Exercise Class : Linear Regression

Exercise 1 :

1) Soit $y \in \mathbb{R}^n$ alors $\text{var}_n(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2$ avec $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$

$$\text{var}_n(y) = 0 \iff \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2 = 0 \iff \sum_{i=1}^{n}\underbrace{(y_i - \bar{y}_n)^2}_{\geq 0} = 0$$

$$\iff \forall i = 1, \dots, n \quad y_i = \bar{y}_n$$

$$\boxed{\text{var}_n(y) = 0 \iff y \in \text{Span}(\mathbb{1}_n) \iff \exists c \in \mathbb{R} \,/\, \forall i, y_i = c}$$

2) Montrons que $\text{Ker}(X) = \text{Ker}(X^TX)$ avec $X \in \mathbb{R}^{n \times p}$

[C]. Soit $u \in \text{Ker}(X)$ alors $Xu = 0$ donc $X^TXu = 0$ et $u \in \text{Ker}(X^TX)$

[I]. Soit $u \in \text{Ker}(X^TX)$ alors $X^TXu = 0$ puis $u^TX^TXu = 0$, i.e $\|Xu\|^2 = 0$ donc $Xu = 0$ et $u \in \text{Ker}(X)$, ainsi $\boxed{\text{Ker}(X) = \text{Ker}(X^TX)}$

3) $\mathbb{1}_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^{n \times 1}$, $\tilde{X} \in \mathbb{R}^{n \times p}$, $X = (\mathbb{1}_n, \tilde{X}) \in \mathbb{R}^{n \times (p+1)}$

$$\tilde{X} = \begin{pmatrix} \cdots\cdots & \tilde{x}_1^T & \cdots\cdots \\ & \vdots & \\ \cdots\cdots & \tilde{x}_n^T & \cdots\cdots \end{pmatrix} \in \mathbb{R}^{n \times p} \qquad \hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i$$

(X^TX) non inversible $\iff \text{Ker}(X^TX) \neq \{0\} \iff \text{Ker}(X) \neq \{0\}$

$$\iff \exists u \in \mathbb{R}^{p+1} \setminus \{0\} \,/\, Xu = 0$$

$$\iff \exists u_0 \in \mathbb{R}, \tilde{u} \in \mathbb{R}^p \setminus \{0\} \,/\, u_0\mathbb{1}_n + \tilde{X}\tilde{u} = 0$$

Posons $y = \tilde{X}\tilde{u} \in \mathbb{R}^n$

(X^TX) non-inversible $\iff \exists u_0 \in \mathbb{R}, \tilde{u} \in \mathbb{R}^p \setminus \{0\} \,/\, y = \tilde{X}\tilde{u} = -u_0\mathbb{1}_n$

$$\iff \exists \tilde{u} \in \mathbb{R}^p \setminus \{0\} \,/\, y \in \text{Span}(\mathbb{1}_n), \text{ i.e } y = \bar{y}\mathbb{1}_n$$

Or $y = \tilde{X}\tilde{u}$ donc $\bar{y} = \bar{\tilde{X}}^T\tilde{u} = \hat{\mu}_n\tilde{u}$

(X^TX) non-inversible $\iff \exists \tilde{u} \in \mathbb{R}^p \setminus \{0\} \,/\, \underbrace{(\tilde{X} - \hat{\mu}_n)}_{= \tilde{X}^c = \begin{pmatrix} \vdots \\ \tilde{x}_i - \hat{\mu}_n \\ \vdots \end{pmatrix}} u = 0$

$$\iff \text{Ker}(\tilde{X}^c) \neq \{0\}$$

$$\iff \text{Ker}(\tilde{X}^{cT}\tilde{X}^c) \neq \{0\}$$

$$\boxed{(X^TX) \notin GL_{p+1}(\mathbb{R}) \iff (\tilde{X}^{cT}\tilde{X}^c) \text{ non-inversible}}$$

$$\boxed{\text{et } \tilde{X}^{cT}\tilde{X}^c = \sum_{i=1}^{n}(\tilde{x}_i - \hat{\mu}_n)(\tilde{x}_i - \hat{\mu}_n)^T}$$

4) Sur les variables centrées $\tilde{X}^c \in \mathbb{R}^{n\times p}$, $Y^c \in \mathbb{R}^n$,

$$\hat{\theta}_n \in \underset{\theta\in\mathbb{R}^p}{\arg\min} \|Y^c - \tilde{X}^c\theta\|_2^2$$

L'estimateur est donné par : (on a supposé $\mathrm{cov}_n(\tilde{X})$ inversible)

$$\boxed{\hat{\theta}_n = (\tilde{X}^{cT}\tilde{X}^c)^{-1}\tilde{X}^{cT}Y^c = \frac{1}{n}\mathrm{cov}_n(\tilde{X})^{-1}\tilde{X}^{cT}Y^c}$$

Exercise 2 :

$X \in \mathbb{R}^{n\times p}$ de rang plein $\tilde{X} = (\tilde{X}_1|\dots|\tilde{X}_p)$, $\tilde{X}_k \in \mathbb{R}^n$ $\forall k = 1,\dots,p$

1) On change d'échelle pour la colonne k : $\tilde{X}_k \leftarrow b\tilde{X}_k$ avec $b>0$

Posons $D = \mathrm{Diag}(1,\dots,1,\underset{\text{position } (k+1)}{b},1,\dots,1)$, $D \in \mathbb{R}^{(p+1)\times(p+1)}$

$$XD = (1_n, \tilde{X})D = \begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}|\tilde{X}_1|\dots|\tilde{X}_p\Big)D = \begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}|\tilde{X}_1\times 1|\dots|\tilde{X}_k\times b|\dots|\tilde{X}_p\times 1\Big)$$

$$XD = \begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}|\tilde{X}_1|\dots|\tilde{X}_k b|\dots|\tilde{X}_p\Big) = (1_n, \tilde{X}_1,\dots,\tilde{X}_k b,\dots,\tilde{X}_p) = X_b$$

$$\boxed{D = \mathrm{Diag}(1,\dots,1,b,1,\dots,1), \quad XD = X_b}$$

2) Par définition :

$$\hat{\theta}_{b,n} \in \underset{\theta\in\mathbb{R}^{p+1}}{\arg\min}\left\{\|Y - X_b\theta\|_2^2 = \|Y - XD\theta\|_2^2\right\}$$

D'après les équations normales (projection, Hilbert)

$$(X_b^T X_b)\hat{\theta}_{b,n} = X_b^T Y$$

i.e.

$$D(X^TX)D\,\hat{\theta}_{b,n} = DX^TY$$

donc

$$(X^TX)D\hat{\theta}_{b,n} = X^TY$$

ainsi,

$$\boxed{\hat{\theta}_n = D\hat{\theta}_{b,n}}$$

3) $\hat{\theta}_{b,n} = D^{-1}\hat{\theta}_n$ donc $\mathrm{var}(\hat{\theta}_{b,n}) = D^{-1}\mathrm{Var}(\hat{\theta}_n)D^{-1}$

model fixed design : $Y = X\theta^* + \varepsilon$, $\mathrm{Var}(\varepsilon) = \sigma^2 I_n$, $\mathrm{Var}(\hat{\theta}_n) = (X^TX)^{-1}\sigma^2$

$$\boxed{\mathrm{Var}(\hat{\theta}_{b,n}) = \sigma^2 D^{-1}(X^TX)^{-1}D^{-1}}$$

4) On regarde la valeur prédite par le modèle OLS

. Pour le modèle initial (X,Y), la fonction de prédiction est

$$g : \mathbb{R}^p \to \mathbb{R}, \quad g(x) = x^T \hat{\theta}_n$$

. Pour le modèle modifié, (X_b, Y), la fonction de prédiction est

$$g_b : \mathbb{R}^p \to \mathbb{R}, \quad g_b(x) = x_b^T \hat{\theta}_{b,n} = x^T (DD^{-1}) \hat{\theta}_n = x^T \hat{\theta}_n$$

Pour la prédiction, on a $\boxed{g_b(x) = g(x)}$

5) On regarde la valeur prédite par le modèle RIDGE

. Pour le modèle modifié (X_b, Y) la fonction de prédiction est

$$g_b^{(ridg)}(x) = x_b^T \hat{\theta}_{b,n}^{(ridg)} = x^T D \hat{\theta}_{b,n}^{(ridg)}$$

avec $\hat{\theta}_{b,n}^{(ridg)} \in \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \|Y - X_b\theta\|_2^2 + \lambda \|\theta\|_2^2 \right\} = (X_b^T X_b + \lambda I)^{-1} X_b^T Y$

donc $\boxed{g_b^{(ridg)}(x) \neq g^{(ridg)}(x)}$ $\qquad \hat{\theta}_{b,n}^{(ridg)} \neq D^{-1} \hat{\theta}_n^{(ridg)}$

Il n'y a plus cette propriété d'invariance, c'est pourquoi on doit renormaliser les variables avec l'estimateur RIDGE

Exercise 3:

1) La fonction de prédiction OLS est : $\boxed{\hat{p}(x) = x^T \hat{\theta}_n}$

2) $\hat{\theta}_n = (X^T X)^{-1} X^T Y$ avec $Y = X\theta^* + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$

ainsi $Y \sim \mathcal{N}(X\theta^*, \sigma^2 I_n)$

puis $\hat{\theta}_n$ est aussi gaussien avec $E[\hat{\theta}_n] = (X^T X)^{-1} X^T (X\theta^*) = \theta^*$

et variance $\sigma^2 (X^T X)^{-1}$, $\hat{\theta}_n \sim \mathcal{N}(\theta^*, \sigma^2 (X^T X)^{-1})$

$$\boxed{\hat{p}(x) \sim \mathcal{N}(x^T \theta^*, x^T (X^T X)^{-1} x \sigma^2)}$$

3) $\sum_{i=1}^{n} (y_i - x_i^T \hat{\theta}_n)^2 / \sigma^2 \sim \chi^2_{n-(p+1)}$; $\hat{\sigma}_n^2 (n-(p+1)) = \sum_{i=1}^{n} (y_i - x_i^T \hat{\theta}_n)^2$

$$\frac{\hat{p}(x) - p(x)}{\hat{\sigma}_n \sqrt{(x^T (X^T X)^{-1} x)}} = \underbrace{\frac{\hat{p}(x) - p(x)}{\sqrt{v(x)}}}_{\mathcal{N}(0,1)} \times \underbrace{\frac{1}{\sqrt{\hat{\sigma}_n^2 / \sigma^2}}}_{= 1 \Big/ \sqrt{\frac{\chi^2_{n-(p+1)}}{n-(p+1)}}} \sim T_{n-(p+1)}$$

4) Define $\text{var}(x) = x^T (X^T X)^{-1} x$

$$\frac{y - \hat{p}(x)}{\hat{\sigma}_n \sqrt{1+\text{var}(x)}} = \frac{(y - p(x)) + (p(x) - \hat{p}(x))}{\hat{\sigma}_n \sqrt{1+\text{var}(x)}} = \frac{\varepsilon + (p(x) - \hat{p}(x))}{\hat{\sigma}_n \sqrt{1+\text{var}(x)}}$$

Au numérateur, somme de deux gaussiennes indépendantes, donc gaussienne moyenne nulle et variance $\sigma^2(1+\text{var}(x))$, c'est la même écriture que Q3.

5)
- Pour $p(x)$: $\left[\hat{p}(x) - \hat{\sigma}_n \sqrt{\text{var}(x)}\, t_{1-\alpha/2} \; ; \; \hat{p}(x) + \hat{\sigma}_n \sqrt{\text{var}(x)}\, t_{1-\alpha/2} \right]$

avec $t_{1-\alpha/2}$ quantile $(1-\frac{\alpha}{2})$ de la Student $(n-(p+1))$.
- Pour Y, on remplace $\text{var}(x)$ par $(1+\text{var}(x))$

Exercise 4:

1) $$\hat{\theta}_n^{rdg} \in \underset{\theta \in \mathbb{R}^p}{\text{argmin}} \left\{ \|Y - X\theta\|_2^2 + \lambda \|\theta\|_2^2 \right\} = \underset{\theta \in \mathbb{R}^p}{\text{argmin}} \; f_\lambda(\theta)$$

$\theta \mapsto f_\lambda(\theta)$ strictement convexe, coercive, existence et unicité du minimum.
Condition du premier ordre ($\nabla f(\theta) = 0$) fournit : $\boxed{\hat{\theta}_n^{rdg} = (X^T X + \lambda I_p)^{-1} X^T Y}$

2) $Y = X\theta^* + \varepsilon$ avec $E[Y] = X\theta^*$ donc

$$\begin{cases} \bullet\; E[\hat{\theta}_n^{rdg}] = (X^T X + \lambda I_p)^{-1} X^T (X\theta^*) = \theta^* - \lambda (X^T X + \lambda I_p)^{-1} \theta^* \\ \bullet\; \text{Var}(\hat{\theta}_n^{rdg}) = \sigma^2 (X^T X + \lambda I_p)^{-1} X^T X (X^T X + \lambda I_p)^{-1} \end{cases}$$

3) L'estimateur ridge s'écrit : $\hat{\theta}_n^{rdg} = AY$ avec Y gaussien donc $\hat{\theta}_n^{rdg}$ est gaussien avec moyenne et variance ci-dessus.

4) $$\frac{\hat{\theta}_{n,k}^{rdg} - \theta_k^*}{\sqrt{\text{Var}(\hat{\theta}_{n,k}^{rdg})}} = \frac{\hat{\theta}_{n,k}^{rdg} - \theta_k^*}{\sqrt{v_k}} \sim \mathcal{N}(0,1)$$

avec $\boxed{\begin{array}{l} v_k = e_k^T (X^T X + \lambda I_p)^{-2} (X^T X) e_k \\ e_k^T = (0, \dots, 0, 1, 0, \dots, 0) \end{array}}$

L'intervalle de confiance de niveau α est alors

$$\boxed{\hat{\theta}_{n,k}^{(rdg)} \pm \sqrt{v_k}\, \Phi^{-1}(1 - \alpha/2)}$$