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'esentiel, être résolus à l'aide d'une machine puisqu'il s'agit d'applications fondées sur des données. À cette fin vous utiliserez le language 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 puiqu'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 librairies 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 et disponible sur la page de Kevin Sheppard :

`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. Déterminez les nombres d'années d'étude minimum et maximum dans l'échantillon ?

ii Calculez le salaire horaire moyen dans l'échantillon.

iii Corrigez le salaire horaire moyen calculé dans la question précédente en exprimant celui-ci en termes de dollars 2018 (plutôt que dollars 1976). Pour répondre à cette question il vous faudra utiliser un indice des prix pour les Étatsz 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'exactement 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 :

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

pour calculer la corrélation entre les variables X et Y à partir des réalisations $\left\{(X_i, Y_i)\middle| i = 1, \ldots, N\right\}$. 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 poucentage 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.

Solution 1.1

(i) Le niveau d'étude moyen dans l'échantillon est 12,56 années. On trouve deux individus déclarant 0 années d'études et 19 individus déclarant 18 années d'études. (ii) Le salaire horaire moyen en 1976 était environ 5,9\$. Ce nombre peut paraître faible, mais il ne faut pas oublier qu'il s'agit de dollar de 1976. (iii) Pour exprimer le salaire horaire moyen de 1976 en termes de dollar de 2018, nous devons corriger de l'inflation. Nous utilisons pour cela une série temporelle d'indice des prix à la cosommation (CPI). Plusieurs institutions publient cette information, nous utilisons les données fournies par le FMI. Nous récupérons les données auprès de l'agrégateur de base de données db.nomics (le code de la série est `IMF/CPI/A.US.PCPI_IX`). L'indice des prix à la consommation vaut 26,0981 en 1976 et 115,1573 en 2018. Pour calculer le salaire horaire moyen en termes de dollar 2018, on fait :

$$w_{1976,2018} = w_{1976,1976}\frac{CPI_{2018}}{CPI_{2018}}$$

et on trouve 26,02\$. (iv) On sait que l'échantillon contient 526 individus. Pour déterminer le nombre de femmes on utilise la variable *dummy* `female` qui vaut 1 si l'individu est une femme, 0 sinon. En sommant les valeurs dans le vecteur `female` on obtient 252, le nombre de femmes. On obtient le nombre d'hommes par complémentarité : $274 = 526 - 252$. (v) Pour calculer le salaire horaire conditionnellement au genre, on doit créer deux sous échantillons selon les valeurs de la variable *dummy* `female`. On trouve que le salaire moyen des femmes $w_f = 20,2429$\$ est nettement inférieur au salaire horaire moyen des hommes $w_h = 31,3263$\$.

Solution 1.2

(i) Les valeurs extrêmes sont 0 et 100. (ii) Dans l'échantillon, 38 écoles obtiennent le taux de réussite maximal en mathématiques, soit 2,08%. (iii) 17 écoles obtiennent un taux de réussite exactement égal à 50%. (iv) Le taux de réussite moyen en mathématiques est 71,91%. Le taux de réussite moyen est sensiblement inférieur en lecture (60,06%), ce qui laisse penser que cette épreuve est plus difficile. (v) La corrélation entre les taux de réussite en mathématiques et en lecture est 0,84. Sachant qu'une corrélation est nécessairement comprise entre -1 et 1, on voit que celle-ci est importante. Ce résultat nous dit, qu'en moyenne, les écoles où les résultats sont bons en mathématiques sont aussi les écoles ont les

résultats sont bons en lecture. (vi) La dépense moyenne par élève est de 5194,87$, l'écart-type de la dépense par élève de 1091,89 est relativement important. (vii) On calcule le pourcentage de la façon suivante :

$$p = 100 \times \left(\frac{6000}{5500} - 1 \right) \simeq 8,7$$

En utilisant la formule alternative donnée dans l'ennoncé, on trouve :

$$\hat{p} \simeq 9,1$$

Ces deux pourcentages sont sensiblement différents. La formule alternative est une approximation du pourcentage, elle repose sur une approximation de la fonction logarithme. En utilisant un développement de Taylor à l'ordre 1 dans un voisinage de 0, on sait que :

$$\log(1+x) = x + o(x^n)$$

Posons $x = \left(\frac{6000}{5500} - 1\right)$, on a alors :

$$1 + x = \frac{6000}{5500}$$

et en utilisant l'approximation :

$$\log\left(\frac{6000}{5500}\right) \simeq \left(\frac{6000}{5500} - 1\right)$$

c'est-à-dire :

$$100 \times \log\left(\frac{6000}{5500}\right) \simeq 100 \times \left(\frac{6000}{5500} - 1\right)$$

ou encore :

$$\hat{p} \simeq p$$

Mais cette approximation n'est « bonne » que pour x, c'est-à-dire ${}^{6000}/_{5500} - 1$, proche de 0. Cette approximation n'est valide que pour des « petits » pourcentages. (viii) On trouve une corrélation quasi nulle, $\rho = -0,03$.

Solution 1.3

(i) Le nombre maximal d'enfants par femme, dans cet échantillon, est 13. Le nombre minimal d'enfants par femme est 0. (ii) Le nombre moyen d'enfants par femme est 2,27. (iii) Le pourcentage de femmes disposant de l'électricité à la maison est 14,01%. (iv) Le nombre moyen d'enfants par femme disposant de l'électricité est 1,9. Le nombre moyen d'enfants par femme ne disposant pas de l'électricité est 2,33.

Références

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