Covariance, corrélation, régression linéaire

1 Définition et propriétés

1.1 Variance

La variance d'une variable aléatoire X est une **mesure de sa dispersion** autour de sa moyenne $\mathbb{E}[X]$.

Définition 1 (*Variance*)

La variance d'une *v.a.* X s'écrit :

$$\mathrm{Var}(X) = \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)^2\Big]$$

(sous réserve de convergence.)

Proposition 2 (*Propriétés*)

1. On a la **formule de Kœnig-Huygens** :
$$\mathrm{Var}(X) = \mathbb{E}\big[X^2\big] - \big(\mathbb{E}[X]\big)^2$$
2. Pour $a, b \in \mathbb{R}$ constantes déterministes, on a : $\mathrm{Var}(aX+b) = a^2 \cdot \mathrm{Var}(X)$.
3. $\big[\mathrm{Var}(X) = 0\big] \iff \big[X$ **déterministe**$\big]$.

(et alors : $X \equiv \mathbb{E}[X]$.)

1.2 Covariance

La covariance est une **mesure de la co-dispersion** d'un couple de *va* (X, Y).

Définition 3 (*Covariance*)

La covariance de deux *v.a.* X, Y s'écrit :

$$\mathrm{Cov}(X,Y) = \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big) \cdot \big(Y - \mathbb{E}[Y]\big)\Big]$$

(sous réserve de convergence.)

Proposition 4 (*Kœnig-Huygens*)

Soient X, Y deux variables aléatoires.
Alors :

$$\mathrm{Cov}(X,Y) = \mathbb{E}\big[X \cdot Y\big] - \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

(sous réserve de convergence)

Des variables indépendantes sont **décorrélées** :

Proposition 5 (*Cas de l'indépendance*)

Si X et Y sont **indépendantes**, alors :

$$\mathrm{Cov}(X,Y) = 0.$$

Proposition 6 (*Propriétés*)

La covariance vérifie :

- **Symétrie** $\mathrm{Cov}(Y,X) = \mathrm{Cov}(X,Y)$.
- **Variance** $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$

Distributivité dans la covariance

La covariance est **bilinéaire** :

$$\mathrm{Cov}(X,Y) = \mathbb{E}\Big[\underbrace{\big(\overbrace{X - \mathbb{E}[X]}^{\text{linéaire en } X}\big) \cdot \big(\overbrace{Y - \mathbb{E}[Y]}^{\text{linéaire en } Y}\big)}_{\text{bilinéaire en } (X,Y)}\Big]$$

Ainsi, l'expression $\mathrm{Cov}(\lambda_1 X_1 + \lambda_2 X_2, \mu_1 Y_1 + \mu_2 Y_2)$ se développe par **double distributivité**.
On a donc les « mêmes règles de calcul » pour $\mathrm{Cov}(X,Y)$ que pour xy où $x, y \in \mathbb{R}$:

- **Formules** Var, Cov
  - $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + 2\,\mathrm{Cov}(X,Y) + \mathrm{Var}(Y)$
  - $\mathrm{Var}(X-Y) = \mathrm{Var}(X) - 2\,\mathrm{Cov}(X,Y) + \mathrm{Var}(Y)$
  - $\mathrm{Cov}(X+Y, X-Y) = \mathrm{Var}(X) - \mathrm{Var}(Y)$
  - $\mathrm{Var}(Y - aX) = \mathrm{Var}(Y) - 2a\,\mathrm{Cov}(X,Y) + a^2\,\mathrm{Var}(X)$
- **Homologues numériques**
  - $(x+y)^2 = x^2 + 2xy + y^2$
  - $(x-y)^2 = x^2 - 2xy + y^2$
  - $(x+y)(x-y) = x^2 - y^2$
  - $(y-ax)^2 = y^2 - 2axy + a^2x^2$

- **Formules de polarisation**
  - $\mathrm{Cov}(X,Y) = \frac{1}{2} \cdot \big[\mathrm{Var}(X+Y) - \mathrm{Var}(X) - \mathrm{Var}(Y)\big]$ — $xy = \frac{1}{2} \cdot \big[(x+y)^2 - x^2 - y^2\big]$
  - $\mathrm{Cov}(X,Y) = \frac{1}{4} \cdot \big[\mathrm{Var}(X+Y) - \mathrm{Var}(X-Y)\big]$ — $xy = \frac{1}{4} \cdot \big[(x+y)^2 - (x-y)^2\big]$

2 L'encadrement de Cauchy-Schwarz

Proposition 7 (*Pour la covariance*)

Soient X, Y deux variables aléatoires.
On note σ_X, σ_Y leurs écarts-type.

($\sigma_X = \sqrt{\mathrm{Var}(X)}$, $\sigma_Y = \sqrt{\mathrm{Var}(Y)}$)

Alors :

$$-\sigma_X \cdot \sigma_Y \leqslant \mathrm{Cov}(X,Y) \leqslant \sigma_X \cdot \sigma_Y$$

Définition 8 (*Corrélation*)

On appelle **corrélation** de X et Y :

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

*(ou **coefficient de corrélation linéaire**)*

Par la Proposition 7, on a donc :

$$\rho(X,Y) \in [-1\,;1].$$

Démonstration de l'encadrement *(Prop. 7)* :

On considère la fonction $T : \begin{cases} \mathbb{R} \to \mathbb{R} \\ a \mapsto \mathrm{Var}(Y - aX). \end{cases}$

- **Signe de T** La variance d'une variable aléatoire $\mathrm{Var}(Z) = \mathbb{E}\Big[\underbrace{\big(Z - \mathbb{E}[Z]\big)^2}_{\geqslant 0}\Big] \geqslant 0$.

  Ainsi, on a $\forall a \in \mathbb{R}, \quad T(a) \geqslant 0$.

- **T est un trinôme**

  On développe par double distributivité : $T(a) = \mathrm{Var}(Y - aX)$ *(bilinéarité puis symétrie)*

  $$= \mathrm{Var}(Y) - 2a \cdot \mathrm{Cov}(X,Y) + a^2 \cdot \mathrm{Var}(X)$$

  On reconnaît un trinôme : $T(a) = r + qa + pa^2$.

- **Discriminant**

  Calculons le discriminant du trinôme : $T(a) = \mathrm{Var}(X) \cdot a^2 - 2 \cdot \mathrm{Cov}(X,Y) \cdot a + \mathrm{Var}(Y)$.

  On trouve : $\Delta = \big[-2\,\mathrm{Cov}(X,Y)\big]^2 - 4\,\mathrm{Var}(X) \cdot \mathrm{Var}(Y)$.

  $$= 4\big[\mathrm{Cov}(X,Y)^2 - \mathrm{Var}(X) \cdot \mathrm{Var}(Y)\big]$$

- **Conclusion** Le trinôme T ne change pas de signe. Son discriminant Δ est donc $\leqslant 0$.

  Il vient : $\mathrm{Cov}(X,Y)^2 - \mathrm{Var}(X) \cdot \mathrm{Var}(Y) \leqslant 0$, soit : $\mathrm{Cov}(X,Y)^2 \leqslant \mathrm{Var}(X) \cdot \mathrm{Var}(Y)$.

  L'inéquation $u^2 \leqslant v$

  > Pour $v \geqslant 0$, on résout l'inéquation : $\big[\, u^2 \leqslant v \,\big] \iff \big[\, -\sqrt{v} \leqslant u \leqslant \sqrt{v} \,\big]$

  On en déduit bien : $-\sqrt{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)} \leqslant \mathrm{Cov}(X,Y) \leqslant \sqrt{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)}$

  soit : $-\sigma_X \cdot \sigma_Y \leqslant \mathrm{Cov}(X,Y) \leqslant \sigma_X \cdot \sigma_Y$

■

3 Régression linéaire

Soient deux variables aléatoires X et Y.

On suppose que :
- la valeur de X est **connue**, *(X est la **variable explicative**)*
- la valeur de Y **inconnue**, et **à estimer**. *(Y est la **variable expliquée**)*

On recherche une estimation de Y en fonction de X, estimation **aussi bonne que possible**.

- **Forme de l'estimation : ajustement affine**

  Plus précisément, on cherche à estimer Y par une expression affine $\widehat{Y} = aX + b$. où a, b des constantes déterministes.

- **Résidu de l'estimation**

  C'est la variable aléatoire ϵ définie comme « l'erreur » : $\epsilon = Y - \widehat{Y}$. *(soit $Y = \widehat{Y} + \epsilon$)*

- **Optimisation au sens des moindres carrés**

  On cherche le couple (a,b) tel que $\widehat{Y} = aX + b$ minimise l'**erreur quadratique** *(moyenne)* :

  $$r(a,b) = \mathbb{E}[\epsilon^2] = \mathbb{E}\big[(Y - \widehat{Y})^2\big]$$

Proposition 9 (*Régression linéaire*)

1. L'optimisation des moindres carrés de l'estimation de Y par $\widehat{Y} = aX + b$ est réalisée par un unique couple (a,b) de réels déterministes.

2. Pour cette valeur optimale de (a,b) le résidu $\epsilon = Y - \widehat{Y}$ vérifie :

- le résidu ϵ est nul en moyenne *($\widehat{Y}$ **non-biaisé**)* : $\mathbb{E}[\epsilon] = 0$,
- le résidu ϵ et X sont décorrélés : $\mathrm{Cov}(X,\epsilon) = 0$.

De plus le coefficient directeur a est du signe de $\mathrm{Cov}(X,Y)$. *(sens de la tendance)*

3. Si $\rho(X,Y) = \pm 1$, alors, pour cette valeur de (a,b), on a $Y = \widehat{Y}$. *(avec probabilité 1)*

Remarques

1. Cette proposition précise le sens en lequel la covariance $\mathrm{Cov}(X,Y)$ est une mesure de codispersion.
Par exemple, elle donne le signe du coefficient directeur.

- si $\mathrm{Cov}(X,Y) > 0$, les variables X et Y varient « plutôt ensemble, » *(dans le même sens)*
- si $\mathrm{Cov}(X,Y) < 0$, les variables X et Y varient « plutôt en sens opposé »

2. L'ajustement affine $\widehat{Y} = aX + b$ et le résidu $\epsilon = Y - \widehat{Y}$ étant décorrélés, leurs variances s'additionnent :

$$\mathrm{Var}\big(\underbrace{\widehat{Y} + \epsilon}_{Y}\big) = \mathrm{Var}\big(\widehat{Y}\big) + \mathrm{Var}(\epsilon)$$

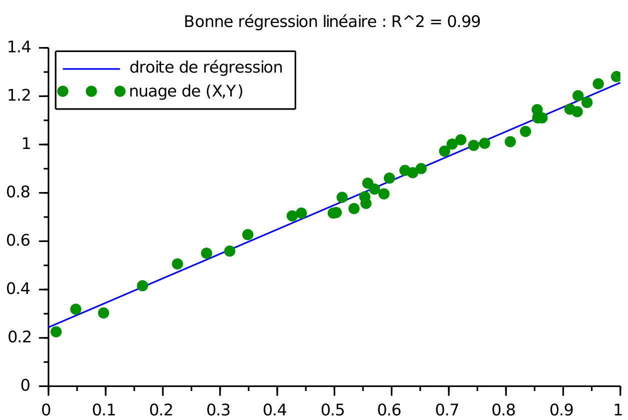
Ainsi on a : $\dfrac{\mathrm{Var}\big(\widehat{Y}\big)}{\mathrm{Var}(Y)} + \dfrac{\mathrm{Var}(\epsilon)}{\mathrm{Var}(Y)} = 1$, et on dit que :

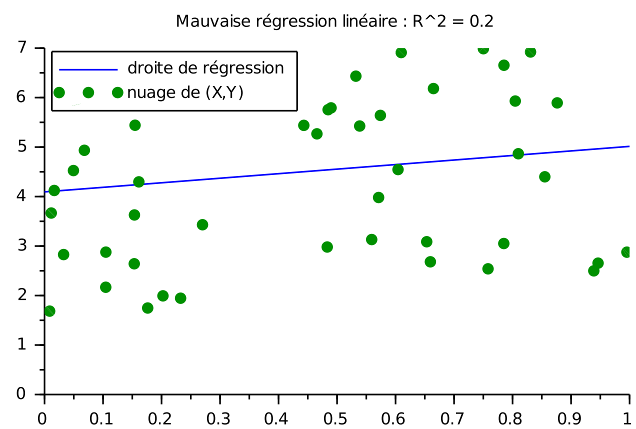
- $\dfrac{\mathrm{Var}\big(\widehat{Y}\big)}{\mathrm{Var}(Y)} = \rho(X,Y)^2$ est la part de variance expliquée, *(ou **coeff^{nt} de détermination**, noté R^2)*
- $\dfrac{\mathrm{Var}(\epsilon)}{\mathrm{Var}(Y)}$ est la part de variance inexpliquée

On a toujours $0 \leqslant R^2 \leqslant 1$, et R^2 indique la pertinence de la régression linéaire de Y par X.

- si le coeff^{nt} de détermination est **élevé** $R^2 \approx 1$, alors X explique bien Y.
- si le coeff^{nt} de détermination est **faible** $R^2 \approx 0$, l'explicabilité linéaire de Y par X est faible.

Interprétation graphique de la qualité de la régression linéaire :





Lemme 10 (*Décomposition biais-variance*)

1. L'erreur quadratique se décompose : $r(a,b) = \Big(\underbrace{\mathbb{E}[Y - aX - b]}_{\mathbb{E}[\epsilon]}\Big)^2 + \underbrace{\mathrm{Var}(Y - aX)}_{T(a)}$.
2. Le trinôme T est minimisé pour $a = \dfrac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$.
3. Le terme $\mathbb{E}[\epsilon]^2$ est minimisé pour $b = \mathbb{E}[Y] - a \cdot \mathbb{E}[X]$.

Démonstration du Lemme 10 :

1. La décomposition biais-variance est une réécriture de la formule de Kœnig-Huygens.
2. Le carré $\mathbb{E}[\epsilon]^2 = \Big(\mathbb{E}[Y - aX - b]\Big)^2$ est minimisé quand $\mathbb{E}[Y - aX - b] = 0$, soit, en développant, pour $b = \mathbb{E}[Y] - a \cdot \mathbb{E}[X]$.
3. On a étudié le trinôme $T(a) = \mathrm{Var}(Y - aX) = \mathrm{Var}(Y) - 2a \cdot \mathrm{Cov}(X,Y) + a^2 \cdot \mathrm{Var}(X)$.
   Il est minimisé quand $T'(a) = -2\,\mathrm{Cov}(X,Y) + 2\,\mathrm{Var}(X) \cdot a = 0$, soit $a = \dfrac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$. ■

Démonstration de la Proposition 9 :

1. On vérifie que minimiser $r(a,b)$, c'est minimiser les deux termes de la décomposition :
$$r(a,b) = \Big(\mathbb{E}[Y - aX - b]\Big)^2 + \mathrm{Var}(Y - aX)$$
et que la minimisation de $r(a,b)$ se réalise donc uniquement pour $\begin{cases} a = \dfrac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \\ b = \mathbb{E}[Y] - a \cdot \mathbb{E}[X] \end{cases}$
On trouve donc l'ajustement affine optimal :
$$\widehat{Y} = \tfrac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \cdot \Big(X - \mathbb{E}[X]\Big) + \mathbb{E}[Y].$$
2.
   - On vérifie que $\mathbb{E}\big[\widehat{Y}\big] = \mathbb{E}[Y]$ et donc que $\mathbb{E}[\epsilon] = 0$.
   - On a : $T'(a) = -2\,\mathrm{Cov}(X,Y) + 2\,\mathrm{Var}(X) \cdot a = -2\,\mathrm{Cov}(X, Y - aX)$.
     Ainsi pour $T'(a) = 0$, il vient : $\mathrm{Cov}(X, \underbrace{Y - aX}_{\epsilon + b}) = 0$, d'où $\mathrm{Cov}(X,\epsilon) = 0$.
3. Si $\rho(X,Y) = \pm 1$, alors, d'après la Remarque **2.**, on trouve $\mathrm{Var}(\epsilon) = 0$.
   Par la Proposition 2-**3.**, la variable ϵ est donc déterministe et $\epsilon \equiv \mathbb{E}[\epsilon] = 0$.
   Ainsi, on a bien $Y - \widehat{Y} \equiv 0$, soit : $Y \equiv \widehat{Y} = aX + b$. ■