Statistiques mathématiques : cours 8

Guillaume Lecué

28 septembre 2016

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 $\mathbf{X}$

Principe : on part de l'observation de n couples

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

Exemple : sur le i-ième marché,

- $Y_i =$ "*Sales*"
- $X_i =$ ("*TV*", "*Radio*", "*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 $r : \mathbb{R}^k \to \mathbb{R}$ telle que

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

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

- Alternative : on modélise ces données avec le modèle

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

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

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

  où

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

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

Régression

- On définit :
$$\xi = Y - \mathbb{E}\left[Y|\,\mathbf{X}\right] \implies \mathbb{E}\left[\xi\right] = 0.$$
- On a alors naturellement la représentation désirée
$$Y = r(\mathbf{X}) + \xi, \quad \mathbb{E}\left[\xi\right] = 0$$
en posant
$$\boxed{r(\mathbf{x}) = \mathbb{E}\left[Y|\,\mathbf{X} = \mathbf{x}\right], \ \ \mathbf{x} \in \mathbb{R}^k}$$
- On observe alors n couples
$$(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$$
où
$$Y_i = r(\mathbf{X}_i) + \xi_i, \ \ \mathbb{E}\left[\xi_i\right] = 0$$
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*

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

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

Régression à design déterministe

- Principe : sur un exemple. On observe

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

  où $r(\theta, \cdot) : [0,1] \to \mathbb{R}$ est une fonction connue au paramètre $\theta \in \Theta \subset \mathbb{R}^d$ près, et les ξ_i sont i.i.d., $\mathbb{E}\left[\xi_i\right] = 0$.
- But : reconstruire $r(\theta, \cdot)$ c'est-à-dire estimer θ.
- Plus généralement, on observe $(Y_i)_{i=1}^n$ où

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

  et $\mathbf{x}_1, \ldots, \mathbf{x}_n$ 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*

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

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

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

- $\mathbf{x}_i$ *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 distribuées.*

Régression linéaire

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

$$\boxed{r(\theta, x) = \langle \theta, x \rangle}$$

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

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

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

EMV en régression gaussienne à design déterministe

Modèle de régression gaussienne à design déterministe :

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

où $g_i \sim \mathcal{N}(0,1)$, i.i.d..
Problème : estimer θ ?
Idée : Expliciter la loi de l'observation $Z = (Y_1, \ldots, Y_n)$ et appliquer le principe du maximum de vraisemblance.
La loi de Y_i : $\mathbb{P}_{Y_i} = f_{\mathbf{x}_i}(\theta, \cdot).\lambda$ où $\forall y \in \mathbb{R}$

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

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

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

EMV pour régression gaussienne à design déterministe

On travail 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

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

$$= \tfrac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\tfrac{1}{2\sigma^2}\sum_{i=1}^n \left(y_i - r(\theta,\mathbf{x}_i)\right)^2\right) := f(\theta,(y_i)_{i=1}^n)$$

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

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

Estimateur des moindres carrés

Maximiser la vraisemblance en régression gaussienne

$$\Updownarrow$$

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

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

Définition

Estimateur des moindres carrés (EMC) : tout estimateur $\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}}$ tel que

$$\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}} \in \arg\min_{\theta \in \mathbb{R}^k} \sum_{i=1}^{n} \big(Y_i - r(\theta, \mathbf{x}_i)\big)^2$$

$$\boxed{\text{En régression Gaussienne : } \underline{\text{EMV} = \text{EMC}}}$$

Droite de régression ($k = 1$)

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

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

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

$$\boxed{Y_i = a + b\, x_i + \xi_i, \quad i = 1, \ldots, n}$$

où les $x_1, \ldots, x_n$ sont des réels donnés et $\xi_1, \ldots, \xi_n$ 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 :

$$\widehat{\theta}_{\mathtt{n}}^{\mathtt{mc}} = \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \arg \min_{(a,b)^\top \in \mathbb{R}^2} \sum_{i=1}^{n} \left(Y_i - a - bx_i \right)^2$$

Estimateur des moindres carrés (1/2)

On peut réécrire la fonction objectif sous forme matricielle :

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

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

$$\boxed{\mathbb{X}^\top\mathbb{X}\,\widehat{\theta}_{\mathbf{n}}^{\mathtt{mc}} = \mathbb{X}^\top\mathbb{Y}}$$

Estimateur des moindres carrés (2/2)

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

$$\boxed{\widehat{\theta}_{\mathrm{n}}^{\,\mathtt{mc}} = \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbb{Y}}$$

- Résidu : si $\widehat{\theta}_n$ est un estimateur de θ alors $\hat{y}_i = r(\widehat{\theta}_n, x_i)$ 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)

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

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

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

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

- Problème : estimer θ
- le cas du design aléatoire se traite identiquement quand on conditionne par rapport au design (il y a cependant des différences entre les deux modèles)

Ecriture matricielle des données

Matriciellement, on réécrit ces données comme

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

où

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

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 $\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}}$ minimisant

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

- En notation matricielle :

$$\|\mathbb{Y} - \mathbb{X}\,\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}}\|^2 = \min_{\theta \in \mathbb{R}^k} \|\mathbb{Y} - \mathbb{X}\theta\|^2 = \min_{v \in V} \|\mathbb{Y} - v\|^2$$

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

Géométrie de l'EMC

- L'EMC vérifie
$$\boxed{\mathbb{X}\widehat{\theta}_n^{\,\mathtt{mc}} = P_V \mathbb{Y}}$$
où P_V est le projecteur orthogonal sur V.
- Mais $\mathbb{X}^\top P_V = \mathbb{X}^\top P_V^\top = (P_V\mathbb{X})^\top = \mathbb{X}^\top$. On en déduit les équations normales des moindres carrés :
$$\boxed{\mathbb{X}^\top\mathbb{X}\widehat{\theta}_n^{\,\mathtt{mc}} = \mathbb{X}^\top\mathbb{Y}} \tag{1}$$
- <u>Remarques.</u>
  - L'EMC est un Z-estimateur (bonnes propriétés quand 1 a une unique solution càd $\mathbb{X}^\top\mathbb{X} \succ 0$).
  - Pas d'unicité de $\widehat{\theta}_n^{\,\mathtt{mc}}$ si la matrice $\mathbb{X}^\top\mathbb{X}$ n'est pas inversible.
  - (1) est équivalente à $\nabla F(\widehat{\theta}_n^{\,\mathtt{mc}}) = 0$

Géométrie de l'EMC

Proposition

Si $\mathbb{X}^\top \mathbb{X}$ (matrice $k \times k$) est inversible, alors $\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}}$ est unique et

$$\boxed{\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}} = \left(\mathbb{X}^\top \mathbb{X}\right)^{-1} \mathbb{X}^\top \mathbb{Y}}$$

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

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

En particulier, $\mathbb{X}^\top \mathbb{X} \succ 0 \quad \Longrightarrow \quad n \geq k$

Régression linéaire multiple sur les données Advertising.csv

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

 = Modèle linéaire gaussien

On suppose que le vecteur bruit est tel que

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

dans le modèle (sous forme matricielle)

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

On a alors plusieurs propriétés remarquables :

- l'EMC $\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}}$ = EMV (dans le modèle à variance connue)
- On sait expliciter la loi (non-asymptotique !) de $\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}}$

Cadre gaussien : loi des estimateurs

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

Proposition (2)

(i) $\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}} \sim \mathcal{N}\left(\theta, \sigma^2 \left(\mathbb{X}^\top \mathbb{X}\right)^{-1}\right)$

(ii) $\|\mathbb{Y} - \mathbb{X}\,\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}}\|_2^2 \sim \sigma^2 \chi^2(n-k)$

(iii) $\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}}$ *et* $\mathbb{Y} - \mathbb{X}\,\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}}$ *sont indépendants*

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

1. $A_j\,\boldsymbol{\xi} \sim \mathcal{N}\left(0, A_j\right)$ sont indépendants,
2. $\|A_j \boldsymbol{\xi}\|_2^2 \sim \chi^2(\mathrm{Rang}(A_j))$

Preuve de la proposition 2 (directe, sans Cochran)

(i) $\widehat{\theta}_{\mathsf{n}}^{\,\mathtt{mc}} = \theta + \left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}^\top\boldsymbol{\xi}$ est une transformation affine d'un vecteur Gaussien donc $\widehat{\theta}_{\mathsf{n}}^{\,\mathtt{mc}}$ est aussi un vecteur Gaussien ; sa moyenne et matrice de covariance sont :

1. $\mathbb{E}[\widehat{\theta}_{\mathsf{n}}^{\,\mathtt{mc}}] = \theta$
2. $\mathrm{Cov}(\widehat{\theta}_{\mathsf{n}}^{\,\mathtt{mc}}) = \mathbb{E}\left[\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}^\top\boldsymbol{\xi}\left(\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}^\top\boldsymbol{\xi}\right)^\top\right] = \sigma^2\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}$

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

$$\begin{aligned}\mathbb{Y} - \mathbb{X}\,\widehat{\theta}_{\mathsf{n}}^{\,\mathtt{mc}} &= \mathbb{X}\left(\theta - \widehat{\theta}_{\mathsf{n}}^{\,\mathtt{mc}}\right) + \boldsymbol{\xi} \\ &= -\mathbb{X}\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}^\top\boldsymbol{\xi} + \boldsymbol{\xi} = \sigma(\mathrm{Id}_n - P_V)\boldsymbol{\xi}'\end{aligned}$$

(iii) le vecteur $(\widehat{\theta}_{\mathsf{n}}^{\,\mathtt{mc}}, \mathbb{Y} - \mathbb{X}\,\widehat{\theta}_{\mathsf{n}}^{\,\mathtt{mc}})$ est gaussien (transformation linéaire de $\boldsymbol{\xi}$). On calcule sa matrice de covariance.

Modèle linéaire Gaussien – variance inconnue

Dans le modèle linéaire Gaussien

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

où θ et σ sont inconnus on a :

$$\mathrm{EMV} = \begin{pmatrix} \widehat{\theta}_{\mathsf{n}}^{\,\mathtt{mc}} \\ \hat{\sigma}_n^2 \end{pmatrix} \text{ où } \hat{\sigma}_n^2 = \frac{\|\mathbb{Y} - \mathbb{X}\,\widehat{\theta}_{\mathsf{n}}^{\,\mathtt{mc}}\,\|_2^2}{n}$$

car la log-vraisemblance

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

est maximale en ce point

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

$$\mathrm{EMV} = \begin{pmatrix} \widehat{\theta}_{\mathrm{n}}^{\,\mathtt{mc}} \\ \hat{\sigma}_n^2 \end{pmatrix}$$

où

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

D'après Proposition 2 :

- $\widehat{\sigma}_n^2$ est indépendant de $\widehat{\theta}_{\mathrm{n}}^{\,\mathtt{mc}}$
- $\widehat{\theta}_{\mathrm{n}}^{\,\mathtt{mc}} \sim \mathcal{N}\left(\theta, \sigma^2 (\mathbb{X}^\top \mathbb{X})^{-1}\right)$
- $n\widehat{\sigma}_n^2/\sigma^2 \sim \chi^2(n-k)$

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

Lois des coordonnées de $\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}}$:

$$(\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}})_j - \theta_j \sim \mathcal{N}(0, \sigma^2 b_j)$$

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

$$\frac{(\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}})_j - \theta_j}{\widetilde{\sigma}_n \sqrt{b_j}} \sim t_{n-k} \text{ pour } \widetilde{\sigma}_n = \frac{\|\mathbb{Y} - \mathbb{X}\,\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}}\|_2^2}{n-k}$$

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 variablité de Y).

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

Problème : on ne veut garder que les variables pertinantes, 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 $\mathbb{X} \to \mathbb{X}_{-j}$) et on construit $\widehat{\theta}_{\mathrm{n}}^{\,\mathtt{mc}}(-j)$ à partir de $\mathbb{Y}$ et $\mathbb{X}_{-j}$
2. on choisi j_1 pour lequel

$$RSS(\widehat{\theta}_{\mathrm{n}}^{\,\mathtt{mc}}(-j_1)) = \min_{1\leq j\leq k} RSS(\widehat{\theta}_{\mathrm{n}}^{\,\mathtt{mc}}(-j)) := RSS_{k-1}$$

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

$$\mathrm{RSS}_m \approx \mathrm{RSS}_{m-1}$$

4. à la fin, seules les colonnes restantes de $\mathbb{X}$ sont des features pertinantes : 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)

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

où $\theta = (\theta_1, \dots, \theta_k)^T \in \mathbb{R}^k$, $\mathbb{X} \in \mathbb{R}^{n \times k}$ et $\mathbb{X}^\top \mathbb{X} \succ 0$.
Problème de test : $a \in \mathbb{R}$, $j \in \{1, \dots, k\}$ donné

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

On a vu que, sous $\mathbb{P}_\theta$,

$$\frac{\left(\widehat{\theta}_{\mathrm{n}}^{\,\mathtt{mc}}\right)_j - \theta_j}{\widetilde{\sigma}_n \sqrt{(\mathbb{X}^\top \mathbb{X})^{-1}_{jj}}} \overset{d}{=} \text{Student}(n-k) \text{ où } \widetilde{\sigma}_n = \frac{\|\mathbb{Y} - \mathbb{X}\,\widehat{\theta}_{\mathrm{n}}^{\,\mathtt{mc}}\|_2^2}{n-k}$$

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)

$$t_n := \frac{\left|\left(\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}}\right)_j - a\right|}{\widetilde{\sigma}_n \sqrt{(\mathbb{X}^\top \mathbb{X})_{jj}^{-1}}}$$

En particulier, pour $a = 0$, on test si le coefficient associè à la j-ième feature est nul. Si on rejete le test, alors cette feature sera <u>sélectionnée</u> (avec un niveau de confiance de $1 - \alpha$). 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 :

$$\boxed{H_0 : \theta_\ell = 0, \quad \forall \ell = k_0, \ldots, k}$$

contre

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

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 :

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

contre

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

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)

$$\mathbb{G}\,\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}} \sim \mathcal{N}\big(\mathbf{b}, \sigma^2 \mathbb{G}(\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{G}^\top\big)$$

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

$$(\mathbb{G}\,\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}} - \mathbf{b})^\top \mathbf{U}^{-1} (\mathbb{G}\,\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}} - \mathbf{b}) \sim \chi^2(m)$$

Si σ^2 est inconnue, on remplace σ^2 par $\widetilde{\sigma}_n^2 = \frac{\|\mathbf{Y} - \mathbb{X}\,\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}}\|_2^2}{n-k}$. Alors la loi de

$$\frac{(\mathbb{G}\,\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}} - \mathbf{b})^\top \widehat{\mathbf{U}}^{-1} (\mathbb{G}\,\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}} - \mathbf{b})}{m}$$

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*

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

On a alors un test de niveau α pour

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

où

$$T_n = \frac{(\mathbb{G}\,\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}} - \mathbf{b})^T \widehat{\mathbf{U}}^{-1} (\mathbb{G}\,\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}} - \mathbf{b})}{m} \text{ et } \widehat{\mathbf{U}} = \widetilde{\sigma}_n^2 \mathbb{G}(\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{G}^\top$$

Information de Fisher dans le modèle linéaire Gaussien

Information de Fisher et régression (1/3)

Cadre : $\mathcal{E}^n$ expérience engendrée par $(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n)$ 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 $\mathbf{x}_1, \ldots, \mathbf{x}_n$ sont déterministes.
Observation : $Z^n = (Y_1, \ldots, Y_n)$ de densité (par rapport à Lebesgue sur $\mathbb{R}^n$)

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

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

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

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 $\widehat{\theta}_{\mathsf{n}}^{\,\mathtt{mc}}$, est Gaussien de matrice de covariance $\mathbb{I}(\theta|\mathcal{E}^n)^{-1}$:

$$\boxed{\widehat{\theta}_{\mathsf{n}}^{\,\mathtt{mc}} \sim \mathcal{N}\left(\theta, \mathbb{I}(\theta|\mathcal{E}^n)^{-1}\right)}$$

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

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

alors

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

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

$$\operatorname{cov} \begin{pmatrix} \widehat{\theta}_n^{\,\mathtt{mv}} \\ \hat{\sigma}_n^2 \end{pmatrix} = \begin{pmatrix} \frac{\mathbb{X}^\top \mathbb{X}}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \frac{n-k}{n} \end{pmatrix}$$

Prévision dans le modèle linéaire Gaussien

Prévision

Modèle linéaire Gaussien

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

où $r(\theta, \mathbf{x}_i) = \langle \theta, \mathbf{x}_i \rangle$ et $\xi_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$.
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, càd prédire la valeur de la fonction de régression en $\mathbf{x}_0$: $r(\theta, \mathbf{x}_0) = \langle \theta, \mathbf{x}_0 \rangle$
- Soit $\widehat{\theta}_n$ un estimateur de θ. Prévision par substitution :
$$\boxed{\widehat{y} = r(\widehat{\theta}_n, \mathbf{x}_0)}$$
- Question statistique : quelle est la qualité de la prévision ? Intervalle de confiance pour $r(\theta, \mathbf{x}_0)$ basé sur $\widehat{y}$?

Prévision : modèle linéaire gaussienne

- On prend $\widehat{\theta}_n = \widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}}$ alors la prédiction est $\boxed{\widehat{y} = \left\langle \mathbf{x}_0, \widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}} \right\rangle}$
- <u>Hyp. 1</u> : $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_n)$
- <u>Hyp. 2</u> : $\mathbb{X}^\top \mathbb{X} \succ 0$

Proposition

(i) $\widehat{y} \sim \mathcal{N}\left(\left\langle \mathbf{x}_0, \theta \right\rangle, \sigma^2 \mathbf{x}_0^T \left(\mathbb{X}^\top \mathbb{X}\right)^{-1} \mathbf{x}_0\right)$

(ii) $\widehat{y} - \left\langle \mathbf{x}_0, \theta \right\rangle$ *et* $\mathbb{Y} - \mathbb{X}\,\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}}$ *sont indépendants*

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

Prévision : modèle linéaire gaussienne

- D'après Proposition 2,

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

- On remplace σ^2 inconnu par $\widetilde{\sigma}_n^2 = \|\mathbb{Y} - \mathbb{X}\,\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}}\|^2/(n-k)$.
- *t*-statistique :

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

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

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

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 $r(\theta, \mathbf{x}_0) = \left\langle \mathbf{x}_0, \theta \right\rangle$:

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

avec probabilité $1-\alpha$.

Régression linéaire non-gaussienne

Régression linéaire non-gaussienne

Modèle de régression linéaire

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

- Hyp. 1' : ξ_i i.i.d., $\mathbb{E}[\xi_i] = 0$, $\mathbb{E}[\xi_i^2] = \sigma^2 > 0$
- Hyp. 2' : $\mathbb{X}^\top \mathbb{X} > 0$, $\lim_n \max_{1 \leq i \leq n} \mathbf{x}_i^T \left(\mathbb{X}^\top \mathbb{X}\right)^{-1} \mathbf{x}_i = 0$

Proposition (Normalité asymptotique de l'EMC)

Quand $n \to \infty$,

$$\sigma^{-1} \left(\mathbb{X}^\top \mathbb{X}\right)^{1/2} (\widehat{\theta}_{\mathrm{n}}^{\,\mathtt{mc}} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \mathrm{Id}_k\right).$$

A comparer avec le cadre gaussien : pour tout n,

$$\sigma^{-1} \left(\mathbb{X}^\top \mathbb{X}\right)^{1/2} (\widehat{\theta}_{\mathrm{n}}^{\,\mathtt{mc}} - \theta) \sim \mathcal{N}\left(0, \mathrm{Id}_k\right)$$

Théorème de Gauss-Markov

Cadre : modèle linéaire (notation matricielle)

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

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

Théorème (Gauss-Markov)

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

$$\mathbb{E}\left\|\widehat{\theta}_{\mathsf{n}}^{\,\mathtt{mc}} - \theta\right\|_2^2 \leq \mathbb{E}\left\|\widehat{\theta}_n - \theta\right\|_2^2$$

Régression non-linéaire

Régression non-linéaire

- On observe
$$(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n),$$
où
$$\boxed{Y_i = r(\theta, \mathbf{x}_i) + \xi_i, \quad i = 1, \ldots, n}$$
avec
$$\mathbf{x}_i \in \mathbb{R}^k, \quad \text{et} \quad \theta \in \Theta \subset \mathbb{R}^d.$$
- Si $\xi_i \sim_{\text{i.i.d.}} \mathcal{N}(0, \sigma^2)$,
$$\mathcal{L}_n(\theta, Y_1, \ldots, Y_n) \propto \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n \big(Y_i - r(\theta, \mathbf{x}_i)\big)^2\Big)$$
et l'estimateur du maximum de vraisemblance est obtenu en minimisant la fonction
$$\theta \mapsto \sum_{i=1}^n \big(Y_i - r(\theta, \mathbf{x}_i)\big)^2.$$

Moindre carrés non-linéaires

Définition

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

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

- *Estimateur des moindres carrés non-linéaires : associé au contraste* $\psi(a, \mathbf{x}, y) = -\big(y - r(a, \mathbf{x})\big)^2$.

- Extension des résultats dans le modèle d'échantillonnage dominé au cas cas de v.a. indépendantes non-équidistribuées.

Modèle à réponse binaire

- On observe

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

- Modélisation via la fonction de régression

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

- Représentation

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

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

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

Modèle à réponse discrète

- Y_i v.a. de Bernoulli de paramètre $p_{\mathbf{x}_i}(\theta)$.
  Vraisemblance
  $$\mathcal{L}_n(\theta, Y_1, \ldots, Y_n) = \prod_{i=1}^{n} p_{\mathbf{x}_i}(\theta)^{Y_i} \big(1 - p_{\mathbf{x}_i}(\theta)\big)^{1-Y_i}$$
  $\rightarrow$ méthodes de résolution numérique.
- Régression logistique (très utile dans les applications)
  $$p_{\mathbf{x}}(\theta) = \psi(\langle \mathbf{x}, \theta \rangle),$$
  $$\psi(t) = \frac{e^t}{1 + e^t}, \; t \in \mathbb{R} \quad \text{fonction logistique}$$

Régression logistique et modèles latents

Représentation équivalente de la régression logistique : on observe

$$\boxed{Y_i = I\big(Y_i^\star > 0\big), \quad i = 1, \ldots, n}$$

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

$$\boxed{Y_i^\star = \big\langle \theta, \mathbf{x}_i \big\rangle + U_i, \ \ i = 1, \ldots, n}$$

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

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

car, pour la fonction logistique ψ,

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