

FONDEMENTS STATISTIQUES ET MATHÉMATIQUES

PRÉPARATION DU TERRAIN

APERÇU

1. Modélisation
2. Distributions
3. Théorème de la limite centrale
4. Estimation
5. Théorème de Bayes
6. Algèbre matricielle
7. Valeurs propres et vecteurs propres
8. Régression
9. Optimisation

Monde réel



Théorie

Repérage des détails
pertinents pour la
description et la
traduction d'objets
du monde réel en
variables de modèle

Modèle



LES MODÈLES EN GÉNÉRAL

Principes de base de la modélisation

- Examiner un système
- Écrire un ensemble de règles et d'équations qui décrivent l'essence du système
- Ignorer les détails qui compliquent les choses et qui sont « moins » importants

Modélisation statistique

- Habituellement, un ensemble d'équations comprenant des paramètres
- Les paramètres sont appris (le modèle est « entraîné ») à l'aide de multiples observations de données
- Échantillon de données c. population

HEURISTIQUE DE LA MODÉLISATION

Voici les étapes de base de l'élaboration d'un modèle statistique :

- **Définition des objectifs**
- **Collecte des données**
- **Choix de la structure du modèle**
- **Préparation des données**
- **Sélection et suppression d'attributs**
- **Élaboration des modèles admissibles**
- **Finalisation du modèle**
- **Mise en œuvre et surveillance**

DONNÉES ET DISTRIBUTIONS

Si un attribut de données peut être caractérisé par une distribution, poser ces **quatre questions fondamentales** :

1. La variable ne peut-elle prendre que des valeurs **discrètes**?
2. La distribution des données est-elle **symétrique**?
3. La variable a-t-elle des limites **supérieures** et **inférieures** théoriques?
4. Quelle est la probabilité d'avoir des **valeurs extrêmes** dans la distribution?

Distribution	Fonction de densité $f(x)$	Moyenne	Variance	Notes
uniforme $U(a, b)$	$\frac{1}{b-a}$ pour $a \leq x \leq b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	La plupart des langages fournissent des générateurs de valeurs aléatoires pour $U(a, b)$; sert à générer des v.a. avec d'autres distributions
de Gauss $N(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ pour $x \in \mathbb{R}$	μ	σ^2	Si $X \sim N(\mu, \sigma^2)$, alors $\frac{X-\mu}{\sigma} \sim N(0,1)$ (et <i>vice-versa</i>); utilisée très souvent
de Poisson $P(\lambda), \lambda \geq 0$	$\frac{\lambda^x}{x!} e^{-\lambda}$ pour $x = 0, 1, 2, \dots$	λ	λ	Estime le nombre d'événements qui se produisent dans un intervalle de temps continu (nombre d'appels reçus dans des intervalles d'une heure)
binomiale $\mathcal{B}(N, p), N \in \mathbb{N},$ $p \in [0, 1]$	$\binom{N}{x} p^x (1-p)^{N-x}$ pour $x = 0, 1, \dots, N$	Np	$Np(1-p)$	Décrit la probabilité exacte de x succès dans N essais indépendants si la probabilité de succès d'un seul essai est p (nombre de faces dans N tirages à pile ou face)
log-normale $\Lambda(\mu, \sigma^2)$	$\frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2}$ pour $x > 0$	$e^{(\mu + \sigma^2/2)}$	$e^{(2\mu + \sigma^2)} [e^{\sigma^2} - 1]$	Si $\ln X \sim N(\mu, \sigma^2)$, alors $X \sim \Lambda(\mu, \sigma^2)$ (et <i>vice-versa</i>); désaxée vers la droite

DISTRIBUTIONS À PLUSIEURS VARIABLES

Les distributions univariées sont des outils de modélisation utiles, surtout lorsque les variables considérées sont **indépendantes**.

Dans la pratique, ce n'est généralement pas le cas. Une **distribution à plusieurs variables** $P(X_1, \dots, X_n)$ donne la probabilité que chaque valeur X_1, \dots, X_n se situe dans une aire de distribution donnée. La **distribution normale à plusieurs variables** $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ comprend une fonction de densité

$$f(x_1, \dots, x_n) := f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

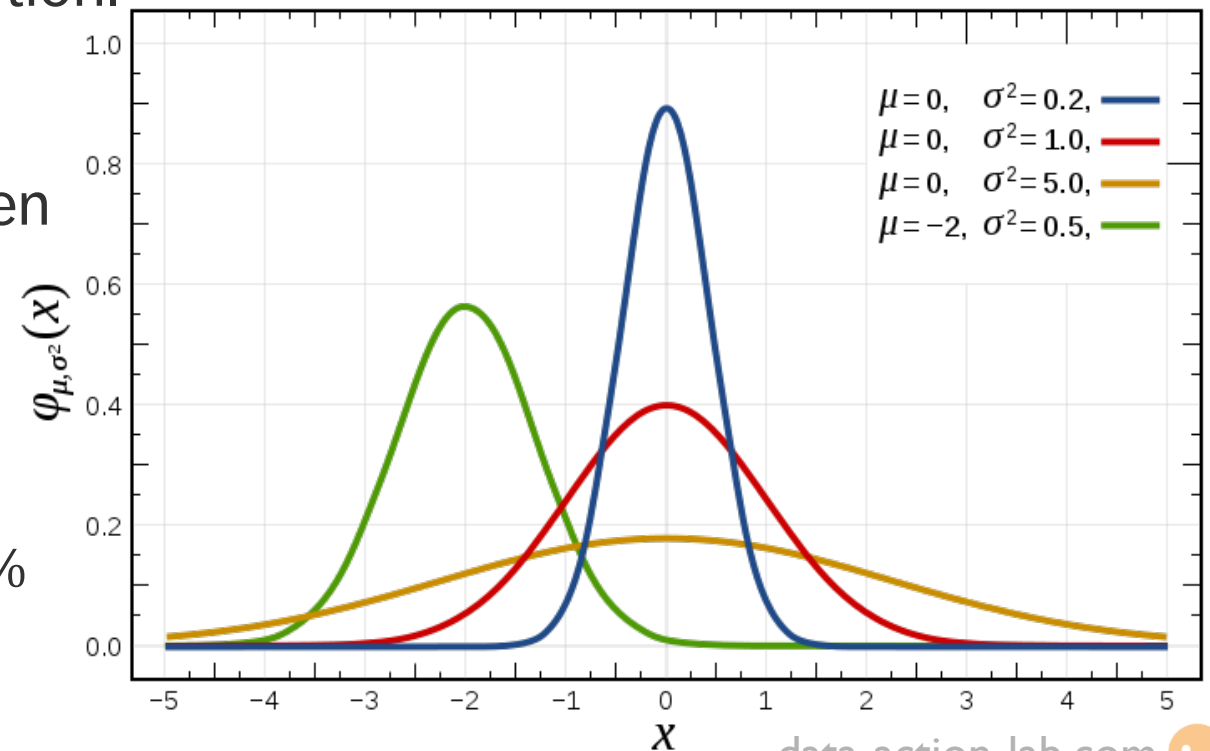
où $\boldsymbol{\mu}$ est le vecteur moyen et $\boldsymbol{\Sigma}$ la matrice de covariance.

DISTRIBUTION NORMALE

L'équation $N(\mu, \sigma^2)$ est **entièrement caractérisée** par la moyenne μ et par l'écart-type σ , ce qui réduit les besoins d'estimation.

La probabilité qu'une valeur soit extraite peut être obtenue si nous savons combien de multiples de σ la séparent de μ

- à l'intérieur de σ par rapport à μ : $\approx 68\%$
- à l'intérieur de 2σ par rapport à μ : $\approx 95\%$
- à l'intérieur de 3σ par rapport à μ : $\approx 99.7\%$



DISTRIBUTION NORMALE

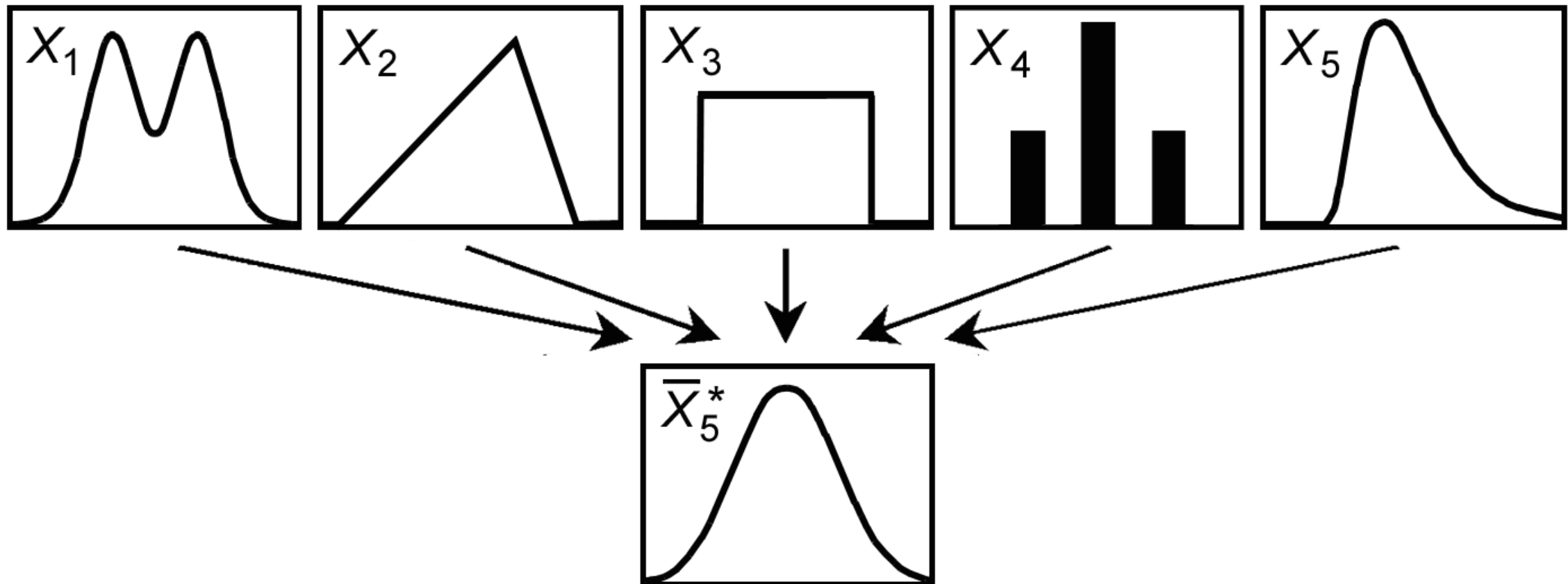
La distribution normale est la mieux adaptée aux données répondant aux exigences minimales suivantes :

- Forte tendance des données à prendre une valeur centrale
- Les écarts positifs et négatifs par rapport à cette valeur centrale sont également probables
- La fréquence des écarts diminue rapidement à mesure que l'on s'éloigne de la valeur centrale

La symétrie des écarts conduit à une **asymétrie** égale à zéro; une faible probabilité d'écarts importants par rapport à la valeur centrale n'entraîne aucun **aplatissement**.

Son omniprésence dans les affaires humaines est liée au **théorème de la limite centrale**.

THÉORÈME DE LA LIMITE CENTRALE EN ACTION



ESTIMATION

L'un des objectifs des statistiques est d'essayer de **comprendre une grande population** sur la base des informations disponibles dans un petit échantillon.

En particulier, nous nous intéressons aux **paramètres** de population, qui sont estimés à l'aide de statistiques d'échantillonnage appropriées.

Par exemple, nous pouvons utiliser la **moyenne de l'échantillon** $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ comme estimation de la **moyenne réelle de la population** μ .

ESTIMATION

L'**estimateur** est une variable aléatoire; l'**estimation** est un nombre.

Comme autre exemple, l'**écart-type de l'échantillon** S est un estimateur de l'**écart-type de la population** réelle σ et de la valeur calculée de S

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

est une estimation de σ .

Un estimateur W de ω est sans biais si $E(W) = \omega$.

CONCEPTS MATHÉMATIQUES DE BASE

Si l'estimation $\hat{\beta}$ est sans biais, $E(\hat{\beta} - \beta) = 0$ alors un **intervalle de confiance** approximatif à 95 % (IC à 95 %) pour β est donné approximativement par

$$\hat{\beta} \pm 2\sqrt{\hat{V}(\hat{\beta})},$$

où $\hat{V}(\hat{\beta})$ est une estimation **spécifique au plan d'échantillonnage** de $V(\hat{\beta})$.

Mais qu'est-ce qu'un IC à 95 % exactement?

PROBABILITÉS CONDITIONNELLES

La **probabilité conditionnelle** est la probabilité qu'un événement se produise en fonction de l'occurrence d'un autre événement.

La probabilité conditionnelle de A étant donné B , $P(A|B)$ est définie par

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

La probabilité que deux événements A et B se produisent est obtenue en appliquant la règle de multiplication

$$P(A \cap B) = P(B) P(A|B) = P(A) P(B|A)$$

THÉORÈME DE BAYES

La règle de la somme et la règle du produit sont les **règles de base en probabilité**.

Le théorème de Bayes et la **règle de marginalisation** sont de simples corollaires de ces règles de base.

Le théorème de Bayes est parfois écrit sous une forme légèrement différente

$$P(X|Y, I) = \frac{P(Y|X, I) \times P(X|I)}{P(Y|I)}$$

THÉORÈME DE BAYES

Mise en place : Supposons qu'une expérience a été menée pour déterminer le degré de validité d'une hypothèse particulière et que des données expérimentales ont été recueillies.

Question d'analyse des données centrales : Étant donné tout ce que l'on savait *avant* l'expérience, les données recueillies appuient-elles (ou invalident-elles) l'hypothèse?

Tout au long de l'expérience, X indique que l'hypothèse en question est vraie, Y indique que l'expérience a produit les données réelles observées, et I indique (comme toujours) l'information de base pertinente.

THÉORÈME DE BAYES

Question d'analyse des données centrales (reprise) :

Quelle est la valeur de $P(\text{l'hypothèse est vraie} \mid \text{données observées}, I)$?

Problème : Cette valeur est presque toujours impossible à calculer directement.

Solution : À l'aide du théorème de Bayes,

$$P(\text{hypothèse} \mid \text{données}, I) = \frac{P(\text{données} \mid \text{hypothèse}, I) \times P(\text{hypothèse} \mid I)}{P(\text{données} \mid I)},$$

il se peut que les termes à droite soient plus faciles à calculer.

THÉORÈME DE BAYES

En termes familiers, la probabilité

- $P(\text{hypothèse} \mid I)$ que l'hypothèse soit vraie avant l'expérience est la **probabilité a priori**
- $P(\text{hypothèse} \mid \text{données}, I)$ que l'hypothèse soit vraie une fois que les données expérimentales sont prises en compte est la **probabilité a posteriori**
- $P(\text{données} \mid \text{hypothèse}, I)$ que les données expérimentales soient observées en supposant que l'hypothèse est vraie est la **vraisemblance**
- $P(\text{données} \mid I)$ que les données expérimentales soient observées indépendamment de toute hypothèse est la **donnée probante**, ou **preuve**

Une hypothèse donnée comprend un modèle (potentiellement implicite) qui peut être utilisé pour calculer ou établir approximativement la **vraisemblance**.

THÉORÈME DE BAYES

La détermination de la **probabilité a priori** est une source de controverse considérable

La **preuve** est plus difficile à calculer sur des bases théoriques. Pour évaluer la probabilité de l'observation des données, il faut un accès à un modèle dans le cadre de *I*.

THÉORÈME DE BAYES

Heureusement, les données probantes sont rarement requises sur les problèmes d'estimation des paramètres (bien qu'elles soient essentielles pour le choix du modèle) :

- avant l'expérience, il existe de nombreuses hypothèses concurrentes
- les données antérieures et les probabilités seront différentes, mais pas les données probantes
- les données probantes ne sont pas nécessaires pour différencier les différentes hypothèses

Le théorème de Bayes est souvent présenté comme suit :

$$P(\text{hypothèse} \mid \text{données}, I) \propto P(\text{données} \mid \text{hypothèse}, I) \times P(\text{hypothèse} \mid I)$$

ou simplement comme $\text{postérieur} \propto \text{probabilité} \times \text{antérieur}$, c'est-à-dire que les **croyances devraient être mises à jour en présence de nouveaux renseignements**.

ALGÈBRE LINÉAIRE

Une **matrice** est un outil mathématique important qui permet d'organiser facilement l'information, de simplifier la notation et de faciliter l'application d'algorithmes aux données.

La plupart des outils statistiques nécessitent des données **rectangulaires** :

- chaque colonne contient une **variable** (caractéristique, champ, attribut)
 - indicateur, cible, question dans une enquête, etc.
- chaque ligne contient une **observation** (cas, unité, article)
 - pays, répondant à l'enquête, sujet d'une expérience, etc.
- chaque cellule contient une **valeur** (mesure) pour une variable et une observation particulières
 - PIB par habitant pour le Canada, réponse à une question précise, âge, etc.

OPÉRATIONS MATRICIELLES

Une matrice est une grille rectangulaire d'**éléments** disposés en **lignes** et en **colonnes**.

Les matrices sont souvent utilisées en algèbre pour résoudre des valeurs inconnues dans les équations linéaires, ainsi qu'en géométrie.

Vous devriez savoir comment effectuer

- l'addition matricielle, la multiplication par un scalaire, la transposition
- le produit de matrices
- l'inversion de matrice, le calcul du déterminant et de la trace d'une matrice carrée

VECTEURS PROPRES ET VALEURS PROPRES

Un **vecteur propre** d'une matrice A est un vecteur $\mathbf{v} \neq \mathbf{0}$ de sorte que, pour certains scalaires λ , $A\mathbf{v} = \lambda\mathbf{v}$.

La valeur λ s'appelle une **valeur propre** de A associée à \mathbf{v} .

Les valeurs propres d'une matrice $n \times n$ A répondent à $\det(A - \lambda I_n) = 0$. Le côté gauche est un polynôme dans λ ; il est appelé **polynôme caractéristique** de A , représenté par $p_A(\lambda)$.

Pour trouver les valeurs propres de A , nous trouvons les racines de $p_A(\lambda)$.


DÉCOMPOSITION EN ÉLÉMENTS PROPRES

Si une matrice A $n \times n$ possède des vecteurs propres n linéairement indépendants, alors A peut être **décomposé** de la manière suivante :

$$A = B\Lambda B^{-1},$$

où Λ est une matrice diagonale dont les entrées diagonales sont les valeurs propres de A et les colonnes de B sont les vecteurs propres correspondants de A .

MODÉLISATION PAR RÉGRESSION

Structure de données d'une tâche de modélisation générale est représenté par 

Nous tenons compte des variables p indépendantes X_i pour essayer de prédire la variable dépendante Y .

X_1	X_2	\cdots	X_p	Y
x_{11}	x_{12}	\cdots	x_{1p}	y_1
x_{21}	x_{22}	\cdots	x_{2p}	y_2
\cdots	\cdots	\cdots	\cdots	\cdots
x_{n1}	x_{n2}	\cdots	x_{np}	y_n

Afin de simplifier l'analyse qui suit, nous présentons la notation matricielle $\mathbf{X}[n \times p]$, $\mathbf{Y}[n \times 1]$, $\boldsymbol{\beta}[p \times 1]$, où n est le nombre d'observations et p est le nombre de variables indépendantes.

RÉGRESSION LINÉAIRE

L'hypothèse de base de la régression linéaire est que la variable dépendante y peut être **approximée** par une combinaison linéaire des variables indépendantes comme suit :

$$Y = X\beta + \varepsilon,$$

où $\beta \in \mathbb{R}^p$ doit être déterminé en fonction de l'ensemble d'apprentissage, et pour lequel

$$E(\varepsilon|X) = 0, \quad E(\varepsilon\varepsilon^T|X) = \sigma^2 I.$$

En règle générale, les erreurs sont également supposées être normalement distribuées, c'est-à-dire :

$$\varepsilon|X \sim N(0, \sigma^2 I).$$

RÉGRESSION LINÉAIRE

Si $\hat{\beta}_i$ est l'estimation du coefficient β_i réel, le modèle de **régression linéaire** associé aux données est le suivant

$$\hat{Y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

Sous forme matricielle, le problème de régression nécessite une solution $\hat{\beta}$ à l'**équation normale** $X^T X \beta = X^T Y$.

Lorsque la matrice symétrique positive définie $X^T X$ est inversable, le coefficient rajusté est simplement $\hat{\beta} = (X^T X)^{-1} (X^T Y)$. Notez que $X^T X$ est une matrice $p \times p$, ce qui rend l'inversion relativement « plus facile » à calculer, lorsque n a une valeur élevé.

RÉGRESSION LINÉAIRE GÉNÉRALISÉE

Les modèles linéaires généralisés (MLG) accroissent la portée des modèles statistiques linéaires en acceptant les variables de réponse ayant une distribution conditionnelle **non normale**.

Sauf pour la **structure d'erreur**, un MLG est essentiellement le même que pour un modèle linéaire :

$$Y_i \sim \text{une certaine distribution avec moyenne } \mu_i, \text{ où } g(\mu_i) = x_i^T \beta$$

Ainsi, un MLG compte trois parties :

- une composante **systematique** $x_i^T \beta$
- une composante **aléatoire** - distribution spécifiée pour Y_i
- une fonction de **liaison** g

OPTIMISATION

Supposons que nous devions **optimiser** une fonction **coût** $f: \mathbb{R}^n \rightarrow \mathbb{R}$ (économique) (la fonction de vraisemblance maximale de régression linéaire, par exemple).

La recherche d'un maximum pour f équivaut à la recherche d'un minimum pour $-f$.

L'objectif consiste à trouver les valeurs des paramètres \mathbf{x} qui minimisent cette fonction :

$$\mathbf{x}^* = \arg \min_{\mathbf{x}} f(\mathbf{x})$$

La fonction coût pourrait être soumise à un certain nombre de contraintes

$$c_i(\mathbf{x}) = 0, i = 1, \dots, m; c_j(\mathbf{x}) \geq 0, j = 1, \dots, k; \mathbf{x} \in \Omega \subseteq \mathbb{R}^n.$$

OPTIMISATION

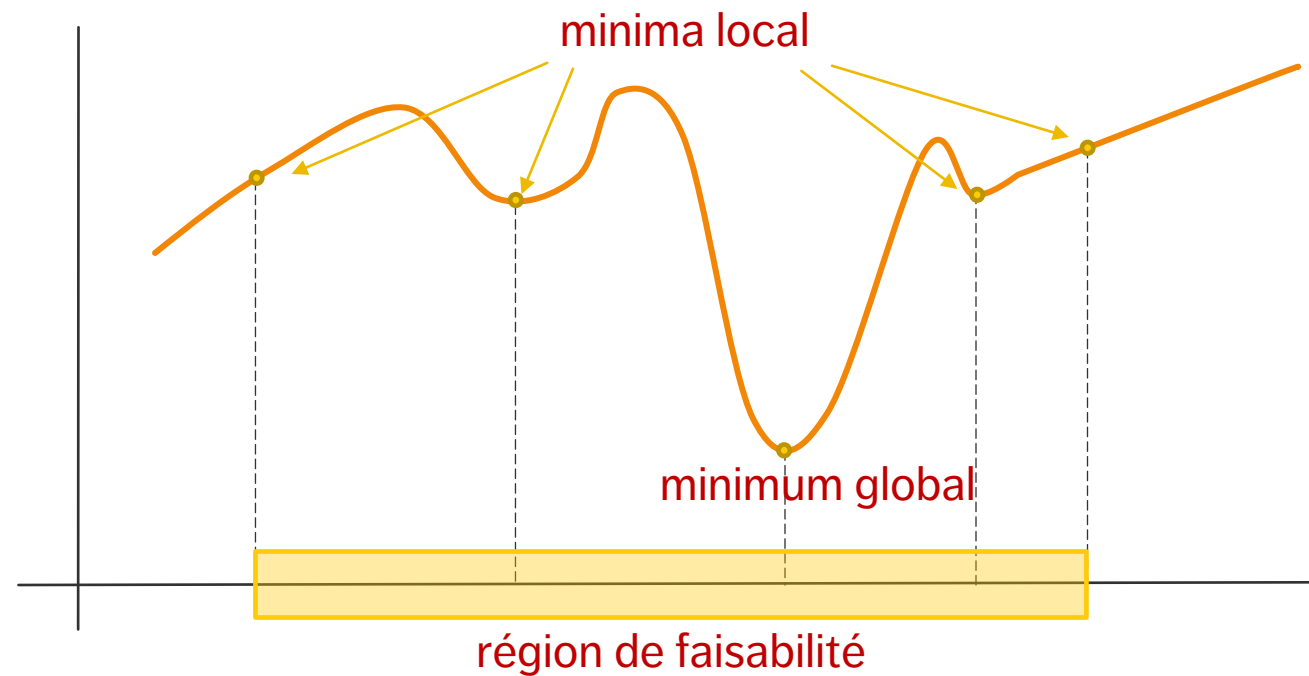
Le problème d'optimisation peut être considéré comme un **problème de décision** qui consiste à trouver le « meilleur » vecteur \mathbf{x} parmi tous les vecteurs possibles dans $\Omega \subseteq \mathbb{R}^n$.

Ce vecteur est appelé le **minimiseur** de f parmi Ω . Il peut y avoir plusieurs minimiseurs, ou aucun.

Si $\Omega = \mathbb{R}^n$, alors nous désignons le problème comme un problème d'optimisation **sans contrainte**.

En général, il ne s'agit pas d'un problème banal (consulter la littérature).

TYPE DE MINIMA



Dans de nombreux cas, l'optimisation est une entreprise **numérique**. Le minima trouvé dépend du **point de départ** de l'algorithme.