Chapitre 3. Corrélation et ajustement affine

R2.13 – Techniques quantitatives et représentations 2

D. Fourer

Département TC-Juvisy

Semestre 2 – 2023/2024

Contexte et notations

VARIABLES STATISTIQUES

- Variables X, Y continues non regroupées en classes sur une même population (taille N).

▶ *Définitions brutes des quantités à partir des individus.*

Contexte et notations

Variables statistiques

- Variables X, Y continues non regroupées en classes sur une même population (taille N).
- *Définitions brutes des quantités à partir des individus.*

Paramètres de position et de dispersion

- Moyennes marginales : $\overline{x}$ et $\overline{y}$.
- Variances et écarts-types marginaux : $\text{var}\,X, \text{var}\,Y$ et σ_X, σ_Y.

Covariance – Construction

Idées directrices

- Obtenir un indicateur numérique (plutôt que le « oui/non » de l'indépendance statistique).
- Obtenir une notion plus facilement réalisée que l'indépendance statistique.

Covariance – Construction

Idées directrices

- Obtenir un indicateur numérique (plutôt que le « oui/non » de l'indépendance statistique).
- Obtenir une notion plus facilement réalisée que l'indépendance statistique.

Définition

Covariance $\text{cov}(X, Y)$: Moyenne statistique des produits écarts des valeurs à leur moyenne marginale respective.

Covariance – Construction

Idées directrices

- Obtenir un indicateur numérique (plutôt que le « oui/non » de l'indépendance statistique).
- Obtenir une notion plus facilement réalisée que l'indépendance statistique.

Définition

Covariance $\mathrm{cov}(X, Y)$: Moyenne statistique des produits écarts des valeurs à leur moyenne marginale respective.

Formules

$$\frac{1}{N}\sum_{i=1}^{N}\big(X(i)-\overline{x}\big)\big(Y(i)-\overline{y}\big), \quad \sum_{i,j} f_{ij}(x_i-\overline{x})(y_j-\overline{y}), \quad \ldots$$

Covariance – Propriétés

Lien avec la variance

Extension et généralisation de la notion de variance :

$$\text{cov}(X, X) = \text{var}\, X.$$

Covariance – Propriétés

Lien avec la variance

Extension et généralisation de la notion de variance :

$$\operatorname{cov}(X, X) = \operatorname{var} X.$$

Calcul effectif

- Diférence entre la moyenne statistique des produits des valeurs et le produit des moyennes marginales.
- Formules :

$$\left(\frac{1}{N} \sum_{i=1}^{N} X(i) Y(i) \right) - \overline{xy}, \quad \left(\sum_{i,j} f_{ij} x_i y_j \right) - \overline{xy}, \quad \ldots$$

Corrélation et lien avec l'indépendance statistique

Définition

On dit que X et Y sont corrélées lorsque $\text{cov}(X, Y) \neq 0$ et non corrélées lorsque $\text{cov}(X, Y) = 0$.

Corrélation et lien avec l'indépendance statistique

Définition

On dit que X et Y sont corrélées lorsque $\text{cov}(X, Y) \neq 0$ et non corrélées lorsque $\text{cov}(X, Y) = 0$.

Proposition

Si X et Y sont indépendantes, alors elles sont non corrélées.
La réciproque est fausse.

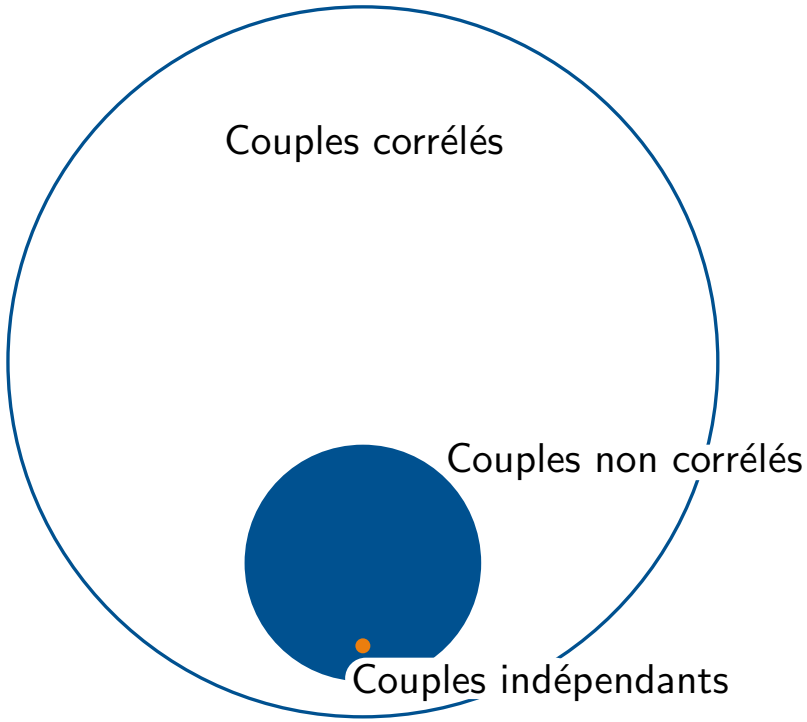
Corrélation et lien avec l'indépendance statistique

Définition

On dit que X et Y sont corrélées lorsque $\text{cov}(X, Y) \neq 0$ et non corrélées lorsque $\text{cov}(X, Y) = 0$.

Proposition

Si X et Y sont indépendantes, alors elles sont non corrélées.
La réciproque est fausse.



ATTENTION !

Les couples non corrélés sont « rares ». (Cf. site Spurious correlations.)

Coefficient de corrélation

Définition

Coefficient de corrélation (ou de régression) de X et Y :

$$r(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}.$$

Coefficient de corrélation

Définition

Coefficient de corrélation (ou de régression) de X et Y :

$$r(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}.$$

- Il discrimine la non corrélation : $r(X,Y) \neq 0 \iff \text{cov}(X,Y) \neq 0$.
- Il est borné : $-1 \leqslant r(X,Y) \leqslant 1$. (Inégalité de Cauchy-Schwarz.)

Coefficient de corrélation

Définition

Coefficient de corrélation (ou de régression) de X et Y :

$$r(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}.$$

- Il discrimine la non corrélation : $r(X,Y) \neq 0 \iff \mathrm{cov}(X,Y) \neq 0$.
- Il est borné : $-1 \leqslant r(X,Y) \leqslant 1$. (Inégalité de Cauchy-Schwarz.)

VALEURS REMARQUABLES

- $r(X,Y) = 1$: points du nuage alignés sur une droite croissante.
- $r(X,Y) = 0$: X et Y non corrélées, nuage « loin » de toute droite.
- $r(X,Y) = -1$: points du nuage alignés sur une droite décroissante.

Qu'est-ce que l'ajustement affine ?

Définition

Ajustement affine : Méthode de construction d'une droite modélisant (au mieux) le nuage de points.

Qu'est-ce que l'ajustement affine ?

Définition

Ajustement affine : Méthode de construction d'une droite modélisant (au mieux) le nuage de points.

Points essentiels

- Comment évaluer la fiabilité de la modélisation ?
- Existe-t-il une droite meilleure que les autres ?
  - *Le cas échéant, en quel sens est-elle meilleure ?*

Qu'est-ce que l'ajustement affine ?

Définition

Ajustement affine : Méthode de construction d'une droite modélisant (au mieux) le nuage de points.

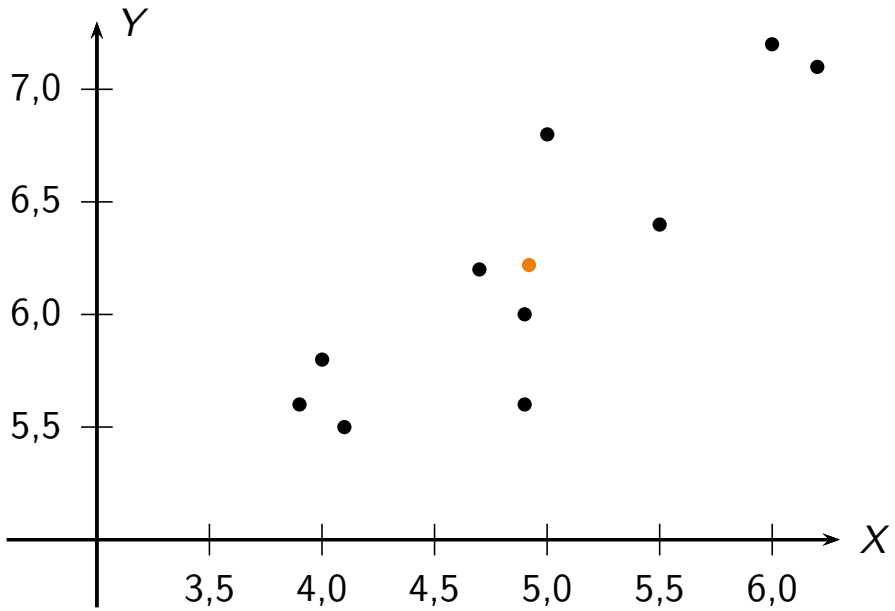
Points essentiels

- Comment évaluer la fiabilité de la modélisation ?
- Existe-t-il une droite meilleure que les autres ?
  - *Le cas échéant, en quel sens est-elle meilleure ?*

Définition (utile)

Point moyen G : Point de coordonnées $(\overline{x}, \overline{y})$ dans le nuage.

Exemple fondamental (TD 3 – Exercice 1)



COORDONNÉES DU POINT MOYEN G

$$\overline{x} = 4,92 \quad \text{et} \quad \overline{y} = 6,22.$$

Construction

Principe

- Variable statistique des écarts verticaux : $E_{a,b} = Y - (aX + b)$.
  - *Choisir a et b pour que $E_{a,b}$ soit proche de la variable nulle.*
- Minimisation de la variance de $E_{a,b}$ (moindres carrés) :

$$\operatorname{var} E_{a,b} = a^2 \operatorname{var} X - 2a \operatorname{cov}(X, Y) + \operatorname{var} Y.$$

- Annulation de la moyenne de $E_{a,b}$: $\quad \overline{e}_{a,b} = \overline{y} - (a\overline{x} + b)$.

Construction

Principe

- Variable statistique des écarts verticaux : $E_{a,b} = Y - (aX + b)$.
  - *Choisir a et b pour que $E_{a,b}$ soit proche de la variable nulle.*
- Minimisation de la variance de $E_{a,b}$ (moindres carrés) :

$$\operatorname{var} E_{a,b} = a^2 \operatorname{var} X - 2a\operatorname{cov}(X, Y) + \operatorname{var} Y.$$

- Annulation de la moyenne de $E_{a,b}$: $\quad \overline{e}_{a,b} = \overline{y} - (a\overline{x} + b)$.

Définition

Droite de régression de Y en X : équation réduite $y = \widetilde{a}x + \widetilde{b}$ avec :

$$\widetilde{a} = \frac{\operatorname{cov}(X, Y)}{\operatorname{var} X} = \frac{\operatorname{cov}(X, Y)}{\sigma_X^2} = r(X, Y)\frac{\sigma_Y}{\sigma_X} \quad \text{et} \quad \widetilde{b} = \overline{y} - \widetilde{a}\overline{x}.$$

Exemple (TD 3 – Exercice 1)

Paramètres statistiques

- Point moyen $G(4,92; 6,22)$.
- Variances et écart-types :

$$\operatorname{var} X = 0,5756 \quad \text{et} \quad \sigma_X = 0,7587 \text{ à } 10^{-4} \text{ près},$$
$$\operatorname{var} Y = 0,3616 \quad \text{et} \quad \sigma_Y = 0,6013 \text{ à } 10^{-4} \text{ près}.$$

- Covariance : $\operatorname{cov}(X, Y) = 0,3966$.

Exemple (TD 3 – Exercice 1)

Paramètres statistiques

- Point moyen $G(4,92; 6,22)$.
- Variances et écart-types :

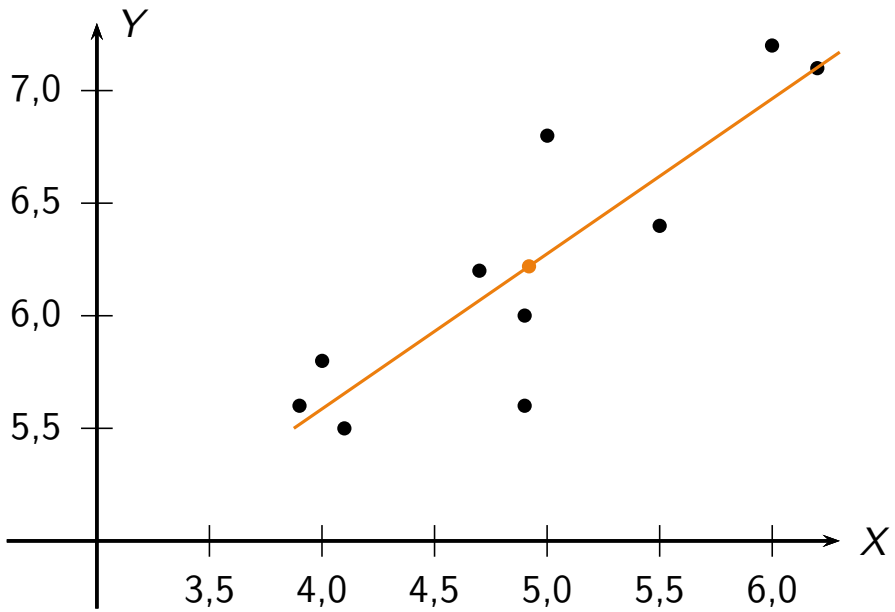
$$\operatorname{var} X = 0,5756 \quad \text{et} \quad \sigma_X = 0,7587 \text{ à } 10^{-4} \text{ près,}$$
$$\operatorname{var} Y = 0,3616 \quad \text{et} \quad \sigma_Y = 0,6013 \text{ à } 10^{-4} \text{ près.}$$

- Covariance : $\operatorname{cov}(X, Y) = 0,3966$.

Droite de régression

- Paramètres : $\widetilde{a} = 0,6890$ et $\widetilde{b} = 2,8301$ à 10^{-4} près.
- Coefficient de corrélation : $r(X, Y) = 0,8693$ à 10^{-4} près.

Exemple (TD 3 – Exercice 1)



Avantages

- Meilleure droite possible (au sens des écarts verticaux).
- Les projections sont fiables lorsque :

$$|r(X,Y)| \geqslant 0,9 \iff \Big(r(X,Y) \geqslant 0,9 \quad \text{ou} \quad r(X,Y) \leqslant -0,9\Big).$$

FIN DU CHAPITRE 3