

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

$$\mathcal{M} = \{\mathbb{P}_\theta : \theta \in \mathbb{R}^6, \sum_{i=1}^6 \theta_i = 1\}.$$

4) Soit x_1, x_2, \dots, 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: $\hat{\mu} = \underset{\mu}{argmin} \sum_{i=1}^n (x_i - \mu)^2$. La condition de premier ordre est:

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

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

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

Sa variance est donnée par:

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

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:

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

Il découle ensuite du théorème de Cochran que:

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

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:

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

Le biais est donc de:

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

6) On suppose que l'on observe y_1, \dots, 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:

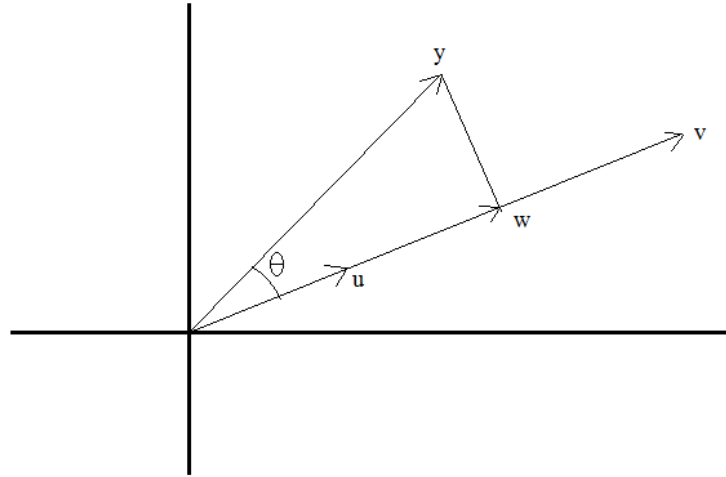
$$\text{var}\left(\frac{n}{\sigma^2} \tilde{\sigma}^2\right) = 2(n-1) \Leftrightarrow \frac{n^2}{\sigma^4} \text{var}(\tilde{\sigma}^2) = 2(n-1) \Leftrightarrow \text{var}(\tilde{\sigma}^2) = \frac{2\sigma^4(n-1)}{n^2}$$

On utilise alors la définition du risque quadratique pour obtenir:

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

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

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



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

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

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

On obtient que $\text{var}_n(\mathbf{y}) = 0$ si et seulement si $\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_n)^2 = 0$. Cela 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 $y_1 = y_2 = \dots = y_n = y$, pour y un scalaire donné. On conclut donc que ces vecteurs sont de la forme $\mathbf{y} = y \cdot \mathbf{1}_n$.

2. Moindres carrés unidimensionnels:

On observe $\mathbf{y} = (y_1, \dots, y_n)^T$ et $\mathbf{x} = (x_1, \dots, x_n)^T$.

1) La fonction $(\theta_0, \theta_1) \rightarrow \frac{1}{2} \sum_{i=1}^n (y_i - \theta_0 - \theta_1 x_i)^2$ 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:

$$\begin{aligned}\frac{\partial f}{\partial \theta_0} &= -2(y_i - \theta_0 - \theta_1 x_i) \\ \frac{\partial^2 f}{\partial \theta_0^2} &= 2 \\ \frac{\partial f}{\partial \theta_1} &= -2x_i(y_i - \theta_0 - \theta_1 x_i) \\ \frac{\partial^2 f}{\partial \theta_1^2} &= 2x_i^2 \\ \frac{\partial^2 f}{\partial \theta_1 \partial \theta_0} &= 2x_i\end{aligned}$$

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 fonction f est convexe. Comme la fonction $(\theta_0, \theta_1) \rightarrow \frac{1}{2} \sum_{i=1}^n (y_i - \theta_0 - \theta_1 x_i)^2$ est une somme de fonctions convexes, elle est convexe elle-même.

3. Moindres carrés:

10) On suppose que X est de rang plein et on note $\hat{\theta}$ l'estimateur OLS. On note $\tilde{X} = (X_1, \dots, X_p)$. On change l'échelle d'une des variables: X_k est remplacé par $X_k b$, où $b > 0$.

a) Soit $X_b = (1, X_1, \dots, X_k b, \dots, X_p)$. 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:

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

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

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:

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

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 $\Omega = \text{diag}(w_1, \dots, w_n)$ définie positive, dans le cas où X est de plein rang.

On commence par développer la forme quadratique:

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

(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:

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

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.

$$R_{pred}(\hat{\theta}_n) = E [\|Y^* - \hat{Y}\|_2^2]$$

$$R_{pred}(\hat{\theta}_n) = E [\|X(\hat{\theta}_n - \theta^*)\|_2^2]$$

$$R_{pred}(\hat{\theta}_n) = E [\|X(X^T X)^{-1} X^T \epsilon\|_2^2]$$

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

$$\begin{aligned} R_{pred}(\hat{\theta}_n) &= E [\epsilon^T H_x \epsilon] \\ &= E [\text{tr}(H_x \epsilon \epsilon^T)] \end{aligned}$$

$$= tr(H_x E(\epsilon \epsilon^T))$$

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 ont 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 orthogonal sur $Vect(X)$, on obtient $rang(H_x) = p$ donc

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

7. Test:

1) Pour des X_1, \dots, X_n indépendamment 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:

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

On choisit comme statistique de test

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

avec l'estimateur 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:

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

avec $t_{1-\frac{\alpha}{2}}$ le quantile de la loi normale centrée réduite

On a la région de rejet R :

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