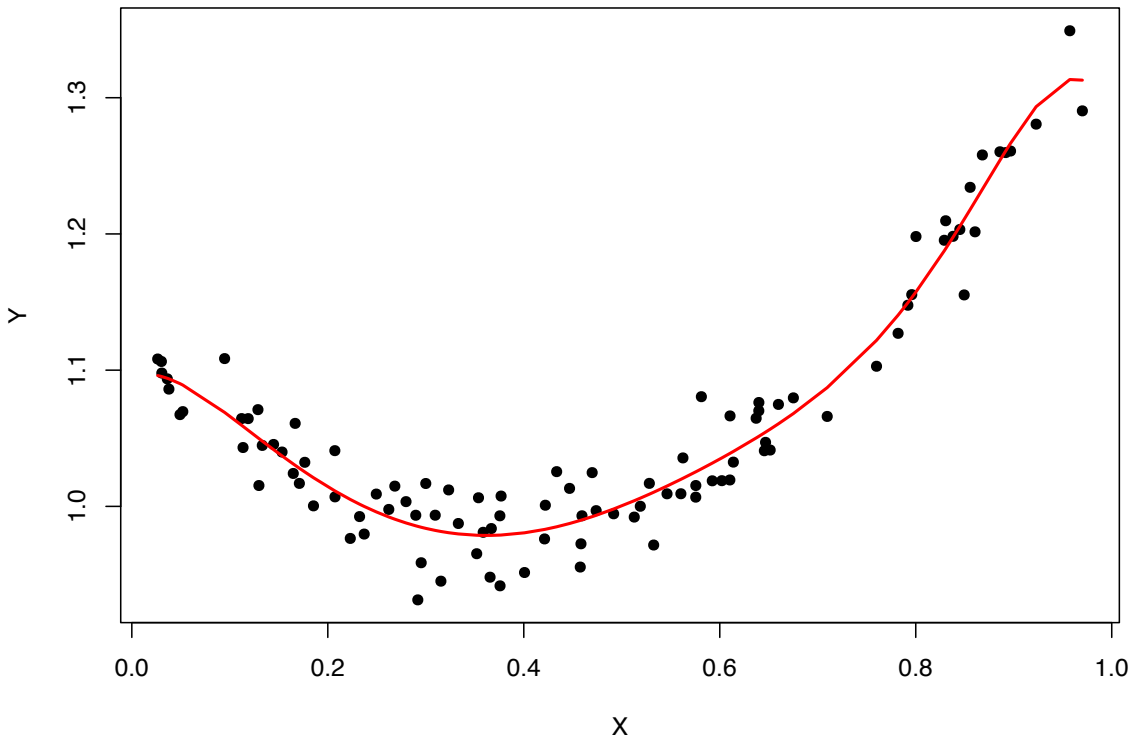
Exemple : régression linéaire (VIII)

Modèle linéaire polynomial ($p = 7$) : $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_i^j + \varepsilon_i$.



Exemple : régression linéaire (IX)

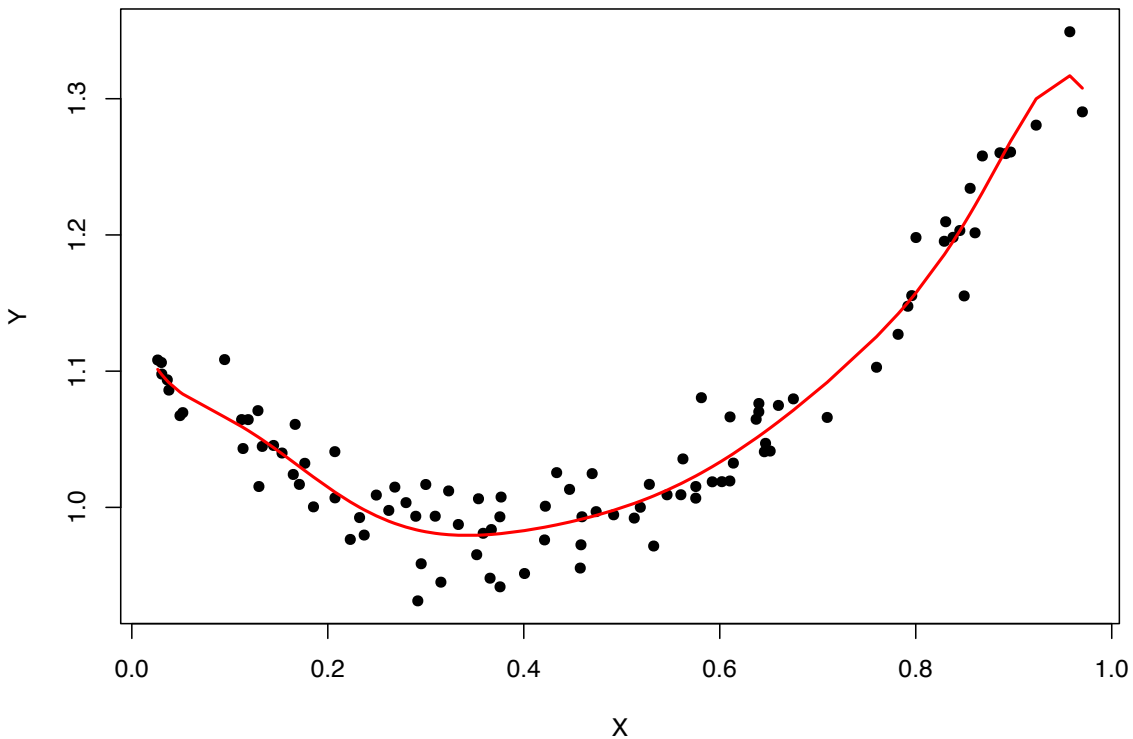
Modèle linéaire polynomial ($p = 8$) : $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_i^j + \varepsilon_i$.



Modèle trop complexe : sur-apprentissage.

Exemple : régression linéaire (X)

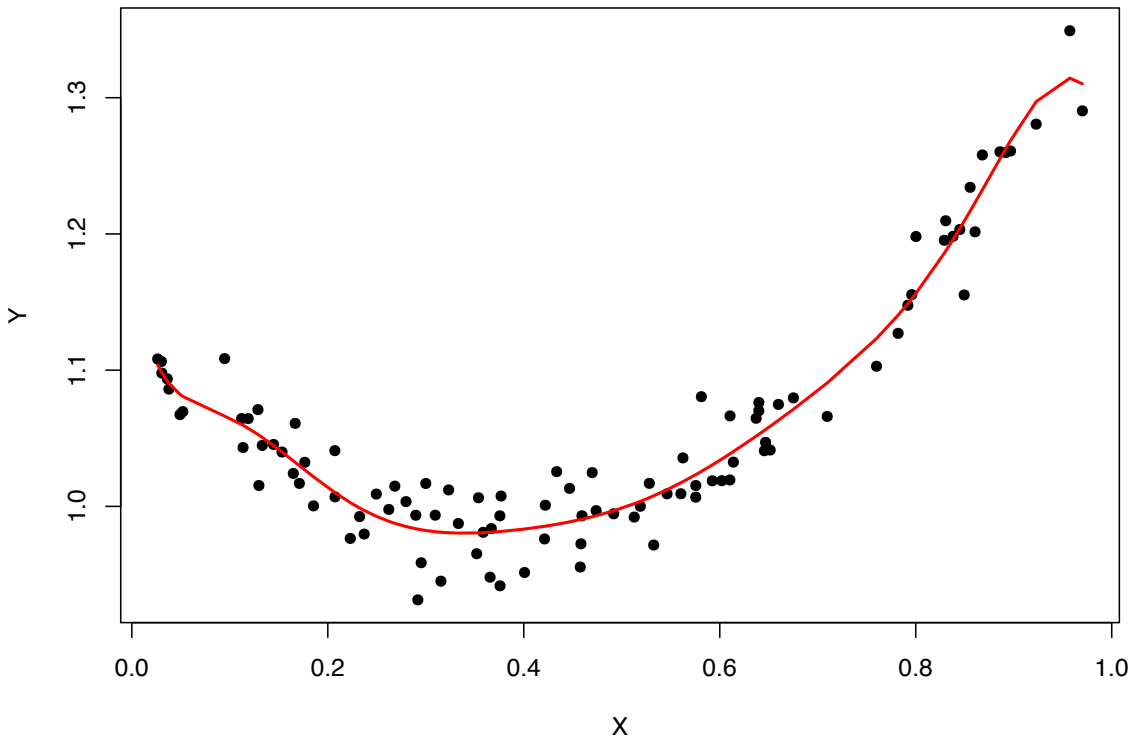
Modèle linéaire polynomial ($p = 9$) : $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_i^j + \varepsilon_i$.



Modèle trop complexe : sur-apprentissage.

Exemple : régression linéaire (XI)

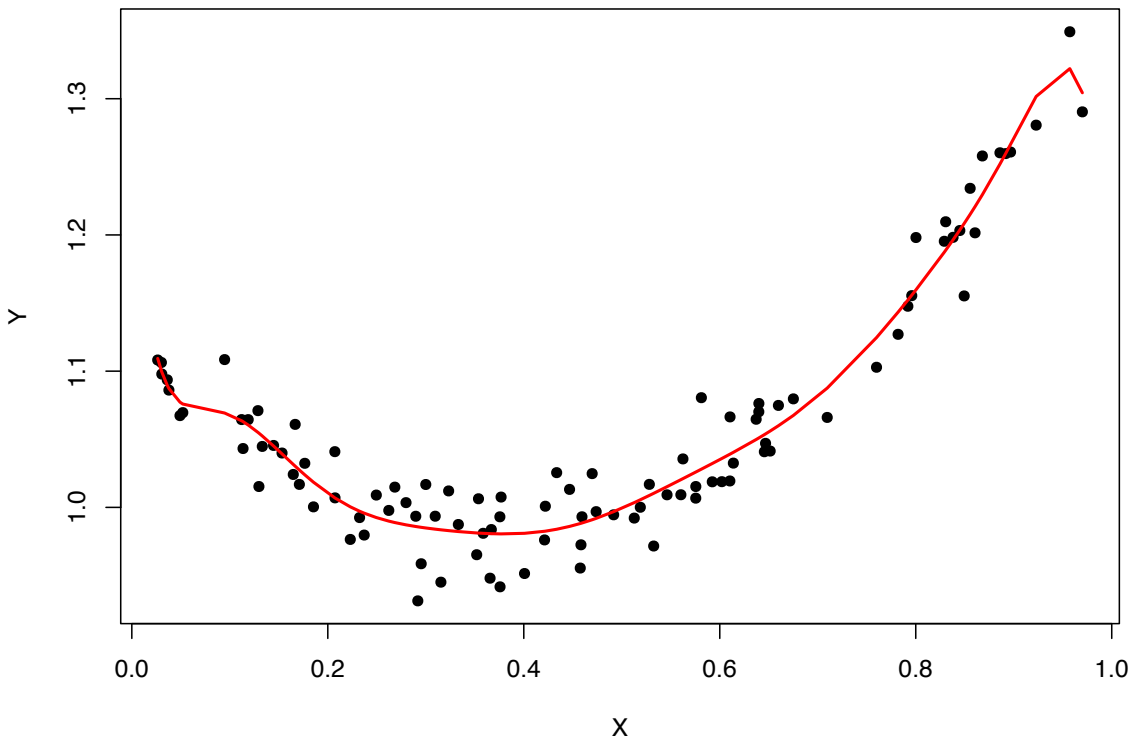
Modèle linéaire polynomial ($p = 10$) : $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_i^j + \varepsilon_i$.



Modèle trop complexe : sur-apprentissage.

Exemple : régression linéaire (XII)

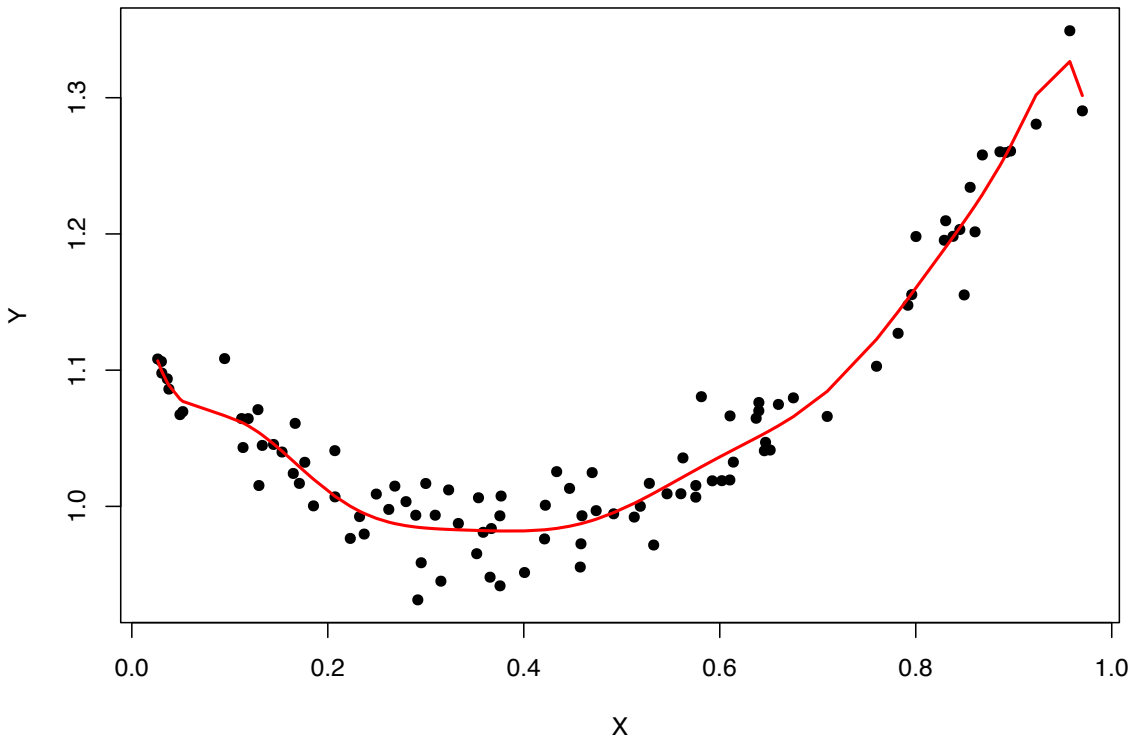
Modèle linéaire polynomial ($p = 11$) : $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_i^j + \varepsilon_i$.



Modèle trop complexe : sur-apprentissage.

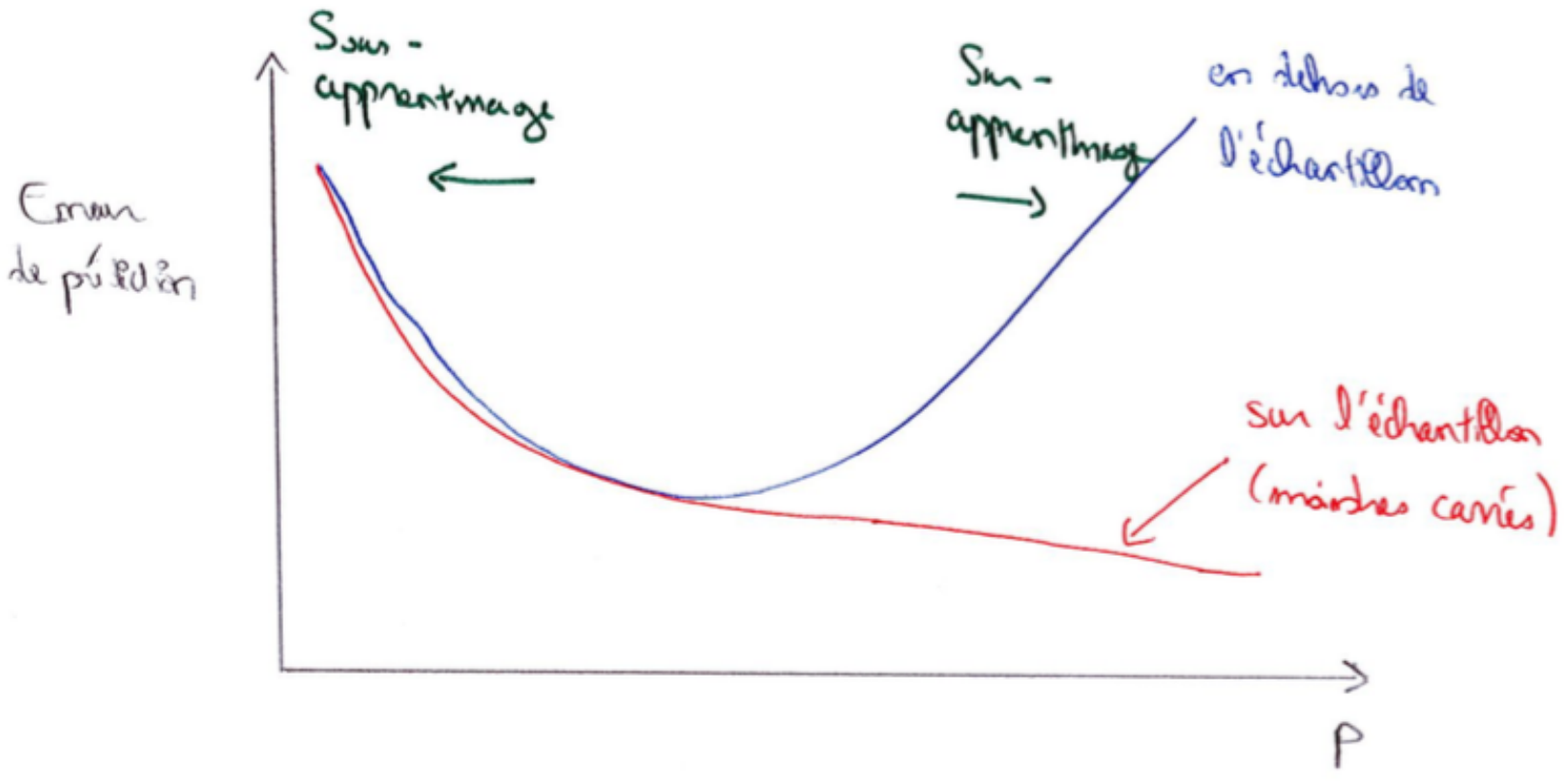
Exemple : régression linéaire (XIII)

Modèle linéaire polynomial ($p = 12$) : $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_i^j + \varepsilon_i$.



Modèle trop complexe : sur-apprentissage.

Sur et sous-apprentissage



Échantillon d'apprentissage et échantillon de test (I)

Comment choisir p ?

- Soit, pour $i = 1, ..., n$, $\widehat{Y}_i^{(p)}$ la prédiction faite à partir du modèle $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_i^j + \varepsilon_i$. Minimiser en p le critère des moindres carrés

$$\frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i^{(p)} \right)^2 ,$$

sélectionne le plus grand $p \Rightarrow$ sur-apprentissage.

Échantillon d'apprentissage et échantillon de test (II)

En pratique, pour calibrer les paramètres du modèle, on sépare l'échantillon en deux sous-parties disjointes : l'échantillon d'apprentissage I_{app} et l'échantillon de validation I_{valid}.

1. On apprend le modèle, c'est-à-dire que l'on estime les coefficients $\beta_0, ..., \beta_p$, pour plusieurs valeurs de p à l'aide de l'échantillon d'apprentissage.
2. On prédit, pour tout i dans l'échantillon de validation et pour tout p, $\widehat{Y}_i^{(p)}$.
3. On sélectionne la valeur de p minimisant

$$\frac{1}{n_{valid}} \sum_{i \in I_{valid}} \left(Y_i - \widehat{Y}_i^{(p)} \right)^2 .$$

Échantillon d'apprentissage et échantillon de test (II)

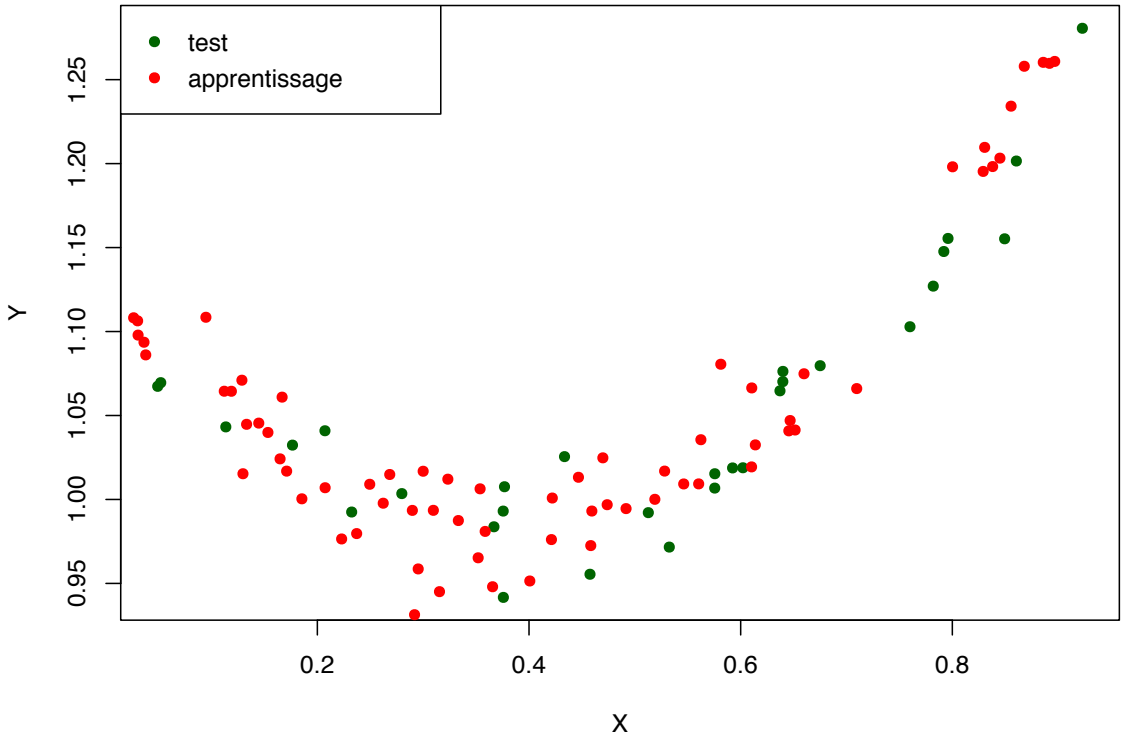
En pratique, pour calibrer les paramètres du modèle, on sépare l'échantillon en deux sous-parties disjointes : l'échantillon d'apprentissage I_{app} et l'échantillon de validation I_{valid}.

1. On apprend le modèle, c'est-à-dire que l'on estime les coefficients $\beta_0, ..., \beta_p$, pour plusieurs valeurs de p à l'aide de l'échantillon d'apprentissage.
2. On prédit, pour tout i dans l'échantillon de validation et pour tout p, $\widehat{Y}_i^{(p)}$.
3. On sélectionne la valeur de p minimisant

$$\frac{1}{n_{valid}} \sum_{i \in I_{valid}} \left(Y_i - \widehat{Y}_i^{(p)} \right)^2 .$$

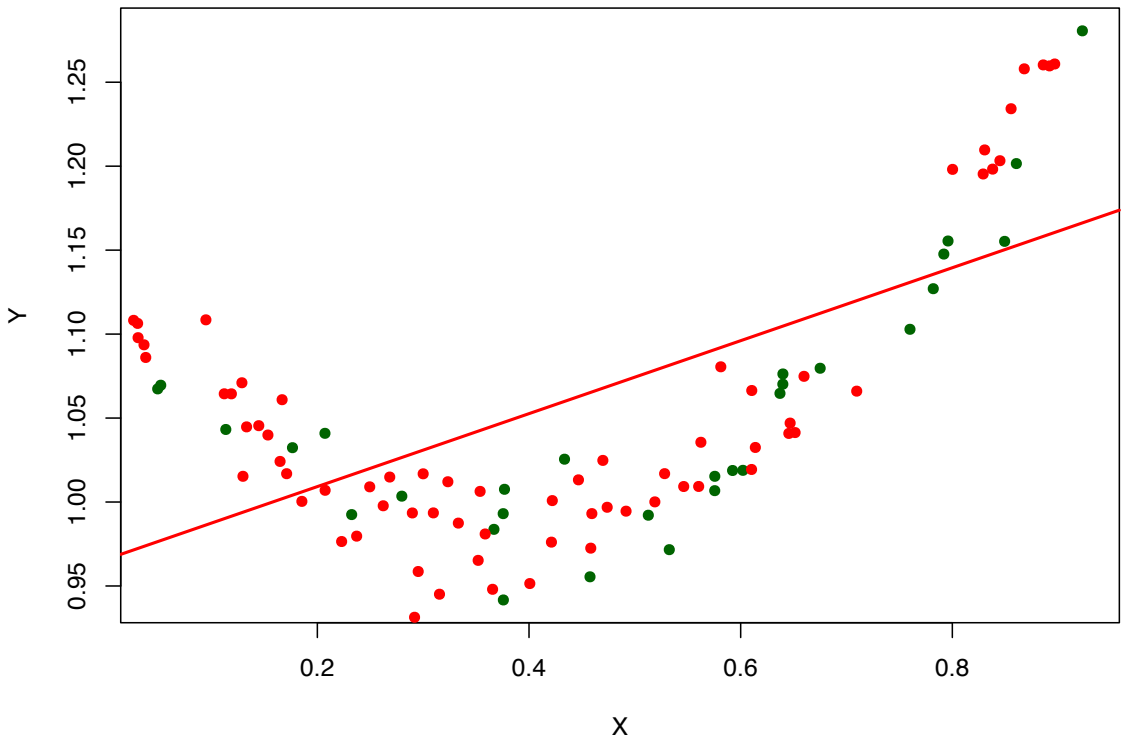
Ce principe est valable aussi dans d'autres contextes d'apprentissage supervisé, par exemple en classification.

Exemple : régression linéaire



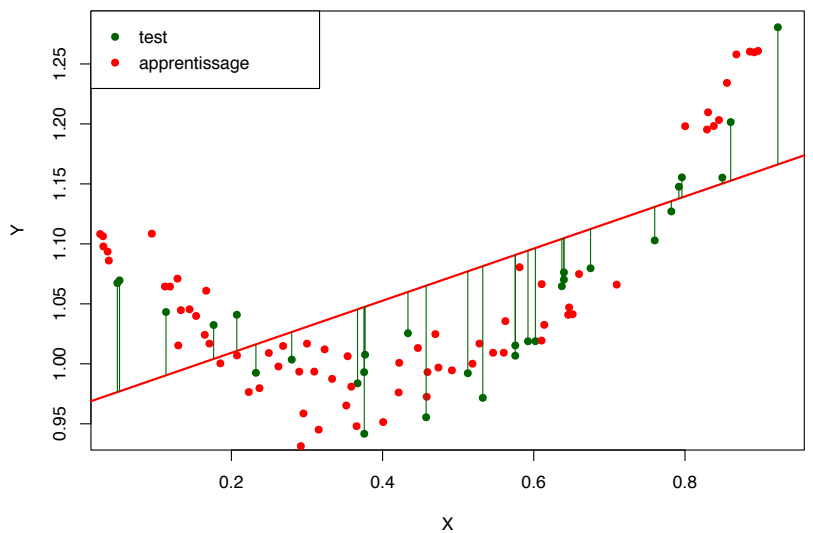
Exemple : régression linéaire (II)

Modèle linéaire simple ($p = 1$) : $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$.



Exemple : régression linéaire (III)

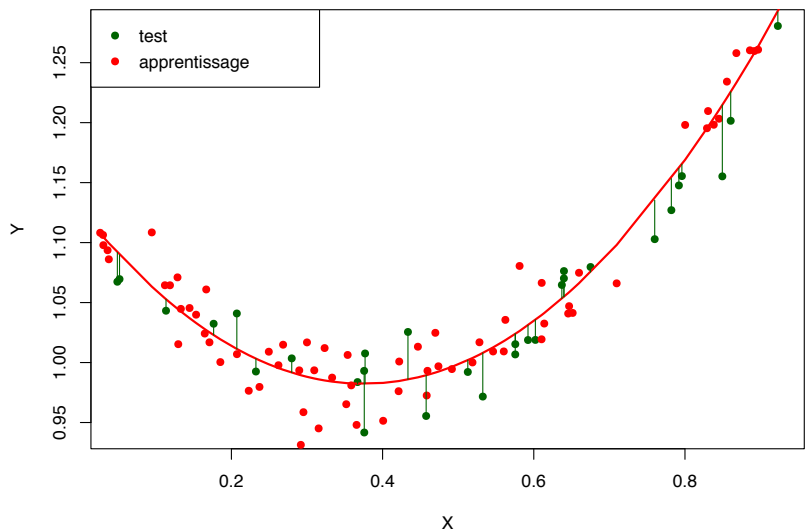
Modèle linéaire simple ($p = 1$) : $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$.



Erreur moyenne sur l'échantillon d'apprentissage :
$\frac{1}{n_{app}} \sum_{i \in I_{app}} \left(Y_i - \widehat{Y}_i \right)^2 = 5.10^{-3}$
Erreur moyenne sur l'échantillon de validation :
$\frac{1}{n_{valid}} \sum_{i \in I_{valid}} \left(Y_i - \widehat{Y}_i \right)^2 = 4.10^{-3}$

Exemple : régression linéaire (IV)

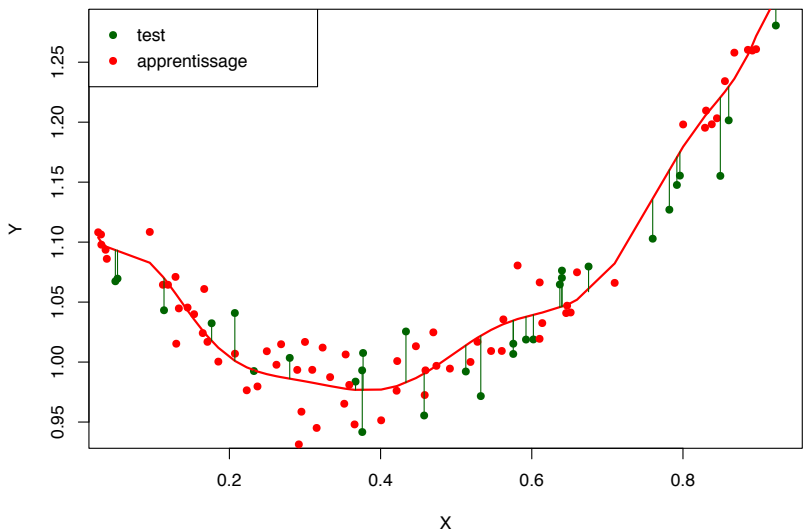
Modèle linéaire quadratique ($p = 2$) : $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$.



Erreur moyenne sur l'échantillon d'apprentissage :
$\frac{1}{n_{app}} \sum_{i \in I_{app}} \left(Y_i - \widehat{Y}_i \right)^2 = 5.10^{-4}$
Erreur moyenne sur l'échantillon de validation :
$\frac{1}{n_{valid}} \sum_{i \in I_{valid}} \left(Y_i - \widehat{Y}_i \right)^2 = 6.10^{-4}$

Exemple : régression linéaire (V)

Modèle linéaire quadratique ($p = 12$) : $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_i^j + \varepsilon_i$.



Erreur moyenne sur l'échantillon d'apprentissage :
$\frac{1}{n_{app}} \sum_{i \in I_{app}} \left(Y_i - \widehat{Y}_i \right)^2 = 4.10^{-4}$
Erreur moyenne sur l'échantillon de validation :
$\frac{1}{n_{valid}} \sum_{i \in I_{valid}} \left(Y_i - \widehat{Y}_i \right)^2 = 9.10^{-4}$

Plan

Régression linéaire en grande dimension

Rappels sur la régression linéaire

- Écriture matricielle

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

avec $\mathbf{Y} = (Y_1, ..., Y_n)^t$, $\boldsymbol{\beta}^* \in \mathbb{R}^{d+1}$, $\mathbf{X}$ matrice non aléatoire de taille $n \times (d+1)$ et $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.

- Estimateur des moindres carrés :

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

- Or, si $d \geq n$, X^tX non inversible : l'estimateur $\widehat{\boldsymbol{\beta}}$ n'est pas calculable !
- En pratique, même lorsque $d < n$, les performances de l'estimateur $\widehat{\boldsymbol{\beta}}$ se dégradent rapidement avec la dimension.

Régression linéaire en grande dimension

On a

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}).$$

Erreur quadratique moyenne de l'estimateur de $\boldsymbol{\beta}$

$$\mathbb{E}\left[\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2\right] = \sigma^2 \mathrm{tr}((\mathbf{X}^t\mathbf{X})^{-1}).$$

Problème

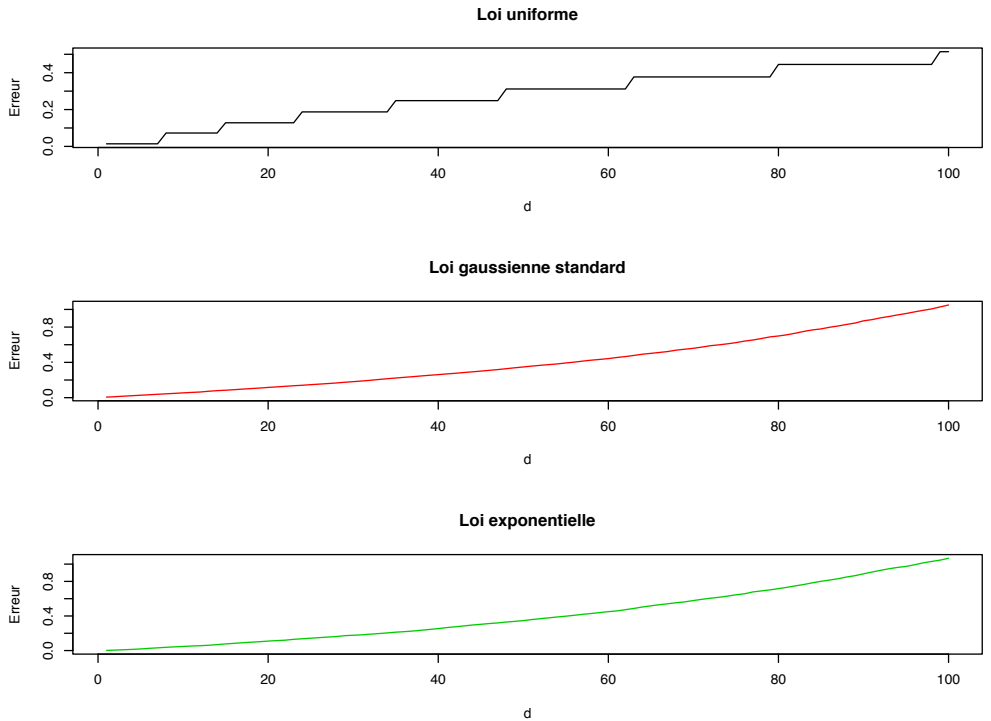
Par exemple, si les colonnes de X sont orthonormales (i.e. $X^tX = I$) alors

$$\mathbb{E}\left[\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2\right] = \sigma^2(d+1).$$

Régression linéaire en grande dimension

Erreur quadratique moyenne de l'estimateur de β

Tracé de $\mathrm{tr}((\mathbf{X}^t\mathbf{X})^{-1})$ en fonction de d lorsque $\{X_{i,j}\}_{i=1,\dots,n;j=2,\dots,d+1}$ i.i.d. ($n = 200$)



Plan

Plan du cours

1. Introduction
2. Pénalisations Ridge et Lasso
3. Quelques approches non-linéaires en apprentissage supervisé
4. Apprentissage non supervisé : introduction aux k-moyennes et à la CAH, classification mixte.