Algorithmique des données
Régression

Charlotte Pelletier
Univ. Bretagne Sud – IRISA Vannes
Basé sur le cours de Chloé Friguet.

12 février 2020

Objectifs du cours

- Introduction aux méthodes informatiques permettant d'exploiter des données dans le cadre de plusieurs problèmes fondamentaux :
  - Description, exploration des données, visualisation
  - Discrimination, classification.
  - Régression, prédiction.
- Objectifs :
  - Comprendre et distinguer les grandes catégories de problèmes se posant avec les données
  - Programmer et étudier des algorithmes permettant de classer ou prédire des données
  - Appréhender la complexité de certains problèmes ainsi que les outils mathématiques nécessaires

Objectifs du cours

- Introduction aux méthodes informatiques permettant d'exploiter des données dans le cadre de plusieurs problèmes fondamentaux :
  - Description, exploration des données, visualisation
  - Discrimination, classification.
  - Régression, prédiction.
- Objectifs :
  - Comprendre et distinguer les grandes catégories de problèmes se posant avec les données
  - Programmer et étudier des algorithmes permettant de classer ou prédire des données
  - Appréhender la complexité de certains problèmes ainsi que les outils mathématiques nécessaires
- **Partie Régression**
  - Connaitre le principe et les propriétés de l'algorithme de descente du gradient en régression
  - Compétence : Être capable de mettre en œuvre l'algorithme d'optimisation avec Python

Plan

Introduction

Modèle et fonction de coût

- Notations et modèle
- Fonction-coût

Algorithme de descente du gradient pour la régression linéaire

- Principe et propriétés
- Cas de la régression linéaire simple
- Cas de la régression linéaire multiple

Autres approches de résolution

- Estimateur des moindres carrés
- Approche du maximum de vraisemblance
- Comparaison

Application

La modélisation statistique

- **Données** : Recueil, présentation, analyse et restitution de l'information
  - ➜ Extraire des connaissances à partir de gros volumes de données observées
  - ➜ Biologie, médecine, marketing, géographie, psychologie, agroalimentaire, océanographie, etc.
- **Modéliser** : Concevoir une simplification de la réalité (observée) à un niveau d'approximation maîtrisé
- **Inférer** : Généraliser un résultat à partir d'observations
- Utilisation **conjointe** dans une démarche de compréhension ou de prédiction d'un phénomène à partir de l'application de théories

- **Modélisation** de la relation entre plusieurs variables

La modélisation statistique

- **Modélisation** de la relation entre plusieurs variables
  - Expliquer un phénomène, interpréter les liens entre des mesures
  - Prédire de nouvelles données
- Variable **à expliquer**, notée y
  - **quantitative** ou **qualitative**
  - variable à prédire, variable d'intérêt, variable endogène, variable dépendante, réponse

La modélisation statistique

- **Modélisation** de la relation entre plusieurs variables
  - Expliquer un phénomène, interpréter les liens entre des mesures
  - Prédire de nouvelles données
- Variable **à expliquer**, notée y
  - **quantitative** ou **qualitative**
  - variable à prédire, variable d'intérêt, variable endogène, variable dépendante, réponse
- Variables **explicatives**, notées $X^1, X^2, \dots X^d$
  - quantitatives, qualitatives ou les deux
  - variables prédictrices, variables exogènes, variables indépendantes, facteurs
  - $\mathbf{X}^j = \{x_i^j\}_{i=1}^m$ pour j allant de 1 à d

La modélisation statistique

- **Modélisation** de la relation entre plusieurs variables
  - Expliquer un phénomène, interpréter les liens entre des mesures
  - Prédire de nouvelles données
- Variable **à expliquer**, notée y
  - **quantitative** ou **qualitative**
  - variable à prédire, variable d'intérêt, variable endogène, variable dépendante, réponse
- Variables **explicatives**, notées $X^1, X^2, \dots X^d$
  - quantitatives, qualitatives ou les deux
  - variables prédictrices, variables exogènes, variables indépendantes, facteurs
  - $\mathbf{X}^j = \{x_i^j\}_{i=1}^m$ pour j allant de 1 à d
- L'analyse de la relation entre y et $X^1, X^2, \dots X^d$ consiste à définir une fonction f telle que pour toute observation i :

$$y_i \approx f(x_i^1, x_i^2, \cdots, x_i^d)$$

Données

- **Apprentissage supervisé** : dans les données observées, on connaît la "vraie" valeur de la variable de sortie et on cherche à comprendre/prédire le lien supposé entre les variables d'entrée et de sortie
- **Nature de la variable de sortie** (Y) ?
  - **quantitative** : **régression**
  - qualitative (à 2 ou >2 modalités) : classification (binaire / multiclasses)
- **Nature et nombre de variables d'entrée** (X) ?
  - nature : **qualitatives** et/ou **quantitatives**
  - **Une** seule variable
    - Peu fréquent en pratique, mais utile pour bien comprendre ce qu'il se passe $\Rightarrow$ visualisation
  - **Plusieurs** variables
    - Plusieurs = de quelques dizaines à plusieurs (dizaines de) milliers $\Rightarrow$ sélection de variables

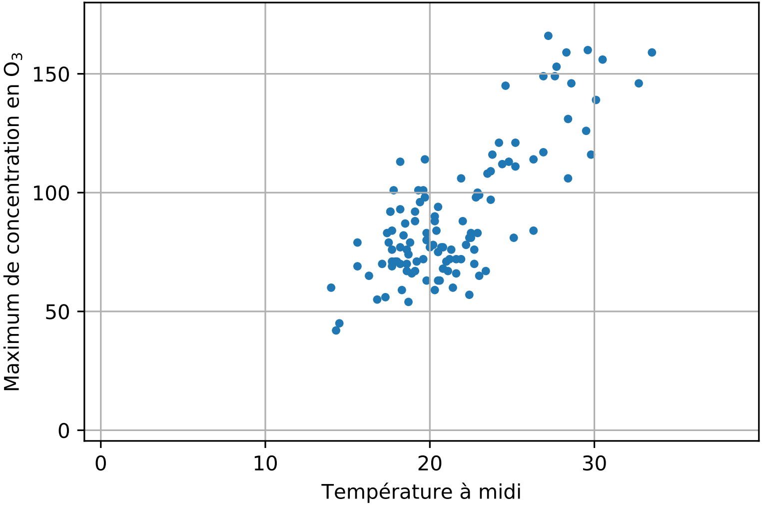
Exemple A

Pour des raisons de santé publique, on s'intéresse à la concentration d'ozone O_3 dans l'air. On cherche en particulier à savoir si on peut expliquer le taux maximal d'ozone de la journée (en $\mu g/ml$) à partir d'autres variables météo mesurées dans la station de Rennes.

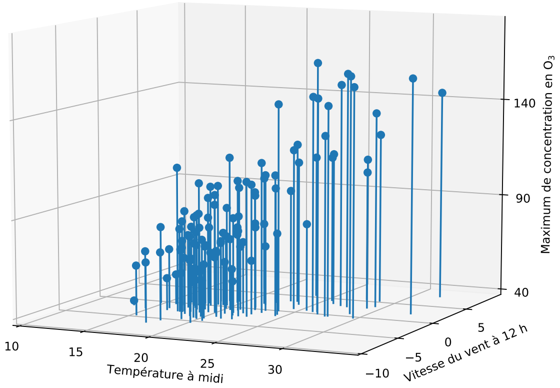
Extrait des données :

O_3 max	Température à midi	Vitesse du vent à 12h	$\cdots$
87	18.5	-1.7101	$\cdots$
82	18.4	-4.0000	$\cdots$
92	17.6	1.8794	$\cdots$
114	19.7	0.3473	$\cdots$
94	20.5	-2.9544	$\cdots$
80	19.8	-5.0000	$\cdots$
79	15.6	-1.8794	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Les données (Air Breizh - 2001) sont issues de Régression : Théorie et applications, Cornillon P.A. et Matzner-Lober E. (2006) Springer



Maximum de concentration en O_3
150
100
50
0
0
10
20
30
Température à midi



Maximum de concentration en O_3
140
90
40
10
15
20
25
30
Température à midi
−10
−5
0
5
Vitesse du vent à 12 h

Plan

Introduction

Modèle et fonction de coût

- Notations et modèle
- Fonction-coût

Algorithme de descente du gradient pour la régression linéaire

- Principe et propriétés
- Cas de la régression linéaire simple
- Cas de la régression linéaire multiple

Autres approches de résolution

- Estimateur des moindres carrés
- Approche du maximum de vraisemblance
- Comparaison

Application

Notations

- Données d'apprentissage : $\{\mathbf{x}_i, y_i\}_{i=1}^m$
  - observations (entrées) : $\mathbf{x}_i \in \mathbb{R}^d$
  - mesure d'intérêt (sortie à prédire) : $y_i \in \mathcal{Y}$
- Fonction de prédiction : $f : \mathbb{R}^d \mapsto \mathcal{Y}$
  - régression : f prédit un réel ($\mathcal{Y} = \mathbb{R}$)
  - classification multi-classes : f prédit un entier entre 1 et k ($\mathcal{Y} = \{1, ..., k\}$)

Notations

- Données d'apprentissage : $\{\mathbf{x}_i, y_i\}_{i=1}^m$
  - observations (entrées) : $\mathbf{x}_i \in \mathbb{R}^d$
  - mesure d'intérêt (sortie à prédire) : $y_i \in \mathcal{Y}$
- Fonction de prédiction : $f : \mathbb{R}^d \mapsto \mathcal{Y}$
  - **régression : f prédit un réel ($\mathcal{Y} = \mathbb{R}$)**
  - classification multi-classes : f prédit un entier entre 1 et k ($\mathcal{Y} = \{1, ..., k\}$)

Modèle

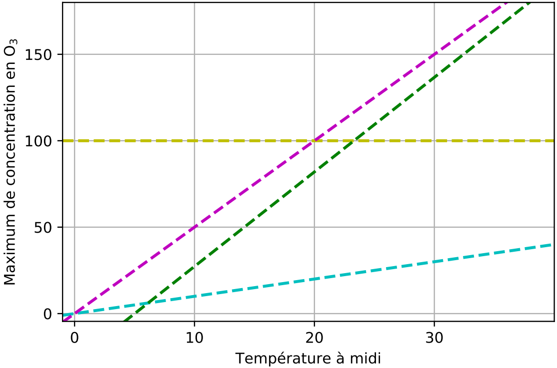
- Un modèle est une équation mathématique qui va permettre de décrire le lien entre la variable d'intérêt $\mathbf{Y} = [y_1, y_2, \cdots, y_m]$ et les variables explicatives $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m]^T$ (observations en ligne, variables en colonnes) :
$$y_i \approx f(\mathbf{x}_i) = f(x_i^1, x_i^2, \cdots, x_i^d)$$
avec $x_i^j \in \mathbb{R}$ la valeur de la variable j pour l'observation i
- La forme de f dépend du contexte (type de données) et du niveau de simplification souhaité
  - On fixe une classe de fonctions : linéaires, polynomiales, *etc*
  - Cas du modèle linéaire :
$$y_i \approx f_{\boldsymbol{\beta}}(\mathbf{x}_i) = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + \ldots + \beta_d x_i^d$$
- D'une faccon générale, un modèle est définit à partir de **paramètres** (β_j)
  - ils sont inconnus
  - il faut les estimer
  - mais, comment ?

Analyse univariée

- une seule variable explicative : $f_{\boldsymbol{\beta}}$ est une fonction affine (droite)

$$f_{\boldsymbol{\beta}}(\mathbf{x}_i) = \beta_0 + \beta_1 x_i^1$$

  - β_0 est l'ordonnée à l'origine
  - β_1 est la pente



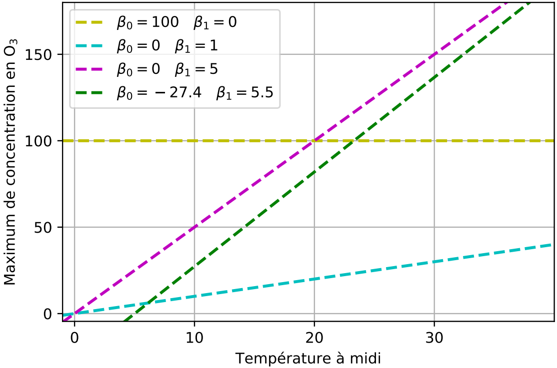
Modèle linéaire

Analyse univariée

- une seule variable explicative : $f\boldsymbol{\beta}$ est une fonction affine (droite)

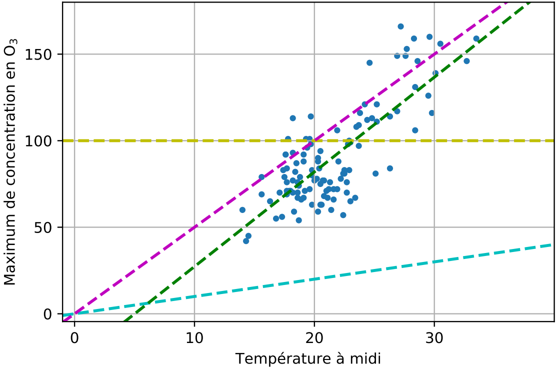
$$f_{\boldsymbol{\beta}}(\mathbf{x}_i) = \beta_0 + \beta_1 x_{i1}$$

  - β_0 est l'ordonnée à l'origine
  - β_1 est la pente



Analyse univariée

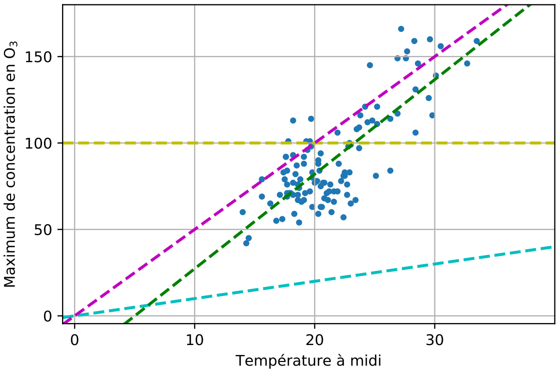
- Comment choisir $\boldsymbol{\beta} = (\beta_0, \beta_1)$?



Modèle linéaire

Analyse univariée

- Comment choisir $\boldsymbol{\beta} = (\beta_0, \beta_1)$?
  - $\boldsymbol{\beta}$ tel que $f_{\boldsymbol{\beta}}(\mathbf{x})$ est proche de y pour **toutes** les données d'apprentissage $\{\mathbf{x}_i, y_i\}_{i=1}^m$



- **Fonction-coût** (ou fonction-objectif)
  - fonction qui sert de critère pour répondre à notre problématique et qu'on va chercher à minimiser (ou maximiser)
  - notée $\mathbf{J}(\boldsymbol{\beta})$
- On veut $\boldsymbol{\beta}$ tel que $f_{\boldsymbol{\beta}}(\mathbf{x})$ soit proche de y pour **toutes** les données d'apprentissage $\{\mathbf{x}_i, y_i\}_{i=1}^{m}$:

$$\hat{y}_i = f_{\boldsymbol{\beta}}(\mathbf{x}_i) \approx y_i \quad \forall i \in \{1, \cdots, m\}$$

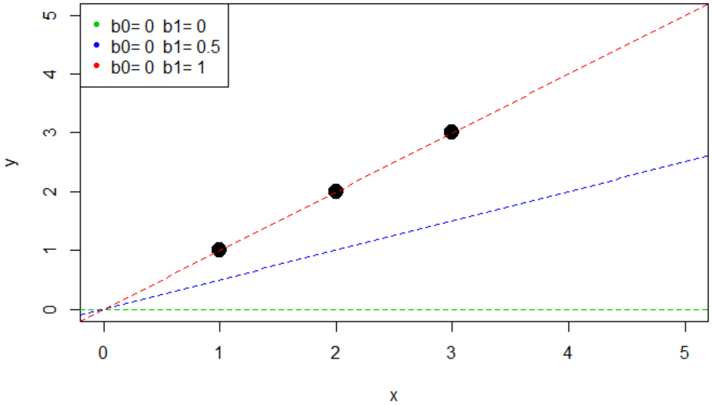
- Trouver le meilleur couple $(\beta_0; \beta_1)$ équivaut à minimiser le coût (quadratique) **global** des erreurs :

$$\mathbf{J}(\boldsymbol{\beta}) = \frac{1}{2m} \sum_{i=1}^{m} \Big(f_{\boldsymbol{\beta}}(\mathbf{x}_i) - y_i \Big)^2$$

➜ Meilleur couple (β_0, β_1) = solution de :

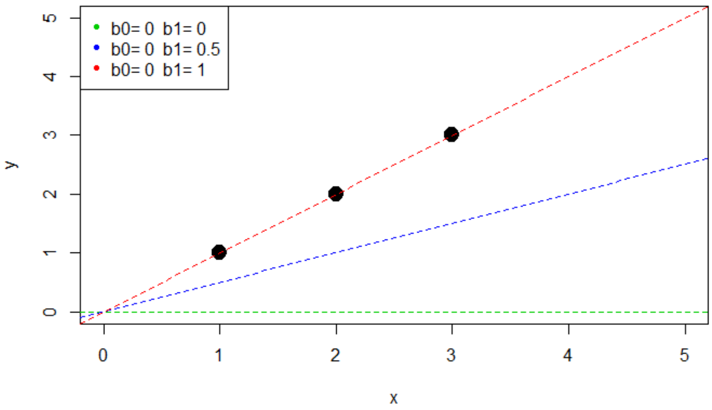
$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \Big(\mathbf{J}(\boldsymbol{\beta}) \Big)$$

Parmi toutes les droites possibles, on cherche la droite pour laquelle la somme des carrés des écarts verticaux des points à la droite est minimale.



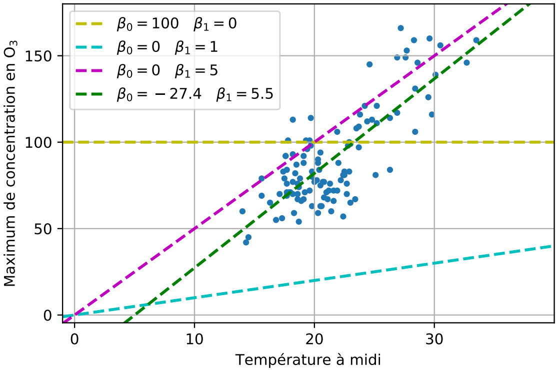
- Calcul de $\mathbf{J}(\beta)$?

$\boldsymbol{\beta}$	(0,0)	(0,0.5)	(0,1)
$\mathbf{J}(\boldsymbol{\beta})$			



- Calcul de $\mathbf{J}(\beta)$?

$\boldsymbol{\beta}$	(0,0)	(0,0.5)	(0,1)
$\mathbf{J}(\boldsymbol{\beta})$	2.33	0.58	0



$\boldsymbol{\beta}$	(100,0)	(0,1)	(0,5)	(27.42,5.47)
$\mathbf{J}(\boldsymbol{\beta})$	440.72	2678.37	303.50	151.80

Plan

Introduction

Modèle et fonction de coût

Notations et modèle

Fonction-coût

Algorithme de descente du gradient pour la régression linéaire

Principe et propriétés

Cas de la régression linéaire simple

Cas de la régression linéaire multiple

Autres approches de résolution

Estimateur des moindres carrés

Approche du maximum de vraisemblance

Comparaison

Application

Principe

- Objectif : trouver le minimum d'une fonction-coût
- Principe : algorithme itératif
  1. initialisation : $\beta^{(0)}$
  2. à chaque étape k, modifier $\boldsymbol{\beta}^{(k-1)}$ pour faire diminuer $\mathbf{J}(\boldsymbol{\beta}^{(k)})$
  3. arrêt lorsque le minimum est atteint

Itération k de l'algorithme de descente du gradient

Pour le paramètre β_j

$$\beta_j^{(k)} \quad := \quad \beta_j^{(k-1)} - \alpha \frac{\partial}{\partial \beta_j} \mathbf{J}(\boldsymbol{\beta}^{(k-1)})$$

avec :

- $\frac{\partial}{\partial \beta_j}$:
- α :

Principe

- Objectif : trouver le minimum d'une fonction-coût
- Principe : algorithme itératif
  1. initialisation : $\beta^{(0)}$
  2. à chaque étape k, modifier $\boldsymbol{\beta}^{(k-1)}$ pour faire diminuer $\mathbf{J}(\boldsymbol{\beta}^{(k)})$
  3. arrêt lorsque le minimum est atteint

Itération k de l'algorithme de descente du gradient

Pour le paramètre β_j

$$\beta_j^{(k)} \quad := \quad \beta_j^{(k-1)} - \alpha \frac{\partial}{\partial \beta_j} \mathbf{J}(\boldsymbol{\beta}^{(k-1)})$$

avec :

- $\frac{\partial}{\partial \beta_j}$: dérivée partielle
- α : pas d'apprentissage

Remarques et propriétés

- Signe de la dérivée : augmentation ou diminution de $\boldsymbol{\beta}^{(k)}$
- Critère de convergence : diminution de $\mathbf{J}(\beta) < \varepsilon$ lors d'une itération (par ex. $\varepsilon = 10^{-3}$)
- Choix de α ?
  - Si trop petit, alors algorithme lent.
  - Si trop grand, alors non convergence possible.
  - En pratique, on teste plusieurs valeurs.
  - Le gradient va diminuer à l'approche du minimum
    amélioration = grands pas au début puis plus petits pas quand on approche du minimum
- Initialisation : attention si proche d'un minimum local
- Échelle des X_j similaire : non divergence / convergence plus rapide
  - **Normalisation** (données centrées-réduites)
    $\forall j \in \{1, \cdots, d\})$: $\mathbf{X}^j := \frac{\mathbf{X}^j - \bar{\mathbf{X}}^j}{r^j}$, avec r^j l'écart-type des $\mathbf{X}^j$

- Analyse de la relation entre $\mathbf{Y}$ et $\mathbf{X}^1$ avec une fonction f linéaire telle que $\forall i \in \{1, \cdots, m\}$:

$$y_i \approx f_{\boldsymbol{\beta}}(\mathbf{x}_i) = \beta_0 + \beta_1 x_i^1$$

- Itération k de l'algorithme de descente du gradient ?

- Analyse de la relation entre $\mathbf{Y}$ et $\mathbf{X}^1$ avec une fonction f linéaire telle que $\forall i \in \{1, \cdots, m\}$:

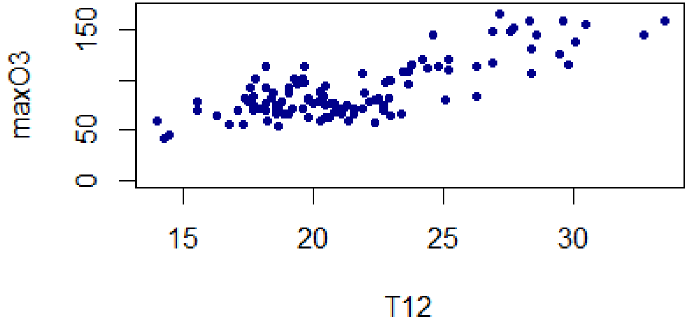
$$y_i \approx f_{\boldsymbol{\beta}}(\mathbf{x}_i) = \beta_0 + \beta_1 x_i^1$$

- Itération k de l'algorithme de descente du gradient ?

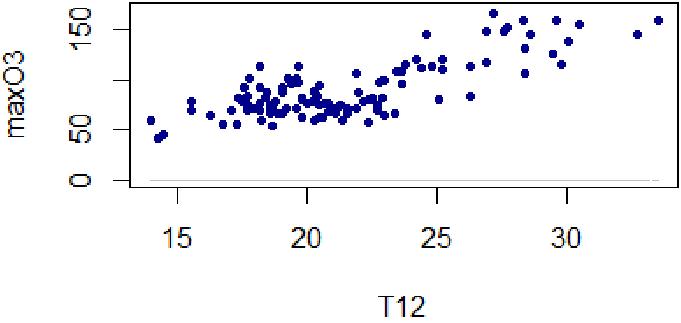
Itération k de l'algorithme de descente du gradient - régression linéaire simple

$$\begin{cases} \beta_0^{(k)} := \beta_0^{(k-1)} - \alpha \dfrac{1}{m} \displaystyle\sum_{i=1}^{m} \left(f_{\boldsymbol{\beta}^{(k-1)}}(\mathbf{x}_i) - y_i \right) \\ \beta_1^{(k)} := \beta_1^{(k-1)} - \alpha \dfrac{1}{m} \displaystyle\sum_{i=1}^{m} \left(f_{\boldsymbol{\beta}^{(k-1)}}(\mathbf{x}_i) - y_i \right) x_i^i \end{cases}$$

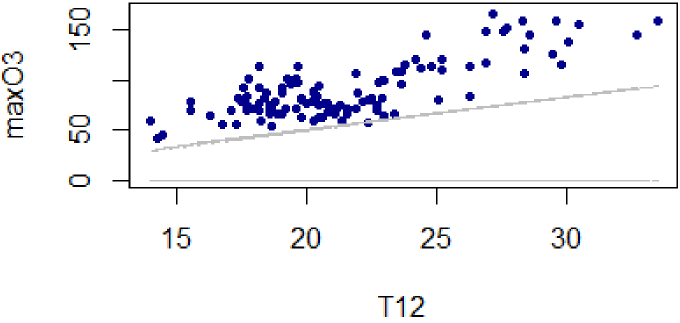
La régression linéaire (simple)



it. β	1	10	20	30	40	50	100
J(β)							

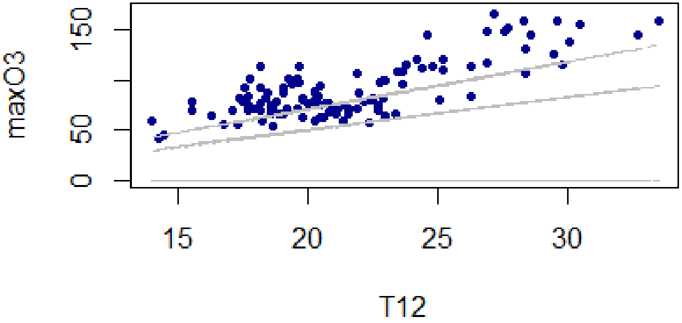


it.	1	10	20	30	40	50	100
β	(0,0)						
$J(\beta)$	3650.76						



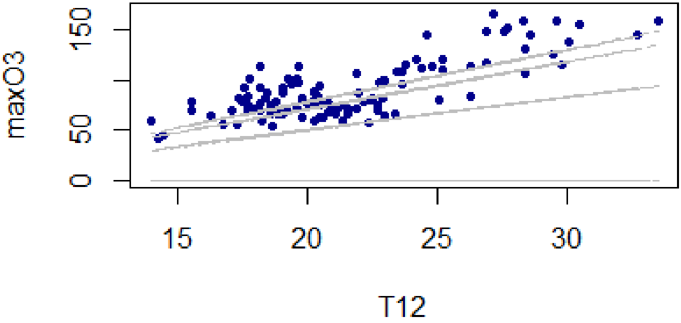
it.	1	10	20	30	40	50	100
β	(0,0)	(-16.39,3.33)					
$J(\beta)$	3650.76	677.30					

La régression linéaire (simple)

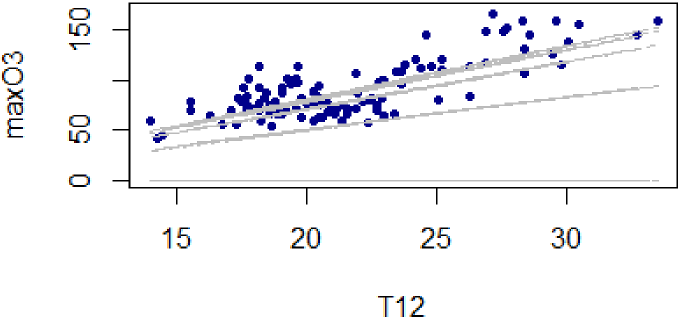


it.	1	10	20	30	40	50	100
β	(0,0)	(-16.39,3.33)	(-23.41,4.72)				
$J(\beta)$	3650.76	677.30	215.54				

La régression linéaire (simple)

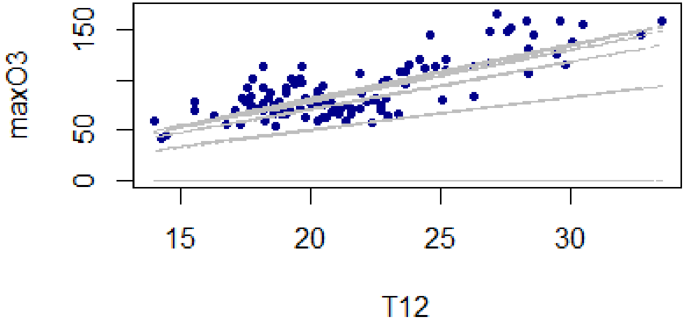


it.	1	10	20	30	40	50	100
β	(0,0)	(-16.39,3.33)	(-23.41,4.72)	(-25.97,5.20)			
$J(\beta)$	3650.76	677.30	215.54	159.34			



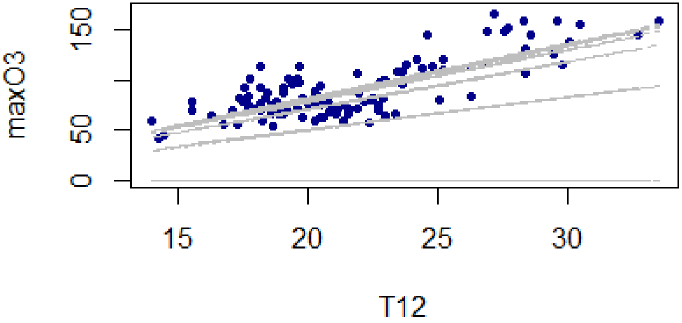
it.	1	10	20	30	40	50	100
β	(0,0)	(-16.39,3.33)	(-23.41,4.72)	(-25.97,5.20)	(-26.89,5.38)		
$J(\beta)$	3650.76	677.30	215.54	159.34	152.50		

La régression linéaire (simple)

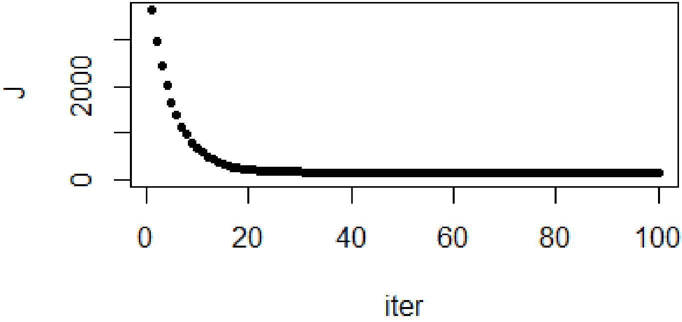


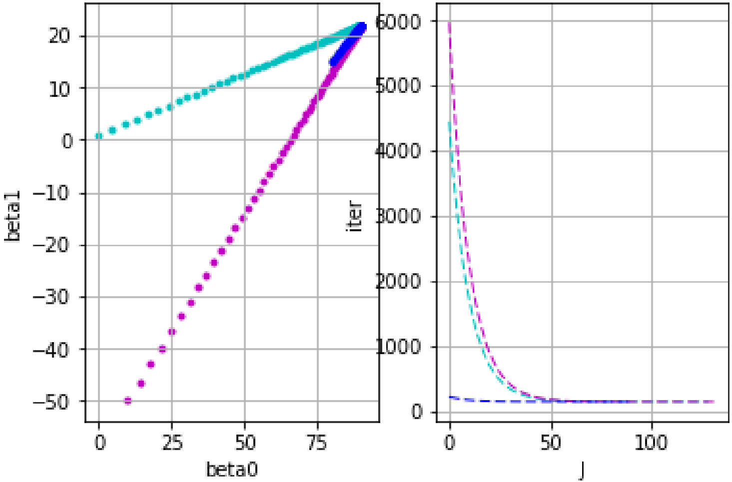
it.	1	10	20	30	40	50	100
β	(0,0)	(-16.39,3.33)	(-23.41,4.72)	(-25.97,5.20)	(-26.89,5.38)	(-27.23,5.44)	
$J(\beta)$	3650.76	677.30	215.54	159.34	152.50	151.67	

La régression linéaire (simple)



it.	1	10	20	30	40	50	100
β	(0,0)	(-16.39,3.33)	(-23.41,4.72)	(-25.97,5.20)	(-26.89,5.38)	(-27.23,5.44)	(-27.42,5.47)
$J(\beta)$	3650.76	677.30	215.54	159.34	152.50	151.67	151.55


J
0
2000
0
20
40
60
80
100
iter



20
10
0
−10
−20
−30
−40
−50
beta1
0
25
50
75
beta0
6000
5000
4000
3000
2000
1000
0
iter
0
50
100
J

- Analyse de la relation entre $\mathbf{Y}$ et toutes les variables $[\mathbf{X}^1, \mathbf{X}^2, \cdots, \mathbf{X}^d]$ avec une fonction f linéaire telle que $\forall i \in \{1, \cdots, m\}$:

$$y_i \approx f_{\boldsymbol{\beta}}(\mathbf{x}_i) = f_{\boldsymbol{\beta}}(x_i^1, x_i^2 \cdots, x_i^d) = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + \ldots + \beta_d x_i^d$$

- Notation matricielle :

$$f_{\boldsymbol{\beta}}(\mathbf{X}) \approx \tilde{\mathbf{X}}\boldsymbol{\beta}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ \\ y_m \end{pmatrix} \approx \begin{pmatrix} 1 & x_1^1 & x_1^2 & \ldots & x_1^d \\ 1 & x_2^1 & x_2^2 & \ldots & x_2^d \\ \vdots & \vdots & \vdots & \vdots & \\ \\ 1 & x_m^1 & x_m^2 & \ldots & x_m^d \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \\ \beta_d \end{pmatrix}$$

avec $\mathbf{X} \in \mathbb{R}^{m \times d}$, $\tilde{\mathbf{X}} \in \mathbb{R}^{m \times (d+1)}$ et $\boldsymbol{\beta} \in \mathbb{R}^{d+1}$

- Itération k de l'algorithme de descente du gradient ?

La régression linéaire (multiple)

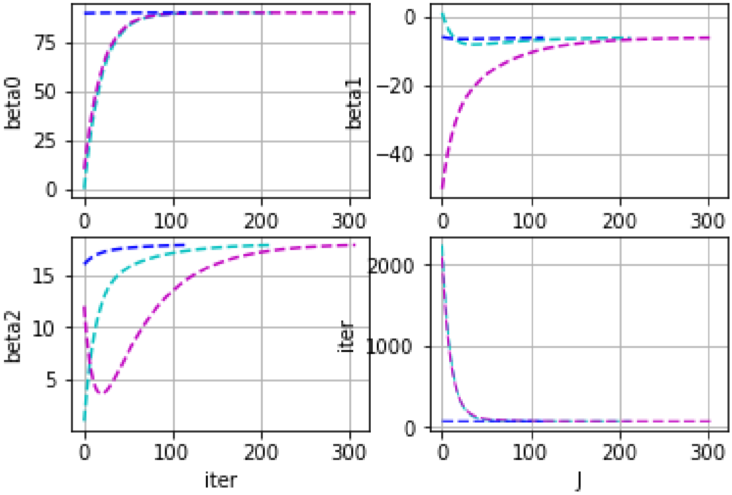
- Itération k de l'algorithme de descente du gradient ?

Itération k de l'algorithme de descente du gradient - régression linéaire multiple

Pour chaque variable β_j :

$$\beta_j^{(k)} \quad := \quad \beta_j^{(k-1)} - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(f_{\boldsymbol{\beta}^{(k-1)}}(\mathbf{x}_i) - y_i \right) x_i^j$$

Remarque : $\forall i, x_i^0 = 1$


beta0
75
50
25
0
0
100
200
300
beta1
0
−20
−40
0
100
200
300
beta2
15
10
5
0
100
200
300
iter
iter
2000
1000
0
0
100
200
300
J

Pour résumer

- Algorithme itératif d'optimisation
  - à chaque itération, on fait diminuer la fonction-coût dans la direction opposée du gradient
- ⚠ mimimum local/global $\Rightarrow$ nécessite plusieurs initialisations
- Pas d'apprentissage α fixe
  - amélioration possible avec des variantes (par exemple gradient conjugué, BFGS [1]),
  - mais variantes plus complexe
- Autres modèles (hypothèses), autres fonctions-pertes

1. Broyden-Fletcher-Goldfarb-Shanno

Plan

Introduction

Modèle et fonction de coût

Notations et modèle

Fonction-coût

Algorithme de descente du gradient pour la régression linéaire

Principe et propriétés

Cas de la régression linéaire simple

Cas de la régression linéaire multiple

Autres approches de résolution

Estimateur des moindres carrés

Approche du maximum de vraisemblance

Comparaison

Application

Dans le cas de la régression linéaire : la fonction-coût est toujours **convexe** $\Rightarrow$ il existe donc un **minimum global**.

La régression linéaire

Dans le cas de la régression linéaire : la fonction-coût est toujours **convexe** $\Rightarrow$ il existe donc un **minimum global**.

Pour trouver ce minimum, il faut trouver le point où le gradient de la fonction s'annule :

$$\frac{\partial \mathcal{J}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$$

On peut montrer pour le cas de la régression linéaire simple que ce minimum est atteint pour :

$$\begin{aligned} \beta_0 &= \bar{y} - \beta 1 \bar{x} \\ \beta_1 &= \frac{cov(x,y)}{var(x)} \end{aligned}$$

Estimateur des moindres carrés

- **Moindres carrés** : minimisation des *carrés des écarts* entre les observations et le modèle

$$\begin{aligned}\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{m} \left(y_i - f_{\boldsymbol{\beta}}(\mathbf{x}_i) \right)^2 &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{m} \left(y_i - (\beta_0 + \beta_1 x_i^1 + \ldots + \beta_d x_i^d) \right)^2 \\ &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \, ||\mathbf{Y} - \tilde{\mathbf{X}}\boldsymbol{\beta}||^2\end{aligned}$$

Théorème de Gauss-Markov

- **Hypothèse** : les erreurs sont centrées (espérance nulle), non-corrélées et de même variance (homoscédasticité)
- **Solution explicite** pour $\hat{\boldsymbol{\beta}}$:

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

- C'est le meilleur estimateur non-biaisé

Le modèle linéaire gaussien

Hypothèses

- $\mathbf{X}$ est de plein rang (variables explicatives non colinéaires)
- les erreurs sont centrées, non-corrélées et de même variance σ^2

$$\epsilon_i = f_{\boldsymbol{\beta}}(\mathbf{x}_i) - y_i$$

- les erreurs suivent une loi $\mathcal{N}$ormale (de paramètres 0 et σ^2) :

$$\begin{cases} \epsilon_i \sim \mathcal{N}(0, \sigma^2) \\ \epsilon_i \text{ indépendants} \end{cases}$$

et donc : $Y \sim \mathcal{N}(\tilde{\mathbf{X}}\boldsymbol{\beta}, \sigma^2\mathbb{I})$

Le modèle linéaire gaussien

Hypothèses

- $\mathbf{X}$ est de plein rang (variables explicatives non colinéaires)
- les erreurs sont centrées, non-corrélées et de même variance σ^2

$$\epsilon_i = f_{\boldsymbol{\beta}}(\mathbf{x}_i) - y_i$$

- les erreurs suivent une loi $\mathcal{N}$ormale (de paramètres 0 et σ^2) :

$$\begin{cases} \epsilon_i \sim \mathcal{N}(0, \sigma^2) \\ \epsilon_i \text{ indépendants} \end{cases}$$

et donc : $Y \sim \mathcal{N}(\tilde{\mathbf{X}}\boldsymbol{\beta}, \sigma^2\mathbb{I})$

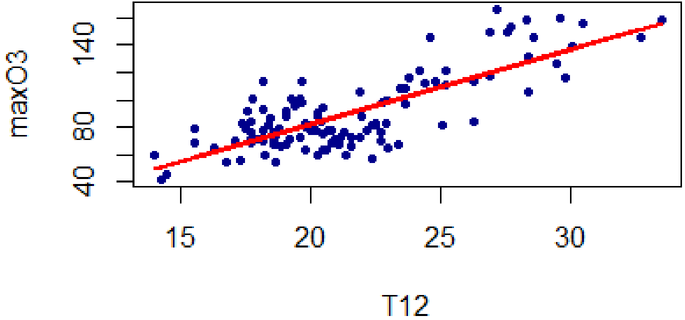
Remarque : plus d'hypothèses sont nécessaires ici

- Maximisation de la (log-)vraisemblance du modèle

$$\begin{aligned}\underset{\boldsymbol{\beta},\sigma^2}{\text{argmax}}\, \mathcal{L}(\mathbf{Y}, \boldsymbol{\beta}, \sigma^2) &= \underset{\boldsymbol{\beta}\sigma^2}{\text{argmax}}\, log\left(\prod_{i=1}^{m} \phi(\mathbf{Y}, \boldsymbol{\beta})\right) \\ &= \underset{\boldsymbol{\beta},\sigma^2}{\text{argmax}}\left(-\frac{m}{2} log\sigma^2 - \frac{m}{2} log 2\pi - \frac{1}{2\sigma^2}||\mathbf{Y} - \tilde{\mathbf{X}}\boldsymbol{\beta}||\right)\end{aligned}$$

- **Solution explicite** (et identique) pour $\hat{\boldsymbol{\beta}}$:

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



$$\beta = (-27.420; 5.469)$$

it.	1	10	20	30	40	50	100
β	(0 ;0)	(-16.39 ;3.33)	(-23.41 ;4.72)	(-25.97 ; 5.20)	(-26.89 ; 5.38)	(-27.23 ; 5.44)	(-27.42 ;5.47)
$J(\beta)$	3650.76	77.30	215.54	159.34	152.50	51.67	151.55

Comparaison

- **Algorithme de descente du gradient**
  - paramètre α à fixer par l'utilisateur
  - algorithme itératif, solution approchée
  - rapide pour de grand d (100,1000,10000...)
- **Estimateur des moindres carrés / Modèle linéaire gaussien**
  - Pas de paramètre à fixer
  - Pas d'itération, solution exacte
  - Complexité en d^3 (calcul de $X^T X^{-1}$) (lent ou impossible pour de grand d) : sélection de variables, régularisation

Plan

Introduction

Modèle et fonction de coût

- Notations et modèle
- Fonction-coût

Algorithme de descente du gradient pour la régression linéaire

- Principe et propriétés
- Cas de la régression linéaire simple
- Cas de la régression linéaire multiple

Autres approches de résolution

- Estimateur des moindres carrés
- Approche du maximum de vraisemblance
- Comparaison

Application

Application

Les données portent sur des informations collectées au début des années 1970 par les services de la ville de Boston (USA) au sujet du logement dans divers quartiers :

- ...
- 6- AGE : proportion de logements occupés par leur propriétaires et construits avant 1940
- 7- DIS : distance (pondérée) à 5 bassins d'emplois
- ...
- 12- LSTAT : % de la population de milieu socio-économique plus défavorisé
- 13- MEDV : valeur médiane des logements occupés par leur propriétaires ($\times \$1,000$)

Objectif : étudier le lien entre la valeur des logements d'un quartier et l'ancienneté, la distance aux bassins d'emploi et le niveau socio-économique du quartier.

The data was originally published by Harrison, D. and Rubinfeld, D.L. 'Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978.

TP / CR à rendre

- Importer les données
- Construire les fonctions nécessaires : f, coût, gradient puis implémenter l'algorithme de descente de gradient
  - Cas de la régression linéaire simple
  - Etendre à la régression linéaire multiple
  - Retourner les coefficients du modèle et les valeurs de la fonction-coût pour toutes les itérations
- Faire varier les paramètres de l'algorithme (initialisation, pas) et commenter.
- Comparer avec la solution du modèle linéaire gaussien