

Sujet de mémoire

On dispose de données concernant un échantillon de 807 individus, fumeurs et non-fumeurs ; les variables sont :

educ	le nombre d'années d'études
pcig	le prix moyen du paquet de cigarettes dans l'État
age	l'âge
rev	le revenu annuel
cigs	le nombre moyen de cigarettes fumées quotidiennement
nres	une variable binaire signalant des restrictions à la tabagie dans les restaurants de l'État considéré

(pcig et rev sont exprimées dans des unités convenables).

On créera les variables dérivées :

lpcig	le logarithme népérien du prix moyen du paquet
age2	le carré de l'âge
fum	une variable binaire indiquant si l'individu est fumeur

(en Eviews, on pourra utiliser la formule : « $fum = (cigs > 0)$ »)

et on triera le fichier de manière à séparer les deux groupes : non-fumeurs et fumeurs.

1° On examine tout d'abord brièvement les données par des méthodes d'analyse élémentaires.

Calculer les statistiques usuelles : moyenne, écart-type, minimum et maximum, pour chaque variable, d'une part pour la population complète et d'autre part pour chacun des deux groupes : fumeurs et non-fumeurs et commenter brièvement les résultats obtenus.

2° Effectuer un test de Student d'égalité de deux moyennes pour comparer respectivement l'âge et le revenu entre les deux groupes. Indiquer si les écarts sont significatifs.

3° On cherche à présent à mettre en évidence les déterminants de la consommation de cigarettes : *cigs*, parmi les variables disponibles.
On considère les trois modèles (linéaires et avec la constante) :

$$\text{cigs} = f(\text{educ}, \text{pcig}, \text{age}, \text{rev}, \text{nres}) \quad (\text{eq.1})$$

$$\text{cigs} = f(\text{educ}, \text{age}, \text{rev}, \text{nres}, \text{lpcig}) \quad (\text{eq.2})$$

$$\text{cigs} = f(\text{educ}, \text{age}, \text{age2}, \text{rev}, \text{nres}, \text{lpcig}) \quad (\text{eq.3})$$

Les estimer par les mco sur l'ensemble des observations, présenter dans un tableau classique les résultats des trois estimations et les commenter succinctement.

Les résultats sont-ils contradictoires ?

Quelles variables paraissent à conserver ?

4° On décide de retenir le modèle :

$$\text{cigs} = f(\text{educ}, \text{nres}, \text{age}, \text{age2}) \quad (\text{eq.4})$$

Ce choix paraît-il pertinent ?

Estimer ce modèle.

Que signifie le signe du coefficient de la variable *age2* ?

Quelle serait, toutes choses égales par ailleurs, l'incidence d'une année de plus sur la consommation journalière de cigarettes d'un individu de 30 ans ?

À quel âge la consommation – toutes choses égales par ailleurs – commencerait-elle à décroître ?

Le modèle a-t-il une forte valeur prédictive ?

5° Supposant que l'emploi de la variable *cigs*, bornée inférieurement par zéro, a pu poser un problème d'estimation, réestimer l'équation précédente en utilisant uniquement la sous-population des fumeurs.

Commenter les résultats.

6° Considérant que les données concernant les non-fumeurs sont néanmoins porteuses d'une part d'information qu'il serait maladroit de ne pas utiliser, réestimer une nouvelle fois l'équation 4 en utilisant le modèle *tobit* et commenter les résultats.

7° Construire un tableau présentant les résultats des trois dernières procédures d'estimation (questions 4°, 5° et 6°)

8° On tente maintenant d'expliquer le fait de fumer, c'est à dire la variable *fum*, par un modèle de choix binaire.

Estimer le modèle *logit* :

$$\text{fum} = f(\text{educ}, \text{age}, \text{age2}, \text{rev}, \text{nres}, \text{lpcig}) \quad (\text{eq.5})$$

et commenter les résultats obtenus.

9° En abandonnant les variables les moins significatives, quel modèle *logit* pourrait-on retenir ?

Estimer ce modèle.

Ces résultats paraissent-ils cohérents avec ceux des questions précédentes ?

10° Revenant sur le modèle estimé en 4°, on imagine que le revenu n'est peut-être pas exogène mais dépend lui-même de la consommation de cigarettes (par exemple via l'absentéisme éventuel consécutif à des problèmes de santé) et que les grandeurs *cigs* et *rev* sont liées dans un modèle linéaire à équations simultanées :

$$\text{rev} = f(\text{cigs}, \text{educ}, \text{age}, \text{age2}) \quad (\text{eq.6})$$

$$\text{cigs} = f(\text{rev}, \text{educ}, \text{age}, \text{age2}, \text{nres}) \quad (\text{eq.7})$$

(dont *rev* et *cigs* sont donc les endogènes).

Estimer l'équation 6 par les mco et commenter les résultats obtenus.

La méthode est-elle adaptée ?

11° Examiner l'identifiabilité des deux équations.

12° Réestimer l'équation 6 par les doubles moindres carrés en utilisant comme instruments l'ensemble des exogènes : *educ*, *pcig*, *age*, *nres* et *age2*.

Commenter les résultats et les comparer à ceux de la question 10°.

13° À quel âge – toutes choses égales par ailleurs - le revenu commencerait-il à décroître ?



Directives et conseils

Les données sont dans le fichier Excel « tabac.xls » (disponible sur le site du mastère). Le prétraitement des données comme les questions 1° et 2° pourront être faits en Excel ou avec le logiciel économétrique employé.

Il est impérativement demandé d'effectuer le travail proprement économétrique en Eviews (ou exceptionnellement en utilisant les logiciels libres gretl ou R), à l'exclusion de tout autre produit.

Le mémoire rendu, d'une trentaine de pages au maximum, répondra avec précision aux questions posées, en respectant soigneusement les notations et numérotations de l'énoncé.

Les rappels de cours sont inutiles et non souhaités.

Les régressions, et autres résultats demandés et commentés, figureront, convenablement mis en forme, dans le corps du texte (utiliser par exemple une police de largeur fixe pour les tableaux).

La collaboration n'est ni interdite ni déconseillée, mais les mémoires rendus n'auront qu'un seul rédacteur et signataire.

Les mémoires sont à rendre, imprimés et non sous forme de fichier informatique, au secrétariat du mastère le 7 juin 2007 au plus tard.

