

# Exercices d'économétrie

## TD 1 : La structure des données.

Janvier, 2020

Cette suite d'exercices suit très largement les exercices proposés dans WOOLDRIDGE (2018), les sections qui suivent correspondent aux chapitres de cette référence. Ces exercices doivent, pour l'essentiel, être résolus à l'aide d'une machine puisqu'il s'agit d'applications fondées sur des données. À cette fin vous utiliserez le langage Python (version 3.x) pour répondre aux questions. Ce logiciel est disponible sur les ordinateurs de l'université, vous pouvez aussi facilement l'installer sur votre ordinateur personnel puisqu'il s'agit d'un logiciel libre. Le plus simple est probablement d'utiliser la distribution Anaconda :

<https://www.anaconda.com/distribution/>

qui en plus de Python installera des centaines de bibliothèques utiles (pour lire des données, calculer des statistiques, construire des graphiques, ...). Pour travailler sur ces exercices, il vous faudra utiliser les modules `pandas`, `statsmodels`, `numpy`, `scipy` et `matplotlib`. Vous trouverez facilement de la documentation pour Python et ces modules sur internet. Un point d'entrée général pour python est :

<https://docs.python.org/fr/3.7/>

Une introduction pour l'économétrie et les statistiques est disponible sur la page de Kevin Sheppard :

[https://www.kevinsheppard.com/files/teaching/python/notes/python\\_introduction\\_2019.pdf](https://www.kevinsheppard.com/files/teaching/python/notes/python_introduction_2019.pdf)

### Exercice 1.1

On utilise la base de données `wage1` pour cet exercice.

- i Calculez le niveau d'étude moyen dans l'échantillon (en utilisant la variable `educ`). Déterminez les nombres d'années d'étude minimum et maximum dans l'échantillon ?
- ii Calculez le salaire horaire moyen dans l'échantillon (en utilisant la variable `wage`).
- iii Corrigez le salaire horaire moyen calculé dans la question précédente en exprimant celui-ci en termes de dollars 2019 (plutôt que dollars 1976). Pour répondre à cette question il vous faudra utiliser un indice des prix pour les États-Unis, qui n'est pas disponible dans la base de données `wage1`, mais que vous pouvez obtenir sur le site `db.nomics`.
- iv Déterminez la répartition par sexe dans l'échantillon, i.e. quel est le nombre de femmes ?
- v Comparez le salaire horaire moyen des femmes et le salaire horaire moyen des hommes.

### Exercice 1.2

On utilise la base de données `meap01` pour cet exercice, une base de données du *Michigan Department of Education* concernant la réussite des élèves (en mathématiques et en lecture).

- i Déterminez l'étendue de la variable `math4`, c'est-à-dire les valeurs extrêmes de cette variable. Commentez.
- ii Combien d'écoles atteignent le taux de réussite maximal à l'examen de mathématiques ? Quel pourcentage de l'échantillon cela représente-t-il ?
- iii Combien d'écoles ont un taux de réussite à l'examen de mathématiques d'exactly 50% ?
- iv Comparez les taux de réussite moyen en mathématiques et en lecture. Lequel des deux examens est le plus dur ?

v Calculez la corrélation entre les variables `math4` et `read4`. On pourra utiliser la formule suivante :

$$\rho = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^N (X_i - \bar{X})^2 \sum_{i=1}^N (Y_i - \bar{Y})^2}}$$

pour calculer la corrélation entre les variables  $X$  et  $Y$  à partir des réalisations  $\{(X_i, Y_i) \mid i = 1, \dots, N\}$ . Commentez le résultat.

- vi La variable `exppp` représente les dépenses effectuées par élève. Calculez la moyenne et l'écart-type de `exppp`. Diriez-vous que l'hétérogénéité des dépenses par élève est importante ?
- vii Soient deux écoles  $A$  et  $B$  : l'école  $A$  dépense 6000\$ par élève et l'école  $B$  5500\$ par élève. De quel pourcentage les dépenses de l'école  $A$  dépassent celles de l'école  $B$  ? Expliquez pourquoi on pourrait vouloir approximer cette quantité par :

$$100 \times \log \left( \frac{6000}{5500} \right)$$

Évaluez l'erreur d'approximation.

- viii Quelle est la corrélation entre `math4` et `exppp` ?

### Exercice 1.3

On utilise la base de données `fertil2` pour cet exercice. Ces données collectées en 1988 au Botswana décrivent la fécondité des femmes. la variable `children` renseigne le nombre d'enfants vivants, la variable `electric` est une variable *dummy* qui vaut 1 si le foyer est raccordé à l'électricité.

- i Quelles sont les valeurs extrêmes de la variable `children` ? Calculez la moyenne.
- ii Quel est le pourcentage de femmes disposant de l'électricité à la maison ?
- iii Calculez la moyenne de la variable `children` pour les femmes disposant de l'électricité à la maison. Faire le même calcul pour les femmes ne disposant pas de l'électricité.

## Références

WOOLDRIDGE, Jeffrey M. (2018). *Introduction à l'économétrie : une approche moderne*. Seconde édition. De Boeck Supérieur.