

Économétrie

Séance 2 — Série d'exercices

Licence 3 Économie · Année universitaire 2026–2027

Les exercices portent sur les notions des séances 1 et 2. 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	une loi jointe, calculs à la main	20 min	TD
Exercice 2	démonstration	20 min	travail personnel
Exercice 3	lecture d'une table de régression	15 min	TD
Exercice 4	raisonner sur les hypothèses	15 min	TD
Exercice 5	données réelles et code	30 min	travail personnel
Total		110 min	60 min en TD

Quiz

Une seule réponse est correcte par question.

Question 1. Après une régression simple par les MCO, on constate que $\sum_{i=1}^n x_i \hat{u}_i = 0$. Que peut-on en conclure sur l'hypothèse RLS.4, $E(u \mid x) = 0$?

- a) Elle est satisfaite.
- b) Elle est violée.
- c) Rien, car cette égalité est vraie dans tout échantillon.
- d) Elle est satisfaite, à condition que n soit grand.

Question 2. « $\hat{\beta}_1$ est un estimateur sans biais de β_1 » signifie que

- a) $\hat{\beta}_1 = \beta_1$ dans l'échantillon dont on dispose
- b) en moyenne sur tous les échantillons possibles, $\hat{\beta}_1$ vaut β_1
- c) $\hat{\beta}_1$ se rapproche de β_1 quand n augmente
- d) la variance de $\hat{\beta}_1$ est nulle

Question 3. Dans une régression de la note sur les heures de travail salarié, RLS.4 est violée, car les étudiants qui travaillent beaucoup sont en moyenne moins motivés. On passe de 200 à 20 000 étudiants tirés dans la même population. Le biais de $\hat{\beta}_1$

- a) reste le même
- b) disparaît
- c) est divisé par 100
- d) change de signe

Question 4. L'hypothèse d'homoscédasticité RLS.5 est nécessaire pour que

- a) l'on puisse calculer $\hat{\beta}_1$
- b) $\hat{\beta}_1$ soit sans biais
- c) les résidus somment à zéro
- d) la formule $\text{Var}(\hat{\beta}_1) = \sigma^2 / \text{SCT}_x$ soit exacte

Question 5. Deux variables aléatoires X et Y ont une covariance nulle. Alors

- a) il n'existe pas de liaison linéaire entre X et Y
- b) X et Y sont indépendantes
- c) $E(Y | X)$ ne dépend pas de X
- d) $\text{Var}(X - Y) = \text{Var}(X) - \text{Var}(Y)$

Question 6. Dans une promotion, 40 % des étudiants sont boursiers. Leur note moyenne est 12, celle des autres étudiants 10. La note moyenne de la promotion vaut

- a) 11
- b) 10,8
- c) 11,2
- d) on ne peut pas la calculer sans les notes individuelles

Question 7. Deux chercheurs estiment la même pente, avec le même nombre d'observations et des erreurs de même variance. Les x du premier vont de 10 à 12, ceux du second de 5 à 20.

- a) Les deux estimations sont également précises.
- b) La première est plus précise, car ses x sont plus homogènes.
- c) La seconde est biaisée, car ses x sont trop dispersés.
- d) La seconde est plus précise, car ses x sont plus dispersés.

Question 8. L'estimateur $\hat{\sigma}^2$ divise la somme des carrés des résidus par $n - 2$ parce que

- a) les résidus vérifient deux contraintes, si bien que SCR/n sous-estime σ^2 en moyenne
- b) $\hat{\sigma}$ doit s'exprimer dans l'unité de y
- c) les deux premières observations servent à calculer $\hat{\beta}_0$ et $\hat{\beta}_1$
- d) $\hat{\sigma}^2$ doit être plus grand que la variance empirique de y

Question 9. Dans une table de régression, la pente vaut 2,5 et le nombre entre parenthèses au-dessous vaut 0,4. Ce nombre est

- a) l'écart-type de la variable explicative
- b) l'écart-type des résidus de la régression
- c) une estimation de l'écart-type de la distribution d'échantillonnage de $\hat{\beta}_1$
- d) l'écart entre $\hat{\beta}_1$ et β_1

Question 10. Une université tire au hasard 500 étudiants dans ses fichiers et compare la note finale

de ceux qui ont choisi de suivre un tutorat à celle des autres.

- a) L'échantillon est tiré au hasard, donc RLS.4 est satisfaite.
- b) RLS.2 est violée, car tous les étudiants n'ont pas suivi le tutorat.
- c) RLS.3 est violée, car x ne prend que deux valeurs.
- d) RLS.4 est douteuse, car le choix du tutorat dépend de facteurs qui agissent aussi sur la note.

Exercice 1. Une loi jointe, à la main

Dans une population d'étudiants, la note sur 20 suit le modèle $y = 8 + x + u$, où x est le nombre d'heures de tutorat suivies par semaine et u résume les autres facteurs, notamment la motivation. Le tableau donne la loi jointe de x et u , c'est-à-dire la probabilité de chaque couple de valeurs.

	$u = -2$	$u = 0$	$u = 2$
$x = 0$	0,10	0,15	0,05
$x = 2$	0,10	0,20	0,10
$x = 4$	0,05	0,15	0,10

1. Vérifier que la somme des probabilités vaut 1. Donner les lois de x et de u .
2. Calculer $E(x)$, $\text{Var}(x)$ et $E(u)$.
3. Calculer $E(u | x)$ pour $x = 0, 2$ et 4 . Vérifier sur cet exemple la loi des espérances itérées. RLS.4 est-elle satisfaite ?
4. Calculer $\text{Cov}(x, u)$. Le résultat est-il cohérent avec la question précédente ?
5. Calculer $E(y | x)$ pour chaque valeur de x . Montrer que $E(y | x)$ est une fonction affine de x et donner sa pente. La comparer à l'effet *ceteris paribus* d'une heure de tutorat. Interpréter l'écart.
6. On admet que la pente des MCO, calculée sur des échantillons tirés dans cette population, vaut en moyenne la pente trouvée à la question 5. Dire en une phrase ce que cela implique pour un chercheur qui estimerait β_1 par les MCO.
7. Construire une loi jointe qui a les mêmes lois de x et de u , mais pour laquelle x et u sont indépendants. RLS.4 est-elle alors satisfaite ? Justifier par une propriété du cours.

Exercice 2. La pente des extrêmes

On se place dans le modèle $y_i = \beta_0 + \beta_1 x_i + u_i$, sous les hypothèses RLS.1 à RLS.5, et on raisonne à $x_1, \dots, x_n$ donnés. Les observations sont rangées de sorte que x_1 soit la plus petite valeur de x et x_n la plus grande, avec $x_1 < x_n$. On propose un autre estimateur de la pente, qui n'utilise que les deux observations extrêmes,

$$\tilde{\beta}_1 = \frac{y_n - y_1}{x_n - x_1}.$$

1. Montrer que $\tilde{\beta}_1 = \beta_1 + \frac{u_n - u_1}{x_n - x_1}$.
2. En déduire que $\tilde{\beta}_1$ est sans biais. Quelles hypothèses avez-vous utilisées ?

3. Montrer que $\text{Var}(\tilde{\beta}_1) = \frac{2\sigma^2}{(x_n - x_1)^2}$.
4. Montrer que, pour tous réels a et b , $a^2 + b^2 \geq (a - b)^2/2$. En déduire que $\text{SCT}_x \geq (x_n - x_1)^2/2$, puis que $\text{Var}(\hat{\beta}_1) \leq \text{Var}(\tilde{\beta}_1)$, où $\hat{\beta}_1$ est l'estimateur des MCO.
5. Avec les cinq étudiants de l'exemple à la main de la séance 1 ($x = 2, 4, 6, 8, 10$ et $y = 7, 10, 11, 14, 13$), calculer $\tilde{\beta}_1$ et $\hat{\beta}_1$, puis le rapport $\text{Var}(\tilde{\beta}_1)/\text{Var}(\hat{\beta}_1)$.
6. Que montre cet exercice sur l'absence de biais comme critère de choix d'un estimateur ?

Exercice 3. Lire une table de régression

Les données NBASAL (Wooldridge) décrivent 269 joueurs de la ligue américaine de basket-ball (NBA), d'après un annuaire publié en 1995. *wage* est le salaire annuel en milliers de dollars, *points* le nombre moyen de points marqués par match au cours de la carrière, *exper* le nombre d'années comme joueur professionnel. On obtient les régressions suivantes par les MCO.

Variable dépendante	(1) wage	(2) points	(3) wage
points	111,33*** (7,82)		
exper		0,329*** (0,104)	120,32*** (16,42)
Constante	287,10*** (92,14)	8,528*** (0,641)	807,93*** (100,85)
n	269	269	269
R^2	0,432	0,036	0,167
$\hat{\sigma}$	755,07	5,80	913,96

Erreurs-types entre parenthèses. *** significatif à 1 %, ** à 5 %, * à 10 %.
 $\hat{\sigma}$ est l'erreur-type de la régression.

1. Écrire l'équation estimée de la colonne (1) et interpréter la pente en une phrase, avec les unités.
2. Quel salaire la colonne (1) prédit-elle pour un joueur qui marque 10 points par match ? Et 20 points ?
3. À partir de $\hat{\sigma}$ et de l'erreur-type de la pente en colonne (1), retrouver SCT_x pour la variable *points*, puis son écart-type empirique s_x .
4. Quelle serait, approximativement, l'erreur-type de la pente avec quatre fois plus de joueurs, de même dispersion de points et de même $\hat{\sigma}$?
5. L'écart-type des résidus de la colonne (1) vaut 567, 762, 839 et 814 dans les quatre quarts de points, du plus faible au plus fort. Quelle hypothèse est en cause ? Quelles conséquences pour $\hat{\beta}_1$ et pour son erreur-type ?
6. À l'aide des colonnes (2) et (3), discuter l'hypothèse RLS.4 en colonne (1). Dans quel sens la pente de la colonne (1) est-elle vraisemblablement faussée si l'on cherche l'effet *ceteris paribus* des points sur le salaire ?

Exercice 4. Raisonner sur les hypothèses

Pour chaque situation, dire quelle hypothèse parmi RLS.1 à RLS.5 est en cause, s'il y en a une, et la conséquence pour la pente des MCO. Peut-on la calculer ? Est-elle sans biais ? La formule usuelle de sa variance est-elle exacte ?

1. Pour étudier l'effet des heures passées en bibliothèque sur la note, on n'interroge que les étudiants qui ont validé leur semestre.
2. Dans un échantillon de salariés, tous ont exactement 12 années d'études. On veut l'effet de l'éducation sur le salaire.
3. Une ville tire au sort les quartiers qui reçoivent un éclairage public renforcé, puis régresse le nombre de cambriolages sur l'indicatrice d'éclairage.
4. On régresse l'épargne annuelle des ménages sur leur revenu. L'épargne des ménages aisés est bien plus dispersée que celle des ménages modestes.
5. Un lycée envoie en priorité ses élèves les plus faibles en soutien scolaire. On régresse la note au baccalauréat sur les heures de soutien.
6. On estime $\text{note} = \beta_0 + \beta_1 \log(\text{heures}) + u$.
7. Un collègue affirme que le modèle est faux, car la moyenne des facteurs omis vaut 3 points et non 0.

Exercice 5. Données réelles, éducation et salaire

Les données HTV (Wooldridge) décrivent 1 230 hommes de 26 à 34 ans interrogés en 1991 aux États-Unis. *wage* est le salaire horaire en dollars, *educ* le nombre d'années d'études, *abil* une mesure d'aptitude cognitive. Pour les charger, on utilise l'un des trois codes suivants. Pour Stata, le fichier *htv.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
htv = wd.data("htv") # 1 230 hommes de 26 à 34 ans, en 1991
print(htv[["wage", "educ", "abil"]].describe())
```

```
[R] ex5-depart
if (!requireNamespace("wooldridge", quietly = TRUE)) install.packages("wooldridge")
library(wooldridge)
data("htv") # 1 230 hommes de 26 à 34 ans, en 1991
summary(htv[, c("wage", "educ", "abil")])
```

```
[Stata] ex5-depart
use htv.dta, clear // 1 230 hommes de 26 à 34 ans, en 1991
summarize wage educ abil
```

1. Estimer la régression de *wage* sur *educ*. Écrire l'équation estimée avec n , R^2 et les erreurs-types, et interpréter la pente.
2. Calculer avec le code SCR, $\hat{\sigma}$, SCT_x et l'erreur-type de la pente à partir des formules du cours. Comparer aux résultats du logiciel.

3. Tracer les résidus en fonction de `educ` et calculer leur écart-type pour les hommes qui ont 12 années d'études ou moins, 13 ou 14 années, 15 années ou plus. Commenter au regard de RLS.5.
4. On traite les 1 230 hommes comme la population. Tirer 2 000 échantillons de 100 hommes avec remise. Pour chacun, calculer la pente des MCO et son erreur-type usuelle. Comparer la moyenne des pentes à la pente calculée sur les 1 230 hommes, puis l'écart-type des pentes à la moyenne des erreurs-types. Que conclure ?
5. La variable `abi1` fait partie de u dans la régression de la question 1. Calculer sa corrélation avec `educ` et avec `wage`. Qu'en déduire sur RLS.4 et sur la lecture de la pente ?