Éléments de corrigé

1° Population totale

*	AGE	AGE2	CIGS	EDUC	FUM	LPCIG	NRES	PCIG	REV
Mean	41.23792	1990.135	8.686493	12.47088	0.384139	4.096032	0.246592	60.30041	19304.83
Median	38.00000	1444.000	0.000000	12.00000	0.000000	4.111742	0.000000	61.05300	20000.00
Maximum	88.00000	7744.000	80.00000	18.00000	1.000000	4.250336	1.000000	70.12900	30000.00
Minimum	17.00000	289.0000	0.000000	6.000000	0.000000	3.784281	0.000000	44.00400	500.0000
Std. Dev.	17.02729	1577.166	13.72152	3.057161	0.486693	0.082919	0.431295	4.738469	9142.958
Observations	807	807	807	807	807	807	807	807	807

Fumeurs

*	AGE	AGE2	CIGS	EDUC	FUM	LPCIG	NRES	PCIG	REV
Mean	39.11290	1753.997	22.61290	11.99677	1.000000	4.093111	0.196774	60.13808	19256.45
Median	36.50000	1332.500	20.00000	12.00000	1.000000	4.108658	0.000000	60.86500	20000.00
Maximum	85.00000	7225.000	80.00000	18.00000	1.000000	4.247051	1.000000	69.89900	30000.00
Minimum	17.00000	289.0000	1.000000	6.000000	1.000000	3.792811	0.000000	44.38100	500.0000
Std. Dev.	14.99677	1313.337	13.23543	2.622097	0.000000	0.085773	0.398203	4.888720	9101.791
Observations	310	310	310	310	310	310	310	310	310

Non-fumeurs

*	AGE	AGE2	CIGS	EDUC	FUM	LPCIG	NRES	PCIG	REV
Mean	42.56338	2137.425	0.000000	12.76660	0.000000	4.097853	0.277666	60.40167	19335.01
Median	39.00000	1521.000	0.000000	12.00000	0.000000	4.114262	0.000000	61.20700	20000.00
Maximum	88.00000	7744.000	0.000000	18.00000	0.000000	4.250336	1.000000	70.12900	30000.00
Minimum	17.00000	289.0000	0.000000	6.000000	0.000000	3.784281	0.000000	44.00400	500.0000
Std. Dev.	18.06765	1706.154	0.000000	3.267455	0.000000	0.081123	0.448299	4.644401	9177.569
Observations	497	497	497	497	497	497	497	497	497

2° Tests d'égalité de deux moyennes : ce test opère de manière différente selon que l'on suppose les écarts types des deux sous-populations égaux ou non. Évidemment, si leurs estimations empiriques sont proches, la conclusion est la même.

Ayant trouvé la formule appropriée dans un manuel de statistique on peut faire le calcul « à la main. » ; on peut également utiliser l'utilitaire d'analyse approprié d'Excel.

Une manière plus subtile de procéder recourt au test de stabilité de Chow (disponible dans tous les logiciels économétriques) : on sait que la régression d'une variable sur la seule constante donne sa moyenne, il suffit donc de mettre en œuvre le test de stabilité pour cette régression entre les deux groupes (préalablement triés si nécessaire) : les fumeurs et les non-fumeurs. Voici les résultats par cette méthode.

Âge

Chow Breakpoint Test: 311

F-statistic	7.907112	Prob. F(1,805)	0.005044
Log likelihood ratio	7.888134	Prob. Chi-Square(1)	0.004976

Les non-fumeurs sont donc significativement plus âgés que les fumeurs.

Revenu

Chow Breakpoint Test: 311

F-statistic	0.014077	Prob. F(1,805)	0.905583
Log likelihood ratio	0.014166	Prob. Chi-Square(1)	0.905259

Il n'y a pas de différence significative quant au revenu.

Une remarque rassurante : la statistique de Fisher calculée est simplement le carré du t de Student utilisé dans la version classique avec variances égales, il s'agit mathématiquement du même test.

3° et 4° Les trois régressions montrent des R^2 vraiment très faibles, cela est fréquent avec les modèles terriblement simplificateurs de comportement humain, les bons ratios de Student permettent néanmoins d'identifier des variables ayant une incidence réelle. Négligeant certainement beaucoup d'autres déterminants de la variable à expliquer, ces modèles sont de très modeste valeur prédictive.

*	Equation 1		Equation 2		Equation 3		Equation 4	
Variable	Coefficient	Student	Coefficient	Student	Coefficient	Student	Coefficient	Student
C	13.01946	1.987332	14.01835	0.584048	2.379099	0.100335	0.152140	0.043427
EDUC	-0.368047	-2.175577	-0.367800	-2.174081	-0.494339	-2.940396	-0.450400	-2.789102
PCIG	0.004674	0.045609						
AGE	-0.043899	-1.529196	-0.043817	-1.526651	0.784559	4.923301	0.822327	5.333323
AGE2					-0.009169	-5.281805	-0.009589	-5.714632
REV	0.000131	2.332421	0.000131	2.334547	5.45E-05	0.956610		
NRES	-2.982192	-2.636786	-2.970217	-2.622862	-2.831490	-2.541286	-2.746372	-2.503872
LPCIG			-0.177913	-0.030344	-0.484952	-0.084083		
N	807		807		807		807	
R2	0.019035		0.019034		0.052089		0.051000	
Adj.R2	0.012912		0.012910		0.044980		0.046267	
F-stat.	3.108583		3.108347		7.326848		10.77509	
Prob.F	0.008701		0.008705		0.000000		0.000000	

L'ajout de la variable age2 (équation 3), qui modifie la prise en compte de l'influence de l'âge, améliore notablement le coefficient de corrélation multiple, et rend non significative et inutile la variable rev, le prix du paquet de cigarettes paraît également sans influence sur la consommation, qu'il soit mesuré en niveau ou par son logarithme.

L'abandon de ces deux dernières explicatives (équation 4) améliore en principe la qualité des estimations des coefficients des variables conservées, qui au reste ne changent guère et ont les signes a priori attendus.

Le coefficient négatif d'age2 traduit un phénomène classique de « rendement décroissant » ; toutes choses égales par ailleurs, on a :

$$cigs = - 0.009589*age^2 + 0.822327*age + constante$$

La courbe est un parabole concave (en « U renversé ») qui croit de manière ralentie jusqu'au sommet, atteint pour 42,88 ans, avant de décroître de manière accélérée.

On calcule encore l'accroissement théorique de la consommation entre 30 et 31 ans :

$$cigs(31) – cigs(30) = 0{,}24 \text{ cigarette}$$

4°, 5°, 6° et 7°

Dependent Variable: CIGS
Method: ML - Censored Normal (TOBIT) (Quadratic hill climbing)
Date: 06/04/07 Time: 15:38
Sample: 1 807
Included observations: 807
Left censoring (value) at zero
Convergence achieved after 6 iterations
Covariance matrix computed using second derivatives

	Coefficient	Std. Error	z-Statistic	Prob.
C	-20.16996	9.073732	-2.222895	0.0262
EDUC	-1.747149	0.420576	-4.154186	0.0000
NRES	-7.949835	2.814057	-2.825044	0.0047
AGE	2.166077	0.418392	5.177152	0.0000
AGE2	-0.026738	0.004727	-5.656466	0.0000
Error Distribution				
SCALE:C(6)	28.04717	1.297537	21.61571	0.0000

R-squared	0.042168	Mean dependent var	8.686493
Adjusted R-squared	0.036189	S.D. dependent var	13.72152
S.E. of regression	13.47095	Akaike info criterion	4.357934
Sum squared resid	145354.6	Schwarz criterion	4.392828
Log likelihood	-1752.426	Hannan-Quinn criter.	4.371333
Avg. log likelihood	-2.171532		

Left censored obs	497	Right censored obs	0
Uncensored obs	310	Total obs	807

La variable cigs ne pouvant être négative est dite « censurée à gauche en zéro ». Il est à craindre que cela n'altère le caractère linéaire du modèle précédemment supposé et estimé, et par suite la qualité des estimations par les mco obtenues.

Une solution grossière consiste à ne conserver que les fumeurs. Le modèle ainsi estimé voit l'influence de l'éducation changer de signe et de sens et celle de la législation anti-tabac perdre sa significativité.

La solution précédente introduit ce que l'on appelle un « biais de sélection » et l'estimation d'un modèle tobit est préférable. Les résultats redonnent son sens initial à l'influence de la variable educ et modifient quelque peu la forme de la courbe en U traduisant l'influence de l'âge.

*	Equation 4		Equation 4 fumeurs		Equation 4 tobit	
Variable	Coefficient	Student	Coefficient	Student	Coefficient	Z-stat
C	0.152140	0.043427	-9.457366	-1.602343	-20.16996	-2.222895
EDUC	-0.450400	-2.789102	0.655841	2.266940	-1.747149	-4.154186
NRES	-2.746372	-2.503872	-2.483702	-1.355513	-7.949835	-2.825044
AGE	0.822327	5.333323	1.184290	4.307444	2.166077	5.177152
AGE2	-0.009589	-5.714632	-0.012332	-3.890953	-0.026738	-5.656466
N	807		310		807	
R2	0.051000		0.093001		0.042168	
Adj.R2	0.046267		0.081105		0.036189	
F-stat.	10.77509		7.818410			
Prob.F	0.000000		0.000005			

8° et 9°

Dependent Variable: FUM
Method: ML - Binary Logit (Quadratic hill climbing)
Date: 06/04/07 Time: 15:40
Sample: 1 807
Included observations: 807
Convergence achieved after 7 iterations
Covariance matrix computed using second derivatives

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	0.895577	3.723274	0.240535	0.8099
EDUC	-0.133502	0.027930	-4.779927	0.0000
AGE	0.104807	0.027921	3.753649	0.0002
AGE2	-0.001375	0.000317	-4.331098	0.0000
REV	1.66E-06	9.02E-06	0.184234	0.8538
NRES	-0.446005	0.182095	-2.449297	0.0143
LPCIG	-0.307308	0.906663	-0.338944	0.7347

Mean dependent var	0.384139	S.D. dependent var	0.486693
S.E. of regression	0.473425	Akaike info criterion	1.282245
Sum squared resid	179.3049	Schwarz criterion	1.322955
Log likelihood	-510.3858	Hannan-Quinn criter.	1.297878
Restr. log likelihood	-537.5055	Avg. log likelihood	-0.632448
LR statistic (6 df)	54.23944	McFadden R-squared	0.050455
Probability(LR stat)	6.60E-10		

Obs with Dep=0	497	Total obs	807
Obs with Dep=1	310		

Dependent Variable: FUM
Method: ML - Binary Logit (Quadratic hill climbing)
Date: 06/04/07 Time: 15:41
Sample: 1 807
Included observations: 807
Convergence achieved after 5 iterations
Covariance matrix computed using second derivatives

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-0.360331	0.576574	-0.624951	0.5320
EDUC	-0.132359	0.026966	-4.908385	0.0000
AGE	0.105720	0.027056	3.907478	0.0001
AGE2	-0.001385	0.000308	-4.498526	0.0000
NRES	-0.451770	0.179332	-2.519183	0.0118

Mean dependent var	0.384139	S.D. dependent var	0.486693
S.E. of regression	0.472881	Akaike info criterion	1.277471
Sum squared resid	179.3405	Schwarz criterion	1.306550
Log likelihood	-510.4595	Hannan-Quinn criter.	1.288637
Restr. log likelihood	-537.5055	Avg. log likelihood	-0.632540
LR statistic (4 df)	54.09206	McFadden R-squared	0.050318
Probability(LR stat)	5.03E-11		

Obs with Dep=0	497	Total obs	807
Obs with Dep=1	310		

Les modèles logit à présent envisagés ne cherchent plus qu'à expliquer la probabilité qu'un individu a d'être un fumeur (variable fum), indépendamment de l'intensité de sa consommation éventuelle de tabac.

Le second modèle estimé (équation 6) ne conserve que les variables paraissant significatives à la lumière de leurs « z-Statisques » (il s'agit d'un test de la log-vraisemblance, via une loi N(0,1), qui remplace le test de Student des mco classiques - son interprétation est similaire).

On observe, ce qui est rassurant, que les variables retenues pour leur influence sur la probabilité de fumer sont les mêmes que celle qui déterminaient l'intensité de cette consommation dans les questions précédentes : educ, age, age2 et nres.

Pour un commentaire plus précis, les coefficients estimés ne mesurent pas l'influence linéaire des variables explicatives sur la probabilité de fumer, mais sur le logit de celle-ci. Les « odds ratio » (non donnés ici) mesureraient l'influence multiplicative sur, précisément, l'odd ratio de la probabilité, (cad le rapport p/(1-p)). Il serait également possible de calculer les influences

marginales sur la probabilité en un point donné, par exemple au point moyen de la population; ce calcul n'est pas demandé ni effectué ici.

10°

Dependent Variable: REV
Method: Least Squares
Date: 06/04/07 Time: 15:43
Sample: 1 807
Included observations: 807

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-4372.294	2180.847	-2.004860	0.0453
CIGS	15.65748	21.92939	0.713995	0.4754
EDUC	834.7494	101.0701	8.259113	0.0000
AGE	685.1320	97.81030	7.004702	0.0000
AGE2	-7.598659	1.066990	-7.121586	0.0000

R-squared	0.169182	Mean dependent var	19304.83
Adjusted R-squared	0.165038	S.D. dependent var	9142.958
S.E. of regression	8354.492	Akaike info criterion	20.90516
Sum squared resid	5.60E+10	Schwarz criterion	20.93424
Log likelihood	-8430.233	F-statistic	40.82840
		Prob(F-statistic)	0.000000

On observe des coefficients bien estimés pour les variables educ, age et age2 ; le niveau d'étude ayant comme attendu une incidence positive sur le revenu, tandis que celle de l'âge traduit une coube en U renversée de maximum à 45 ans. La variable cigs qui mesure la consommation de tabac montre un signe contraire à l'intuition, mais est mal estimée comme le donne à penser son très mauvais ratio de Student.

De toute manière, on peut supposer que cette dernière variable est endogène, auquel cas la méthode des mco n'était pas appropriée et le test précédent sans valeur.

11° Écrivons de manière plus détaillée et sous une forme un peu différente (mais équivalente et qui simplifiera les calculs) le modèle structurel proposé :

$$rev = a_1 + a_2.cigs + a_3.educ + a_4.age + a_5.age2 + \varepsilon_{rev} \qquad (éq.6)$$

$$rev = b_1 + b_2.cigs + b_3.educ + b_4.age + b_5.age2 + b_6.nres + \varepsilon_{cigs} \qquad (éq.7)$$

et la forme réduite peut s'écrire a priori :

$$rev = \alpha_1 + \alpha_2.educ + \alpha_3.age + \alpha_4.age2 + \alpha_5.nres + \nu_{rev}$$

$$cigs = \beta_1 + \beta_2.educ + \beta_3.age + \beta_4.age2 + \beta_5.nres + \nu_{cigs}$$

L'identifiabilité d'une équation (structurelle) signifie la possibilité d'exprimer ses coefficients en fonction des coefficients de la forme réduite.

On connaît une condition nécessaire d'identifiabilité d'une équation structurelle : c'est que le nombre de variables manquantes soit égal au nombre d'endogènes moins un. La première équation structurelle satisfait ce critère et est donc peut-être identifiable. La seconde, n'ayant aucune variable manquante, est sous-indentifiée.

On peut en rester là, ou tenter « à la main » l'identification de la première équation. Pour ce faire, il faut exprimer la forme réduite en fonction de la forme structurelle ; cela peut être fait en éliminant respectivement rev et cigs entre les équations 6 et 7 :

$$cigs = [(a_1-b_1) + (a_3-b_3).educ + (a_4-b_4).age \ldots + (a_5-b_5).age2 - b_6.nres + aléa] / (b_2-a_2)$$

$$rev = [(b_2.a_1-a_2.b_1) + (b_2.a_3-a_2.b_3).educ + (b_2.a_4-a_2.b_4).age \ldots + (b_2.a_5-a_2.b_5).age2 - a_2.b_6.nres + aléa] / (b_2-a_2)$$

En identifiant les coefficients ci-dessus aux α_i et β_i, on obtient 10 équations dont les a_i et les b_i sont les 11 inconnues :

$$(a_1-b_1) / (b_2-a_2) = \alpha_1 \qquad (1')$$

$$(b_2.a_1-a_2.b_1) / (b_2-a_2) = \beta_1 \qquad (2')$$

$$\ldots \qquad \ldots$$

$$- b_6 / (b_2-a_2) = \alpha_5 \qquad (9')$$

$$- a_2.b_6 / (b_2-a_2) = \beta_5 \qquad (10')$$

En examinant la situation avec attention, on voit qu'on peut d'abord obtenir a_2 en utilisant les relations (9') et (10'), puis ensuite, en multipliant (1') par a_2, désormais connu, et en soustrayant (2'), on a l'heureuse surprise de voir disparaître b_1 et b_2 pour obtenir a_1, et les autres termes a_i s'obtiennent de manière

comparable, ce qui montre que la première équation structurelle est réellement identifiable.

Il n'y a en revanche aucun espoir d'obtenir les b_i…

12° L'estimation part les doubles moindres carrés de l'équation 6 corrige le problème rencontré en 10° : l'influence la variable cigs est maintenant significative et du signe voulu.

Dependent Variable: REV
Method: Two-Stage Least Squares
Date: 06/04/07 Time: 18:28
Sample: 1 807
Included observations: 807
White Heteroskedasticity-Consistent Standard Errors & Covariance
Instrument list: C EDUC PCIG AGE NRES AGE2

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-4596.170	3283.752	-1.399670	0.1620
CIGS	-659.1373	339.9929	-1.938680	0.0529
EDUC	516.4646	221.4447	2.332250	0.0199
AGE	1241.006	314.1063	3.950911	0.0001
AGE2	-14.06471	3.594452	-3.912894	0.0001

R-squared	-0.811712	Mean dependent var	19304.83
Adjusted R-squared	-0.820748	S.D. dependent var	9142.958
S.E. of regression	12337.06	Sum squared resid	1.22E+11
t		Second-stage SSR	5.55E+10

13° L'influence de l'âge sur le revenu se manifeste comme pour la consommation de tabac par une courbe en U renversé.

Le maximum de la fonction :

$$rev = -14.06471*age^2 + 1241.006*age + constante$$

est atteint pour :

$$age = 1241{,}006 / (2*14{,}06471) = 42{,}12 \text{ ans.}$$

(25 juin 2007)