

# Apprentissage Statistique

**Nicolas Verzelen, Joseph Salmon (Pierre Pudlo)**

INRA / Université de Montpellier



# Plan

Introduction

Apprentissage statistique

Régression linéaire

# Problèmes d'apprentissage statistique

1. Identifier les facteurs de risque du cancer de la prostate
2. Classifier des phonèmes à partir de périodogrammes
3. Prédire si une personne est sujette aux crises cardiaques, à partir de mesures cliniques, son régime et des données démographiques
4. Personnaliser un système de détection de spam email
5. Lecture de codes postaux écrits à la main
6. Classification d'échantillons de tissus dans différents types de cancer, en fonction de données d'expression de gènes
7. Établir une relation entre salaires et variables démographiques
8. Classifier les pixels d'une image satellite

# Problème d'apprentissage supervisé

## Point de départ :

$y$  : réponse, variable dépendante, cible

$x$  : variables explicatives, co-variables

**Régression** :  $y$  quantitatif, continue

**Classification** (ou discrimination) :  $y$  qualitatif, discrète

## Données d'apprentissage :

$$D_1^n := \{(x_1, y_1), \dots, (x_n, y_n)\}$$

avec  $x_i \in \mathcal{X}$  quelconque (souvent  $\mathbb{R}^p$ ),

$y_i \in \mathcal{Y}$  pour  $i = 1, \dots, n$ .

**Notations** :  $\mathbf{x} \in \mathcal{X}^n$ ,  $\mathbf{y} \in \mathcal{Y}^n$  vecteurs des données d'apprentissage

## Objectifs :

À partir de la base de données, on voudrait

- ▶ prédire le plus précisément possible la sortie  $y$  pour une nouvelle entrée  $x$ .
- ▶ comprendre quelle(s) co-variables influe sur la réponse, et comment
- ▶ (évaluer la qualité des inférences et prédictions)

# Motivation

- ▶ Un point fondamental dans la formation moderne d'un statisticien
- ▶ Comprendre d'abord les méthodes les plus simples, avant de passer aux méthodes les plus sophistiquées
- ▶ Applications en sciences, industrie, finance, . . .
- ▶ “*Machine Learning*” : issu de la communauté d'intelligence artificielle

# Apprentissage non supervisé

- ▶ Données d'apprentissage  $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$  avec  $x_i \in \mathcal{X}$  ;  
Ici pas de réponse  $y$  !
- ▶ Buts plus flous : description de données, analyse de données
- ▶ Difficulté de l'évaluation du résultat
- ▶ Différent de l'apprentissage supervisé, mais peut être une étape préliminaire à un apprentissage supervisé

Exemple : Regrouper des observations / clients / patients similaires (marketing,...)

# Plan

Introduction

Apprentissage statistique

Régression linéaire

# Modèle statistique non paramétrique

On suppose que  $D_1^n$  est l'observation d'un  $n$ -échantillon  $D_1^n = \{(x_1, y_1), \dots, (x_n, y_n)\}$  d'une loi conjointe  $P$  sur  $\mathcal{X} \times \mathcal{Y}$  inconnue

On suppose que  $x$  est une nouvelle observation,  $(x, y)$  étant un couple aléatoire de loi conjointe  $P$  indépendante de  $D_1^n$ .

---

---

## Définition

---

---

Une **règle de prédiction/ régression ou discrimination** est une fonction (mesurable)  $f : \mathcal{X} \mapsto \mathcal{Y}$  qui associe la sortie  $f(x)$  à l'entrée  $x \in \mathcal{X}$ .

---

---

**Attention à l'abus de notation** :  $x$  et  $y$  désigne parfois des éléments déterministe de  $\mathcal{X}$  et  $\mathcal{Y}$ , parfois des variables aléatoires. *Qu'en est-il dans la définition ci-dessus ?*



# Qualité de prédiction

---

---

## Définition

---

---

Une fonction  $l : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$  est une fonction de perte si  $l(y, y) = 0$  et  $l(y, y') > 0$  pour  $y \neq y'$

---

---

Exemples :

- ▶  $l(y, y') = |y - y'|^q$  en régression réelle  
(perte absolue si  $q = 1$ , perte quadratique si  $q = 2$ )
- ▶  $l(y, y') = 1_{y \neq y'}$  en discrimination binaire.

---

---

## Définition

---

---

Le **risque** ou l'**erreur de généralisation** - d'une règle de prédiction  $f$  est défini par

$$R_P(f) = \mathbb{E}_{(x,y) \sim P}[l(y, f(x))]$$

---

---

Si  $\mathcal{F}$  désigne l'ensemble des règles de prédiction possibles, quelles sont les règles de prédiction optimales au sens la minimisation du risque sur  $\mathcal{F}$ , c'est-à-dire les règles  $f^*$  t.q.  $R_P(f^*) = \inf_{f \in \mathcal{F}} R_P(f)$  ?

# Régression réelle et discrimination

---

---

## Définition

---

---

On appelle **fonction de régression** la fonction  $\eta^* : \mathcal{X} \rightarrow \mathcal{Y}$  définie par  $\eta^*(x_0) = \mathbb{E}[y|x = x_0]$

---

---

(quelle hypothèse sur  $\mathcal{Y}$  ?)

### Cas de la régression réelle

$$\mathcal{Y} = \mathbb{R}, \quad l(y, y') = (y - y')^2$$

---

---

## Théorème

---

---

La fonction de régression  $\eta^*$  vérifie  $R_P(\eta^*) = \inf_{f \in \mathcal{F}} R_P(f)$

---

---

$$\mathcal{Y} = \mathbb{R}, \quad l(y, y') = |y - y'|$$

---

---

## Théorème

---

---

La règle de régression définie par  $\mu^*(x_0) = \text{mediane}[y|x = x_0]$  vérifie  $R_P(\mu^*) = \inf_{f \in \mathcal{F}} R_P(f)$

---

---

## Cas de la discrimination binaire

$$\mathcal{Y} = \{-1, 1\}, \quad l(y, y') = 1_{y \neq y'} = |y - y'|/2 = (y - y')^2/4$$

---

---

### Définition

---

---

On appelle **règle de Bayes** toute fonction  $\phi^*$  de  $\mathcal{F}$  telle que pour tout  $x_0 \in \mathcal{X}$ ,

$$\mathbb{P}(y = \phi^*(x_0) | x = x_0) = \max_{y' \in \mathcal{Y}} \mathbb{P}(y = y' | x = x_0).$$

---

---

### Théorème

---

---

Si  $\phi^*$  est une règle de Bayes, alors  $R_P(\phi^*) = \inf_{f \in \mathcal{F}} R_P(f)$ .

---

---

## De la régression réelle à la discrimination binaire

---

---

### Théorème

---

---

Pour tout règle de régression  $\eta$ , si  $\phi_\eta(x) = \text{sign}(\eta(x))$  alors

$$\begin{aligned} & \mathbb{E}_{(x,y) \sim P} [1_{y \neq \phi_\eta(x)} - 1_{y \neq \phi_{\eta^*}(x)}] \\ & \leq \mathbb{E}_{x \sim P_x} |\eta(x) - \eta^*(x)| \leq \sqrt{\mathbb{E}_{(x,y) \sim P} [(y - \eta(x))^2 - (y - \eta^*(x))^2]} \end{aligned}$$

# Plug-in et risque moyen

---

---

## Théorème

---

---

La règle de discrimination *plug-in* définie par

$$\phi_{\eta^*}(x) = \text{signe}(\eta^*(x)) = 1_{\eta^*(x) \geq 0} - 1_{\eta^*(x) < 0}$$

est une règle de Bayes

---

---

Ces règles de prédiction optimales dépendent de  $P$  !

---

---

## Définition

---

---

**Risque moyen** de  $\hat{f}$ , construit avec  $D_1^n$ , est

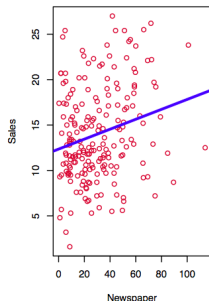
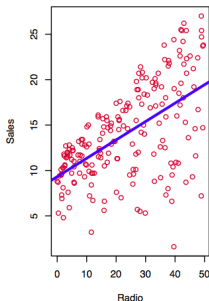
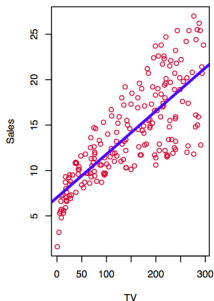
$$\mathcal{R}_P(\hat{f}) = \mathbb{E}_{D_1^n \sim P^{\otimes n}} [\mathbb{E}_{(x,y) \sim P} [l(y, \hat{f}(x))]]$$

---

---

Objectif : proposer  $\hat{f}$  t.q.  $\mathcal{R}_P(\hat{f})$  le plus petit possible  
Le risque moyen d'un algorithme dépend de  $P$  inconnu !  
Comment faire en pratique ?

## En plus concret



Sur le graphique :  
ventes (Sales) en fonction de

- ▶ pub TV (TV)
- ▶ pub radio (Radio)
- ▶ pub presse (Newspaper)

Objectif : prédire Sales en  
utilisant les trois variable

*i.e.*, construire une fonction  $f$   
tel que  
$$\text{Sales} \approx f(\text{TV}, \text{Radio}, \text{Newspaper})$$

Les trois lignes **bleues** sont trois  
régression linéaire simple

# Cadre

## Notations

- ▶ Sales est la réponse. On la note  $y$ .  $\mathcal{Y} = \mathbb{R}$ .
- ▶ TV, Radio, Newspaper sont les co-variables. On les notes  $x^{(1)}, x^{(2)}, x^{(3)}$ .
- ▶ Le vecteur ligne des co-variables  
$$x = (x^{(1)}, x^{(2)}, x^{(3)})$$

On a  $\mathcal{X} = \mathbb{R}^3$ .

La fonction de régression

$\eta^*(x_0) = \mathbb{E}(y|x = x_0)$  minimise le risque quadratique  $l(y, y') = (y - y')^2$ .

Si on définit  $\varepsilon = y - \eta^*(x)$ , alors on a  
$$y = \eta^*(x) + \varepsilon,$$

avec  $\mathbb{E}(\varepsilon|x) = 0$ .

$\varepsilon$  capture l'information de  $y$  non prédictible par  $x$ .

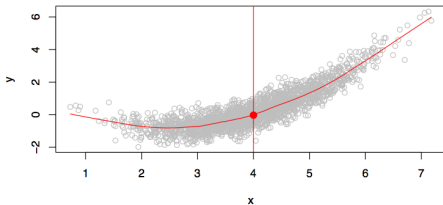
Cette décomposition ne **repose sur aucune hypothèse** sur le lien entre  $y$  et  $x$  (hormis l'intégrabilité de  $y$ ).

*Question* : Si on utilise la perte  $l_1$ , quelle décomposition obtient-on ?

## Pourquoi estimer une règle $\hat{f}$ proche de $\eta^*$ ?

- ▶ (*par définition*) faire des **prédictions** précises de Sales ( $y$ ) pour une nouvelle valeurs de (TV, Radio, Newspaper) à de nouveaux points  $x$
- ▶ On peut comprendre quelles composantes de  $x = (x^{(1)}, \dots, x^{(p)})$  sont importantes pour expliquer  $y$  (exemple : Experience et Diplome ont une grande influence sur le Salaire, mais le Statut marital en a peu, ou pas)
- ▶ Suivant la complexité de  $\eta^*$ , on peut comprendre comment chaque composante  $x^{(j)}$  de  $x$  influe  $y$

## Comment estimer $\eta^*$ à partir de $D_1^n$ ?



Un premier estimateur

$$\hat{f}(x) = \frac{\sum_{i=1}^n y_i \mathbb{1}_{\{x_i=x\}}}{\sum_{i=1}^n \mathbb{1}_{\{x_i=x\}}},$$

avec la convention  $0/0 = 0$ .

Exemple : Si  $\mathcal{X} = \{0\}$ .

$$R_P(\hat{f}) = \mathbb{E}_{(x,y) \sim P} (y - \hat{f}(x))^2 = R_P(\eta^*) + \left( \frac{1}{n} \sum_{i=1}^n \varepsilon_i \right)^2$$

$$\mathcal{R}_P(\hat{f}) = R_P(\eta^*) + \frac{\text{var}(\varepsilon)}{n} = \text{var}(\varepsilon) \left( 1 + \frac{1}{n} \right)$$

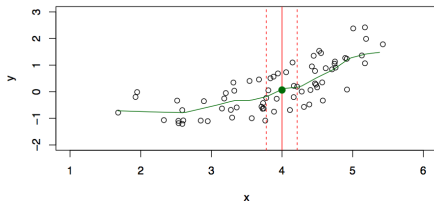


# Estimer $\eta^*$ par moyennes mobiles

- Pour  $\mathcal{X} = \mathbb{R}$ ,  $\hat{f}$  est nulle sauf en un nombre fini de valeurs.
- On relâche la définition et on prend (convention  $0/0 = 0$ )

$$\hat{f}_h(x) = \frac{\sum_{i=1}^n y_i 1_{x_i \in \mathcal{N}_h(x)}}{\sum_{i=1}^n 1_{x_i \in \mathcal{N}_h(x)}},$$

où  $\mathcal{N}_h(x)$  est un **voisinage** de  $x$  de rayon  $h$ .



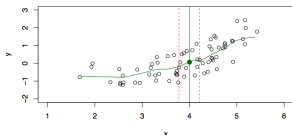
$$\mathbb{E}_{(y,x) \sim P} \left( (y - \hat{f}_h(x))^2 \middle| x \right) = \underbrace{\left( \eta^*(x) - \hat{f}_h(x) \right)^2}_{\text{réductible}} + \underbrace{\text{Var}(\varepsilon | x)}_{\text{irréductible}}$$

$$R_P(\hat{f}_h) = \mathbb{E}_{x \sim P_x} \left( (y - \hat{f}_h(x))^2 \right) = \mathbb{E}_{x \sim P_x} \left( \eta^*(x) - \hat{f}_h(x) \right)^2 + R_P(\eta^*)$$

# Décomposition Biais-variance

Considérons le risque moyen

$$\begin{aligned} \mathcal{R}_P(\hat{f}_h) = & \underbrace{\mathbb{E}_{x \sim P_x} \left( \eta^*(x) - \mathbb{E}_{D_1^n \sim P^{\otimes n}}(\hat{f}_h(x)) \right)^2}_{\text{Biais}} \\ & + \underbrace{\mathbb{E}_{x \sim P_x} [\text{Var}_{D_1^n \sim P^{\otimes n}}(\hat{f}_h(x))]}_{\text{Variance}} + R_P(\eta^*) \end{aligned}$$



- ↪ Pour  $h$  petit, la variance est grande (peu de points moyennés).
- ↪ Pour  $h$  grand, le biais est trop important.

**Problème :** Un bon  $h$  dépend de la distribution  $P$  des données.

# La moyenne mobile est-elle un bon choix ?

- ▶ Méthodes, basées sur des moyennes autour des voisins plutôt bonnes si :
  - petite dimension  $p \leq 4$
  - grand échantillon  $n \gg p$
- ▶ Plus tard dans le cours : version lissées, obtenues par
  - méthodes à noyaux
  - lissage par splines,
  - ...

Ces méthodes peuvent être **mauvaises** quand  $p$  est grand.

## Raison : fléau de la dimension

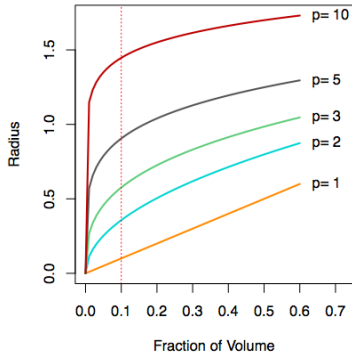
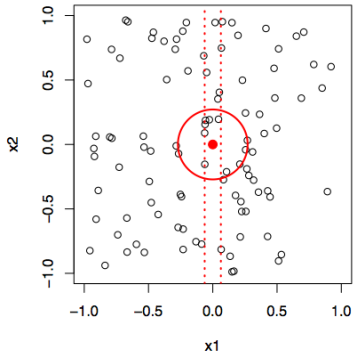
Les voisins proches peuvent être éloignés en grande dimension

- ▶ Il faut suffisamment de  $y_i$  à moyenner pour que  $\hat{f}_h(x)$  ait une faible variance
- ▶ En grande dimension, besoin de s'éloigner de  $x$  pour cela.

On perd l'idée de moyenne **locale** autour de  $x$ .

# Le fléau de la dimension

10% Neighborhood



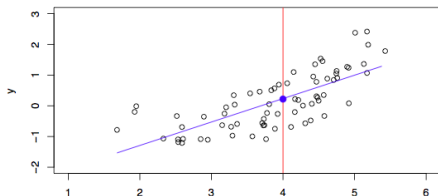
# Modèles linéaires

Construire une règle de régression dans la famille  $\mathcal{F}_L \subset \mathcal{F}$ .

$$\mathcal{F}_L := \{f(x) = \beta_0 + \beta_1 x^{(1)} + \dots + \beta_p x^{(p)}; \beta \in \mathbb{R}^{p+1}\}$$

- ▶ La dimension de  $\mathcal{F}_L$  vaut  $p + 1$ .
- ▶ Lorsque la perte est quadratique, la règle  $\hat{f}_L$  est généralement ajustée par critères des moindres carrés sur  $D_1^n$ .

Exemple :  $\hat{f}_L(x) = \hat{\beta}_0 + \hat{\beta}_1 x$

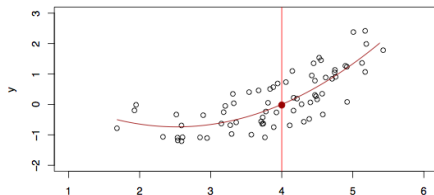


Les modèles linéaires sont **rarement** corrects, *i.e.*,  $\eta^* \notin \mathcal{F}_L$  mais peuvent fournir de bonnes approximations interprétables  $\eta^*(x)$

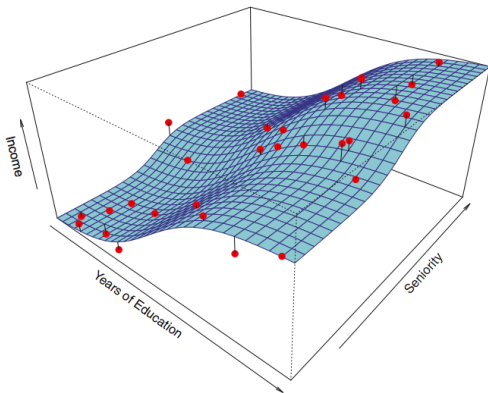
# Modèle paramétrique (suite)

Modèle quadratique :

$$\mathcal{F}_Q := \{f(x) = \beta_0 + \beta_1 x + \beta_2 x^2\}$$



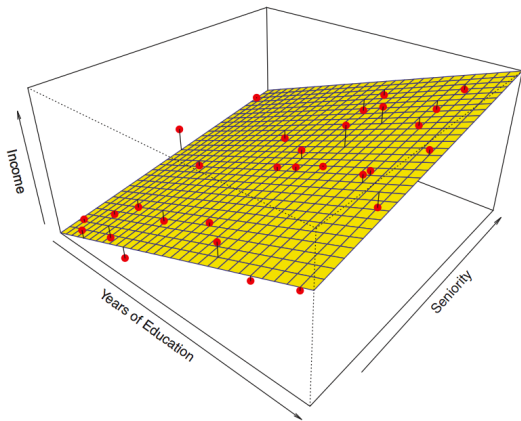
## Un exemple simulé



$$\text{income} = \eta^*(\text{education}, \text{seniority}) + \varepsilon$$

La véritable fonction  $\eta^*$  est la surface bleue

## Exemple simulé : modèle linéaire

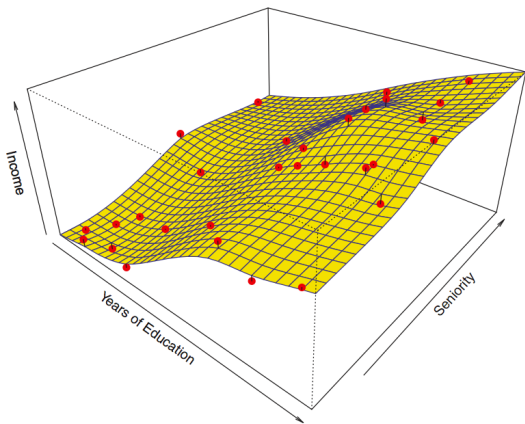


$$\widehat{f}_L(\text{education}, \text{seniority}) = \widehat{\beta}_0 + \widehat{\beta}_1 \times \text{education} + \widehat{\beta}_2 \times \text{seniority}$$

ajustée aux données

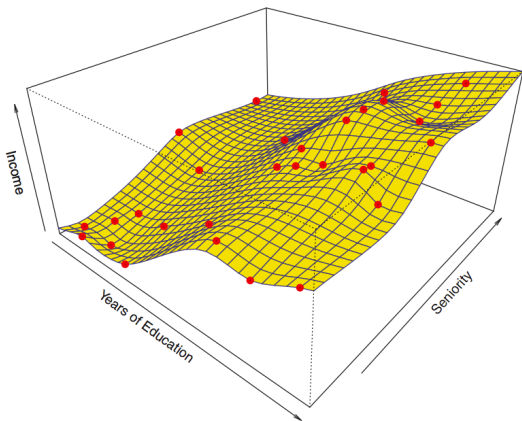


## Exemple simulé : modèle de splines



Voir plus tard dans le cours. . .

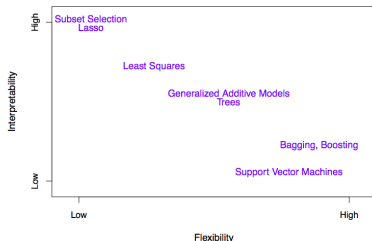
## Exemple simulé : second modèle de splines



**Sur-apprentissage** : la surface passe par tous les points observés, mais généralise très mal.

# Des méthodes, des compromis

- ▶ précision vs. interprétation
  - modèle linéaire sont faciles (?) à interpréter ; les splines en plaques minces non.
- ▶ ajustement vs. sur-ajustement vs. sous-ajustement
  - Comment savoir ?
- ▶ Parcimonie vs. boîte noire
  - modèles simples impliquent peu de paramètres → modèles non-paramétriques



Toutes une gamme de méthodes

## Conclusion temporaire

- ▶ Il n'existe pas de méthode universellement meilleure que les autres.
- ▶ Pour chaque approche : de nouveaux paramètres à ajuster.
- ▶ Sélectionner une approche nécessite de savoir les comparer, *i.e.*, estimer le risque de plusieurs règles.
- ▶ Le choix de la méthode dépend aussi des objectifs du statisticiens (interprétation).

**Une implication nécessaire de l'utilisateur dans l'analyse**

# Un exemple en discrimination (classification)

Ici, la réponse  $Y$  est

**qualitative.**

Exemple :  $\mathcal{Y} = \{\text{spam}, \text{ham}\}$   
(ham=e-mail correct) ou bien  
 $\mathcal{Y} = \{0, 1, \dots, 9\}$ . Les objectifs  
peuvent être

- ▶ construire un classifieur  $\phi(x)$  qui assigne une nouvelle observation  $x$  à l'une des classes dans  $\mathcal{Y}$
- ▶ évaluer l'incertitude sur cette classification
- ▶ comprendre les rôles de  $x^{(1)}, \dots, x^{(p)}$

Le **classifieur de Bayes**  
(optimal)

$$\phi^*(x_0) = \arg \max_{y \in \mathcal{Y}} P(y|x = x_0)$$

pour la perte  $l(y, y') = 1_{y \neq y'}$ .

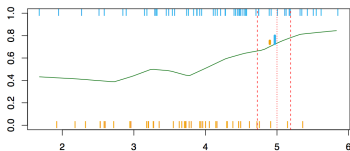
On peut toujours utiliser une méthode des plus proches voisins pour la construction d'une règle de classification  $\hat{\phi}(x)$  à partir de  $D_1^n$ .

# Problèmes de classification

**classifieur de Bayes** optimal

$$\phi^*(x = x_0) = \arg \max_{y \in \mathcal{Y}} P(y|x = x_0)$$

pour la perte  $l(y, y') = 1_{y \neq y'}$ .

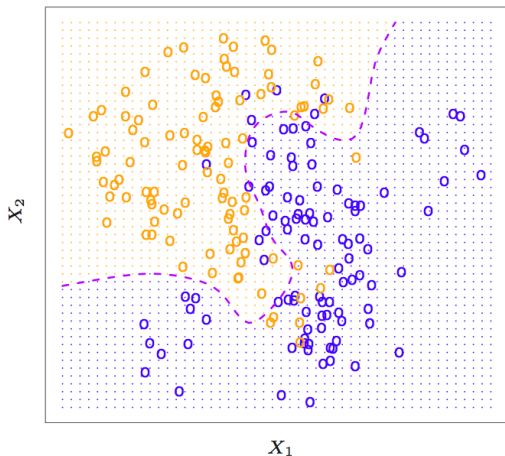


On peut toujours utiliser une méthode des plus proches voisins pour la construction d'une règle de classification  $\hat{\phi}(x)$  à partir de  $D_1^n$ .

Autres méthodes :

- ▶ régression logistique / analyse linéaire discriminante
- ▶ modèles additifs généralisés
- ▶ *Support Vector Machine* (SVM)

## Exemple : $K$ -plus proches voisins en 2D

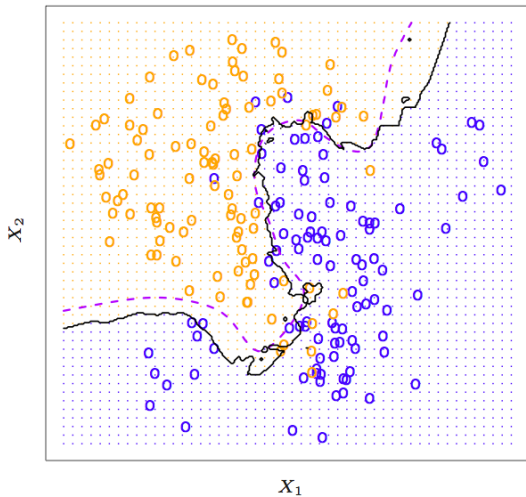


Données simulées.

La courbe en pointillée correspond à

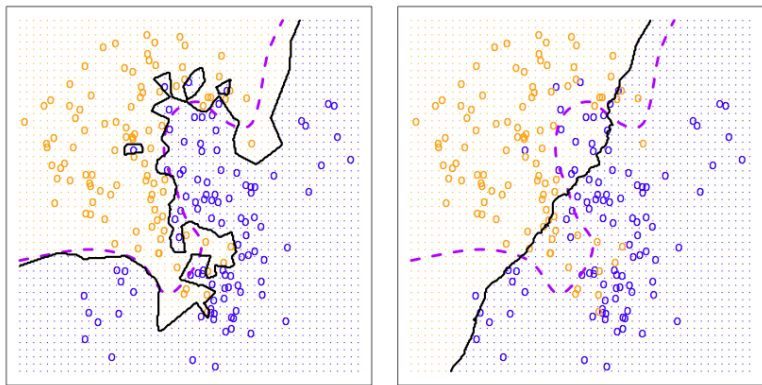
$$\mathbb{P}(y = -1|x) = \mathbb{P}(y = 1|x) = 1/2$$

(-1 : bleu ; 1 : orange)



Prédictions avec  $K = 10$





Prédictions avec  $K = 1$  et  $K = 100$

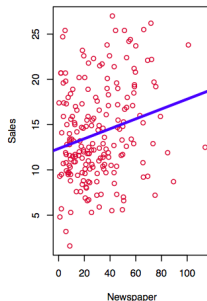
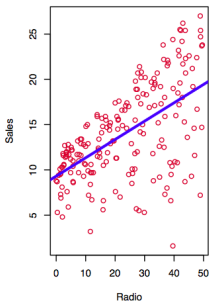
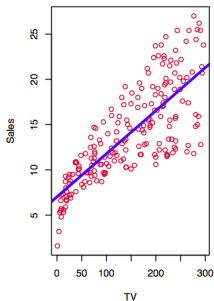
# Plan

Introduction

Apprentissage statistique

Régression linéaire

# Exemple de données de ventes : des questions

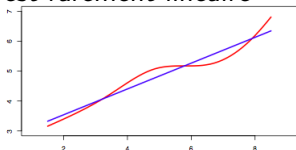


- Existence d'un lien entre le budget publicitaire et les ventes ?
- Quelle est la force de ce lien entre budget publicitaire et ventes ?
- Quel média contribue le plus aux ventes ?
- Précision des prédictions, linéarité du lien ?
- Existe-t-il des synergies entre médias ?

# Régression linéaire

- Approche simple pour l'apprentissage supervisé, supposant que  $y$  dépend linéairement de  $x^{(1)}, \dots, x^{(p)}$

- La fonction de régression est rarement linéaire



- Bien que simple, la régression linéaire est utile à la fois conceptuellement et en pratique.

- **Apprentissage** : On cherche à créer une règle de régression dans la classe

$$\mathcal{F}_L := \{f(x) = \beta_0 + \sum_{j=1}^p \beta_j x^{(j)}, \beta \in \mathbb{R}^{p+1}\}$$

Ici,  $\eta^*$  n'appartient pas forcément à  $\mathcal{F}_L$ .

- **Paramétrique** : on suppose qu'il existe  $\beta$  (inconnu) t.q.

$$y = \beta_0 + \sum_{j=1}^p \beta_j x^{(j)} + \varepsilon,$$

avec  $\mathbb{E}[\varepsilon|x] = 0$ .

...mais des méthodes **communes** aux deux perspectives.

# Régression linéaire (II)

## Notations pour l'échantillon d'apprentissage

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \dots \\ \beta_p \end{pmatrix} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \dots \\ \epsilon_n \end{pmatrix}$$

Matrice de design :



$$X = \begin{pmatrix} x_{1,1} & \dots & x_{1,p} \\ \dots & & \\ x_{n,1} & \dots & x_{n,p} \end{pmatrix}$$

# Régression linéaire (III)

## Régression linéaire

- ▶ On pose un modèle de la forme

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

où  $\beta_0, \beta_1$  inconnus sont ordonnée à l'origine ( : *intercept*) et pente ( : *slope*).

Ce sont les **coefficients** du modèle

- ▶ Étant estimés ces coefficients par  $\hat{\beta}_0$  et  $\hat{\beta}_1$ , on prédit  $y$  sachant  $x$  avec

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

## Moindres carrés

- ▶ Soit  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$  la prédiction de  $y$  sur la  $i$ -ème valeur de  $x$ , et  $e_i = y_i - \hat{y}_i$  est le  $i$ -ème **résidu**

- ▶ Somme des carrés résiduels :
$$SSE = e_1^2 + e_2^2 + \dots + e_n^2$$

à minimiser pour avoir  $\hat{\beta}_0, \hat{\beta}_1$

- ▶ Ce qui donne
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- ▶ Hypothèse : loi de  $\varepsilon$  sachant  $x$  est gaussienne

# Régression linéaire simple

## Moindres carrés

- ▶ Soit  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$  la prédiction de  $y_i$  sur la  $i$ -ème valeur de  $x$ , et  $e_i = y_i - \hat{y}_i$  est le  $i$ -ème **résidu**

- ▶ Somme des carrés résiduels :  
$$SSE = e_1^2 + e_2^2 + \dots + e_n^2$$

à minimiser pour avoir  $\hat{\beta}_0, \hat{\beta}_1$

- ▶ Ce qui donne  
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
  
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- ▶ On suppose maintenant que la loi de  $\varepsilon$  sachant  $x$  est gaussienne centrée

- ▶ Erreur d'échantillonnage en terme d'écart-type

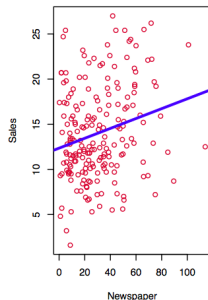
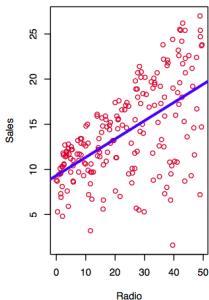
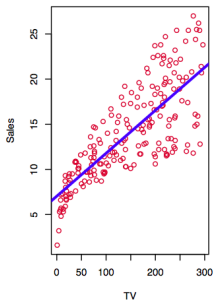
$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left( 1 + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$$

$\sigma^2 = \text{Var}(\varepsilon|x)$  supposée constante  
(homoscédasticité)

- ▶ Permet de construire des **intervalles de confiances** pour l'inférence  $\beta_j$
- ▶ Des **tests d'hypothèses** ( $t$ -test par exemple sur la nullité d'un coefficient)

# Exemple de données de ventes (suite)



	Coefficient	Std.Err	<i>t</i> -statistic	<i>p</i> -value
Intercept	7.0325	0.4578	15.36	<0.0001
TV	0.0475	0.0027	17.67	<0.0001



# Évaluer la précision complète du modèle linéaire simple

- ▶ Erreur standard résiduelle

$$\hat{\sigma} = \sqrt{\frac{1}{n-2}SSE}$$

où  $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$

- ▶ Fraction de variance expliquée

$$R^2 = \frac{SST - SSE}{SST}$$

où  $SST = \sum_{i=1}^n (y_i - \bar{y})^2$  est la somme des carrés totale

- ▶  $R^2$  est le carré du coefficient de corrélation entre  $X$  et  $Y$

## Exemple des données de ventes

Quantité	Valeur
$SSE$	3.26
$R^2$	0.612
$F$ -stat	312.1

# Régression linéaire multiple

$$\text{Modèle : } y = \beta_0 + \beta_1 x^{(1)} + \dots + \beta_p x^{(p)} + \varepsilon$$

- ▶ On interprète  $\beta_j$  comme l'effet moyen sur  $y$  d'un accroissement de  $x^{(j)}$  d'une unité **lorsque tous les autres prédicteurs sont fixés.**
- ▶ On ne peut faire aucune affirmation en terme de **causalité.**

Exemple :  $y$  = nombre de tacles,  $w$  = poids et  $h$  = taille  
 $\hat{y} = b_0 + .50w - .10h$

Comment s'interprète  $-.10$  ?

## Interprétation des coefficients

- ▶ Scenario idéal : prédicteurs indépendants et design équilibré
  - ▶ chaque  $\beta_j$  peut être estimé / testé séparément
  - ▶ interprétation de gauche OK
- ▶ Les corrélations entre  $x^{(j)}$  posent problème
  - ▶ la variance des estimateurs s'accroît
  - ▶ interprétation hasardeuse (lorsque  $x^{(j)}$  change, tout change !)

## Citations

*“Essentially, all models are wrong, but some are useful”*

George Box

*“All models are wrong but some come with good open source implementation and good documentation so use those.”*

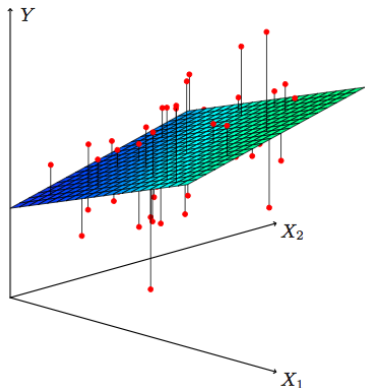
Alexandre Gramfort

*“The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively”*

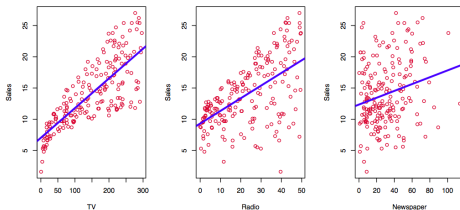
Fred Mosteller and John Tukey

# Estimation et prédiction en régression multiple

- ▶ À partir d'estimation des coef  $\widehat{\beta}_0, \widehat{\beta}_1, \dots, \widehat{\beta}_p$ , on peut prédire avec
$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x^{(1)} + \dots + \widehat{\beta}_p x^{(p)}$$
- ▶ Comme en dimension 1, on estime les  $\beta$  en minimisant la somme des carrés résiduelle.  
Formule théorique qui dépend d'une inversion de matrice produit. → utiliser un logiciel de statistique
- ▶ De même,  $SE(\beta_j)$  pour chaque coefficient,  $t$ -test de nullité, test de Fisher,...



# Résultats pour les données de ventes



	Coeff	Std.Err	t-stat	p-value
Intercept	2.939	0.3119	9.42	<0.0001
TV	0.046	0.0014	32.81	<0.0001
radio	0.189	0.0086	21.89	<0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Corrélations				
	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

# Des questions importantes

1. Y a t-il au moins un des  $x^{(j)}$  utile pour prédire  $y$  ?
2. Sont-ils vraiment tous utiles ?
3. Comment le modèle s'ajuste aux données ?
4. Avec une nouvelle valeur de  $x$ , quelle réponse doit-on prédire ?  
Précision de la prédiction ?

# Des questions importantes

1. Y a t-il au moins un des  $x^{(j)}$  utile pour prédire  $y$  ?
2. Sont-ils vraiment tous utiles ?
3. Comment le modèle s'ajuste aux données ?
4. Avec une nouvelle valeur de  $x$ , quelle réponse doit-on prédire ?  
Précision de la prédiction ?

Pour la première question, on utilise la  $F$ -statistique

$$F = \frac{(SST - SSE)/p}{SSE/(n - p - 1)} \sim F_{p, n-p-1}$$

Quantité	Valeur
Residual Std.Err	1.69
$R^2$	0.897
$F$ -stat	570

# Des questions importantes

1. Y a t-il au moins un des  $x^{(j)}$  utile pour prédire  $y$  ?
  2. Sont-ils vraiment tous utiles ?
  3. Comment le modèle s'ajuste aux données ?
  4. Avec une nouvelle valeur de  $x$ , quelle réponse doit-on prédire ?  
Précision de la prédiction ?
- ▶ Choix de co-variables :  
approche complète  
Comparer les modèles linéaires avec tous les sous-ensembles possibles de co-variables
  - ▶ Souvent  $2^p$  trop grand  
( $\log_{10}(2^{40}) \approx 12.0$ )  
On utilise une méthode que ne parcourt que certains sous-ensembles. Deux approches standard  
Sélections progressive, ou rétrograde
  - ▶ Nécessite de répondre à la question suivante pour effectuer la comparaison.



# Choix de co-variables

## Méthode progressive

(forward)

1. Commencer par le modèle nul (à zéro co-variables)
2. Ajuster les  $p$  régressions linéaires simples et ajouter au modèle nul la co-variable qui à le plus petit  $SSE$
3. Ajouter à ce modèle à une co-variable la co-variable qui fait baisser le plus le  $SSE$
4. Continuer jusqu'à un critère d'arrêt (par exemple sur la  $p$ -value du  $t$ -test)

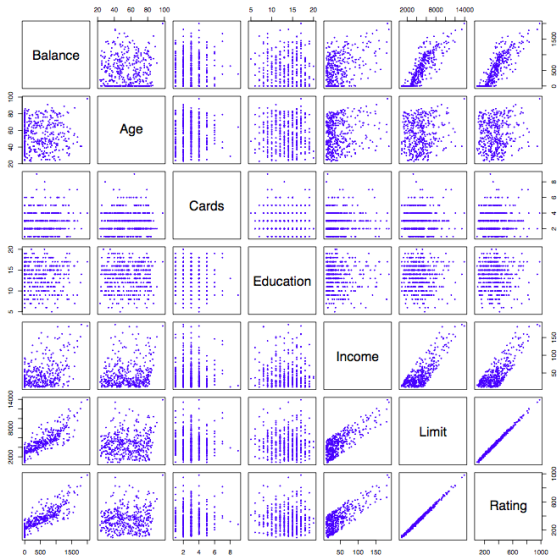
## Méthode rétrograde (backward)

1. Commencer par le modèle avec tous les co-variables
2. Supprimer la variable avec la plus grande  $p$ -value —*i.e.*, la co-variable la moins significative pour le modèle
3. Ré-ajuster le modèle, et enlever de nouveau la co-variable de plus grande  $p$ -value
4. Continuer jusqu'à un critère d'arrêt (par exemple portant sur la valeur de la  $p$ -value de la co-variable que l'on enlèverait)

## Plus tard

- ▶ critère plus systématique pour choisir le modèle « optimal » dans ceux que l'on parcourt
- ▶ Avec  $C_p$  de Mallows, Akaike information criterion (AIC), Bayesian information criterion (BIC),  $R^2$  ajusté et validation croisée (CV)

# Autre problème



- co-variables qualitatives (ou discrètes)

## Prédicteurs qualitatifs

Exemple : Solde de la carte de crédit en fonction du genre (Homme/Femme), en ignorant les autres co-variables. On crée une nouvelle variable

$$x_i = \begin{cases} 1 & \text{si la } i\text{-ème pers. est une femme} \\ 0 & \text{si la } i\text{-ème pers. est un homme} \end{cases}$$

Le modèle résultant

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{si la } i\text{-ème pers. est une femme} \\ \beta_0 + \varepsilon_i & \text{si la } i\text{-ème pers. est un homme} \end{cases}$$

Résultats

	coeff	Std.Err	t-stat	p-value
Intercept	509.80	33.13	15.389	<0.0001
gender [Female]	19.73	46.05	0.429	0.6690

## Prédicteurs à plus de deux modalités

- ▶ On crée plus de variables binaires.

Exemple : ethnicity dans le jeu de données cartes de crédit  
 $\in \{\text{Asian}, \text{Caucasian}, \text{African.American}\}$

$$x_{i,1} = \begin{cases} 1 & \text{si la } i\text{-ème pers. est Asian} \\ 0 & \text{si la } i\text{-ème pers. N'est PAS Asian} \end{cases}$$

$$x_{i,2} = \begin{cases} 1 & \text{si la } i\text{-ème pers. est Caucasian} \\ 0 & \text{si la } i\text{-ème pers. N'est PAS Caucasian} \end{cases}$$

Modèle résultant :  $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i$

$$\begin{cases} y_i = \beta_0 + \beta_1 + \varepsilon_i & \text{si la } i\text{-ème pers. est Asian} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{si la } i\text{-ème pers. est Caucasian} \\ \beta_0 + \varepsilon_i & \text{si la } i\text{-ème pers. est African.American} \end{cases}$$

- ▶ Nécessaire : une variable binaire de moins que le nombre de modalités (ici African.American)

# Extensions du modèle linéaire

Supprimer l'hypothèse additive : **interactions** et **non-linéarité**

## Interaction

- ▶ Avec les données de ventes, nous avons supposé que les effets sur sales produits par l'accroissement de publicité dans un médium est indépendant des effets des autres média.

- ▶ Ainsi
$$\begin{aligned}\widehat{\text{sales}} &= \hat{\beta}_0 \\ &+ \hat{\beta}_1 \times \text{TV} \\ &+ \hat{\beta}_2 \times \text{radio} \\ &+ \hat{\beta}_3 \times \text{newspaper}\end{aligned}$$

- ▶ Il se peut que dépenser de l'argent à la radio accroisse l'effet de la pub TV
- ▶ À budget fixé, en dépenser la moitié à la radio, l'autre à la TV est peut-être plus bénéfique que tout miser sur la TV ou la radio
- ▶ Marketing : synergie ;  
Statistique : interaction

## Modéliser les interactions

Le modèle prend alors la forme

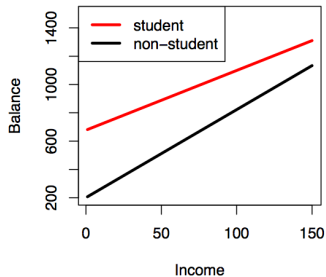
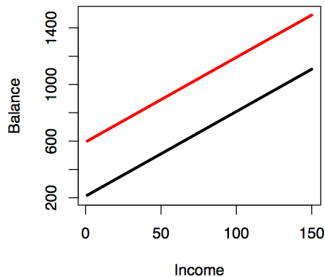
$$\begin{aligned}\text{sales} &= \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times (\text{radio} \times \text{TV}) + \varepsilon \\ &= \beta_0 + (\beta_1 + \beta_3' \times \text{radio}) \times \text{TV} + \beta_2 \times \text{radio} + \varepsilon.\end{aligned}$$

### Résultats

	coeff	Std.Err	t-stat	p-value
Intercept	6.7502	0.248	27.23	<0.0001
TV	0.0191	0.002	12.70	<0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	<0.0001

**Principe de hiérarchie** : on garde toujours l'effet principal lorsque l'effet d'interaction est significatif, quelque soit la  $p$ -value de l'effet principal !

# Interaction qualitatif - quantitatif

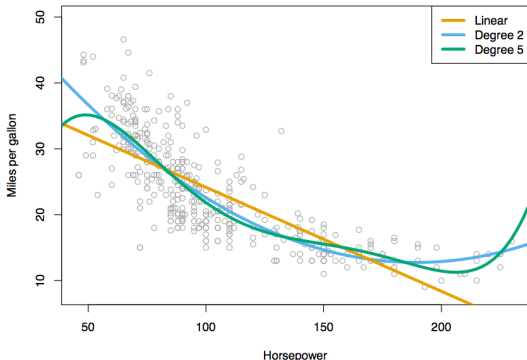


- Ordonnée à l'origine et pente changent tous les deux !



# Effets non linéaires

Régression polynomiale sur les données Auto



Suggère un modèle

$$\text{mpg} = \beta_0 + \beta_1 \times \text{horsepower} + \beta_2 \times \text{horsepower}^2 + \varepsilon$$

	coeff	Std.Err	t-stat	p-value
Intercept	56.9001	1.8004	31.6	<0.0001
horsepower	-0.4662	0.0311	-15.0	<0.0001
horsepower <sup>2</sup>	0.0012	0.0001	10.2	<0.0001

# Généralisation des modèles linéaires

- ▶ **Problèmes de classification** : régression logistique, support vector machine
- ▶ **Non-linéarité** : lissage à noyau, splines, modèles additifs généralisés, méthode de plus proches voisins
- ▶ **Interactions** : méthode basée sur des arbres, bagging, forêts aléatoires (random forests), et boosting (capturent aussi les non-linéarités)
- ▶ **Ajustement régularisé** : régression ridge et lasso