Cours 3 : Corrélation, régression linéaire

Essentiel du cours 2

Définitions

- Moyenne :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Variance :

$$\mathrm{var}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- Écart-type : $\sigma_{\mathbf{x}} = \sqrt{\mathrm{var}(\mathbf{x})}$

Propositions

- Transformation linéaire $y_i = a + bx_i$ $(1 \leq i \leq n)$:

$$\overline{y} = a + b\overline{x} \qquad \mathrm{var}(\mathbf{y}) = b^2\,\mathrm{var}(\mathbf{x}).$$

- Formule alternative :

$$\mathrm{var}(\mathbf{x}) = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i)^2\right) - (\overline{x})^2$$

Définitions

- $x_{(i)}$ est la statistique d'ordre i après avoir ré-ordonné les x_i :

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(i)} \leq \cdots \leq x_{(n)}$$

- Médiane :

$$m = x_{(k+1)} \text{ si } n = 2k+1$$

$$m = \frac{x_{(k)} + x_{(k+1)}}{2} \text{ si } n = 2k$$

- Déviation absolue (par rapport à la médiane) :

$$\mathrm{mad}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}|x_i - m|$$

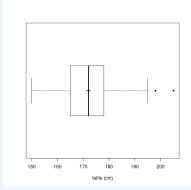
Exemple

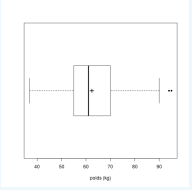
Taille et poids d'une population d'étudiants (AgroParisTech, 1993)

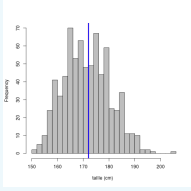
- $n = 731$ étudiant(e)s
- x_i = taille (cm) du i-ème étudiant
- y_i = poids (kg) du i-ème étudiant
- Données :

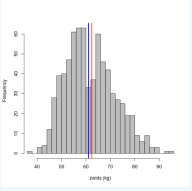
i	Taille	Poids
1	168	60
2	178	57
3	165	46
4	160	50
5	162	56
6	160	58
⋮	⋮	⋮

Taille et poids d'une population d'étudiants (AgroParisTech, 1993)







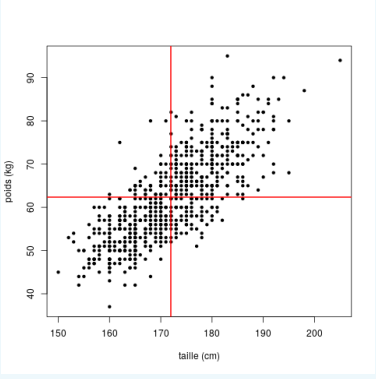


	moyenne	médiane	variance	écart-type
taille	171.96	172	80.11	8.95
poids	62.36	61	100.71	10.04

Exemple

Taille et poids d'une population d'étudiants (AgroParisTech, 1993)

Lien entre les deux variables (x = taille, y = poids)



Covariance

Données :

- Ensemble des observations $\{(x_i, y_i)\}_{1 \leq i \leq n}$
- Vecteurs des observations $\mathbf{x} = [x_1 \ \dots \ x_i \ \dots \ x_n]$, $\mathbf{y} = [y_1 \ \dots \ y_i \ \dots \ y_n]$

Définition 1 (Covariance)

$$\text{cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$$

Remarques :

- la covariance peut être négative, positive ou nulle ;
- elle s'exprime dans l'unité produit des deux variables $\mathbf{x}$ et $\mathbf{y}$;
- $\text{var}(\mathbf{x}) = \text{cov}(\mathbf{x}, \mathbf{x})$.

Transformation de la covariance

Proposition 1 (Transformation linéaire de la covariance)

Si on applique les transformations linéaires $u_i = a + bx_i$ et $v_i = c + dy_i$ pour $1 \leq i \leq n$, la covariance des données tranformées vaut

$$\operatorname{cov}(\mathbf{u}, \mathbf{v}) = bd \operatorname{cov}(\mathbf{x}, \mathbf{y}).$$

Proposition 2 (Écriture vectorielle de la covariance)

$$\operatorname{cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \langle \mathbf{x} - \overline{x}\mathbf{1}, \mathbf{y} - \overline{y}\mathbf{1} \rangle .$$

Transformation de la covariance

Proposition 3 (Encadrement de la covariance)

Pour tous $\mathbf{x}$ *et* $\mathbf{y}$*, on a*

$$(\mathrm{cov}(\mathbf{x},\mathbf{y}))^2 \leq \mathrm{var}(\mathbf{x})\,\mathrm{var}(\mathbf{y})$$

c'est-à-dire

$$-\sqrt{\mathrm{var}(\mathbf{x})\,\mathrm{var}(\mathbf{y})} \leq \mathrm{cov}(\mathbf{x},\mathbf{y}) \leq \sqrt{\mathrm{var}(\mathbf{x})\,\mathrm{var}(\mathbf{y})}.$$

Proposition 4 (Inégalité de Cauchy-Schwartz)

Pour tous vecteurs $\mathbf{u}$ *et* $\mathbf{v}$*, on a*

$$\langle \mathbf{u},\mathbf{v}\rangle^2 \leq \|\mathbf{u}\|^2\|\mathbf{v}\|^2.$$

Démonstration de l'inégalité de Cauchy Schwartz en TD.

Encadrement de la covariance

Coefficient de corrélation (linéaire)

La proposition 3 fournit un moyen de pallier les difficultés d'interprétation de la valeur de la covariance $\mathsf{cov}(\mathbf{x}, \mathbf{y})$.

Définition 2 (Coefficient de corrélation)

On définit donc le *coefficient de corrélation* comme

$$\mathsf{cor}(\mathbf{x}, \mathbf{y}) = \frac{\mathsf{cov}(\mathbf{x}, \mathbf{y})}{\sqrt{\mathsf{var}(\mathbf{x})\,\mathsf{var}(\mathbf{y})}}.$$

Proposition 5 (Encadrement de la corrélation)

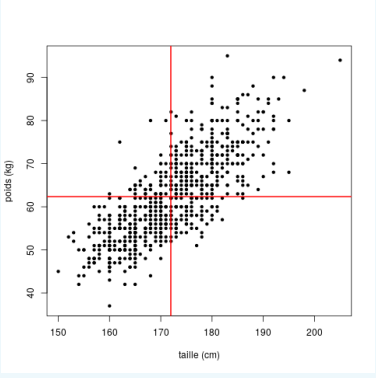
$$-1 \leq \frac{\mathsf{cov}(\mathbf{x}, \mathbf{y})}{\sqrt{\mathsf{var}(\mathbf{x})\,\mathsf{var}(\mathbf{y})}} \leq +1.$$

Le coefficient de corrélation est donc une grandeur

- sans dimension et
- bornée par deux valeurs de références (-1 et $+1$).

Taille et poids d'une population d'étudiants (AgroParisTech, 1993)

Lien entre les deux variables (x = taille, y = poids)



$\text{var}(\mathbf{x})$	$\text{var}(\mathbf{y})$	$\text{cov}(\mathbf{x},\mathbf{y})$	$\text{cor}(\mathbf{x},\mathbf{y})$
80.11	100.71	70.78	0.79

On souhaite maintenant :

- décrire la relation entre les variables x et y ;
- on recherche une transformation simple de x (en l'occurrence linéaire) qui donnerait une valeur proche de y
- il s'agit donc de déterminer deux nombres a et b tels que pour chaque $1 \leq i \leq n$:

$$y_i \simeq a + bx_i$$

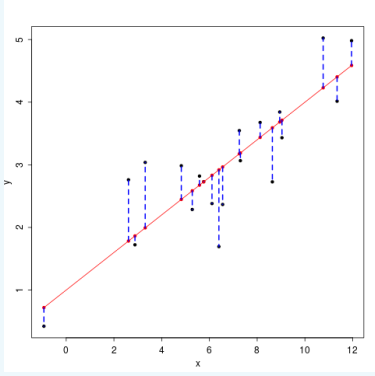
Définition 3 (Critère des moindres carrés)

$$C(a, b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Critère des moindres carrés

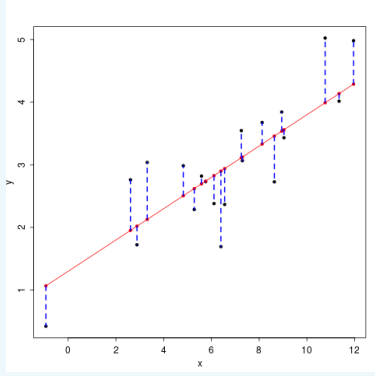
Exemple fictif ($n = 20$)

$a = 1, b = .3$



$C(a, b) = 6.62$

$a = 1.3, b = .25$



$C(a, b) = 6.84$

Remarques :

- le problème est posé de façon asymétrique ($y_i \simeq a + bx_i$ et non $x_i = a' + b'y_i$) ;
- le critère $\sum_{i=1}^{n} |y_i - a - bx_i|$ donne lieu à des calculs beaucoup moins simples.

Proposition 6 (Minimisation du critère des moindres carrés)

Les valeurs a^* et b^* qui minimisent le critère des moindres carrés $C(a, b)$ sont

$$b^* = \frac{\text{cov}(\mathbf{x}, \mathbf{y})}{\text{var}(\mathbf{x})}, \qquad a^* = \overline{y} - b^* \overline{x}.$$

Remarques :

- La droite d'équation $y = a^* + b^* x$ est appelée *droite de régression* de y sur x.
- Le coefficient b^* est sa *pente* ou *coefficient directeur*.
- Le coefficient a^* est son *ordonnée à l'origine* (en anglais *intercept*).

Minimisation du critère

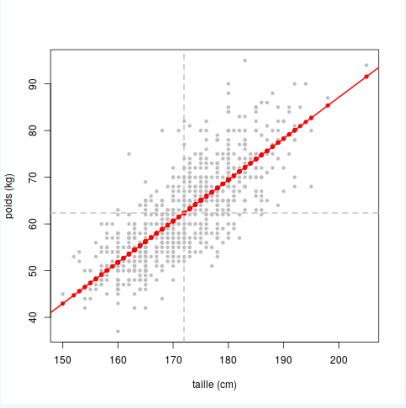
Minimisation du critère

Minimisation du critère

Exemple

Taille et poids.

Régression du poids sur la taille :



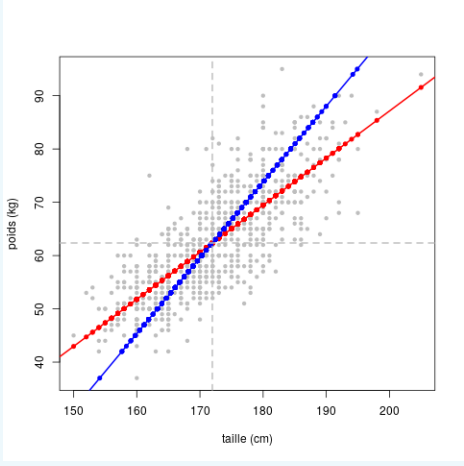
$x =$ taille, $y = poids$

a^*	b^*
-89.58	0.88

Remarques :

1. La droite de régression passe par la point moyen $(\overline{x}, \overline{y})$.
2. Les formules de a^* et b^* ne sont pas symétriques en x et y.

Taille et poids.



	a^*	b^*
poids / taille	-89.58 kg	0.88 kg/cm
taille / poids	128.13 cm	0.7 cm/kg

On souhaite souvent évaluer la qualité de l'approximation

$$y_i \simeq a + bx_i.$$

Définition 4 (Critère du $R^2(a, b)$)

Soit $\widetilde{\mathbf{y}}$ le vecteur de coordonnées $\widetilde{y}_i = a + bx_i$ $(1 \leq i \leq n)$, on mesure l'ajustement de l'approximation linéaire $y_i \simeq a + bx_i$ par le carré du coefficient de corrélation entre $\widetilde{\mathbf{y}}$ et $\mathbf{y}$:

$$R^2(a, b) = \left(\text{cor}(\widetilde{\mathbf{y}}, \mathbf{y})\right)^2.$$

Proposition 7 ($R^2(a^*, b^*)$)

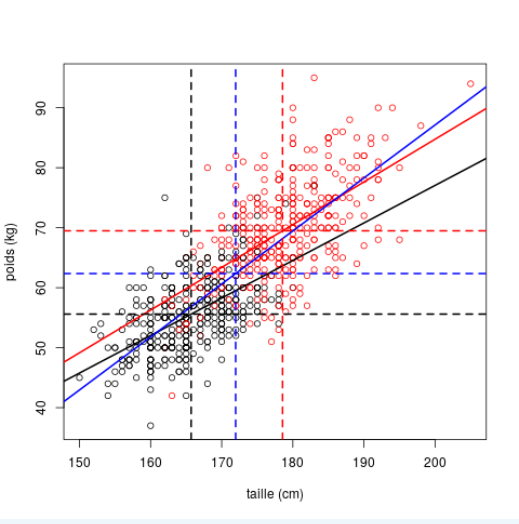
$$R^2(a^*, b^*) = \left(\text{cor}(\mathbf{x}, \mathbf{y})\right)^2.$$

Taille et poids.

$$\text{cor}(\mathbf{x}, \mathbf{y}) = 0.79, \qquad R^2 = 0.62$$

Mesure de l’ajustement

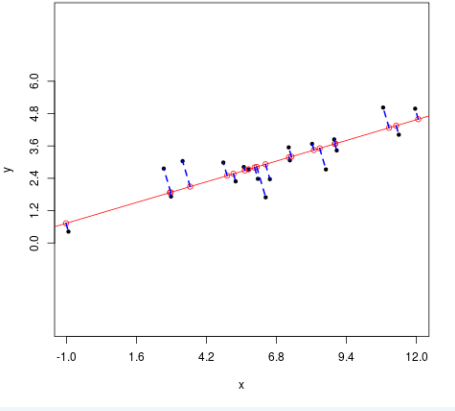
Taille et poids selon le sexe.



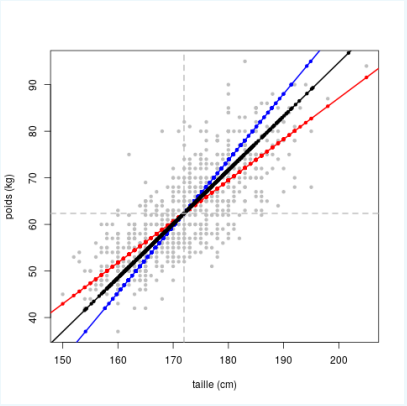
	n	a^*	b^*
F	375	-48.08	0.63
M	356	-57.67	0.71
F+M	731	-89.58	0.88

La droite de régression est définie de façon asymétrique : $y_i \simeq a + bx_i$ ou $x_i \simeq a' + b'y_i$.
L'analyse en composante principale détermine un droite optimale selon un critère symétrique.

Exemple fictif ($n = 20$)



Taille et poids



poids / taille
taille / poids
ACP

Remarques :

- Trouver les a^* et b^* optimaux nécessite de recourir soit à l'optimisation numérique (cours 5), soit à de l'algèbre linéaire hors programme.
- Réduction de dimension : problème général, surtout utile pour résumer $p \gg 2$ variables.

Définition 5 (Produit d'une matrice par un vecteur)

Le produit de la matrice $\mathbf{U}$ de dimension $m \times p$ par le vecteur $\mathbf{v}$ de dimension p :

$$\mathbf{U} = \begin{bmatrix} u_{11} & \dots & u_{1j} & \dots & u_{1p} \\ \vdots & & \vdots & & \vdots \\ u_{i1} & \dots & u_{ij} & \dots & u_{ip} \\ \vdots & & \vdots & & \vdots \\ u_{m1} & \dots & u_{mj} & \dots & u_{mp} \end{bmatrix}, \qquad \mathbf{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_j \\ \vdots \\ v_p \end{bmatrix}$$

est le vecteur $\mathbf{w}$ de dimension m :

$$\mathbf{w} = \mathbf{U}\mathbf{v} = \begin{bmatrix} w_1 \\ \vdots \\ v_i \\ \vdots \\ v_m \end{bmatrix} \text{ tel que } w_i = \sum_{j=1}^{p} u_{ij} v_j.$$

Proposition 8 (Écriture vectorielle de la régression simple)

Soit $\mathbf{X}$ *la matrice* $(n \times 2)$ *et* $\mathbf{b}$ *le vecteur (de dimension 2) définis par*

$$\mathbf{X} = [\mathbf{1}\ \mathbf{x}] = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_i \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} a \\ b \end{bmatrix},$$

on a

$$\widetilde{\mathbf{y}} = \mathbf{X}\mathbf{b} \qquad \text{et} \qquad C(a,b) = \|\mathbf{y} - \widetilde{\mathbf{y}}\|^2.$$

Écriture vectorielle de la régression simple : démonstration

Régression linéaire multiple : exemple

Fréquentation de salles de cinéma

source CNC 2003 / package R `ade4`

- $n = 94$ départements métropolitains (hors Corse)
- $y =$ `ticketRatio` = nombre moyen de billets par habitants
- $x_1 =$ `theatreRatio` = nombre de salles / nombre d'habitants
- $x_2 =$ `showRatio` = nombre de séances / nombre de salles
- $x_3 =$ `theatreSize` = taille moyenne des salles
- $x_4 =$ `artesRatio` = proportion de salles d'arts et d'essais

theatreRatio	showRatio	theatreSize	artesRatio	ticketRatio
67.96	0.74	179.66	0.34	1.49
70.9	0.74	194.82	0.21	1.36
81.16	0.61	141.29	0.14	1.45
164.29	0.74	151.3	0.3	3.24
289.26	0.6	172.94	0.14	4.31
92.98	1.18	178.34	0.09	3.48

On définit les vecteurs

$$\mathbf{y} = [y_1 \dots y_i \dots y_n]^\intercal,$$
$$\mathbf{x}_j = [x_{1j} \dots x_{ij} \dots x_{nj}]^\intercal \qquad \text{pour chaque } 1 \leq j \leq p, \mathbf{x}_0 = \mathbf{1} = [1 \dots 1 \dots 1]^\intercal.$$

(c'est à dire $x_{i0} = 1$, pour tout $1 \leq i \leq n$)
On définit la matrice $\mathbf{X}$ de dimensions $n \times (1 + p)$:

$$\mathbf{X} = [\mathbf{x}_0\ \mathbf{x}_1\ \dots\ \mathbf{x}_j\ \dots\ \mathbf{x}_p] = \begin{bmatrix} 1 & x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & x_{i1} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix}.$$

On veut cette fois

$$y_i \simeq b_0 + b_1 x_{i1} + \cdots + b_j x_{ij} + \cdots + b_p x_{ip} = \sum_{j=0}^{p} b_j x_{ij}$$

c'est-à-dire

$$\mathbf{Xb} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & x_{i1} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_j \\ \vdots \\ b_p \end{bmatrix} = \begin{bmatrix} \sum_{j=0}^{p} b_j x_{1j} \\ \vdots \\ \sum_{j=0}^{p} b_j x_{ij} \\ \vdots \\ \sum_{j=0}^{p} b_j x_{nj} \end{bmatrix} \simeq \mathbf{y},$$

en notant

$$\mathbf{b} = [b_0 \ b_1 \ \dots \ b_j \ \dots \ b_p]^\intercal.$$

On définit le vecteur $\mathbf{b}^*$ optimal à partir du critère des moindres carrés.

Définition 6

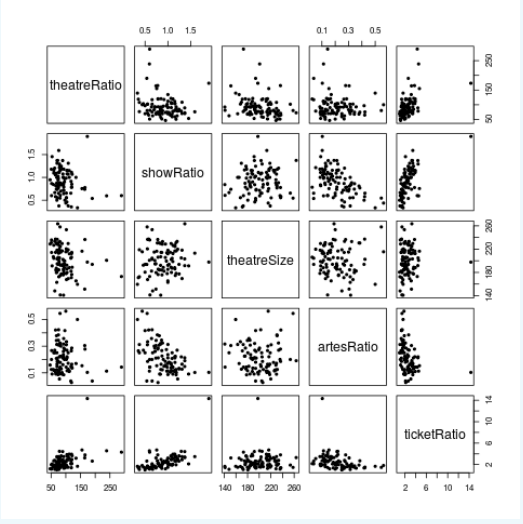
Critère des moindres carrés

$$C(b_0, b_1, \dots b_p) = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} b_j x_{ij} \right)^2 = \|\mathbf{y} - \mathbf{Xb}\|^2.$$

Solution :

- La technique utilisée pour trouver (a^*, b^*) ne fonctionne plus ici pour déterminer le vecteur $\mathbf{b}^*$ optimal.
- Comme pour l'ACP, $\mathbf{b}^*$ est obtenu soit par optimisation numérique (cours #5) soit par des techniques d'algèbre linéaire hors programme.

Fréquentation de salles de cinéma



(Intercept)	-5.45
theatreRatio	0.0268
showRatio	4.09
theatreSize	0.00664
artesRatio	2.02

La régression multiple n'est *pas équivalente* à un ensemble de p régressions simples.

Fréquentation de salles de cinéma

b_j^*	theatreRatio	showRatio	theatreSize	artesRatio
régressions simples	0.0189	3.12	0.00272	-4.65
régression multiple	0.0268	4.09	0.00664	2.02

Essentiel du cours 3

Définitions

- Covariance :

$$\mathrm{cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$$

- Coefficient de corrélation :

$$\mathrm{cor}(\mathbf{x}, \mathbf{y}) = \frac{\mathrm{cov}(\mathbf{x}, \mathbf{y})}{\sqrt{\mathrm{var}(\mathbf{x})\,\mathrm{var}(\mathbf{y})}}.$$

- Critère des moindres carrés :

$$C(a, b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

- Critère du R^2 : $\widetilde{y}_i = a + bx_i$

$$R^2(a, b) = \left(\mathrm{cor}(\widetilde{\mathbf{y}}, \mathbf{y})\right)^2.$$

Propositions

- Bilinéarité : $u_i = a + bx_i$ et $v_i = c + dy_i$ pour $1 \leq i \leq n$

$$\mathrm{cov}(\mathbf{u}, \mathbf{v}) = bd\,\mathrm{cov}(\mathbf{x}, \mathbf{y}).$$

- Inégalité de Cauchy-Schwartz :

$$\langle \mathbf{u}, \mathbf{v} \rangle^2 \leq \|\mathbf{u}\|^2 \|\mathbf{v}\|^2.$$

- Encadrement :

$$\left(\mathrm{cov}(\mathbf{x}, \mathbf{y})\right)^2 \leq \mathrm{var}(\mathbf{x})\,\mathrm{var}(\mathbf{y})$$

- Minimisation de $C(a, b)$:

$$b^* = \frac{\mathrm{cov}(\mathbf{x}, \mathbf{y})}{\mathrm{var}(\mathbf{x})}, \qquad a^* = \overline{y} - b^* \overline{x}.$$

- R^2 des moindres carrés :

$$R^2(a^*, b^*) = \left(\mathrm{cor}(\mathbf{x}, \mathbf{y})\right)^2.$$