FONCTIONS DE GAINS : une approche non paramétrique

Michel Simioni

Introduction

La mesure des rendements de l'éducation peut faire l'objet d'une estimation directe par l'intermédiaire d'une fonction de gains. Celle-ci consiste en la régression du logarithme du salaire perçu par un individu sur un ensemble de variables qui caractérisent ce dernier et dont on pense qu'elles ont un impact sur la détermination des revenus (ancienneté dans l'entreprise, nombre d'années d'études initiales, expérience professionnelle ...). Il s'agit alors de dégager les profils âge-gains selon la formation des individus. L'objet du travail présenté ici est l'estimation d'une telle fonction à partir de données individuelles extraites de l'enquête Formation Qualification Professionnelle effectuée par l'Insee en 1993. Les résultats obtenus à partir d'estimations paramétriques et non paramétriques sont ainsi comparés et commentés en vue de dégager les profils évoqués ci-dessus.

Une première section présente deux approches possibles quant à la mesure des rendements de l'éducation : l'approche classique basée sur une spécification paramétrique de la fonction de gains, et l'approche que nous nous proposons d'utiliser ici et qui repose sur l'utilisation de techniques non paramétriques en vue d'estimer cette fonction. Cette section introduit les principaux concepts utilisés dans la littérature économétrique sur la mesure des rendements de l'éducation. Une deuxième section, consacrée aux données dont nous disposons, décrit les variables que nous avons utilisées. Les résultats des estimations paramétriques et non paramétriques sont présentées et comparés dans la dernière section.

Mesure des rendements de l'éducation

L'approche classique

La littérature économétrique sur les déterminants du salaire d'un individu repose essentiellement sur la spécification d'un modèle de régression du type suivant[1] :

$$\ln w_i = m(s_i, x_i, z_i) + \varepsilon_i$$

où w_i représente le salaire perçu par l'individu i, s_i est une mesure du capital scolaire, x_i est une mesure de l'expérience professionnelle, et z_i est un vecteur d'autres variables telles que le sexe de l'individu, son ancienneté dans l'entreprise, qui peuvent avoir un impact sur le salaire d'un individu. Ce modèle de régression est appelé fonction de gains. Une forme fonctionnelle paramétrique pour $m(.,.,.)$ peut être déduite dans le cadre de la théorie du capital humain (voir BECKER [1964]) moyennant certaines hypothèses. Ainsi supposons que le rendement de l'éducation puisse être mesuré en termes de différence de salaires. Le rendement provenant d'une année d'éducation, noté r_1 peut alors être calculé de la sorte :

$$r_1 = \frac{w_1 - w_0}{w_0}$$

où w_1 est le salaire résultant d'une année d'éducation et w_0 est le salaire pour aucune année d'éducation. D'une façon similaire, définissons alors le rendement d'une s ième année d'éducation comme $r_s = (w_s - w_{s-1}) / w_{s-1}$, alors on peut exprimer le salaire perçu après S années d'éducation de la sorte :

$$w_s = w_0 (1 + r_1) \times (1 + r_2) \times \ldots \times (1 + r_s)$$

Sous l'hypothèse que le taux de rendement de l'éducation est le même quel que soit le nombre d'années, soit $r_1 = r_2 = \ldots = r_s = r$, et si l'on utilise e^r comme approximation de $(1 + r)$, l'équation précédente devient :

$$w_s = w_0 \ e^{r \times s}$$

qui, après linéarisation et addition d'un terme d'erreur ε, s'écrit :

$$\ln w_s = \ln w_0 + r \times s + \varepsilon$$

1. Cette section reprend en partie la présentation de ce thème telle qu'elle est faite dans l'ouvrage de Berndt (1991).

Cette équation est la forme la plus simple de fonction de gains rencontrées dans la littérature. Elle peut être aisément estimée par une technique du type moindres carrés si l'on dispose d'observations sur les salaires et le nombre d'années d'études dans une population d'individus. La valeur estimée du coefficient de *S* fournit ainsi une mesure du rendement de l'éducation. Quant au terme constant, sa valeur estimée indique quel est le gain d'un individu n'ayant pas réalisé d'études. Remarquons alors que la spécification proposée ci-dessus n'incorpore aucune autre variable comme déterminant du salaire. Elle a ainsi été généralisée par MINCER (1974) en vue d'y incorporer les effets possibles de l'expérience acquise par les individus lors de leur vie professionnelle. Une des formalisations alors retenues de la fonction de gains s'écrit :

$$\ln w_i = \ln w_0 + \beta_1 \; Si + \beta_2 \; x_i + \beta_3 \; x_i^2 + \varepsilon_i$$

Cette spécification paramétrique est celle qui est le plus souvent utilisée dans la littérature économétrique sur la mesure des rendements de l'éducation (voir, entre autres, BOUHMADI et PLASSARD [1992)]). Elle est toujours supposée être une fonction linéaire du nombre d'années d'études S_i, mais elle est aussi supposée être une fonction quadratique du nombre d'années d'expérience x_i. Il s'agit alors de capturer une des suggestions de la théorie du capital humain selon laquelle un individu a une plus forte incitation à s'investir dans son travail les années qui suivent sa sortie de la scolarité et à acquérir ainsi un plus fort capital dû à l'expérience. De plus, toujours dans ce cadre théorique, cette incitation tend à décroître avec l'âge. Ainsi, la fonction de gains devrait être concave par rapport à l'expérience, ou encore, la valeur estimée de β_2 devrait être positive et celle de β_3 négative. Remarquons qu'on peut alors calculer le nombre d'années d'expérience pour lequel le gain est maximal, comme étant la valeur estimée de $-\beta_2 / \beta_3$.

Une approche non paramétrique

Quoique faisant appel au cadre formalisé de la théorie du capital humain, la littérature économétrique sur la mesure des rendements de l'éducation est une littérature essentiellement empirique basée sur des formalisations paramétriques de la fonction de gains telles que celles présentées ci-dessus. De telles spécifications peuvent être aisément estimées par des techniques du type moindres carrés. Remarquons maintenant que la formalisation générale d'une fonction de gains donnée ci-dessus peut être écrite comme un modèle de régression général où si l'on suppose que, conditionnellement aux (s_i, x_i, z_i), $i = 1, \ldots, n$, les termes d'erreur β_i sont des variables indépendantes de moyenne nulle,

$$m(s_i, x_i, z_i) = E\,(\ln w_0 \mid (s, x, z) = (s_i, x_i, z_i))$$

Cette espérance conditionnelle peut être estimée en ayant recours aux techniques d'estimation paramétrique lorsqu'on suppose que

$$E(\ln w \mid (s, x, z) = (s_i, x_i, z_i)) = \ln w_0 + \beta_1 si + \beta_2 xi + \beta_3 x_i^2$$

Mais, dans une perspective de trouver une certaine structure dans un jeu de données sans avoir recours à une hypothèse paramétrique du type précédent, on peut aussi utiliser les techniques d'estimation non paramétrique en vue d'estimer cette espérance conditionnelle. Ici, nous proposons d'utiliser la technique dite des polynômes locaux (voir FAN et GIJBELS [1995]) en vue d'estimer dans une première étape la relation existant entre le logarithme du salaire et l'expérience et de vérifier si, comme le voudrait la théorie du capital humain, cette relation est concave. Dans une deuxième étape, nous utiliserons aussi cette technique d'estimation en vue de dégager divers profils quant à cette relation selon le niveau d'étude, le sexe .... Ce second point sera précisé ci-dessous lors de la présentation du jeu de données dont nous disposons.

Description des données

Les données utilisées ici proviennent de l'enquête *Formation Qualification Professionnelle* réalisée par l'*Insee* en 1993. D'un échantillon comprenant un peu plus de 18 000 individus nous avons extrait tous les individus qui étaient salariés à temps plein au moment de l'enquête et dont le salaire mensuel n'excéde pas 80 000 francs[1]. A l'issue de cette extraction, nous disposons d'un échantillon de 7 948 individus pour lesquels nous avons construit les différentes variables présentées dans le *tableau 1*. Pour construire ces différentes variables, nous nous sommes inspirés du travail de BLANC et LAGRIFFOUL (1995).

Tableau 1

Dictionnaire des variables

Salaire :	salaire mensuel perçu par l'individu
Expérience :	nombre d'années depuis le début du premier emploi
Ancienneté :	nombre d'années dans le dernier emploi
Dip1 :	aucun dipôme ou CEP
Dip2 :	BEPC, CAP, BEP, brevets
Dip3 :	baccaulauréat, brevets professionnels
Dip4 :	diplôme supérieur au bac ou équivalent
Pubpri :	emploi dans le public ou le privé
Sexe :	sexe de l'individu

1. Seuls trois individus salariés à plein temps présentent un salaire mensuel largement supérieur à cette valeur.

Les variables que nous utilisons, sont de deux types : *i.)* continue : salaire, expérience et ancienneté mesurées en années (quelques caractéristiques quant à leurs distributions empiriques sont données dans le tableau 2), et *ii.)* discrète : le type de diplôme possédé par l'individu comme mesure de son niveau d'éducation, la nature de l'emploi (public, privé) et le sexe (les effectifs correspondant aux modalités de ces variables sont récapitulés dans le tableau 2bis). Ces variables qualitatives vont définir des sous échantillons pour lesquels il sera possible d'estimer la relation existant entre le logarithme du salaire et l'expérience. Divers profils quant à cette relation seront ainsi obtenus et comparés. Remarquons ainsi que nous ne disposons pas d'une mesure continue quant au niveau d'éducation d'un individu. Seule une mesure discrète est disponible : l'indication du diplôme le plus élevé obtenu par l'individu (voir les variables *Dip1* à *Dip4*). Les écarts entre gains estimés que l'on observera selon le niveau de diplôme donneront ainsi une mesure des rendements de l'éducation.

Tableau 2

Variables continues

Variables	Moyenne	Ecart-type	Minimum	Maximum
Salaire	9437.97	5458.56	3766.67	77583.33
Expérience	19.80	11.10	0.00.	49.00
Ancienneté	11.85	9.71	0.00	46.00
Taille de l'échantillon : 7948 salariés				

Tableau 2bis

Variables qualitatives

Variables	nombre d'observations
Dip1	2032 individus
Dip2	3356 individus
Dip3	924 individus
Dip4	1636 individus
Pubpri	public : 2638 ind., privé : 5310 ind.
Sexe	femme : 3109 ind., homme : 4839 ind.
Taille de l'échantillon: 7948 salariés	

Gains et expérience : relation quadratique ?

La relation existant entre le logarithme du salaire mensuel et l'expérience est-elle quadratique ? Pour répondre à cette question, deux types de résultats sont comparés :

i.) Les premiers proviennent de la régression par moindres carrés ordinaires du logarithme du salaire mensuel sur l'expérience et l'expérience élevée au carré. Ils sont présentés dans le tableau 3 dans les colonnes correspondant au modèle restraint, i.e.,

$$\ln w_i = \beta_0 + \beta_1 \exp_i + \beta_2 \exp_i^2 + \varepsilon_i$$

Remarquons premièrement que ces deux variables n'expliquent qu'une faible partie de la variabilité observée des logarithmes des salaires mensuels. En effet, le coefficient de détermination ajusté pour tenir compte du nombre de variables, ou $\overline{R}^2$, est très petit, 0.063. Quant aux coefficients des deux variables explicatives, ils sont significativement différents de zéros et possèdent les signes attendus ; ainsi, le coefficient de l'expérience est positif et celui de l'expérience élevée au carré est négatif, quoique relativement petit.

Tableau 3

Modèles paramétriques

	Modèle restraint		Modèle général	
Variables	**Coef. est.**	**Ecart-type**	**Coef. est.**	**Ecart-type**
Constante	8.7356	0.0147	8.6785	0.0200
Exp	0.0287	0.0015	0.0261	0.0012
Exp^2	-0.0005	0.0000	-0.0004	0.0000
Anc.	...	...	0.0086	0.0005
Dip2	...	...	0.1761	0.0094
Dip3	...	...	0.3672	0.0135
Dip4	...	...	0.6519	0.0116
Pubpri	...	...	0.0363	0.0083
Sexe	...	...	-0.2149	0.0076
R^2 ajusté	0.063		0.407	

ii.) Le second type de résultats présentés provient de la régression non paramétrique du logarithme du salaire mensuel sur l'expérience[1]. En vue d'obtenir un intervalle de confiance uniforme en chacun des points où est effectué la régression, nous utilisons ici l'estimateur de Nadaraya et Watson. Cet estimateur correspond au cas particulier dans le cadre de la méthode des polynômes locaux où l'ordre du polynôme est $p = 0$. Il s'agit là du seul estimateur de type polynômes locaux pour lequel il est actuellement possible de construire ce type d'intervalle de confiance. Le noyau utilisé est le noyau quadratique dont la formule est :

$$K(u) = \frac{15}{16} (1 - u^2)^2 \; 1(|u| < 1)$$

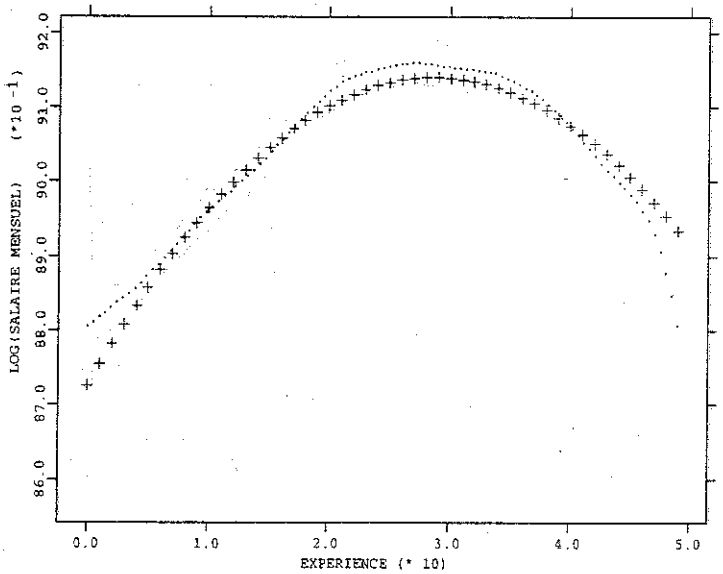
Quant à la fenêtre, elle est choisie par validation croisée : la valeur ainsi obtenue est égale à 5.7392. L'estimation de l'espérance conditionnelle est réalisée pour 100 valeurs équiréparties entre le minimum et le maximum observés de la variable *expérience* (ces

1. Toutes les estimations présentées ici ont été réalisées avec le logiciel XploRe (voir Härdle et al. (1995)).

valeurs sont notées $\overset{o}{X}_j$, $j = 1, ..., 100$)). Les valeurs estimées obtenues et les intervalles de confiance correspondants sont représentés sur la figure 1 : les premières par des points, les seconds par les deux lignes continues. Remarquons que la courbe obtenue correspond aussi à une fonction concave de l'expérience.

Figure 1

Relation quadratique



La comparaison des résultats des estimations paramétrique et non paramétrique peut alors être faite graphiquement en portant sur la figure 1 les valeurs estimées à partir de l'estimation paramétrique, du logarithme du salaire mensuel pour les valeurs des $\overset{o}{x}_j$, $j = 1, ..., 100$) ou

$$\ln \hat{w}_j = \hat{\beta}_o + \hat{\beta}_i x_j^o + \hat{\beta}_2 (x_j^o)^2$$

Ces valeurs sont représentée par des $+$. Remarquons alors que celles-ci appartiennent aux intervalles de confiance obtenues à partir de la régression non paramétrique. Cette constatation indique que le choix de la spécification paramétrique exprimant le logarithme du salaire mensuel comme une fonction quadratique de l'expérience "semble" correct. Remarquons qu'une telle conclusion visuelle pourrait faire l'objet d'un test statistique formel du type de celui développé par HÄRDLE et MAMMEN (1993).

Gains et expérience : divers profils

Nous allons maintenant nous intéresser aux profils gains/expérience en fonction du niveau de formation du salarié, du caractère public ou privé de son secteur d'activité et de son sexe. Pour cela, nous avons estimé en utilisant la méthode des polynômes locaux l'espérance conditionnelle du logarithme du salaire mensuel sachant le nombre d'années d'expérience pour chacune des sous-populations définies par les modalités des variables *Dip*, *Pubpri* et *Sexe*. D'un point de vue technique, notons que nous avons choisi comme ordre pour le polynôme local $p = 3$. Un tel choix peut, en effet, permettre de mieux capturer la courbure de l'espérance conditionnelle et donc sa possible concavité. Les fenêtres ont été choisies par validation croisée. Notons que le critère à minimiser comme fonction de la fenêtre ne fait intervenir que l'espérance conditionnelle et pas ses dérivées. Avant d'entrer dans le commentaire des résultats de ces estimations (voir les figures 2, 3 et 4), précisons que nous avons aussi estimé le modèle paramétrique le plus souvent utilisé pour analyser de tels profils, i.e.,

$$\ln wi = \beta_0 + \beta_1 \exp_i + \beta_2 \exp_i^2 + \beta_3 anc_i$$
$$+ \beta_4 Dip2_i + \beta_5 Dip3_i + \beta_6 Dip4_i$$
$$+ \beta_7 Pubpri_i + \beta_8 sexe_i + \varepsilon_i$$

Les résultats de l'estimation de ce modèle sont reportés dans le tableau 3 aux colonnes correspondant au modèle général.

En fonction du diplôme

Les valeurs estimées des coefficients des trois variables *Dip2*, *Dip3* et *Dip4* dans le modèle général indiquent que le passage du niveau de diplôme représenté par la variable *Dip1*, i.e., aucun diplôme ou CEP, à l'un des trois autres niveaux de diplôme entraîne des gains significatifs en termes de salaire. Une telle conclusion peut aussi être faite en regardant les quatre courbes tracées sur la figure 2. Elles correspondent aux estimations non paramétriques de l'espérance conditionnelle du logarithme du salaire sachant le niveau d'expérience pour les quatre sous populations définies par le type de diplômes qui y est possédé : ainsi, + correspond à *Dip1*, X à *Dip2*, ° à *Dip3* et * à *Dip4*. Notons aussi qu'à la vue de la figure 3, il apparaît que le niveau de diplôme semble être la variable qui capture le mieux la variabilité existant entre les salaires mensuels observés. L'évolution du $\overline{R}^2$ entre le modèle paramétrique restreint et le modèle paramétrique général va aussi dans le sens d'une telle conclusion.

Les gains en termes de salaire mis en évidence ci-dessus peuvent être calculés pour différentes valeurs de l'expérience. Le tableau 4 contient de telles mesures pour 5, 10, 20 et 30 années d'expérience. La comparaison des valeurs qui y sont reportées pour les deux types d'estimation, met en évidence le fait que la spécification paramétrique tend à surévaluer ces gains par rapport aux mesures de ceux-ci obtenues à partir des estimations non paramétriques pour les petites valeurs de l'expérience et à les sous-évaluer pour de grandes valeurs de cette dernière.

Tableau 4

Quelques mesures non paramétriques du rendement de l'éducation

Expérience	Référence	Dip2	Dip3	Dip4
5 années	5767.36	689.92	1466.97	3902.30
	6621.26	*1234.44*	*2938.24*	*6084.69*
10 années	6232.25	754.23	1876.76	4942.49
	7300.22	*1370.94*	*3293.54*	*6708.69*
20 années	6787.15	1727.15	3284.70	7095.25
	8311.13	*1560.79*	*3688.14*	*7636.68*
30 années	7425.86	2226.65	5612.74	8307.05
	8760.14	*1628.21*	*3847.45*	*7967.61*
Référence = salaire mensuel estimé pour Dipl, En italiques, résultats obtenus à partir de l'estimation paramétrique				

En fonction du secteur, du sexe

Comme pour le niveau de diplôme, les valeurs estimées des coefficients des variables *Pubpri* et *Sexe* dans le modèle général indiquent que, quoique le coefficient de *Pubpri* soit significativement différent de zéro, sa valeur est relativement faible. Il n'existe pas de différences importantes entre salaires dans le public et le privé et ceci, quel que le nombre d'années d'expérience. Par contre, une différence significative semble exister entre les salaires pour les hommes (modalité de référence de la variable *Sexe* lors de l'estimation du modèle général) et les salaires pour les femmes. De plus, le signe du coefficient indique que cette différence joue en défaveur de celles-ci.

Considérons maintenant les résultats des estimations non paramétriques. Ces résultats sont présentés sur les figures 2 et 3 où X correspond au privé (fig. 2) ou aux hommes (fig. 3) et + au public (fig. 2) ou aux femmes (fig. 3). A la vue de ces deux figures, aucune différence n'apparaît entre les estimations de l'espérance conditionnelle du logarithme du salaire sachant le niveau d'expérience pour les deux sous-populations définies par les modalités de la variable *Pubpri* (voir fig. 3). Par contre, les deux courbes obtenues pour les deux modalités de la variable *Sexe* ne sont pas confondues. Il existe ainsi une

différence entre les deux estimateurs de l'espérance conditionnelle du logarithme du salaire mensuel sachant le niveau d'expérience obtenus pour les deux modalités de cette variable.

Figure 2

Profils Formation

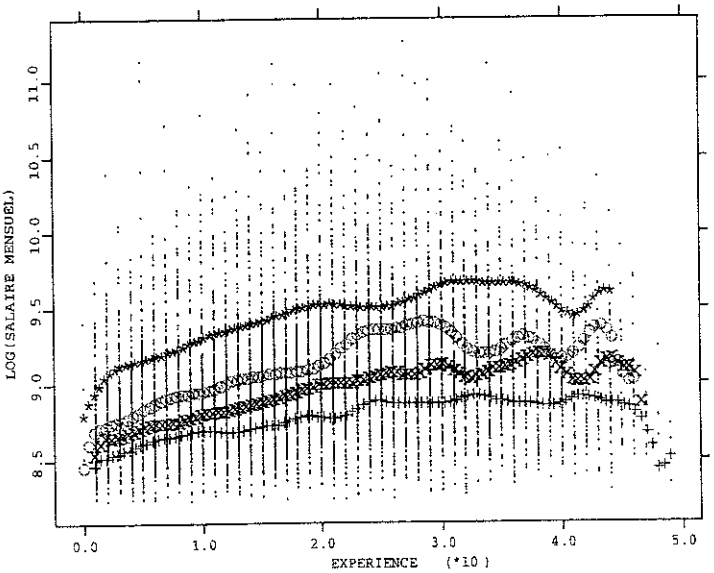


Figure 3

Profils Secteur

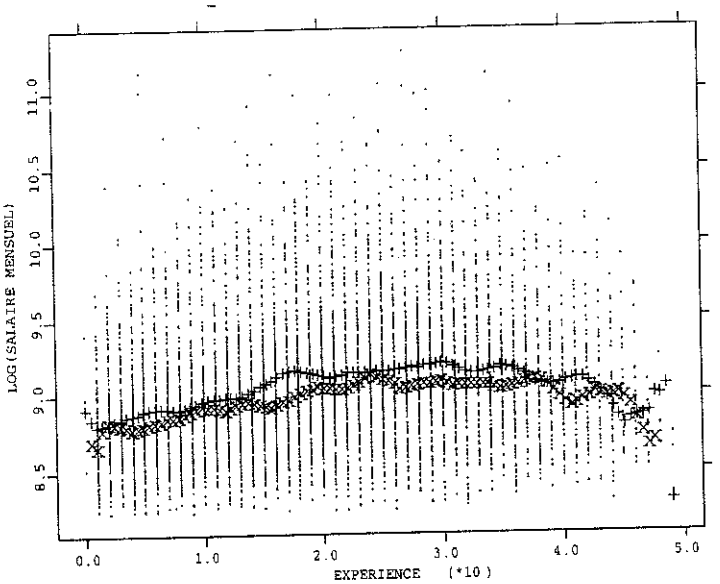
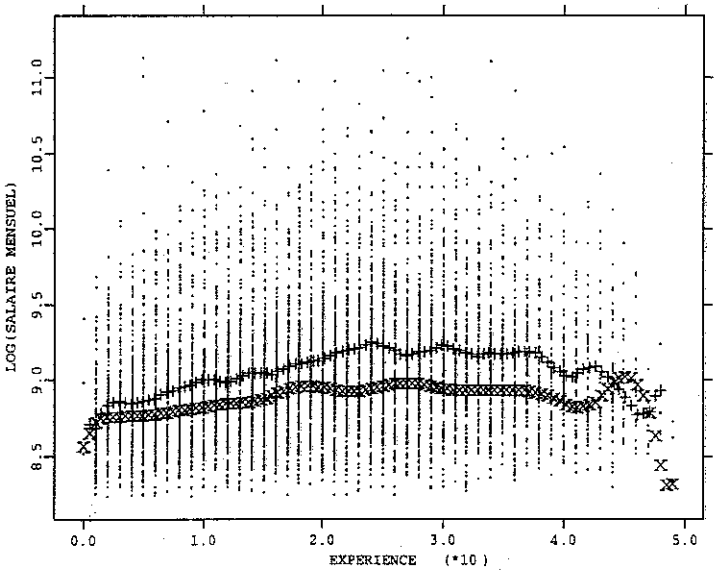


Figure 4

Profils Sexe



Les deux types de résultats, paramétriques et non paramétriques, sont ainsi cohérents. Comme ci-dessus, il est possible de mesurer les gains en termes de salaires dus au passage de la modalité public (resp. femme) de la variable *Pubpri* (resp. *Sexe*) à la modalité privé (resp. homme) de la même variable. Du fait du biais mis en évidence ci-dessus pour les mesures obtenues à partir des estimations paramétriques, nous ne reportons ici que les mesures obtenues à partir des estimations non paramétriques (voir tableau 5). Plus précisèment, ce tableau donne quelques statistiques quant à ces gains.

Tableau 5

Mesures nonparamétriques des écarts de salaires selon le secteur et le sexe

Variable	Moyenne	Ecart-type	Minimum	Maximum
Pubpri	659.69	584.29	-1400.57	2152.78
Sexe	1430.07	915.70	-1534.99	3034.77

BIBLIOGRAPHIE

BECKER G.S. (1964), *Human Capital : A Theoretical and Empirical Analysis, with Special Reference to Education*, National Bureau of Economic Research, New-York.

BERNDT E.R. (1991), *The Practice of Econometrics*, Addison-Wesley Publishing Company.

BLANC M. et LAGRIFFOUL **C.** (1995), "Mobilité et marchés du travail ruraux : une approche en termes de segmentation", à paraître dans *Revue d'Economie Régionale et Urbaine*.

BOUHMADI R. et PLASSARD J.M. (1992), "Note à propos du caractère endogène de la variable éducation dans la fonction de gains", *Revue Economique*, 43 (1), pp.145-156.

FAN J. et GIJBELS I. (1995), *Local Polynomial Modeling and Its Application - Theory and Methodologies*, Chapman et Hall.

HÄRDLE W. (1990), *Applied Nonparametric Regression*, Cambridge University Press.

HÄRDLE W. et MAMMEN E. (1993), Comparing Nonparametric versus Parametric Regression Fits, *Annals of Statistics*, 21 (4), 1926-1947.

HÄRDLE W., KINKLE S. et TURLACH B.A. (1995), *XploRe : An Interactive Statistical Computing Environment*, Springer-Verlag.

MINCER J. (1974), *Schooling, Experience and Earnings*, Columbia University Press.