

Économétrie

Séance 1 — Introduction et régression linéaire simple

Licence 3 — Économie

Anderson VIL

vladerinson@gmail.com

THEMA

CY Cergy Paris Université

Année universitaire 2026–2027

Au programme

1. Pourquoi l'économétrie ?
2. Feuille de route du cours
3. Les données économiques
4. Causalité et raisonnement *ceteris paribus*
5. Les étapes d'une analyse empirique
6. Outils de base
7. Le modèle de régression linéaire simple

Un même jeu de données nous accompagne tout au long de la séance. Il décrit 856 étudiants d'un cours d'introduction à l'économie.

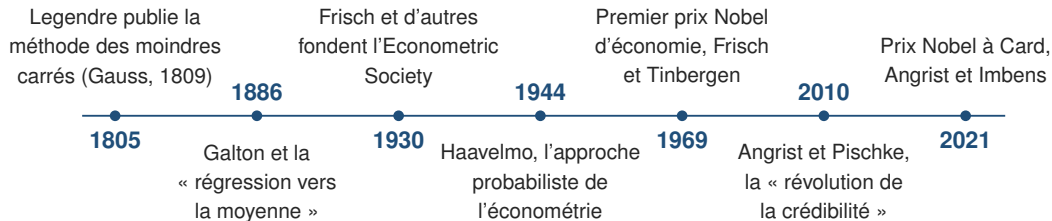
1. Pourquoi l'économétrie ?

Trois questions

- ▶ Travailler à côté de ses études fait-il baisser les notes ?
- ▶ Une formation destinée aux personnes éloignées de l'emploi augmente-t-elle leurs revenus ?
- ▶ Quand le chômage baisse, l'inflation augmente-t-elle ?

Pour y répondre, il faut des données, un modèle et une méthode. C'est l'objet de l'économétrie.

Deux siècles en sept dates



Le mot est plus ancien que la société savante. Ciompa emploie « Oekonometrie » en 1910, Frisch « économétrie » en 1926. La revue *Econometrica* paraît à partir de 1933.

À quoi sert l'économétrie ?

Mesurer

De combien la note finale augmente-t-elle avec la moyenne universitaire ?

Tester une théorie

Chômage et inflation évoluent-ils en sens inverse, comme l'annonce la courbe de Phillips ?

Évaluer une politique

Un programme de formation augmente-t-il les revenus de ceux qui le suivent ?

Prévoir

Quel sera le taux de chômage l'an prochain ?

Dans les quatre cas, on part d'une relation économique¹ et on la confronte à des données qui ne sont presque jamais issues d'une expérience contrôlée².

► Partie 3, données expérimentales et observationnelles

Corrélation et causalité

- ▶ Deux variables sont *corrélées* quand elles varient ensemble, dans le même sens ou en sens contraire. On mesurera cette liaison dans la partie 6
- ▶ x *cause* y quand une variation de x , tout le reste étant inchangé, entraîne une variation de y
- ▶ **Corrélation n'est pas causalité.** Deux variables peuvent varier ensemble parce qu'un troisième facteur agit sur les deux, ou parce que la causalité va dans l'autre sens

Un troisième facteur. Les ventes de glaces et les noyades augmentent ensemble en été. La chaleur explique les deux.

Une causalité en sens inverse. Les villes qui ont plus de policiers ont souvent plus de crimes. La criminalité pousse aussi à recruter des policiers.

Dans nos données, la corrélation entre les heures d'étude hebdomadaires et la note finale vaut 0,009. Faut-il en conclure qu'étudier ne sert à rien ? Rien ne l'autorise. Les étudiants en difficulté étudient peut-être davantage.

▶ Code

Ce que les données ne disent pas seules

- ▶ Les données sont le plus souvent *observationnelles*. On observe ce que les gens ont choisi, pas ce qui se serait passé s'ils avaient choisi autrement
- ▶ Un modèle est une approximation utile, pas la réalité. Il retient quelques facteurs et relègue tous les autres dans un terme d'erreur
- ▶ La statistique décrit des régularités. Leur lecture causale repose sur des hypothèses concernant ce que l'on n'observe pas

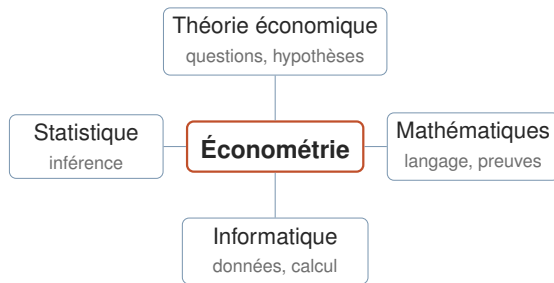
Qu'est-ce que l'économétrie ?

Définition

L'économétrie développe et applique des méthodes statistiques pour quantifier des relations économiques à partir de données, le plus souvent non expérimentales.

D'après Wooldridge (2020, section 1-1)
et Verbeek (2017, section 1.1).

Ces estimations servent à mesurer, à tester une théorie, à évaluer une politique et à prévoir. L'économétrie se distingue de la statistique par ses questions (économiques) et par ses données (observationnelles), qui ont appelé des méthodes propres.



Les grandes familles de l'économétrie

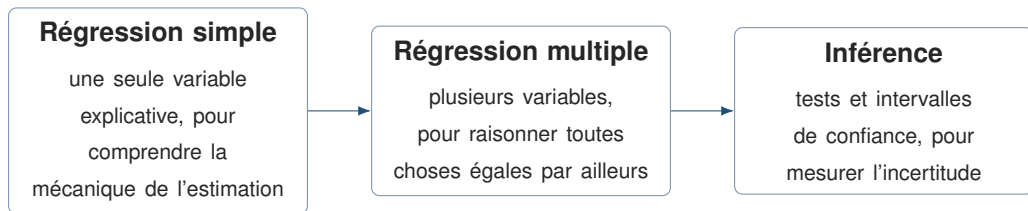
On peut classer les travaux selon trois critères, qui se combinent.

Ce qu'on étudie	microéconométrie individus, ménages, entreprises	macroéconométrie croissance, inflation	finance marchés, actifs
Comment sont les données	coupes transversales	séries temporelles	données de panel
Ce qu'on cherche	mesurer une relation	un effet causal, évaluer une politique	prévoir

Chaque case a un versant *théorique* (construire des méthodes, établir leurs propriétés) et un versant *appliqué* (répondre à des questions avec des données). Toutes reposent sur le même socle, le modèle de régression linéaire. C'est l'objet de ce cours.

2. Feuille de route du cours

Trois étapes, dans cet ordre



- ▶ Chaque étape réutilise la précédente. La régression multiple généralise la simple, l'inférence s'appuie sur les propriétés des estimateurs
- ▶ Des rappels de probabilités précèdent l'étude de ces propriétés

Trois langages, pour les mêmes résultats

- ▶ *Python* et *R*, gratuits, utilisables en ligne sans installation (notebooks Google Colab)
- ▶ *Stata*, logiciel payant très répandu en économie appliquée

Les données viennent du manuel de Wooldridge. En Python et en R, chaque bloc de code installe au besoin le paquet `wooldridge`, puis charge les données.

En Stata, on charge les fichiers `.dta` du cours, placés dans le dossier de travail.

Le code est toujours en annexe

- ▶ Le bouton [▶ Code](#) mène au code de la slide
- ▶ Le bouton [◀ Retour](#) ramène à la slide de départ
- ▶ Les onglets Python, R et Stata passent d'un langage à l'autre

Essayez avec le bouton ci-dessous.

Chaque code est autonome et peut être copié seul. Des notebooks Python et R sont fournis avec le support.

[▶ Code](#)

3. Les données économiques

Coupe transversale

<i>i</i>	score	colgpa	work	study	econhs
1	84,43	3,49	15	10	0
2	57,38	2,10	0	22,5	1
3	66,39	3,09	25	12	0
4	81,15	2,68	30	40	0
5	95,90	3,75	25	15	1
		⋮			
856			...		

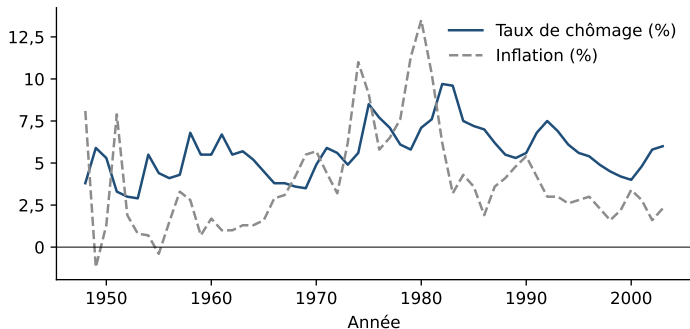
ECONMATH. 856 étudiants d'un cours d'introduction à la microéconomie (Michigan State University)

- ▶ score note finale du cours, en %
- ▶ colgpa moyenne universitaire en début de semestre, sur 4
- ▶ work, study heures hebdomadaires de travail salarié et d'étude
- ▶ econhs vaut 1 si un cours d'économie a été suivi au lycée

Des unités observées à une même date. L'ordre des lignes n'a aucune importance.

Série temporelle

year	unem	inf
1948	3,8	8,1
1949	5,9	-1,2
1950	5,3	1,3
	⋮	
2002	5,8	1,6
2003	6,0	2,3



PHILLIPS. États-Unis, 1948 à 2003, taux de chômage et inflation annuels (%). Une même unité observée dans le temps. L'ordre chronologique porte l'information et une année ressemble à la précédente.

Coupes transversales empilées

year	educ	exper	female	lwage
1978	12	8	0	1,215
1978	12	30	1	1,609
		⋮		
1985	10	27	0	2,197
1985	12	20	0	1,705
		⋮		

CPS78_85. Deux enquêtes américaines sur l'emploi, 550 personnes en 1978 et 534 en 1985

- ▶ educ, exper années d'études et d'expérience
- ▶ lwage logarithme du salaire horaire

Deux échantillons de *personnes différentes*, réunis en un seul fichier. La variable year dit à quelle enquête appartient chaque ligne.

Données de panel

nr	year	educ	exper	hours	lwage
13	1980	14	1	2672	1,198
13	1981	14	2	2320	1,853
			⋮		
17	1980	13	4	2484	1,676
17	1981	13	5	2804	1,518
			⋮		

WAGEPAN. 545 jeunes hommes américains suivis chaque année de 1980 à 1987, soit 4 360 lignes

- `nr` identifiant de la personne
- `hours` heures travaillées dans l'année

Les *mêmes* personnes sont observées plusieurs fois. On range les lignes par personne, puis par année.

Ce que chaque structure permet d'étudier

	Permet	Ne permet pas
Coupe transversale	comparer des unités à une date donnée	suivre une évolution
Série temporelle	étudier une dynamique, des retards, prévoir	traiter les années comme des tirages indépendants
Coupes empilées	comparer une relation avant et après une réforme, grossir l'échantillon	suivre une même personne
Panel	suivre les mêmes unités, neutraliser ce qui est propre à chacune et ne change pas	éviter les coûts de collecte et les sorties d'échantillon

Les premières séances portent sur les coupes transversales. Les autres structures demandent des outils supplémentaires.

Données expérimentales et données observationnelles

Expérimentales. Le chercheur fixe la variable d'intérêt, par exemple par tirage au sort.

JTRAIN2. 445 hommes en difficulté sur le marché du travail, dont 185 tirés au sort pour suivre une formation (1976–1977)

Moyennes	Formés	Non formés
Revenu 1978 (k\$ de 1982)	6,35	4,55
Années d'études	10,35	10,09
Âge	25,82	25,05

Observationnelles. Le chercheur enregistre ce que les agents ont choisi.

Dans ECONMATH, chaque étudiant a choisi ses heures de travail salarié et d'étude. Rien ne garantit que ceux qui travaillent beaucoup ressemblent aux autres.

Le tirage au sort rend les deux groupes comparables en moyenne. L'écart de revenu peut alors être attribué à la formation.

4. Causalité et raisonnement *ceteris paribus*

De combien baisserait la note d'un étudiant
s'il travaillait 15 heures ou plus par semaine au lieu de moins,
tout le reste étant inchangé?

- ▶ *Ceteris paribus* signifie « les autres facteurs pertinents étant égaux ». La plupart des questions économiques sont de cette nature
- ▶ Vous le faites déjà en microéconomie. L'effet du prix sur la quantité demandée se lit à revenu, prix des autres biens et goûts inchangés
- ▶ Ici, il faudrait tenir fixes les capacités, la motivation, les heures d'étude, le sommeil. . . dont certains ne sont pas observés

Raisonner par contrefactuel

Imaginons chaque étudiant i dans deux états du monde

- ▶ $y_i(0)$ sa note s'il travaille moins de 15 heures
- ▶ $y_i(1)$ sa note s'il travaille 15 heures ou plus

Effet causal³ pour l'étudiant i

$$y_i(1) - y_i(0)$$

Le contrefactuel fixe automatiquement « tout le reste », puisqu'il s'agit de la même personne. La difficulté est qu'une des deux notes manque toujours.

Étudiant	$y_i(0)$	$y_i(1)$	Effet
A	78	?	?
B	?	64	?
C	71	?	?

Illustration. Chaque étudiant n'est observé que dans un seul des deux états.

Ce que montrent les données

ECONMATH	Moins de 15 h	15 h ou plus
<i>Moyennes</i>		
Note finale (%)	73,95	69,17
Moyenne universitaire (sur 4)	2,86	2,70
Part des mères diplômées du supérieur	0,66	0,54
Nombre d'étudiants	615	241

- ▶ L'écart de note est de $-4,8$ points
- ▶ Mais les deux groupes diffèrent aussi par leur moyenne universitaire (2,86 contre 2,70) et par leur milieu familial. L'écart mélange l'effet du travail et ces différences

Pourquoi c'est difficile avec des données observationnelles

L'expérience idéale

- ▶ Tirer au sort les étudiants qui travaillent 15 heures ou plus
- ▶ Les deux groupes seraient comparables en moyenne, sur l'observé comme sur l'inobservé

Irréalizable ici, pour des raisons éthiques et pratiques.

Avec des données observationnelles

- ▶ Chacun choisit, et ses choix dépendent de facteurs qui influencent aussi la note
- ▶ On peut tenir compte des facteurs observés. Pour les autres, il faut faire des hypothèses

Énoncer ces hypothèses, comprendre à quoi elles servent et comment les discuter est le fil conducteur du cours.

5. Les étapes d'une analyse empirique

De la question à l'interprétation



- ▶ Tout commence par une question formulée avec soin
- ▶ Le modèle économique peut être formel (maximisation sous contrainte) ou reposer sur un raisonnement plus informel
- ▶ Le passage au modèle économétrique est l'étape décisive. C'est elle que nous détaillons

Un exemple suivi, le modèle économique

Question. Le travail salarié réduit-il la réussite dans le cours ?

Raisonnement. Le temps est rare. Une heure de travail salarié est une heure de moins pour étudier, assister aux cours ou se reposer. La réussite dépend aussi des acquis et des capacités.

$$\text{note} = f(\text{travail salarié}, \text{acquis}, \text{temps d'étude}, \text{capacités}, \dots)$$

Prédiction. À acquis, temps d'étude et capacités donnés, la note diminue avec le travail salarié.

La forme de f n'est pas précisée et certains arguments, comme les capacités, ne sont pas observables. On ne peut pas encore l'estimer.

Du modèle économique au modèle économétrique

$$\text{score} = \beta_0 + \beta_1 \text{work} + \beta_2 \text{colgpa} + \beta_3 \text{study} + u$$

Forme fonctionnelle

Linéaire. Une heure de travail en plus a le même effet, β_1 , quel que soit le point de départ.

Terme d'erreur u

Tout ce qui agit sur la note et n'est pas dans le modèle.
Capacités, motivation, sommeil, erreurs de mesure.

Variables définies

score en %, work et study en heures par semaine déclarées, colgpa sur 4.

Les β_j sont les *paramètres*, inconnus. Une hypothèse économique devient une hypothèse sur eux. « Le travail salarié est sans effet sur la note » s'écrit $\beta_1 = 0$.

Pause

6. Outils de base

$$\sum_{i=1}^n x_i \equiv x_1 + x_2 + \cdots + x_n$$

Trois propriétés (pour des constantes a, b, c)

$$\sum_{i=1}^n c = n c$$

$$\sum_{i=1}^n c x_i = c \sum_{i=1}^n x_i$$

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

Deux pièges

$$\sum_{i=1}^n \frac{x_i}{y_i} \neq \frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n y_i}$$

$$\sum_{i=1}^n x_i^2 \neq \left(\sum_{i=1}^n x_i \right)^2$$

Avec $n = 2$, $x_1^2 + x_2^2 \neq x_1^2 + 2x_1x_2 + x_2^2$.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{moyenne empirique (sample average)}$$

Les écarts à la moyenne se compensent toujours

$$\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0.$$

Exemple. $x = (2; 4; 6; 8; 10)$ donne $\bar{x} = 6$ et des écarts $(-4; -2; 0; 2; 4)$, de somme nulle.

Deux identités utiles

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i (x_i - \bar{x}) = \sum_{i=1}^n x_i^2 - n\bar{x}^2$$

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i (y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x}) y_i$$

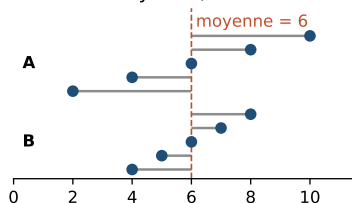
Preuve de la seconde. On développe

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i (y_i - \bar{y}) - \bar{x} \underbrace{\sum_{i=1}^n (y_i - \bar{y})}_{=0}.$$

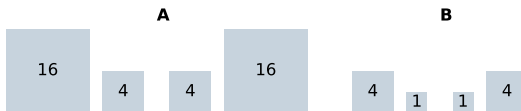
La première s'obtient avec $y_i = x_i$. On peut centrer une seule des deux variables.

Construire une mesure de dispersion

Écarts à la moyenne, de somme nulle



Carrés des écarts, d'aire $(x - \text{moyenne})$ au carré



1. **Écarts à la moyenne.** Ils disent où est chaque point par rapport au centre. Leur somme est toujours nulle, leur moyenne ne mesure donc rien
2. **Carrés des écarts.** Élever au carré rend tout positif et donne plus de poids aux points éloignés. Chaque carré est l'aire d'un carré de côté $|x_i - \bar{x}|$
3. **Moyenne des carrés.** On divise la somme par $n - 1 = 4$ plutôt que par $n = 5$ (voir la slide suivante). A donne $(16 + 4 + 0 + 4 + 16)/4 = 10$, B donne $(4 + 1 + 0 + 1 + 4)/4 = 2,5$. A est plus dispersée que B

[► Code](#)

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \qquad s_x = \sqrt{s_x^2}$$

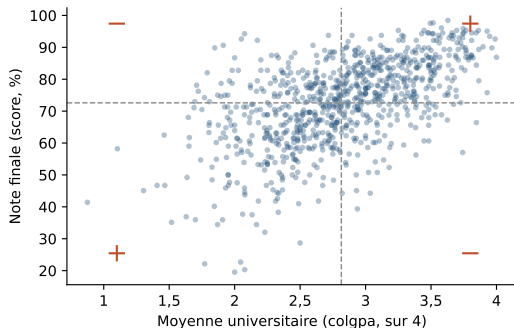
- ▶ La variance mesure la dispersion autour de la moyenne. Elle s'exprime dans le carré de l'unité de x , l'écart-type dans l'unité de x
- ▶ $s_x^2 = 0$ si et seulement si toutes les valeurs x_i sont égales
- ▶ On divise par $n - 1$ et non par n . La raison est statistique et sera donnée avec les rappels de probabilités. Certains ouvrages divisent par n , ce qui ne change rien quand n est grand

Covariance et corrélation empiriques

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \in [-1, 1]$$

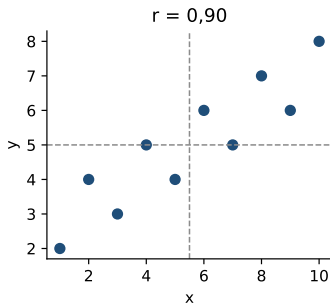
- ▶ Le signe de chaque produit dépend du quadrant où tombe le point
- ▶ s_{xy} dépend des unités, r_{xy} non
- ▶ r_{xy} mesure une liaison *linéaire*



ECONMATH, $s_{xy} = 4,13$ et $r_{xy} = 0,57$.

▶ Code

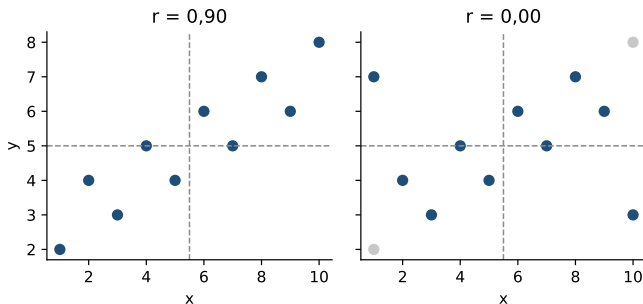
Déplacer des points change le signe



Départ. Presque tous les points sont dans les quadrants $+$, $s_{xy} = 5$ et $r_{xy} = 0,90$.

La moyenne de y reste égale à 5. Une version interactive, où l'on déplace soi-même les points, est fournie avec le support.

Déplacer des points change le signe

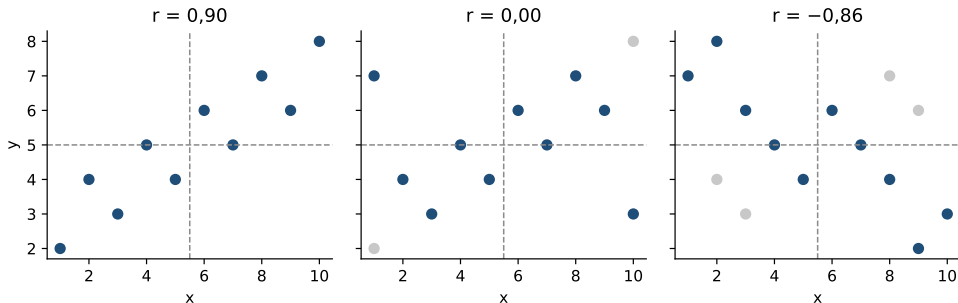


On déplace deux points (en gris, leur ancienne place) vers des quadrants $-$. Les produits positifs et négatifs se compensent, $s_{xy} = 0$.

La moyenne de y reste égale à 5. Une version interactive, où l'on déplace soi-même les points, est fournie avec le support.

[► Code](#)

Déplacer des points change le signe



On en déplace quatre autres. Les quadrants – dominant, $s_{xy} = -4,78$ et $r_{xy} = -0,86$.

La moyenne de y reste égale à 5. Une version interactive, où l'on déplace soi-même les points, est fournie avec le support.

Un exemple à la main

Cinq étudiants fictifs. x heures de travail personnel par semaine, y note sur 20.

i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	2	7	-4	-4	16	16	16
2	4	10	-2	-1	2	4	1
3	6	11	0	0	0	0	0
4	8	14	2	3	6	4	9
5	10	13	4	2	8	16	4
Σ	30	55	0	0	32	40	30

$$\bar{x} = 6 \quad \bar{y} = 11 \quad s_x^2 = 40/4 = 10 \quad s_y^2 = 30/4 = 7,5 \quad s_{xy} = 32/4 = 8$$
$$r_{xy} = 8/\sqrt{10 \times 7,5} = 0,924 \text{ Le code donne les mêmes valeurs.}$$

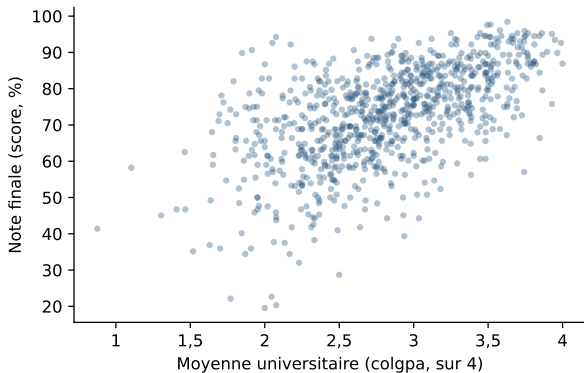
7. Le modèle de régression linéaire simple

La question du fil rouge

Les étudiants qui arrivent avec une meilleure moyenne réussissent-ils mieux ce cours ?

De combien ?

856 étudiants de ECONMATH. Chaque point est un étudiant.



► Code

Modèle de régression linéaire simple

$$y = \beta_0 + \beta_1 x + u$$

Il est supposé valoir dans la population étudiée. β_0 et β_1 sont des paramètres inconnus.

	Français	Anglais
y	variable dépendante, expliquée	<i>dependent, explained variable</i>
x	variable indépendante, explicative, régresseur	<i>independent, explanatory variable</i>
u	terme d'erreur, perturbation	<i>error term, disturbance</i>
β_0	constante	<i>intercept</i>
β_1	pente	<i>slope</i>

« Indépendante » n'a ici rien à voir avec l'indépendance statistique.

La pente

$$\Delta y = \beta_1 \Delta x \quad \text{si } \Delta u = 0$$

β_1 est la variation de y quand x augmente d'une unité, les autres facteurs contenus dans u restant fixes. L'effet est le même quel que soit le niveau de départ de x .

Pour que l'estimation de β_1 ait une lecture *ceteris paribus*, il faudra une hypothèse sur le lien entre u et x . Ce sera l'objet de la séance suivante.

Le terme d'erreur

Tous les facteurs autres que x qui agissent sur y . Dans le fil rouge

$$\text{score} = \beta_0 + \beta_1 \text{colgpa} + u$$

u contient la motivation, le niveau en mathématiques, les heures d'étude et de travail salarié. . .

Dans la population

$$y = \beta_0 + \beta_1 x + u$$

β_0 et β_1 sont des nombres fixes et inconnus.

Dans l'échantillon

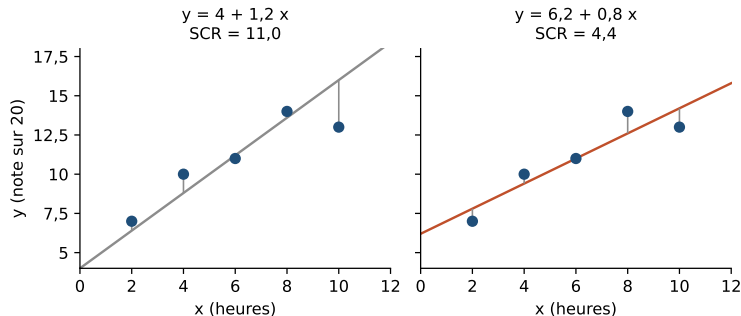
$$\{(x_i, y_i), i = 1, \dots, n\}$$

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

On observe x_i et y_i , jamais u_i .

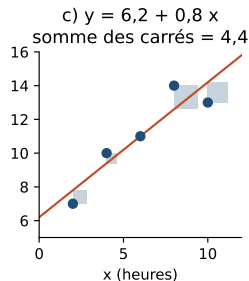
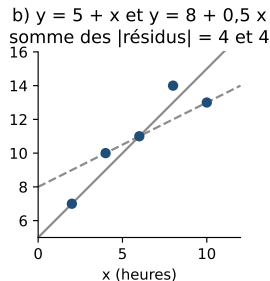
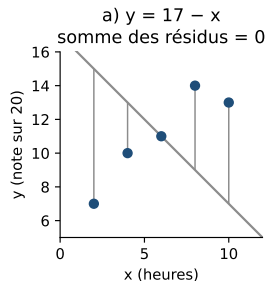
On cherche à calculer, à partir des seules données, des valeurs $\hat{\beta}_0$ et $\hat{\beta}_1$ (« bêta chapeau ») qui approchent β_0 et β_1 . Graphiquement, il s'agit de faire passer une droite dans le nuage de points.

Quelle droite choisir ?



Pour une droite $b_0 + b_1 x$, le *résidu* de l'observation i est l'écart vertical $y_i - b_0 - b_1 x_i$. On retient la droite qui rend la *somme des carrés des résidus* (SCR) la plus petite possible.

Pourquoi des carrés ?



a) Somme des résidus. Toute droite qui passe par $(\bar{x}, \bar{y})$ donne une somme nulle, même une droite de pente négative. Ce critère ne départage rien.

b) Somme des valeurs absolues. Elle a du sens et s'utilise parfois. Ici, deux droites différentes atteignent le même minimum, 4. Il n'existe pas de formule explicite.

c) Somme des carrés. Chaque résidu compte pour l'aire d'un carré, ce qui pénalise les grands écarts. Le minimum est unique et donné par une formule.

C'est la méthode des **moindres carrés ordinaires**, MCO (*ordinary least squares* OLS).

[► Code](#)

Le problème de minimisation

La somme des carrés des résidus de la droite retenue s'écrit

$$\sum_{i=1}^n \hat{u}_i^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

Pour une droite quelconque $b_0 + b_1 x$, on note cette somme

$Q(b_0, b_1) = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$. Les MCO choisissent le couple $(\hat{\beta}_0, \hat{\beta}_1)$ qui minimise Q .

On y annule les deux dérivées partielles. Ce sont les *conditions du premier ordre* (CPO)

$$\frac{\partial Q}{\partial b_0} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (\text{CPO 1})$$

$$\frac{\partial Q}{\partial b_1} = -2 \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (\text{CPO 2})$$

Deux équations, deux inconnues. La solution est bien le minimum, et elle est unique⁴.

CPO 1. En divisant par $-2n$, on obtient $\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} = 0$, donc

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

CPO 2. En remplaçant $\hat{\beta}_0$,

$$\sum_{i=1}^n x_i (y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^n x_i (x_i - \bar{x}).$$

Avec les deux identités utiles, les deux sommes s'écrivent en écarts à la moyenne

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})^2.$$

Résultat

Si $\sum_{i=1}^n (x_i - \bar{x})^2 > 0$,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- ▶ La condition exclut seulement le cas où tous les x_i sont égaux. On ne peut alors tracer aucune pente
- ▶ $\hat{\beta}_1 = r_{xy} s_y / s_x$. La pente a **toujours** le signe de la corrélation
- ▶ La régression simple est donc, au fond, une **analyse de corrélation**. D'où la prudence avant toute lecture causale

Avec les sommes déjà calculées

$$\hat{\beta}_1 = \frac{32}{40} = 0,8$$

$$\hat{\beta}_0 = 11 - 0,8 \times 6 = 6,2$$

$$\hat{y} = 6,2 + 0,8x$$

Une heure de travail personnel en plus par semaine est associée à une note supérieure de 0,8 point sur 20.

C'est la droite orange de la slide « [Quelle droite choisir ?](#) ». Le code retrouve les mêmes valeurs.

Valeur prédite (*fitted value*)

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Résidu (*residual*)

$$\hat{u}_i = y_i - \hat{y}_i$$

Si $\hat{u}_i > 0$, la droite sous-estime y_i .

Le résidu $\hat{u}_i$ n'est pas l'erreur u_i . Le résidu se calcule avec les données, l'erreur n'est jamais observée.

Droite de régression de l'échantillon

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Elle estime la droite de la population, qui est fixe et inconnue. Un autre échantillon donnerait une autre droite.

Trois propriétés algébriques

Elles valent pour *tout* échantillon, par construction.

1. $\sum_{i=1}^n \hat{u}_i = 0$ (CPO 1)

2. $\sum_{i=1}^n x_i \hat{u}_i = 0$ (CPO 2)
la covariance empirique entre x et $\hat{u}$ est nulle

3. Le point $(\bar{x}, \bar{y})$ est sur la droite $(\hat{\beta}_0)$

x_i	y_i	$\hat{y}_i$	$\hat{u}_i$	$x_i \hat{u}_i$
2	7	7,8	-0,8	-1,6
4	10	9,4	0,6	2,4
6	11	11,0	0,0	0,0
8	14	12,6	1,4	11,2
10	13	14,2	-1,2	-12,0
30	55	55,0	0,0	0,0

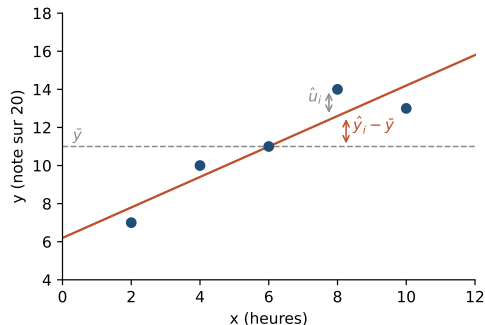
$$6,2 + 0,8 \times 6 = 11 = \bar{y}$$

Décomposer la variation de y

Pour chaque observation

$$y_i - \bar{y} = \underbrace{(\hat{y}_i - \bar{y})}_{\text{expliqué par } x} + \underbrace{\hat{u}_i}_{\text{non expliqué}}$$

Sur la figure, l'étudiant 4 ($x = 8$, $y = 14$). Sa note dépasse la moyenne de 3 points, dont 1,6 sont prédits par ses heures de travail et 1,4 restent inexpliqués.



[Code](#)

SCT = SCE + SCR

$$\text{SCT} = \sum_{i=1}^n (y_i - \bar{y})^2 \quad \text{SCE} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad \text{SCR} = \sum_{i=1}^n \hat{u}_i^2$$

Sommes des carrés totale, expliquée, résiduelle (*SST*, *SSE*, *SSR*).

Preuve. On élève au carré la décomposition et on somme

$$\text{SCT} = \text{SCE} + \text{SCR} + 2 \sum_{i=1}^n \hat{u}_i (\hat{y}_i - \bar{y}).$$

Le dernier terme est nul, car $\hat{y}_i - \bar{y} = \hat{\beta}_0 - \bar{y} + \hat{\beta}_1 x_i$, d'où

$$\sum_{i=1}^n \hat{u}_i (\hat{y}_i - \bar{y}) = (\hat{\beta}_0 - \bar{y}) \sum_{i=1}^n \hat{u}_i + \hat{\beta}_1 \sum_{i=1}^n x_i \hat{u}_i = 0 \quad (\text{propriétés 1 et 2}).$$

Le coefficient de détermination

R^2

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT} \in [0,1]$$

Part de la variation de y dans l'échantillon qui est expliquée par x .

- ▶ $R^2 = 1$ si tous les points sont sur la droite, R^2 proche de 0 si la droite n'explique presque rien
- ▶ En régression simple, $R^2 = r_{xy}^2$, d'où son nom
- ▶ Exemple à la main. $SCT = 30$, $SCE = 25,6$, $SCR = 4,4$, donc $R^2 = 0,853 = 0,924^2$

Ce que le R^2 mesure, et ce qu'il ne mesure pas

Il mesure

- ▶ la qualité de l'ajustement *dans l'échantillon*
- ▶ la part de la dispersion de y associée à x

Il ne mesure pas

- ▶ si $\hat{\beta}_1$ est proche de β_1
- ▶ si la relation est causale
- ▶ la qualité d'un modèle en soi

Dans l'expérience JTRAIN2, la régression du revenu de 1978 sur la formation donne $R^2 = 0,018$. Le tirage au sort permet pourtant d'estimer l'effet de la formation, 1 790 dollars de 1982 en moyenne. Beaucoup d'autres facteurs expliquent le revenu, ce qui est normal avec des données individuelles.

Une première estimation

	(1)	(2)
colgpa	14,32*** (0,70)	
actmth		1,44*** (0,11)
Constante	32,31*** (2,00)	39,21*** (2,66)
<i>n</i>	856	814
<i>R</i> ²	0,329	0,167

Variable dépendante score, note finale du cours en %. Estimation par les MCO.

actmth score au test ACT de mathématiques, passé avant l'université.

Erreurs-types entre parenthèses. *** significatif à 1 %, ** à 5 %, * à 10 %.

Erreurs-types et étoiles seront étudiées avec l'inférence. En colonne (2), 42 étudiants sans score ACT sont écartés.

Une table de régression se présente ainsi dans les articles. Chaque colonne est une régression.

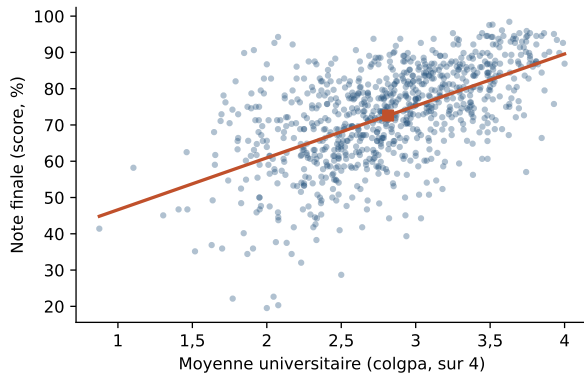
Colonne (1), $\widehat{\text{score}} = 32,31 + 14,32 \text{ colgpa}$

- ▶ Un point de moyenne universitaire en plus (sur 4) est associé à une note finale supérieure de 14,3 points de pourcentage
- ▶ Pour $\text{colgpa} = 3$, la note prédite est $32,31 + 14,32 \times 3 = 75,25$
- ▶ La constante est la note prédite pour $\text{colgpa} = 0$. Aucun étudiant n'en est proche (minimum 0,88), elle n'a pas d'interprétation utile
- ▶ $R^2 = 0,329$. La moyenne universitaire explique un tiers de la dispersion des notes

Colonne (2). Un point de plus au test ACT de mathématiques est associé à 1,44 point de note en plus, avec $R^2 = 0,167$.

Ces pentes décrivent des associations. Chacune mélange l'effet de sa variable et celui des facteurs restés dans u .

La droite des MCO sur le nuage



$$\widehat{\text{score}} = 32,31 + 14,32 \text{ colgpa}$$

La droite passe par le point moyen (2,815 ; 72,60), marqué d'un carré.

À moyenne donnée, les notes restent très dispersées autour de la droite. C'est ce que traduit $R^2 = 0,33$.

Valeurs prédites et résidus dans les données

Étudiant	colgpa	score	$\widehat{\text{score}}$	$\hat{u}$
1	3,49	84,43	82,28	2,15
2	2,10	57,38	62,37	-4,99
3	3,09	66,39	76,47	-10,08
4	2,68	81,15	70,68	10,47
5	3,75	95,90	85,92	9,98

- ▶ L'étudiant 4 obtient 10,5 points de plus que ce que prédit sa moyenne, l'étudiant 3 en obtient 10,1 de moins
- ▶ Sur les 856 étudiants, $\sum \hat{u}_i = 0$ et $153\,539,3 = 50\,502,7 + 103\,036,6$, soit $SCT = SCE + SCR$

L'essentiel de la séance

- ▶ L'économétrie mesure des relations économiques à partir de données, le plus souvent observationnelles
- ▶ Une question causale est une question *ceteris paribus*. Avec des données observationnelles, y répondre exige des hypothèses
- ▶ Les MCO choisissent la droite qui minimise la somme des carrés des résidus
 $\hat{\beta}_1 = \sum_i (x_i - \bar{x})(y_i - \bar{y}) / \sum_i (x_i - \bar{x})^2$ et $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- ▶ Les résidus somment à zéro et sont non corrélés à x . $SCT = SCE + SCR$,
 $R^2 = SCE/SCT$

- ▶ Rappels de probabilités. Espérance, variance, covariance, espérance conditionnelle
- ▶ Sous quelles hypothèses $\hat{\beta}_1$ approche-t-il β_1 ? Le rôle de l'hypothèse sur u et x
- ▶ Absence de biais et variance des estimateurs des MCO

Pour préparer, Wooldridge, annexe B (probabilités) et section 2-5.

Travaux dirigés

Série d'exercices de la séance 1

10 questions de quiz et 5 exercices

(support distinct)

Annexes

Explications

Annexe 1. Les relations économiques des quatre exemples

Usage	Relation économique	Forme estimée
Mesurer	La réussite dans un cours dépend des acquis antérieurs	$\text{score} = \beta_0 + \beta_1 \text{colgpa} + u$, avec β_1 les points de note par point de moyenne
Tester une théorie	Courbe de Phillips. Un chômage plus bas va de pair avec une inflation plus forte	$\text{inf}_t = \beta_0 + \beta_1 \text{unem}_t + u_t$. La théorie prédit $\beta_1 < 0$
Évaluer une politique	Le revenu dépend de la formation suivie	$\text{re78} = \beta_0 + \beta_1 \text{train} + u$, avec β_1 l'effet moyen de la formation
Prévoir	Le chômage d'une année dépend de celui de l'année précédente	$\text{unem}_t = \beta_0 + \beta_1 \text{unem}_{t-1} + u_t$, d'où la prévision pour l'an prochain

Une relation économique dit quelles variables sont liées et, souvent, dans quel sens. L'économétrie lui donne une forme estimable et un ordre de grandeur.

[◀ Retour](#)

Annexe 2. Expérience contrôlée

Expérience contrôlée

Dans une expérience contrôlée (*controlled experiment*), le chercheur décide lui-même, le plus souvent par tirage au sort, quelles unités reçoivent le traitement, puis il observe le résultat.

Exemple. Dans JTRAIN2, 445 hommes éloignés de l'emploi ont été tirés au sort, 185 pour suivre une formation et 260 pour servir de témoins. On compare ensuite leurs revenus de 1978.

Le tirage au sort garantit que rien d'autre que la formation ne distingue les deux groupes, en moyenne. C'est ce qui manque aux autres sources.

Annexe 2. Les autres sources de données

Enquêtes	questionnaires auprès d'un échantillon, comme l'enquête Emploi de l'Insee ou la CPS américaine (CPS78_85)
Données administratives	produites par la gestion d'un dispositif, comme les fichiers fiscaux ou les déclarations des employeurs
Statistiques agrégées	comptabilité nationale, prix, chômage, comme PHILLIPS
Expériences naturelles <i>natural experiments</i>	un événement extérieur crée des groupes comparables. La hausse du salaire minimum au New Jersey en 1992, comparé à la Pennsylvanie voisine (Card et Krueger, 1994)
Données numériques	transactions, requêtes en ligne, images satellitaires

Ces données sont *observationnelles*. Seule l'expérience naturelle s'approche d'un tirage au sort, sans que le chercheur le décide.

[◀ Retour](#)

Annexe 3. Qu'est-ce qu'un effet causal ?

Effet causal

L'effet causal de x sur y est la variation de y provoquée par une variation de x , tout le reste étant inchangé.

Exemples. L'effet d'un médicament sur la tension, d'une formation sur le revenu, d'une baisse de prix sur les ventes.

Pourquoi $y_i(1) - y_i(0)$ est l'effet causal. Les deux notes concernent la même personne, au même moment, dans deux situations qui ne diffèrent que par le traitement (travailler 15 heures ou plus). Capacités, motivation et milieu sont identiques par construction. La différence ne peut venir que du traitement. C'est le *ceteris paribus* appliqué à un individu.

La difficulté. On n'observe jamais les deux notes d'une même personne. C'est le « problème fondamental de l'inférence causale » (*fundamental problem of causal inference*, Holland, 1986).

Annexe 3. Un exemple, avec les deux mondes connus

Étudiant	$y_i(0)$	$y_i(1)$	Effet	Choix
A	78	72	-6	< 15 h
B	60	56	-4	$\geq$ 15 h
C	72	67	-5	< 15 h
D	58	55	-3	$\geq$ 15 h
Effet moyen			-4,5	

L'effet varie d'une personne à l'autre. On résume souvent par sa moyenne, l'effet moyen du traitement (*average treatment effect*, ATE).

Chiffres fictifs. En réalité, seule la note correspondant au choix est observée.

On observe 78 et 72 chez ceux qui travaillent moins, 56 et 55 chez les autres. L'écart des moyennes vaut $55,5 - 75 = -19,5$, bien plus que l'effet moyen de $-4,5$. Il se décompose en

$$-19,5 = \underbrace{(55,5 - 59)}_{\text{effet chez B et D}} + \underbrace{(59 - 75)}_{\text{écart de départ}}.$$

Ceux qui travaillent avaient de toute façon des notes plus basses. Le tirage au sort supprimerait, en moyenne, cet écart de départ.

Annexe 4. Les MCO donnent bien le minimum, et il est unique

Pour une droite quelconque, on écrit $y_i - b_0 - b_1 x_i = \hat{u}_i + d_i$ avec $d_i = (\hat{\beta}_0 - b_0) + (\hat{\beta}_1 - b_1)x_i$. En élevant au carré et en sommant,

$$Q(b_0, b_1) = \sum_{i=1}^n \hat{u}_i^2 + \sum_{i=1}^n d_i^2 + 2(\hat{\beta}_0 - b_0) \underbrace{\sum_{i=1}^n \hat{u}_i}_{=0 \text{ (CPO 1)}} + 2(\hat{\beta}_1 - b_1) \underbrace{\sum_{i=1}^n x_i \hat{u}_i}_{=0 \text{ (CPO 2)}}.$$

Résultat

$Q(b_0, b_1) = \sum_{i=1}^n \hat{u}_i^2 + \sum_{i=1}^n d_i^2 \geq \sum_{i=1}^n \hat{u}_i^2$, avec égalité seulement si tous les d_i sont nuls. Les x_i n'étant pas tous égaux, la droite $(\hat{\beta}_0 - b_0) + (\hat{\beta}_1 - b_1)x$ s'annule alors en deux valeurs distinctes de x . Elle est donc nulle, et $b_0 = \hat{\beta}_0$, $b_1 = \hat{\beta}_1$.

Exemple à la main. Pour la droite $4 + 1,2x$, les $d_i = 2,2 - 0,4x_i$ valent 1,4 ; 0,6 ; -0,2 ; -1 ; -1,8. Leurs carrés somment à 6,6, et $Q(4 ; 1,2) = 4,4 + 6,6 = 11,0$, comme sur la slide « Quelle droite choisir ? ». Démonstration adaptée de Wooldridge (2020, annexe 2A), qui ne détaille pas l'unicité.

Annexes

Code Python, R et Stata

```
import subprocess, sys
subprocess.run([sys.executable, "-m", "pip", "install", "-q",
               "--disable-pip-version-check", "wooldridge"])      # installe si absent
import wooldridge as wd
econ = wd.data("econmath")                                       # 856 étudiants
r = econ["study"].corr(econ["score"])                             # corrélation
print("corrélation study-score :", round(r, 3))
```

Code R, heures d'étude et note

Python

R

Stata

```
if (!requireNamespace("wooldridge", quietly = TRUE)) install.packages("wooldridge")
library(wooldridge)
data("econmath") # 856 étudiants
round(cor(econmath$study, econmath$score), 3) # corrélation
```

[◀ Retour](#)

Code Stata, heures d'étude et note

[Python](#)[R](#)[Stata](#)

```
use econmath.dta, clear  
correlate study score
```

```
// 856 étudiants
```

[◀ Retour](#)

```
import subprocess, sys
subprocess.run([sys.executable, "-m", "pip", "install", "-q",
               "--disable-pip-version-check", "wooldridge"])    # installe si absent
import wooldridge as wd
econ = wd.data("econmath")                                     # charge ECONMATH
print(econ.shape)                                              # lignes, colonnes
print(econ.head())                                             # premières lignes
wd.data("econmath", description=True)                          # description des variables
```

Code R, charger des données de Wooldridge

Python

R

Stata

```
if (!requireNamespace("wooldridge", quietly = TRUE)) install.packages("wooldridge")
library(wooldridge)
data("econmath")           # charge ECONMATH
dim(econmath)              # lignes, colonnes
head(econmath)             # premières lignes
# ?econmath affiche la description des variables
```

[◀ Retour](#)

Code Stata, charger des données de Wooldridge

Python

R

Stata

```
use econmath.dta, clear
```

```
describe
```

```
list in 1/5
```

```
// charge ECONMATH
```

```
// variables et libellés
```

```
// premières lignes
```

[◀ Retour](#)

```
import subprocess, sys
subprocess.run([sys.executable, "-m", "pip", "install", "-q",
               "--disable-pip-version-check", "wooldridge"])    # installe si absent
import wooldridge as wd
econ = wd.data("econmath")
print(econ[["score", "colgpa", "work", "study", "econhs"]].head(5))
print("n =", len(econ))
```

```
if (!requireNamespace("wooldridge", quietly = TRUE)) install.packages("wooldridge")
library(wooldridge)
data("econmath")
head(econmath[, c("score", "colgpa", "work", "study", "econhs")], 5)
nrow(econmath)
```

```
use econmath.dta, clear           // fichier du cours, dossier de travail  
list score colgpa work study econhs in 1/5  
count
```

```
import subprocess, sys
subprocess.run([sys.executable, "-m", "pip", "install", "-q",
    "--disable-pip-version-check", "wooldridge"])    # installe si absent
import wooldridge as wd
import matplotlib.pyplot as plt
phil = wd.data("phillips")
print(phil[["year", "unem", "inf"]].head(3))
print(phil[["year", "unem", "inf"]].tail(2))
fig, ax = plt.subplots(figsize=(6, 3))
ax.plot(phil["year"], phil["unem"], color="#1F4E79", label="Taux de chômage (%)")
ax.plot(phil["year"], phil["inf"], color="#8C8C8C", ls="--", label="Inflation (%)")
ax.axhline(0, color="black", lw=0.5)
ax.set_xlabel("Année")
ax.legend(frameon=False)
ax.spines[["top", "right"]].set_visible(False)
ax.yaxis.set_major_formatter(lambda v, p: f"{v:g}".replace(".", ","))
fig.tight_layout()
fig.savefig("fig_phillips.pdf")
plt.show()
```

```
if (!requireNamespace("wooldridge", quietly = TRUE)) install.packages("wooldridge")
library(wooldridge)
data("phillips")
head(phillips[, c("year", "unem", "inf")], 3)
tail(phillips[, c("year", "unem", "inf")], 2)
plot(phillips$year, phillips$unem, type = "l", col = "#1F4E79", lwd = 1.5,
      ylim = range(phillips$unem, phillips$inf), xlab = "Année", ylab = "", yaxt = "n")
axis(2, at = axTicks(2), labels = sub(".", ", ", axTicks(2)), fixed = TRUE, las = 1)
lines(phillips$year, phillips$inf, col = "#8C8C8C", lty = 2)
abline(h = 0, lwd = 0.5)
legend("topleft", c("Taux de chômage (%)", "Inflation (%)"),
      col = c("#1F4E79", "#8C8C8C"), lty = c(1, 2), bty = "n")
```

```
use phillips.dta, clear           // fichier du cours, dossier de travail
list year unem inf in 1/3
list year unem inf in -2/1
twoway (line unem year, lcolor("31 78 121")) (line inf year, lcolor(gs9) lpattern(dash)), ///
       yline(0, lwidth(vthin)) xtitle("Année") ytitle("") ///
       legend(order(1 "Taux de chômage (%)" 2 "Inflation (%)") position(11) ring(0) cols(1))
graph export "fig_phillips_stata.pdf", replace
```

```
import subprocess, sys
subprocess.run([sys.executable, "-m", "pip", "install", "-q",
               "--disable-pip-version-check", "wooldridge"])    # installe si absent
import wooldridge as wd
cps = wd.data("cps78_85")
var = ["year", "educ", "exper", "female", "lwage"]
print(cps[cps["year"] == 78][var].head(2))    # enquête de 1978
print(cps[cps["year"] == 85][var].head(2))    # enquête de 1985
print(cps["year"].value_counts())
```

Code R, coupes transversales empilées

[Python](#)**R**[Stata](#)

```
if (!requireNamespace("wooldridge", quietly = TRUE)) install.packages("wooldridge")
library(wooldridge)
data("cps78_85")
var <- c("year", "educ", "exper", "female", "lwage")
head(cps78_85[cps78_85$year == 78, var], 2)      # enquête de 1978
head(cps78_85[cps78_85$year == 85, var], 2)      # enquête de 1985
table(cps78_85$year)
```

[◀ Retour](#)

```
use cps78_85.dta, clear           // fichier du cours, dossier de travail
list year educ exper female lwage in 1/2      // enquête de 1978
list year educ exper female lwage in 551/552   // enquête de 1985
tabulate year
```

```
import subprocess, sys
subprocess.run([sys.executable, "-m", "pip", "install", "-q",
               "--disable-pip-version-check", "wooldridge"])    # installe si absent
import wooldridge as wd
pan = wd.data("wagepan")
choix = pan["nr"].isin([13, 17]) & (pan["year"] <= 1981)
print(pan[choix][["nr", "year", "educ", "exper", "hours", "lwage"]])
print("individus :", pan["nr"].nunique(), " années :", pan["year"].nunique())
```

```
if (!requireNamespace("wooldridge", quietly = TRUE)) install.packages("wooldridge")
library(wooldridge)
data("wagepan")
subset(wagepan, nr %in% c(13, 17) & year <= 1981,
       select = c(nr, year, educ, exper, hours, lwage))
length(unique(wagepan$nr)); length(unique(wagepan$year))
```

```
use wagepan.dta, clear           // fichier du cours, dossier de travail
list nr year educ exper hours lwage if inlist(nr, 13, 17) & year <= 1981
xtset nr year                   // déclare la structure de panel
```

```
import subprocess, sys
subprocess.run([sys.executable, "-m", "pip", "install", "-q",
               "--disable-pip-version-check", "wooldridge"])      # installe si absent
import wooldridge as wd
jt = wd.data("jtrain2")
print(jt.groupby("train")[["re78", "educ", "age"]].mean().round(2))
print(jt["train"].value_counts())                                # effectifs
```

```
if (!requireNamespace("wooldridge", quietly = TRUE)) install.packages("wooldridge")
library(wooldridge)
data("jtrain2")
round(aggregate(cbind(re78, educ, age) ~ train, data = jtrain2, FUN = mean), 2)
table(jtrain2$train)                                # effectifs
```

```
use jtrain2.dta, clear           // fichier du cours, dossier de travail
tabstat re78 educ age, by(train) statistics(mean) format(%9.2f)
tabulate train                  // effectifs
```

Code Python, travailler 15 heures ou plus

[Python](#)[R](#)[Stata](#)

```
import subprocess, sys
subprocess.run([sys.executable, "-m", "pip", "install", "-q",
               "--disable-pip-version-check", "wooldridge"])      # installe si absent
import wooldridge as wd
econ = wd.data("econmath")
econ["w15"] = (econ["work"] >= 15).astype(int)      # 1 si 15 h ou plus
print(econ.groupby("w15")[["score", "colgpa", "mothcoll"]].mean().round(2))
print(econ["w15"].value_counts())                  # effectifs
```

[◀ Retour](#)

Code R, travailler 15 heures ou plus

[Python](#)**R**[Stata](#)

```
if (!requireNamespace("wooldridge", quietly = TRUE)) install.packages("wooldridge")
library(wooldridge)
data("econmath")
econmath$w15 <- as.integer(econmath$work >= 15) # 1 si 15 h ou plus
round(aggregate(cbind(score, colgpa, mothcoll) ~ w15, data = econmath, FUN = mean), 2)
table(econmath$w15) # effectifs
```

[◀ Retour](#)

Code Stata, travailler 15 heures ou plus

[Python](#)[R](#)[Stata](#)

```
use econmath.dta, clear           // fichier du cours, dossier de travail
generate w15 = (work >= 15)       // 1 si 15 h ou plus
tabstat score colgpa mothcoll, by(w15) statistics(mean) format(%9.2f)
tabulate w15                      // effectifs
```

[◀ Retour](#)

```
import numpy as np
import matplotlib.pyplot as plt
a = np.array([2, 4, 6, 8, 10]); b = np.array([4, 5, 6, 7, 8])      # même moyenne, 6
print("écarts A :", a - 6, " somme :", (a - 6).sum())
print("carrés A :", (a - 6)**2, " variance A :", a.var(ddof=1))
print("carrés B :", (b - 6)**2, " variance B :", b.var(ddof=1))
fig, (g, d) = plt.subplots(1, 2, figsize=(8, 2.2), width_ratios=[1, 1.6])
ha = [1.9, 2.2, 2.5, 2.8, 3.1]; hb = [0.2, 0.5, 0.8, 1.1, 1.4]  # hauteurs des points
g.hlines(ha, 6, a, color="#8C8C8C"); g.hlines(hb, 6, b, color="#8C8C8C") # écarts
g.scatter(a, ha, color="#1F4E79", zorder=3); g.scatter(b, hb, color="#1F4E79", zorder=3)
g.axvline(6, color="#C0502B", ls="--", lw=1); g.text(6.15, 3.3, "moyenne = 6", color="#C0502B")
g.text(0.3, 2.4, "A", fontweight="bold"); g.text(0.3, 0.7, "B", fontweight="bold")
g.set_xlim(0, 11.5); g.set_ylim(0, 3.7); g.set_yticks([])
g.spines[["top", "right", "left"]].set_visible(False)
g.set_title("Écarts à la moyenne, de somme nulle")
cotes = list(np.abs(a - 6)) + list(np.abs(b - 6))                # côtés des carrés, A puis B
pos = [0, 4.6, 7.2, 7.8, 10.4, 16.5, 19.1, 20.7, 21.3, 22.9]    # positions horizontales
[d.add_patch(plt.Rectangle((p, 0), c, c, fc="#1F4E79", alpha=0.25)) for p, c in zip(pos, cotes)]
[d.text(p + c / 2, c / 2, c * c, ha="center", va="center") for p, c in zip(pos, cotes) if c > 0]
d.text(7.2, 4.4, "A", fontweight="bold"); d.text(20.7, 4.4, "B", fontweight="bold")
d.set_xlim(-0.5, 25.5); d.set_ylim(-0.3, 5.2); d.set_aspect("equal"); d.axis("off")
d.set_title("Carrés des écarts, d'aire (x - moyenne) au carré")
fig.tight_layout(); fig.savefig("fig_variance.pdf"); plt.show()
```

```
a <- c(2, 4, 6, 8, 10)      # série A, dispersée
b <- c(4, 5, 6, 7, 8)      # série B, même moyenne, plus resserrée
a - mean(a); sum(a - mean(a))      # écarts, de somme nulle
(a - mean(a))^2; var(a)      # carrés et variance de A
(b - mean(b))^2; var(b)      # carrés et variance de B
layout(matrix(1:2, 1), widths = c(1, 1.6)); par(mar = c(3, 1, 2, 1))
ha <- c(1.9, 2.2, 2.5, 2.8, 3.1); hb <- c(0.2, 0.5, 0.8, 1.1, 1.4)
plot(a, ha, pch = 16, col = "#1F4E79", xlim = c(0, 11.5), ylim = c(0, 3.7),
     yaxt = "n", xlab = "", ylab = "", main = "Écarts à la moyenne, de somme nulle",
     cex.main = 0.9, font.main = 1)
points(b, hb, pch = 16, col = "#1F4E79")
segments(6, ha, a, ha, col = "#8C8C8C"); segments(6, hb, b, hb, col = "#8C8C8C")
abline(v = 6, col = "#C0502B", lty = 2)
text(c(0.3, 0.3), c(2.4, 0.7), c("A", "B"), font = 2)
cotes <- c(abs(a - 6), abs(b - 6))      # côtés des carrés, A puis B
pos <- c(0, 4.6, 7.2, 7.8, 10.4, 16.5, 19.1, 20.7, 21.3, 22.9)
plot.new(); plot.window(xlim = c(-0.5, 25.5), ylim = c(-0.3, 5.2), asp = 1)
title("Carrés des écarts, d'aire (x - moyenne) au carré", cex.main = 0.9, font.main = 1)
rect(pos, 0, pos + cotes, cotes, col = adjustcolor("#1F4E79", 0.25), border = "#1F4E79")
k <- cotes > 0; text(pos[k] + cotes[k] / 2, cotes[k] / 2, cotes[k]^2, cex = 0.8)
text(c(7.2, 20.7), c(4.4, 4.4), c("A", "B"), font = 2)
```

```
clear
input a b
  2 4
  4 5
  6 6
  8 7
10 8
end
generate ea = a - 6           // écarts à la moyenne de A
generate ca = (a - 6)^2       // carrés des écarts de A
generate cb = (b - 6)^2       // carrés des écarts de B
tabstat ea ca cb, statistics(sum) // somme des écarts nulle
tabstat a b, statistics(variance) // variances 10 et 2,5
generate ha = 2.7 + 0.3*_n
generate hb = 0.3 + 0.3*_n
generate m = 6
twoway (pcspike ha m ha a, lcolor(gs9)) (scatter ha a, mcolor("31 78 121")) ///
      (pcspike hb m hb b, lcolor(gs9)) (scatter hb b, mcolor("31 78 121")), ///
      xline(6, lpattern(dash) lcolor("192 80 43")) text(3.6 0.5 "A" 1.2 0.5 "B") ///
      ylabel(none) ytitle("") xtitle("") legend(off) title("Écarts à la moyenne") name(G, replace)
graph bar (asis) ca cb, over(a) legend(order(1 "A" 2 "B")) ///
      title("Carrés des écarts") name(D, replace)
graph combine G D, rows(1)
graph export "fig_variance_stata.pdf", replace
```

```
import subprocess, sys
subprocess.run([sys.executable, "-m", "pip", "install", "-q",
               "--disable-pip-version-check", "wooldridge"]) # installe si absent
import wooldridge as wd
import matplotlib.pyplot as plt
econ = wd.data("econmath")
fig, ax = plt.subplots(figsize=(5.2, 3.4))
ax.scatter(econ["colgpa"], econ["score"], s=8, color="#1F4E79", alpha=0.35)
ax.axvline(econ["colgpa"].mean(), color="#8C8C8C", ls="--", lw=1)
ax.axhline(econ["score"].mean(), color="#8C8C8C", ls="--", lw=1)
ax.text(3.8, 97, "+", fontsize=18, color="#C0502B", ha="center", va="center")
ax.text(1.1, 25, "+", fontsize=18, color="#C0502B", ha="center", va="center")
ax.text(3.8, 25, "\u2212", fontsize=18, color="#C0502B", ha="center", va="center")
ax.text(1.1, 97, "\u2212", fontsize=18, color="#C0502B", ha="center", va="center")
ax.set_xlabel("Moyenne universitaire (colgpa, sur 4)")
ax.set_ylabel("Note finale (score, %)")
ax.spines[["top", "right"]].set_visible(False)
ax.xaxis.set_major_formatter(lambda v, p: f"{v:g}".replace(".", ","))
fig.tight_layout()
fig.savefig("fig_quadrants.pdf")
plt.show()
print("covariance :", round(econ["colgpa"].cov(econ["score"]), 3))
print("corr lation :", round(econ["colgpa"].corr(econ["score"]), 3))
```

```
if (!requireNamespace("wooldridge", quietly = TRUE)) install.packages("wooldridge")
library(wooldridge)
data("econmath")
plot(econmath$colgpa, econmath$score, pch = 16, cex = 0.5,
     col = adjustcolor("#1F4E79", 0.35), xaxt = "n",
     xlab = "Moyenne universitaire (colgpa, sur 4)", ylab = "Note finale (score, %)")
axis(1, at = axTicks(1), labels = sub(".", ",", axTicks(1), fixed = TRUE))
abline(v = mean(econmath$colgpa), h = mean(econmath$score), col = "#8C8C8C", lty = 2)
text(c(3.8, 1.1, 3.8, 1.1), c(97, 25, 25, 97), c("+", "+", "\u2212", "\u2212"),
     col = "#C0502B", cex = 1.6)
round(cov(econmath$colgpa, econmath$score), 3)
round(cor(econmath$colgpa, econmath$score), 3)
```

```
use econmath.dta, clear           // fichier du cours, dossier de travail
quietly summarize colgpa
local mx = r(mean)
quietly summarize score
local my = r(mean)
twoway scatter score colgpa, msize(vsmall) mcolor("31 78 121%35") ///
    xline(`mx', lpattern(dash) lcolor(gs9)) yline(`my', lpattern(dash) lcolor(gs9)) ///
    text(97 3.8 "+" 25 1.1 "+" 25 3.8 "-" 97 1.1 "-", color("192 80 43") size(large)) ///
    xtitle("Moyenne universitaire (colgpa, sur 4)") ytitle("Note finale (score, %)")
graph export "fig_quadrants_stata.pdf", replace
correlate colgpa score, covariance
correlate colgpa score
```

```
import numpy as np
import matplotlib.pyplot as plt
x = np.arange(1, 11)                                # x = 1, 2, ..., 10
yA = np.array([2, 4, 3, 5, 4, 6, 5, 7, 6, 8])        # nuage de départ
yB = np.array([7, 4, 3, 5, 4, 6, 5, 7, 6, 3])        # on déplace 2 points
yC = np.array([7, 8, 6, 5, 4, 6, 5, 4, 2, 3])        # on en déplace 4 autres
rA = np.corrcoef(x, yA)[0, 1]; rB = np.corrcoef(x, yB)[0, 1]; rC = np.corrcoef(x, yC)[0, 1]
print("r :", round(rA, 3), round(rB, 3) + 0, round(rC, 3))
print("s_xy :", np.cov(x, yA)[0, 1].round(3), np.cov(x, yB)[0, 1].round(3) + 0, np.cov(x, yC)[0, 1].round(3))
fig, axs = plt.subplots(1, 3, figsize=(9, 3), sharey=True)
axs[1].scatter(x[[0, 9]], yA[[0, 9]], color="#C8C8C8", s=40)                # anciennes places
axs[2].scatter(x[[1, 2, 7, 8]], yB[[1, 2, 7, 8]], color="#C8C8C8", s=40)
axs[0].scatter(x, yA, color="#1F4E79", s=40); axs[1].scatter(x, yB, color="#1F4E79", s=40)
axs[2].scatter(x, yC, color="#1F4E79", s=40)
[a.set_title(f"r = {round(r, 2) + 0:.2f}".replace(".", ",").replace("-", "\u2212")) for a, r in zip(axs, [rA, rB, rC])]
[a.axvline(5.5, color="#8C8C8C", ls="--", lw=1) for a in axs]                # moyenne de x
[a.axhline(5, color="#8C8C8C", ls="--", lw=1) for a in axs]                # moyenne de y
[a.set_xlabel("x") for a in axs]; axs[0].set_ylabel("y")
[a.spines[["top", "right"]].set_visible(False) for a in axs]
fig.tight_layout(); fig.savefig("fig_deplacer.pdf"); plt.show()
```

```
x <- 1:10
yA <- c(2, 4, 3, 5, 4, 6, 5, 7, 6, 8)      # nuage de départ
yB <- c(7, 4, 3, 5, 4, 6, 5, 7, 6, 3)      # on déplace 2 points
yC <- c(7, 8, 6, 5, 4, 6, 5, 4, 2, 3)      # on en déplace 4 autres
round(c(cor(x, yA), cor(x, yB), cor(x, yC)), 3)
round(c(cov(x, yA), cov(x, yB), cov(x, yC)), 3)
par(mfrow = c(1, 3))
plot(x, yA, pch = 16, col = "#1F4E79", ylim = c(2, 8), ylab = "y",
      main = paste("r =", sub(".", ",", format(round(cor(x, yA), 2), nsmall = 2), fixed = TRUE)))
abline(v = 5.5, h = 5, col = "#8C8C8C", lty = 2)
plot(x, yB, pch = 16, col = "#1F4E79", ylim = c(2, 8), ylab = "y",
      main = paste("r =", sub(".", ",", format(round(abs(cor(x, yB)), 2), nsmall = 2), fixed = TRUE)))
points(x[c(1, 10)], yA[c(1, 10)], pch = 16, col = "#C8C8C8") # ancienne place
abline(v = 5.5, h = 5, col = "#8C8C8C", lty = 2)
plot(x, yC, pch = 16, col = "#1F4E79", ylim = c(2, 8), ylab = "y",
      main = paste("r =", sub(".", ",", format(round(cor(x, yC), 2), nsmall = 2), fixed = TRUE)))
points(x[c(2, 3, 8, 9)], yB[c(2, 3, 8, 9)], pch = 16, col = "#C8C8C8")
abline(v = 5.5, h = 5, col = "#8C8C8C", lty = 2)
```

```
clear
input x yA yB yC
 1 2 7 7
 2 4 4 8
 3 3 3 6
 4 5 5 5
 5 4 4 4
 6 6 6 6
 7 5 5 5
 8 7 7 4
 9 6 6 2
10 8 3 3
end

correlate x yA yB yC           // 1re colonne, r entre x et chaque y
correlate x yA yB yC, covariance // 1re colonne, covariances

twoway (scatter yA x, mcolor("31 78 121")), xline(5.5, lpattern(dash)) yline(5, lpattern(dash)) ///
      title("Départ") legend(off) name(A, replace)

twoway (scatter yA x if inlist(x, 1, 10), mcolor(gs12)) (scatter yB x, mcolor("31 78 121")), ///
      xline(5.5, lpattern(dash)) yline(5, lpattern(dash)) title("2 points déplacés") legend(off) name(B, replace)

twoway (scatter yB x if inlist(x, 2, 3, 8, 9), mcolor(gs12)) (scatter yC x, mcolor("31 78 121")), ///
      xline(5.5, lpattern(dash)) yline(5, lpattern(dash)) title("4 autres déplacés") legend(off) name(C, replace)

graph combine A B C, rows(1) ycommon
graph export "fig_deplacer_stata.pdf", replace
```

```
import numpy as np
x = np.array([2, 4, 6, 8, 10])      # heures de travail personnel par semaine
y = np.array([7, 10, 11, 14, 13])  # note sur 20
print("moyennes :", x.mean(), y.mean())
print("variances :", x.var(ddof=1), y.var(ddof=1))  # ddof=1, division par n-1
print("covariance :", np.cov(x, y, ddof=1)[0, 1])
print("corrélation :", round(np.corrcoef(x, y)[0, 1], 4))
```

Code R, exemple jouet, statistiques descriptives

[Python](#)**R**[Stata](#)

```
x <- c(2, 4, 6, 8, 10)      # heures de travail personnel par semaine
y <- c(7, 10, 11, 14, 13)   # note sur 20
c(mean(x), mean(y))
c(var(x), var(y))          # var() divise par n - 1
cov(x, y)
round(cor(x, y), 4)
```

[◀ Retour](#)

```
clear
input x y
 2  7
 4 10
 6 11
 8 14
10 13
end
summarize x y           // écarts-types calculés avec  $n - 1$ 
correlate x y, covariance
correlate x y
```

```
import subprocess, sys
subprocess.run([sys.executable, "-m", "pip", "install", "-q",
               "--disable-pip-version-check", "wooldridge"])    # installe si absent
import wooldridge as wd
import matplotlib.pyplot as plt
econ = wd.data("econmath")
fig, ax = plt.subplots(figsize=(5.2, 3.4))
ax.scatter(econ["colgpa"], econ["score"], s=8, color="#1F4E79", alpha=0.35)
ax.set_xlabel("Moyenne universitaire (colgpa, sur 4)")
ax.set_ylabel("Note finale (score, %)")
ax.spines[["top", "right"]].set_visible(False)
ax.xaxis.set_major_formatter(lambda v, p: f"{v:g}".replace(".", ","))
fig.tight_layout()
fig.savefig("fig_nuage.pdf")
plt.show()
```

```
if (!requireNamespace("wooldridge", quietly = TRUE)) install.packages("wooldridge")
library(wooldridge)
data("econmath")
plot(econmath$colgpa, econmath$score, pch = 16, cex = 0.5,
     col = adjustcolor("#1F4E79", 0.35), xaxt = "n",
     xlab = "Moyenne universitaire (colgpa, sur 4)", ylab = "Note finale (score, %)")
axis(1, at = axTicks(1), labels = sub(".", ",", axTicks(1), fixed = TRUE))
```

Code Stata, nuage de points du fil rouge

[Python](#)[R](#)[Stata](#)

```
use econmath.dta, clear           // fichier du cours, dossier de travail
twoway scatter score colgpa, msize(vsmall) mcolor("31 78 121%35") ///
    xtitle("Moyenne universitaire (colgpa, sur 4)") ytitle("Note finale (score, %)")
graph export "fig_nuage_stata.pdf", replace
```

[◀ Retour](#)

```
import numpy as np
import matplotlib.pyplot as plt
x = np.array([2, 4, 6, 8, 10]); y = np.array([7, 10, 11, 14, 13])
g = np.linspace(0, 12, 50)                                # grille pour tracer
scr_1 = ((y - (4 + 1.2 * x))**2).sum()                      # droite 1,  $4 + 1,2 x$ 
scr_2 = ((y - (6.2 + 0.8 * x))**2).sum()                    # droite 2,  $6,2 + 0,8 x$ 
print("SCR :", round(scr_1, 2), round(scr_2, 2))
fig, axs = plt.subplots(1, 2, figsize=(6.4, 3), sharey=True)
axs[0].plot(g, 4 + 1.2 * g, color="#8C8C8C"); axs[0].vlines(x, y, 4 + 1.2 * x, color="#8C8C8C", lw=1)
axs[1].plot(g, 6.2 + 0.8 * g, color="#C0502B"); axs[1].vlines(x, y, 6.2 + 0.8 * x, color="#8C8C8C", lw=1)
axs[0].set_title(f"y = 4 + 1,2 x\nSCR = {scr_1:.1f}".replace(".", ","), fontsize=10)
axs[1].set_title(f"y = 6,2 + 0,8 x\nSCR = {scr_2:.1f}".replace(".", ","), fontsize=10)
[a.scatter(x, y, color="#1F4E79", zorder=3) for a in axs]
[a.set_xlim(0, 12) for a in axs]; [a.set_ylim(4, 18) for a in axs]
[a.set_xlabel("x (heures)") for a in axs]; axs[0].set_ylabel("y (note sur 20)")
[a.spines[["top", "right"]].set_visible(False) for a in axs]
[a.yaxis.set_major_formatter(lambda v, p: f"{v:g}".replace(".", ",")) for a in axs]
fig.tight_layout(); fig.savefig("fig_mco_illus.pdf"); plt.show()
```

```
x <- c(2, 4, 6, 8, 10)
y <- c(7, 10, 11, 14, 13)
scr_1 <- sum((y - (4 + 1.2 * x))^2)           # droite 1 :  $4 + 1,2 x$ 
scr_2 <- sum((y - (6.2 + 0.8 * x))^2)       # droite 2 :  $6,2 + 0,8 x$ 
round(c(scr_1, scr_2), 2)
par(mfrow = c(1, 2))
plot(x, y, pch = 16, col = "#1F4E79", xlim = c(0, 12), ylim = c(4, 18),
     xlab = "x (heures)", ylab = "y (note sur 20)",
     main = paste0("y = 4 + 1,2 x\nSCR = ", sub(".", ", ", format(round(scr_1, 1), nsmall = 1), fixed = TRUE)))
abline(a = 4, b = 1.2, col = "#8C8C8C", lwd = 1.5)
segments(x, y, x, 4 + 1.2 * x, col = "#8C8C8C") # résidus
plot(x, y, pch = 16, col = "#1F4E79", xlim = c(0, 12), ylim = c(4, 18),
     xlab = "x (heures)", ylab = "y (note sur 20)",
     main = paste0("y = 6,2 + 0,8 x\nSCR = ", sub(".", ", ", format(round(scr_2, 1), nsmall = 1), fixed = TRUE)))
abline(a = 6.2, b = 0.8, col = "#C0502B", lwd = 1.5)
segments(x, y, x, 6.2 + 0.8 * x, col = "#8C8C8C")
```

```
clear
input x y
 2 7
 4 10
 6 11
 8 14
10 13
end
generate f1 = 4 + 1.2*x           // droite 1
generate f2 = 6.2 + 0.8*x       // droite 2
generate c1 = (y - f1)^2
generate c2 = (y - f2)^2
tabstat c1 c2, statistics(sum)   // SCR 11 et 4,4
twoway (rspike y f1 x, lcolor(gs9)) (function y = 4 + 1.2*x, range(0 12) lcolor(gs9)) ///
      (scatter y x, mcolor("31 78 121")), title("y = 4 + 1,2 x   SCR = 11,0") legend(off) ///
      xtitle("x (heures)") ytitle("y (note sur 20)") name(A, replace)
twoway (rspike y f2 x, lcolor(gs9)) (function y = 6.2 + 0.8*x, range(0 12) lcolor("192 80 43")) ///
      (scatter y x, mcolor("31 78 121")), title("y = 6,2 + 0,8 x   SCR = 4,4") legend(off) ///
      xtitle("x (heures)") ytitle("") name(B, replace)
graph combine A B, rows(1) ycommon
graph export "fig_mco_illus_stata.pdf", replace
```

```
import numpy as np
import matplotlib.pyplot as plt
x = np.array([2, 4, 6, 8, 10]); y = np.array([7, 10, 11, 14, 13])
g = np.linspace(0, 12, 50)
u_a = y - (17 - x)                                # a) droite fausse, pente -1
u_b1 = y - (5 + x); u_b2 = y - (8 + 0.5 * x)      # b) deux droites
u_c = y - (6.2 + 0.8 * x)                          # c) droite des MCO
print("a) somme des résidus :", u_a.sum(), " MCO :", round(u_c.sum(), 10) + 0)
print("b) sommes des valeurs absolues :", np.abs(u_b1).sum(), np.abs(u_b2).sum())
print("c) somme des carrés :", round((u_c**2).sum(), 2))
fig, axs = plt.subplots(1, 3, figsize=(8, 2.6))
axs[0].plot(g, 17 - g, color="#8C8C8C"); axs[0].vlines(x, y, 17 - x, color="#8C8C8C", lw=1)
axs[1].plot(g, 5 + g, color="#8C8C8C"); axs[1].plot(g, 8 + 0.5 * g, color="#8C8C8C", ls="--")
axs[2].plot(g, 6.2 + 0.8 * g, color="#C0502B")
[axs[2].add_patch(plt.Rectangle((xi, min(yi, yi - ui)), abs(ui), abs(ui),
    fc="#1F4E79", alpha=0.25)) for xi, yi, ui in zip(x, y, u_c)] # carrés des résidus
axs[0].set_title("a) y = 17 - x \nsomme des résidus = 0")
axs[1].set_title("b) y = 5 + x et y = 8 + 0,5 x \nsomme des |résidus| = 4 et 4")
axs[2].set_title("c) y = 6,2 + 0,8 x \nsomme des carrés = 4,4")
[a.scatter(x, y, color="#1F4E79", zorder=3) for a in axs]
[a.set_xlim(0, 12) for a in axs]; [a.set_ylim(5, 16) for a in axs]
[a.set_aspect("equal") for a in axs] # carrés non déformés
[a.set_xlabel("x (heures)") for a in axs]; axs[0].set_ylabel("y (note sur 20)")
[a.spines[["top", "right"]].set_visible(False) for a in axs]
[a.yaxis.set_major_formatter(lambda v, p: f"{v:g}".replace(".", ",")) for a in axs]
fig.tight_layout(); fig.savefig("fig_carres.pdf"); plt.show()
```

```
x <- c(2, 4, 6, 8, 10)
y <- c(7, 10, 11, 14, 13)
u_a <- y - (17 - x)           # droite fausse, pente -1
u_b1 <- y - (5 + x); u_b2 <- y - (8 + 0.5 * x) # deux droites
u_c <- y - (6.2 + 0.8 * x)     # droite des MCO
c(sum(u_a), round(sum(u_c), 10)) # a) sommes des résidus
c(sum(abs(u_b1)), sum(abs(u_b2))) # b) sommes des valeurs absolues
round(sum(u_c^2), 2)          # c) somme des carrés
par(mfrow = c(1, 3))
plot(x, y, pch = 16, col = "#1F4E79", xlim = c(0, 12), ylim = c(5, 16), asp = 1,
     main = "a) y = 17 - x\nsomme des résidus = 0")
abline(a = 17, b = -1, col = "#8C8C8C"); segments(x, y, x, 17 - x, col = "#8C8C8C")
plot(x, y, pch = 16, col = "#1F4E79", xlim = c(0, 12), ylim = c(5, 16), asp = 1,
     main = "b) y = 5 + x et y = 8 + 0,5 x\nsomme des |résidus| = 4 et 4")
abline(a = 5, b = 1, col = "#8C8C8C"); abline(a = 8, b = 0.5, col = "#8C8C8C", lty = 2)
plot(x, y, pch = 16, col = "#1F4E79", xlim = c(0, 12), ylim = c(5, 16), asp = 1,
     main = "c) y = 6,2 + 0,8 x\nsomme des carrés = 4,4")
abline(a = 6.2, b = 0.8, col = "#C0502B")
rect(x, pmin(y, y - u_c), x + abs(u_c), pmin(y, y - u_c) + abs(u_c),
     col = adjustcolor("#1F4E79", 0.25), border = NA)
```

```
clear
input x y
  2  7
  4 10
  6 11
  8 14
10 13
end
generate ua = y - (17 - x)           // a) droite fausse, pente -1
generate ab1 = abs(y - (5 + x))       // b) deux droites
generate ab2 = abs(y - (8 + 0.5*x))
generate uc = y - (6.2 + 0.8*x)       // c) droite des MCO
generate uc2 = uc^2
tabstat ua uc ab1 ab2 uc2, statistics(sum)
generate fa = 17 - x
generate fc = 6.2 + 0.8*x
twoway (rspike y fa x, lcolor(gs9)) (function y = 17 - x, range(0 12) lcolor(gs9)) ///
      (scatter y x, mcolor("31 78 121")), title("a) somme des résidus = 0") legend(off) name(A, replace)
twoway (function y = 5 + x, range(0 12) lcolor(gs9)) (function y = 8 + 0.5*x, range(0 12) lpattern(dash)) ///
      (scatter y x, mcolor("31 78 121")), title("b) somme des |résidus| = 4 et 4") legend(off) name(B, replace)
twoway (rspike y fc x, lcolor(gs9)) (function y = 6.2 + 0.8*x, range(0 12) lcolor("192 80 43")) ///
      (scatter y x, mcolor("31 78 121")), title("c) somme des carrés = 4,4") legend(off) name(C, replace)
graph combine A B C, rows(1) ycommon
graph export "fig_carres_stata.pdf", replace
```

```
import subprocess, sys
subprocess.run([sys.executable, "-m", "pip", "install", "-q",
               "--disable-pip-version-check", "statsmodels"])    # installe si absent
import numpy as np
import pandas as pd
import statsmodels.formula.api as smf
x = np.array([2, 4, 6, 8, 10])
y = np.array([7, 10, 11, 14, 13])
b1 = ((x - x.mean()) * (y - y.mean())).sum() / ((x - x.mean())**2).sum()
b0 = y.mean() - b1 * x.mean()
print("à la main :", round(b0, 4), round(b1, 4))
jouet = pd.DataFrame({"x": x, "y": y})
print("par le code :", smf.ols("y ~ x", data=jouet).fit().params.round(4).tolist())
```

Code R, exemple jouet, estimateurs des MCO

[Python](#)**R**[Stata](#)

```
x <- c(2, 4, 6, 8, 10)
y <- c(7, 10, 11, 14, 13)
b1 <- sum((x - mean(x)) * (y - mean(y))) / sum((x - mean(x))^2)
b0 <- mean(y) - b1 * mean(x)
c(b0, b1)                                # à la main
coef(lm(y ~ x))                          # par le code
```

[◀ Retour](#)

```
clear
input x y
  2  7
  4 10
  6 11
  8 14
10 13
end
quietly summarize x
scalar xbar = r(mean)
quietly summarize y
scalar ybar = r(mean)
generate num = (x - xbar)*(y - ybar)
generate den = (x - xbar)^2
quietly summarize num
scalar snum = r(sum)
quietly summarize den
scalar b1 = snum/r(sum)
scalar b0 = ybar - b1*xbar
display "à la main : b0 = " b0 "    b1 = " b1
regress y x                               // par le code
```

Code Python, exemple jouet, propriétés algébriques

[Python](#)[R](#)[Stata](#)

```
import numpy as np
x = np.array([2, 4, 6, 8, 10])
y = np.array([7, 10, 11, 14, 13])
b1 = ((x - x.mean()) * (y - y.mean())).sum() / ((x - x.mean())**2).sum()
b0 = y.mean() - b1 * x.mean()
yhat = b0 + b1 * x                    # valeurs prédites
uhat = y - yhat                       # résidus
print("prédites :", yhat.round(2), " résidus :", uhat.round(2))
print("somme des résidus :", round(uhat.sum(), 10))
print("somme x * résidus :", round((x * uhat).sum(), 10))
print("droite en x moyen :", round(b0 + b1 * x.mean(), 10), " y moyen :", y.mean())
```

[◀ Retour](#)

Code R, exemple jouet, propriétés algébriques

[Python](#)**R**[Stata](#)

```
x <- c(2, 4, 6, 8, 10)
y <- c(7, 10, 11, 14, 13)
b1 <- sum((x - mean(x)) * (y - mean(y))) / sum((x - mean(x))^2)
b0 <- mean(y) - b1 * mean(x)
yhat <- b0 + b1 * x           # valeurs prédites
uhat <- y - yhat              # résidus
rbind(yhat, uhat)
round(sum(uhat), 10)          # somme des résidus
round(sum(x * uhat), 10)      # somme x * résidus
c(b0 + b1 * mean(x), mean(y)) # la droite passe par le point moyen
```

[◀ Retour](#)

```
clear
input x y
  2  7
  4 10
  6 11
  8 14
10 13
end
quietly regress y x
predict yhat, xb           // valeurs prédites
predict uhat, residuals    // résidus
generate xu = x*uhat
list x y yhat uhat xu, clean
tabstat uhat xu, statistics(sum) // sommes nulles
quietly summarize x
display "droite en x moyen : " _b[_cons] + _b[x]*r(mean)
```

```
import numpy as np
import matplotlib.pyplot as plt
x = np.array([2, 4, 6, 8, 10])
y = np.array([7, 10, 11, 14, 13])
b1 = ((x - x.mean()) * (y - y.mean())).sum() / ((x - x.mean())**2).sum()
b0 = y.mean() - b1 * x.mean()
yhat = b0 + b1 * x
g = np.linspace(0, 12, 50)
fig, ax = plt.subplots(figsize=(4.6, 3.2))
ax.scatter(x, y, color="#1F4E79", zorder=3)
ax.plot(g, b0 + b1 * g, color="#C0502B")
ax.axhline(y.mean(), color="#8C8C8C", ls="--", lw=1)
ax.annotate("", xy=(8.25, yhat[3]), xytext=(8.25, y.mean()),
            arrowprops=dict(arrowstyle="<->", color="#C0502B")) # 4e étudiant, x = 8
ax.annotate("", xy=(7.75, y[3]), xytext=(7.75, yhat[3]),
            arrowprops=dict(arrowstyle="<->", color="#8C8C8C"))
ax.text(8.45, (yhat[3] + y.mean()) / 2, r"$\hat{y}_i - \bar{y}$", color="#C0502B")
ax.text(7.55, (y[3] + yhat[3]) / 2, r"$\hat{u}_i$", color="#8C8C8C", ha="right")
ax.text(0.3, y.mean() + 0.3, r"$\bar{y}$", color="#8C8C8C")
ax.set_xlim(0, 12); ax.set_ylim(4, 18)
ax.set_xlabel("x (heures)"); ax.set_ylabel("y (note sur 20)")
ax.spines[["top", "right"]].set_visible(False)
fig.tight_layout()
fig.savefig("fig_decomp.pdf")
plt.show()
```

```
x <- c(2, 4, 6, 8, 10)
y <- c(7, 10, 11, 14, 13)
b1 <- sum((x - mean(x)) * (y - mean(y))) / sum((x - mean(x))^2)
b0 <- mean(y) - b1 * mean(x)
yhat <- b0 + b1 * x
plot(x, y, pch = 16, col = "#1F4E79", xlim = c(0, 12), ylim = c(4, 18),
      xlab = "x (heures)", ylab = "y (note sur 20)")
abline(a = b0, b = b1, col = "#C0502B", lwd = 1.5)
abline(h = mean(y), col = "#8C8C8C", lty = 2)
arrows(8.25, mean(y), 8.25, yhat[4], code = 3, length = 0.05, col = "#C0502B") # 4e étudiant
arrows(7.75, yhat[4], 7.75, y[4], code = 3, length = 0.05, col = "#8C8C8C")
text(8.45, (yhat[4] + mean(y)) / 2, expression(hat(y)[i] - bar(y)), col = "#C0502B", adj = 0)
text(7.55, (y[4] + yhat[4]) / 2, expression(hat(u)[i]), col = "#8C8C8C", adj = 1)
```

```
clear
input x y
 2  7
 4 10
 6 11
 8 14
10 13
end
quietly regress y x
predict yhat, xb
list x y yhat in 4           // 4e étudiant, x = 8
twoway (scatter y x, mcolor("31 78 121")) (function y = 6.2 + 0.8*x, range(0 12) lcolor("192 80 43")) ///
      (pcarrowi 11 8.25 12.6 8.25, lcolor("192 80 43") mcolor("192 80 43")) ///
      (pcarrowi 12.6 7.75 14 7.75, lcolor(gs9) mcolor(gs9)), ///
      yline(11, lpattern(dash) lcolor(gs9)) legend(off) xtitle("x (heures)") ytitle("y (note sur 20)")
graph export "fig_decomp_stata.pdf", replace
```

```
import numpy as np
x = np.array([2, 4, 6, 8, 10])
y = np.array([7, 10, 11, 14, 13])
b1 = ((x - x.mean()) * (y - y.mean())).sum() / ((x - x.mean())**2).sum()
b0 = y.mean() - b1 * x.mean()
yhat = b0 + b1 * x
sct = ((y - y.mean())**2).sum()
sce = ((yhat - y.mean())**2).sum()
scr = ((y - yhat)**2).sum()
print("SCT, SCE, SCR :", sct, round(sce, 4), round(scr, 4))
print("R2 :", round(sce / sct, 4), " r2 :", round(np.corrcoef(x, y)[0, 1]**2, 4))
```

Code R, exemple jouet, SCT, SCE, SCR et R2

[Python](#)**R**[Stata](#)

```
x <- c(2, 4, 6, 8, 10)
y <- c(7, 10, 11, 14, 13)
b1 <- sum((x - mean(x)) * (y - mean(y))) / sum((x - mean(x))^2)
b0 <- mean(y) - b1 * mean(x)
yhat <- b0 + b1 * x
SCT <- sum((y - mean(y))^2); SCE <- sum((yhat - mean(y))^2); SCR <- sum((y - yhat)^2)
c(SCT, SCE, SCR)
round(c(R2 = SCE / SCT, r2 = cor(x, y)^2), 4)
```

[◀ Retour](#)

```
clear
input x y
 2  7
 4 10
 6 11
 8 14
10 13
end
quietly regress y x
display "SCT = " e(mss) + e(rss) "    SCE = " e(mss) "    SCR = " e(rss) "    R2 = " e(r2)
quietly correlate x y
display "r2 = " r(rho)^2
```

```
import subprocess, sys
subprocess.run([sys.executable, "-m", "pip", "install", "-q",
               "--disable-pip-version-check", "wooldridge", "statsmodels"])
import wooldridge as wd
import statsmodels.formula.api as smf
jt = wd.data("jtrain2")
m_jt = smf.ols("re78 ~ train", data=jt).fit()
print(m_jt.params.round(2).tolist(), " R2 :", round(m_jt.rsquared, 3))
```

Code R, un R2 faible dans une expérience

[Python](#)**R**[Stata](#)

```
if (!requireNamespace("wooldridge", quietly = TRUE)) install.packages("wooldridge")
library(wooldridge)
data("jtrain2")
m_jt <- lm(re78 ~ train, data = jtrain2)
round(coef(m_jt), 2); round(summary(m_jt)$r.squared, 3)
```

[◀ Retour](#)

Code Stata, un R2 faible dans une expérience

Python

R

Stata

```
use jtrain2.dta, clear  
regress re78 train
```

// fichier du cours, dossier de travail

[◀ Retour](#)

```
import subprocess, sys
subprocess.run([sys.executable, "-m", "pip", "install", "-q",
               "--disable-pip-version-check", "wooldridge", "statsmodels"])
import wooldridge as wd
import statsmodels.formula.api as smf
econ = wd.data("econmath")
m1 = smf.ols("score ~ colgpa", data=econ).fit()
m2 = smf.ols("score ~ actmth", data=econ).fit()      # 42 valeurs manquantes écartées
print(m1.params.round(3).tolist(), m1.bse.round(3).tolist(), round(m1.rsquared, 3), int(m1.nobs))
print(m2.params.round(3).tolist(), m2.bse.round(3).tolist(), round(m2.rsquared, 3), int(m2.nobs))
print("p-valeurs (étoiles) :", m1.pvalues.tolist(), m2.pvalues.tolist())
print("prédiction pour colgpa = 3 :", round(m1.params["Intercept"] + 3 * m1.params["colgpa"], 2))
print("colgpa min et max :", econ["colgpa"].min(), econ["colgpa"].max())
```

```
if (!requireNamespace("wooldridge", quietly = TRUE)) install.packages("wooldridge")
library(wooldridge)
data("econmath")
m1 <- lm(score ~ colgpa, data = econmath)
m2 <- lm(score ~ actmth, data = econmath)           # 42 valeurs manquantes écartées
summary(m1)$coefficients[, c(1, 2, 4)]; round(summary(m1)$r.squared, 3); nobs(m1)      # 4e colonne : p-valeurs
summary(m2)$coefficients[, c(1, 2, 4)]; round(summary(m2)$r.squared, 3); nobs(m2)
round(predict(m1, newdata = data.frame(colgpa = 3)), 2)
range(econmath$colgpa)
```

```
use econmath.dta, clear           // fichier du cours, dossier de travail
regress score colgpa
display "prédiction pour colgpa = 3 : " _b[_cons] + 3*_b[colgpa]
regress score actmth             // 42 valeurs manquantes écartées
summarize colgpa
```

```
import subprocess, sys
subprocess.run([sys.executable, "-m", "pip", "install", "-q",
               "--disable-pip-version-check", "wooldridge", "statsmodels"])
import wooldridge as wd
import numpy as np
import matplotlib.pyplot as plt
import statsmodels.formula.api as smf
econ = wd.data("econmath")
m1 = smf.ols("score ~ colgpa", data=econ).fit()
g = np.linspace(econ["colgpa"].min(), econ["colgpa"].max(), 50)
fig, ax = plt.subplots(figsize=(5.2, 3.4))
ax.scatter(econ["colgpa"], econ["score"], s=8, color="#1F4E79", alpha=0.35)
ax.plot(g, m1.params["Intercept"] + m1.params["colgpa"] * g, color="#C0502B", lw=2)
ax.scatter(econ["colgpa"].mean(), econ["score"].mean(), color="#C0502B", marker="s", s=40, zorder=4)
ax.set_xlabel("Moyenne universitaire (colgpa, sur 4)")
ax.set_ylabel("Note finale (score, %)")
ax.spines[["top", "right"]].set_visible(False)
ax.xaxis.set_major_formatter(lambda v, p: f"{v:g}".replace(".", ","))
fig.tight_layout()
fig.savefig("fig_droite.pdf")
plt.show()
print("point moyen :", round(econ["colgpa"].mean(), 3), round(econ["score"].mean(), 3))
```

```
if (!requireNamespace("wooldridge", quietly = TRUE)) install.packages("wooldridge")
library(wooldridge)
data("econmath")
m1 <- lm(score ~ colgpa, data = econmath)
plot(econmath$colgpa, econmath$score, pch = 16, cex = 0.5,
     col = adjustcolor("#1F4E79", 0.35), xaxt = "n",
     xlab = "Moyenne universitaire (colgpa, sur 4)", ylab = "Note finale (score, %)")
axis(1, at = axTicks(1), labels = sub(".", ",", axTicks(1), fixed = TRUE))
abline(m1, col = "#C0502B", lwd = 2)
points(mean(econmath$colgpa), mean(econmath$score), pch = 15, col = "#C0502B", cex = 1.2)
round(c(mean(econmath$colgpa), mean(econmath$score)), 3)
```

```
use econmath.dta, clear           // fichier du cours, dossier de travail
quietly summarize colgpa
local mx = r(mean)
quietly summarize score
local my = r(mean)
display "point moyen : " `mx' " " `my'
twoway (scatter score colgpa, msize(vsmall) mcolor("31 78 121%35")) ///
      (lfit score colgpa, lcolor("192 80 43") lwidth(medthick)) ///
      (scatteri `my' `mx', msymbol(S) mcolor("192 80 43")), legend(off) ///
      xtitle("Moyenne universitaire (colgpa, sur 4)") ytitle("Note finale (score, %)")
graph export "fig_droite_stata.pdf", replace
```

```
import subprocess, sys
subprocess.run([sys.executable, "-m", "pip", "install", "-q",
               "--disable-pip-version-check", "wooldridge", "statsmodels"])
import wooldridge as wd
import statsmodels.formula.api as smf
econ = wd.data("econmath")
m1 = smf.ols("score ~ colgpa", data=econ).fit()
econ["score_chap"] = m1.fittedvalues           # valeurs prédites
econ["u_chap"] = m1.resid                     # résidus
print(econ[["colgpa", "score", "score_chap", "u_chap"]].head(5).round(2))
print("somme des résidus :", round(m1.resid.sum(), 8))
sct = ((econ["score"] - econ["score"].mean())**2).sum()
print("SCT, SCE, SCR :", round(sct, 1), round(m1.ess, 1), round(m1.ssr, 1))
```

```
if (!requireNamespace("wooldridge", quietly = TRUE)) install.packages("wooldridge")
library(wooldridge)
data("econmath")
m1 <- lm(score ~ colgpa, data = econmath)
econmath$score_chap <- fitted(m1)           # valeurs prédites
econmath$u_chap <- resid(m1)               # résidus
round(head(econmath[, c("colgpa", "score", "score_chap", "u_chap")], 5), 2)
round(sum(resid(m1)), 8)
SCT <- sum((econmath$score - mean(econmath$score))^2)
round(c(SCT = SCT, SCE = sum((fitted(m1) - mean(econmath$score))^2), SCR = sum(resid(m1)^2)), 1)
```

Code Stata, valeurs prédites et résidus

[Python](#)[R](#)[Stata](#)

```
use econmath.dta, clear           // fichier du cours, dossier de travail
quietly regress score colgpa
predict score_chap, xb           // valeurs prédites
predict u_chap, residuals        // résidus
list colgpa score score_chap u_chap in 1/5, clean
tabstat u_chap, statistics(sum)
display "SCT = " e(mss) + e(rss) "   SCE = " e(mss) "   SCR = " e(rss)
```

[◀ Retour](#)