

PROBLÈME 2: RÉGRESSION LINÉAIRE SIMPLE

MICHAL URDANIVIA, UNIVERSITÉ DE GRENOBLE ALPES, FACULTÉ D'ÉCONOMIE, GAEL

Courriel: michal.wong-urdanivia@univ-grenoble-alpes.fr

4 OCTOBRE 2017

EXERCICE 1

Les données "BWGHT.dta" contiennent des informations sur les naissances par des femmes au États-Unis. Deux variables retiennent ici notre attention : le poids à la naissance des nouveaux-nés(mesuré en onces(variable *bwght*), et le nombre moyen de cigarettes fumées par jour durant la période de grossesse(variable *cigs*). On souhaite estimer la régression suivante sur les données relatives à $N = 1388$ naissances,

$$bwght_i = \beta_0 + \beta_1 cigs_i + U_i, \quad i = 1, \dots, N$$

où l'on suppose que $\mathbb{E}(U_i | cigs_i) = 0$. L'échantillon $\{(bwght_i, cigs_i)\}_{i=1}^N$ est par ailleurs supposé i.i.d.

- (1) Donnez une interprétation de β_0 et de β_1 .
- (2) L'estimation par moindres carrés donne $\hat{\beta}_0 = 119.77$ et $\hat{\beta}_1 = -0.514$. Rappelez les formules pour obtenir ces estimateurs.
- (3) Quel est le poids prédit par ce modèle quand $cigs = 0$, et celui quand $cigs = 20$? Commentez la différence.
- (4) L'hypothèse $\mathbb{E}(U_i | cigs_i) = 0$ vous semble t-elle justifiée?(argumentez votre réponse).
- (5) Présentez un script (ou "notebook jupyter") Python permettant de calculer $\hat{\beta}_1$ et $\hat{\beta}_0$ (avec lequel vous devez trouver les valeurs indiquées plus-haut).

EXERCICE 2

Le site de Joshua Angrist contient les données utilisées dans plusieurs de ses travaux¹. On s'intéresse ici à celle relatives à [Angrist and Lavy \(1999\)](#). Dans ce travail les auteurs s'intéressent à l'effet causal de la taille des classes sur les résultats des élèves à des tests d'évaluation. Pour identifier cet effet causal ils considèrent le cas des écoles israéliennes dont une des règles de fonctionnement leur permet d'utiliser une stratégie par variables instrumentales. Comme cette méthode sera vue dans une partie plus avancée du cours, on se contentera pour le moment des aspects du travail qui emploient des régressions "de base"

- (1) Lisez l'article jusqu'à la fin de la section 1(au moins), et faites un résumé de deux pages au plus en présentant la problématique, et les données.
- (2) Téléchargez les données et calculez les statistiques descriptives du tableau 1 pour l'échantillon de la cinquième année("5th grade", soit 2019 observations, voir page 539 dans la revue, 7 dans le fichier pdf). Vous fournirez le code Python utilisé. Notez que l'unité d'observation est la classe, que le nombre d'inscrits("enrollment"

1. Allez à (<http://economics.mit.edu/faculty/angrist/data1/data/anglavy99>)

en anglais dans l'article) est appelé *c_size*, que le pourcentage d'élèves désavantagés("percent disadvantaged" en anglais dans l'article) est appelé *tipuach*.

En outre, pour reproduire exactement les nombres du tableau 4 vous devez suivre les indications de la note de bas de page 11 et restreindre l'échantillon aux écoles avec au moins 5 inscriptions et des classes de taille inférieures à 45. Il y a aussi certaines corrections peu évidentes à faire sur les données. Ainsi il y a des observations qui présentent une moyenne au test de maths(variable *avgmath*) et une moyenne au test en expression écrite *avgverb* supérieures à 100, en raison d'erreurs dans l'enregistrement des données. Les valeurs correctes sont 87.606 et 81.246 (et non 187.606 et 181.246). Enfin, il y a un score en maths non manquant pour une observation avec *mathsize* = 0 ce qui est impossible(car pas d'élèves en maths). Il faut ici remplacer cette valeur non manquante par une valeur manquante. Le code qui accompagne ce TD règle les problèmes ci-dessus mentionnés.

- (3) Un débat courant entre économistes, et d'autres chercheurs en sciences sociales concerne l'arbitrage entre les coûts et les bénéfices associés à la réduction de la taille des classes. Quel serait le signe de l'effet de la réduction de la taille des classes sur les résultats scolaires(tels que mesurés par des tests comme ceux de l'article) si cette réduction valait la peine ?
- (4) Estimez la régression des résultats en maths sur la taille des classes(et une constante), et des résultats en expression écrite sur la taille des classes(et une constante). Quel est le signe de ces relations ? Sont-elles significativement différentes de zéro ?

RÉFÉRENCES

Angrist, J.D. and V. Lavy. 1999. "Using Maimonides' rule to estimate the effect of class size on scholastic achievement." *The Quarterly Journal of Economics* 114 (2) :533–575. URL <http://qje.oxfordjournals.org/content/114/2/533.short>.