Cours du 03/04/2020.

Application 2 : Moindre carrés, régression linéaires + variantes

1. Projection orthogonale.

Rappel. E, F supplémentaires si $\mathbb{R}^n = E \oplus F$
$\rightsquigarrow p_{E/F}$ = la proj. sur E parallèlement à F.

(i) Produit scalaire et norme

$\vec{u} = (x_1, \dots x_n)$, $\vec{v} = (y_1, \dots y_n) \in \mathbb{R}^n$,

- $\langle \vec{u}, \vec{v} \rangle = x_1 y_1 + x_2 y_2 + \dots + x_n y_n \in \mathbb{R}$
  produit scalaire de $\vec{u}$ et $\vec{v}$.
- $(\langle \vec{u}, \vec{u} \rangle)^{1/2} = \sqrt{x_1^2 + x_2^2 + \dots x_n^2} \in \mathbb{R}^+$
  norme (euclidienne) de $\vec{u}$.

(ii) $\vec{u} = \vec{0} \iff \langle \vec{u}, \vec{u} \rangle = 0$,
$\vec{u} \perp \vec{v} \iff \langle \vec{u}, \vec{v} \rangle = 0$.

Déf E un sev de $\mathbb{R}^n$, on définit le sev orthogonal à E, $E^\perp$, comme l'ensemble des vecteurs orthogonaux à E :

$$E^\perp = \{ \vec{v} \in \mathbb{R}^n \mid \forall \vec{u} \in E,\ \langle \vec{u}, \vec{v} \rangle = 0 \}.$$

Thm E sev de $\mathbb{R}^n$. $\mathcal{F} = (\vec{u}_1, \dots \vec{u}_p)$

1. $E^\perp$ sev de $\mathbb{R}^n$,
2. Si $\mathcal{F}$ engendre E (i.e $E = \mathrm{vect}(\mathcal{F})$)
$$E^\perp = \{ \vec{v} \in \mathbb{R}^n \mid \forall \vec{u}_i \in \mathcal{F},\ \langle \vec{u}_i, \vec{v} \rangle = 0 \}.$$
3. $\mathbb{R}^n = E \oplus E^\perp$ ($E, E^\perp$ supplémentaires).

dém. 1 et 2 viennent de la linéarité du produit scalaire ($\vec{u}$ fixé, $\vec{v} \longmapsto \langle \vec{u}, \vec{v} \rangle$).
3. admis. □

Déf Soit E un sev, on définit la projection orthogonale de $\mathbb{R}^n$ sur E, notée p_E, la projection de $\mathbb{R}^n$ sur E parallèlement à $E^\perp$.

i.e $p_E = p_{E/E^\perp}$.

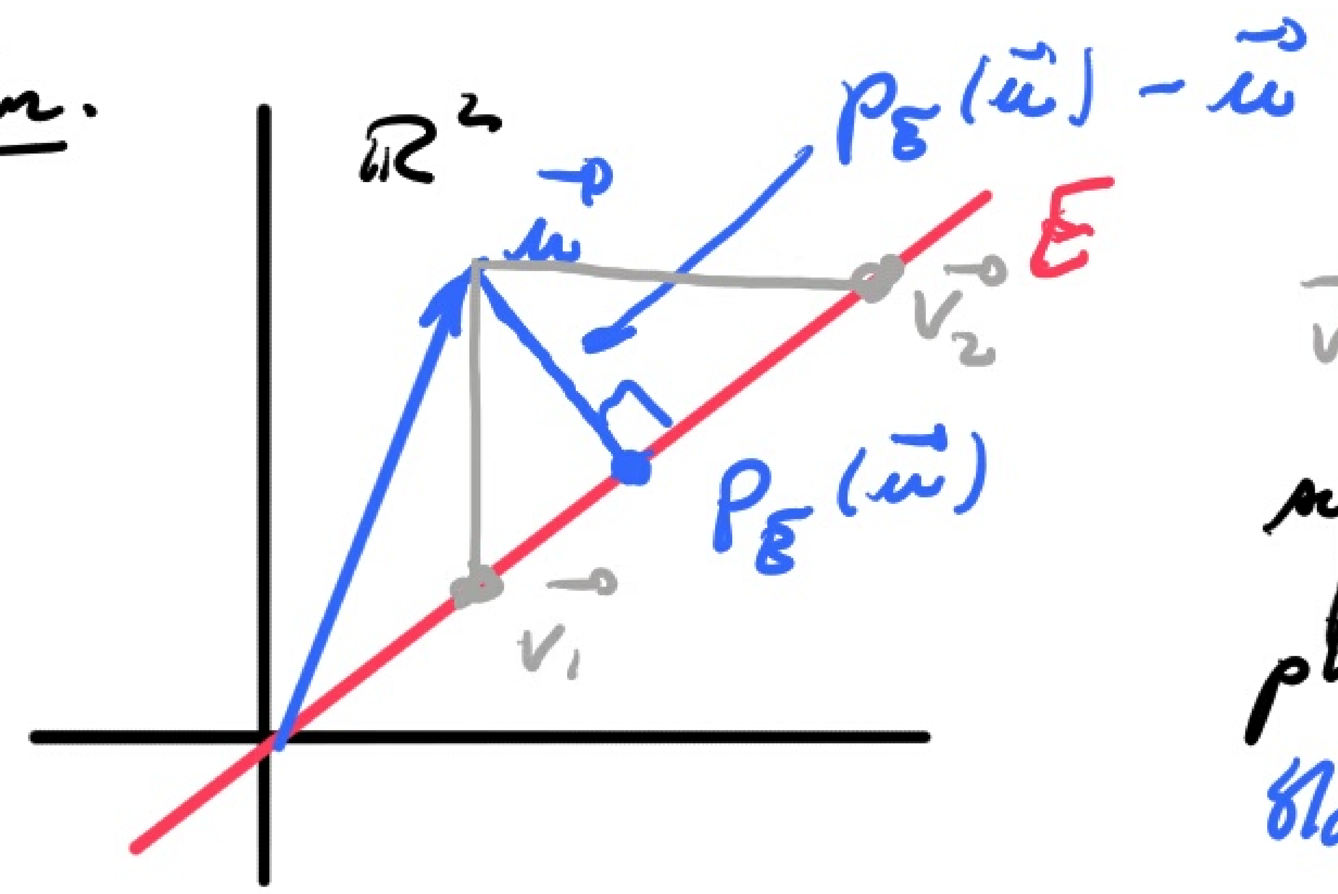
Thm (Pythagore). Si $\vec{u} \perp \vec{v}$, alors
$$\|\vec{u}+\vec{v}\|^2 = \|\vec{u}\|^2 + \|\vec{v}\|^2.$$

__dém.__
$$\|\vec{u}+\vec{v}\|^2 = \langle \vec{u}+\vec{v}, \vec{u}+\vec{v} \rangle$$
$$= \langle \vec{u}, \vec{u} \rangle + \underbrace{\langle \vec{u}, \vec{v} \rangle}_{0} + \underbrace{\langle \vec{v}, \vec{u} \rangle}_{0} + \langle \vec{v}, \vec{v} \rangle$$
$$= \|\vec{u}\|^2 + \|\vec{v}\|^2 \qquad \square$$

__Caractérisation géométrique de la proj. orthogonale.__

Thm E un sev de $\mathbb{R}^n$, p_E proj. orthog. sur E.
Soit $\vec{u} \in \mathbb{R}^n$, $p_E(\vec{u})$ est __le__ vecteur de E le plus proche de $\vec{u}$: (le "moindre carré")
$$\forall \vec{v} \in E, \ \|\vec{v} - \vec{u}\|^2 \geq \boxed{\|p_E(\vec{u}) - \vec{u}\|^2}.$$
avec égalité si et __seulement si__ $\vec{v} = p_E(\vec{u})$.

__dém.__



Par Pythagore, $\vec{v_1}, \vec{v_2} \in E$, les segments gris sont tous plus longs que le segment bleu sauf si $\vec{v} = p_E(\vec{u})$.

$$\vec{v} - \vec{u} = \underbrace{(\vec{v} - p_E(\vec{u}))}_{\in E} + \underbrace{(p_E(\vec{u}) - \vec{u})}_{\in E^\perp}$$

Donc par Pythagore : $\quad \vec{u} = p_E(\vec{u}) + (\vec{u} - p_E(\vec{u}))$
$$\|\vec{v} - \vec{u}\|^2 = \underbrace{\|\vec{v} - p_E(\vec{u})\|^2}_{} + \|p_E(\vec{u}) - \vec{u}\|^2.$$
$$\geq 0 \text{ et } = 0 \text{ ssi } \vec{v} = p_E(\vec{u}) \qquad \square.$$

2. Régression linéaire.

__Idée__ : Étant données 2 séries de données de même taille $(x_1, \dots x_n)$ et $(y_1, \dots y_n)$, on veut déterminer la droite de $\mathbb{R}^2$ la plus proche $\{(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)\}$.

__Si__ $\forall i, \ x_i = x_0$ constante, alors la droite verticale d'équation $x = x_0$ passe par tous les points.

__Sinon__ on cherche $a, b \in \mathbb{R}$ tq $Y - (aX + bU)$ soit minimal, où $U = (1, \text{—} 1)$ en colonne !

On forme la matrice $M = (X \mid U)$ et on veut minimiser $\|M\binom{a}{b} - Y\|^2$.

$C = \begin{pmatrix} a \\ b \end{pmatrix}$ "vecteur coefficients" est cherché tq $\|MC - Y\|^2$ est minimal.

Rmq. MC = élément de l'image de M.

i.e on cherche l'élément de l'image de M le plus proche de Y

i.e on cherche le projeté orthogonal de Y sur l'image de M :

$$\boxed{C \text{ tq } MC = P_{Im(M)}(Y)}$$

Il y a des formules (horribles !) en fonction $x_1, \dots, x_n, y_1, \dots, y_n$ pour déterminer C.

En pratique, méthode lstsq de np.linalg

(LeaST SQuares = moindres carrés).

- import numpy as np
- X, Y donnés, $M = (X \mid U)$, Y colonne. $\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$
- C = np.linalg.lstsq(M, Y, rcond=None)[0]

éviter mess. error

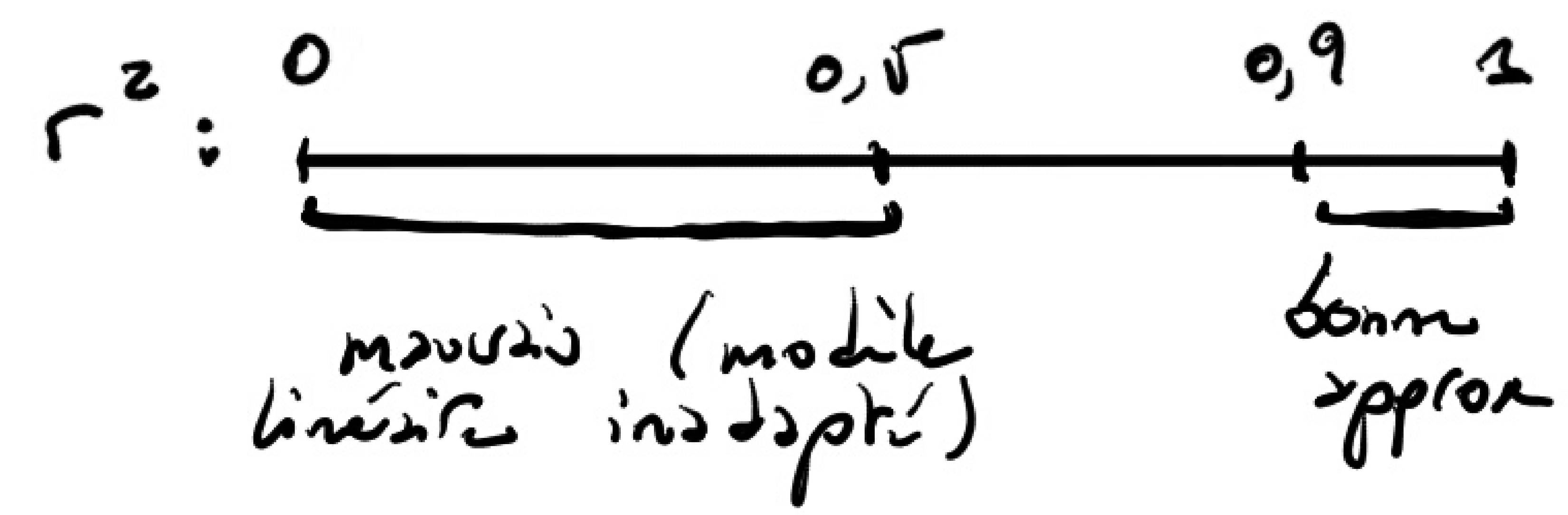
renvoie [[a],[b]].

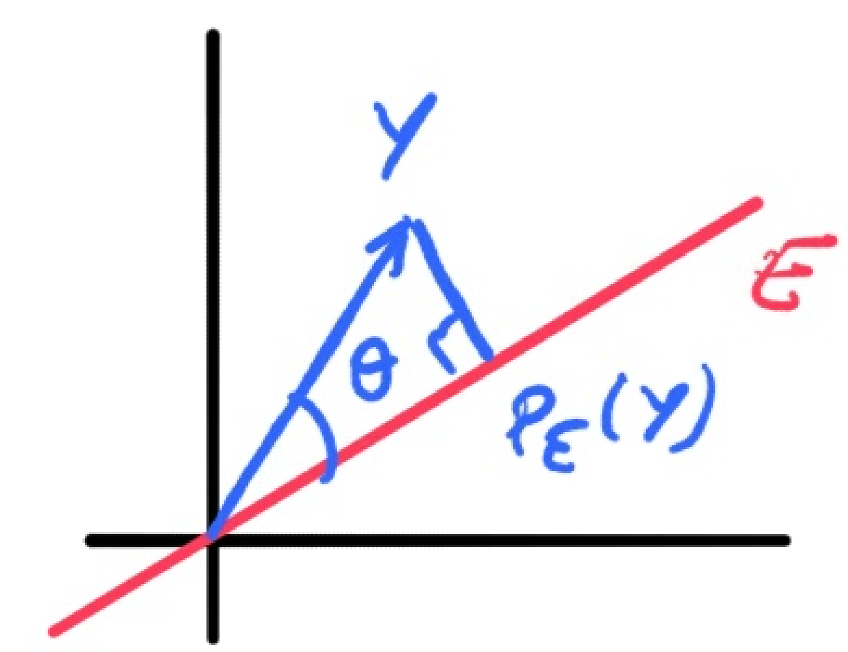
- a, b = C[:,0] ⇝ $y = ax + b$ est la droite de la régression linéaire.

E l'ensemble des $\tilde{Y}$ obtenus de X via une transformation affine $aX + bU$.

3. Qualité de l'approximation.

déf. Le coefficient de corrélation de Y par rapport à E est le nombre $r = \cos\theta \in [0,1]$ où $\cos\theta = \dfrac{\|P_E(Y)\|}{\|Y\|}$.





Dans notre cas, X, Y ⇝ C tq $MC = P_E(Y)$ Im M

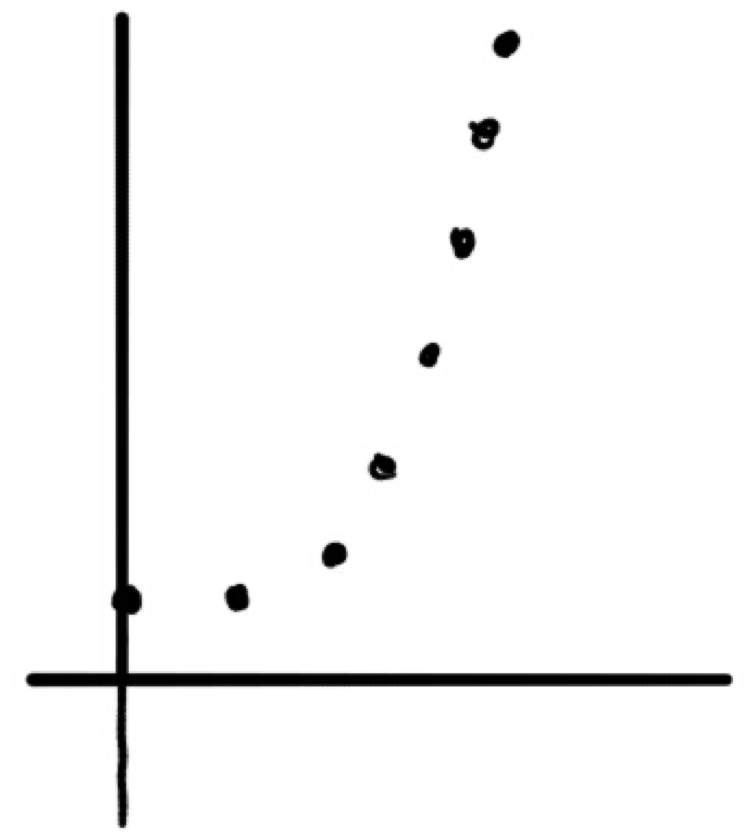
Donc $P_E(Y) = M \cdot C$ (P = np.dot(M,C))

et r^2 = np.dot(P.transpose(), P) / np.dot(Y.transpose(), Y).

$\left(\langle \vec{u}, \vec{u} \rangle = U^T \cdot U = (x_1, \dots, x_n) \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \dots \right)$

4. Variantes.

a. Vous représentez les données, s'il est clair que la régression linéaire est inadaptée.



b. On trouve a, b, et $r^2 < 0,5$

$\to$ changer de modèle.

oui mais en gardant la puissance des moindres carrés.

(i) Approximation quadratique. $y_i \simeq a x_i^2 + b x_i + c$

On forme $M = (X^2 \,|\, X \,|\, U)$.

$X^2 = \begin{pmatrix} x_1^2 \\ \vdots \\ x_n^2 \end{pmatrix}$

$(aX^2 + bX + cU) - Y$

Trouver $C = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$ tq $\|MC - Y\|^2$ est minimal.

(ii) Approximation sinusoïdale. $y_i \simeq a + b\cos(\omega x_i) + c\sin(\omega x_i)$.

$M = (U \,|\, \text{"}\cos(\omega X)\text{"} \,|\, \text{"}\sin(\omega X)\text{"})$

$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ $\cos(\omega X) = \begin{pmatrix} \cos(\omega x_1) \\ \vdots \\ \cos(\omega x_n) \end{pmatrix}$ $\sin(\omega X) = \begin{pmatrix} \sin(\omega x_1) \\ \vdots \\ \sin(\omega x_n) \end{pmatrix}$

On cherche $C = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$ tq $\|MC - Y\|^2$ minimal.

(iii) Approximation exponentielle. $y_i \simeq k e^{\alpha x_i}$

- On applique ln : $\ln(y_i) = \ln(k) + \ln(e^{\alpha x_i})$
  $= \underbrace{\ln k}_{=b} + \underbrace{\alpha x_i}_{=a}$

$M = (X \,|\, U)$

$C = \begin{pmatrix} a \\ b \end{pmatrix}$ avec $Y \to$ "$\ln Y$" $= \begin{pmatrix} \ln y_1 \\ \vdots \\ \ln y_n \end{pmatrix}$.

$C =$ lstsq ✓ et $\alpha = a$ et $k = e^b$.

(iv) Plus de variables. $P_i \simeq a + bT_i + cA_i + dG_i$

poids, taille, âge, genre

$M = (U \,|\, T_i \,|\, A_i \,|\, G_i)$, on cherche $C = \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}$

tq $\|MC - Y\|^2$ minimal. ✓