École nationale
de la statistique
et de l'administration
économique

université
PARIS-SACLAY

Timothé KRAUTH, Johan MACQ

Estimation des rendements privés et sociaux de l'enseignement supérieur

1 Rendements privés de l'éducation et équation de Mincer

Avant propos :

Pour cette première partie nous avons réalisé plusieurs transformations sur les variables d'intérêt des bases de données fournies. Premièrement, nous avons sélectionné les variables suivantes :

- annee : année de l'enquête
- actop : statut actif (1) ou non (2)
- salmee : salaire mensuel
- salmet : salaire mensuel en tranche
- dip : diplôme obtenu
- datgen, datsup, dattec, datdip : année d'obtention du diplôme correspondant
- naia : année de naissance
- sexe : sexe
- age : age

Puisqu'on s'intéresse à l'effet de l'éducation sur le salaire, nous avons sélectionné uniquement les individus déclarés actifs. Ensuite, la variable catégorielle "dip" a été modifiée pour faire correspondre au diplôme le nombre d'années d'études (après le brevet).
La variable "datediplome" a été créée, elle correpond à la date du diplôme le plus élevé obtenu : max(datgen,datsup,dattec) ou directement datdip. Lorsque l'individu n'a pas renseigné de diplôme, nous avons considéré que sa date de "diplomation" était son année de naissance + 16 ans. Nous faisons donc l'hypothèse qu'il n'est pas possible de travailler avant 16 ans.
En soustrayant à l'année de l'enquête la date de diplomation de l'individu, nous avons créé la variable "experience" qui représente l'expérience potentielle.

Les données utlisées proviennent des résultats des enquêtes de l'INSEE sur la période 2003-2014.

Question 1)

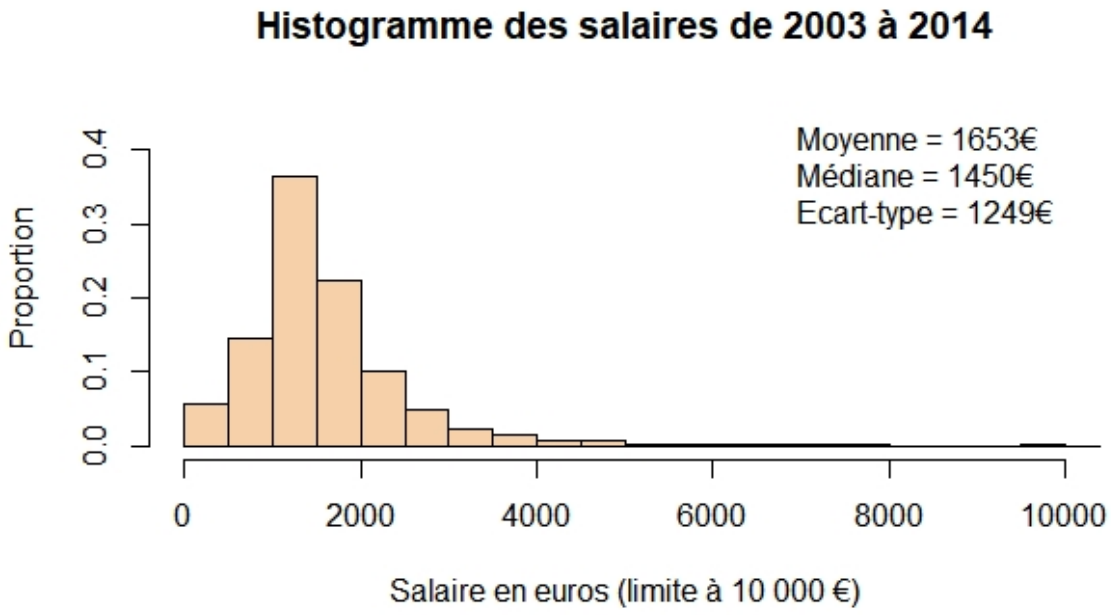
L'éducation fait partie d'un concept plus général qui est le capital humain. Il regroupe l'ensemble de ses aptitudes, talents et compétences d'un individu qui déterminent en partie sa capacité à travailler ou à produire. Dans cette optique, l'éducation peut avoir un retentissement à un niveau microéconomique en économie de l'information, mais aussi à un niveau macroéconomique avec l'économie du développement.
D'une part, le fait de faire des études afin d'augmenter son salaire futur peut-être vu en microéconomie comme un mécanisme de signaling. En effet, l'individu souhaite payer un coût supérieur au moment de faire des études, afin de pouvoir se démarquer des autres candidats sur le marché du travail. La scolarité en elle-même n'améliorerait pas ou peu la capacité productive des futurs salariés, mais aurait plutôt pour rôle de sélectionner, classer et signaler les individus selon leurs aptitudes. De ce fait, une personne ayant fait de longues études envoie un signal garantissant ses aptitudes sur le marché du travail.
D'autre part, dans un cadre macroéconomique, il est possible d'introduire dans les différents modèles standardisés de la croissance le capital humain (dans le modèle de Solow Swann par exemple). Le capital humain est une analogie du capital financier conventionnel. Il a été introduit dans la Théorie du Capital Humain par Gary Becker. Il est composé des compétences, des experiences et des savoirs, qui déterminent une certaine aptitude de l'individu à travailler. Ainsi, comme le capital physique, il

peut s'acquérir (par l'éducation par exemple), se préserver et se développer. De ce fait, l'individu peut-être considéré comme son propre investisseur, le rendant acteur sur le marché du travail. Les détenteurs de main d'oeuvre louent donc leur capital humain aux détenteurs de capital physique. Ces derniers en ont besoin car le capital humain permet la valorisation du capital physique. Par exemple, un travailleur décide d'investir dans son capital humain en prolongeant ses études. De ce fait, il préfère payer le coût supplémentaire dû au manque à gagner et à l'effort de faire de longues études, afin de profiter d'un capital humain plus grand. Il s'agit donc pour l'investisseur d'accroître son potentiel productif, et donc son salaire. Le salaire est alors considéré comme le rendement du capital humain, la rémunération de l'investissement dans l'éducation.

Question 2)

Les variables d'intérêt suivantes ont été sélectionnées : salaire, diplôme, expérience et taux d'actifs. Nous avons rapidement détecté la présence d'observations aberrantes, que nous avons choisi d'enlever de la base car elles ne reflètent pas le "phénomène économique" que nous souhaitons mettre en lumière. Par exemple, les individus se déclarant actif mais ne renseignant aucun salaire ne sont d'aucune utilité ici. De même, ceux gagnant plus de 75 000 euros mensuel sont considérés comme "outliers" car leur salaire n'est pas la conséquence d'un quelconque "phénomène économique". De la même manière, les actifs âgés de 100 ans ou plus sont écartés.

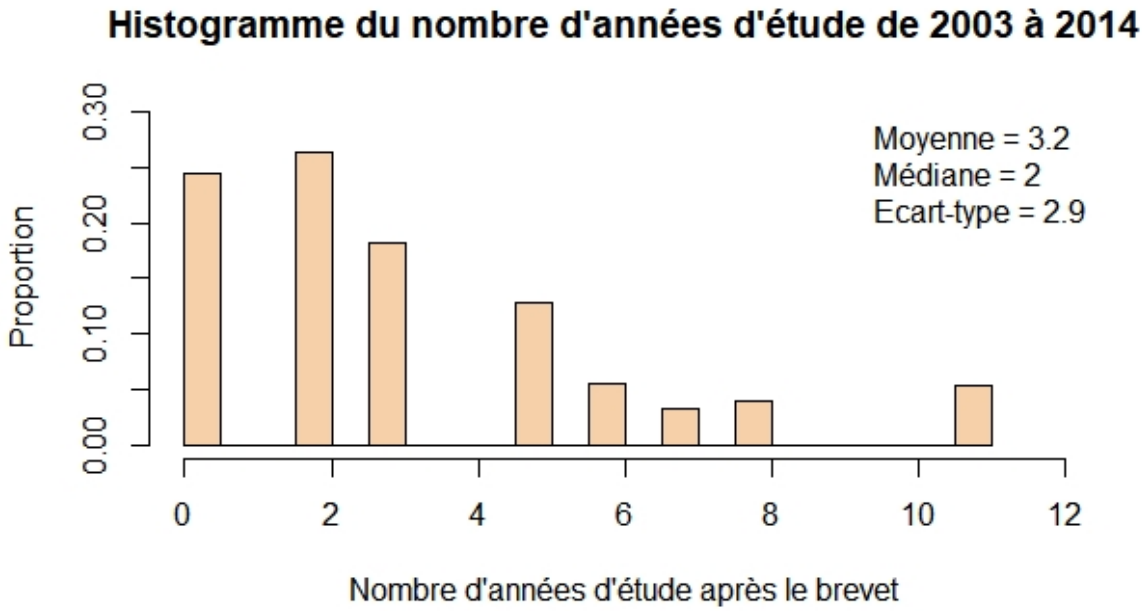
Ci-dessous est représenté l'histogramme des salaires que nous avons choisi de couper à 10 000 euros pour plus de visibilité :



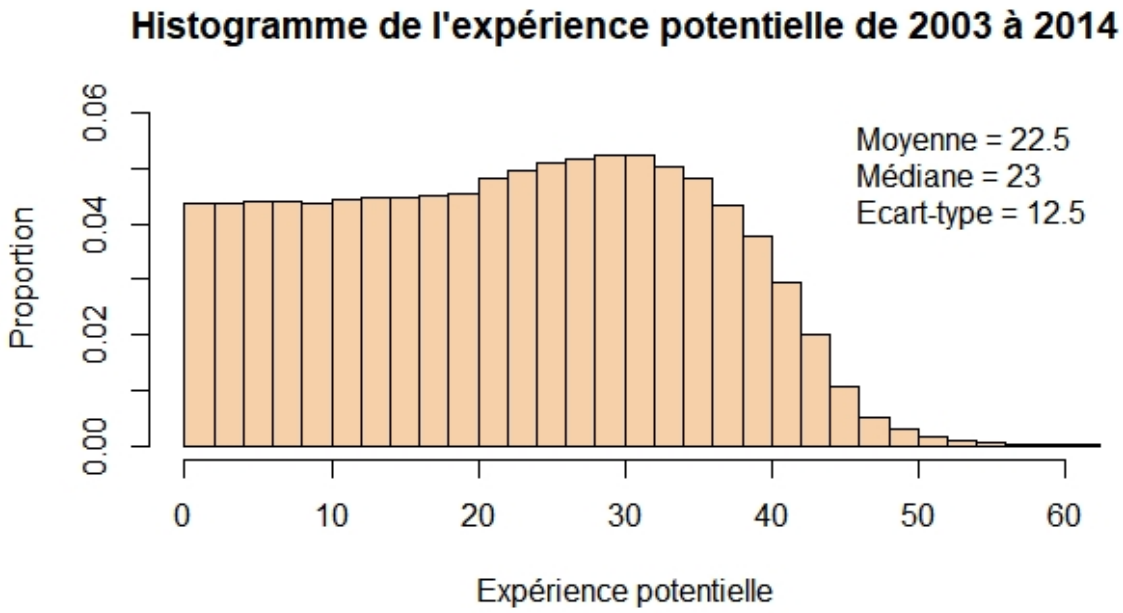
On remarque que l'écrasante majorité des salaires se situe en dessous de 4000 euros mensuel, avec presque 40% des individus gagant entre 1000 et 1500 euros. Cependant,

la présence de quelques revenus très élevés placent la moyenne au dessus de la médiane avec un écart-type quasiment de l'ordre de ces dernières.

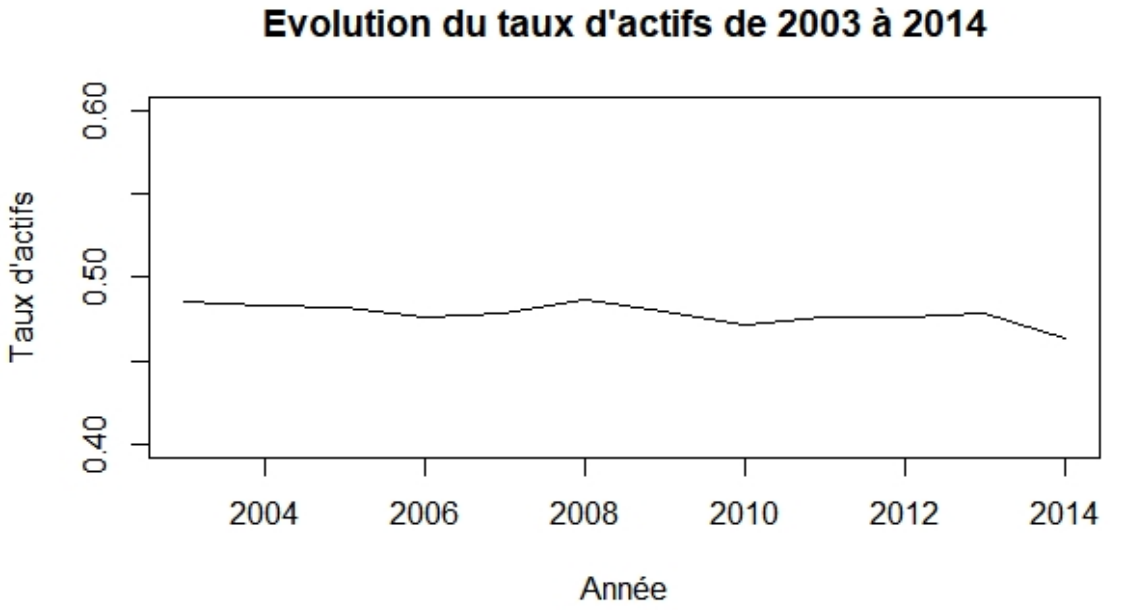
On s'intéresse maintenant à la distribution des diplômes, plus précisement au nombre d'années d'étude après le brevet. Le graphique suivant nous dit qu'un quart des personnes sondées n'ont pas de diplôme ou uniquement le brevet. Dans une proportion légèrement supérieure arrivent les diplômes professionnels type CAP et BEP. Environ 20% des individus obtiennent un BAC. Ensuite, 15% d'entre eux sont diplômés d'un diplôme type BTS, paramédical équivalent à BAC+2. Enfin, on observe entre 0 et 5% d'individus diplômés pour chaque catégorie après BAC+3 (licence, maîtrise, master et doctorat).



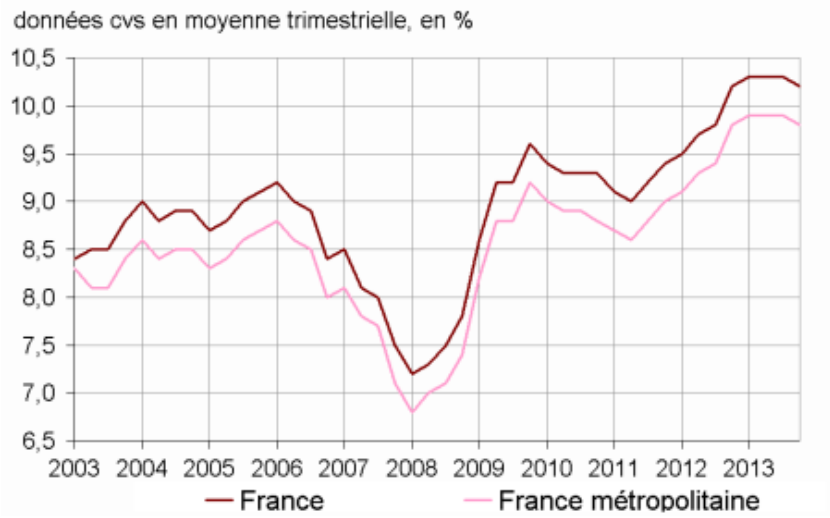
L'expérience professionnelle est une variable essentielle dans la détermination du salaire d'un individu. Nous la matérialisons dans le modèle par l'expérience potentielle qui est l'expérience qu'aurait un individu s'il avait travaillé depuis sa diplomation sans interruption. L'histogramme nous montre que la répartition de cette variable est assez stable au sein de la base.



Enfin, il nous a semblé intéressant de regarder l'évolution du taux d'actifs au cours de la période considérée. Bien que non présente dans le modèle, cette variable permet de dégager certaines tendances économiques. Puisque la catégorie "actif" regroupe une population bien plus large que "au chômage"/"en emploi", les tendances du taux de chômage sont faiblement visibles.



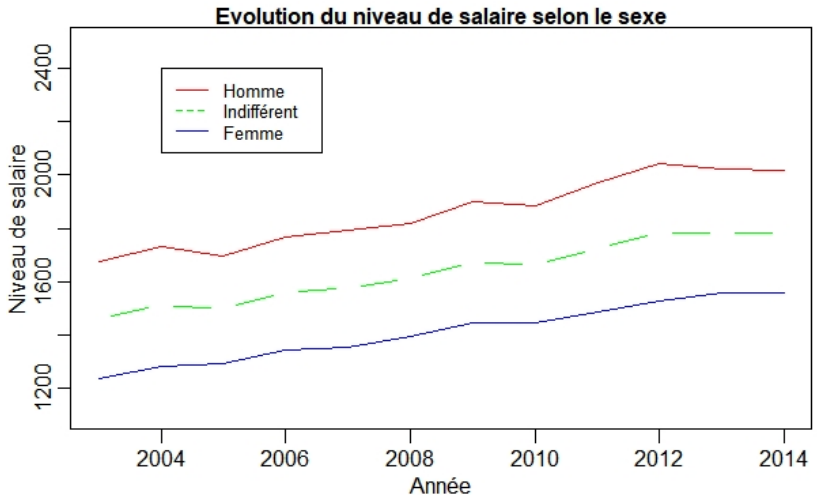
Evolution du taux de chômage en France, 2003-2014 (source : Insee)

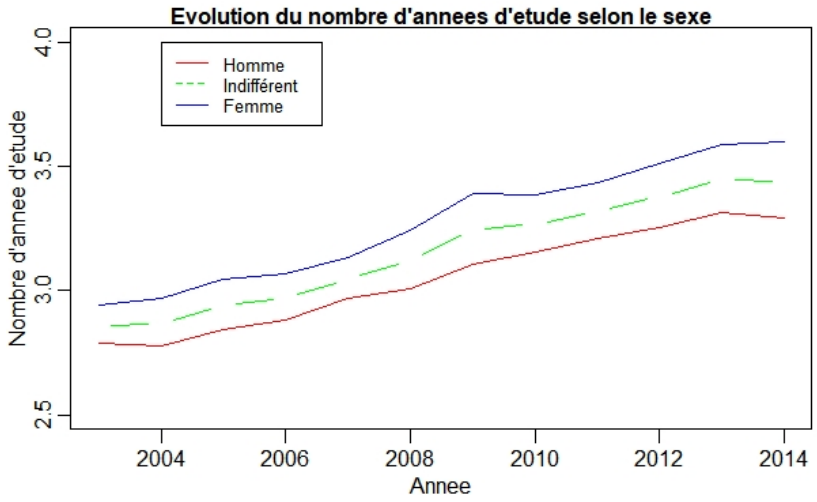


Cependant, lorsqu'on compare le premier graphe au deuxième, on remarque une légère corrélation inverse. En effet, on peut voir que sous l'effet de la crise, à partir de 2008, la proportion d'actifs diminue, alors que le chômage explose. Cette tendance baissière se poursuit avec l'installation du chômage de masse.

Question 3)

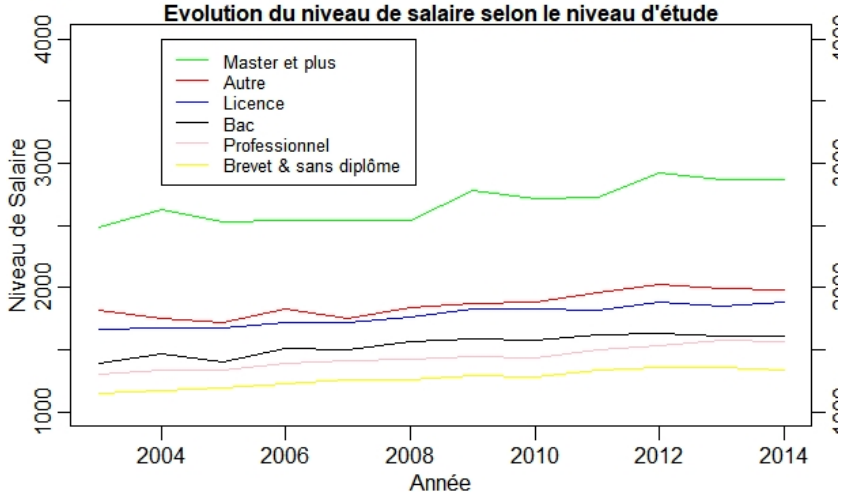
Lorsque l'on représente l'évolution des salaires dans le temps, on voit qu'il y a une croissance relativement constante. De même pour le nombre d'année d'étude.





La distinction selon le sexe permet de mettre en lumière une inégalité de revenus entre les hommes et les femmes. Malgré des études plus longues en moyenne, ces dernières ont un salaire plus faible en moyenne que ces derniers, et ce sur toute la période considérée. On remarque aussi que l'écart de salaire ne diminue pas avec le temps.

Le niveau de salaire augmente dans le temps pour tous les individus, peu importe leur diplôme. Cependant, il apparaît que les diplômés d'un Master ou plus ont un salaire nettement supérieur aux autres. Ils profitent aussi d'un accroissement de revenus plus élevé :

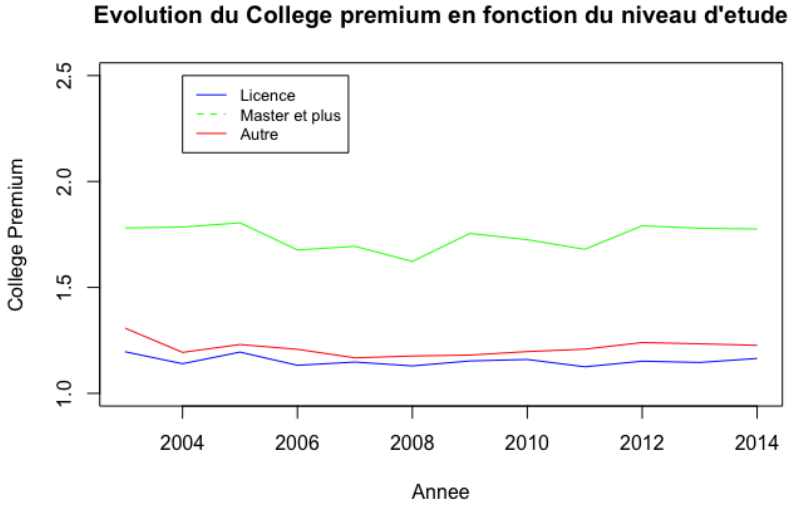


	Brevet ou sans diplôme	Bac	Master ou plus
Salaire.2014 - Salaire.2003	189	220	384

(Salaire.k mensuel de l'année k exprimé en euros)

Il est intéressant de remarquer qu'à chaque fois que le chômage augmente brusquement (2008-2009 et 2011-2012), le salaire des diplômés d'un Master+ augmente lui aussi plus rapidement.

Le College Premium, défini comme le ratio entre le salaire moyen des individus diplômés de l'enseignement supérieur et le salaire moyen des individus ayant uniquement le baccalauréat, reste relativement stable dans le temps. Cela entraîne donc l'accroissement des écarts de rémunération entre les diplomés du supérieur et les autres, puisque les salaires augmentent dans le temps.



Question 4)

Une première approche consiste à évaluer l'effet de l'éducation sur le salaire via le modèle de Mincer (1974) :

$$ln(w_{it}) = \beta_0 + \beta_1 s_{it} + \beta_2 x_{it} + \beta_3 x_{it}^2 + \epsilon_{it}$$

Où w_{it} est le salaire mensuel de l'individu "i" de l'année "t", s_{it} le nombre d'années d'étude et x_{it} l'expérience potentielle. Il est préférable ici de prendre l'expérience potentielle plutôt que l'expeience véritable car cette dernière peut-être d'une part liée aux choix de l'individu, et d'autre part aussi dépendre du salaire. Ainsi, cette variable serait potentiellement endogène par double causalité. Au contraire, l'expérience potentielle dépend uniquement de l'âge, et non de choix.
L'application de la méthode des MCO nous donne les résultats suivants :

TABLE 1 – Modèle de Mincer estimé par MCO

Observations	133,494			
R^2	0.202			
Adjusted R^2	0.202			
Residual Std. Error	0.572 (df = 133490)			
F Statistic	11,242.050*** (df = 3; 133490)			
	Estimate	Std. Error	t value	Pr(>|t|)
β_0	6.4155	0.0054	1188.36	0.0000
β_1	0.0956	0.0006	161.40	0.0000
β_2	0.0471	0.0005	99.08	0.0000
β_3	-0.0008	0.0000	-75.96	0.0000

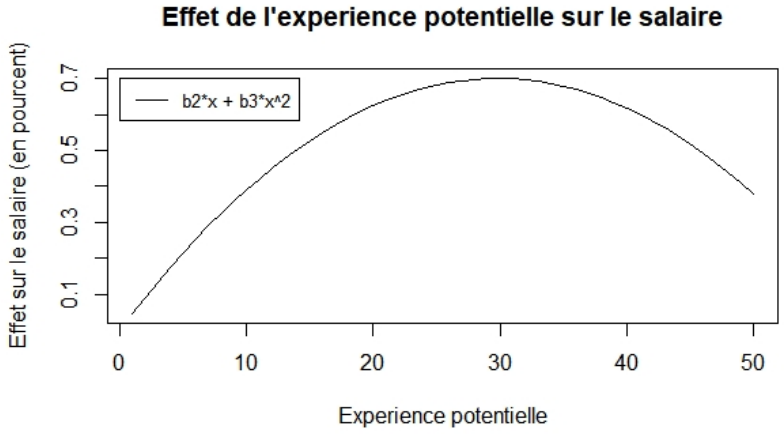
Le modèle est significatif dans son ensemble à 1% d'après la F-statistique. De même, les coefficients sont significatifs un à un avec des p-valeurs très inférieures à 1%.
Le R2 du modèle est égal à 0.2, il manque donc beaucoup de variables explicatives.

Puisque l'on régresse le salaire en échelle logarithmique, l'interprétation des coefficients se fait de la manière suivante :

— Chaque année d'étude en plus après le brevet fait varier le salaire de 9.5% à la hausse. Le signe positif du coefficient est logique et nous nous attendions à un tel résultat.

— Chaque année d'expérience potentielle supplémentaire fait varier le salaire de $(\beta_2 + 2\beta_3)$*100%

On remarque que le signe de β_3 est négatif, l'effet de l'expérience sur le salaire est donc une parabole concave (représenté sur le graphe ci-dessous). Le maximum est atteint en $\frac{-\beta_2}{2\beta_3} = 29.8$ années d'expérience.



Ces coefficients sont sans biais sous les hyothèses imposées par les MCO :
MLR1 : La variable expliquée doit être linéaire en les coefficients. Cette hypothèses est repectée.
MLR2 : Les variables (Y_i, X_i) sont iid. Cette hypothèse n'est pas respectée car premièrement

nous avons un biai de sélection : les individus sélectionnés sont actifs donc on ne peut pas généraliser à l'ensemble de la population. Il est possible de corriger ce biai en utilisant un modèle d'auto-sélection de type Tobit généralisé (ou Tobit II), en supposant que les hypothèses sous-jacentes soient vérifiées. Pour ce faire, on pourrait utliser la méthode d'Heckit. De plus, on a négligé l'aspect panel des données. Donc d'une année à l'autre et pour un même individu, son salaire et ses caractéristiques sont directement corrélés. En prenant en compte cette caractéristique, on corrigerait notre erreur.
MLR3 : Les variables explicatives doivent être exogènes. Ici on a clairement un problème d'endogénéité. En effet, le modèle ne prend pas en compte un nombre important de régresseur qui influent sur le salaire. Ces derniers, qui se retrouvent dans le terme d'erreur, ont de grandes chances d'être corrélés aux variables explicatives. Ceci est d'autant plus vrai que les régresseurs oubliés dépendent de l'individu. Le fait de prendre l'expérience potentielle plutôt que l'expérience à proprement parler participe à réduire les chances d'avoir cette variable endogène. Encore une fois, considérer l'aspect panel des données permettrait d'appliquer des méthodes telles que la méthode de la "first difference" ou encore "within" qui aide à corriger le problème d'endogénéité. Cette problématique est traitée dans la question 6.
MLR 4 : Les variables explicatives ne sont pas constantes et il n'y a pas de colinéarité parfaite entre elles. Cette hypothèse est sans aucun doute respectée.
MLR 5 : Les résidus sont homoscedastiques : $Var(\epsilon_{it}|s_{it}, x_{it}) = \sigma^2$. Pour vérifier la validité de cette hypothèse on réalise un test de Breusch-Pagan. L'hypothèse nulle d'homoscédasticité est rejetée avec une p-valeur inférieure à $2e^{-16}$. Pour corriger ce problème et obtenir des écarts-type convergents, on pourrait utiliser l'estimateur robuste de la variance :

$$\hat{V} = (\frac{1}{n}\sum_{i=1}^{n} X_i X_i')^{-1}(\frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon_i}^2 X_i X_i')(\frac{1}{n}\sum_{i=1}^{n} X_i X_i')^{-1}$$

Question 5)

Il y a surement des caractéristiques inobservées intrasèques à l'individu qui peuvent influencer sur le niveau de salaire ; l'abilité par exemple. De plus, il peut y avoir aussi des effets de conjecture économique, qui influent sur les niveaux de salaire. Nous avons donc certainement un problème de variables omises qui peut conduire à de l'endogénéité. Néanmoins, hormis les conjectures économiques, on peut raisonnablement penser que les caractéristiques individuelles inobservées sont constantes dans le temps.

$$Y_{it} = X_{it}'\beta_0 + \nu_{it} \text{ avec } \nu_{it} = \alpha_i + \epsilon_{it}$$

α_i représente l'effet individuel. Comme il est corrélé à X, on a un problème d'endogénéité qui peut être cependant résolu grâce aux données de panel. Sous la condition d'exogénéité forte, il est possible de se débarrasser du problème d'endogénéité avec les estimateurs des différences premières ou des estimateurs within. Le premier est recommandable dans le cas de chocs idiosyncratiques persistants (résidus très corrélés entre eux), alors que le second est utilisé dans le cas de chocs indépendants (résidus très peu correlés). La première option n'est pas envisageable car le niveau d'éducation de l'individu est constant dans le temps. Donc en utilisant les FD, on le supprimerait tout

simplement et on ne pourrait plus mesurer son effet. On préferera donc utiliser l'estimateur Within.

Cependant, cette solution repose sur des hypothèses fortes : l'exogénéité forte $E[X_{it}\epsilon_{it'}] = 0 \ \forall(t,t')$. Cela signifie que les régresseurs ne sont correlés ni aux chocs passés, ni aux chocs futurs. On peut suspecter une défaillance de l'hypothèse dans le cas où les 2 estimateurs sont très différents.

Dans notre cas, l'hypothèse d'exogénéité forte est peu plausible. En effet, il est fort probable que le salaire présent d'un individu dépende du salaire de l'année passée. Il y a donc en plus de l'hétérogénéité individuelle de la dépendance d'état par l'effet de Y_{it-1}. Dans ce cas, les estimateurs des différences premières et within ne sont pas consistants. De plus, l'estimateur within est préférable dans le cas de résidus peu corrélés, ce qui ici est aussi peu probable.

On peut alors se ramener à une estimation des doubles moindres carrés sur les données empilées : on utilise X_{it-1} comme instrument de ΔX_t. Ainsi, on obtient un estimateur convergent sous l'hypothèse d'exogénéité faible.

On ne peut néanmoins pas résoudre le problème de biais de sélection avec les données de panel. En effet, on observe uniquement les salaires des personnes qui sont actives.

Question 6)

On dispose ici du salaire par "tranche" des individus, stocké dans la variable "salmet", et dont les seuils sont référencés par l'INSEE. Ces derniers sont :
$[0, 500, 1000, 1250, 1500, 2000, 2500, 3000, 5000, 8000, Inf]$, exprimés en euros. Afin d'avoir plus d'observations disponibles, nous avons ajouté les individus déclarant leur salaire "réel" à la tranche correspondante.
Il s'agit ici d'un modèle dit polytomique ordonné, qui est estimé par maximum de vraisemblance. On a $Y^* = X'\beta + \sigma\epsilon$ et on observe $Y = \sum_{k=1}^{K} k * \mathbf{1}(\alpha_{0k} < Y^* \leq \alpha_{0k+1})$ avec α_{0k} les seuils définis ci-dessus et ϵ de loi F, indépendant de X. On choisit ici $F = \Phi$, soit un modèle probit. La vraisemblance du modèle s'écrit :

$$l_n(Y|X,\beta,\sigma) = \sum_{i=1}^{n} ln[F(\frac{\alpha_{0Yi+1} - X_i'\beta}{\sigma}) - F(\frac{\alpha_{0Yi} - X_i'\beta}{\sigma})]$$

que l'on maximise en (β, σ).

Dans un premier modèle, on considère $X = (1, s_{it}, x_{it}, x_{it}^2)$, soit l'équation de Mincer de la question 4. Les résultats sont présentés dans la table 2 ci-dessous.

TABLE 2 – Modèle de Mincer estimé par maximum de vraisemblance (salaire par tranche)

Interval regression				
Maximum Likelihood estimation				
Log-Likelihood				-302360.7
Observations				157872
Free parameters (df = 157867)				5
	Estimate	Std. Error	t value	Pr(>|t|)
β_0	6.487	0.004107	1579.36	<2e-16
β_1	0.0956	0.0004	235.06	<2e-16
β_2	0.0428	0.0003	126.70	<2e-16
β_3	-0.00068	7.018e-06	-96.89	<2e-16
σ	0.4645	0.00069	669.80	<2e-16

Les résultats montrent un modèle très similaire à celui estimé avec le salaire réel. Tous les coefficients sont significatifs. Les rendements privés de l'éducation sont exactement égaux au premier modèle. On remarque un léger changement dans l'effet de l'expérience sur le salaire : $\beta_2 = 4.28\%$ et $\beta_3 = -6.8 * 10^{-4}$ contre 4.71% et $-8 * 10^{-4}$ respectivement pour le modèle de la question 4. Cela amène l'expérience potentielle optimale à $\frac{-\beta_2}{2\beta_3} = 31.4$ années contre 29.8 précédemment.

On rajoute maintenant une variable explicative au modèle : f_{it} qui vaut 1 si l'individu est une femme et 0 sinon. On a donc $X = (1, s_{it}, x_{it}, x_{it}^2, f_{it})$. On estime le modèle toujours par maximum de vraisemblance dont les résultats sont présentés dans la table 3.

TABLE 3 – Modèle de Mincer complété du sexe de l'individu estimé par maximum de vraisemblance (salaire par tranche)

Interval regression				
Maximum Likelihood estimation				
Log-Likelihood				-293702
Observations				157872
Free parameters (df = 157866)				6
	Estimate	Std. Error	t value	Pr(>|t|)
β_0	6.626	0.003956	1674.92	<2e-16
β_1	0.09898	3.818e-04	259.27	<2e-16
β_2	0.04198	3.121e-04	134.52	<2e-16
β_3	-6.496e-04	6.523e-06	-99.59	<2e-16
β_4	-0.3068	2.283e-03	-134.35	<2e-16
σ	0.4384	6.475e-04	677.05	<2e-16

Tous les coefficients du modèle sont significatifs et on remarque même que ceux déjà présents dans les modèles précédents ont tous eu une significativité renforcée avec des t-valeurs plus extrêmes. L'effet du nombre d'années d'étude a augmenté passant de 9.56% à 9.90%. L'effet de l'expérience potentielle a aussi évolué avec une expérience

optimale désormais à 32.3 années. Enfin, l'ajout de la variable "sexe" dans le modèle nous dit que le fait d'être une femme engendre une baisse moyenne de 30.7% du salaire mensuel. Ceci est cohérent avec les études réalisées en question 3 qui montrent un écart moyen de 24% entre les hommes et les femmes.

2 Les rendements sociaux de l'éducation

Question 7)

Dans cette partie, on se propose d'estimer les rendements sociaux de l'éducation. Autrement dit, on veut regarder l'influence de la proportion de salariés diplomés du supérieur dans le département sur le niveau de salaire des individus. Pour se faire, on introduit le modèle suivant :

$$\ln(w_{it}) = \alpha + \beta s_{it}^{ind} + \delta S^{dep_{it}} + \gamma X_{it}^{ind} + \rho X^{dep_{it}} + \eta_{it}$$

— s_{it}^{ind} représente le niveau de diplôme de l'individu. C'est une variable catégorielle à 4 niveaux : 0 sans diplôme, 1 baccalauréat, 2 licence, et 3 master et plus. On prendra le niveau 0 comme catégorie de référence dans la régression.

— $S^{dep_{it}}$ est la proportion des salariés diplômés du supérieur dans le département.

— X_{it}^{ind} est le vecteur des caractéristiques socio-démographiques de l'individu. On y a fait apparaître l'âge, le sexe, le nombre d'enfants (catégorielle 9 niveaux) et le type de ménage (catégorielle 5 niveaux).La description des variables catégorielles est fournie en annexe. Dans la régression, on prendra respectivement "sans enfants" et "ménage d'une personne seule" pour les catégories de références.

— $X^{dep_{it}}$ représente le vecteur des caractéristiques du département. Nous avons choisi d'y faire paraître la proportion que représente le nombre d'étudiants du département, par rapport au nombre d'étudiants en France. Cette variable reflète la tendance démographique et l'attractivité du département pour les étudiants et donc d'une certaine manière pour les actifs.

On utilise donc la méthode des moindres carrés pour estimer les coefficients de la régression.

Observations	125,505
R^2	0.217
Adjusted R^2	0.217
Residual Std. Error	0.624 (df = 125484)
F Statistic	1,742.646*** (df = 20; 125484)
Note :	*p<0.1; **p<0.05; ***p<0.01

	Estimate	Std. Error	t value	Pr(>|t|)
α	6.6716	0.0104	644.29	0.0000
β_{dip1}	0.2912	0.0042	70.15	0.0000
β_{dip2}	0.3973	0.0064	62.54	0.0000
β_{dip3}	0.6787	0.0067	101.58	0.0000
δ	1.2435	0.0245	50.66	0.0000
γ_{age}	0.0152	0.0002	91.92	0.0000
γ_{sexe}	-0.3206	0.0036	-88.84	0.0000
γ_{nbenf1}	0.0844	0.0064	13.14	0.0000
γ_{nbenf2}	0.1483	0.0107	13.88	0.0000
γ_{nbenf3}	0.1483	0.0099	15.02	0.0000
γ_{nbenf4}	0.1436	0.0070	20.58	0.0000
γ_{nbenf5}	0.1648	0.0100	16.52	0.0000
γ_{nbenf6}	0.1542	0.0108	14.33	0.0000
γ_{nbenf7}	0.0566	0.0113	4.99	0.0000
γ_{nbenf8}	0.0932	0.0132	7.08	0.0000
γ_{nbenf9}	0.1170	0.0151	7.72	0.0000
γ_{men2}	-0.0652	0.0083	-7.82	0.0000
γ_{men3}	0.0174	0.0059	2.94	0.0033
γ_{men4}	-0.0236	0.0065	-3.63	0.0003
γ_{men5}	-0.1221	0.0108	-11.29	0.0000
ρ_{petu}	-0.7041	0.1645	-4.28	0.0000

Le modèle est significatif dans son ensemble avec une F-statistique très élevée. En effectuant une succession de t-tests sur les coefficients de la régression, on remarque qu'ils sont tous significatifs à 1%. En regardant les rendements privés de l'éducation (les coefficients β), on observe comme prévu que plus le diplôme obtenu est élevé, plus le salaire augmente. Dans le cas d'un diplôme de niveau master ou plus, le salaire attendu augmente de 68% par rapport à une personne n'ayant aucun diplôme. On remarque que comparé à l'estimation de la première partie, les rendements privés sont inférieurs. En effet, on avait un rendement de 9,5% par année d'étude et un master correspond à 8 années, ce qui fait un rendement total de 76%.
Concernant les rendements sociaux de l'éducation, lorsqu'on augmente de 1% la population de personnes diplomées du supérieur dans le département, on augmente le salaire de 1,24%. Ainsi, comme attendu, les rendements sociaux du capital humain sont positifs. Dans un environnement ou la proportion de travailleurs qualifiés est plus grande, le salaire de l'ensemble de travailleurs augmente.
De la même manière, on constate que plus le département représente une proportion d'étudiants élevée par rapport au nombre d'étudiants en France, moins les salaires sont hauts. Ils baissent de 0,7% pour chaque augmentation de 1% de cette variable. Cela pourrait s'expliquer par le fait que plus il y a d'étudiants dans le département, plus il y a un effet de concurrence.

De plus, comme nous l'avons vu avec les statistiques descriptives, la différence de salaire entre les hommes et les femmes est marquée. Une femme aura un salaire d'environ 32% inférieur à un homme, ce qui est comparable aux résultats des précédents modèles.
Enfin, fait surprenant, il semblerait que les parents ayant des enfants ont tendance à gagner plus que les parents sans enfants. On pourrait expliquer que les coefficients $\gamma_{nbenf2}, \gamma_{nbenf3}, \gamma_{nbenf5}, \gamma_{nbenf6}, \gamma_{nbenf8}, \gamma_{nbenf9}$ sont plus élevés que $\gamma_{nbenf1}, \gamma_{nbenf4}, \gamma_{nbenf7}$ car ils mentionnent la présence d'un jeune enfant (maximum 5 ans) dans le ménage. Dans certains cas, un des deux parents s'arrête de travailler pour s'occuper de l'enfant, et il s'agit le plus souvent du parent avec le salaire le moins élevé à la base. On mesure alors le salaire du parent resté actif, donc avec un salaire plus élevé.
En revanche, on ne peut pas donner de conclusions générales pour le type de ménage, les coefficients allant à l'encontre de ce que l'on vient d'énoncer pour le nombre d'enfant. Par exemple, le fait que γ_{men2} soit négatif nous dit que le parent dans une famille monoparentale gagne moins qu'une personne seule. On peut seulement dire qu'une "complexification" du ménage entraine une baisse de salaire.

Question 8)

Comme dans la première partie, il y a probablement un problème d'hétérogénéité inobservée et donc d'endogénéité. Il y a beaucoup de caractéristiques inobservées sur le département où le travailleur travaille qui peuvent influencer le niveau de salaire, tout en étant corrélées avec la proportion de travaileurs qualifiés dans le département (le capital humain). On peut en distinguer de deux sortes : les chocs de demande qui affectent la productivité relative des travailleurs, puis les chocs d'offres qui eux ont un effet sur l'attractivité relative du déparement pour un travailleur avec un fort capital humain. Par ce fait, il y a un biais dans l'estimation par MCO des coefficients.

En réalité, le coefficient d'erreur peut-être décomposé en 3 parties :

$$\eta_{idt} = \mu_d \theta_i + \nu_{dt} + \epsilon_{idt}$$

où d représente le département, θ_i est une composante individuelle permante, telle que l'abilité ou le milieu familial, μ_d un facteur de poids propore à chaque département, ν_{dt} représentant les chocs de demande de travail dans le département d au temps t, et enfin ϵ_{idt} les chocs idiosyncratiques.

La première source de biais est due aux chocs de demande variant dans le temps et étant corrélés au stock de capital humain dans les départements. En effet, ces derniers diffèrent par leur localisation géographique, leurs conditions météorologiques, ou leur industrialisation par exemple. Même si les effets permanents dans le temps peuvent être supprimés grâce aux premières différences, certains ne sont pas constants ; par exemple les chocs de demande transitoires qui attirent les profils qualifiés et augmentent les salaires.

La seconde source de biais de variables omises est la présence de caractéristiques inobservées des individus. Il y a des variables qui jouent sur le fait que les individus observés dans les départements avec un haut degré de capital humain sont de meilleurs travailleurs que des individus équivalents observés dans des départements avec un faible degré de capital humain. En considérant le modèle de Roy, où les différents départements récompensent les capacités des travailleurs différemment, les décisions de mobilité sont basées sur la comparaison des avantages. Ainsi, dans un tel modèle, les travailleurs ne sont pas distribués aléatoirement dans les départements.

Question 9)

Dans cette question, les données utilisées se situent sur la période 2003-2012 pour des raisons techniques.

Comme on l'a vu dans la question 8, on suspecte dans le modèle plusieurs formes d'endogénéités pour la variable d'intérêt ; la proportion de diplomés du supérieur dans le département. Une partie de cette endogénéité est due à des effets constants, comme par exemple les caractéristiques inobservées du départements qui attirent les travailleurs qualifiés (commodités, aménagements, climat, etc..). Grâce aux données de panel, on peut résoudre ce problème avec l'estimateur des différences premières. L'idée est alors ici de faire la régression de $log(w_{it}) - log(w_{it-1})$ sur $X_{it} - X_{it-1}$, où X_{it} représente l'ensemble des variables explicatives du modèle. Grâce à cette méthode, il est possible de se débarrasser des effets constants dans le temps, en supposant l'hyphothèse d'exogénéité forte.
Néanmoins, on a aussi vu qu'une partie de cette endogénéité était due à des effets qui n'étaient pas nécessairement constants dans le temps. Par exemple, un choc positif temporaire dans l'offre de travail qui va attirer plus de travailleurs qualifiés et faire monter les salaires. Pour essayer de résoudre ce problème d'endogénéité, on peut utiliser la méthode des variables instrumentales. Le but est ici de trouver des variables qui jouent directement sur la proportion de travailleurs qualifiés dans le département, sans directement jouer sur le niveau de salaire. Autrement dit, si Z_{it} est la variable instrumentale, on doit avoir $cov(Z_{it}, S_{it}^{dep}) \neq 0$ et $cov(Z_{it}, \eta_{it}) = 0$.

Nous avons alors sélectionné une variable instrumentale qui pourrait remplir ces conditions : la création d'établissements du supérieur dans le département à la date t-5. A priori, cette variable est corrélée avec proportion de travailleurs qualifiés dans le département. En effet, il est légitime de penser que plus il y a d'opportunités de faire des études supérieures dans le département, plus il y a de chances qu'il y ait des travailleurs qualifiés. De plus, nous supposons qu'il faut environ 5 ans pour obtenir un diplôme du supérieur. L'hypothèse de rang ici présentée pourra être vérifiée ultérieurement lors du calcul des estimateurs "Two Stage Least Square". La première étape revient à régresser la variable à instrumenter sur les variables instrumentales. Si les coefficients dans la régression sont significatifs, les instruments choisis sont pertinents. En revanche, il n'est pas possible de tester l'hypothèse de non corrélation des variables instrumentales avec

les résidus. Néanmoins, il n'y a aucune raison de penser qu'elle influe directement sur le niveau de salaire. Le fait d'augmenter le nombre d'établissements du supérieur 5 ans auparavant ne conduit pas à une augmentation des salaires des travailleurs qui s'y trouvent. Il est donc probable que la cette dernière variable ne soit pas correlée avec les variables non observées du salaire.
Ainsi, on instrumentera ΔS_{it}^{dep} par ΔZ_{it}, où Z_{it} est le vecteur des instruments définis plus haut. Il se composera des variables de contrôle (les variables supposées exogènes du modèle général), et des instruments eux-même. Cela permet d'éviter d'induire une endogénéité qui n'existait pas avant l'instrumentation.

Au moment de construire les variables en premières différences, nous avons remarqué que peu d'individus étaient présents lors de 11 années consécutives. Nous avons alors regardé ceux qui étaient présents d'une année sur l'autre. C'est en faisant la différence des caractéristiques de ces derniers que nous constituons notre jeu de données en FD.
De plus, on supprime du modèle toutes les variables qui deviennent constantes en différences premières (sexe, age, etc..), ainsi que toutes les variables catégorielles. Il reste donc dans la régression le taux de diplômés du supérieur dans le département, et la proportion d'étudiants dans le département par rapport à la population d'étudiants en France.

• Première étape : Régression de la variation du taux de diplomés du supérieur sur le nombre d'établissements créés à la date T-5 :

$$\Delta S_{it}^{dep} = \mu_1 \Delta N_{it}^{etab} + \mu_2 \Delta X_{it}^{dep} + \Delta \lambda_{it}$$

Où l'on suppose ΔN_{it}^{etab} et ΔX_{it}^{dep} exogènes par rapport à $\Delta \lambda_{it}$ le terme d'erreur.

TABLE 4 – Instrumentation de ΔS_{it}^{dep} par ΔN_{it}^{etab}

Observations	92,513
R^2	0.133
Adjusted R^2	0.133
Residual Std. Error	0.033 (df = 92511)
F Statistic	7,068.091*** (df = 2 ; 92511)
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01

	Estimate	Std. Error	t value	Pr(>|t|)
μ_1	0.0005	0.0000	35.29	0.0000
μ_2	2.6341	0.0278	94.64	0.0000

On constate dans dans la table 4 que le coefficient devant l'instrument est significatif. La condition de rang est donc vérifiée. De plus, il est positif ce qui paraît cohérent économiquement car l'ouverture d'un établissement ne peut qu'augmenter la proportion de salariés diplômés du supérieur dans le département. Enfin, lorsque la première promotion de cet établissement arrive sur le marché du travail 5 ans plus tard, l'ordre de grandeur de $\mu_1 = 0.05\%$ semble correct. En revanche, on est ici en présence d'un

instrument faible car μ_1 est proche de zéro. Il faudra donc être vigilant à la valeur de la F-statistique de la deuxième étape. On obtient donc $\widehat{\Delta S_{it}^{dep}} = \widehat{\mu_1}\Delta N_{it}^{etab} + \widehat{\mu_2}\Delta X_{it}^{dep}$.

• Etape 2 : Régression de la variation du logarithme du salaire sur la valeur estimée de la proportion de diplomés dans le supérieur, et sur la proportion d'étudiants dans le département par rapport au nombre d'étudiants en France :

$$\Delta log(w_{it}) = \delta\widehat{\Delta S_{it}^{dep}} + \rho\Delta X_{it}^{dep} + \Delta\eta_{it}$$

TABLE 5 – Estimation du modèle en différences premières

Observations	2,215
R^2	0.001
Adjusted R^2	−0.0003
Residual Std. Error	0.499 (df = 2213)
F Statistic	0.660 (df = 2 ; 2213)
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01

	Estimate	Std. Error	t value	Pr(>|t|)
δ	-2.7117	2.7479	-0.99	0.3238
ρ	6.4502	8.5005	0.76	0.4480

Les résultats de ce modèle ne sont pas satisfaisants car le modèle global n'est pas significatif ni aucune des deux variables une à une. Cela est surprenant d'autant plus que l'instrumentation était correcte. Plusieurs raisons peuvent être attribuées à cet échec. D'une part, la variation de la proportion de diplômés du supérieur est en général très proche de zéro. Malgré une instrumentation significative, on peut avoir du mal à estimer correctement cette variable et avoir des changements de signes fréquents. Ici la F-stat est très inférieure à 10, on a donc toujours un biai significatif du fait de l'instrument faible. D'autre part, la création d'établissements du supérieur est souvent égale à 0 d'une année sur l'autre. Cela rajoute donc encore des difficultés d'estimer correctement la variation de la proportion de salariés diplômés du supérieur.

Nous avons pensé à introduire d'autres variables instrumentales. Par manque de temps ou à cause de la difficulté de trouver les informations correspondantes dans nos bases, nous n'avons pas pu les implémenter ici. Mais nous allons néanmoins discuter leur validité.

Tout d'abord, nous avons pensé à instrumenter la variation de la proportion de salariés diplômés du supérieur du département par la variation de la proportion de foyers pauvres dans le département. En effet, le nombre de personnes vivant sous le seuil de pauvreté est un indicateur de l'attractivité économique et du dynamisme de la région. Il est donc raisonnable de penser que si le nombre de personnes pauvres augmentent dans le département, ce dernier est moins compétitif, et donc la proportion de travailleurs qualifiés risque de diminuer. Cependant, elle doit aussi être décorrélée avec l'erreur sur les salaires ; à savoir les variables que l'on observe pas. La proportion de pauvres dans

le département ne constituant pas un choix des individus, ni des entreprises qui s'y trouvent, on pourrait donc penser qu'elle soit exogène. En revanche, cette proportion peut aussi être liée à une mauvaise conjecture économique. Or cette dernière affecte négativement les salaires et n'est pas observée dans notre modèle donc se retrouve dans le terme d'erreur. L'instrument serait donc lui aussi corrélé à l'erreur et donc endogène. De plus, une augmentation (resp. une diminution) de la proportion de foyers pauvres dans le département pourrait être dûe à une "fuite" (resp. une arrivée) des ménages aisés du département. Cela reflète des qualités intrasèques du territoire, inobservées ici et donc cela entrainerait probablement des problèmes d'endogénéité.

Nous avons aussi pensé à instrumenter sur la proportion de ménages urbains ou ruraux. En effet, les travailleurs qualifiés ont plus tendance à s'installer dans les grandes aires urbaines, car c'est en général ici que l'on trouve les grandes entreprises. Il est donc aussi raisonnable de penser que la proportion de ménages urbains soit correlée positivement avec la proportion de travailleurs diplômés du supérieur. En effet, les deux traduisent un dynamisme de la région qui attire les travailleurs qualifiés. De plus, il y a de fortes chances pour que cet instrument soit aussi exogène. En effet, il ne constitue pas un choix de l'individu, et le taux de ménages urbain ne risque probablement pas de dépendre de quelques effets de conjectures économiques qui pourraient aussi avoir un impact sur les niveaux de salaires. Avec plusieurs instruments, nous aurions pu effectuer un test de Sargan pour essayer de voir s'il n'y avait pas de preuves supplémentaires contre l'exogénéité des instruments.

Question 10)

Dans cette question, on souhaite estimer simultanément les rendements privés et sociaux de l'éducation. Cependant, comme le niveau d'étude est constant dans le temps, on ne peut plus utiliser la méthode des différences premières pour essayer de se débarrasser de la potentielle endogénéité. On procède donc cette fois-ci à une estimation 2SLS où l'on instrumente à la fois le niveau d'étude et la proportion de travailleurs qualifiés du département. Il s'agissait donc de trouver des instruments pour les deux variables. On rappelle le modèle à estimer :

$$\ln(w_{it}) = \alpha + \beta s_{it}^{ind} + \delta S^{dep_{it}} + \gamma X_{it}^{ind} + \rho X_{it}^{dep} + \eta_{it}$$

• Concernant la proportion de travailleurs qualifiés, nous avons choisi d'utiliser le même instrument que précédemment. A savoir le nombre d'établissements du supérieur présents dans le département à l'année T-5. Les considérations liées à la justification de l'instrument restent les mêmes par rapport au cas en premières différences. On régresse donc ici la proportion de travailleurs diplomés du supérieur sur le nombre d'établissements présents dans le département à l'année T-5, ainsi que tous les autres régresseurs exogènes du modèle, à savoir le nombre d'enfants, le type de ménage, la

population d'étudiants, l'age et le sexe.

$$S_{it}^{dep} = \mu_0 + \mu_1 N_{it}^{etab} + \mu_2 X_{it}^{ind} + \mu_3 X_{it}^{dep} + \lambda_{it}$$

Les résultats de la régression sont visibles sur la table 6 ci-après. Ils nous permettent d'obtenir $\widehat{S_{it}^{dep}} = \widehat{\mu_0} + \widehat{\mu_1} N_{it}^{etab} + \widehat{\mu_2} X_{it}^{ind} + \widehat{\mu_3} X_{it}^{dep}$.

TABLE 6 – Instrumentation de S_{it}^{dep} sur N_{it}^{etab}

Observations	467,738
R^2	0.430
Adjusted R^2	0.430
Residual Std. Error	0.067 (df = 467720)
F Statistic	20,768.000*** (df = 17 ; 467720)
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01

	Estimate	Std. Error	t value	Pr(>|t|)
μ_0	0.1304	0.0006	232.62	0.0000
μ_1	0.0014	0.0000	298.58	0.0000
μ_{2age}	0.0001	0.0000	7.69	0.0000
μ_{2sexe}	-0.0057	0.0002	-28.68	0.0000
$\mu_{2nbenf1}$	0.0020	0.0003	5.77	0.0000
$\mu_{2nbenf2}$	0.0066	0.0006	10.74	0.0000
$\mu_{2nbenf3}$	0.0115	0.0006	19.97	0.0000
$\mu_{2nbenf4}$	0.0049	0.0004	12.88	0.0000
$\mu_{2nbenf5}$	0.0071	0.0005	12.95	0.0000
$\mu_{2nbenf6}$	0.0125	0.0006	20.50	0.0000
$\mu_{2nbenf7}$	0.0012	0.0006	1.97	0.0488
$\mu_{2nbenf8}$	0.0047	0.0007	6.38	0.0000
$\mu_{2nbenf9}$	0.0079	0.0008	9.31	0.0000
μ_{2men2}	-0.0055	0.0005	-11.71	0.0000
μ_{2men3}	-0.0073	0.0003	-21.42	0.0000
μ_{2men4}	-0.0113	0.0004	-31.21	0.0000
μ_{2men5}	-0.0069	0.0006	-11.33	0.0000
μ_{3etu}	1.3925	0.0112	124.19	0.0000

Le coefficient de l'instrument est significatif. Il est bien positif et son ordre de grandeur est cohérent avec ce qu'on avait précédement. Cependant, on est toujours en présence d'un instrument faible, il faudra donc être vigilant à la F-stat de la deuxième étape.

• Pour instrumenter le niveau de salaire qui est une variable catégorielle à 4 niveaux, nous avons d'abord pensé à utiliser un modèle polytomique non ordonné. Nous avons choisi un logit multinomial car les régresseurs sont des caractéristiques de l'individu et non de la modalité de la variable de l'intérêt. Néanmoins, au regard de la difficulté d'utiliser un tel modèle dans un 2SLS, nous avons préféré transformer le niveau d'étude en nombre d'années d'étude, et de nous ramener à un cas classique. Nous avons choisi

comme instrument la catégorie socio-professionnelle du père (CSPP) qui est une variable catégorielle à 8 niveaux. La description est fournie en annexe. Cette variable est très certainement correlée positivement au nombre d'années d'étude de l'individu. Si son père a fait de longues études, il est probable qu'il en fasse de même. De plus, la catégorie socio-professionnelle du père n'a que très peu de chance d'être endogène. Ce n'est pas un choix de l'individu. Elle ne va pas être correlée aux variables non observée du salaire de l'individu. Cependant, on pourrait observer des cas d'endogénéité si la CSPP reflétait une autre caractéristique de l'individu comme son origine par exemple. Prenons l'exemple d'un enfant d'immigré, dont le père est ouvrier. Pour le père, le fait d'être issu directement de l'immigration est probablement corrélé à sa CSP. Ensuite, le salaire de l'enfant est corrélé avec le fait qu'il soit de la première génération issue de l'immigration, ce qui est inobservé ici. Donc, cette caractéristique inobservée est corrélée avec la CSP du père, ce qui la rend endogène. Ainsi, on estime donc le nombre d'années d'étude par la catégorie socio-professionnelle, et tous les autres régresseurs exogènes du modèle :

$$s_{it}^{ind} = \phi_0 + \phi_1 CSSP_{it}^{ind} + \phi_2 X_{it}^{ind} + \phi_3 X_{it}^{dep} + \psi_{it}$$

Les résultats sont disponibles en table 7 et nous permettent d'obtenir $\widehat{s_{it}^{ind}} = \widehat{\phi_0} + \widehat{\phi_1} CSSP_{it}^{ind} + \widehat{\phi_2} X_{it}^{ind} + \widehat{\phi_3} X_{it}^{dep}$.

Les coefficients associés à l'intrument sont tous significatifs, donc la condition de rang est respectée. Ils ne sont pas proches de zéro, nous n'avons donc pas d'instruments faibles ici. De plus on peut voir que par rapport à la catégorie de référence (agriculteur), seul un enfant d'ouvrier aura des études moins longues en comparaison. On remarque aussi qu'un enfant de cadre a des études bien plus longues qu'un enfant d'agriculteur et même de toutes les autres catégories.

TABLE 7 – Instrumentation de s_{it}^{ind} sur $CSSP_{it}^{ind}$

Observations	467,738			
R^2	0.198			
Adjusted R^2	0.198			
Residual Std. Error	0.863 (df = 467714)			
F Statistic	5,017.666*** (df = 23 ; 467714)			
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01			
	Estimate	Std. Error	t value	Pr(>|t|)
ϕ_0	0.7031	0.0085	82.61	0.0000
ϕ_{1cssp2}	0.2739	0.0055	49.76	0.0000
ϕ_{1cssp3}	1.0278	0.0057	180.17	0.0000
ϕ_{1cssp4}	0.4844	0.0054	90.07	0.0000
ϕ_{1cssp5}	0.1787	0.0056	31.79	0.0000
ϕ_{1cssp6}	-0.1456	0.0046	-31.41	0.0000
ϕ_{1cssp7}	0.1651	0.0177	9.33	0.0000
ϕ_{1cssp8}	0.0194	0.0115	1.69	0.0919
ϕ_{2age}	-0.0063	0.0001	-53.54	0.0000
ϕ_{2sexe}	0.0983	0.0026	38.48	0.0000
$\phi_{2nbenf1}$	0.0091	0.0045	2.02	0.0436
$\phi_{2nbenf2}$	0.2101	0.0079	26.56	0.0000
$\phi_{2nbenf3}$	0.3338	0.0074	44.99	0.0000
$\phi_{2nbenf4}$	0.1296	0.0049	26.56	0.0000
$\phi_{2nbenf5}$	0.2527	0.0071	35.66	0.0000
$\phi_{2nbenf6}$	0.3279	0.0079	41.51	0.0000
$\phi_{2nbenf7}$	0.1716	0.0080	21.58	0.0000
$\phi_{2nbenf8}$	0.1942	0.0095	20.51	0.0000
$\phi_{2nbenf9}$	0.1979	0.0109	18.19	0.0000
ϕ_{2men2}	-0.3090	0.0061	-50.84	0.0000
ϕ_{2men3}	-0.0643	0.0044	-14.67	0.0000
ϕ_{2men4}	-0.2183	0.0047	-46.54	0.0000
ϕ_{2men5}	-0.2813	0.0079	-35.82	0.0000
ϕ_{3etu}	6.7012	0.0987	67.91	0.0000

- Nous allons maintenant réaliser la deuxième étape du 2SLS, c'est à dire estimer le modèle :

$$\ln(w_{it}) = \alpha + \beta\widehat{s_{it}^{ind}} + \delta\widehat{S^{dep_{it}}} + \gamma X_{it}^{ind} + \rho X_{it}^{dep} + \eta_{it}$$

Les résultats sont donnés dans la table 8 ci-après.

TABLE 8 – Estimation du modèle

Observations	467,738
R^2	0.430
Adjusted R^2	0.430
Residual Std. Error	0.067 (df = 467720)
F Statistic	20,768.000*** (df = 17 ; 467720)
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01

	Estimate	Std. Error	t value	Pr(>|t|)
α	6.6864	0.0135	494.61	0.0000
β	0.2875	0.0051	56.66	0.0000
δ	1.0049	0.0634	15.85	0.0000
γ_{age}	0.0148	0.0002	83.83	0.0000
γ_{sexe}	-0.3223	0.0038	-83.90	0.0000
γ_{nbenf1}	0.0798	0.0068	11.73	0.0000
γ_{nbenf2}	0.1311	0.0114	11.53	0.0000
γ_{nbenf3}	0.1210	0.0106	11.39	0.0000
γ_{nbenf4}	0.1337	0.0074	18.00	0.0000
γ_{nbenf5}	0.1508	0.0107	14.15	0.0000
γ_{nbenf6}	0.1245	0.0116	10.77	0.0000
γ_{nbenf7}	0.0329	0.0120	2.73	0.0064
γ_{nbenf8}	0.0722	0.0140	5.17	0.0000
γ_{nbenf9}	0.0963	0.0161	5.99	0.0000
γ_{men2}	-0.0467	0.0090	-5.19	0.0000
γ_{men3}	0.0257	0.0063	4.08	0.0000
γ_{men4}	-0.0107	0.0071	-1.52	0.1289
γ_{men5}	-0.1083	0.0116	-9.37	0.0000
ρ_{etu}	-0.5877	0.2843	-2.07	0.0387

On remarque tout d'abord que tous les coefficients sont significatifs, et que la nullité jointe des coefficients est fortement rejetée par la F-statistique. La valeur de cette dernière permet de ne pas se soucier d'avoir eu un instrument faible sur la première étape. Les signes des coefficients paraissent par ailleurs raisonnables. Par exemple, comme attendu, l'augmentation des années d'étude d'un individu, ou l'augmentation du capital humain dans le département ont toutes les deux des effets positifs sur le niveau de salaire. Comme précédemment, les individus ayant des enfants ont des salaires plus élevés que les individus sans enfants. De même, une complexification du type de ménage à plutôt un effet négatif. La proportion d'étudiant a toujours un effet négatif, ce qui pourrait s'expliquer par un effet de concurrence. Plus il y a d'étudiants dans le département, plus l'offre de travail est élevée.
L'effet du sexe sur le salaire n'a pas changé. Il le diminue de 32% pour les femmes. Il en est de même pour l'effet de l'age, qui reste identique. Un gain d'expérience augmente légèrement le salaire, même si comme on l'a vu dans l'équation de Mincer, il semble avoir un effet négatif au delà d'un certain seuil. Ce qui n'est pas pris en compte ici car on omet l'effet quadratique de l'âge.
Concernant les rendements privés de l'éducation, ils sont plus élevés que dans l'équation de Mincer. On pourrait l'expliquer simplement parce que la précision de la variable utilisée ici est bien moindre que celle que nous avions utilisé dans la première partie. En

effet, par soucis de simplicité, nous avions transformé les 4 niveaux d'étude en nombe d'années d'étude. Dans l'équation de Mincer, nous avions utilisé 11 niveaux. Ici, tous les diplomes sous le bac (CAP, etc..) passent au niveau 0 ; au delà du bac (comme BTS, IUT) passent directement dans la catégorie bac, les bac+4 sont considérés comme des licences bac+3, etc.... Il est donc cohérent ici que les rendements privés soient surestimés, car le nombre d'années est sous-estimé. Pour la même raison, ils sont surestimés par rapport à l'équation de la question 7. Nous avons un biais car la variable "nombre d'années d'étude" est en réalité mal observée par le fait de considérer faussement une variables catégorielle comme continue.
Les rendements sociaux néanmoins ont eux un effet moindre par rapport à la question 7. En instrumentant la variable du capital humain du département, nous avons essayé de réduire l'endogénéité de cette dernière. Comme nous l'avons expliqué dans la question 8, il pouvait y avoir des effets dus aux chocs de demande qui affectent la productivité relative des travailleurs, mais aussi des chocs d'offres, qui eux affectent l'attractivité du département. Il était donc difficile dans un premier temps d'estimer le biais induit par l'endogénéité, qui pouvait aussi bien être positif que négatif. Ici visiblement, on avait dans la première estimation du rendement social un biais positif. On avait tendance à la surestimer. Cela voulait dire que la variable du capital humain était correlée positivement avec les variables inobservées du résidu du niveau de salaire. Ici donc, une augmentation de 1% de la proportion de salariés diplômés du supérieur dans le département conduit à une augmentation de 1% du niveau de salaire.

Question 11)

L'insatisfaction grandissante concernant les systèmes éducatifs des pays développés ont amené les économiste à s'intéresser de près aux rendements privés de l'éducation. Cela permet de mesurer quantitativement l'efficacité et l'équité des systèmes, mais aussi de pouvoir les comparer d'un pays à l'autre. Comme on a pu le voir dans les statistiques descriptives, le nombre d'étudiant effectuant des études longues n'a pas arrêté de grandir sur les dernières décennies. Il est donc important de pouvoir mesurer le retour sur investissement de ces futurs travailleurs. Le premier modèle estimé est celui proposé par Mincer (1974) qui est assez simple mais explique bien l'effet de l'éduction sur le salaire. Notre estimation du modèle de Mincer explique environ 20% de la dispersion des salaires par la dispersion de l'éducation et de l'expérience. Dans la littérature, on observe en général environ 30%. Son avantage est donc sa simplicité d'interprétation et son pouvoir explicatif universellement vérifié, ce qui est rare en économie.
Le second modèle consiste en l'équation de Mincer complétée du sexe de l'individu. Les résultats sont très proches de ceux estimés avec l'équation du Mincer : des rendements privés égaux à 9% par année d'étude supplémentaire. De plus, la variable "sexe" permet de mettre en lumière un écart de 30% des salaires entre les hommes et les femmes, ce qui s'observe empiriquement.
Néanmoins, la contrepartie de la simplicité de l'équation de Mincer est ses limites. L'éducation par exemple est une variable fortement endogène. Il y a notamment le biais talent, qui provient de facteurs non observables corrélés avec l'éducation et le salaire, résultant en une surestimation du rendement de l'éducation. Plusieurs méthodes

peuvent être utilisées pour traîter le problème, comme le fait d'utiliser des différences premières dans le cas de données de panel. Cependant, les économistes Angrist et Krueger instrumentent le temps passé à l'école par le trimestre de naissance, en raison des lois qui interdisent de quitter l'école avant le 16e anniversaire. Une personne née en fin d'année est donc suceptible d'être scolarisé plus longtemps. D'autres variables instrumentales peuvent aussi être utilisées, comme la distance du foyer par rapport à l'université la plus proche.
En réalité, les économistes ont remarqué que les résultats obtenus avec l'équation simple de Mincer ne diffèrent pas grandement des résultats instrumentés. Guille et Skalli (1999) ont remarqué qu'en fait, le biais positif du talent était compensé avec le biais négatif d'erreur de mesure dans les enquêtes.
Pour finir, la limite la plus importante de l'approche Mincerienne est d'imposer un rendement marginal de l'éducation commun à tous les individus et constant pour tous les niveaux d'étude.
Une seconde méthode bien plus globale pour estimer les rendements de l'éducation s'appuie sur la théorie du capital humain. L'éducation est considérée comme un investissment qui engendre des coûts et des bénéfices. On cherche alors à estimer les gains nets de l'investissement éducatif de l'individu, et non plus l'effet causal des études sur le salaire.

Dans la seconde partie, on souhaite mesurer les rendements sociaux de l'éducation. Dans un premier modèle, on estime simplement par OLS les rendements sociaux de l'éducation, en prenant en compte des caractéristiques socio-démographique de l'individu et des caractéristiques propres au département. Les rendements privés sont différenciés selon le type de diplôme, à l'aide de variable catégorielles. On a cependant toujours un problème d'endogénéité qui entraine un biais positif des rendements sociaux. Une fois de plus, des caractéristiques propres au département et inobservées, liées à l'attractivité de ce dernier. Ces variables sont corrélées positivement à la proportion des salariés diplômés du supérieur car un département attractif aura tendance à attirer des grosses entreprises et des individus qualifiés.
Pour tenter de résoudre les problèmes d'endogénéité, nous avons estimé ce modèle en différences premières en utilisant l'aspect panel des données. De plus, nous avons instrumenté la variation de proportion de diplômés du supérieur par la création d'établissements du supérieur dans le département à l'année N-5. Le but étant d'estimer les rendements sociaux par un 2SlS. Nous nous sommes confronté à un instrument faible et à une difficulté d'estimer correctement la première étape du 2SlS. Celà entraine une deuxième étape non concluante.
Enfin, nous avons réalisé dans un dernier modèle une instrumentation du niveau individuel et général de diplôme. Le niveau individuel est instrumenté par la catégorie socio-proffessionnelle du père et le niveau général par le nombre d'établissement du supérieurs à l'année N-5 dans le département. En comparaison avec le modèle de la question 7, on voit que les rendements sociaux sont plus faibles. Ce qui confirme notre hypothèse de biais positif.
Estimer les rendements sociaux de l'éducation est aussi important car ce sont les gouver-

nements qui disposent des leviers nécessaires afin d'élever le niveau moyen de l'éducation de la population. La question est donc de savoir si l'éducation reçue par l'individu augmente sa productivité et son utilité à la vie en collectivité. Les rendements privés ne peuvent pas servir de base à la prise de décisions politiques. En effet, le salaire n'est pas uniquement dû à la productivité, mais aussi des modes de négociations, de la présence d'un marché du travail dual, etc... De plus, la théorie du signaling admet effectivement un cadre où la relation croissante entre le salaire et l'éducation est compatible avec l'absence d'effet de l'éducation sur la productivité, à cause de l'asymétrie d'information. Les rendements sociaux suggèrent qu'il y a une externalité positive, où l'éducation des travailleurs qualifiés a un effet bénéfique sur le bien-être des autres.

3 Annexe

3.1 Description des variables Nombre d'enfants et Type de ménage

Type de Ménage
1 - Ménage d'une seule personne
2 - Familles monoparentales
3 - Couples sans enfant
4 - Couple avec enfant(s)
5 - Ménages complexes de plus d'une personne

Nombre d'enfants
0 - Pas d'enfant de moins de 18 ans
1 - Un enfant de 6 a 17ans
2 - Un enfant de 3 a 5ans
3 - Un enfant de moins de 3 ans
4 - Deux enfants, dont le plus jeune a de 6 a 17 ans
5 - Deux enfants, dont le plus jeune a de 3 a 5 ans
6 - Deux enfants, dont le plus jeune a moins de 3 ans
7 - Trois enfants ou plus, dont le plus jeune a de 6 a 17 ans
8 - Trois enfants ou plus, dont le plus jeune a de 3 a 5 ans
9 - Trois enfants ou plus, dont le plus jeune a moins de 3 ans

Catégorie socio-professionnelle du père
1 - Agriculteurs
2 - Artisans/commercants
3 - Cadres
4 - Professeurs, Clergé
5 - Fonctionnaires
6 - Ouvriers
7 - Retraités
8 - Inactifs

3.2 Logit multinomial du niveau de diplôme sur les CSSP

Modalité de base : sans diplôme.

Table 9 – Estimation du modèle

Observations	467,738
LR chi2(21)	79536.08
Prob > chi2	0.0000
Log likelihood	-499056.88
Pseudo R2	0.0738
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01

Diplome		Estimate	Std. Error	t value	Pr(>|t|)
0	(base outcome)				
1					
	$Icspp_2$	.5290749	.0143821	36.79	0.000
	$Icspp_3$	1.646637	.0174442	94.39	0.000
	$Icspp_4$	1.04279	.0142064	73.40	0.000
	$Icspp_5$	.5353428	.0145097	36.90	0.000
	$Icspp_6$	-.1652303	.012131	-13.62	0.000
	$Icspp_7$	.3863669	.0452763	8.53	0.000
	$Icspp_8$	.0458413	.0306799	1.49	0.135
	cons	-.7906726	.0108536	-72.85	0.000
2					
	$Icspp_2$	.9500108	.0262193	36.23	0.000
	$Icspp_3$	2.660339	.0266588	99.79	0.000
	$Icspp_4$	1.727309	.0248643	69.47	0.000
	$Icspp_5$	.9048469	.0265658	34.06	0.000
	$Icspp_6$	-.1403685	.0243037	-5.78	0.000
	$Icspp_7$	.8509855	.0717718	11.86	0.000
	$Icspp_8$	.4254832	.052781	8.06	0.000
	cons	-2.470528	.021715	-113.77	0.000
3					
	$Icspp_2$	1.179574	.0267004	44.18	0.000
	$Icspp_3$	3.335172	.0266488	125.15	0.000
	$Icspp_4$	1.776214	.0258908	68.60	0.000
	$Icspp_5$	.7910224	.028268	27.98	0.000
	$Icspp_6$	-.4834631	.0264493	-18.28	0.000
	$Icspp_7$	.4953889	.0863971	5.73	0.000
	$Icspp_8$	.5109385	.0535412	9.54	0.000
	cons	-2.572513	.0227645	-113.01	0.000