Exercices d'économétrie

TD 5. Régression multiple : résultats asymptotiques des MCO.

Janvier, 2020

Exercice 5.1

On considère les données contenues dans le fichier `wage1` pour cet exercice.

i Estimez l'équation :

$$\texttt{wage}_i = \beta_0 + \beta_1 \texttt{educ}_i + \beta_2 \texttt{exper}_i + \beta_3 \texttt{tenure}_i + u_i$$

où `wage` est le salaire horaire, `educ` l'éducation en nombre d'années, `exper` l'expérience en nombre d'années et `tenure` le nombre d'années chez l'employeur actuel. Récupérez les résidus et réalisez un histogramme.

ii Répétez l'étape (i), mais avec log(`wage`) comme variable dépendante.

iii Diriez-vous que l'hypothèse de normalité des résidus est plus proche d'être satisfaite pour le modèle en niveau ou en log-niveau ?

Exercice 5.2

Utilsez les données contenues dans `gpa2` pour cet exercice.

i Utilisez les 4 137 observations, estimez l'équation :

$$\texttt{colgpa}_i = \beta_0 + \beta_1 \texttt{hsperc}_i + \beta_2 \texttt{sat}_i + u_i$$

où `colgpa` est le Grade Point Average (la moyenne des notes) du premier semestre, `hsperc` High School Percentile from Top (le classement de l'école en centiles à partir du haut), `sat` le score du test d'entrée à l'université.

ii Réestimez l'équation de la question (i), en utilisant que les 2 070 premières observations.

iii Calculez le ratio des écarts-types estimés de $\hat{\beta}_1$ dans le cadre des questions (i) et (ii).

Exercice 5.3

Utilisez les données contenues dans `bwght` pour cet exercice.

i Estimez l'équation suivante :

$$\texttt{bwght}_i = \beta_0 + \beta_1 \texttt{cigs}_i + \beta_2 \texttt{parity}_i + \beta_3 \texttt{faminc}_i + \beta_4 \texttt{motheduc}_i + \beta_5 \texttt{fatheduc}_i + u_i$$

où `bwght` est le poids du bébé à la naissance (baby weight), `cigs` la consommation de tabac de la mère durant la grossesse, `parity` rang de naissance, `faminc` le revenu familial (family income), `motheduc` le niveau d'éducation de la mère (mother education) et `fatheduc` le niveau d'éducation du père (father education).

ii Calculez la statistique LM pour tester si `motheduc` et `fatheduc` sont conjointement significatifs. En récupérant les résidus du modèle contraint, prenez garde à ce que le modèle contraint soit estimé avec les seules observations disponibles pour toutes les variables du modèle non contraint.

Exercice 5.4

Plusieurs statistiques sont couramment utilisées pour détecter une éventuelle non normalité dans les distributions de la population sous-jacente. Nous allons nous concentrer dans cet exercice sur l'une d'entre elles qui mesure la quantité d'asymétrie (ou coefficient de « skewness ») de la distribution. Rappelez-vous que toute variable aléatoire caractérisée par une distributon normale est symétrique autour de sa valeur moyenne ; dès lors si nous standardisons une variable aléatoire caractérisée par une distribution symétrique, disons $z = (y - \mu_y)/\sigma_y$, avec $\mu_y = \mathbb{E}[y]$ et σ_y l'écart-type de y, alors z est de moyenne nulle, de variance un, et $\mathbb{E}[z^3] = 0$. Soit un échantillon de données observées $\{y_i; i = 1, \ldots, n\}$, nous pouvons standardiser y_i dans l'échantillon en utilisant $z_i = (y_i \hat{\mu}_y)/\hat{\sigma}_y$ avec $\hat{\mu}_y$ et $\hat{\sigma}_y$ les moments empirique. Une

statistique empirique mesurant le coefficient d'asymétrie est donnée par $n^{-1}\sum_{i=1}^{n} z_i^3$ (ou par la même somme divisée par $n-1$ plutôt que n de façon à ajuster le nombre de degrés de liberté. Si y est caractérisée par une distribution normale dans la population, le coefficient d'aymétrie mesuré sur la base de l'échantillon observé, ne devrait pas être significativement différent de zéro.

i Utilisez tout d'abord les données contenues dans `401ksubs`, en ne conservant que les données pour lequeslles `fsize` (la taille de la famille) est égal à 1. Identifiez la mesure du coefficient d'asymétrie pour `inc` (le revenu annuel en milliers de dollars). Faites de même pour log(`inc`). Quelle est, parmi ces deux variables celle dont le coefficient d'asymétrie est le plus élevé et de ce fait, semble le moins corespondre à une variable normalement distribuée ?

ii Dans un second temps, utilisez les données issues de la base `bwght2`. Identifiez les mesures du coefficient d'asymétrie pour `bwght` (le poids du bébé à la naissance) et log(`bwght`). Que pouvez-vous en conclure ?

iii Évaluez l'assertion suivante : « La transformation logarithmique permet à des variables positives de se rapprocher d'une distribution normale. »

iv Dans la mesure où nous nous intéressons à l'hypothèse de normalité dans le contexte de régression linéaire, devrions-nous évaluer les distributions non conditionnelles de y et $\log(y)$?

Exercice 5.5

Dans cet exercice nous n'utiliserons pas de données mais des données engendrées aléatoirement, afin d'illustrer les propriétés de l'estimateur des Moindres Carrés Ordinaires.

i Créez 500 000 observations pour une variable explicative, x_i, en tirant dans une loi uniforme entre 0 et 10. Utilisez la fonction `rand` dans le module Python `numpy.random`. Générez 500 000 erreurs, u_i, en tirant dans une loi normale d'espérance nulle et de variance 36 (et donc d'écart-type 6). Vous utiliserez la fonction `randn` du même module `numpy.random`. Créez la variable y selon le modèle :

$$y_i = b_0 + b_1 x_i + u_i$$

avec $b_0 = 1$ et $b_1 = 2$. Représentez graphiquement l'échantillon (utilisez le module `matplotlib`).

ii En utilisant les données engendrées dans la question précédente, estimez le modèle :

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

par les Moindres Carrés Ordinaires pour 10 échantillons contenant les 10, 50, 100, 500, 1 000, 5 000, 10 000, 50 000, 100 000 et 500 000 premières observations. Que remarquez-vous ?

iii Reprenez l'étape précédente en simulant 1 000 échantillons et en stockant les valeurs obtenues des estimateurs. Pour chaque échantillon vous utiliserez les mêmes réalisations de la variable explicative, *i.e.* les différences entre les échantillons ne proviennent que des erreurs. Pour chaque dimension d'échantillon (10, 50, 100, 500, 1 000, 5 000, 10 000, 50 000, 100 000 et 500 000) calculez la variance de $\hat{\beta}_1$ et représentez sa distribution à l'aide d'un histogramme. Que remarquez-vous ?