

Statistiques mathématiques : cours 8

Guillaume Lécué

28 septembre 2017

Aujourd'hui : Mise en oeuvre des méthodes statistiques des cours précédants dans le modèle de régression

Présentation des modèles de régression

Méthodes d'estimation en régression

Tests et sélection de variables

Données : publicités et ventes d'un même produit sur 200 marchés

fichier Advertising.csv

id-market	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
...
200	232.1	8.6	8.7	13.4

Questions :

1. Quelle est l'influence des campagnes "TV" sur les "Sales" ?
2. Etant donné un budget publicité, où faut-il investir ? et combien de "Sales" peut-on espérer en retirer ?

Présentation des modèles de régression

Expliquer une variable Y par une autre X

Principe : on part de l'observation de n couples

$$(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n) \text{ où } Y_i \in \mathbb{R} \text{ et } \mathbf{X}_i \in \mathbb{R}^k$$

Exemple : sur le i -ième marché,

- ▶ $Y_i = \text{"Sales"}$
- ▶ $X_i = (\text{"TV"}, \text{"Radio"}, \text{"Newspaper"}) \in \mathbb{R}^3$

Idée : On **pense** que \mathbf{X}_i peut **expliquer** la " majeure partie de la variabilité des Y_i " ; càd que Y_i est " presque " fonction de \mathbf{X}_i (à quelque chose près).

Modélisation de "l'influence"

- ▶ Si \mathbf{X}_i contient toute la variabilité de Y_i , alors Y_i est fonction de \mathbf{X}_i : il existe $r : \mathbb{R}^k \rightarrow \mathbb{R}$ telle que

$$Y_i = r(\mathbf{X}_i)$$

mais peu réaliste (ou alors problème d'interpolation numérique).

- ▶ Alternative : on modélise ces données avec le modèle

$$Y_i = r(\mathbf{X}_i) + \xi_i$$

où ξ_i est un terme aléatoire qui explique le reste de la variabilité de Y_i et $r(\cdot)$ une fonction qu'on va estimer. On suppose que $\mathbb{E} \xi_i = 0$ (pour l'identifiabilité).

prédiction et influence des features

Dans le modèle

$$Y_i = r(\mathbf{X}_i) + \xi_i$$

pour $\mathbf{X}_i \in \mathbb{R}^k$, les coordonnées des \mathbf{X}_i sont appelées les **features**

Exemple : "TV", "Radio" et "Newspaper" sont les features du problème.

- ▶ Si $\hat{r}(\cdot)$ est un estimateur de $r(\cdot)$ alors la variabilité de $\hat{r}(\cdot)$ en la j -ième coordonnée ($1 \leq j \leq k$) mesure l'**influence de la feature j sur la variable à expliquer Y**
- ▶ Si $x \in \mathbb{R}^k$ alors $\hat{y} = \hat{r}(x)$ **prédit** la valeur de la variable expliquée associée à x .

Motivation : meilleure approximation L^2

- ▶ Meilleure approximation L^2 : si $\mathbb{E}[Y^2] < +\infty$, la meilleure approximation de Y par une variable aléatoire \mathbf{X} -mesurable est donnée par l'**espérance conditionnelle** $\mathbb{E}[Y|\mathbf{X}]$:

$$\mathbb{E}[(Y - r(\mathbf{X}))^2] = \min_h \mathbb{E}[(Y - h(\mathbf{X}))^2]$$

où

$$r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}], \quad \mathbf{x} \in \mathbb{R}^k$$

- ▶ On appelle $r(\cdot)$ **fonction de régression de Y sachant \mathbf{X}** .

Régression

- ▶ On définit :

$$\xi = Y - \mathbb{E}[Y|\mathbf{X}] \implies \mathbb{E}[\xi] = 0$$

- ▶ On a alors naturellement la représentation désirée

$$Y = r(\mathbf{X}) + \xi, \quad \mathbb{E}[\xi] = 0$$

en posant

$$r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}], \quad \mathbf{x} \in \mathbb{R}^k$$

- ▶ On observe alors n couples

$$(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$$

où

$$Y_i = r(\mathbf{X}_i) + \xi_i, \quad \mathbb{E}[\xi_i] = 0$$

avec comme paramètre la fonction de régression $r(\cdot)$ + un jeu d'hypothèses sur la loi des ξ_i .

Modèle de régression à design aléatoire

Définition

Modèle de régression paramétrique à design aléatoire = observation d'un n -échantillon de couples

$$(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$$

avec $(\mathbf{X}_i, Y_i) \in \mathbb{R}^k \times \mathbb{R}$ *i.i.d.* $\sim (\mathbf{X}, Y)$, et

$$Y = r(\theta, \mathbf{X}) + \xi, \quad \mathbb{E}[\xi | \mathbf{X}] = 0, \quad \theta \in \Theta \subset \mathbb{R}^d.$$

- ▶ $\mathbf{x} \mapsto r(\theta, \mathbf{x})$ *fonction de régression* de Y sachant \mathbf{X} (inconnue, car θ est inconnu : paramètre du modèle)
- ▶ \mathbf{X}_i : *variables explicatives, co-variables, input*
- ▶ $(\mathbf{X}_1, \dots, \mathbf{X}_n)$: *design*
- ▶ Y_i : *variables expliquées, output*

Régression à design déterministe

- ▶ Principe : sur un exemple. On observe

$$Y_i = r(\theta, i/n) + \xi_i, \quad i = 1, \dots, n$$

où $r(\theta, \cdot) : [0, 1] \rightarrow \mathbb{R}$ est une fonction connue au paramètre $\theta \in \Theta \subset \mathbb{R}^d$ près, et les ξ_i sont i.i.d., $\mathbb{E} [\xi_i] = 0$.

- ▶ But : reconstruire $r(\theta, \cdot)$ c'est-à-dire estimer θ .
- ▶ Plus généralement, on observe $(Y_i)_{i=1}^n$ où

$$Y_i = r(\theta, \mathbf{x}_i) + \xi_i, \quad i = 1, \dots, n$$

et $\mathbf{x}_1, \dots, \mathbf{x}_n$ sont des points de \mathbb{R}^k **déterministes**.

Modèle de régression à design déterministe

Définition

Modèle de régression à **design déterministe** = donnée de l'observation

$$(\mathbf{x}_1, Y_1), \dots, (\mathbf{x}_n, Y_n) \text{ ou plus simplement } Y_1, \dots, Y_n$$

avec $Y_i \in \mathbb{R}$, $\mathbf{x}_i \in \mathbb{R}^k$, et

$$Y_i = r(\boldsymbol{\theta}, \mathbf{x}_i) + \xi_i, \quad \mathbb{E}[\xi_i] = 0, \quad \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d.$$

- ▶ \mathbf{x}_i déterministes, donnés (ou choisis) : plan d'expérience, points du "design".
- ▶ Hypothèses sur les ξ_i : par exemple : i.i.d., gaussien, etc.
- ▶ **Attention !** Les Y_i ne sont *pas identiquement distribués*.

Régression linéaire

On parle de **modèle de régression linéaire** quand la fonction de régression $r(\theta, \cdot)$ est supposée linéaire : pour tout $x \in \mathbb{R}^d$

$$r(\theta, x) = \langle \theta, x \rangle$$

On a alors pour les modèles :

- ▶ $Y_i = \langle \theta, \mathbf{X}_i \rangle + \zeta_i$: modèle linéaire à design aléatoire,
- ▶ $Y_i = \langle \theta, \mathbf{x}_i \rangle + \zeta_i$: modèle linéaire à design déterministe,

et pour un bruit gaussien : $g_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$,

- ▶ $Y_i = \langle \theta, \mathbf{X}_i \rangle + \sigma g_i$: modèle linéaire gaussien à design aléatoire (on suppose de plus que les g_i sont indépendants des \mathbf{X}_i),
- ▶ $Y_i = \langle \theta, \mathbf{x}_i \rangle + \sigma g_i$: modèle linéaire gaussien à design déterministe,

Méthodes d'estimation en régression à design déterministe et bruit gaussien

Modèle de régression gaussienne à design déterministe :

$$Y_i = r(\theta, \mathbf{x}_i) + \sigma g_i, \quad \theta \in \Theta \subset \mathbb{R}^d$$

où $g_i \sim \mathcal{N}(0, 1)$, i.i.d..

Problème : estimer θ ?

Idée : Expliciter la loi de l'observation $Z = (Y_1, \dots, Y_n)$ et appliquer le principe du maximum de vraisemblance.

La loi de Y_i : $\mathbb{P}_{Y_i} = f_{\mathbf{x}_i}(\theta, \cdot) \cdot \lambda$ où $\forall y \in \mathbb{R}$

$$f_{\mathbf{x}_i}(\theta, y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - r(\theta, \mathbf{x}_i))^2\right)$$

Loi de (Y_1, \dots, Y_n) : $\mathbb{P}_{(Y_1, \dots, Y_n)} = f(\theta, \cdot) \cdot \lambda^n$ où

$$f(\theta, (y_1, \dots, y_n)) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - r(\theta, \mathbf{x}_i))^2\right)$$

On travaille alors dans le modèle $\{\mathbb{P}_\theta^n = \mathbb{P}_{(Y_1, \dots, Y_n)} : \theta \in \mathbb{R}^d\}$, dominé par $\mu = \lambda^n$, ayant pour densités

$$\begin{aligned} \frac{d\mathbb{P}_\theta^n}{d\mu}(y_1, \dots, y_n) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - r(\theta, \mathbf{x}_i))^2\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - r(\theta, \mathbf{x}_i))^2\right) := f(\theta, (y_i)_{i=1}^n) \end{aligned}$$

La fonction de vraisemblance vaut en $\theta \in \mathbb{R}^d$,

$$\mathcal{L}_n(\theta, Y_1, \dots, Y_n) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - r(\theta, \mathbf{x}_i))^2\right)$$

Estimateur des moindres carrés

Maximiser la **vraisemblance** en régression gaussienne



Minimiser la somme des carrés : trouver les $\theta \in \mathbb{R}^d$ minimisant

$$\theta \in \mathbb{R}^d \longrightarrow \sum_{i=1}^n (Y_i - r(\theta, \mathbf{x}_i))^2$$

Définition

Estimateur des moindres carrés (EMC) : tout estimateur $\hat{\theta}_n^{\text{mc}}$ tel que
 $\hat{\theta}_n^{\text{mc}} \in \arg \min_{\theta \in \mathbb{R}^k} \sum_{i=1}^n (Y_i - r(\theta, \mathbf{x}_i))^2$

En régression Gaussienne : **EMV = EMC**

Droite de régression ($k = 1$)

Modèle le plus simple : on suppose que la fonction de régression est une fonction affine de la forme

$$r(\theta, x) = a + bx$$

alors le modèle de régression à design déterministe s'écrit ici :

$$Y_i = a + bx_i + \xi_i, \quad i = 1, \dots, n$$

où les x_1, \dots, x_n sont des réels donnés et ξ_1, \dots, ξ_n sont i.i.d. centrées et de variances finies.

- ▶ on paramétrise par $\theta = (a, b)^T \in \Theta = \mathbb{R}^2$;
a est appelé l'intercept.
- ▶ L'estimateur des moindres carrés :

$$\hat{\theta}_n^{\text{mc}} = \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \arg \min_{(a,b)^T \in \mathbb{R}^2} \sum_{i=1}^n (Y_i - a - bx_i)^2$$

Estimateur des moindres carrés (1/2)

On peut réécrire la **fonction objectif** sous forme matricielle :

$$F(a, b) = \sum_{i=1}^n (Y_i - a - bx_i)^2 = \left\| \mathbb{Y} - \mathbb{X} \begin{pmatrix} a \\ b \end{pmatrix} \right\|_2^2$$

où

$$\mathbb{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \text{ et } \mathbb{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

et comme

$$\nabla F(a, b) = -2\mathbb{X}^\top (\mathbb{Y} - \mathbb{X}(a, b)^\top) \text{ et } \nabla^2 F(a, b) = 2\mathbb{X}^\top \mathbb{X} \succeq 0$$

l' (ou les) EMC est (sont) solution(s) de

$$\mathbb{X}^\top \mathbb{X} \hat{\theta}_n^{\text{mc}} = \mathbb{X}^\top \mathbb{Y}$$

Estimateur des moindres carrés (2/2)

- ▶ Unique solution quand $\mathbb{X}^T \mathbb{X}$ est inversible :

$$\hat{\theta}_n^{\text{mc}} = \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{Y}$$

- ▶ Résidu : si $\hat{\theta}_n$ est un estimateur de θ alors $\hat{y}_i = r(\hat{\theta}_n, x_i)$ est la valeur prédite par l'estimateur au point x_i et

$$Y_i - \hat{y}_i : \text{résidu au point } i$$

- ▶ RSS : (Residual Sum of Squares)

$$RSS := \sum_{i=1}^n (Y_i - \hat{y}_i)^2$$

Régression linéaire simple sur les données Advertising.csv

http://localhost:8888/notebooks/linear_regression.ipynb

Régression linéaire multiple (=Modèle linéaire)

La fonction de régression est $r(\theta, \mathbf{x}_i) = \langle \theta, \mathbf{x}_i \rangle$. On observe

$$(\mathbf{x}_1, Y_1), \dots, (\mathbf{x}_n, Y_n)$$

sous le modèle

$$Y_i = \langle \theta, \mathbf{x}_i \rangle + \xi_i, \quad i = 1, \dots, n$$

où $\theta \in \Theta = \mathbb{R}^k$, $\mathbf{x}_i \in \mathbb{R}^k$.

- ▶ Problème : estimer θ
- ▶ l'analyse des estimateurs pour un **design aléatoire** est un plus délicate

Écriture matricielle des données

Matriciellement, on réécrit ces données comme

$$\mathbb{Y} = \mathbb{X}\theta + \xi$$

où

$$\mathbb{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \in \mathbb{R}^n, \mathbb{X} = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} \in \mathbb{R}^{n \times k} \text{ et } \xi = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} \in \mathbb{R}^n$$

On parle de régression linéaire **avec intercept** quand

$$\mathbb{X} = \begin{pmatrix} 1 & \mathbf{x}_1^\top \\ \vdots & \vdots \\ 1 & \mathbf{x}_n^\top \end{pmatrix} \in \mathbb{R}^{n \times (k+1)}$$

EMC en régression linéaire multiple

- ▶ Estimateur des **moindres carrés** en régression linéaire multiple : tout estimateur $\hat{\theta}_n^{\text{mc}}$ minimisant

$$\theta \in \mathbb{R}^k \mapsto F(\theta) := \min_{\theta \in \mathbb{R}^k} \sum_{i=1}^n (Y_i - \langle \theta, \mathbf{x}_i \rangle)^2$$

- ▶ En notation matricielle :

$$\|\mathbb{Y} - \mathbb{X}\hat{\theta}_n^{\text{mc}}\|^2 = \min_{\theta \in \mathbb{R}^k} \|\mathbb{Y} - \mathbb{X}\theta\|^2 = \min_{v \in V} \|\mathbb{Y} - v\|^2$$

où $V = \text{Im}(\mathbb{X}) = \{v \in \mathbb{R}^n : v = \mathbb{X}\theta, \theta \in \mathbb{R}^k\}$. Donc $\mathbb{X}\hat{\theta}_n^{\text{mc}}$ est la projection orthogonale de \mathbb{Y} sur V .

Géométrie de l'EMC

- ▶ L'EMC vérifie

$$\mathbb{X} \hat{\theta}_n^{\text{mc}} = P_V \mathbb{Y}$$

où P_V est le projecteur orthogonal sur V .

- ▶ Mais $\mathbb{X}^T P_V = \mathbb{X}^T P_V^T = (P_V \mathbb{X})^T = \mathbb{X}^T$. On en déduit **les équations normales des moindres carrés** :

$$\mathbb{X}^T \mathbb{X} \hat{\theta}_n^{\text{mc}} = \mathbb{X}^T \mathbb{Y} \quad (1)$$

- ▶ Remarques.

- ▶ L'EMC est un Z -estimateur (bonnes propriétés quand (1) a une unique solution c-à-d $\mathbb{X}^T \mathbb{X} \succ 0$).
- ▶ Pas d'**unicité** de $\hat{\theta}_n^{\text{mc}}$ si la matrice $\mathbb{X}^T \mathbb{X}$ n'est pas inversible.
- ▶ (1) est équivalente à $\nabla F(\hat{\theta}_n^{\text{mc}}) = 0$

Proposition

Si $\mathbb{X}^T \mathbb{X}$ (matrice $k \times k$) est inversible, alors $\hat{\theta}_n^{\text{mc}}$ est unique et

$$\hat{\theta}_n^{\text{mc}} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{Y}$$

- ▶ Contient le cas précédent de la droite de régression simple.
- ▶ Résultat géométrique, **non stochastique**.
- ▶ on a toujours $\mathbb{X}^T \mathbb{X} \succeq 0$; de plus :

$$\mathbb{X}^T \mathbb{X} \text{ inversible} \Leftrightarrow \mathbb{X}^T \mathbb{X} \succ 0 \Leftrightarrow \text{rang}(\mathbb{X}) = k \Leftrightarrow \dim(V) = k$$

En particulier, $\mathbb{X}^T \mathbb{X} \succ 0 \implies n \geq k$ (statistiques en petites dimensions)

Régression linéaire multiple sur les données Advertising.csv

http://localhost:8888/notebooks/linear_regression.ipynb

Régression linéaire gaussienne = Modèle linéaire gaussien

On suppose que le vecteur bruit est tel que

$$\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \text{Id}_n)$$

dans le modèle (sous forme matricielle)

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\xi}$$

On a alors plusieurs propriétés remarquables :

- ▶ l'EMC $\hat{\boldsymbol{\theta}}_n^{\text{mc}} = \text{EMV}$ (dans le modèle à variance connue)
- ▶ On sait expliciter la loi (non-asymptotique!) de $\hat{\boldsymbol{\theta}}_n^{\text{mc}}$

Cadre gaussien : loi des estimateurs

- ▶ Hyp. 1 : $\xi \sim \mathcal{N}(0, \sigma^2 \text{Id}_n)$
- ▶ Hyp. 2 : $\mathbb{X}^\top \mathbb{X} \succ 0$

Proposition (2)

- (i) $\hat{\theta}_n^{\text{mc}} \sim \mathcal{N}(\theta, \sigma^2 (\mathbb{X}^\top \mathbb{X})^{-1})$
- (ii) $\|\mathbb{Y} - \mathbb{X} \hat{\theta}_n^{\text{mc}}\|_2^2 \sim \sigma^2 \chi^2(n - k)$
- (iii) $\hat{\theta}_n^{\text{mc}}$ et $\mathbb{Y} - \mathbb{X} \hat{\theta}_n^{\text{mc}}$ sont indépendants

Preuve : **Thm. de Cochran** : Si $\xi \sim \mathcal{N}(0, \text{Id}_n)$ et P_j matrices $n \times n$ de projection t.q. $P_j P_i = 0$ pour $i \neq j$, alors :

1. $P_j \xi \sim \mathcal{N}(0, P_j)$ sont **indépendants**,
2. $\|P_j \xi\|_2^2 \sim \chi^2(\text{Rang}(P_j))$

Preuve de la proposition 2 (directe, sans Cochran)

(i) $\hat{\theta}_n^{\text{mc}} = \theta + (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \boldsymbol{\xi}$ est une transformation affine d'un vecteur Gaussien donc $\hat{\theta}_n^{\text{mc}}$ est aussi un vecteur Gaussien ; sa moyenne et matrice de covariance sont :

1. $\mathbb{E}[\hat{\theta}_n^{\text{mc}}] = \theta$

2. $\text{Cov}(\hat{\theta}_n^{\text{mc}}) = \mathbb{E} [(\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \boldsymbol{\xi} ((\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \boldsymbol{\xi})^\top] = \sigma^2 (\mathbb{X}^\top \mathbb{X})^{-1}$

(ii) pour $P_V = \mathbb{X}(\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top$: matrice de projection sur $V = \text{Im}(\mathbb{X})$ et $\boldsymbol{\xi}' = \sigma^{-1} \boldsymbol{\xi} \sim \mathcal{N}(0, \text{Id}_n)$

$$\begin{aligned} \mathbb{Y} - \mathbb{X} \hat{\theta}_n^{\text{mc}} &= \mathbb{X}(\theta - \hat{\theta}_n^{\text{mc}}) + \boldsymbol{\xi} \\ &= -\mathbb{X}(\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \boldsymbol{\xi} + \boldsymbol{\xi} = \sigma(\text{Id}_n - P_V) \boldsymbol{\xi}' \end{aligned}$$

(iii) le vecteur $(\hat{\theta}_n^{\text{mc}}, \mathbb{Y} - \mathbb{X} \hat{\theta}_n^{\text{mc}})$ est gaussien (transformation linéaire de $\boldsymbol{\xi}$). On calcule sa matrice de covariance.

Modèle linéaire Gaussien – variance inconnue

Dans le modèle linéaire Gaussien

$$Y = X\theta + \sigma\mathcal{N}(0, I_n)$$

où θ et σ sont inconnus on a :

$$\text{EMV} = \begin{pmatrix} \hat{\theta}_n^{\text{mc}} \\ \hat{\sigma}_n^2 \end{pmatrix} \text{ où } \hat{\sigma}_n^2 = \frac{\|Y - X\hat{\theta}_n^{\text{mc}}\|_2^2}{n}$$

car la log-vraisemblance

$$\ell_n(\theta, \sigma^2) = \frac{-n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|Y - X\theta\|_2^2$$

est maximale en ce point

Propriétés de l'EMV : cadre gaussien variance inconnue (1/2)

$$\text{EMV} = \begin{pmatrix} \hat{\theta}_n^{\text{mc}} \\ \hat{\sigma}_n^2 \end{pmatrix}$$

où

$$\hat{\theta}_n^{\text{mc}} = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbb{Y} \text{ et } \hat{\sigma}_n^2 = \frac{\|\mathbb{Y} - \mathbb{X} \hat{\theta}_n^{\text{mc}}\|_2^2}{n}$$

D'après Proposition 2 :

- ▶ $\hat{\sigma}_n^2$ est **indépendant** de $\hat{\theta}_n^{\text{mc}}$
- ▶ $\hat{\theta}_n^{\text{mc}} \sim \mathcal{N}(\theta, \sigma^2 (\mathbb{X}^\top \mathbb{X})^{-1})$
- ▶ $n\hat{\sigma}_n^2/\sigma^2 \sim \chi^2(n - k)$

Propriétés de l'EMV : cadre gaussien variance inconnue (2/2)

Lois des coordonnées de $\hat{\theta}_n^{\text{mc}}$:

$$(\hat{\theta}_n^{\text{mc}})_j - \theta_j \sim \mathcal{N}(0, \sigma^2 b_j)$$

où b_j est le j ème élément diagonal de $(\mathbb{X}^\top \mathbb{X})^{-1}$ et

$$\frac{(\hat{\theta}_n^{\text{mc}})_j - \theta_j}{\tilde{\sigma}_n \sqrt{b_j}} \sim t_{n-k} \text{ pour } \tilde{\sigma}_n = \frac{\|\mathbb{Y} - \mathbb{X} \hat{\theta}_n^{\text{mc}}\|_2^2}{n - k}$$

Définition

La *loi de Student à $n - k$ degrés de liberté* est la loi de

$$t_{n-k} = \frac{g}{\sqrt{\eta/(n-k)}}$$

où $g \sim \mathcal{N}(0, 1)$, $\eta \sim \chi^2(n - k)$ et g indépendant de η .

Tests et sélection de variables dans le modèle linéaire Gaussien

Features selection = Sélection de variables

Problème : On cherche à expliquer une variable $Y \in \mathbb{R}$ en fonction d'une autre variable $X \in \mathbb{R}^k$. Certaines coordonnées de X n'ont peut-être aucun intérêt pour ce problème (elles n'expliquent en rien la variabilité de Y).

Exemple : peut-être que la variable "Newspaper" n'explique en rien "Sales" (?)

Problème : on ne veut garder que les variables pertinentes, c'est le problème de **features selection**

Features selection via backward elimination

1. On retire la j -ième feature (= on retire la j -ième colonne de $\mathbb{X} \rightarrow \mathbb{X}_{-j}$) et on construit $\hat{\theta}_n^{\text{mc}}(-j)$ à partir de \mathbb{Y} et \mathbb{X}_{-j}
2. on choisi j_1 pour lequel

$$RSS(\hat{\theta}_n^{\text{mc}}(-j_1)) = \min_{1 \leq j \leq k} RSS(\hat{\theta}_n^{\text{mc}}(-j)) := RSS_{k-1}$$

3. on réitère jusqu'à la stabilisation de RSS :

$$RSS_m \approx RSS_{m-1}$$

4. à la fin, seules les colonnes restantes de \mathbb{X} sont des features pertinentes : ceux sont celles qui expliquent le plus la variabilité de Y

Autres idées : Forward procédures, critères AIC et BIC, LASSO, tests, etc.

Feature selection via test (1/2)

Cadre : **Modèle linéaire gaussien** (à design déterministe)

$$Y = X\theta + \xi, \quad \xi \sim \mathcal{N}(0, \sigma^2 \text{Id}_n),$$

où $\theta = (\theta_1, \dots, \theta_k)^T \in \mathbb{R}^k$, $X \in \mathbb{R}^{n \times k}$ et $X^T X \succ 0$.

Problème de test : $a \in \mathbb{R}$, $j \in \{1, \dots, k\}$ donné

$$H_0 : \theta_j = a \text{ contre } H_1 : \theta_j \neq a$$

On a vu que, sous \mathbb{P}_θ ,

$$\frac{(\hat{\theta}_n^{\text{mc}})_j - \theta_j}{\tilde{\sigma}_n \sqrt{(X^T X)^{-1}_{jj}}} \stackrel{d}{=} \text{Student}(n - k) \text{ où } \tilde{\sigma}_n = \frac{\|Y - X\hat{\theta}_n^{\text{mc}}\|_2}{n - k}$$

Feature selection via test (2/2)

On peut alors construire un test de niveau α par :

$$\varphi_\alpha = \begin{cases} H_0 & \text{quand } t_n \leq q_{1-\alpha/2}^{Student(n-k)} \\ H_1 & \text{sinon} \end{cases}$$

pour la t-statistique (de la feature j)

$$t_n := \frac{|(\hat{\theta}_n^{\text{mc}})_j - a|}{\tilde{\sigma}_n \sqrt{(\mathbb{X}^\top \mathbb{X})_{jj}^{-1}}}$$

En particulier, pour $a = 0$, on test si le coefficient associé à la j -ième feature est nul. Si on rejete le test (petite p-value), alors cette feature sera sélectionnée (avec un niveau de confiance de $1 - \alpha$ ou $\alpha = p - \text{value}$). On répète la procédure de test pour les k features : pour chaque feature, on calcul sa t-statistique et la p-value associée

Sélection de groupes de variables

Cadre : modèle linéaire Gaussien (à design déterministe) et paramètre $\theta \in \mathbb{R}^k$

Problème de test : $1 \leq k_0 < k$ fixé. On souhaite savoir si au moins une des $k - k_0$ dernières features a une influence.

On choisit alors les hypothèses :

$$H_0 : \theta_\ell = 0, \quad \forall \ell = k_0, \dots, k$$

contre

$$H_1 : \text{il existe } \ell \in \{k_0, \dots, k\} \text{ t.q. } \theta_\ell \neq 0$$

(choix des hypothèses tel que le rejet répond à la question : "rejet" = "oui il y a au moins une feature influente")

Formulation plus générale du problème : F-tests

Soit $\mathbb{G} \in \mathbb{R}^{m \times k}$ et $\mathbf{b} \in \mathbb{R}^m$ donné. On considère le problème de test :

$$H_0 : \mathbb{G}\theta = \mathbf{b}$$

contre

$$H_1 : \mathbb{G}\theta \neq \mathbf{b}$$

Ici : on prend

$$\mathbb{G} = \begin{pmatrix} 0 & \dots & 0 & 1 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 1 \end{pmatrix} \in \mathbb{R}^{k_0 \times k} \text{ et } \mathbf{b} = \mathbf{0} \in \mathbb{R}^{k_0}$$

F-tests (1/2)

Sous H_0 (càd pour θ t.q. $\mathbb{G}\theta = \mathbf{b}$) on a (cf. Proposition 2)

$$\mathbb{G}\hat{\theta}_n^{\text{mc}} \sim \mathcal{N}(\mathbf{b}, \sigma^2 \mathbb{G}(\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{G}^\top)$$

et donc en posant $\mathbf{U} = \sigma^2 \mathbb{G}(\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{G}^\top$ (et si \mathbf{U} est inversible), on a

$$(\mathbb{G}\hat{\theta}_n^{\text{mc}} - \mathbf{b})^\top \mathbf{U}^{-1} (\mathbb{G}\hat{\theta}_n^{\text{mc}} - \mathbf{b}) \sim \chi^2(m)$$

Si σ^2 est inconnue, on pose $\tilde{\sigma}_n^2 = \frac{\|\mathbf{Y} - \mathbb{X}\hat{\theta}_n^{\text{mc}}\|_2^2}{n-k}$ et $\hat{\mathbf{U}} = \tilde{\sigma}_n^2 \mathbb{G}(\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{G}^\top$, alors, la loi de

$$\frac{(\mathbb{G}\hat{\theta}_n^{\text{mc}} - \mathbf{b})^\top \hat{\mathbf{U}}^{-1} (\mathbb{G}\hat{\theta}_n^{\text{mc}} - \mathbf{b})}{m}$$

ne dépend pas de θ ni de σ^2 sous H_0 et suit la loi de Fisher-Snedecor à $(m, n - k)$ degrés de liberté.

F-tests (2/2)

Définition

Si $X \sim \chi^2(m)$, $Y \sim \chi^2(n - k)$ et X est indépendante de Y alors

$$\frac{X/m}{Y/(n - k)} \sim \text{Fisher - Snedecor}(m, n - k) := F(m, n - k)$$

On a alors un **test de niveau α** pour le problème de test

$$H_0 : \mathbb{G}\theta = \mathbf{b} \text{ contre } H_1 : \mathbb{G}\theta \neq \mathbf{b}$$

donné par

$$\varphi_\alpha = \begin{cases} H_0 & \text{si } T_n \leq q_{1-\alpha}^{F(m, n-k)} \\ H_1 & \text{sinon} \end{cases}$$

où

$$T_n = \frac{(\mathbb{G}\hat{\theta}_n^{\text{mc}} - \mathbf{b})^T \hat{\mathbf{U}}^{-1} (\mathbb{G}\hat{\theta}_n^{\text{mc}} - \mathbf{b})}{m} \text{ et } \hat{\mathbf{U}} = \tilde{\sigma}_n^2 \mathbb{G}(\mathbf{X}^T \mathbf{X})^{-1} \mathbb{G}^T$$

Information de Fisher dans le modèle linéaire Gaussien

Information de Fisher et régression (1/3)

Cadre : \mathcal{E}^n expérience engendrée par $(\mathbf{x}_1, Y_1), \dots, (\mathbf{x}_n, Y_n)$ avec

$$Y_i = r(\theta, \mathbf{x}_i) + \xi_i,$$

où les ξ_i sont i.i.d. admettant une densité g par rapport à la mesure de Lebesgue et $\mathbf{x}_1, \dots, \mathbf{x}_n$ sont déterministes.

Observation : $Z^n = (Y_1, \dots, Y_n)$ de densité (par rapport à Lebesgue sur \mathbb{R}^n)

$$f_n(\theta, Z^n) = \prod_{i=1}^n g(Y_i - r(\theta, \mathbf{x}_i))$$

Information de Fisher :

$$\mathbb{I}(\theta | \mathcal{E}^n) = - \mathbb{E}_\theta [\nabla_\theta^2 \log f_n(\theta, Z^n)]$$

Information de Fisher et régression (2/3)

Quand le bruit est Gaussien :

$$g(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-t^2}{2\sigma^2}\right)$$

et donc, pour le problème d'estimation de θ à σ connue, on a

$$\mathbb{I}(\theta|\mathcal{E}^n) = \sigma^{-2} \mathbb{X}^\top \mathbb{X}$$

On a $\mathbb{I}(\theta|\mathcal{E}^n) \succ 0$ si et seulement si $\mathbb{X}^\top \mathbb{X} \succ 0$. Dans ce cas, l'EMV qui est ici l'EMC $\hat{\theta}_n^{\text{mc}}$, est Gaussien de matrice de covariance $\mathbb{I}(\theta|\mathcal{E}^n)^{-1}$:

$$\hat{\theta}_n^{\text{mc}} \sim \mathcal{N}(\theta, \mathbb{I}(\theta|\mathcal{E}^n)^{-1})$$

Ce résultat est **non-asymptotique**. D'une autre côté, c'est le comportement qu'on obtient **asymptotiquement** pour les EMV dans les modèles d'échantillonnage réguliers.

Information de Fisher et régression (3/3)

Dans le modèle linéaire Gaussien avec variance inconnue (et design déterministe), on peut calculer l'information de Fisher pour le problème d'estimation du paramètre (θ, σ^2) . On a

$$\nabla_{(\theta, \sigma^2)}^2 \ell_n \begin{pmatrix} \theta \\ \sigma^2 \end{pmatrix} = \begin{pmatrix} \frac{-\mathbf{X}^\top \mathbf{X}}{\sigma^2} & \frac{-\mathbf{X}(\mathbf{Y} - \mathbf{X}\theta)}{\sigma^4} \\ \left[\frac{-\mathbf{X}(\mathbf{Y} - \mathbf{X}\theta)}{\sigma^4} \right]^\top & \frac{n}{2\sigma^4} - \frac{\|\mathbf{Y} - \mathbf{X}\theta\|_2^2}{\sigma^6} \end{pmatrix}$$

alors

$$\mathbb{I}((\theta, \sigma^2) | \mathcal{E}^n) = \begin{pmatrix} \frac{\mathbf{X}^\top \mathbf{X}}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$$

Rem. : la covariance de l'EMV est ici :

$$\text{cov} \begin{pmatrix} \hat{\theta}_n^{\text{mv}} \\ \hat{\sigma}_n^2 \end{pmatrix} = \begin{pmatrix} \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \frac{n}{n-k} \end{pmatrix} \neq \mathbb{I}((\theta, \sigma^2) | \mathcal{E}^n)^{-1}$$

Prévision dans le modèle linéaire Gaussien

Prévision

Modèle linéaire Gaussien

$$Y_i = r(\theta, \mathbf{x}_i) + \xi_i, \quad i = 1, \dots, n$$

où $r(\theta, \mathbf{x}_i) = \langle \theta, \mathbf{x}_i \rangle$ et $\xi_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$.

Exemple : \mathbf{x}_i vecteur de 3 variables explicatives (TV, RADIO, Newspaper) pour le marché i .

- ▶ **Problème de prévision** : On investit dans un nouveau marché avec $\mathbf{x}_0 \in \mathbb{R}^3$. On souhaite estimer les "SALES" attendus, c-à-d prédire la valeur de la fonction de régression en \mathbf{x}_0 : $r(\theta, \mathbf{x}_0) = \langle \theta, \mathbf{x}_0 \rangle$

- ▶ Soit $\hat{\theta}_n$ un estimateur de θ . **Prévision par substitution** :

$$\hat{y} = r(\hat{\theta}_n, \mathbf{x}_0)$$

- ▶ Question statistique : quelle est la qualité de la prévision ? **Intervalle de confiance** pour $r(\theta, \mathbf{x}_0)$ basé sur \hat{y} ?

Prévision : modèle linéaire gaussienne

- ▶ On prend $\hat{\theta}_n = \hat{\theta}_n^{\text{mc}}$ alors la prédiction est $\hat{y} = \langle \mathbf{x}_0, \hat{\theta}_n^{\text{mc}} \rangle$
- ▶ Hyp. 1 : $\xi \sim \mathcal{N}(0, \sigma^2 \text{Id}_n)$
- ▶ Hyp. 2 : $\mathbb{X}^\top \mathbb{X} \succ 0$

Proposition

- (i) $\hat{y} \sim \mathcal{N}(\langle \mathbf{x}_0, \theta \rangle, \sigma^2 \mathbf{x}_0^\top (\mathbb{X}^\top \mathbb{X})^{-1} \mathbf{x}_0)$
- (ii) $\hat{y} - \langle \mathbf{x}_0, \theta \rangle$ et $\mathbb{Y} - \mathbb{X} \hat{\theta}_n^{\text{mc}}$ sont indépendants

Rem. : $\langle \mathbf{x}_0, \theta \rangle = r(\theta, x_0)$ est la quantité qu'on cherche à prédire

Prévision : modèle linéaire gaussienne

- ▶ D'après Proposition 2,

$$\eta := \frac{\hat{y} - \langle \mathbf{x}_0, \theta \rangle}{\sqrt{\sigma^2 \mathbf{x}_0^T (\mathbb{X}^T \mathbb{X})^{-1} \mathbf{x}_0}} \sim \mathcal{N}(0, 1)$$

- ▶ On remplace σ^2 inconnu par $\tilde{\sigma}_n^2 = \|\mathbb{Y} - \mathbb{X} \hat{\theta}_n^{\text{mc}}\|^2 / (n - k)$.
- ▶ **t-statistique :**

$$t := \frac{\hat{y} - \langle \mathbf{x}_0, \theta \rangle}{\sqrt{\tilde{\sigma}_n^2 \mathbf{x}_0^T (\mathbb{X}^T \mathbb{X})^{-1} \mathbf{x}_0}} \sim \frac{g}{\sqrt{\frac{\chi(n-k)}{n-k}}} \sim \text{Student}(n - k),$$

Prévision : intervalle de confiance

Pour $q_{1-\frac{\alpha}{2}}^{t_{n-k}}$, le quantile d'ordre $1 - \alpha/2$ d'une Student(n-k) et la t -statistique

$$t := \frac{\hat{y} - \langle \mathbf{x}_0, \theta \rangle}{\sqrt{\hat{\sigma}_n^2 \mathbf{x}_0^\top (\mathbb{X}^\top \mathbb{X})^{-1} \mathbf{x}_0}}$$

on a

$$\mathbb{P} \left[|t| \leq q_{1-\frac{\alpha}{2}}^{t_{n-k}} \right] = 1 - \alpha$$

On obtient ainsi un **intervalle de confiance de niveau $1 - \alpha$** (non-asymptotique) pour $r(\theta, \mathbf{x}_0) = \langle \mathbf{x}_0, \theta \rangle$:

$$r(\theta, \mathbf{x}_0) \in \left[\hat{y} \pm q_{1-\frac{\alpha}{2}}^{t_{n-k}} \sqrt{\hat{\sigma}_n^2 \mathbf{x}_0^\top (\mathbb{X}^\top \mathbb{X})^{-1} \mathbf{x}_0} \right]$$

avec probabilité $1 - \alpha$.

Prévision : bande de confiance

On peut encadrer la droite de régression par **deux arcs d'hyperboles** donnant ainsi une région de confiance pour la droite de régression. Sous les hypothèses :

- ▶ Hyp. 1 : $\xi \sim \mathcal{N}(0, \sigma^2 \text{Id}_n)$
- ▶ Hyp. 2 : $\mathbb{X}^\top \mathbb{X} \succ 0$

La Proposition 2 assure que

$$\hat{\theta}_n^{\text{mc}} \sim \mathcal{N}(\theta, \sigma^2 (\mathbb{X}^\top \mathbb{X})^{-1})$$

De plus $\hat{\sigma}_n^2 \xrightarrow{\mathbb{P}} \sigma^2$, on en déduit que

$$\frac{\left\| (\mathbb{X}^\top \mathbb{X})^{1/2} (\hat{\theta}_n^{\text{mc}} - \theta) \right\|_2^2}{\hat{\sigma}_n^2} \xrightarrow{d} \chi^2(k).$$

Prévision : bande de confiance

On obtient ainsi une zone de confiance asymptotique de niveau $1 - \alpha$ pour θ donnée par $\hat{\theta}_n^{\text{mc}} + \widehat{\mathcal{E}}_\alpha$ où

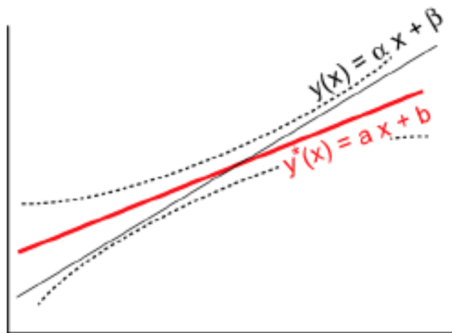
$$\widehat{\mathcal{E}}_\alpha := \left\{ x \in \mathbb{R}^k : \left\| (\mathbb{X}^\top \mathbb{X})^{1/2} x \right\|_2 \leq \hat{\sigma}_n \sqrt{q_{1-\alpha}^{\chi^2(k)}} \right\}$$

et $q_{1-\alpha}^{\chi^2(k)}$ est le quantile d'ordre $1 - \alpha$ d'une $\chi^2(k)$.

$\hat{\theta}_n^{\text{mc}} + \widehat{\mathcal{E}}_\alpha$ est une ellipsoïde centrée en $\hat{\theta}_n^{\text{mc}}$ d'axes et rayons donnés par la décomposition spectrale de $(\mathbb{X}^\top \mathbb{X})$.

A chaque point $\hat{\theta} \in \hat{\theta}_n^{\text{mc}} + \widehat{\mathcal{E}}_\alpha$, on peut associer la droite de régression $x \rightarrow \langle \hat{\theta}, x \rangle$. Ainsi en traçant l'ensemble de toutes ses droites, on obtient une bande de confiance autour de la droite de régression.

Prévision : bande de confiance



Régression linéaire non-gaussienne

Régression linéaire non-gaussienne

Modèle de régression linéaire

$$Y_i = \langle \theta, \mathbf{x}_i \rangle + \xi_i, \quad i = 1, \dots, n.$$

- ▶ Hyp. 1' : ξ_i i.i.d., $\mathbb{E}[\xi_i] = 0$, $\mathbb{E}[\xi_i^2] = \sigma^2 > 0$
- ▶ Hyp. 2' : $\mathbb{X}^\top \mathbb{X} > 0$, $\lim_n \max_{1 \leq i \leq n} \mathbf{x}_i^\top (\mathbb{X}^\top \mathbb{X})^{-1} \mathbf{x}_i = 0$

Proposition (Normalité asymptotique de l'EMC)

Quand $n \rightarrow \infty$,

$$\sigma^{-1}(\mathbb{X}^\top \mathbb{X})^{1/2}(\hat{\theta}_n^{\text{mc}} - \theta) \xrightarrow{d} \mathcal{N}(0, \text{Id}_k).$$

A comparer avec le cadre gaussien : pour tout n ,

$$\sigma^{-1}(\mathbb{X}^\top \mathbb{X})^{1/2}(\hat{\theta}_n^{\text{mc}} - \theta) \sim \mathcal{N}(0, \text{Id}_k)$$

Théorème de Gauss-Markov

Cadre : modèle linéaire (notation matricielle)

$$Y = X\theta + \xi$$

où $\mathbb{E} \xi = 0$, $\mathbb{E} \xi \xi^\top = \sigma^2 I_n$ et $X^\top X \succ 0$.

Théorème (Gauss-Markov)

L'estimateur des moindres carrés $\hat{\theta}_n^{\text{mc}}$ est optimal (au sens du risque quadratique) parmi tous les estimateurs linéaires sans biais : si $\hat{\theta}_n$ est un estimateur de la forme $\hat{\theta}_n = AY$ tel que $A \in \mathbb{R}^{n \times k}$ et $\mathbb{E} \hat{\theta}_n = \theta$ alors

$$\mathbb{E} \left\| \hat{\theta}_n^{\text{mc}} - \theta \right\|_2^2 \leq \mathbb{E} \left\| \hat{\theta}_n - \theta \right\|_2^2$$

Régression non-linéaire

Régression non-linéaire

- ▶ On observe

$$(\mathbf{x}_1, Y_1), \dots, (\mathbf{x}_n, Y_n),$$

où

$$Y_i = r(\theta, \mathbf{x}_i) + \xi_i, \quad i = 1, \dots, n$$

avec

$$\mathbf{x}_i \in \mathbb{R}^k, \quad \text{et } \theta \in \Theta \subset \mathbb{R}^d.$$

- ▶ Si $\xi_i \sim_{\text{i.i.d.}} \mathcal{N}(0, \sigma^2)$,

$$\mathcal{L}_n(\theta, Y_1, \dots, Y_n) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - r(\theta, \mathbf{x}_i))^2\right)$$

et l'estimateur du **maximum de vraisemblance** est obtenu en minimisant la fonction

$$\theta \mapsto \sum_{i=1}^n (Y_i - r(\theta, \mathbf{x}_i))^2.$$

Moindre carrés non-linéaires

Définition

- ▶ *M-estimateur associé à la fonction de contraste*
 $\psi : \Theta \times \mathbb{R}^k \times \mathbb{R} \rightarrow \mathbb{R}$: tout estimateur $\hat{\theta}_n$ satisfaisant

$$\sum_{i=1}^n \psi(\hat{\theta}_n, \mathbf{x}_i, Y_i) = \max_{a \in \Theta} \sum_{i=1}^n \psi(a, \mathbf{x}_i, Y_i).$$

- ▶ Estimateur des *moindres carrés non-linéaires* : associé au contraste $\psi(a, \mathbf{x}, y) = -(y - r(a, \mathbf{x}))^2$.
- ▶ **Extension** des résultats dans le modèle d'échantillonnage dominé au cas cas de v.a. indépendantes **non-équidistribuées**.

Modèle à réponse binaire

- ▶ On observe

$$(\mathbf{x}_1, Y_1), \dots, (\mathbf{x}_n, Y_n), \quad Y_i \in \{0, 1\}, \quad \mathbf{x}_i \in \mathbb{R}^k.$$

- ▶ Modélisation **via la fonction de régression**

$$\mathbf{x} \mapsto p_{\mathbf{x}}(\theta) = \mathbb{E}_{\theta} [Y | \mathbf{X} = \mathbf{x}] = \mathbb{P}_{\theta} [Y = 1 | \mathbf{X} = \mathbf{x}]$$

- ▶ **Représentation**

$$\begin{aligned} Y_i &= p_{\mathbf{x}_i}(\theta) + (Y_i - p_{\mathbf{x}_i}(\theta)) \\ &= r(\theta, \mathbf{x}_i) + \xi_i \end{aligned}$$

avec $r(\theta, \mathbf{x}_i) = p_{\mathbf{x}_i}(\theta)$ et $\xi_i = Y_i - p_{\mathbf{x}_i}(\theta)$.

- ▶ $\mathbb{E}_{\theta} [\xi_i] = 0$ mais structure des ξ_i **compliquée** (dépendance en θ).

Modèle à réponse binaire

- ▶ Y_i v.a. de Bernoulli de paramètre $p_{\mathbf{x}_i}(\theta)$.

Vraisemblance

$$\mathcal{L}_n(\theta, Y_1, \dots, Y_n) = \prod_{i=1}^n p_{\mathbf{x}_i}(\theta)^{Y_i} (1 - p_{\mathbf{x}_i}(\theta))^{1 - Y_i}$$

→ méthodes de résolution numérique.

- ▶ **Régression logistique** (très utile dans les applications)

$$p_{\mathbf{x}}(\theta) = \psi(\langle \mathbf{x}, \theta \rangle),$$

$$\psi(t) = \frac{e^t}{1 + e^t}, \quad t \in \mathbb{R} \quad \text{fonction logistique}$$

Régression logistique et modèles latents

Représentation équivalente de la régression logistique : on observe

$$Y_i = I(Y_i^* > 0), \quad i = 1, \dots, n$$

(les \mathbf{x}_i sont donnés), et Y_i^* est une **variable latente** ou cachée,

$$Y_i^* = \langle \theta, \mathbf{x}_i \rangle + U_i, \quad i = 1, \dots, n$$

avec $U_i \stackrel{i.i.d.}{\sim} F$, où

$$F(t) = \frac{1}{1 + e^{-t}}, \quad t \in \mathbb{R}.$$

car, pour la fonction logistique ψ ,

$$\mathbb{P}_\theta [Y_i^* > 0] = \psi(\langle \mathbf{x}_i, \theta \rangle) = \mathbb{P}[Y_i = 1]$$

Modèle à réponse discrète multiples : modèle de Poisson

- ▶ On observe

$$(\mathbf{x}_1, Y_1), \dots, (\mathbf{x}_n, Y_n), \quad Y_i \in \mathbb{N}, \mathbf{x}_i \in \mathbb{R}^k.$$

- ▶ Modélisation via la densité de $Y|X = \mathbf{x}$:

$$k \in \mathbb{N} \mapsto p_{\mathbf{x}}(\theta, k) = \mathbb{P}_{\theta} [Y = k | \mathbf{X} = \mathbf{x}]$$

- ▶ **Modèle de Poisson** $Y|X = \mathbf{x} \sim \text{Poisson}(\exp(\langle \theta, \mathbf{x} \rangle))$: pour tout $k \in \mathbb{N}$,

$$\mathbb{P}_{\theta}[Y = k | X = \mathbf{x}] = \frac{\lambda^k}{k!} \exp(-\lambda) \text{ où } \lambda = \exp(\langle \theta, \mathbf{x} \rangle).$$

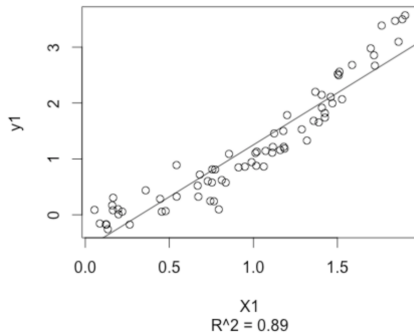
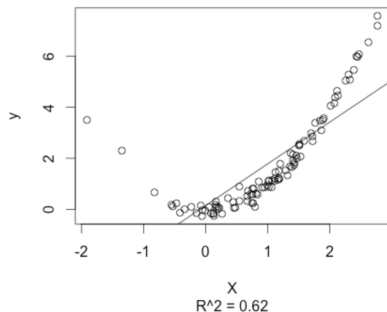
- ▶ $\mathbb{E}_{\theta}[Y | X = \mathbf{x}] = \exp(\langle \theta, \mathbf{x} \rangle)$, $\text{var}(Y | X = \mathbf{x}) = \exp(\langle \theta, \mathbf{x} \rangle)$.

Test empirique pour le modèle linéaire

Le Rainbow test

Idée : Même si la vraie relation entre Y et les covariables n'est pas linéaire, localement on peut imaginer qu'elle l'est (approximation d'ordre de 1 de Taylor). Si on construit un estimateur par moindres carrés à partir d'un sous-ensemble de données autour de \bar{X}_n , alors cette régression devrait être assez bonne.

Par exemple : $Y = X^2 + \mathcal{N}(0, 1)$



Le *Rainbow test*

On note $\tilde{\theta}$ l'estimateur construit à partir de m données d'indices $I \subset \{1, \dots, n\}$ autour de \bar{X}_n et par $\tilde{y}_i = \langle X_i, \tilde{\theta} \rangle$ la valeur prédite en X_i . On a donc un R^2 (*coefficient de détermination*) donné par

$$\tilde{R}_I^2 = 1 - \frac{\sum_{i \in I} (y_i - \tilde{y}_i)^2}{\sum_{i \in I} (y_i - \bar{y}_I)^2}$$

Idée : L'idée centrale du *Rainbow test* est que si le modèle est vraiment linéaire alors l'ajout de données au sous-échantillon $(y_i, X_i)_{i \in I}$ ne devrait pas trop modifier le R^2 . Par contre, si le modèle n'est pas linéaire alors l'ajout de donnée loin de \bar{X}_n devrait dégrader le R^2 . La comparaison entre le R^2 local autour de $\bar{X}_n : \tilde{R}_I^2$; et le R^2 de tout l'échantillon est à la base du *Rainbow test*.

Statistic de test du Rainbow test :

$$T = \frac{(R^2 - \tilde{R}_I^2) (m - k)}{\tilde{R}_I^2 (n - m)}.$$

Sous hypothèse de linéarité (modèle linéaire gaussien), on a

$$T \sim F(n - m, m - k)$$

(loi de Fisher de degrés $(n - m, m - k)$).

Le *Rainbow test*

Le choix du sous-échantillon pour le *Rainbow test* se fait généralement en prenant les $m > k$ données les plus proche de \bar{X}_n pour la **distance de Mahalanobis** :

$$d(x, y) = \sqrt{(x - y)^\top (\mathbb{X}^\top \mathbb{X})(x - y)} = \|\mathbb{X}(x - y)\|_2.$$

On choisit donc pour sous-ensemble de données $(y_i, X_i)_{i \in I}$ l'ensemble de m données telles que $d(X_i, \bar{X}_n)$ est la plus petite.

Autre tests

- ▶ Ramsey's RESET test : "Regression Specification Error Test"
- ▶ Harvey and Collier test : for a convex or concave alternative
- ▶ Test de Breusch-Pagan sur l'homoscédasticité du terme d'erreur.
- ▶ test de Durbin-Watson : tester l'autocorrélation des résidus dans un modèle de régression linéaire.
- ▶ F-test (ou test de Fisher) et ANOVA : test d'égalité de variance et de fit du modèle.