Exercice I

I. Désirant étudier quelques déterminants du salaire dans le cadre de la théorie du capital humain, on dispose de données (fichier « travail85.sas », en mode ASCII, section *Exemples pour SAS*) concernant un échantillon représentatif de 534 travailleurs américains en 1985 :

LSAL	le logarithme népérien du salaire horaire moyen
ETU	le nombre d'années d'études
FEM	variable indicatrice du sexe féminin
MAR	variable indicatrice du statut marié avec cohabitation
EXP	le nombre d'année d'expérience professionnelle
SYN	variable indicatrice indiquant l'appartenance à un syndicat

1° Calculer la moyenne des variables quantitatives pour l'ensemble de l'échantillon considéré, puis pour le groupe des femmes et celui des hommes. Calculer la proportion de femmes, de personnes mariées cohabitant avec leur conjoint et de syndiqués pour l'ensemble de l'échantillon, puis les deux dernières proportions pour les femmes et pour les hommes.

2° Estimer par les mco le modèle simple :

$$(1) \qquad LSAL = \alpha + \beta.ETU + \varepsilon$$

Commenter brièvement les résultats obtenus, et indiquer quelle serait selon ce modèle l'influence d'une année d'étude supplémentaire.

3° Estimer par les mco le modèle :

$$(2) \qquad LSAL = \alpha + \beta.ETU + \gamma.FEM + \delta.EXP + \zeta.SYN + \varepsilon$$

Commenter les résultats obtenus, et indiquer quelle est selon ce modèle l'incidence sur le salaire du fait d'être une femme, celle d'être syndiqué ; combien d'années d'expérience professionnelle supplémentaires seraient nécessaires pour doubler son salaire.

4° On introduit la nouvelle variable : EXP2, égale au carré de EXP, et on cherche à perfectionner le modèle (2) par l'ajout des variables EXP2 et MAR. Estimer par les mco le modèle (3) ainsi défini et commenter les résultats obtenus.

Corrigé

I.1° Comme toujours, avant de mettre en œuvre des méthodes économétriques plus ou moins élaborées, il est souhaitable de connaître ses données et leurs caractéristiques élémentaires.

La variable LSAL désigne le logarithme népérien du salaire horaire (sur l'année) des individus membres du panel observé, la moyenne Moy(LSAL) = 2,059 est donc la moyenne sur le panel du logarithme du salaire horaire, mais c'est une erreur d'en déduire que son exponentielle Exp(2,059) = 7,84 est donc le salaire horaire moyen ; l'exponentielle de la moyenne n'est en effet pas la moyenne de l'exponentielle (cette propriété n'est vérifiée que par les fonctions linéaires, c'est à dire du premier degré).

Table of FEM by SYN

FEM (variable indicatrice du sexe féminin)
SYN (indicatrice de l'appartenance à un syndicat)

Frequency Percent **Row Pct** Col Pct	0	1	Total
0	221 41.39 **76.47** 50.46	68 12.73 **23.53** 70.83	289 54.12
1	217 40.64 **88.57** 49.54	28 5.24 **11.43** 29.17	245 45.88
Total	438 82.02	96 17.98	534 100.00

Les tris croisés sont une méthode élémentaire permettant de mettre en évidence les liaisons éventuelles entre deux variables qualitatives ou à items (ainsi le sexe et le fait d'être syndiqué ou non). Traditionnellement sont éditables pour chaque couple d'items possible, l'effectif croisé, la fréquence du couple croisé parmi l'ensemble, et les deux fréquences conditionnelles (ici syndicalisation conditionnée par le sexe et l'inverse), il convient d'identifier ces quatre quantités et de commenter celle qui présente un intérêt explicatif (ici la syndicalisation

conditionnée par le sexe, le choix inverse dépendant évidemment de la structure par sexe du panel).

I.2 On tente désormais d'expliquer la variable LSAL par des modèles économétriques de complexité allant croissant, estimés et testés sur les données fournies.

Il n'est pas inutile de se rappeler qu'il existe, pour simplifier, deux grands types de modèles : ceux destinés à des prévisions ou des simulations très précises, tels ceux mis au point et employés par les instituts de prévision, et ceux, plus académiques et moins exigeants quant à la précision de leurs prédictions, qui visent plutôt à éclairer la compréhension d'un domaine, à révéler les déterminants d'une variable et à les hiérarchiser, à préciser à l'inverse les facteurs de peu d'influence… Il est clair que les modèles qui vont être envisagés ici appartiennent à cette seconde famille, c'est fréquemment le cas en des modèles de type socio-économiques ou encore en sciences humaines, il est en effet illusoire d'espérer prévoir avec précision le salaire d'une personne à partir de quelques variables la décrivant !

Le modèle proposé : $LSAL = \alpha + \beta.ETU + \varepsilon$, est un modèle semi-logarithmique, l'une des variable (LSAL) étant en logarithme et l'autre (ETU) en niveau. Bien d'autres formulations fonctionnelles différentes pourraient *a priori* être imaginées et on peut se demander d'où provient celle-ci. Sans perdre de vue que les modèles retenus sont priés *in fine* de se conformer à la réalité des données observées, ils peuvent être suggérés sinon imposés par la théorie (ou les pratiques antérieures) ou résulter simplement de divers tâtonnements empiriques. Dans le problème, ces étapes antérieures sont supposées achevées et la formulation semi-logarithmique ci-dessus n'est plus remise en cause.

On ne présente pas les résultats numériques de cette question.

I.3 L'équation estimée du modèle proposée peut s'écrire par exemple :

```
LSAL = 0,972.ETU - 0,229.FEM + 0,0118.EXP + 0,210.SYN + 0,651
       (12,31)     (-5,81)     (6,97)       (4,09)      (5,48)
```

la valeur estimée éditée du coefficient de EXP étant 0,01176. En utilisant la forme multiplicative du modèle en passant par l'exponentielle, on voit qu'une personne dont le salaire suivrait exactement cette loi et dont les autres caractéristiques ne changeraient pas pendant la période verrait chaque année d'expérience professionnelle supplémentaire le logarithme de son salaire

augmenté de 0,01176, ou celui-ci multiplié par Exp(0,01176) soit 1,01183, c'est à dire augmenté d'environ 1,2%.

Le nombre N d'années nécessaires pour doubler son salaire, ou aussi bien augmenter son logarithme : LSAL, de log2 est donc simplement N = log2/0,01176 = 58,94, soit environ 59 ans.

On peut juger cette valeur peu crédible, mais elle n'est qu'une valeur théorique et dépend des hypothèses faites, la plus importante et la plus discutable étant la forme de l'influence de l'expérience, qui est précisément mise en cause dans la question suivante…

I.4 On commente donc la démarche et les résultats numériques complets uniquement pour le dernier modèle proposé, à la suite d'une démarche ascendante procédant par ajouts successifs de variables (une *stepwise regression* à la main, en quelque sorte…) de la question I.2 à la question I.4.

Comme toujours dans une telle procédure, l'acceptation d'un modèle plus complexe met en doute la pertinence de ceux qui le précédaient, c'est la démarche habituelle, comme dans la théorie des tests qui en est l'un des outils, il n'y a pas lieu de s'en émouvoir…

Il ne faut enfin pas oublier, dans un tel type d'étude, qu'après la phase technique, le commentaire doit essentiellement viser à donner la signification, l'interprétation économique, des faits mis en évidence par l'économétrie ; les propos sans rapport avec ces derniers ne sont pas nécessairement dépourvus d'intérêt mais n'ont pas leur place à cet endroit.

On commence par examiner brièvement R^2 et F, plus par tradition que pour leur intérêt véritable.

Ici, le coefficient de détermination : $R^2 = 0,32$, quoique plus élevé que dans les régressions précédentes reste relativement bas et aurait été considéré comme « mauvais » par les statisticiens d'autrefois. En dépit de la formulation traditionnelle selon laquelle « 32% de la variable LSAL est expliquée par le modèle », on sait qu'il est de peu de signification. Un modèle donnant un R^2 très proche de 1 peut être parfaitement illusoire et résulter par exemple de la simple présence d'un trend commun, tandis qu'un modèle de faible R^2 peut à l'inverse refléter une liaison réelle, mais affectée d'un fort aléa… De modestes R^2 - jusqu'à inférieurs à 0,10 ! - peuvent se rencontrer dans les modèles de type socio-économiques par exemple, et ne doivent pas décourager l'économètre.

The REG Procedure

Model: MODEL1
Dependent Variable: LSAL logarithme népérien du salaire horaire moyen

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	47.31747	7.88625	41.10	<.0001
Error	527	101.12442	0.19189		
Corrected Total	533	148.44189			

Root MSE	0.43805	R-Square	0.3188
Dependent Mean	2.05918	Adj R-Sq	0.3110
Coeff Var	21.27297		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	constante	1	0.57189	0.11825	4.84	<.0001
ETU	nombre d'années d'études	1	0.09055	0.00790	11.47	<.0001
FEM	indicatrice du sexe féminin	1	-0.23075	0.03872	-5.96	<.0001
EXP	nombre d'années d'expérience	1	0.03287	0.00561	5.85	<.0001
EXP2	carré de EXP	1	-0.00049838	0.00012111	-4.12	<.0001
SYN	indicatrice de l'appartenance à un syndicat	1	0.19945	0.05055	3.95	<.0001
MAR	indicatrice du statut marié avec cohabitation	1	0.04151	0.04266	0.97	0.3310

L'excellent test F, de significativité globale de la régression, ici F = 41,10 soit un niveau de risque inférieur à 0,01%, doit également être reçu avec modestie, on sait en effet qu'hormis en des exemples artificiels construits exprès pour cela, c'est quasiment toujours le cas.

N'ayant pour l'occasion plus d'indicateurs globaux à examiner (telle une éventuelle statistique de Durbin-Watson en présence de données temporelles), on passe à l'étape essentielle : aux variables explicatives, à leurs coefficients estimés et aux statistiques qui les accompagnent.

Une bonne manière de procéder est d'examiner en premier lieu les ratios de Student, de manière à identifier les variables significatives et celles qui ne le paraissent pas ou peu, cela sans accorder une valeur excessive au risque 5% et au seuil, approximativement égal à 2 en valeur absolue, pour les t-ratios.

A cette aune, on observe que toutes les variables paraissent intervenir significativement à l'exception de MAR (le statut marié ou équivalent avec

cohabitation). Tous les autres t-ratios excédant très largement 2, il n'y a pas lieu d'établir une hiérarchie entre les autres variables (ou encore entre des certitudes au risque 1 pour 10.000 et 1 pour 100.000 !)

En accordant naturellement plus d'importance aux variables significatives, on examine ensuite les signes des estimations, regardant en particulier s'ils vont dans le sens des idées que l'on pouvait avoir a priori ou à l'inverse les contredisent (ce qui est heureusement moins fréquent).

Suivant ce programme, on peut d'abord signaler que la faible significativité de MAR s'accorde bien avec l'embarras dans lequel on aurait été pour prévoir l'influence de cette variable (la vie en couple produit généralement des enfants qui peuvent diminuer la productivité au travail et donc le revenu d'un au moins des ascendants, mais à l'inverse les employeurs préfèrent les individus plus stables, tels ceux engagés dans une famille, etc. Que conclure ? Rien justement, c'est ce que montrent les chiffres…)

Quant aux autres variables, à l'exception de EXP2 dont le signe pourrait troubler un économètre peu averti mais dont on va reparler, toutes les influences sont conformes à ce que l'on pouvait attendre, la condition féminine est un élément défavorable au salaire, contrairement à la durée des études, à l'expérience professionnelle et à l'adhésion à un syndicat.

L'étape suivante s'intéresse enfin à la valeur elle-même des coefficients. On rappelle avant toute chose qu'on ne peut comparer directement en intensité que des coefficients associés à des variables homogènes, faute de quoi ils sont mesurés en des unités différentes et on peut aisément faire varier l'un d'eux d'un facteur arbitraire par un simple changement d'unité. Ajoutons enfin que la grandeur LSAL étant un logarithme et les explicatives des niveaux ou des dummy, les coefficients mesurent non des élasticités, mais l'influence d'une variation unitaire de chacune de ces variables sur le logarithme du salaire.

Avec ces restrictions, les coefficients -0,231 et 0,199 des dummy variables associées au sexe féminin et à l'appartenance syndicale montrent que ces effets sont opposés et du même ordre de grandeur sur LSAL ; si l'on souhaite exprimer cela d'une manière plus parlante par l'effet sur le salaire proprement dit, on passe à nouveau en exponentielle : le salaire est multiplié par Exp(-0,23075) soit environ 0,794, ou encore baisse d'à peu près 20%, si l'on est une femme (toutes choses égales par ailleurs), et par Exp(0,19945) si l'on est syndiqué, c'est à dire augmente d'environ 22%.

Par parenthèse, si l'on avait voulu malgré tout accorder quelque crédit à l'estimation de faible qualité du coefficient de la troisième dummy, MAR, on

aurait dit que son effet est positif et de l'ordre de 5% sur le salaire (Exp(0,04151) = 1,04238).

Bien qu'homogènes au sens physique du terme, car toutes les deux temporelles et mesurées en années, les variables ETU et EXP ne le sont pas tout à fait quant à leurs échelles respectives, la durée des études excédant rarement 15 ans, alors qu'un carrière laborieuse peut durer 50 ans et plus, pour cette raison, sans la présence du terme EXP2, il serait sage de se limiter à calculer comme ci-dessus l'incidence d'une année d'étude et d'une année d'expérience professionnelle de plus.

En ce qui concerne le terme EXP2, l'introduction de ce terme carré destiné à affiner l'influence de l'expérience professionnelle proposée dans le modèle du I.3, peut donner lieu à deux contresens.

Le premier consiste à juger le coefficient estimé : -0,0000498, très faible et donc cette influence négligeable. C'est oublier l'hétérogénéité des variables et comparer EXP2 à tort à d'autres d'ordres de grandeur différents. EXP2 est non un temps mais un temps au carré (de même qu'un mètre carré n'est pas un mètre, même s'il est plus aisément concevable qu'une année au carré…), et si EXP peut raisonnablement atteindre 40 voire 50 ans, EXP2 vaut alors 1600 et 2500 années au carré, et le terme correspondant de l'équation n'a plus rien de négligeable devant les autres.

Le phénomène est usuel : lorsqu'apparaissent des variables à croissance rapide (carrés, cubes, exponentielles, etc.) ou simplement pouvant prendre de grandes valeurs, il est normal que leurs coefficients soient relativement faibles, sinon leur effet écraserait les autres.

La seconde erreur est de s'étonner du signe négatif du terme EXP2. Au contraire, supposant les variables autres que EXP et EXP2 fixées, la grandeur LSAL est une fonction du second degré de l'expérience, ou graphiquement une parabole, et le terme du second degré, le coefficient de EXP2, étant négatif c'est une parabole à la concavité tournée vers le bas ; si l'on est, comme il conviendrait de le vérifier, dans sa partie montante, la croissance est de plus en plus lente, il s'agit tout simplement d'un phénomène classique de *rendement décroissant* (de l'expérience sur le logarithme du salaire précisément), et il n'apparaît pas nécessaire d'invoquer la démotivation des vieux travailleurs, ou encore leurs difficultés à assimiler les nouvelles technologies...

Pour terminer, sous une forme plus littérale l'équation estimée s'écrit donc

```
LSAL = 0,572 + 0,0906.ETU - 0,231.FEM + 0,0329.EXP - 0,000498.EXP2 + 0,199.SYN + 0,0415.MAR
       (4,84)  (11,47)      (-5,96)     (5,85)       (4,12)          (3,95)      (0,97)
```

en figurant les ratios de Student sous chaque coefficient estimé.

Tableau récapitulatif

*	Constante	ETU	FEM	EXP	EXP2	SYN	MAR	R-deux	SCR
(1)	1.05986 9.87***	0.07676 9.49***						0.1447	126.960
(2)	0.65123 5.48***	0.09721 12.31***	-0.22881 -5.81***	0.01176 6.97***		0.21013 4.09***		0.2921	105.085
(3)	0.57189 4.84***	0.09055 11.47***	-0.23075 -5.96***	0.03287 5.85***	-0.00049838 -4.12***	0.19945 3.95***	0.04151 0.97	0.3188	101.124

(*, ** et *** : Student significatifs à 10%, 5% et 1%)

----===ooOoo===-----

(27.05.2004)