MAP 433 : Introduction aux méthodes statistiques. Cours 5

25 Septembre 2015

Aujourd'hui

1. Méthode d'estimation dans le modèle de régression
   - Modèle de régression
   - Régression à design déterministe
   - La droite des moindres carrés
   - Régression linéaire multiple
   - Propriétés de l'estimateur des Moindres Carrés
   - Le cas gaussien
   - Modèle linéaire gaussien
2. Sélection de variables
   - Sélection rétrograde
   - LASSO
3. Régression non-linéaire
4. Bilan provisoire : modèles paramétriques dominés

Influence d'une variable sur une autre

- Principe : on part de l'observation d'un n-échantillon

$$Y_1, \ldots, Y_n \quad (Y_i \in \mathbb{R})$$

- A chaque observation Y_i est associée une observation auxiliaire $\boldsymbol{X}_i \in \mathbb{R}^k$.
- On suspecte l'échantillon

$$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \quad (\boldsymbol{X}_i \in \mathbb{R}^k)$$

de contenir la ≪ majeure partie de la variabilité des Y_i ≫.

Modélisation de l'influence

- Si $\boldsymbol{X}_i$ contient toute la variabilité de Y_i, alors Y_i est mesurable par rapport à $\boldsymbol{X}_i$: il existe $r : \mathbb{R}^k \to \mathbb{R}$ telle que
$$Y_i = r(\boldsymbol{X}_i),$$
mais peu réaliste (ou alors problème d'interpolation numérique).
- Alternative : représentation précédente avec erreur additive : on postule
$$Y_i = r(\boldsymbol{X}_i) + \xi_i,$$
ξ_i erreur aléatoire centrée (pour des raisons d'identifiabilité).

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 $\boldsymbol{X}$-mesurable est donnée par l'espérance conditionnelle $\mathbb{E}\left[Y | \boldsymbol{X}\right]$:

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

- où

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

- On appelle $r(\cdot)$ fonction de régression de Y sur $\boldsymbol{X}$.

Régression

- On définit :

$$\xi = Y - \mathbb{E}\left[Y | \boldsymbol{X}\right] \implies \mathbb{E}\left[\xi\right] = 0.$$

- On a alors naturellement la représentation désirée

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

si l'on pose

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

- On observe alors un n-échantillon

$$(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$$

où

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

avec comme paramètre la fonction $r(\cdot)$+ un jeu d'hypothèses

régresseurs aléatoires

Definition

Modèle de régression à design aléatoire = donnée de l'observation

$$(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$$

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

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

- $\boldsymbol{x} \rightsquigarrow r(\beta, \boldsymbol{x})$ *fonction de régression, connue au paramètre* β *près.*
- $\boldsymbol{X}_i$ *= variables explicatives, co-variables, prédicteurs ;* $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ *= design.*

Modèle de régression à design déterministe

Definition

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

$$(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_n, Y_n)$$

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

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

- $\boldsymbol{x}_i$ *déterministes, donnés (ou choisis) : plan d'expérience, points du « design ».*
- *Hypothèses sur les* ξ_i *: à débattre. Pour simplifier, les variables* ξ_i *sont centrées,* $\mathbb{E}_\theta[\xi_i] = 0$, *décorrélées,* $\mathbb{E}_\theta[\xi_i\xi_j] = 0$ *si* $i \neq j$ *et de variance unité* $\mathbb{E}[\xi_i^2] = 1$ *(homoscédasticité).*
- *Attention ! Les* Y_i *ne sont pas identiquement distribuées.*

Régression gaussienne

- Modèle de régression à design déterministe :

$$Y_i = r(\boldsymbol{\beta}, \boldsymbol{x}_i) + \ \sigma xi_i, \quad \theta \in \Theta \subset \mathbb{R}^d \times \mathbb{R}_+.$$

- Supposons : $\xi_i \sim \mathcal{N}(0,1)$, i.i.d.
- On a alors le modèle de régression gaussienne. Comment estimer θ ? On sait expliciter la loi de l'observation $Z = (Y_1, \ldots, Y_n) \Longrightarrow$ appliquer le principe du maximum de vraisemblance.
- La loi de Y_i :

$$\mathbb{P}^{Y_i}(dy) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\big(-\frac{1}{2\sigma^2}(y - r(\boldsymbol{\beta}, \boldsymbol{x}_i))^2\big)\, dy \ll dy.$$

EMV pour régression gaussienne

- Le modèle $\{\mathbb{P}_\theta^n = \text{loi de } (Y_1, \ldots, Y_n), \theta \in \mathbb{R}^k\}$ est dominé par $\mu^n(dy_1 \ldots dy_n) = dy_1 \ldots dy_n$.
- D'où

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

$$= \tfrac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\big(-\tfrac{1}{2\sigma^2}\sum_{i=1}^n \big(y_i - r(\boldsymbol{\beta}, \boldsymbol{x}_i)\big)^2\big).$$

- La fonction de vraisemblance

$$\boxed{\mathcal{L}_n(\theta, Y_1, \ldots, Y_n) \propto \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n \big(Y_i - r(\boldsymbol{\beta}, \boldsymbol{x}_i)\big)^2\Big)}$$

Estimateur des moindres carrés

Maximiser la vraisemblance en régression gaussienne = minimiser la somme des carrés :

$$\sum_{i=1}^{n} e\big(Y_i - r(\beta, \boldsymbol{x}_i)\big)^2 \to \min_{\theta \in \Theta}.$$

Definition

Estimateur des moindres carrés : tout estimateur $\widehat{\beta}_n$ t.q.

$$\widehat{\beta}_n \in \arg\min_{\beta \in \Theta} \sum_{i=1}^{n} \big(Y_i - r(\beta, \boldsymbol{x}_i)\big)^2.$$

- L'EMC est un M-estimateur. Pour le modèle de régression gaussienne : $\boxed{\mathrm{EMV} = \mathrm{EMC}}$.
- Existence, unicité.

Droite de régression

- Modèle le plus simple $\boxed{r(\beta, x) = \beta_0 + \beta_1 x}$

$$\boxed{Y_i = \beta_0 + \beta_1 x_i + \xi_i, \quad i = 1, \ldots, n}$$

avec $\beta = (\beta_0, \beta_1)^T \in \mathbb{R}^2$ et les $(x_1, \ldots, x_n)$ données.

- L'estimateur des moindres carrés :

$$\widehat{\beta}_n = (\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{(b_0, b_1) \in \mathbb{R}^2} \sum_{i=1}^{n} \big(Y_i - b_0 - b_1 x_i\big)^2.$$

- Solution explicite

Droite de régression

- Le minimum est caractérisé par les équations

$$\begin{cases} b_0 + b_1 n^{-1} \sum_{i=1}^n x_i & = n^{-1} \sum_{i=1}^n Y_i \\ b_0 n^{-1} \sum_{i=1}^n x_i + b_1 n^{-1} \sum_{i=1}^n x_i^2 & = n^{-1} \sum_{i=1}^n x_i Y_i \,. \end{cases}$$

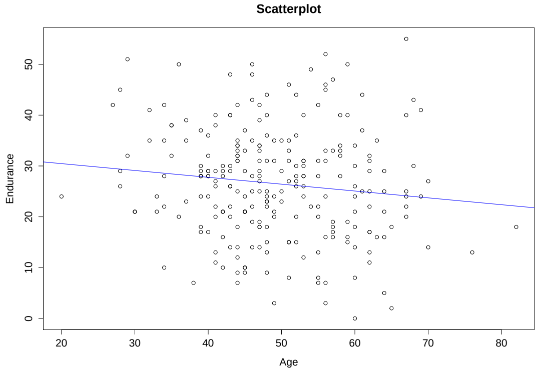
- Notons $\bar{x}_n = n^{-1} \sum_{i=1}^n x_i$. Si le déterminant $\Delta_n \neq 0$ où

$$\Delta_n = \begin{vmatrix} 1 & n^{-1} \sum_{i=1}^n x_i \\ n^{-1} \sum_{i=1}^n x_i & n^{-1} \sum_{i=}^n x_i^2 \end{vmatrix} = S_{xx} = n^{-1} \sum_{i=1}^n (x_i^2 - \bar{x}_n)^2 \,, \qquad ,$$

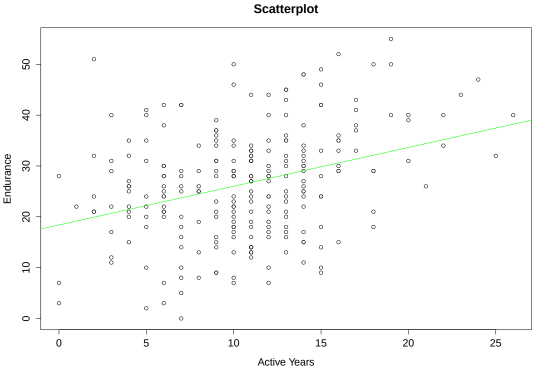
alors ce système d'équations a une solution unique :

$$\begin{cases} \widehat{\beta}_{n0} & = \bar{Y}_n - \widehat{\beta}_{n1} \bar{x}_n \\ \widehat{\beta}_{n1} & = \frac{S_{xY}}{S_{xx}} \,, \quad S_{xY} = n^{-1} \sum_{i=1}^n (x_i - \bar{X}_n)(Y_i - \bar{x}_n) \,. \end{cases}$$

Régression linéaire simple



Régression linéaire simple



Régression linéaire multiple (=Modèle linéaire)

- La fonction de régression est $r(\boldsymbol{\beta}, \boldsymbol{x}_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$. On observe

$$(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_n, Y_n)$$

avec

$$\boxed{Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \sigma \xi_i, \quad i = 1, \ldots, n}$$

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

- Matriciellement

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

avec

  - $\boldsymbol{Y} = (Y_1 \cdots Y_n)^T$,
  - $\boldsymbol{\xi} = (\xi_1 \cdots \xi_n)^T$
  - $\mathbb{X}$ la matrice $(n \times k)$ dont la i-ème ligne est $\mathbb{X}_{i,\cdot} = \boldsymbol{x}_i^T$.

Régression linéaire multiple (=Modèle linéaire)

- La fonction de régression est $r(\beta, \boldsymbol{x}_i) = \boldsymbol{x}_i^T \beta$. On observe

$$(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_n, Y_n)$$

  avec

$$\boxed{Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \sigma \xi_i, \quad i = 1, \ldots, n}$$

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

- Matriciellement

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

  avec
  - $\boldsymbol{Y} = (Y_1 \cdots Y_n)^T$,
  - $\boldsymbol{\xi} = (\xi_1 \cdots \xi_n)^T$
  - $\mathbb{X}$ la matrice $(n \times k)$ dont la i-ème ligne est $\mathbb{X}_{i,\cdot} = \boldsymbol{x}_i^T$.

Régression linéaire multiple (=Modèle linéaire)

- La fonction de régression est $r(\beta, \boldsymbol{x}_i) = \boldsymbol{x}_i^T \beta$. On observe

$$(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_n, Y_n)$$

  avec

$$\boxed{Y_i = \boldsymbol{x}_i^T \beta + \sigma \xi_i, \quad i = 1, \ldots, n}$$

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

- Matriciellement

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

  avec
  - $\boldsymbol{Y} = (Y_1 \cdots Y_n)^T$,
  - $\boldsymbol{\xi} = (\xi_1 \cdots \xi_n)^T$
  - $\mathbb{X}$ la matrice $(n \times k)$ dont la i-ème ligne est $\mathbb{X}_{i,\cdot} = \boldsymbol{x}_i^T$.

EMC en régression linéaire multiple

- Estimateur des moindres carrés en régression linéaire multiple : tout estimateur $\widehat{\theta}_n^{\,\mathtt{mc}}$ satisfaisant

$$\sum_{i=1}^{n}\left(Y_i - \boldsymbol{x}_i^T\,\widehat{\beta}_n\right)^2 = \min_{\mathbf{b}\in\mathbb{R}^k}\sum_{i=1}^{n}\left(Y_i - \boldsymbol{x}_i^T\,\mathbf{b}\right)^2.$$

- En notation matricielle :

$$\begin{aligned}\|\boldsymbol{Y} - \mathbb{X}\,\widehat{\beta}_n\|^2 &= \min_{\mathbf{b}\in\mathbb{R}^k}\|\,\boldsymbol{Y} - \mathbb{X}\beta\|^2\\ &= \min_{v\in V}\|\boldsymbol{Y} - v\|^2\end{aligned}$$

où $V = \mathrm{Im}(\mathbb{X}) = \{v\in\mathbb{R}^n : v = \mathbb{X}\mathbf{b},\ \mathbf{b}\in\mathbb{R}^k\}$. Projection orthogonale sur V.

Géométrie de l'EMC

- L'EMC vérifie

$$\boxed{\mathbb{X}\,\widehat{\boldsymbol{\beta}}_{\mathrm{n}} = P_V\,\boldsymbol{Y}}$$

où P_V est le projecteur orthogonal sur V.

- Comme $\boldsymbol{Y} - P_V\,\boldsymbol{Y} \perp V$, on en déduit les équations normales des moindres carrés :

$$\boxed{\mathbb{X}^T\mathbb{X}\widehat{\boldsymbol{\beta}}_{\mathrm{n}} = \mathbb{X}^T\,\boldsymbol{Y}.}$$

- Remarques.
  - L'EMC est un Z-estimateur.
  - unicité de $\widehat{\boldsymbol{\beta}}_{\mathrm{n}}$ si la matrice de Gram $\mathbb{X}^T\mathbb{X}$ est inversible (la matrice $\mathbb{X}$ est de rang complet).

Géométrie de l'EMC

Proposition

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

$$\boxed{\widehat{\theta}_n^{\,\mathtt{mc}} = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\boldsymbol{Y}}$$

- Contient le cas précédent de la droite de régression simple.
- Résultat géometrique, non stochastique.
- $\mathbb{X}^T\mathbb{X} \geq 0$; $\mathbb{X}^T\mathbb{X}$ inversible $\Longleftrightarrow \mathbb{X}^T\mathbb{X} > 0$;

$$\mathbb{X}^T\mathbb{X} > 0 \iff \mathrm{rang}(\mathbb{X}) = k \iff \dim(V) = k.$$

$$\mathbb{X}^T\mathbb{X} > 0 \implies n \geq k.$$

Géométrie de l'EMC

Soit $\mathbb{X}^T\mathbb{X} > 0$. Alors, la matrice $n \times n$

$$A = \mathbb{X}\big(\mathbb{X}^T\mathbb{X}\big)^{-1}\mathbb{X}^T$$

est dite matrice chapeau (`hat matrix`).

Proposition

Si $\mathbb{X}^T\mathbb{X} > 0$, alors A est le projecteur sur V : $A = P_V$ et $\mathrm{rang}(A) = k$.

Démonstration.

$A = A^T$, $A = A^2$, donc A est un projecteur. $\mathrm{Im}(A) = V$, donc $A = P_V$; $\mathrm{rang}(P_V) = \dim(V) = k$. $\square$

Géométrie de l'EMC

Soit $\mathbb{X}^T\mathbb{X} > 0$. Alors, la matrice $n \times n$

$$A = \mathbb{X}\big(\mathbb{X}^T\mathbb{X}\big)^{-1}\mathbb{X}^T$$

est dite matrice chapeau (`hat matrix`).

Proposition

Si $\mathbb{X}^T\mathbb{X} > 0$, alors A est le projecteur sur V : $A = P_V$ et $\mathrm{rang}(A) = k$.

Chapeau, car A génère la prévision de $\mathbb{X}\theta$ notée $\widehat{\boldsymbol{Y}}$:

$$\widehat{\boldsymbol{Y}} = \mathbb{X}\,\widehat{\theta}_n^{\,\mathtt{mc}} = A\boldsymbol{Y}.$$

Hypothèses

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \sigma\boldsymbol{\xi}$$

1. $\mathbb{X}$ est de rang complet.
2. $\mathbb{E}_\theta[\boldsymbol{\xi}] = 0$ pour tout $\theta \in \Theta$ (les erreurs sont centrées)
3. La variance des erreurs est constante et les erreurs sont décorrélées $\mathbb{E}_\theta[\boldsymbol{\xi}\boldsymbol{\xi}^T] = \sigma^2 I$ (homoscédasticité)
4. (optionnel) les erreurs sont gaussiennes.

Hypothèses

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \sigma\boldsymbol{\xi}$$

1. $\mathbb{X}$ est de rang complet.
2. $\mathbb{E}_\theta[\boldsymbol{\xi}] = 0$ pour tout $\theta \in \Theta$ (les erreurs sont centrées)
3. La variance des erreurs est constante et les erreurs sont décorrélées $\mathbb{E}_\theta[\boldsymbol{\xi}\boldsymbol{\xi}^T] = \sigma^2 I$ (homoscédasticité)
4. (optionnel) les erreurs sont gaussiennes.

Hypothèses

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \sigma\boldsymbol{\xi}$$

1. $\mathbb{X}$ est de rang complet.
2. $\mathbb{E}_\theta[\boldsymbol{\xi}] = 0$ pour tout $\theta \in \Theta$ (les erreurs sont centrées)
3. La variance des erreurs est constante et les erreurs sont décorrélées $\mathbb{E}_\theta[\boldsymbol{\xi}\boldsymbol{\xi}^T] = \sigma^2 I$ (homoscédasticité)
4. (optionnel) les erreurs sont gaussiennes.

Hypothèses

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \sigma\boldsymbol{\xi}$$

1. $\mathbb{X}$ est de rang complet.
2. $\mathbb{E}_\theta[\boldsymbol{\xi}] = 0$ pour tout $\theta \in \Theta$ (les erreurs sont centrées)
3. La variance des erreurs est constante et les erreurs sont décorrélées $\mathbb{E}_\theta[\boldsymbol{\xi}\boldsymbol{\xi}^T] = \sigma^2 I$ (homoscédasticité)
4. (optionnel) les erreurs sont gaussiennes.

Estimateur sans biais

Théorème

L'estimateur $\widehat{\beta}_n$ est sans biais, i.e. pour tout $\theta \in \Theta$, $\mathbb{E}_\theta[\widehat{\beta}_n] = \theta$.
De plus, $\mathrm{Cov}_\theta(\widehat{\beta}_n) = \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}$.

Estimateur sans biais

Théorème

L'estimateur $\widehat{\boldsymbol{\beta}}_{\mathsf{n}}$ est sans biais, i.e. pour tout $\theta \in \Theta$, $\mathbb{E}_\theta[\widehat{\boldsymbol{\beta}}_{\mathsf{n}}] = \theta$. De plus, $\mathrm{Cov}_\theta(\widehat{\boldsymbol{\beta}}_{\mathsf{n}}) = \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}$.

Démonstration.

$\widehat{\boldsymbol{\beta}}_{\mathsf{n}} = \theta + \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\boldsymbol{\xi}$.

On vérifie : $\mathbb{E}_\theta[\widehat{\boldsymbol{\beta}}_{\mathsf{n}}] = \theta$,

$$\mathbb{E}_\theta\left[\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\boldsymbol{\xi}\left(\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\boldsymbol{\xi}\right)^T\right]$$
$$= \sigma^2\left(\mathbb{X}^T\mathbb{X}\right)^{-1}.$$

$\square$

Erreur de prédiction

- Erreur de prédiction :

$$\hat{\boldsymbol{\xi}} = \boldsymbol{Y} - \mathbb{X}\,\widehat{\boldsymbol{\beta}}_{\mathrm{n}} = \boldsymbol{Y} - \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}\,\boldsymbol{Y}$$
$$= (I - A)\,\boldsymbol{Y}$$

- Sous $\mathbb{P}_\theta$, $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \sigma\boldsymbol{\xi}$. Donc,

$$\hat{\boldsymbol{\xi}} = (I - A)\mathbb{X}\boldsymbol{\beta} + \sigma(I - A)\boldsymbol{\xi}$$
$$= \sigma(I - A)\boldsymbol{\xi}$$

car $A\mathbb{X} = \mathbb{X}$ (A is the orthogonal projector on the image of $\mathbb{X}$).

Résidus et variance résiduelle

Theorem

Pour tout $\theta \in \Theta$

1. $\mathbb{E}_\theta[\hat{\xi}] = 0.$
2. $\mathrm{Cov}_\theta(\hat{\xi}) = \sigma^2(I - A).$
3. $\mathbb{E}_\theta[\hat{\mathbf{Y}}] = \mathbb{X}\beta.$
4. $\mathrm{Cov}_\theta(\hat{\xi}, \hat{\mathbf{Y}}) = 0.$

Démonstration.

$$\begin{aligned}\mathbb{E}_\theta[\hat{\xi}] &= \sigma\,\mathbb{E}_\theta[(I - A)\xi] \\ &= \sigma(I - A)\,\mathbb{E}_\theta[\xi]\,.\end{aligned}$$

Résidus et variance résiduelle

Theorem

Pour tout $\theta \in \Theta$

1. $\mathbb{E}_\theta[\hat{\xi}] = 0$.
2. $\mathrm{Cov}_\theta(\hat{\xi}) = \sigma^2(I - A)$.
3. $\mathbb{E}_\theta[\hat{\mathbf{Y}}] = \mathbb{X}\beta$.
4. $\mathrm{Cov}_\theta(\hat{\xi}, \hat{\mathbf{Y}}) = 0$.

Démonstration.

$$\begin{aligned}\mathrm{Cov}_\theta(\hat{\xi}) &= \sigma^2(I - A)\,\mathbb{E}_\theta[\xi\xi'](I - A)\\ &= \sigma^2(I - A).\end{aligned}$$

Résidus et variance résiduelle

Theorem

Pour tout $\theta \in \Theta$

1. $\mathbb{E}_\theta[\hat{\boldsymbol{\xi}}] = 0$.
2. $\mathrm{Cov}_\theta(\hat{\boldsymbol{\xi}}) = \sigma^2(I - A)$.
3. $\mathbb{E}_\theta[\hat{\mathbf{Y}}] = \mathbb{X}\boldsymbol{\beta}$.
4. $\mathrm{Cov}_\theta(\hat{\boldsymbol{\xi}}, \hat{\mathbf{Y}}) = 0$.

Démonstration.

$$\begin{aligned}\mathbb{E}_\theta[\hat{\mathbf{Y}}] &= \mathbb{E}_\theta[A(\mathbb{X}\boldsymbol{\beta} + \sigma\boldsymbol{\xi})] \\ &= A\mathbb{X}\boldsymbol{\beta} + \sigma\,\mathbb{E}_\theta[\boldsymbol{\xi}] \\ &= \mathbb{X}\boldsymbol{\beta}.\end{aligned}$$

Résidus et variance résiduelle

Theorem

Pour tout $\theta \in \Theta$

1. $\mathbb{E}_\theta[\hat{\xi}] = 0$.
2. $\mathrm{Cov}_\theta(\hat{\xi}) = \sigma^2(I - A)$.
3. $\mathbb{E}_\theta[\hat{\mathbf{Y}}] = \mathbb{X}\beta$.
4. $\mathrm{Cov}_\theta(\hat{\xi}, \hat{\mathbf{Y}}) = 0$.

Démonstration.

On a $\hat{\mathbf{Y}} - \mathbb{E}_\theta[\hat{\mathbf{Y}}] = \sigma A \xi$ et donc

$$\begin{aligned}\mathrm{Cov}_\theta(\hat{\xi}, \hat{\mathbf{Y}}) &= \sigma^2\, \mathbb{E}_\theta[(I - A)\xi\xi' A] \\ &= \sigma^2(I - A)A = 0.\end{aligned}$$

Estimateur sans bians de la variance de l'erreur de prédiction

Théorème

$\hat{\sigma}_n^2 = (n-p)^{-1}\|\hat{\boldsymbol{\xi}}\|^2$ *est un estimateur sans biais de la variance de l'erreur.*

Démonstration.

Comme $(I-A)^2 = (I-A)$, nous avons

$$\begin{aligned}
\mathbb{E}_\theta[\hat{\sigma}_n^2] &= (n-p)^{-1}\,\mathbb{E}_\theta[\boldsymbol{\xi}^T(I-A)\boldsymbol{\xi}] \\
&= (n-p)^{-1}\,\mathbb{E}[\mathrm{Tr}((I-A)\boldsymbol{\xi}\boldsymbol{\xi})] \\
&= \sigma^2(n-p)^{-1}\mathrm{Tr}(I-A) = \sigma^2\,.
\end{aligned}$$

$\square$

Prevision

- Un des buts de la régression est de proposer des prédictions pour la réponse lorsque nous avons de nouvelles valeurs des variables explicatives.
- Un prédicteur naturel de la réponse est $\hat{Y}_n(\boldsymbol{x}) = \boldsymbol{x}^T \widehat{\beta}_{\mathrm{n}}$

Moyenne et variance de la prévision

Theorem

- $\mathbb{E}_\theta[\hat{Y}_n(\boldsymbol{x})] = \boldsymbol{x}^T \beta$
- $\mathrm{Var}_\theta(\hat{Y}_n(\boldsymbol{x})) = \sigma^2 \boldsymbol{x}^T (\mathbb{X}^T \mathbb{X})^{-1} \boldsymbol{x}$
- $\mathbb{E}_\theta[(Y(\boldsymbol{x}) - \hat{Y}_n(\boldsymbol{x}))^2] = \sigma^2 (1 + \boldsymbol{x}' ({}^T \mathbb{X})^{-1} \boldsymbol{x})$

$\hat{Y}_n(\boldsymbol{x}) = \boldsymbol{x}^T \widehat{\beta}_{\mathrm{n}}$ et $\mathbb{E}_\theta[\widehat{\beta}_{\mathrm{n}}] = \beta$

Moyenne et variance de la prévision

Theorem

- $\mathbb{E}_\theta[\hat{Y}_n(\boldsymbol{x})] = \boldsymbol{x}^T \boldsymbol{\beta}$
- $\mathrm{Var}_\theta(\hat{Y}_n(\boldsymbol{x})) = \sigma^2 \boldsymbol{x}^T(\mathbb{X}^T\mathbb{X})^{-1} \boldsymbol{x}$
- $\mathbb{E}_\theta[(Y(\boldsymbol{x}) - \hat{Y}_n(\boldsymbol{x}))^2] = \sigma^2(1 + \boldsymbol{x}'(^T\mathbb{X})^{-1} \boldsymbol{x})$

$$
\begin{aligned}
\hat{Y}_n(\boldsymbol{x}) - \boldsymbol{x}^T \boldsymbol{\beta} &= \boldsymbol{x}^T \mathbb{X}^T\mathbb{X}^{-1}\mathbb{X}^T \boldsymbol{Y} - \boldsymbol{x}^T \boldsymbol{\beta} \\
&= \boldsymbol{x}^T \mathbb{X}^T\mathbb{X}^{-1}\mathbb{X}^T(\mathbb{X}\boldsymbol{\beta} + \sigma\boldsymbol{\xi}) - \boldsymbol{x}^T \boldsymbol{\beta} \\
&= \sigma\, \boldsymbol{x}^T \mathbb{X}^T\mathbb{X}^{-1}\mathbb{X}^T\boldsymbol{\xi}
\end{aligned}
$$

Moyenne et variance de la prévision

Theorem

- $\mathbb{E}_\theta[\hat{Y}_n(\boldsymbol{x})] = \boldsymbol{x}^T \boldsymbol{\beta}$
- $\mathrm{Var}_\theta(\hat{Y}_n(\boldsymbol{x})) = \sigma^2 \boldsymbol{x}^T(\mathbb{X}^T\mathbb{X})^{-1} \boldsymbol{x}$
- $\mathbb{E}_\theta[(Y(\boldsymbol{x}) - \hat{Y}_n(\boldsymbol{x}))^2] = \sigma^2(1 + \boldsymbol{x}'({}^T\mathbb{X})^{-1} \boldsymbol{x})$

$$
\begin{aligned}
\mathbb{E}_\theta[(Y(\boldsymbol{x}) - \hat{Y}_n(\boldsymbol{x}))^2] &= \mathbb{E}_{\boldsymbol{\beta}}[(Y(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{\beta}}[\hat{Y}_n(\boldsymbol{x})])^2] + \mathrm{Var}_\theta(\hat{Y}_n(\boldsymbol{x})) \\
&= \mathbb{E}_{\boldsymbol{\beta}}[(Y(\boldsymbol{x}) - \boldsymbol{x}^T \boldsymbol{\beta})^2] + \sigma^2 \boldsymbol{x}^T({}^T\mathbb{X})^{-1} \boldsymbol{x}
\end{aligned}
$$

Diagnostic de régression

```
> summary(model1)

Call:
lm(formula = endur$endurance ~ endur$age)

Residuals:
     Min       1Q   Median       3Q      Max
-25.0734  -7.6331   0.0974   6.7710  30.8696

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.15667    3.42033   9.694   <2e-16 ***
endur$age   -0.13472    0.06812  -1.978   0.0491 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.76 on 243 degrees of freedom
Multiple R-squared:  0.01584, Adjusted R-squared:  0.01179
F-statistic: 3.911 on 1 and 243 DF,  p-value: 0.04911
```

FIGURE – Régression à un facteur : endurance / âge

Diagnostic de régression

```
> summary(model2)

Call:
lm(formula = endur$endurance ~ endur$activeyears)

Residuals:
     Min       1Q   Median       3Q      Max 
-23.7296  -7.0671   0.5579   5.7454  31.0829 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        18.3921     1.5998  11.496  < 2e-16 ***
endur$activeyears   0.7625     0.1369   5.571  6.7e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.21 on 243 degrees of freedom
Multiple R-squared:  0.1133,	Adjusted R-squared:  0.1096 
F-statistic: 31.04 on 1 and 243 DF,  p-value: 6.697e-08
```

FIGURE – Régression à un facteur : endurance / nombre d'années de pratique

Diagnostic de régression

```
> summary(model3)

Call:
lm(formula = endur$endurance ~ endur$age + endur$activeyears)

Residuals:
     Min       1Q   Median       3Q      Max
-21.7994  -6.9040   0.5701   5.6326  27.2279

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        29.3952     3.2054   9.171  < 2e-16 ***
endur$age          -0.2571     0.0655  -3.925 0.000113 ***
endur$activeyears   0.9163     0.1386   6.610 2.44e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.919 on 242 degrees of freedom
Multiple R-squared:  0.1663, Adjusted R-squared:  0.1594
F-statistic: 24.14 on 2 and 242 DF,  p-value: 2.754e-10
```

FIGURE – Régression à un deux facteurs : endurance / âge + nombre d'années de pratique

Régression gaussienne

Régression gaussienne : on suppose $\xi \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_n)$. Alors on a plusieurs proriétés remarquables :

- Estimateur des moindres carrés $\widehat{\theta}_n^{\,\mathtt{mc}}$ et estimateur du maximum de vraisemblance coïncident.

  Preuve : écriture de la fonction de vraisemblance.
- On sait expliciter la loi exacte (non-asymptotique !) de $\widehat{\theta}_n^{\,\mathtt{mc}}$.

  Ingrédient : loi des vecteurs gaussiens sont caractérisés par leur moyenne et matrice de variance-covariance.

Cadre gaussien : loi des estimateurs

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

Proposition

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

(ii) $\| \boldsymbol{Y} - \mathbb{X}\,\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}} \|^2 \sim \sigma^2 \chi^2(n-k)$ *loi du Chi 2 à $n-k$ degrés de liberté*

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

- Preuve : Thm. de Cochran (Poly, page 18). Si $\xi \sim \mathcal{N}(0, \mathrm{Id}_n)$ et A_j matrices $n \times n$ projecteurs t.q. $A_j A_i = 0$ pour $i \neq j$, alors : $A_j\,\xi \sim \mathcal{N}\big(0, A_j\big)$, indépendants, $\|A_j \xi\|^2 \sim \chi^2(\mathrm{Rang}(A_j))$.

Preuve de la proposition

- (ii)

$$\begin{aligned}\boldsymbol{Y} - \mathbb{X}\widehat{\theta}_n^{\,\mathrm{mc}} &= \mathbb{X}\big(\theta - \widehat{\theta}_n^{\,\mathrm{mc}}\big) + \boldsymbol{\xi} \\ &= -\mathbb{X}\big(\mathbb{X}^T\mathbb{X}\big)^{-1}\mathbb{X}^T\boldsymbol{\xi} + \boldsymbol{\xi} \\ &= \sigma(\mathrm{Id}_n - A)\boldsymbol{\xi}',\ \boldsymbol{\xi}' \sim \mathcal{N}(0, \mathrm{Id}_n).\end{aligned}$$

- (iii) le vecteur $(\widehat{\theta}_n^{\,\mathrm{mc}}, \boldsymbol{Y} - \mathbb{X}\widehat{\theta}_n^{\,\mathrm{mc}})$ est gaussien. On calcule explicitement sa matrice de variance-covariance.

Propriétés de l'EMC : cadre gaussien

Estimateur de la variance σ^2 :

$$\boxed{\widehat{\sigma}_n^2 = \frac{\|\boldsymbol{Y} - \mathbb{X}\,\widehat{\theta}_{\mathrm{n}}^{\,\mathtt{mc}}\|^2}{n-k} = \frac{1}{n-k}\sum_{i=1}^{n}\left(Y_i - \left(\widehat{\theta}_{\mathrm{n}}^{\,\mathtt{mc}}\right)^T \boldsymbol{x}_i\right)^2}$$

D'après la dernière Proposition :

- $\widehat{\sigma}_n^2/\sigma^2 \sim \chi^2(n-k)$ loi du Chi 2 à $n-k$ degrés de liberté
- C'est un estimateur sans biais :

$$\mathbb{E}_\theta\left[\widehat{\sigma}_n^2\right] = \sigma^2.$$

- $\widehat{\sigma}_n^2$ est indépendant de $\widehat{\theta}_{\mathrm{n}}^{\,\mathtt{mc}}$.

Propriétés de l'EMC : cadre gaussien

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

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

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

$$\frac{(\widehat{\theta}_{\mathrm{n}}^{\,\mathtt{mc}})_j - \theta_j}{\widehat{\sigma}_n\sqrt{b_j}} \sim t_{n-k}$$

loi de Student à $n-k$ degrés de liberté.

$$t_q = \frac{\xi}{\sqrt{\eta/q}}$$

où $q \geq 1$ un entier, $\xi \sim \mathcal{N}\big(0,1\big)$, $\eta \sim \chi^2(q)$ et ξ indépendant de η.

Exemple de données de régression

Résultats de traitement statistique initial

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	152.133	2.576	59.061	$< 2e-16 ***$
age	-10.012	59.749	-0.168	0.867000
sex	-239.819	61.222	-3.917	$0.000104 ***$
bmi	519.840	66.534	7.813	$4.30e-14 ***$
map	324.390	65.422	4.958	$1.02e-06 ***$
tc	-792.184	416.684	-1.901	0.057947
ldl	476.746	339.035	1.406	0.160389
hdl	101.045	212.533	0.475	0.634721
tch	177.064	161.476	1.097	0.273456
ltg	751.279	171.902	4.370	$1.56e-05 ***$
glu	67.625	65.984	1.025	0.305998

Questions statistiques

- Sélection de variables. Lesquelles parmi les 10 variables :

  `age,sex,bmi,map,tc,ldl,hdl,tch,ltg,glu`

  sont significatives ? Formalisation mathématique : trouver (estimer) l'ensemble $N = \{j : \theta_j \neq 0\}$.
- Prévison. Un nouveau patient arrive avec son vecteur des 10 variables $x_0 \in \mathbb{R}^{10}$. Donner la prévison de la réponse Y =état du patient dans 1 an.

RSS (Residual Sum of Squares)

Modèle de régression

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

- Résidu : si $\widehat{\theta}_n$ est un estimateur de θ,

$$\widehat{\xi}_i = Y_i - r(\widehat{\theta}_n, x_i) \text{ résidu au point } i.$$

- RSS : Residual Sum of Squares, somme résiduelle des carrés. Caractérise la qualité d'approximation.

$$\mathrm{RSS}(= \mathrm{RSS}_{\widehat{\theta}_n}) = \|\widehat{\xi}\|^2 = \sum_{i=1}^{n} \big(Y_i - r(\widehat{\theta}_n, x_i)\big)^2.$$

- En régression linéaire : $\boxed{\mathrm{RSS} = \| Y - \mathbb{X}\,\widehat{\theta}_n \|^2.}$

Sélection rétrograde

- On se donne un critère d'élimination de variables (plusieurs choix de critère possibles...).
- On élimine une variable, la moins significative du point de vue du critère choisi.
- On calcule l'EMC $\widehat{\theta}^{\mathrm{mc}}_{n,k-1}$ dans le nouveau modèle, avec seulement les $k-1$ paramétres restants, ainsi que le RSS :
$$\mathrm{RSS}_{k-1} = \|\mathbf{Y} - \mathbb{X}\widehat{\theta}^{\mathrm{mc}}_{n,k-1}\|^2.$$
- On continue à éliminer des variables, une par une, jusqu'à la stabilisation de RSS : $\mathrm{RSS}_m \approx \mathrm{RSS}_{m-1}$.

Données de diabète : sélection rétrograde

- Sélection "naïve" : {`sex,bmi,map,ltg`}
- Sélection par Backward Regression :
  Critère d'élimination : plus grande valeur de $\Pr(>|t|)$.

	Estimate	Std. Error	t value	$\Pr(>\lvert t\rvert)$
(Intercept)	152.133	2.576	59.061	$< 2e-16 ***$
age	-10.012	59.749	-0.168	0.867000
sex	-239.819	61.222	-3.917	$0.000104 ***$
bmi	519.840	66.534	7.813	$4.30e-14 ***$
map	324.390	65.422	4.958	$1.02e-06 ***$
tc	-792.184	416.684	-1.901	0.057947
ldl	476.746	339.035	1.406	0.160389
hdl	101.045	212.533	0.475	0.634721
tch	177.064	161.476	1.097	0.273456
ltg	751.279	171.902	4.370	$1.56e-05 ***$
glu	67.625	65.984	1.025	0.305998

Données de diabète : Backward Regression

Backward Regression : Itération 2.

Critère d'élimination : plus grande valeur de $\Pr(>|t|)$.

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	152.133	2.573	59.128	$< 2e-16$
sex	-240.835	60.853	-3.958	0.000104
bmi	519.905	64.156	5.024	$8.85e-05$
map	322.306	65.422	4.958	$7.43e-07$
tc	-790.896	416.144	-1.901	0.058
ldl	474.377	338.358	1.402	0.162
hdl	99.718	212.146	0.470	0.639
tch	177.458	161.277	1.100	0.272
ltg	749.506	171.383	4.373	$1.54e-05$
glu	67.170	65.336	1.013	0.312

Données de diabète : Backward Regression

Backward Regression : Itération 5 (dernière).

Variables sélectionnées :

$$\{\texttt{sex},\texttt{bmi},\texttt{map},\texttt{tc},\texttt{ldl},\texttt{ltg}\}$$

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	152.133	2.572	59.159	$< 2e-16$
sex	-226.511	59.857	-3.784	0.000176
bmi	529.873	65.620	8.075	$6.69e-15$
map	327.220	62.693	5.219	$2.79e-07$
tc	-757.938	160.435	-4.724	$3.12e-06$
ldl	538.586	146.738	3.670	0.000272
ltg	804.192	80.173	10.031	$< 2e-16$

Sélection de variables : Backward Regression

Discussion de `Backward Regression` :

- Méthode de sélection purement empirique, pas de justification théorique.
- Application d'autres critères d'élimination en `Backward Regression` peut amener aux résultats différents.
  Exemple. Critère C_p de Mallows–Akaike : on élimine la variable j qui réalise

$$\min_j \left(\mathrm{RSS}_{m,(-j)} + 2\widehat{\sigma}_n^2 m \right).$$

Sélection de variables : LASSO

LASSO = Least Absolute Shrinkage and Selection Operator

- Estimateur LASSO : tout estimateur $\widehat{\theta}_n^L$ vérifiant

$$\widehat{\theta}_n^L \in \arg\min_{\theta \in \mathbb{R}^k} \left(\sum_{i=1}^n \left(Y_i - \theta^T \boldsymbol{x}_i \right)^2 + \lambda \sum_{j=1}^k |\theta_j| \right) \quad \text{avec } \lambda > 0.$$

- Si $\mathbb{X}^T\mathbb{X} > 0$, l'estimateur LASSO $\widehat{\theta}_n^L$ est unique.
- Estimateur des moindres carrés pénalisé. Pénalisation par $\sum_{j=1}^k |\theta_j|$, la norme ℓ_1 de θ.

Sélection de variables : LASSO

- Deux utilisations de LASSO :
  - Estimation de θ : alternative à $\widehat{\theta}_{\mathrm{n}}^{\mathtt{mc}}$ si $k > n$.
  - Sélection de variables : on ne retient que les variables qui correspondent aux coordonnées non-nulles du vecteur $\widehat{\theta}_n^L$.
- LASSO admet une justification théorique : sous certaines hypothèses sur la matrice $\mathbb{X}$,

$$\lim_{n\to\infty} \mathbb{P}\{\widehat{N}_n = N\} = 1,$$

où $N = \{j : \theta_j \neq 0\}$ et $\widehat{N}_n = \{j : \widehat{\theta}_{n,j}^L \neq 0\}$.

Application de LASSO : "regularization path"

Données de diabète : LASSO

Application aux données de diabète.

- L'ensemble de variables sélectionné par LASSO :

$$\{\texttt{sex},\texttt{bmi},\texttt{map},\texttt{tc},\texttt{hdl},\texttt{ltg},\texttt{glu}\}$$

- `Backward Regression` :

$$\{\texttt{sex},\texttt{bmi},\texttt{map},\texttt{tc},\texttt{ldl},\texttt{ltg}\}$$

- Sélection naïve :

$$\{\texttt{sex},\texttt{bmi},\texttt{map},\texttt{tc}\}$$

Prévision

Modèle de régression
$$Y_i = r(\theta, x_i) + \xi_i, \quad i = 1, \ldots, n.$$

Régression linéaire : $r(\theta, x_i) = \theta^T \mathbf{x}_i$. Exemple : x_i vecteur de 10 variables explicatives (`age`,`sex`,`bmi`,...) pour patient i.

- Problème de prévision : Un nouveau patient arrive avec son vecteur des 10 variables $\mathbf{x}_0 \in \mathbb{R}^{10}$. Donner la prévison de la valeur de fonction de régression $r(\theta, x_0) = \theta^T x_0$ (=état du patient dans 1 an).
- Soit $\widehat{\theta}_n$ un estimateur de θ. Prévision par substitution :
$$\boxed{\widehat{Y} = r(\widehat{\theta}_n, x_0).}$$
- Question statistique : quelle est la qualité de la prévision ? Intervalle de confiance pour $r(\theta, x_0)$ basé sur $\widehat{Y}$?

Prévision : modèle linéaire gaussienne

- Traitement sur l'exemple : $r(\theta, x) = \theta^T \mathbf{x}$, régression linéaire gaussienne et $\widehat{\theta}_n = \widehat{\theta}_n^{\,\mathtt{mc}}$. $\Longrightarrow$ $\boxed{\widehat{Y} = \mathbf{x}_0^T\, \widehat{\theta}_n^{\,\mathtt{mc}}}$
- Hyp. 1 : $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_n)$.
- Hyp. 2 : $\mathbb{X}^T\mathbb{X} > 0$.

Proposition

(i) $\widehat{Y} \sim \mathcal{N}\left(\mathbf{x}_0^T\theta, \sigma^2 \mathbf{x}_0^T \left(\mathbb{X}^T\mathbb{X}\right)^{-1} x_0\right)$

(ii) $\widehat{Y} - \mathbf{x}_0^T\theta$ *et* $\boldsymbol{Y} - \mathbb{X}\,\widehat{\theta}_n^{\,\mathtt{mc}}$ *sont indépendants.*

Rappel : $\|\boldsymbol{Y} - \mathbb{X}\,\widehat{\theta}_n^{\,\mathtt{mc}}\|^2 \sim \sigma^2\chi^2(n-k)$ loi du Chi 2 à $n-k$ degrés de liberté.

Prévision : modèle linéaire gaussienne

- D'après la Proposition,

$$\eta := \frac{\widehat{Y} - x_0^T\theta}{\sqrt{\sigma^2 \mathbf{x}_0^T\left(\mathbb{X}^T\mathbb{X}\right)^{-1}x_0}} \sim \mathcal{N}(0,1).$$

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

$$t := \frac{\widehat{Y} - x_0^T\theta}{\sqrt{\widehat{\sigma}_n^2 x_0^T\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbf{x}_0}} = \frac{\eta}{\sqrt{\chi/(n-k)}} \sim t_{n-k},$$

loi de Student à $n-k$ degrés de liberté, car $\eta \sim \mathcal{N}(0,1)$, $\chi := \|\boldsymbol{Y} - \mathbb{X}\,\widehat{\theta}_\mathrm{n}^{\,\mathtt{mc}}\|^2/\sigma^2 \sim \chi^2(n-k)$ et $\eta\chi$.

Prévision : intervalle de confiance

$$\mathbb{P}\left(-q_{1-\frac{\alpha}{2}}(t_{n-k}) \leq \frac{\widehat{Y} - x_0^T\theta}{\sqrt{\widehat{\sigma}_n^2 \mathbf{x}_0^T \left(\mathbb{X}^T\mathbb{X}\right)^{-1} x_0}} \leq q_{1-\frac{\alpha}{2}}(t_{n-k})\right)$$

$$= \mathbb{P}(-q_{1-\frac{\alpha}{2}}(t_{n-k}) \leq t \leq q_{1-\frac{\alpha}{2}}(t_{n-k})) = 1-\alpha.$$

$\Longrightarrow$ intervalle de confiance de niveau $1-\alpha$ pour $r(\theta, x_0) = x_0^T\theta$ est $[r_L, r_U]$, où :

$$\begin{aligned} r_L &= \widehat{Y} - q_{1-\frac{\alpha}{2}}(t_{n-k})\sqrt{\widehat{\sigma}_n^2 x_0^T \left(\mathbb{X}^T\mathbb{X}\right)^{-1} x_0}, \\ r_U &= \widehat{Y} + q_{1-\frac{\alpha}{2}}(t_{n-k})\sqrt{\widehat{\sigma}_n^2 \mathbf{x}_0^T \left(\mathbb{X}^T\mathbb{X}\right)^{-1} x_0}. \end{aligned}$$

Limites des moindres carrés et du cadre gaussien

- Calcul explicite (et efficace) de l'EMC limité à une fonction de régression linéaire.
- Modèle linéaire donne un cadre assez général :
  - Modèle polynomial,
  - Modèles avec interactions...
- Hypothèse de gaussianité = cadre asymptotique implicite.
- Besoin d'outils pour les modèles à réponse Y discrète.

Régression linéaire non-gaussienne

Modèle de régression linéaire

$$Y_i = \theta^T x_i + \xi_i, \quad i = 1, \ldots, n.$$

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

Proposition (Normalité asymptotique de l'EMC)

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

- A comparer avec le cadre gaussien :

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

Régression non-linéaire

- On observe

$$(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_n, Y_n),$$

où

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

avec

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

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

$$\mathcal{L}_n(\theta, Y_1, \ldots, Y_n) \propto \exp\Big(- \frac{1}{2\sigma^2} \sum_{i=1}^{n} \big(Y_i - r(\theta, \boldsymbol{x}_i)\big)^2 \Big)$$

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

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

Moindre carrés non-linéaires

Definition

- *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*

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

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

- Extension des résultats en densité → théorèmes limites pour des sommes de v.a. indépendantes non-équidistribuées.

Modèle à réponse binaire

- On observe

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

- Modélisation via la fonction de régression

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

- Représentation

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

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

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

Modèle à réponse discrète

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

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

  $\rightarrow$ méthodes de résolution numérique.

- Régression logistique (très utile dans les applications)

$$p_{\boldsymbol{x}}(\theta) = \psi(\boldsymbol{x}^T\,\theta),$$

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

Régression logistique et modèles latents

- Représentation équivalente de la régression logistique : on observe

$$\boxed{Y_i = 1_{\left\{Y_i^\star > 0\right\}},\quad i = 1, \ldots, n}$$

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

$$\boxed{Y_i^\star = \theta^T \boldsymbol{x}_i + U_i,\quad i = 1, \ldots, n}$$

avec $U_i \sim_{\text{i.i.d.}} F$, où

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

-

$$\begin{aligned}
\mathbb{P}_\theta\left[Y_i^\star > 0\right] &= \mathbb{P}_\theta\left[\boldsymbol{x}_i^T\theta + U_i > 0\right] \\
&= 1 - \mathbb{P}_\theta\left[U_i \leq -\boldsymbol{x}_i^T\theta\right]
\end{aligned}$$

Bilan provisoire : modèles paramétriques dominés

- Modèle de densité : on observe

$$X_1, \ldots, X_n \sim_{\text{i.i.d ?}} \mathbb{P}_\theta, \quad \theta \in \Theta \subset \mathbb{R}^d.$$

Estimateurs : moments, Z- et M-estimateurs, EMV.

- Modèle de régression : on observe

$$Y_i = r(\theta, \boldsymbol{x}_i) + \xi_i, \quad i = 1, \ldots, n, \quad \xi_i \text{ i.i.d.}, \ \theta \in \Theta \subset \mathbb{R}^d.$$

Estimateurs :

  - Si $r(\theta, \boldsymbol{x}) = \theta^T \boldsymbol{x}$, EMC (coïncide avec l'EMV si les ξ_i gaussiens)
  - Sinon, M-estimateurs, EMV...
  - Autres méthodes selon des hypothèses sur le « design »...

Bilan provisoire (cont.) : précision d'estimation

$\widehat{\theta}_n$ estimateur de θ : précision, qualité de $\widehat{\theta}_n$? Approche par région-intervalle de confiance

- Pour $\alpha \in (0,1)$, on construit $\mathcal{C}_{n,\alpha}(\widehat{\theta}_n)$ ne dépendant pas de θ (observable) tel que

$$\mathbb{P}_\theta\left[\theta \in \mathcal{C}_{n,\alpha}(\widehat{\theta}_n)\right] \geq 1-\alpha$$

  asymptotiquement lorsque $n \to \infty$, uniformément en θ... La précision de l'estimateur est le diamètre (moyen) de $\mathcal{C}_{n,\alpha}(\widehat{\theta}_n)$.

- Par exemple : $\mathcal{C}_{n,\alpha}(\widehat{\theta}_n) =$ boule de centre $\widehat{\theta}_n$ et de rayon à déterminer.

En pratique, une information non-asymptotique de type

$$\mathbb{E}\left[\|\,\widehat{\theta}_n - \theta\|^2\right] \leq c_n(\theta)^2,$$

ou bien asymptotique de type

$$v_n(\widehat{\theta}_n - \theta) \xrightarrow{d} Z_\theta, \quad n \to \infty$$

(avec $v_n \to \infty$) permet « souvent » de construire un(e) région-intervalle de confiance.