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
   - La droite des moindres carrés
   - Régression linéaire multiple
   - Propriétés de l'estimateur des Moindres Carrés
   - Modèle linéaire gaussien
2. Prévision
3. Régression non-linéaire

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 $\boldsymbol{X}_i \in \mathbb{R}^k$, les variables explicatives ou régresseurs.
- On suspecte les variables explicatives

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

d'expliquer une part significative de la variabilité des observations 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).

régresseurs aléatoires

Definition

Données : $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$ *avec* $(Y_i, \boldsymbol{X}_i) \in \mathbb{R} \times \mathbb{R}^k$ *i.i.d., et*

$$Y_i = r(\beta, \boldsymbol{X}_i) + \sigma\xi_i, \quad \mathbb{E}_\theta\left[\xi_i \middle| \boldsymbol{X}_i\right] = 0$$

- $\boldsymbol{x} \rightsquigarrow r(\beta, \boldsymbol{x})$ *fonction de régression, connue au paramètre* β *près ;*
- $\boldsymbol{X}_i$ = *variables explicatives, co-variables, prédicteurs ;*
- $\theta = (\beta, \sigma^2) \in \mathbb{R}^d \times \mathbb{R}_+^*$ *paramètres.*

régresseurs aléatoires

Definition

Données : $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$ *avec* $(Y_i, \boldsymbol{X}_i) \in \mathbb{R} \times \mathbb{R}^k$ *i.i.d., et*

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

- $\boldsymbol{x} \rightsquigarrow r(\beta, \boldsymbol{x})$ *fonction de régression, connue au paramètre* β *près ;*
- $\boldsymbol{X}_i$ *= variables explicatives, co-variables, prédicteurs ;*
- $\theta = (\beta, \sigma^2) \in \mathbb{R}^d \times \mathbb{R}_+^*$ *paramètres.*

régresseurs aléatoires

Definition

Données : $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$ *avec* $(Y_i, \boldsymbol{X}_i) \in \mathbb{R} \times \mathbb{R}^k$ *i.i.d., et*

$$Y_i = r(\beta, \boldsymbol{X}_i) + \sigma\xi_i, \quad \mathbb{E}_\theta\left[\xi_i \middle| \boldsymbol{X}_i\right] = 0$$

- $\boldsymbol{x} \rightsquigarrow r(\beta, \boldsymbol{x})$ *fonction de régression, connue au paramètre* β *près ;*
- $\boldsymbol{X}_i$ = *variables explicatives, co-variables, prédicteurs ;*
- $\theta = (\beta, \sigma^2) \in \mathbb{R}^d \times \mathbb{R}_+^*$ *paramètres.*

Modèle de régression à design déterministe

Definition

Données : $(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_n, Y_n)$ *avec* $Y_i \in \mathbb{R}, \boldsymbol{x}_i \in \mathbb{R}^k$, *et*

$$Y_i = r(\boldsymbol{\beta}, \boldsymbol{x}_i) + \sigma\xi_i, \quad \mathbb{E}_\theta\left[\xi_i\right] = 0,$$

- $\boldsymbol{x}_i$ *déterministes, donnés (ou choisis) : plan d'expérience.*
- *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.*
- $\theta = (\boldsymbol{\beta}, \sigma^2) \in \mathbb{R}^d \times \mathbb{R}_+$.

Question : Comment estimer $\theta = (\boldsymbol{\beta}, \sigma)$ dans ce modèle?

Modèle de régression à design déterministe

Definition

Données : $(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_n, Y_n)$ *avec* $Y_i \in \mathbb{R}, \boldsymbol{x}_i \in \mathbb{R}^k$, *et*

$$Y_i = r(\boldsymbol{\beta}, \boldsymbol{x}_i) + \sigma\xi_i, \quad \mathbb{E}_\theta\left[\xi_i\right] = 0,$$

- $\boldsymbol{x}_i$ *déterministes, donnés (ou choisis) : plan d'expérience.*
- *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.*
- $\theta = (\boldsymbol{\beta}, \sigma^2) \in \mathbb{R}^d \times \mathbb{R}_+$.

Question : Comment estimer $\theta = (\boldsymbol{\beta}, \sigma)$ dans ce modèle?

Modèle de régression à design déterministe

Definition

Données : $(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_n, Y_n)$ *avec* $Y_i \in \mathbb{R}, \boldsymbol{x}_i \in \mathbb{R}^k$, *et*

$$Y_i = r(\boldsymbol{\beta}, \boldsymbol{x}_i) + \sigma\xi_i, \quad \mathbb{E}_\theta\left[\xi_i\right] = 0,$$

- $\boldsymbol{x}_i$ *déterministes, donnés (ou choisis) : plan d'expérience.*
- *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.*
- $\theta = (\boldsymbol{\beta}, \sigma^2) \in \mathbb{R}^d \times \mathbb{R}_+$.

Question : Comment estimer $\theta = (\boldsymbol{\beta}, \sigma)$ dans ce modèle?

Modèle de régression à design déterministe

Definition

Données : $(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_n, Y_n)$ *avec* $Y_i \in \mathbb{R}, \boldsymbol{x}_i \in \mathbb{R}^k$, *et*

$$Y_i = r(\boldsymbol{\beta}, \boldsymbol{x}_i) + \sigma\xi_i, \quad \mathbb{E}_\theta\left[\xi_i\right] = 0,$$

- $\boldsymbol{x}_i$ *déterministes, donnés (ou choisis) : plan d'expérience.*
- *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.*
- $\theta = (\boldsymbol{\beta}, \sigma^2) \in \mathbb{R}^d \times \mathbb{R}_+$.

Question : Comment estimer $\theta = (\boldsymbol{\beta}, \sigma)$ dans ce modèle?

Régression gaussienne

- Modèle de régression à design déterministe :

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

- 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 $\{\mathbb{P}_\theta^n = \text{loi de } (Y_1, \ldots, Y_n), \theta = (\boldsymbol{\beta}, \sigma^2) \in \mathbb{R}^k \times \mathbb{R}_+^*\}$ est dominé par $\mu^n(dy_1 \ldots dy_n) = dy_1 \ldots dy_n$.
- D'où

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

- La fonction de vraisemblance

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

Estimateur des moindres carrés

Maximiser la vraisemblance en régression gaussienne

$$\widehat{\beta}_{\mathrm{n}} \in \operatorname{argmin}_{b\in\mathbb{R}^k} \sum_{i=1}^{n} \big(Y_i - r(b, \boldsymbol{x}_i)\big)^2$$

$$\hat{\sigma}_n^2 = n^{-1} \sum_{i=1}^{n} (Y_i - r(\widehat{\beta}_{\mathrm{n}}, \boldsymbol{x}_i))^2$$

- L'estimateur $\widehat{\beta}_{\mathrm{n}}$ est appelé l'estimateur des moindres carrés. Il peut être appliqué même dans un cas non gaussien.
- Existence, unicité.

Droite de régression

- Modèle le plus simple $\boxed{r(\boldsymbol{\beta},x)=\boldsymbol{\beta}_0+\boldsymbol{\beta}_1 x}$

$$\boxed{Y_i=\boldsymbol{\beta}_0+\boldsymbol{\beta}_1 x_i+\xi_i,\quad i=1,\ldots,n}$$

avec $\boldsymbol{\beta}=(\boldsymbol{\beta}_0,\boldsymbol{\beta}_1)^T\in\mathbb{R}^2$ et les $(x_1,\ldots,x_n)$ données.

- L'estimateur des moindres carrés :

$$\widehat{\boldsymbol{\beta}}_{\mathrm{n}}=(\hat{\boldsymbol{\beta}}_0,\hat{\boldsymbol{\beta}}_1)=\arg\min_{(b_0,b_1)\in\mathbb{R}^2}\sum_{i=1}^n\big(Y_i-b_0-b_1x_i\big)^2.$$

- Solution explicite

Droite de régression

- Le minimum est caractérisé par les équations

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

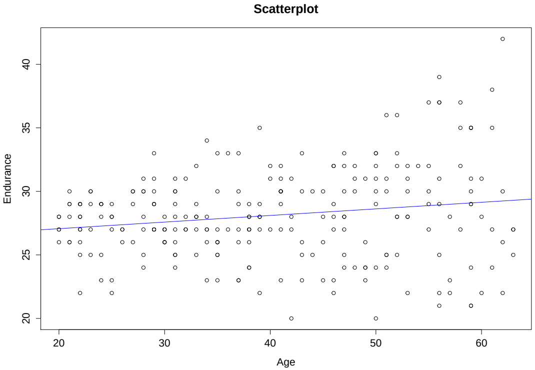
- Notons $\bar{x}_n = n^{-1} \sum_{i=1}^n x_i$. Si le déterminant $\Delta_n \neq 0$ où

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

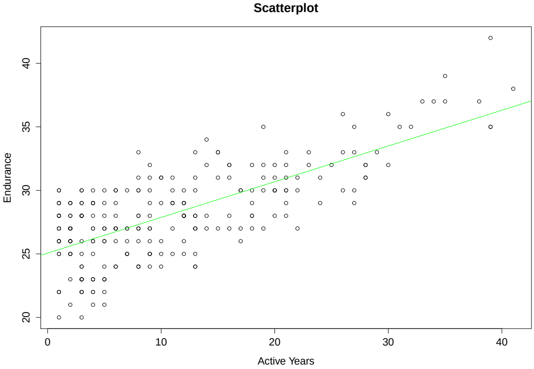
alors ce système d'équations a une solution unique :

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

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

- 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 $r(\boldsymbol{\beta}, \boldsymbol{x}_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$. 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 $\theta \in \Theta = \mathbb{R}^k, \quad \boldsymbol{x}_i \in \mathbb{R}^k$.

- 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 $r(\boldsymbol{\beta}, \boldsymbol{x}_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$. 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 $\theta \in \Theta = \mathbb{R}^k, \quad \boldsymbol{x}_i \in \mathbb{R}^k$.
- 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 $\mathbb{X}_{i,\cdot} = \boldsymbol{x}_i^T$.

EMC en régression linéaire multiple

- Estimateur des moindres carrés en régression linéaire multiple : tout estimateur $\widehat{\beta}_{\mathrm{n}}$ satisfaisant

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

- En notation matricielle :

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

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

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

  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 :

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

- Remarques.
  - L'EMC est un Z-estimateur.
  - unicité de $\widehat{\boldsymbol{\beta}}_{\mathsf{n}}$ 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 $\widehat{\beta}_n$ est unique et

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

Contient le cas précédent de la droite de régression simple.

Géométrie de l'EMC

Proposition

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

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

Résultat géometrique, non stochastique. $\mathbb{X}^T\mathbb{X} \geq 0$; $\mathbb{X}^T\mathbb{X}$ inversible $\Longleftrightarrow \mathbb{X}^T\mathbb{X} > 0$;

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

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

Géométrie de l'EMC

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

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

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

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

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

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 $\widehat{\boldsymbol{Y}}$:

$$\widehat{\boldsymbol{Y}} = \mathbb{X}\widehat{\beta}_{\mathrm{n}} = A\boldsymbol{Y}.$$

Pseudo-inverse de Moore-Penrose

Soit $\mathbb{X}$ une matrice $n \times k$ avec $k \leq n$. On suppose que $\mathbb{X}$ est de rang k.

- $\mathbb{X}^{\#} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$ est la pseudo-inverse de Moore-Penrose.
- $\mathbb{X}^{\#}\mathbb{X} = \mathrm{Id}_k$: $\mathbb{X}$ est un inverse à gauche de la matrice $\mathbb{X}$.
- $\mathbb{X}\mathbb{X}^{\#} = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$ est le projecteur sur l'espace image (l'espace vectoriel engendré par les colonnes de $\mathbb{X}$).
- Méthode de calcul : décomposition QR ou décomposition en valeurs singulières.

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 $\mathbb{E}_\theta[\boldsymbol{\xi}\boldsymbol{\xi}^T] = \mathrm{Id}_n$ (homoscédasticité)

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 $\mathbb{E}_\theta[\boldsymbol{\xi}\boldsymbol{\xi}^T] = \mathrm{Id}_n$ (homoscédasticité)

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 $\mathbb{E}_\theta[\boldsymbol{\xi}\boldsymbol{\xi}^T] = \mathrm{Id}_n$ (homoscédasticité)

Estimateur sans biais

Théorème

L'estimateur $\widehat{\boldsymbol{\beta}}_{\mathsf{n}}$ est sans biais, i.e. pour tout $\theta \in \Theta$,

- $\mathbb{E}_\theta[\widehat{\boldsymbol{\beta}}_{\mathsf{n}}] = \theta$
- $\mathrm{Cov}_\theta(\widehat{\boldsymbol{\beta}}_{\mathsf{n}}) = \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}$.

$$\widehat{\boldsymbol{\beta}}_{\mathsf{n}} = \mathbb{X}^{\#}\,\boldsymbol{Y} = \boldsymbol{\beta} + \sigma\mathbb{X}^{\#}\boldsymbol{\xi}.$$

$$\begin{aligned}\mathbb{E}_\theta[\widehat{\boldsymbol{\beta}}_{\mathsf{n}}] &= \boldsymbol{\beta} + \sigma\mathbb{X}^{\#}\,\mathbb{E}_\theta[\boldsymbol{\xi}] \\ &= \boldsymbol{\beta}\end{aligned}$$

Estimateur sans biais

Théorème

L'estimateur $\widehat{\boldsymbol{\beta}}_{\mathsf{n}}$ est sans biais, i.e. pour tout $\theta \in \Theta$,

- $\mathbb{E}_\theta[\widehat{\boldsymbol{\beta}}_{\mathsf{n}}] = \theta$
- $\mathrm{Cov}_\theta(\widehat{\boldsymbol{\beta}}_{\mathsf{n}}) = \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}$.

$$\widehat{\boldsymbol{\beta}}_{\mathsf{n}} = \mathbb{X}^{\#}\,\boldsymbol{Y} = \boldsymbol{\beta} + \sigma\mathbb{X}^{\#}\boldsymbol{\xi}.$$

$\widehat{\boldsymbol{\beta}}_{\mathsf{n}} - \boldsymbol{\beta} = \sigma\mathbb{X}^{\#}\boldsymbol{\xi}$ ce qui implique

$$\begin{aligned}\mathrm{Cov}_\theta(\widehat{\boldsymbol{\beta}}_{\mathsf{n}}) &= \sigma^2\,\mathbb{E}_\theta\left[\{\mathbb{X}^{\#}\boldsymbol{\xi}\}\{\mathbb{X}^{\#}\boldsymbol{\xi}\}^T\right] \\ &= \sigma^2\{\mathbb{X}^{\#}\}\{\mathbb{X}^{\#}\}^T = \sigma^2\left(\mathbb{X}^T\mathbb{X}\right)^{-1}.\end{aligned}$$

Erreur de prédiction

- Erreur de prédiction :

$$\begin{aligned}\hat{\boldsymbol{\xi}} &= \boldsymbol{Y} - \mathbb{X}\,\widehat{\boldsymbol{\beta}}_{\mathrm{n}} = \boldsymbol{Y} - \mathbb{X}\mathbb{X}^{\#}\,\boldsymbol{Y} \\ &= (\mathrm{Id}_n - A)\,\boldsymbol{Y} \quad \text{car } A = \mathbb{X}\mathbb{X}^{\#} = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\end{aligned}$$

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

$$\begin{aligned}\hat{\boldsymbol{\xi}} &= (\mathrm{Id}_n - A)\mathbb{X}\boldsymbol{\beta} + \sigma(\mathrm{Id}_n - A)\boldsymbol{\xi} \\ &= \sigma(\mathrm{Id}_n - A)\boldsymbol{\xi}\end{aligned}$$

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

Démonstration.

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

Résidus et variance résiduelle

Theorem

Pour tout $\theta \in \Theta$

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

Démonstration.

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

Résidus et variance résiduelle

Theorem

Pour tout $\theta \in \Theta$

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

Démonstration.

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

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

Estimateur sans biais 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.

$\hat{\boldsymbol{\xi}} = (\mathrm{Id}_n - A)\,\boldsymbol{Y} = \sigma(\mathrm{Id}_n - A)\boldsymbol{\xi}$. Comme $(\mathrm{Id}_n - A)^2 = (\mathrm{Id}_n - A)$, nous avons

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

$\square$

Coefficient de détermination

- Pythagore

$$\begin{aligned} \| \boldsymbol{Y} \|^2 &= \|A \boldsymbol{Y} \|^2 + \|(\mathrm{Id}_n - A) \boldsymbol{Y} \|^2 \\ &= \|\hat{\mathbf{Y}}\|^2 + \|\hat{\boldsymbol{\xi}}\|^2 \end{aligned}$$

- Coefficient de détermination

$$\begin{aligned} R^2 &= \frac{\|\hat{\mathbf{Y}}\|^2}{\| \boldsymbol{Y} \|^2} \\ &= 1 - \frac{\|\hat{\boldsymbol{\xi}}\|^2}{\| \boldsymbol{Y} \|^2} = 1 - \frac{\mathrm{SCR}}{\mathrm{SCT}} \end{aligned}$$

où SCR est la somme des carrés résiduels (RSS : residual sum of squares) et SCT est la somme des carrés totaux.

Diagnostic de régression

Residuals :

Min	1Q	Median	3Q	Max
-5.2585	-1.3536	-0.0403	1.4752	5.3874

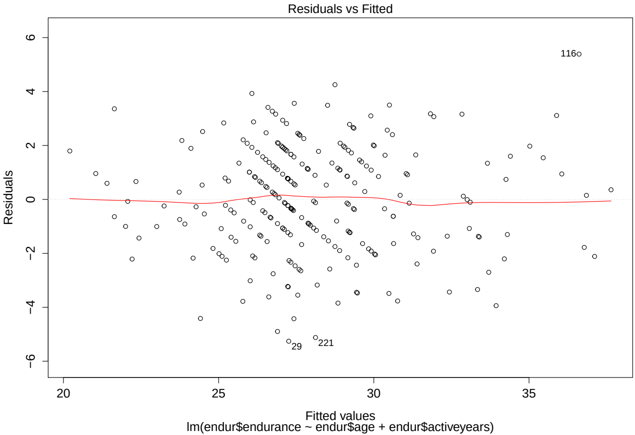
Coefficients :

	Estimate	Std. Error	t value	$\Pr(> \|t\|)$
(Intercept)	30.14524	0.45829	65.78	$< 2e-16 ***$
age	-0.16728	0.01369	-12.22	$< 2e-16 ***$
activeyears	0.43177	0.01823	23.69	$< 2e-16 ***$

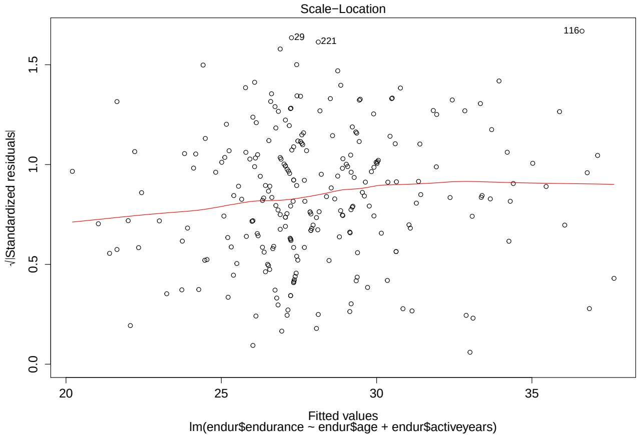
Residual standard error : 1.977 on 247 degrees of freedom

Multiple R-squared : 0.7041, Adjusted R-squared : 0.7017

Diagnostic de régression



Diagnostic de régression



Régression gaussienne

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

- On sait expliciter la loi exacte (non-asymptotique !) de $(\widehat{\boldsymbol{\beta}}_{\mathrm{n}}, \hat{\sigma}^2)$.
- Ingrédient :
  - loi des vecteurs gaussiens sont caractérisés par leur moyenne et matrice de variance-covariance.
  - pour des vecteurs gaussiens, la décorrélation implique l'indépendance.

Cadre gaussien : loi des estimateurs

Proposition

1. $\widehat{\beta}_n \sim \mathcal{N}\left(\beta, \sigma^2 \left(\mathbb{X}^T \mathbb{X}\right)^{-1}\right)$
2. $\|(\mathrm{Id}_n - A)\, \boldsymbol{Y}\, \|^2 \sim \sigma^2 \chi^2(n-k)$ *loi du Chi 2 à $n-k$ degrés de liberté*
3. $\widehat{\beta}_n$ *et* $(\mathrm{Id}_n - A)\, \boldsymbol{Y}$ *sont indépendants.*

Théorème de Cochran

Théorème

Soit $\boldsymbol{Y} \sim N(\mu, \sigma^2 \mathrm{Id}_n)$, $\mathcal{M}$ *un sous espace de* $\mathbb{R}^n$ *de dimension* k, Π *la matrice de projection orthogonale sur* $\mathcal{M}$ *et* $\Pi_\perp = \mathrm{Id}_n - \Pi$ *la matrice de projection orthogonale sur* $\mathcal{M}^\perp$*. Nous avons*

1. $\Pi\, \boldsymbol{Y} \sim \mathcal{N}(\Pi\mu, \sigma^2 \Pi)$, $\Pi_\perp\, \boldsymbol{Y} \sim \mathcal{N}(\Pi_\perp \mu, \sigma^2 \Pi_\perp)$
2. *les vecteurs* $\Pi\, \boldsymbol{Y}$ *et* $\Pi_\perp\, \boldsymbol{Y}$ *sont indépendants*
3. $\|\Pi(\boldsymbol{Y} - \mu)\|^2/\sigma^2 \sim \chi^2_k$ *et* $\Pi_\perp(\boldsymbol{Y} - \mu)\|^2/\sigma^2 \sim \chi^2_{n-k}$.

Théorème de Cochran

Théorème

Soit $\boldsymbol{Y} \sim N(\mu, \sigma^2 \mathrm{Id}_n)$, $\mathcal{M}$ *un sous espace de* $\mathbb{R}^n$ *de dimension* k, Π *la matrice de projection orthogonale sur* $\mathcal{M}$ *et* $\Pi_\perp = \mathrm{Id}_n - \Pi$ *la matrice de projection orthogonale sur* $\mathcal{M}^\perp$*. Nous avons*

1. $\Pi\,\boldsymbol{Y} \sim \mathcal{N}(\Pi\mu, \sigma^2\Pi)$, $\Pi_\perp\,\boldsymbol{Y} \sim \mathcal{N}(\Pi_\perp\mu, \sigma^2\Pi_\perp)$
2. *les vecteurs* $\Pi\,\boldsymbol{Y}$ *et* $\Pi_\perp\,\boldsymbol{Y}$ *sont indépendants*
3. $\|\Pi(\boldsymbol{Y}-\mu)\|^2/\sigma^2 \sim \chi^2_k$ *et* $\Pi_\perp(\boldsymbol{Y}-\mu)\|^2/\sigma^2 \sim \chi^2_{n-k}$.

Théorème de Cochran

Théorème

Soit $\boldsymbol{Y} \sim N(\mu, \sigma^2 \mathrm{Id}_n)$, $\mathcal{M}$ *un sous espace de* $\mathbb{R}^n$ *de dimension* k, Π *la matrice de projection orthogonale sur* $\mathcal{M}$ *et* $\Pi_\perp = \mathrm{Id}_n - \Pi$ *la matrice de projection orthogonale sur* $\mathcal{M}^\perp$. *Nous avons*

1. $\Pi\, \boldsymbol{Y} \sim \mathcal{N}(\Pi\mu, \sigma^2 \Pi)$, $\Pi_\perp\, \boldsymbol{Y} \sim \mathcal{N}(\Pi_\perp \mu, \sigma^2 \Pi_\perp)$
2. *les vecteurs* $\Pi\, \boldsymbol{Y}$ *et* $\Pi_\perp\, \boldsymbol{Y}$ *sont indépendants*
3. $\|\Pi(\boldsymbol{Y} - \mu)\|^2/\sigma^2 \sim \chi^2_k$ *et* $\Pi_\perp(\boldsymbol{Y} - \mu)\|^2/\sigma^2 \sim \chi^2_{n-k}$.

Cadre gaussien : loi des estimateurs

Proposition

1. $\widehat{\beta}_n \sim \mathcal{N}\left(\beta, \sigma^2 (\mathbb{X}^T \mathbb{X})^{-1}\right)$
2. $\|(\mathrm{Id}_n - A)\, \boldsymbol{Y}\|^2 \sim \sigma^2 \chi^2(n-k)$ *loi du Chi 2 à $n-k$ degrés de liberté*
3. $\widehat{\beta}_n$ *et* $(\mathrm{Id}_n - A)\, \boldsymbol{Y}$ *sont indépendants.*

- Définition : $\widehat{\beta}_n = \mathbb{X}^{\#}\, \boldsymbol{Y}$ et $\boldsymbol{Y} \sim N(\mathbb{X}\beta, \sigma^2 \mathrm{Id}_n)$.
- $\widehat{\beta}_n \sim N(\mathbb{X}^{\#}\mathbb{X}\beta, \sigma^2 \{\mathbb{X}^{\#}\}\{\mathbb{X}^{\#}\}^T)$
- On conclut en remarquant $\mathbb{X}^{\#}\mathbb{X} = \mathrm{Id}_k$ et $\{\mathbb{X}^{\#}\}\{\mathbb{X}^{\#}\}^T = (\mathbb{X}^T\mathbb{X})^{-1}$

Cadre gaussien : loi des estimateurs

Proposition

1. $\widehat{\beta}_{\mathrm{n}} \sim \mathcal{N}\big(\beta, \sigma^2 (\mathbb{X}^T \mathbb{X})^{-1}\big)$
2. $\|(\mathrm{Id}_n - A)\,\boldsymbol{Y}\,\|^2 \sim \sigma^2 \chi^2(n-k)$ *loi du Chi 2 à $n-k$ degrés de liberté*
3. $\widehat{\beta}_{\mathrm{n}}$ *et* $(\mathrm{Id}_n - A)\,\boldsymbol{Y}$ *sont indépendants.*

Application directe de Cochran en remarquant que

$$(\mathrm{Id}_n - A)\,\mathbb{E}_\theta[\boldsymbol{Y}] = (\mathrm{Id}_n - A)\,\boldsymbol{X}\,\beta = 0.$$

Cadre gaussien : loi des estimateurs

Proposition

1. $\widehat{\beta}_n \sim \mathcal{N}\left(\beta, \sigma^2 \left(\mathbb{X}^T \mathbb{X}\right)^{-1}\right)$
2. $\|(\mathrm{Id}_n - A)\, \boldsymbol{Y}\,\|^2 \sim \sigma^2 \chi^2(n-k)$ *loi du Chi 2 à $n-k$ degrés de liberté*
3. $\widehat{\beta}_n$ *et* $(\mathrm{Id}_n - A)\, \boldsymbol{Y}$ *sont indépendants.*

- Le théorème de Cochran montre que $A\, \boldsymbol{Y}$ et $(\mathrm{Id}_n - A)\, \boldsymbol{Y}$ sont indépendants.
- On conclut en remarquant que $\widehat{\beta}_n = \mathbb{X}^{\#}\, \boldsymbol{Y} = \mathbb{X}^{\#} A\, \boldsymbol{Y}$.

Estimateur de la variance σ^2 : cadre gaussien

$$\widehat{\sigma}_n^2 = \frac{\|(\mathrm{Id}_n - A)\, \boldsymbol{Y}\,\|^2}{n-k} = \frac{1}{n-k}\sum_{i=1}^{n}\big(Y_i - \boldsymbol{x}_i^T\,\widehat{\boldsymbol{\beta}}_{\mathrm{n}}\big)^2$$

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 $\widehat{\boldsymbol{\beta}}_{\mathrm{n}}$.

Lois des coordonnées de $\widehat{\beta}_{\mathrm{n}}$: cadre gaussien

$$\widehat{\beta}_{\mathrm{n}j} - \beta_j \sim \mathcal{N}\left(0, \sigma^2 b_j\right)$$

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

$$\frac{\widehat{\beta}_{\mathrm{n}j} - \beta_j}{\widehat{\sigma}_n\sqrt{b_j}} \sim t_{n-k}$$

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

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

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

Prévision

Modèle de régression

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

Régression linéaire : $r(\boldsymbol{\beta}, \boldsymbol{x}_i) = \boldsymbol{\beta}^T \boldsymbol{x}_i$.

- Problème de prévision : Donner la prévison de la valeur de fonction de régression $r(\boldsymbol{\beta}, \boldsymbol{x}) = \boldsymbol{\beta}^T \boldsymbol{x}$
- Soit $\widehat{\boldsymbol{\beta}}_n$ un estimateur de $\boldsymbol{\beta}$. Prévision par substitution :
$$\boxed{\widehat{Y}(\boldsymbol{x}) = r(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{x}).}$$
- Question statistique : quelle est la qualité de la prévision ? Intervalle de confiance pour $r(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{x})$?

Moyenne et variance de la prévision

Theorem

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

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

Moyenne et variance de la prévision

Theorem

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

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

car $\mathbb{X}^\# \mathbb{X} = I$. Par conséquent, comme $\boldsymbol{X}^\# \{\boldsymbol{X}^\#\}^T = (\boldsymbol{X}^T \boldsymbol{X})^{-1}$,

$$
\mathrm{Var}_\theta(\hat{Y}_n(\boldsymbol{x})) = \sigma^2 \boldsymbol{x}^T \mathbb{X}^\# \mathbb{E}_\theta[\boldsymbol{\xi}\boldsymbol{\xi}^T] \{\mathbb{X}^\#\}^T \boldsymbol{x}
$$

Moyenne et variance de la prévision

Theorem

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

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

Prévision : modèle linéaire gaussienne

Proposition

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

1. $\widehat{Y}(\boldsymbol{x}) \sim \mathcal{N}\left(\boldsymbol{x}^T \boldsymbol{\beta}, \sigma^2 \boldsymbol{x}^T \left(\mathbb{X}^T \mathbb{X}\right)^{-1} \boldsymbol{x} \right)$
2. $\widehat{Y}(\boldsymbol{x})$ *et* $(\mathrm{Id}_n - A)\, \boldsymbol{Y}$ *sont indépendants.*

Prévision : modèle linéaire gaussienne

- D'après la Proposition,

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

- On replace σ^2 inconnu par $\widehat{\sigma}_n^2 = \|(\mathrm{Id}_n - A)\, \boldsymbol{Y}\|^2/(n-k)$.
- *t*-statistique :

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

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

Prévision : intervalle de confiance

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

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

$\Longrightarrow$ intervalle de confiance de niveau $1-\alpha$ pour $r(\boldsymbol{\beta}, \boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$ est $[r_L, r_U]$, où :

$$r_L = \widehat{Y} - q_{1-\frac{\alpha}{2}}(t_{n-k})\sqrt{\widehat{\sigma}_n^2\, \boldsymbol{x}^T\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\boldsymbol{x}},$$

$$r_U = \widehat{Y} + q_{1-\frac{\alpha}{2}}(t_{n-k})\sqrt{\widehat{\sigma}_n^2\, \boldsymbol{x}^T\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\boldsymbol{x}}.$$

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)

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

- A comparer avec le cadre gaussien :

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

Régression non-linéaire

On observe

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

où

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

avec $\boldsymbol{x}_i \in \mathbb{R}^k$, et $\boldsymbol{\beta} \in \Theta \subset \mathbb{R}^d$, $\sigma \in \mathbb{R}^+$.
Si $\xi_i \sim_{\text{i.i.d.}} \mathcal{N}(0,1)$, la fonction de vraisemblance est donnée par

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

Régression non-linéaire

On observe

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

où

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

avec $\boldsymbol{x}_i \in \mathbb{R}^k$, et $\boldsymbol{\beta} \in \Theta \subset \mathbb{R}^d$, $\sigma \in \mathbb{R}^+$.
L'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\beta}}_{\mathrm{n}}$ de $\boldsymbol{\beta}$ est obtenu en minimisant la fonction

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

L'estimateur du maximum de vraisemblance de σ^2 est donné par

$$\hat{\sigma}_n^2 = n^{-1} \sum_{i=1}^{n} \big(Y_i - r(\widehat{\boldsymbol{\beta}}_{\mathrm{n}}, \boldsymbol{x}_i)\big)^2.$$

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* $\widehat{\beta}_n$ *satisfaisant*

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

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

- Extension des résultats en densité $\to$ 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

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

-

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