

Statistique descriptive

Chapitre 13 : Corrélation et droite d'ajustement

Jaouad Madkour

12 octobre 2026

Plan du chapitre

Normaliser la covariance	2
Le problème posé au chapitre 12	2
Le coefficient sur deux exemples	2
Comment le lire	3
Des seuils conventionnels	3
Ce que r ne dit pas	3
Le quartet d'Anscombe	3
L'avertissement	4
Les autres limites	4
Linéarité et causalité	4
Le paradoxe de Simpson	5
La droite d'ajustement	5
Le principe des moindres carrés	5
La solution	5
La droite des dix salariés, à la main	6
La droite des 500 salariés	6
La droite sur le nuage	6
L'interprétation	7
La qualité de l'ajustement	7
Les résidus	7
Regarder les résidus	7
Le coefficient de détermination	8
r^2 et s^2	8
Les deux droites de régression	8
Ce que l'angle entre les droites signifie	9
Bilan du cours	9
Le chemin parcouru	9
Le fil conducteur, refermé	9
Ce que vous savez faire	10

Ce que le cours n'a pas permis de dire	10
S'entraîner	10
Exercice 1 : une droite à partir de cinq sommes	10
Exercice 1 : solution	10
Exercice 2 : lire les résidus	11
Exercice 2 : solution	11
Ce qu'il faut retenir	11
Le résumé du chapitre	11

Normaliser la covariance

Le problème posé au chapitre 12

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 $-1 \leq r \leq +1$.

Le coefficient sur deux exemples

Les dix salariés du chapitre 12

$\text{cov} = 9,7$, $\sigma_x = 5,48$ ans, $\sigma_y = 1,91$ millier de dirhams :

$$r = \frac{9,7}{5,48 \times 1,91} = \frac{9,7}{10,46} = 0,927$$

Nos 500 salariés

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

Même sens, intensité bien moindre : dans la grande entreprise, l'ancienneté laisse beaucoup plus de place à d'autres facteurs.

Comment le lire

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**.

Des seuils conventionnels

Ils dépendent du domaine

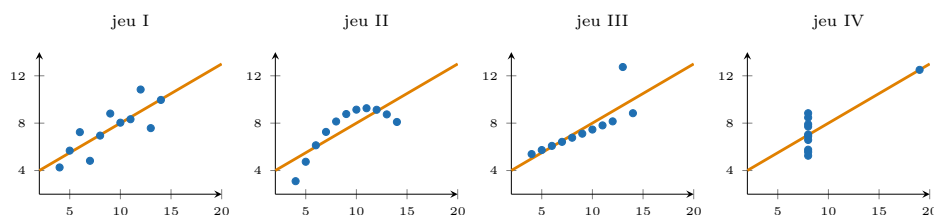
En physique, $r = 0,9$ est décevant. En sciences sociales, $r = 0,5$ est un beau résultat. Notre $r = 0,636$ est une liaison **forte** pour des données individuelles de salaires : l'ancienneté compte réellement.

r ne dépend pas des unités

Salaire en dirhams ou en milliers, ancienneté en années ou en mois : la covariance change, les écarts-types changent dans la même proportion, et r ne bouge pas. C'est exactement ce qu'on cherchait.

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.

L'avertissement

Ce que montre le quartet

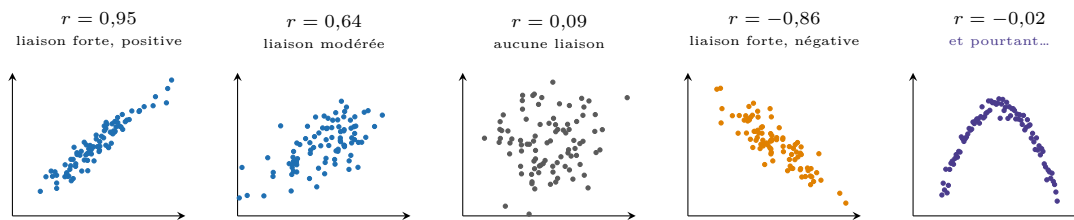
Quatre jeux de onze couples ayant **exactement** les mêmes moyennes, les mêmes variances, le même $r = 0,82$ et la même droite d'ajustement $\hat{y} = 3,00 + 0,500x$. Et quatre situations sans aucun rapport :

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 8 et pour le Gini au chapitre 9 ; ici il est démontré par construction.

Les autres limites



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.

Linéarité et causalité

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 : x cause y ; y cause x ; une **troisième variable** cause les deux ; une **coïncidence**.

Le troisième cas est le plus fréquent et le plus trompeur : les ventes de crème solaire et les noyades sont fortement corrélées, et la chaleur cause les deux. **Aucun calcul ne peut établir une causalité.**

Le paradoxe de Simpson

Piège 3 : une corrélation qui change de signe

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 : c'est le conditionnement du chapitre 11, appliqué à la corrélation.

La droite d'ajustement

Le principe des moindres carrés

Le problème

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

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

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 — l'argument du chapitre 7. La **verticale** parce qu'on cherche à prédire y à partir de x : l'erreur qui compte est celle qu'on commet sur y .

La solution

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 10.

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.

La droite des dix salariés, à la main

Les coefficients

$V(x) = 30,05$, $\text{cov} = 9,7$, $\bar{x} = 8,5$, $\bar{y} = 7,1$:

$$\hat{b} = \frac{9,7}{30,05} = 0,323 \quad \hat{a} = 7,1 - 0,323 \times 8,5 = 4,356$$

$$\hat{y} = 4,356 + 0,323 x \quad (\text{milliers de dirhams})$$

Valeurs ajustées et résidus de quelques salariés

Ancienneté x_i	2	6	8	12	20
Salaire y_i	4,50	7,50	6,00	7,50	11,00
Ajusté $\hat{y}_i$	5,00	6,29	6,94	8,23	10,81
Résidu e_i	-0,50	+1,21	-0,94	-0,73	+0,19

La droite des 500 salariés

Sur nos données

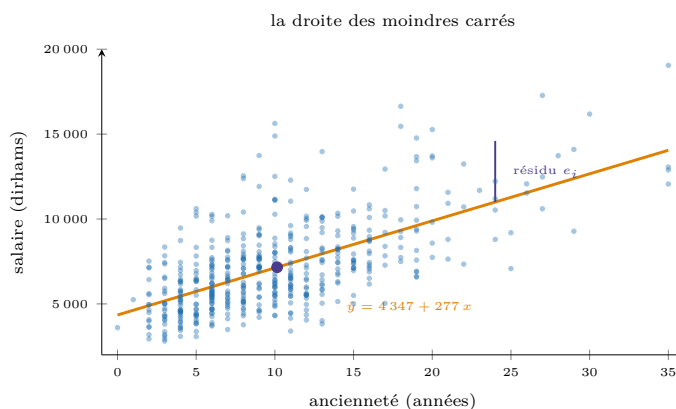
$$\hat{b} = \frac{9860}{35,574} = 277,2 \quad \hat{a} = 7158,26 - 277,17 \times 10,142 = 4347$$

$$\widehat{\text{salaire}} = 4347 + 277 \times \text{ancienneté}$$

Deux prévisions

À 5 ans : $4347 + 277 \times 5 = 5733$ dirhams. À 20 ans : $4347 + 277 \times 20 = 9891$ dirhams.
Ce sont des salaires **moyens** prévus : à ancienneté donnée, les salaires réels s'en écartent en moyenne de 2000 dirhams.

La droite sur le nuage



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

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

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 4347$$

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.

L'interprétation

Les deux coefficients, en français

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

$\hat{a} = 4347$: le salaire prédit pour une ancienneté nulle — un salarié qui vient d'être embauché. La constante a ici un sens, parce que $x = 0$ appartient au domaine observé ; 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 est la seule chose qui sépare une analyse d'une opinion.

La qualité de l'ajustement

Les résidus

Résidu et décomposition

Le résidu $e_i = y_i - \hat{y}_i$ est la part du salaire que la droite n'explique pas. Les moindres carrés garantissent $\sum e_i = 0$, et

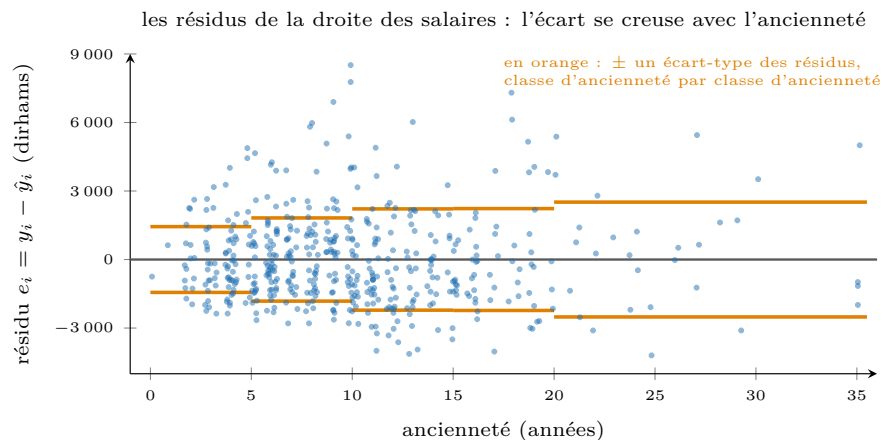
$$V(y) = V(\hat{y}) + V(e)$$

La variance totale se partage entre la variance **expliquée** par la droite et la variance **résiduelle** — comme au chapitre 11.

Les dix salariés

$V(y) = 3,640$; $V(\hat{y}) = \hat{b}^2 V(x) = 0,323^2 \times 30,05 = 3,131$; $V(e) = 0,509$. La droite explique $3,131/3,640 = 86\%$ de la variance.

Regarder les résidus



Une dispersion qui grandit

L'écart-type des résidus passe de 1 440 dirhams avant 5 ans à 2 514 au-delà de 20 ans : **les carrières se différencient avec le temps**. La droite n'en dit rien ; l'économétrie appelle cela l'hétéroscédasticité.

Le coefficient de détermination

Le coefficient de détermination

$$r^2 = \frac{\text{variance expliquée}}{\text{variance totale}} = \frac{V(\hat{y})}{V(y)}$$

C'est le **carré** du coefficient de corrélation : 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.

r^2 et η^2

La même question, deux types de données

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

Les deux droites de régression

Un point souvent ignoré

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

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

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

Vérification

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

Ce que l'angle entre les droites signifie

Une lecture géométrique de r^2

Les deux droites se coupent au point moyen. Elles se confondent si $r^2 = 1$, et deviennent perpendiculaires si $r = 0$ — l'une horizontale, l'autre verticale.

L'angle entre les deux droites mesure donc la qualité de l'ajustement : plus l'ajustement est bon, plus elles se rapprochent.

La bonne droite est celle de la question

Pour prévoir le salaire à partir de l'ancienneté, on utilise y sur x . Pour estimer l'ancienneté d'un salarié dont on ne connaît que le salaire, x sur y — et non l'inverse de la première.

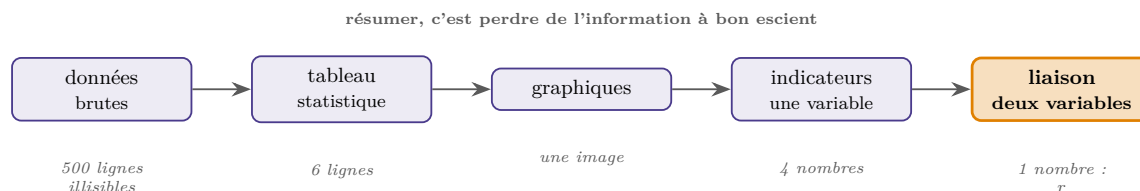
Bilan du cours

Le chemin parcouru

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 fil conducteur, refermé



Ce que vous savez faire

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.

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

Ce que le cours n'a pas permis de dire

Trois questions pour la suite

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 cours suivants du parcours. Ce cours-ci vous a donné les objets sur lesquels ils travaillent.

S'entraîner

Exercice 1 : une droite à partir de cinq sommes

Publicité et ventes sur 8 mois

Budget publicitaire x et chiffre d'affaires y , en milliers de dirhams :

$$\begin{aligned} n = 8 \quad \sum x_i = 40 \quad \sum y_i = 400 \\ \sum x_i^2 = 240 \quad \sum y_i^2 = 21\,000 \quad \sum x_i y_i = 2\,160 \end{aligned}$$

1. Calculer $\bar{x}$, $\bar{y}$, $V(x)$, $V(y)$ et $\text{cov}(x, y)$.
2. En déduire r et r^2 .
3. Déterminer la droite des moindres carrés de y en x , et prévoir les ventes pour un budget de 7 000 dirhams.
4. Peut-on dire que la publicité fait vendre ?

Exercice 1 : solution

Moments et corrélation

1. $\bar{x} = 5$, $\bar{y} = 50$; $V(x) = 240/8 - 25 = 5$; $V(y) = 21\,000/8 - 2\,500 = 125$; $\text{cov} = 2\,160/8 - 5 \times 50 = 20$.
2. $r = 20/\sqrt{5 \times 125} = 20/25 = 0,8$ et $r^2 = 0,64$.

Droite, prévision, interprétation

3. $\hat{b} = 20/5 = 4$ et $\hat{a} = 50 - 4 \times 5 = 30$: $\hat{y} = 30 + 4x$. Pour $x = 7$: $\hat{y} = 58$ milliers de dirhams.
4. Non : chaque millier de dirhams de publicité est **associé** à 4 000 dirhams de ventes en moyenne, mais une saison favorable peut augmenter à la fois le budget et les ventes. La publicité explique 64 % de la variance des ventes, sans que le calcul prouve qu'elle les cause.

Exercice 2 : lire les résidus

Les dix salariés, avec la droite $\hat{y} = 4,356 + 0,323 x$

1. Calculer le salaire ajusté et le résidu du salarié qui a 10 ans d'ancienneté et gagne 8 500 dirhams.
2. Parmi les résidus du tableau de la diapositive « La droite des dix salariés », lequel désigne le salarié le mieux payé **compte tenu de son ancienneté** ?
3. Que vaut la somme des dix résidus ? Pourquoi ?
4. Prévoir le salaire d'un salarié ayant 25 ans d'ancienneté. Quelle réserve formuler ?

Exercice 2 : solution

Résidus et prévision

1. $\hat{y} = 4,356 + 0,323 \times 10 = 7,58$ milliers, soit 7 580 dirhams ; résidu $8\,500 - 7\,580 = +920$ dirhams.
2. Le salarié de 6 ans : résidu +1 210 dirhams, le plus élevé. Il gagne autant que celui de 12 ans d'ancienneté.
3. **Zéro** : la droite passe par le point moyen, et les moindres carrés équilibrent exactement les résidus positifs et négatifs.
4. $\hat{y} = 4,356 + 0,323 \times 25 = 12,43$ milliers. Mais l'ancienneté observée s'arrête à 20 ans : c'est une **extrapolation**, rien ne garantit que la relation reste linéaire au-delà.

Ce qu'il faut retenir

Le résumé du chapitre

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 $\hat{a} = \bar{y} - \hat{b}\bar{x}$; la droite passe par le point moyen.
5. Les **résidus** somment à zéro ; $V(y) = V(\hat{y}) + V(e)$. Il faut aussi les regarder.
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

Les deux derniers chapitres refont **toute la chaîne** sur les 500 salariés — en R, puis en Python — et retrouvent chacun des nombres du cours.