

Économétrie

Séance 1 — Série d'exercices

Licence 3 Économie · Année universitaire 2026–2027

Les exercices portent sur les notions de la séance 1. Les calculs se font à la main, avec une calculatrice, sauf à l'exercice 5.

	Contenu	Durée estimée	Organisation
Quiz	10 questions	10 min	TD
Exercice 1	calcul à la main des MCO et du R^2	20 min	TD
Exercice 2	démonstration	15 min	travail personnel
Exercice 3	lecture d'une table de régression	15 min	TD
Exercice 4	raisonnement causal	15 min	TD
Exercice 5	données réelles et code	25 min	travail personnel
Total		100 min	60 min en TD

Quiz

Une seule réponse est correcte par question.

Question 1. On estime $y_i = \beta_0 + \beta_1 x_i + u_i$ par les MCO. Quelle grandeur peut-on calculer à partir des données ?

- a) u_i , l'erreur de l'observation i
- b) $\hat{u}_i$, le résidu de l'observation i
- c) β_1 , la pente de la population
- d) u_i et $\hat{u}_i$, qui sont égaux

Question 2. Dans un échantillon d'étudiants, la corrélation entre les heures de tutorat suivies et la note finale vaut 0,8. Que peut-on conclure ?

- a) Le tutorat augmente la note de 80 %.
- b) 80 % de la dispersion des notes est expliquée par le tutorat.
- c) Les deux variables sont fortement liées linéairement, ce qui ne prouve pas un effet du tutorat.
- d) Une heure de tutorat en plus augmente la note de 0,8 point.

Question 3. Dans une expérience où la participation à une formation a été tirée au sort, la régression du revenu sur l'indicatrice de formation donne $R^2 = 0,02$.

- a) La formation n'a pas d'effet sur le revenu.
- b) Le tirage au sort a échoué.
- c) La pente estimée est biaisée parce que le R^2 est faible.
- d) La formation explique 2 % de la dispersion des revenus, ce qui n'empêche pas la pente d'estimer son effet moyen.

Question 4. Dans un échantillon, $r_{xy} = 0$. Alors

- a) il n'y a pas de liaison linéaire entre x et y dans l'échantillon
- b) x et y n'ont aucun lien
- c) x et y sont statistiquement indépendants
- d) la pente des MCO n'est pas définie

Question 5. On exprime y en milliers d'euros au lieu d'euros, sans changer x .

- a) s_{xy} est divisée par 1 000 et r_{xy} ne change pas.
- b) s_{xy} et r_{xy} sont divisées par 1 000.
- c) s_{xy} et r_{xy} ne changent pas.
- d) s_{xy} ne change pas et r_{xy} est divisée par 1 000.

Question 6. Un institut interroge 3 000 ménages en 2015 et 3 000 autres ménages en 2022, puis réunit les réponses dans un seul fichier. Ce sont

- a) des données de panel
- b) une série temporelle
- c) des coupes transversales empilées
- d) une coupe transversale unique

Question 7. Dans une régression avec constante estimée par les MCO, l'égalité $\sum_{i=1}^n \hat{u}_i = 0$ est vraie

- a) seulement si le modèle est vrai dans la population
- b) seulement dans les grands échantillons
- c) seulement si les erreurs u_i somment aussi à zéro
- d) pour tout échantillon

Question 8. Une régression du salaire horaire (en euros) sur l'âge (en années), estimée sur des salariés de 25 à 60 ans, donne une constante $\hat{\beta}_0 = -3,2$.

- a) Le modèle prédit un salaire négatif à la naissance, il est donc faux.
- b) C'est le salaire prédit pour un âge de 0, valeur très éloignée des données et sans interprétation utile.
- c) C'est la moyenne des erreurs du modèle.
- d) Une constante négative prouve une erreur de calcul.

Question 9. Pourquoi la variance empirique s_x^2 divise-t-elle la somme des carrés des écarts par $n - 1$?

- a) Pour obtenir une valeur plus grande, donc plus prudente.
- b) Parce que la première observation sert de référence et ne compte pas.
- c) Parce que les données sont plus proches de $\bar{x}$ que de la moyenne de la population, et que diviser par $n - 1$ compense exactement cet écart en moyenne.

- d) Pour que la variance s'exprime dans l'unité de x .

Question 10. Paul a suivi une formation. Son revenu est $y_{\text{Paul}}(1)$ s'il la suit, $y_{\text{Paul}}(0)$ sinon. L'effet causal de la formation pour Paul est

- a) $y_{\text{Paul}}(1) - y_{\text{Paul}}(0)$
- b) son revenu moins le revenu moyen des personnes non formées
- c) l'écart de revenu moyen entre personnes formées et non formées
- d) la pente de la régression du revenu sur l'indicatrice de formation

Exercice 1. Calculs à la main

Six étals d'un marché vendent des paniers de fraises. On relève pour chacun le prix du panier x (en euros) et le nombre de paniers vendus en une heure y .

Étal	1	2	3	4	5	6
x_i (euros)	2	3	4	6	7	8
y_i (paniers)	14	12	13	9	8	4

1. Calculer $\bar{x}$ et $\bar{y}$, puis, à l'aide d'un tableau des écarts à la moyenne, s_x^2 , s_y^2 , s_{xy} et r_{xy} .
2. Calculer $\hat{\beta}_1$ et $\hat{\beta}_0$ dans la régression de y sur x . Interpréter $\hat{\beta}_1$ en une phrase, avec les unités.
3. Calculer les valeurs prédites $\hat{y}_i$ et les résidus $\hat{u}_i$. Vérifier les trois propriétés algébriques des MCO.
4. Calculer SCT, SCE et SCR, puis R^2 . Vérifier que $R^2 = r_{xy}^2$.
5. Quelles ventes la droite prévoit-elle pour un prix de 5 euros, puis de 15 euros? Commenter la seconde valeur.
6. On exprime désormais le prix en centimes. Donner, sans refaire les calculs, les nouvelles valeurs de $\hat{\beta}_0$, $\hat{\beta}_1$ et R^2 . Justifier.
7. Peut-on affirmer qu'un étal qui baisserait son prix d'un euro vendrait 1,5 panier de plus par heure? Expliquer.

Exercice 2. Les deux droites de régression

On dispose d'un échantillon $\{(x_i, y_i), i = 1, \dots, n\}$ avec $s_x > 0$ et $s_y > 0$. On note $\hat{\beta}_0$ et $\hat{\beta}_1$ les estimateurs des MCO dans la régression de y sur x , et $\hat{\gamma}_0$ et $\hat{\gamma}_1$ ceux de la régression de x sur y , c'est-à-dire de la droite $x = \gamma_0 + \gamma_1 y$.

1. Montrer que $\sum_{i=1}^n \hat{y}_i \hat{u}_i = 0$ dans la régression de y sur x .
2. Montrer que $\hat{\beta}_1 \hat{\gamma}_1 = r_{xy}^2$.
3. On suppose $r_{xy} \neq 0$. Dans le plan où x est en abscisse et y en ordonnée, la seconde droite s'écrit $y = (x - \hat{\gamma}_0) / \hat{\gamma}_1$. Montrer que les deux droites passent par le point $(\bar{x}, \bar{y})$, puis qu'elles sont confondues si et seulement si $|r_{xy}| = 1$.

- On pose $z_i = (x_i - \bar{x})/s_x$ et $w_i = (y_i - \bar{y})/s_y$. Montrer que la droite des MCO de w sur z est $\hat{w} = r_{xy} z$. En déduire que, si $|r_{xy}| < 1$, la valeur prédite $\hat{w}$ est toujours plus proche de 0 que z . Relier ce résultat à la « régression vers la moyenne » de Galton.
- Vérifier le résultat de la question 2 avec les données de l'exercice 1.

Exercice 3. Lire une table de régression

Les données ATTEND (Wooldridge) décrivent 680 étudiants d'un cours de principes de microéconomie. `final` est la note à l'examen final (en points), `attend` le nombre de séances suivies sur les 32 du semestre, `priGPA` la moyenne universitaire cumulée avant le semestre (sur 4). On obtient les régressions suivantes par les MCO.

Variable dépendante	(1) final	(2) final	(3) attend
attend	0,121*** (0,033)		
priGPA		3,164*** (0,309)	4,278*** (0,348)
Constante	22,730*** (0,877)	17,707*** (0,817)	15,081*** (0,919)
n	680	680	680
R^2	0,020	0,134	0,182

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

Dans l'échantillon, `attend` varie de 2 à 32.

- Écrire l'équation estimée de la colonne (1). Interpréter la pente en une phrase, avec les unités, puis donner l'écart de note prédit entre deux étudiants dont l'un a suivi 10 séances de plus que l'autre.
- Calculer la note prédite pour un étudiant présent aux 32 séances, puis pour un étudiant présent à 16 séances.
- La constante de la colonne (1) a-t-elle une interprétation utile ?
- Un camarade affirme qu'avec $R^2 = 0,020$, la présence n'a « aucun intérêt » pour expliquer la note et que la colonne (1) est inutilisable. Que lui répondre ?
- On mesure désormais la présence par le pourcentage de séances suivies, `atndrte` = $100 \times \text{attend}/32$. Donner, sans nouvelle estimation, la pente, la constante et le R^2 de la régression de `final` sur `atndrte`.
- Combien de séances la colonne (3) prédit-elle pour un étudiant dont `priGPA` vaut 3 ?
- À l'aide des colonnes (2) et (3), expliquer pourquoi la pente de la colonne (1) ne mesure probablement pas l'effet causal de la présence sur la note. Dans quel sens cette pente est-elle vraisemblablement faussée ?

Exercice 4. Résultats potentiels et tirage au sort

Une université propose en L2 un tutorat facultatif en statistique. On considère huit étudiants. Le tableau donne, pour chacun, la note à l'examen (sur 20) qu'il obtiendrait sans tutorat, $y_i(0)$, et

avec tutorat, $y_i(1)$, ainsi que son choix. Ces valeurs sont fictives. Dans la réalité, seule la note correspondant au choix est observée.

Étudiant	A	B	C	D	E	F	G	H
$y_i(0)$	8	9	7	10	14	13	15	12
$y_i(1)$	11	11	10	12	15	15	16	14
Participe au tutorat	oui	oui	oui	oui	non	non	non	non

1. Calculer l'effet causal du tutorat pour chaque étudiant, puis l'effet moyen sur les huit étudiants.
2. Quelles notes un chercheur observe-t-il ? Calculer l'écart entre la note moyenne observée des participants et celle des non-participants. Que conclurait un observateur naïf ?
3. Décomposer cet écart en un effet moyen chez les participants et un écart de sélection. Interpréter chaque terme.
4. Citer un facteur observable et un facteur difficile à observer qui peuvent expliquer l'écart de sélection.
5. On tire au sort quatre participants parmi les huit, par exemple A, C, E et G. Calculer l'écart observé avec ce tirage. Il y a 70 tirages possibles de quatre étudiants parmi huit. Que vaut, selon vous, l'écart moyen sur ces 70 tirages ? Justifier en une phrase, puis contrôler avec le code du corrigé.
6. Les données de la question 2 sont-elles expérimentales ou observationnelles ? Et celles de la question 5 ?

Exercice 5. Données réelles, vin et maladies cardiaques

Les données WINE (Wooldridge) donnent, pour 21 pays, la consommation d'alcool issu du vin (alcohol, en litres par habitant) et les décès par maladie cardiaque (heart), par maladie du foie (liver) et toutes causes (deaths), pour 100 000 habitants. Elles proviennent d'un article du *New York Times* de 1994. Pour les charger, on utilise l'un des trois codes suivants. Pour Stata, le fichier wine.dta du cours doit se trouver dans le dossier de travail.

```
[Python] ex5-depart
import subprocess, sys
subprocess.run([sys.executable, "-m", "pip", "install", "-q",
                "--disable-pip-version-check", "wooldridge"]) # installe si absent
import wooldridge as wd
wine = wd.data("wine") # 21 pays
print(wine)

[R] ex5-depart
if (!requireNamespace("wooldridge", quietly = TRUE)) install.packages("wooldridge")
library(wooldridge)
data("wine") # 21 pays
wine

[Stata] ex5-depart
use wine.dta, clear // 21 pays
list
```

1. Quelle est la structure de ces données ? Quelle est l'unité d'observation ?

2. Tracer le nuage de points de `heart` en fonction de `alcohol`. Calculer la covariance et la corrélation entre les deux variables.
3. Estimer par les MCO la régression de `heart` sur `alcohol`. Écrire l'équation estimée avec n et R^2 , et interpréter la pente.
4. Calculer la valeur prédite et le résidu de la France et du Japon. Commenter.
5. Vérifier numériquement que les résidus somment à zéro et que $SCT = SCE + SCR$.
6. Estimer les régressions de `liver`, puis de `deaths`, sur `alcohol`. Qu'apportent ces résultats ?
7. Peut-on conclure que boire du vin réduit le risque de maladie cardiaque ? Discuter en citant au moins deux raisons précises, et décrire l'expérience idéale qui permettrait de répondre.