Proposition de correction pour le quizz de statistiques

1. Général:

1) Que vaut $Cov(X+\mu)$ pour tout $\mu \in \mathbb{R}^p$ déterministe, et tout vecteur aléatoire $X \in \mathbb{R}^p$?

On a:

$$\begin{aligned}
& Cov(X+\mu) \\
= \; & \mathbb{E}\left[(X+\mu-\mathbb{E}(X+\mu))\,(X+\mu-\mathbb{E}(X+\mu))^T\right] \\
= \; & \mathbb{E}\left[(X+\mu-\mathbb{E}(X)-\mu)\,(X+\mu-\mathbb{E}(X)-\mu)^T\right] \\
= \; & \mathbb{E}\left[(X-\mathbb{E}(X))\,(X-\mathbb{E}(X))^T\right] \\
= \; & Cov(X)
\end{aligned}$$

2) Que vaut $Cov(AX)$, pour toute matrice $A \in \mathbb{R}^{m\times p}$ et tout vecteur aléatoire $X \in \mathbb{R}^p$?

On a:

$$\begin{aligned}
& Cov(AX) \\
= \; & \mathbb{E}\left[(AX-\mathbb{E}(AX))\,(AX-\mathbb{E}(AX))^T\right] \\
= \; & \mathbb{E}\left[(AX-A\,\mathbb{E}(X))\,(AX-A\,\mathbb{E}(X))^T\right] \\
= \; & \mathbb{E}\left[(A[X-\mathbb{E}(X)])\,(A[X-\mathbb{E}(X)])^T\right] \\
= \; & \mathbb{E}\left[A\,([X-\mathbb{E}(X)])\,([X-\mathbb{E}(X)])^T\,A^T\right] \\
= \; & A\,\mathbb{E}\left[(X-\mathbb{E}(X))(X-\mathbb{E}(X))^T\right]A^T \\
= \; & A\;Cov(X)\;A^T
\end{aligned}$$

3) Donner un modèle naturel pour "un lancer de dé" (non-nécessairement équilibré)?

Un modèle naturel (ou famille de loi naturelle) pour "un lancer de dé" est la distribution catégorielle ou multi-Bernoulli, qui généralise la loi Bernoulli à k catégories. Ici, on aura $k=6$, et $\mathcal{M} = \{\mathbb{P}_\theta : \theta \in \mathbb{R}^6, \sum\limits_{i=1}^{6} \theta_i = 1\}$.

4) Soit $x_1, x_2, \ldots, x_n$ i.i.d. tel que $\mathbb{E}[x_1^2] < \infty$. Quel estimateur $\hat{\mu}$ minimise $\sum_{i=1}^n (x_i-\mu)^2$? Donner son biais et sa variance, pour tout $n > 1$.

On cherche: $\hat{\mu} = \underset{\mu}{argmin} \sum\limits_{i=1}^{n} (x_i-\mu)^2$. La condition de premier ordre est:

$$
\begin{aligned}
& \frac{d}{d\mu}\left(\sum_{i=1}^{n}(x_i-\mu)^2\right)=0 \\
\Leftrightarrow\quad & -2\sum_{i=1}^{n}(x_i-\mu)=0 \\
\Leftrightarrow\quad & \sum_{i=1}^{n}(x_i-\mu)=0 \\
\Leftrightarrow\quad & \sum_{i=1}^{n}x_i=n\mu \\
\Leftrightarrow\quad & \mu=\frac{1}{n}\sum_{i=1}^{n}\ x_i
\end{aligned}
$$

L'estimateur est donc $\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}\ x_i$. Son biais est donné par:

$$
\begin{aligned}
& Biais(\hat{\mu},\mu) \\
= \quad & \mathbb{E}_\theta(\hat{\mu}-\mu) \\
= \quad & \mathbb{E}_\theta\left(\frac{1}{n}\ \sum_{i=1}^{n}x_i-\mu\right) \\
= \quad & \frac{1}{n}\ \sum_{i=1}^{n}\ \mathbb{E}_\theta(x_i)-\mu \\
= \quad & \frac{1}{n}\ n\,\mathbb{E}_\theta(X)-\mu \qquad \text{(car les } x_i \text{ sont identiquement distribuées)} \\
= \quad & \mathbb{E}_\theta(X)-\mu \\
= \quad & \mu-\mu \\
= \quad & 0
\end{aligned}
$$

Sa variance est donnée par:

$$
\begin{aligned}
& var(\hat{\mu}) \\
= \quad & var\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right) \\
= \quad & \frac{1}{n^2}var\left(\sum_{i=1}^{n}x_i\right) \\
= \quad & \frac{1}{n^2}\ \sum_{i=1}^{n}var(x_i) \qquad \text{(par indépendence des } x_i\text{)} \\
= \quad & \frac{1}{n^2}\ n\ var(X) \qquad \text{(car les } x_i \text{ sont identiquement distribuées)} \\
= \quad & \frac{1}{n}\ var(X)
\end{aligned}
$$

5) Que vaut le biais de $\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2$ ($\bar{y}_n$ est la moyenne empirique) pour des y_i i.i.d. gaussiens, centrés et de variance σ^2?

On définit la variance empirique comme $\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2$. On peut ensuite la manipuler pour obtenir:

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2 \quad \Leftrightarrow \quad \frac{n}{\sigma^2}\tilde{\sigma}^2 = \sum_{i=1}^{n}\frac{(y_i - \bar{y}_n)^2}{\sigma^2}$$

Il découle ensuite du théorème de Cochran que:

$$\sum_{i=1}^{n}\frac{(y_i - \bar{y}_n)^2}{\sigma^2} \sim \chi^2(n-1) \quad \Leftrightarrow \quad \frac{n}{\sigma^2}\tilde{\sigma}^2 \sim \chi^2(n-1)$$

Les propriétés de la distribution χ^2 impliquent que $\frac{n}{\sigma^2}\tilde{\sigma}^2$ a une espérance de $(n-1)$ et une variance de $2(n-1)$. On conclut alors:

$$\mathbb{E}(\frac{n}{\sigma^2}\tilde{\sigma}^2) = n-1 \ \Leftrightarrow \ \mathbb{E}(\tilde{\sigma}^2) = \frac{n-1}{n}\sigma^2$$

Le biais est donc de:

$$Biais(\tilde{\sigma}^2, \sigma^2) = \mathbb{E}(\tilde{\sigma}^2) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2$$

6) On suppose que l'on observe $y_1, \ldots, y_n$, des variables réelles i.i.d., gaussiennes, centrées et de variance σ^2. Quel est le risque quadratique de l'estimateur $\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2$ de σ^2 ($\bar{y}_n$ est la moyenne empirique)?

On utilise le même raisonnement qu'à la question précédente, qu'on complète en ajoutant que:

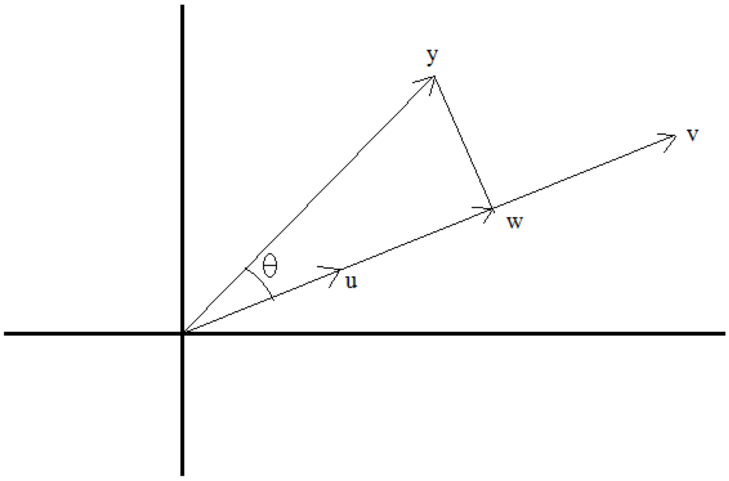
$$var(\frac{n}{\sigma^2}\tilde{\sigma}^2) = 2(n-1) \ \Leftrightarrow \ \frac{n^2}{\sigma^4}var(\tilde{\sigma}^2) = 2(n-1) \ \Leftrightarrow \ var(\tilde{\sigma}^2) = \frac{2\sigma^4(n-1)}{n^2}$$

On utilise alors la définition du risque quadratique pour obtenir:

$$\begin{aligned} & R(\tilde{\sigma}^2) \\ = \ & var(\tilde{\sigma}^2) + (biais(\tilde{\sigma}^2))^2 \\ = \ & \frac{2\sigma^4(n-1)}{n^2} + \frac{\sigma^4}{n^2} \\ = \ & \frac{\sigma^4(2n-1)}{n^2} \end{aligned}$$

7) Quelle est la projection orthogonale du vecteur $\boldsymbol{y} \in \mathbb{R}^n$ sur Vect(1_n), avec $1_n = (1, \ldots, 1)^T \in \mathbb{R}^n$?

Soit $\boldsymbol{y} \in \mathbb{R}^n$ et $\boldsymbol{v} = 1_n = (1, \ldots, 1)^T \in \mathbb{R}^n$. On considère la projection orthogonale $\boldsymbol{w}$ de $\boldsymbol{y}$ sur $\boldsymbol{v}$, que l'on peut représenter par le grahique suivant:



Le vecteur $\boldsymbol{u}$ est un vecteur unitaire dans la direction de $\boldsymbol{w}$ tel que $u = \frac{1}{\sqrt{n}} 1_n$. On note que par construction $\boldsymbol{w} = \|w\|\boldsymbol{u}$. Par definition de la fonction cosinus, on a également $cos(\theta) = \frac{\|w\|}{\|y\|}$ et par définition du produit scalaire on a $cos(\theta) = \frac{\langle u,y \rangle}{\|u\|\|y\|}$. En combinant ces deux dernières expressions, on obtient $\|w\| = \langle u, y\rangle$. En substituant dans la première expression, on obtient $w = \langle u, y\rangle \; u$ et donc:

$$\begin{aligned}
& w \\
= \; & \langle u, y\rangle \; u \\
= \; & \langle \frac{1}{\sqrt{n}} 1_n, y\rangle \; \frac{1}{\sqrt{n}} 1_n \\
= \; & \frac{1}{n} \langle 1_n, y\rangle \; 1_n \\
= \; & \frac{1}{n} \; (\sum_{i=1}^{n} y_i) \; 1_n \\
= \; & \bar{y} \; 1_n \\
= \; & \begin{pmatrix} \bar{y} \\ \bar{y} \\ \vdots \\ \bar{y} \end{pmatrix}
\end{aligned}$$

8) Quels sont les vecteurs $\boldsymbol{y} \in \mathbb{R}^n$ tels que $var_n(\boldsymbol{y}) = 0$ (var_n est la variance empirique)?

On obtient que $var_n(\boldsymbol{y}) = 0$ si et seulement si $\frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y}_n)^2 = 0$. Celà implique que $y_i = \bar{y} \; \forall i$ (car sinon il existe un i tel que $(y_i - \bar{y}_n)^2 > 0$ et donc $\sum_{i=1}^{n} (y_i - \bar{y}_n)^2 > 0$), ce qui n'est possible que si $y_1 = y_2 = \ldots = y_n = y$, pour y un scalaire donné. On conclut donc que ces vecteurs sont de la forme $\boldsymbol{y} = y.1_n$.

2. Moindres carrés unidimensionnels:

On observe $\boldsymbol{y} = (y_1, \dots, y_n)^T$ et $\boldsymbol{x} = (x_1, \dots, x_n)^T$.
1) La fonction $(\theta_0, \theta_1) \to \frac{1}{2}\sum_{i=1}^n (y_i - \theta_0 - \theta_1 x_i)^2$ est-elle convexe ou concave?

On considère la fonction $f = (y_i - \theta_0 - \theta_1 x_i)^2$. Pour estimer sa convexité, on calcule sa Hessienne:

$$\frac{\partial f}{\partial \theta_0} = -2(y_i - \theta_0 - \theta_1 x_i)$$
$$\frac{\partial^2 f}{\partial \theta_0^2} = 2$$
$$\frac{\partial f}{\partial \theta_1} = -2x_i(y_i - \theta_0 - \theta_1 x_i)$$
$$\frac{\partial^2 f}{\partial \theta_1^2} = 2x_i^2$$
$$\frac{\partial^2 f}{\partial \theta_1 \partial \theta_2} = 2x_i$$

La Hessienne H est donc donnée par $H = \begin{pmatrix} 2 & 2x_i \\ 2x_i & 2x_i^2 \end{pmatrix}$. On note que la matrice est singulière (sa seconde colonne est la première multipliée par x_i), donc au moins une de ses valeurs propres est 0. En utilisant le fait que la trace est la somme des valeurs propres, on obtient que la seconde valeur propre est $2(1 + x_i^2)$, qui est toujours positive. Donc la Hessienne H est symétrique semi-définie positive, et la function f est convexe. Comme la fonction $(\theta_0, \theta_1) \to \frac{1}{2}\sum_{i=1}^n (y_i - \theta_0 - \theta_1 x_i)^2$ est une somme de fonctions convexes, elle est convexe elle-même.

2) Donner la formule $(\hat{\theta}_0, \hat{\theta}_1)$ des estimateurs des moindres carrés où $\hat{\theta}_0$ correspond au coefficient des constantes et $\hat{\theta}_1$ correspond à l'influence de x sur y. On les exprimera en fonction des $x_i, y_i, \bar{x}_n, \bar{y}_n$.

On cherche à minimiser la fonction $f(\theta_0, \theta_1) = \frac{1}{2}\sum\limits_{i=1}^n (y_i - \theta_1 x_i - \theta_0)^2$. Pour un jeu de données (x_i, y_i) fixé, c'est une fonction de θ_0 et θ_1. D'après le théorme de Fermat, un minimum de f est à chercher parmi les couples (θ_0, θ_1) qui annulent le gradient de f:
$\frac{\partial f}{\partial \theta_0} = 0 \Leftrightarrow \sum\limits_{i=1}^n (-1)(y_i - \theta_1 x_i - \theta_0) = 0$
$\frac{\partial f}{\partial \theta_1} = 0 \Leftrightarrow \sum\limits_{i=1}^n (-x_i)(y_i - \theta_1 x_i - \theta_0) = 0$

En divisant par n:
$\frac{\partial f}{\partial \theta_0} = 0 \Leftrightarrow \frac{1}{n}\sum\limits_{i=1}^n y_i - \theta_1 \frac{1}{n}\sum\limits_{i=1}^n x_i - \frac{1}{n}\sum\limits_{i=1}^n \theta_0 = 0$
$\frac{\partial f}{\partial \theta_1} = 0 \Leftrightarrow \frac{1}{n}\sum\limits_{i=1}^n (x_i y_i) - \theta_1 \frac{1}{n}\sum\limits_{i=1}^n x_i^2 - \theta_0 \frac{1}{n}\sum\limits_{i=1}^n x_i = 0$

En notant avec le symbole barre les moyennes $\bar{x}$ et $\bar{y}$:
$\frac{\partial f}{\partial \theta_0} = 0 \Leftrightarrow \bar{y} - \theta_1 \bar{x} = \theta_0 \qquad (1)$
$\frac{\partial f}{\partial \theta_1} = 0 \Leftrightarrow \frac{1}{n}\sum\limits_{i=1}^n (x_i y_i) - \theta_1 \frac{1}{n}\sum\limits_{i=1}^n x_i^2 = \theta_0 \bar{x} \qquad (2)$

Ce qui constitute un système de deux équations à deux inconnues θ_0 et θ_1. En multipliant (1) par $\bar{x}$ et en soustrayant la ligne obtenue à l'équation (2), on obtient:
$\frac{\partial f}{\partial \theta_0} = 0 \Leftrightarrow \theta_0 = \bar{y} - \theta_1 \bar{x}$
$\frac{\partial f}{\partial \theta_1} = 0 \Leftrightarrow \theta_1 = \frac{\frac{1}{n}\sum\limits_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum\limits_{i=1}^{n} x_i^2 - \bar{x}^2}$

3. Moindres carrés:

1) Ecrire un pseudo-code de descente de gradient pour résoudre le problème des moindres carrés.
<u>Initialisation</u>
(θ_0^0, θ_1^0): valeurs initiales de θ
T: nombre maximum d'itérations
ε: critère d'arrêt
α: pas de l'algorithme

<u>Boucle</u>
for $1 \leq t \leq T$:

$(\theta_0^{t+1}, \theta_1^{t+1}) := (\theta_0^{t+1}, \theta_1^{t+1}) - \alpha . \nabla f(\theta_0^t, \theta_1^t)$
avec (selon le cours) $\nabla f(\theta_0^t, \theta_1^t) = X^T(X\theta - Y)$

soit (en utilisant la fonction de la question 2.2):
$\theta_0^{t+1} := \theta_0^t + \alpha \sum\limits_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)$
$\theta_1^{t+1} := \theta_1^t + \alpha \sum\limits_{i=1}^{n} x_i(y_i - \theta_0 - \theta_1 x_i)$

Stop si critère inférieur à ε

Fin de la boucle

Return (θ_0^T, θ_1^T)

Critères d'arrêt possibles:
$\|\nabla f(\theta_0^t, \theta_1^t)\| \leq \varepsilon$
$|f(\theta_0^{t+1}, \theta_1^{t+1}) - f(\theta_0^t, \theta_1^t)| \leq \varepsilon$
$\|\theta^{t+1} - \theta^t\| \leq \varepsilon$ avec $\theta = (\theta_0, \theta_1)$
$\frac{\|\theta^{t+1} - \theta^t\|}{\|\theta^t\|} \leq \varepsilon$ avec $\theta = (\theta_0, \theta_1)$

10) On suppose que X est de rang plein et on note $\hat{\theta}$ l'estimateur OLS. On note $\tilde{X} = (X_1, \ldots, X_p)$.On change l'échelle d'une des variables: X_k est remplacé par $X_k b$, où $b > 0$.
a) Soit $X_b = (1, X_1, \ldots, X_k b, \ldots, X_p)$. Montrer que $X_b = XD$ où D est une matrice diagonale que l'on précisera.
On utilise simplement la définition de X_b pour obtenir:
$X_b = \begin{pmatrix} 1 & X_1 & \ldots & X_k b & \ldots & X_p \end{pmatrix} = \begin{pmatrix} 1 \times 1 & 1 \times X_1 & \ldots & b \times X_k & \ldots & 1 \times X_p \end{pmatrix}$

$$\begin{pmatrix}1 & X_1 & \dots & X_k & \dots & X_p\end{pmatrix} \times \begin{pmatrix}1 & & & & & \\ & 1 & & & & \\ & & \ddots & & & \\ & & & b & & \\ & & & & \ddots & \\ & & & & & 1\end{pmatrix} = XD$$

D est donc la matrice identité de dimension $p+1$ dont l'entrée diagonale $k+1$ est remplacée par b.

b) Soit $\hat{\theta}_{b,n}$ l'estimateur OLS associé à X_b. Exprimer $\hat{\theta}_{b,n}$ en fonction de $\hat{\theta}_n$ et de D.

Par les équations normales, l'estimateur OLS $\hat{\theta}_{b,n}$ est égal à:

$\hat{\theta}_{b,n} = (X_b^T X_b)^{-1}(X_b^T y) = [(XD)^T(XD)]^{-1}[(XD)^T y] = [D^T X^T X D]^{-1}[D^T X^T y]$

$= [D^{-1}(X^T X)^{-1}(D^T)^{-1}]\ [D^T X^T y] = D^{-1}(X^T X)^{-1} X^T y = D^{-1}\hat{\theta}_n$

Autrement dit, l'estimateur $\hat{\theta}_{b,n}$ est égal à l'estimateur $\hat{\theta}_n$ dont le coefficient $k+1$ a été multiplié par $1/b$.

c) Donner la variance de $\hat{\theta}_{b,n}$.

On utilise simplement les propriétés de la variance (et le fait que D est diagonale) pour obtenir:

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

d) La prédiction donnée par le modèle est:

$\hat{y}_b = X_b \hat{\theta}_{b,n} = (XD)(D^{-1}\hat{\theta}_n) = X\hat{\theta}_n = \hat{y}$

Autrement dit, la prédiction n'est elle pas affectée par le changement d'échelle d'une des variables.

11) Donner une formule explicite du problème $argmin_\theta\ \frac{1}{2}(y - X\theta)^T \Omega (y - X\theta)$ pour une matrice $\Omega = diag(w_1, \dots, w_n)$ définie positive, dans le cas où X est de plein rang.

On commence par développer la forme quadratique:

$\frac{1}{2}(y - X\theta)^T \Omega (y - X\theta) = \frac{1}{2}\left[y^T \Omega y + \theta^T X^T \Omega X \theta - 2\theta^T X^T \Omega y\right]$

(où pour obtenir le terme $2\theta^T X^T \Omega y$ on a utilisé le fait qu'un scalaire est égal à sa transposée, et que Ω est symétrique).

Pour trouver l'argmin, on utilisera les règles suivantes de dérivées matricielles:

- Si a et b sont des vecteurs, on a: $\frac{\partial b^T a}{\partial b} = a$. Cela implique que: $\frac{\partial 2\theta^T X^T \Omega y}{\partial \theta} = 2X^T \Omega y$.
- Si A est une matrice symétrique et b un vecteur, on a: $\frac{\partial b^T A b}{\partial b} = 2Ab$. Cela implique que: $\frac{\partial \theta^T X^T \Omega X \theta}{\partial \theta} = 2X^T \Omega X \theta$.

On conclut que:

$$\begin{aligned}
&\frac{\partial}{\partial \theta}\left(\frac{1}{2}(y - X\theta)^T \Omega (y - X\theta)\right) = 0 \\
\Leftrightarrow\ &\frac{\partial}{\partial \theta}\left(\frac{1}{2}\left[y^T \Omega y + \theta^T X^T \Omega X \theta - 2\theta^T X^T \Omega y\right]\right) = 0 \\
\Leftrightarrow\ &\frac{1}{2}\left(2X^T \Omega y - 2X^T \Omega X \theta\right) = 0 \\
\Leftrightarrow\ &X^T \Omega y - X^T \Omega X \theta = 0 \\
\Leftrightarrow\ &X^T \Omega X \theta = X^T \Omega y \\
\Leftrightarrow\ &\hat{\theta} = (X^T \Omega X)^{-1}(X^T \Omega y)
\end{aligned}$$

12). Dans le cas du modéle de régression avec désign aléatoire, décrire l'asymptotique

de l'estimateur des moindre carrées. On donnera la loi asymptotique de $\sqrt{n}(\hat{\beta} - \beta^*)$.

(Le modéle *Random Design* n'a pas été traité en cours. En revanche, une question sur la loi asymptotique avec le modéle gaussien peut tomber).

- *Modéle gaussien*

Pour étudier la convergence de $\hat{\beta}$, on fait appel au théorème central limite (TCL). On se base sur le calcul du biais $\hat{\beta}$:

$$\hat{\beta} - \beta^* = (X^T X)^{-1} X^T \varepsilon$$

$\hat{\beta} - \beta^*$ est une combinaison linéaire certaine de lois indépendantes ε_i ce qui permet d'appliquer le TCL (condition 1).

Pour la variance, on suppose que l'hypothèse suivante est vérifié avec V_X une matrice finie definie-positive (les variables explicatives conservent de la variance quand $n \to +\infty$, soit plus d'observations apportent plus d'information ce qui exclut la possibilité d'une multicolinéarité stricte au niveau asymptotique) :

$$\lim_{n\to+\infty} \frac{1}{n}(X^T X)^{-1} = V_X$$

D'où (condition 2) :

$$\lim_{n\to+\infty} \mathbf{V}(\sqrt{n}(\hat{\beta} - \beta^*)) = \lim_{n\to+\infty} n\sigma^2 (X^T X)^{-1} = \lim_{n\to+\infty} \sigma^2 \Big(\frac{X^T X}{n}\Big)^{-1} = \sigma^2 V_X^{-1}$$

Conséquence, en partant du résultat que :

$$\sqrt{n}(\beta - \beta^*) \sim \mathcal{N}(0, \sigma(X^T X)^{-1})$$

On déduit que $\beta - \beta^*$ converge en loi vers :

$$\sqrt{n}(\beta - \beta^*) \to \mathcal{N}(0, \sigma V_X^{-1})$$

- *Modéle désign aléatoire*

(Hors programme)

13) Dans le cas du modéle de régression avec design déterministe et bruit gaussien centré de variance σ^2, donner la loi de l'estimateur des moindre carrées $\hat{\beta}$.

Dans le modéle gaussien, les perturbations (ou le bruit blanc) $(\varepsilon_i)_{i=1,\dots,n}$ sont des variables aléatoires réelles gaussiennes telles que : $\varepsilon_i \sim^{i.i.d} \mathcal{N}(0, \sigma^2)$ et en forme vectorielle $\varepsilon \sim^{i.i.d} \mathcal{N}(0, \sigma^2 I_n)$.

Or les variables à expliquer Y suivent aussi une loi gaussienne puisque $Y = X\beta^* + \varepsilon$ est une combinaison linéaire additive de variables aléatoires gaussiennes. D'où : $Y_i \sim \mathcal{N}(X_i^T \beta^*, \sigma^2)$.

$\hat{\beta}$ est une combinaison linéaire certaine des Y_i, d'où à l'instar de Y, le vecteur $\hat{\beta}$ suit aussi une loi normale d'espérance μ et de variance-covariance Σ. Le calcul du biais et de la variance de l'estimateur $\hat{\beta}$ nous donne ces 2 quantités :

$$\mathbf{E}(\hat{\beta}) = \beta^* \qquad \mathbf{V}(\hat{\beta}) = \sigma^2(X^TX)^{-1}$$

Ainsi :

$$\hat{\beta} \sim \mathcal{N}(\beta^*, \sigma^2(X^TX)^{-1})$$

Et en particulier :

$$j = 1, ..., p \qquad \hat{\beta}_j \sim \mathcal{N}(\beta_j^*, \sigma^2(X^TX)^{-1}_{j,j})$$

14) Dans le cas du modèle de régression avec design déterministe où X est de plein rang p, donner la valeur du risque de prédiction.

$$Rpred(\hat{\theta}_n) = E\left[\|Y^* - \hat{Y}\|_2^2\right]$$
$$Rpred(\hat{\theta}_n) = E\left[\|X(\hat{\theta}_n - \theta^*)\|_2^2\right]$$
$$Rpred(\hat{\theta}_n) = E\left[\|X(X^TX)^{-1}X^T\epsilon\|_2^2\right]$$

On pose $H_x = X(X^TX)^{-1}X^T$, on remarque que H_x est un projecteur orthogonal et on écrit :

$$Rpred(\hat{\theta}_n) = E\left[\epsilon^T H_x \epsilon\right]$$
$$= E\left[tr(H_x\epsilon\epsilon^T)\right]$$
$$= tr(H_x E(\epsilon\epsilon^T)$$

Comme $Cov(\epsilon) = \sigma^2 I_n$

$$= \sigma^2 tr(H_x)$$

Comme H_x est un projecteur orthogonal on a :

$$\begin{cases} H_x^T = H_x \\ H_x^2 = H_x \end{cases}$$

Donc les valeurs propres de λ_i de H_x on pour propriété :

$$\lambda_i^2 = \lambda_i \iff \begin{cases} \lambda_i = 0 \\ \lambda_i = 1 \end{cases}$$

Ainsi :

$$Rpred = \sigma^2 tr(H_x)$$
$$= \sigma^2 \sum_{k=1}^{n} \lambda_k$$

avec λ_k les vp de H_x

$$= \sigma^2 rang(H_x)$$

Avec l'hypothèse de rang plein $Ker(X) = \{0\}$ on a :

$$dim(Vect(X)) = p$$

H_x étant le projecteur orthgonal sur Vect(X), on obtient $rang(H_x) = p$ donc

$$Rpred(\hat{\theta}_n) = \sigma^2 p$$

4. Ridge:

On note $\hat{\theta} = \arg\min_{\theta} \dfrac{1}{2}\|\mathbf{y} - X\theta\|_2^2 + \dfrac{\lambda}{2}\|\theta\|_2^2$ l'estimateur Ridge
Soit $f : \theta \mapsto \frac{1}{2}||Y - X\theta||_2^2 + \frac{\lambda}{2}||\theta||_2^2$

1) Quand $X = Id_n$ on a $n = p + 1$ et $f(\theta) = \frac{1}{2}||Y - Id_n\theta||_2^2 + \frac{\lambda}{2}||\theta||_2^2$

Pour des questions de notations, on utilisera $Id_n = Id_{p+1} = Id$

$f(\theta) = \frac{1}{2}\Sigma_{i=1}^n (Y_i - Id_i\theta)^2 + \frac{\lambda}{2}\Sigma_{i=1}^n \theta_i^2$

Le minimum de la fonction est atteint lorsque $\nabla f(\theta) = 0$

$\nabla f(\theta) = 0 => \forall k = 1, ..., p : \frac{\partial f(\theta)}{\partial \theta_k} = 0$

$\frac{\partial f(\theta)}{\partial \theta_k} = 2\frac{1}{2}(-Id_{i,k})\Sigma_{i=1}^n (Y_i - \Sigma_{j=1}^p Id_{i,j}\theta_j) + 2\frac{\lambda}{2}\theta_k$

$= -\Sigma_{i=1}^n (Id_{i,k}Y_i - Id_{i,k}Id_i\theta) + \lambda\theta_k$

Donc $\nabla f(\theta) = -\Sigma_{i=1}^n (Id_i^T Y_i - Id_i^T Id_i\theta) + \lambda\theta = -\Sigma_{i=1}^n Id_i^T Y_i + \Sigma_{i=1}^n Id_i^T Id_i\theta + \lambda\theta$

$= -Y + \theta + \lambda\theta$

$\nabla f(\theta) = 0 => \theta = \frac{1}{1+\lambda}Y = \hat{\theta}_{n,\lambda}^{Ridge}$

2) Pour $X \in \mathbb{R}^{n\times p}$ quelconque, $f(\theta) = \frac{1}{2}\Sigma_{i=1}^n (Y_i - X_i\theta)^2 + \frac{\lambda}{2}\Sigma_{i=1}^n \theta_i^2$

Comme pour la question 1),
$\nabla f(\theta) = 0 => \forall k = 1, ..., p : \frac{\partial f(\theta)}{\partial \theta_k} = 0$

$\frac{\partial f(\theta)}{\partial \theta_k} = 2\frac{1}{2}(-X_{i,k})\Sigma_{i=1}^n (Y_i - \Sigma_{j=1}^p X_{i,j}\theta_j) + 2\frac{\lambda}{2}\theta_k$

$= -\Sigma_{i=1}^n (X_{i,k}Y_i - X_{i,k}X_i\theta) + \lambda\theta_k$

Donc $\nabla f(\theta) = -\Sigma_{i=1}^n (X_i^T Y_i - X_i^T X_i\theta) + \lambda\theta = -X^T Y + X^T X\theta + \lambda\theta$

Ainsi $\nabla f(\theta) = 0 => -X^T Y + X^T X\theta + \lambda\theta = 0$
$=> (X^T X + \lambda Id_n)\theta = X^T Y => \theta = (X^T X + \lambda Id_n)^{-1} X^T Y = \hat{\theta}_{n,\lambda}^{Ridge}$

3) Donner la variance de l'estimateur Ridge sous l'hypothèse que le bruit $\mathbf{y} - X\theta^*$ est centré et de variance $\sigma^2 \mathrm{Id}_n$.

$$\hat{\theta}_\lambda^{rdg} = (X^\top X + \lambda Id_p)^{-1} X^\top Y$$
$$Var(\hat{\theta}_\lambda^{rdg}) = Var((X^\top X + \lambda Id_p)^{-1} X^\top Y)$$
$$= ((X^\top X + \lambda Id_p)^{-1} X^\top) Var(Y) ((X^\top X + \lambda Id_p)^{-1} X^\top)^\top$$
$$= (X^\top X + \lambda Id_p)^{-1} X^\top Var(X\theta^\star + \epsilon) X (X^\top X + \lambda Id_p)^{-1}$$
$$= (X^\top X + \lambda Id_p)^{-1} X^\top Var(\epsilon) X (X^\top X + \lambda Id_p)^{-1}$$
$$= \sigma^2 (X^\top X + \lambda Id_p)^{-1} X^\top X (X^\top X + \lambda Id_p)^{-1}$$

En remarquant que $X^\top X$ et $(X^\top X + \lambda Id_p)^{-1}$ sont diagonalisables dans la même base, on pose:

$$(X^\top X + \lambda Id_p)^{-1} = PA^{-1}P^{-1}$$
$$X^\top X = PBP^{-1}$$

Avec A^{-1} et B deux matrices diagonales.
On a alors:

$$Var(\hat{\theta}_\lambda^{rdg}) = \sigma^2 PA^{-1}P^{-1}PBP^{-1}PA^{-1}P^{-1}$$
$$= \sigma^2 PA^{-1}BA^{-1}P^{-1}$$
$$= \sigma^2 PA^{-1}A^{-1}BP^{-1}$$
$$= \sigma^2 PA^{-1}Id_p A^{-1} Id_p BP^{-1}$$
$$= \sigma^2 PA^{-1}P^{-1}PA^{-1}P^{-1}PBP^{-1}$$
$$= \sigma^2 (X^\top X + \lambda Id_p)^{-1} (X^\top X + \lambda Id_p)^{-1} X^\top X$$
$$= \sigma^2 (X^\top X + \lambda Id_p)^{-2} X^\top X$$

5. Lasso :

1) Exprimer $\eta_\lambda(z) = argmin_{\ x\in\mathbb{R}}\ x \mapsto \frac{1}{2}(z-x)^2 + \lambda|x|$ en fonction du signe de z et de la partie positive $(.)_+$.

Posons
$$f(x) = \frac{1}{2}(z-x)^2 + \lambda x$$

avec z et λ cte, $\lambda \geqslant 0$

<u>1^{er} cas :</u> $x \geqslant 0$

$$f(x) = \frac{1}{2}(z-x)^2 + \lambda x$$
$$f'(x) = -(z-x) + \lambda$$

alors

$$f'(x) = 0 \Leftrightarrow x = z - \lambda$$

Ce qui est vérifié dans notre cas si $z \geqslant \lambda$.
Sinon, si $z < \lambda$, alors

$$\underset{x}{argmin}\ f(x) = \underset{x}{argmin}\ \frac{1}{2}x^2 - (z - \lambda)x + z^2$$

Cette parabole atteint son minimum sur $\mathbb{R}^*$ en $x = 0$ puisque les racines sont 0 et $2(z - \lambda)$. Ainsi, pour $x \geqslant 0$

$$\begin{cases} x = z - \lambda \ \ si\ z \geqslant \lambda \\ x = 0 \ \ sinon \end{cases}$$

<u>2ème cas :</u> $x \leqslant 0$

$$f(x) = \frac{1}{2}(z - x)^2 - \lambda x$$

$$f'(x) = -(z - x) - \lambda$$

alors

$$f'(x) = 0 \Leftrightarrow x = z + \lambda$$

Ce qui est vérifié dans notre cas si $z \leqslant -\lambda$.
Sinon, si $z > -\lambda$, alors on reprend notre raisonnement sur la parabole qui atteindra de la même manière son minimum en $x = 0$.
Ainsi, pour $x \leqslant 0$

$$\begin{cases} x = z + \lambda \ \ si\ z \leqslant -\lambda \\ x = 0 \ \ sinon \end{cases}$$

<u>Conclusion :</u>

- $\boldsymbol{z \geqslant \lambda}$, $\eta_\lambda(z) = z - \lambda = |z| - \lambda = \boldsymbol{sign(z)\ (|z| - \lambda)}$
- $\boldsymbol{|z| \leqslant \lambda}$, $\boldsymbol{\eta_\lambda(z) = 0}$
- $\boldsymbol{z \leqslant -\lambda}$, $\eta_\lambda(z) = z + \lambda = -|z| + \lambda = -(|z| - \lambda) = sign(z)\ (|z| - \lambda)$ d'où
  $\boldsymbol{\eta_\lambda(z) = sign(z)\ (|z| - \lambda)_+}$

7. Test:

1) Pour des X1,...,Xn identiquement distribuées à valeur dans 0,1 décrire une procédure de test de l'hypothèse $p = P(X_1 = 1) = 1/2$ contre son contraire.
On pose:

$$\begin{cases} H_0 : p = P(X_1) = \frac{1}{2} \\ H_1 : p = P(X_1 = 1) \neq \frac{1}{2} \end{cases}$$

On choisit comme statistique de test

$$T_i = \sqrt{n}\frac{\hat{p} - p}{\hat{\sigma}}$$

avec l'etismateur de l'espérance de X

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

, $p = \frac{1}{2}$
et

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{p})^2 = \hat{p} - \hat{p}^2$$

On suppose n assez grand pour que

$$T_i \sim \mathcal{N}(0,\, 1)\,.$$

On note α notre niveau de précision. Pour ne pas rejeter l'hypothèse H_0, on doit avoir:

$$\sqrt{n}\frac{\hat{p} - p}{\hat{\sigma}} <= t_{1-\frac{\alpha}{2}}$$

avec $t_{1-\frac{\alpha}{2}}$ le quantile de la loi normale centrée réduite
On a la région de rejet R:

$$R = [-t_{1-\frac{\alpha}{2}}; t_{1-\frac{\alpha}{2}}]$$