

Statistique descriptive

Chapitre 14 : Corrélation et droite d'ajustement



Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 27 août 2026

Normaliser la covariance

Ce que r ne dit pas

La droite d'ajustement

La qualité de l'ajustement

Bilan du cours

Ce qu'il faut retenir

Normaliser la covariance

Rappel

La covariance a le bon **signe** mais pas de **valeur** interprétable : elle dépend des unités et n'a pas de bornes.

Cauchy-Schwarz nous a donné la solution : $|\text{cov}(x, y)| \leq \sigma_x \sigma_y$. Il suffit de diviser par le majorant.

Coefficient de corrélation linéaire de Bravais-Pearson

$$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

r est **sans unité** et vérifie toujours $-1 \leq r \leq +1$.

Sur nos données

$$r = \frac{9\,860}{5,964 \times 2\,598,10} = \frac{9\,860}{15\,495} = 0,636$$

L'échelle

$ r $	Intensité de la liaison linéaire
0 à 0,2	négligeable
0,2 à 0,4	faible
0,4 à 0,6	modérée
0,6 à 0,8	forte
0,8 à 1	très forte

Les cas extrêmes

- $r = +1$: tous les points sont **exactement alignés** sur une droite croissante.
- $r = -1$: alignés sur une droite décroissante.
- $r = 0$: aucune liaison **linéaire**.

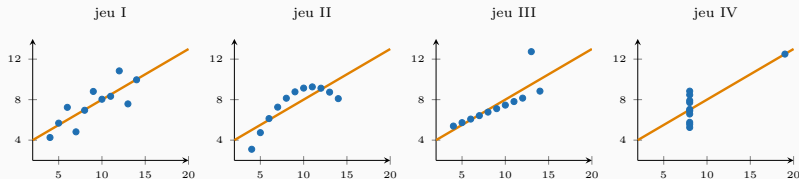
Ces seuils sont conventionnels

Ils dépendent du domaine. En physique, $r = 0,9$ est décevant. En sciences sociales, $r = 0,5$ est un beau résultat.

Notre $r = 0,636$ est une liaison **forte** pour des données individuelles de salaires : l'ancienneté compte réellement.

Ce que r ne dit pas

Le quartet d'Anscombe



Le quartet d'Anscombe (1973) : quatre jeux de onze couples ayant *exactement* les mêmes moyennes, les mêmes variances, le même $r = 0,82$ et la même droite d'ajustement $\hat{y} = 3,00 + 0,500x$. Et pourtant quatre situations sans rapport.

Les indicateurs ne dispensent jamais de regarder les données.

Ce que montre le quartet

Quatre jeux de onze couples ayant **exactement** les mêmes moyennes, les mêmes variances, le même $r = 0,82$ et la même droite d'ajustement $\hat{y} = 3,00 + 0,500x$.

Et quatre situations sans aucun rapport :

1. une vraie liaison linéaire bruitée;
2. une liaison **parabolique** parfaite — la droite n'a aucun sens;
3. une liaison linéaire parfaite **gâchée par un seul point**;
4. aucune liaison, un seul point isolé créant à lui seul toute la corrélation.

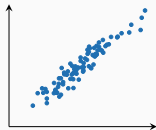
La leçon

Les indicateurs ne dispensent jamais de regarder les données.

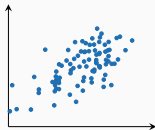
C'est le même avertissement que pour les moments au chapitre 9 et pour le Gini au chapitre 10. Ici il est démontré par construction, ce qui le rend irréfutable.

Les autres limites

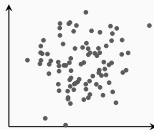
$r = 0,95$
liaison forte, positive



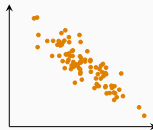
$r = 0,64$
liaison modérée



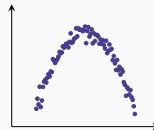
$r = 0,09$
aucune liaison



$r = -0,86$
liaison forte, négative



$r = -0,02$
et pourtant...



r ne mesure que la liaison **linéaire** : le dernier nuage est parfaitement structuré, et pourtant $r \simeq 0$.
Un coefficient de corrélation ne se lit jamais sans le nuage de points.

Trois pièges

Piège 1 : r ne mesure que le linéaire

Un nuage en cloche, en U, en escalier peut être parfaitement structuré et donner $r \simeq 0$.

$r = 0$ signifie « pas de liaison linéaire », jamais « pas de liaison ».

Piège 2 : corrélation n'est pas causalité

Une corrélation entre x et y admet **quatre** explications, et la statistique descriptive ne permet pas de trancher :

1. x cause y ;
2. y cause x ;
3. une **troisième variable** cause les deux;
4. c'est une **coïncidence**.

Le cas 3 est le plus fréquent et le plus trompeur. Les ventes de crème solaire et les noyades sont fortement corrélées : la chaleur cause les deux.

Aucun calcul ne peut établir une causalité. Seuls un raisonnement économique et, si possible, une expérience le peuvent.

Piège 3 : le paradoxe de Simpson

Une corrélation peut **changer de signe** quand on conditionne sur une troisième variable.

Il peut exister une corrélation positive entre ancienneté et salaire dans l'ensemble, et négative à l'intérieur de chaque service — si les services les plus anciens sont aussi les moins bien payés.

Le remède : toujours regarder le nuage par sous-groupes avant de conclure sur l'ensemble.

La droite d'ajustement

Le principe des moindres carrés

Le problème

Parmi toutes les droites $y = a + b x$, on cherche celle qui rend minimale la **somme des carrés des résidus** :

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - a - b x_i)^2$$

où e_i est l'écart **vertical** entre le point observé et la droite.

Pourquoi les carrés, et pourquoi la verticale

Les **carrés** parce qu'ils annulent les signes tout en restant dérivables — exactement l'argument du chapitre 8 pour la variance.

La **verticale** parce qu'on cherche à prédire y à partir de x . L'erreur qui compte est celle qu'on commet sur y , pas la distance géométrique au nuage.

Les deux coefficients

$$\hat{b} = \frac{\text{cov}(x, y)}{V(x)}$$

$$\hat{a} = \bar{y} - \hat{b} \bar{x}$$

Deux lectures immédiates de ces formules

La seconde dit que la droite passe **par le point moyen** $(\bar{x}; \bar{y})$: c'est le contrôle annoncé au chapitre 11.

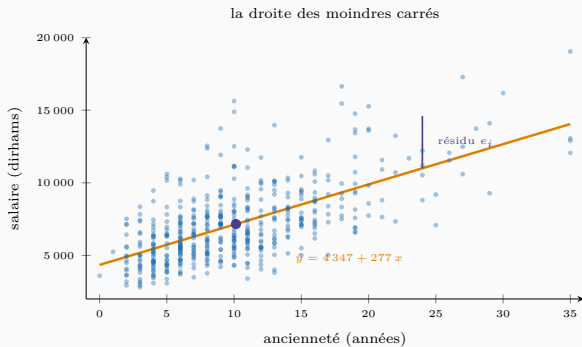
La première dit que la pente a **le signe de la covariance**, puisque $V(x) > 0$. Corrélation positive et pente positive vont donc toujours ensemble.

Sur nos données

$$\hat{b} = \frac{9\,860}{35,574} = 277,2 \qquad \hat{a} = 7\,158,26 - 277,17 \times 10,142 = 4\,347$$

$$\widehat{\text{salaire}} = 4\,347 + 277 \times \text{ancienneté}$$

La solution (fin)



la droite qui rend la **somme des carrés des résidus** minimale.

$$\hat{b} = \frac{\text{cov}(x, y)}{V(x)} = 277$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 4\,347$$

Une année d'ancienneté de plus, c'est 277 dirhams de salaire en plus *en moyenne*.

$r^2 = 0,405$: l'ancienneté explique 40 % de la variance des salaires.

Les deux coefficients, en français

$\hat{b} = 277$: une année d'ancienneté de plus correspond, **en moyenne**, à 277 dirhams de salaire en plus.

$\hat{a} = 4\,347$: c'est le salaire prédit pour une ancienneté nulle — un salarié qui vient d'être embauché.

Ici la constante a un sens économique, parce que $x = 0$ appartient au domaine observé. **Ce n'est pas toujours le cas** : si l'ancienneté allait de 10 à 35 ans, $\hat{a}$ serait une extrapolation sans signification.

Les mots à employer, et ceux à éviter

Dire : « correspond en moyenne à », « est associé à ».

Ne pas dire : « fait augmenter », « provoque », « entraîne ». Ce sont des affirmations causales, que ce calcul ne permet pas.

La prudence de vocabulaire n'est pas de la coquetterie : c'est la seule chose qui sépare une analyse d'une opinion.

La qualité de l'ajustement

Le coefficient de détermination

$$r^2 = \frac{\text{variance expliquée}}{\text{variance totale}}$$

C'est le **carré** du coefficient de corrélation, et il représente la part de la variance de y expliquée par x .

Sur nos données

$$r^2 = 0,636^2 = 0,405$$

L'ancienneté explique 40,5 % de la variance des salaires. Les 59,5 % restants tiennent au diplôme, au poste, à la région, à la performance — et à l'arbitraire.

Un rapprochement avec le chapitre 12

r^2 et η^2 répondent à la même question — quelle part de la variance est expliquée — mais ne s'appliquent pas aux mêmes données :

	r^2	η^2
Variable explicative	quantitative	qualitative
Forme supposée	linéaire	aucune
Sur nos données	40,5 % (ancienneté)	25,8 % (diplôme)

L'ancienneté explique donc davantage que le diplôme. Une régression **multiple**, qui les combinerait, expliquerait davantage encore — c'est l'objet du cours d'économétrie.

Un point souvent ignoré

Il existe **deux** droites, selon la variable qu'on cherche à prédire :

$$y \text{ sur } x : \hat{b}_{y/x} = \frac{\text{cov}(x, y)}{V(x)} \qquad x \text{ sur } y : \hat{b}_{x/y} = \frac{\text{cov}(x, y)}{V(y)}$$

Leur produit

$$\hat{b}_{y/x} \times \hat{b}_{x/y} = \frac{\text{cov}^2}{V(x)V(y)} = r^2$$

Vérification

$$277,17 \times 0,001461 = 0,405 = r^2 \quad \checkmark$$

Ce que cela signifie géométriquement

Les deux droites se coupent au point moyen. Elles se confondent si $r^2 = 1$, et deviennent perpendiculaires si $r = 0$.

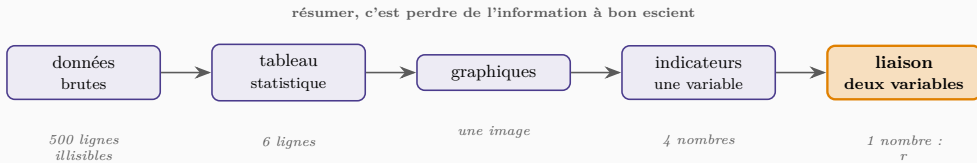
L'angle entre les deux droites mesure donc la qualité de l'ajustement. C'est une manière visuelle de comprendre r^2 : plus l'ajustement est bon, plus les deux droites se rapprochent l'une de l'autre.

Bilan du cours

De 500 lignes à un nombre

Étape	Résultat
Données brutes	500×6 valeurs
Tableau statistique	6 lignes
Graphique	une image
Indicateurs univariés	12 nombres
Liaison bivariée	$r = 0,636$
Ajustement	une droite, 2 coefficients

Le chemin parcouru (suite)



La fiche complète de nos 500 salariés

Un salaire moyen de 7 158 dirhams, médian de 6 770, dispersé d'un écart-type de 2 598 — soit un coefficient de variation de 0,36.

Une distribution **étalée à droite** ($\gamma_1 = 1,16$) et à **queues épaisses** ($\gamma_2 = 4,84$), modérément concentrée ($G = 0,195$).

Le **diplôme** explique 25,8 % de la dispersion, l'**ancienneté** 40,5 %, à raison de 277 dirhams par année.

Six lignes pour cinq cents salariés, et rien d'essentiel n'a été perdu.

Ce que le cours n'a pas permis de dire

Trois choses, et il faut savoir les nommer :

1. Si ces résultats **valent au-delà** de cette entreprise — c'est la statistique **inférentielle**.
2. Si l'ancienneté **cause** le salaire — c'est un problème d'identification, qui exige des hypothèses économiques.
3. Ce que donnerait la prise en compte **simultanée** de plusieurs variables — c'est la régression **multiple**.

Ces trois questions sont l'objet des trois cours suivants du parcours. Ce cours-ci vous a donné les objets sur lesquels ils travaillent.

Ce qu'il faut retenir

Sept points

1. $r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$: sans unité, compris entre -1 et $+1$.
2. r ne mesure que la liaison **linéaire** — le quartet d'Anscombe le démontre.
3. **Corrélation n'est pas causalité** : quatre explications possibles, dont la troisième variable.
4. Moindres carrés : $\hat{b} = \frac{\text{cov}(x, y)}{V(x)}$ et $\hat{a} = \bar{y} - \hat{b} \bar{x}$.
5. La droite passe **toujours** par le point moyen.
6. r^2 est la **part de variance expliquée** : ici 40,5 %.
7. Deux droites de régression, dont le produit des pentes vaut r^2 .

La suite

Tous ces calculs ont été faits ici à la main, sur un petit nombre de valeurs, pour qu'on en comprenne le mécanisme.

Les deux derniers chapitres refont **toute la chaîne** sur les 500 salariés — en R, puis en Python.