Université de Lille

Faculté des Sciences économiques et sociales

Etude prédictive de la capacité d'obtenir un haut salaire

M1 DS

Introduction

Le Centre d'études et de recherches sur les qualifications CEREQ est un pôle d'études et de recherche au service des professionnels, des décideurs, des partenaires sociaux et plus largement de tous les acteurs de la formation, du travail et de l'emploi.

Ainsi, le centre d'études et de recherches sur les qualification a mis en place un dispositif d'enquêtes original qui permet d'étudier l'accès à l'emploi des jeunes à l'issue de leur formation initiale. Tous les trois ans, une nouvelle enquête est réalisée auprès de jeunes qui ont en commun d'être sortis du système éducatif la même année quel que soit le niveau ou le domaine de formation atteint, d'où la notion de « génération ». En effet, les jeunes diplômés sont interrogés sur leur situation actuelle et leur parcours passé à savoir le dernier diplôme obtenu , le type de formation (apprentissage ou initial) leurs caractéristiques personnelles tel que le sexe et, âge.les jeunes salariés sont également intérrogés sur la situation de leurs parents (position professionnelle et niveau d'étude).

Nous sommes amenés à faire une étude de cas pour la mise en place d'un modèle de prévision destiné à mesurer la capacité d'un jeune diplômé arrivé sur le marché de travail depuis 3 ans à obtenir un haut salaire.Dans le cadre de notre étude un jeune salarié sera considéré comme ayant un haut salaire s'il perçoit un salaire net mensuel de plus de 2000€. Notre unité d'observation c'est les 6000 observations qui sont les réponses des jeunes sortis du système scolaire en 2013 trois ans après

Dans un premier temps, nous allons commencer par nettoyer notre base de données des données manquantes et des données atypiques pour ensuite pouvoir faire nos analyses statistiques sur les différentes variables dans le but de déterminer les variables qui influencent le plus le salaire.

Afin de répondre au mieux à notre problématique qui est la mise en place d'un modèle prédictif destiné à mesurer la probabilité qu'un jeune étudiant diplômé depuis 3 ans à obtenir un salaire mensuel net supérieur à 2000 €, plusieurs approches économétriques ont été élaborées. Nous avons procédé par trois modèles : linéaire , logit et probit .

Après l'analyse des statistiques descriptives réalisées , nous allons présenter les résultats de chaque modèle ainsi que des tests sur sa validité et sa capacité explicative. Finalement nous allons procéder à une analyse comparative des différents modèles afin de déterminer le modèle le plus adapté pour répondre à notre problématique.

Etude préalable des données

1) Présentation de la base de données et repérage des valeurs manquantes et aberrantes :

Nous disposons d'une base de données de 6000 observations, c'est-à-dire un échantillon de 6000 étudiants diplômés après trois ans sur le marché du travail regroupant le salaire de ces derniers ainsi que d'autres caractéristiques :

- Age de l'enquêté
- Sexe de l'enquêté
- Diplôme le plus haut obtenu
- Niveau d'études de la mère
- Niveau d'études du père
- Sortant de formation par voie apprentissage
- Position professionnel de la mère à la fin des études
- Position professionnel de la mère à la fin des études

2) Analyse des valeurs manquantes de notre base :

Après les procédures statistiques effectué nous avons repéré des variables manquantes pour les variables suivantes :

- Le niveau d'étude du père CA11 :124 variables manquantes.
- Le niveau d'étude de la mère CA12 : 50 variables manquantes.

Nous avons décidé de supprimer ces valeurs manquantes puisqu'elles représentent moins de 3% de l'échantillon de l'étude.

3) Analyse des valeurs aberrantes de notre base :

Une valeur est considérée comme aberrante est une donnée observée pour une variable qui semble anormale au regard des autres valeurs dont on dispose. La qualité du modèle dépendra largement du soin apporté à cette recherche et à la façon de traiter ces données. Afin de repérer ces éventuelles données nous avons regardé les valeurs maximales et minimales de chaque variable.

Nous avons détecté des variables aberrantes pour les variables suivantes :

- Âge : moins de l'âge de 14 ans et plus de 35 ans.
- Salaire : mois de 1142€ et plus de 3500€.

Afin d'avoir une étude plus pertinente nous avons décidé de concentrer notre étude sur les individus ayant un âge entre 14 ans et 35 ans et un salaire supérieur ou égal au SMIC en 2016,1143€ et inférieur à 3500€ par mois. En effet cette population représente les salariés à temps plein qui est notre population d'intérêt.

4) Analyse des valeurs à faible occurrence :

Pour les variable âge, plus haut diplôme obtenu en 2013 , position professionnelle de la mère et la position professionnelle du père on remarque qu'on a des modalités avec des fréquence de moins de 3% ce qui risque de rendre nos résultats moins significatifs.Nous avons décidé de faire des regroupements pour ces variables:

Nous avons décidé de faire un codage pour la variable âge sous la forme qualitative et de la regrouper en 4 tranches:(modalités)

- tranche 1: les individus ayant un âge inférieur ou égale à 20 ans
- tranche 2: les individus ayant un âge entre 21 ans et 25 ans
- tranche 3: les individus ayant un âge entre 26 ans et 30 ans
- tranche 4: les individus ayant un âge entre 31 ans et 35 ans

Regroupement de la variable diplôme plus haut obtenu :

- Non diplômé
- CAP-BEP-MC
- Baccalauréat
- BTS,DUT ou autre bac+2
- Bac+2 ou Bac +3 santé social
- Bac+2 à Bac+4
- Bac+5
- Doctorat

Regroupement de la variable position professionnelle de la mère à la date de fin d'études :

Ne pas citer : les NSP, les individus décédés ou ceux n'ayant jamais travaillé.

Nous avons effectué ce même regroupement pour la position professionnelle du père à la date de fin d'étude.

5) Analyse de la variable d'intérêt le salaire :

Comparaison des salaires moyens et médians de notre échantillon à ceux de l'ensemble de la population française à la même date

Pour exprimer la représentativité de notre échantillon à l'ensemble des individus français, nous décidons de faire une comparaison des indicateurs statistiques de notre variable d'intérêt,le salaire, avec le salaire de l'ensemble de la population française sur la même période.

caractéristiques statistiques de la variable salaire : SALPRSFIN			
Moyenne	1789.817	Ecart-type	536.43022
Médiane	1633.297	Variance	287757

Nous remarquons que la moitié des diplômés de notre échantillon gagne un salaire mensuel net supérieur à 1633€ et l'autre moitié gagne un salaire mensuel net inférieur à 1633€ .

D'après l'INSEE le salaire médian mensuel net de la population française en 2016 s'élève à 1789€.

Globalement nous pouvons conclure qu'en 2016 le salaire médian des individus de notre échantillon est inférieur à celui du salaire médian des français sur la même période .

En ce qui concerne le salaire moyen des jeunes diplômées de notre échantillon nous constatons qu'il était relativement bas par rapport au salaire moyen de l'ensemble de la population française puisqu'il a atteint 2 238 € net mensuel contre seulement 1789.817€ pour le salaire moyen des individus de notre base.

Nous pouvons conclure que notre échantillon est représentatif au regard de l'ensemble de la population française.

Création de la variable d'intérêt le salaire sous forme dichotomique : « obtention d'un salaire supérieur à 2000€ ».

Lors de notre étude , notre intérêt se porte principalement sur la proportion des étudiants ayant un salaire supérieur à 2000€. Pour cette raison nous avons décidé de faire un regroupement de la variable salaire en variable dichotomique .En effet ce regroupement permettra de calculer facilement les probabilités estimées.

Dans la suite de la modélisation cette variable sera recodée en dummy et prendra la valeur 1 pour un salaire supérieur à 2000 et 0 pour un salaire inférieur 2000.

Y : Tranche de salaire	Fréquence	Pourcentage
Salaire inférieur à 2000€	1515	25.25%
salaire supérieur à 2000€	4485	74.75%

Le tableau ci-dessus nous permet de visualiser le pourcentage des deux tranches selon le salaire. En effet les individus ayant un salaire supérieur à 2000€ représentent 74.75% de notre échantillon contre 25.25% des individus qui ont un salaire inférieur à 2000€. Nous concluons que la majorité des individus de notre échantillon ont un salaire supérieur à 2000€.

3) Analyse du lien entre les variables.

Afin d'étudier l'influence des variables prises en compte sur le salaire . Il est impératif de faire un choix sur les variables à garder dans le modèle dans la mesure où il peut exister des fortes corrélations entre ainsi que de discerner le lien entre le salaire et les autres variables .

Pour cela nous avons décidé de faire un test de V de Cramer qui permet de comparer l'intensité du lien entre les deux variables étudiées et un test du Khi deux afin de déterminer si les variables sont interdépendantes ou indépendantes dans le cas où il n'existe aucun lien statistique entre elles.

			salaire	Sexe	Age en 2013	plus haut diplôme obtenu en 2013	la position professionnelle du père à la date de fin d'études	la position professionnelle de la mère à la date de fin d'études	le niveau d'études de la mère	le niveau d'études du pére	formation dispensée par voie d'apprentissage
Salaire	V de cramer			-0,1043	0.3951	0.5978	0.2275	0.2161	0.2225	0.2633	0.0844
	Khi 2 : prob			<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
Sexe	V de cramer				0.2199	0.5223	0.0519	0.0669	0.0664	0.0302	0.1876
	Khi 2 : prob				<.0001	<.0001	0.0005	<.0001	0.0005	0.5443	<.0001
Age en 2013	V de cramer					0.2909	0.1619	0.1552	0.1539	0.1618	0.2384
	Khi 2 : prob					<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
plus haut diplôme obtenu en 2013	V de cramer						0.2013	0.1935	0.1827	0.1887	0.4550
	Khi 2 : prob						<.0001	<.0001	<.0001	<.0001	<.0001
La position professionnelle du père à la date de fin d'études	V de cramer							0.3288	0.2206	0.3301	0.1146
	Khi 2 : prob							<.0001	<.0001	<.0001	<.0001
La position professionnelle de la mère à la date de fin	V de cramer								0.3117	0.2039	0.1581
	Khi 2 : prob								<.0001	<.0001	<.0001
Le niveau d'études de la mère	V de cramer									0.3667	0.1470
	Khi 2 : prob									<.0001	<.0001
Le niveau d'études du père	V de cramer										0.1538
	Khi 2 : prob										<.0001

Relation faible

Relation moyenne

Relation forte

Test du khi-2 et V de cramer

On ne va pas s'intéresser à la valeur du khi-2 car cette dernière augmente avec l'augmentation du nombre d'observations mais on va plutôt s'intéresser à la statistique du test qui est plus significative. On teste l'indépendance des fréquences des modalités des variables.

On pose deux hypothèses :

· H0 : Deux variables sont indépendantes

· H1 : Deux variables ne sont pas indépendantes

On pose le seuil de significativité égale à α=5% On pose la règle de décision basée sur la p-value est donc de refuser H0 .si le niveau de la p-value est inférieure à un seuil α donnée.

En ce qui concerne l'indépendance du salaire de l'individu et son âge

On pose deux hypothèses :

· H0 : Le salaire de l'individu et son âge sont indépendants

· H1 : Le salaire de l'individu et son âge ne sont pas indépendants

D'après le tableau ci-dessus ,nous avons une p-value<0.0001, inférieur au seuil de significativité « 5% » donc on a un risque très faible de se tromper en rejetant HO.

Conclusion : on rejette HO, le salaire et l'âge des individus ne sont pas indépendants.

En générale d'après le tableau ci-dessus, on remarque que toutes nos variables sont interdépendantes, nous concluons donc qu'il existe une association statistiquement significative entre les variables sauf la relation entre le sexe et le niveau d'étude du père qui n'est pas significative.

En ce qui concerne le V de cramer, qui nous permet de tester l'intensité du lien entre les deux variables, plus il est proche de zéro, plus il y a indépendance entre les deux variables étudiées. Il vaut 1 en cas de complète dépendance.

D'une part par rapport à l'étude de liens entre notre variable d'intérêt qui est le salaire et nos variables explicatives, nous remarquons que la valeur de V de cramer la plus élevé correspond l'intersection entre le salaire et le diplôme le plus haut obtenu, elle est de 0.5968, on peut dire qu'il existe une dépendance forte entre ces deux variables. En effet, le diplôme le plus haut obtenu est la variable qui influence le plus le salaire. En ce qui concerne la variable sortant d'une formation dispensée par voie d'apprentissage nous remarquons que cette dernière a une faible influence sur le salaire puisque la valeur de V de cramer correspondant à l'intersection de ces variables est égale à 0.0844 donc on a une relation très faible entre ces deux variables.

D'autre part pour les liens entre les variables explicatives nous remarquons qu'on a des liens d'intensité moyenne voir forte entre les variables niveau d'étude de la mère et position professionnelle de cette dernière ainsi que pour le niveau d'étude du père et position professionnelle du père, nous avons égalons une relation forte entre le dernier diplôme de l'individu et le fait qu'il soit sortant d'une formation dispensée en voie d'apprentissage.

A l'issue de ces résultats nous avons décidé que la variable position professionnelle du père plutôt que de la position professionnelle de la mère, le niveau d'étude de la mère et le niveau d'étude du père puisque nous estimons que ces variables fournissent la même information. Nous avons également décidé de ne pas introduire la variable sortant d'une formation dispensée en voie d'apprentissage puisque nous avons un lien fort entre cette dernière et le diplôme le plus haut obtenu.

Modélisation

Après étude de la base de données et les diverses statistiques réalisées, on passe à la modélisation.

Pour notre modélisation nous avons décidé de transformer toutes les modalités des variables que nous avons à des variables dummy.(avec deux modalités 0 ou 1)

variables salaire

- SALPES FN : salaire mensuel net en fin de séquence des salaires

- Y Tranche de salaire = 0 => salaire inférieure à 2000€

  Tranche de salaire= 1 => salaire supérieur à 2000€

variable age

age_inf_20 : les individus ayant un âge inférieur ou égale à 20 ans

age_21_25 : les individus ayant un âge entre 21 ans et 25 ans

age_26_30 : les individus ayant un âge entre 26 ans et 30 ans

age_31_35 : les individus ayant un âge entre 31 ans et 35 ans

variable sexe

Femme : si l'individus tes une femme la variable prend la valeurs 1 sinon homme.

variable Apprentissage

non_Apprenti : si sortant de formation par voie d'apprentissage la variable prend la valeur 0

variable plus haut diplôme obtenu

non_diplome, cap_bep, Bac, Bac_plus_2_Bts_Dut, Bac_plus_2_3_sante , Bac_plus_3_4_hors_sante, Bac_plus_5_M2 , Doctorat

variable position professionel du père

P_ouvrier, P_employe, P_Techni_agent_vendeur, P_Cadre_ingenieur, P_Artisan_chef

A présent,nous allons maintenant utiliser trois méthodes de modélisation à savoir le modèle linéaire, le modèle LOGIT et le modèle PROBIT. Nous cherchons à estimer la probabilité d'avoir un salaire supérieur à 2000€ pour un individu sorti du système scolaire depuis trois ans.

Dans la démarche d'une modélisation, un choix de variables de références s'impose. Ces différentes modalités de référence seront maintenues pour les trois modèles

- *Choix de variable de référence*

Pour le choix de la situation de référence nous avons décidé de choisir la modalité qui correspond à la fréquence intermédiaire de chaque variable.

Variable	***Référence***
Âge	Entre 20 ans et 25 ans
Sexe	homme
Apprentissage	Non apprentissage
Diplôme le plus haut	Bac
Position professionnelle du père	Technicien/agent de maîtrise/vendeur

Analyse des résultats de la régression de différents modèles

L'analyse de régression est une méthode statistique de modélisation des relations entre différentes variables. En effet on suppose qu'il existe une relation de causalité entre la variable à expliquer et les variables explicatives. La valeur de la variable explicative affecte la valeur de la variable expliquée. Elle est utilisée pour décrire et analyser les relations entre les données. L'analyse de régression permet de réaliser des prédictions, les relations entre les données étant utilisées comme une base pour la prévision et la conception d'un modèle de prédiction.

Modèle linéaire de détermination de la probabilité d'obtenir un salaire supérieur à 2000€ (Salaire latente) :

➢ *Résultat de l'estimation*

Le tableau ci-dessous présente la modélisation de la variable à expliquer SALPRSFIN de type quantitatif qui représente le salaire mensuel

variable à expliquer : SALPES FN

Variable	DDL	Valeur estimée des paramètres	Pr > |t|
Intercept	1	1667.34396	<.0001
âge inférieur à 20 ans	1	-69.80577	0.0003
âge entre 26 ans et 30 ans	1	19.83514	0.3835
âge entre 31 ans et 35 ans	1	43.09268	0.2517
Femme	1	-116.08561	<.0001
Non Apprentissage	1	-72.48092	<.0001
Père Ouvrier	1	-67.22039	0.0033
Père Employé	1	-46.76598	0.0417
Père Cadre ingénieur	1	11.01641	0.6149
Père Artisan chef d'entreprise	1	-11.20138	0.6236
non diplômé	1	-35.15227	0.3973
Cap Bep	1	-39.79603	0.0977
Bac plus 2/Bts/Dut	1	69.83722	0.0029
Bac Plus 2/3 santé	1	256.74650	<.0001
Bac plus 3/4 hors santé	1	140.58637	<.0001
Bac plus 5/M2	1	568.38535	<.0001
Doctorat	1	850.22953	<.0001
R carré		**0.3767**	
R car. ajust		**0.3748**	

➢ *interprétation des résultats de la régression du modèle*

- Le salaire d'un homme âgé entre 20 ans et 25 ans qui a le bac sortant d'une formation apprentissage et la position professionnelle de son père est de catégorie "technicien/agent de maîtrise/vendeur" est égale à 1667.34€
- Le fait d'être une femme fait baisser le salaire de -116.08€ par rapport à un homme,toutes choses égale par ailleurs

➢ *Test statistique de la pertinence du modèle*

❖ *Test de fisher:*

Afin de tester la significativité globale du modèle nous allons effectuer un test de Fisher

On pose deux hypothèses :

H0 : tous les paramètres sont nuls

H1 : au moins un paramètre est non nul

On pose le seuil de significativité égale à α= 5% On pose la règle de décision basée sur la p-value est donc de refuser H0 si le niveau de la p-value est inférieure à un seuil α donnée

On remarque qu'on a une P-value <.0001 donc on rejette H0.

On conclut que notre modèle est significatif.

❖ *Test de student:*

Afin de tester la significativité partielle c'est à dire de tester la significativité de chaque variable nous allons établir le test de Student

Test de student t=(b^/σ^)

On pose deux hypothèses :

- H0: coefficient estimé est non significative
- H1: coefficient estimé est significativement différent de 0

On pose le seuil de significativité égale à α= 5% On pose la règle de décision basée sur la p-value est donc de refuser H0 si le niveau de la p-value est inférieure à un seuil α donnée

Test de student pour la variable Sexe la p-value < 0,001 , moins de 1% de chance de se tromper en rejetant H0 -> on rejette H0

➔ D'après les résultats obtenus nous constatons que les modalités de la variable"position professionel du père" sont moins significatives par rapport au reste des variables

➢ *Coefficient de détermination du modèle*

D'après le tableau ci-dessus nous remarquons qu'on a un coefficient de détermination égale à 0.37 donc nous pouvons dire que 37% de la variation du salaire des étudiants diplômés après 3 ans sur le marché du travail est expliqué par notre modèle.

➢ *score du modèle*

Le score permet de déterminer le profil le plus favorable ou défavorable pour obtenir un salaire supérieur à 2000€.

Score modèle						
Paramètre	Estimation	Variables	Minimum	Maximum	Ecart	Score
Formation apprentissage	0,00	Sortant de formation par voie apprentissage (CFAD)	-72,48	0,00	72,48	6
Formation initial	-72,48					0
Homme	0,00	Sexe	-116,09	0,00	116,09	10
Femme	-116,09					0
Age entre 20 ans et 25 ans	0,00	Âge	-69,81	43,09	112,90	6
âge inférieur à 20 ans	-69,81					0
âge entre 26 ans et 30 ans	19,84					8
âge entre 31 ans et 35 ans	43,09					9
BAC	0,00	Plus haut diplôme obtenu (PHD)	-39,80	850,23	890,03	3
non diplômé	-35,152					0
CAP_BEP_MC	-39,796					0
Bac plus 2/Bts/Dut	69,84					9

Bac Plus 2/3 santé	256,75					25
Bac plus 3/4 hors santé	140,59					15
Bac plus 5/M2	568,39					51
Doctorat	850,23					75
La somme des écarts		1191,49				

Les résultats de notre scoring nous permettent ainsi de définir deux profils :

- le profil le plus favorable à l'obtention d'un haut salaire :
  - ➔ Homme
  - ➔ Sortant d'une formation par voie d'apprentissage
  - ➔ Age entre 26 ans et 30 ans
  - ➔ plus haut diplôme obtenu est le Doctorat

- le profil le moins favorable à l'obtention d'un haut salaire :
  - ➔ Femme
  - ➔ Non Sortant d'une formation par voie d'apprentissage
  - ➔ Âge inférieur à 20 ans
  - ➔ plus haut diplôme obtenu est CAP/BEP/MC

II. Modèle linéaire de détermination de la probabilité d'obtenir un salaire supérieur à 2000€ (Salaire qualitative) :

On cherche à estimer la probabilité d'obtenir un haut salaire (> 2000€) pour un individu sorti du système scolaire depuis 3 ans. Autrement dit, quelles sont les chances d'obtention d'un salaire supérieur à 2000€, 3 ans après la fin des études ?

La différence entre ce modèle et le modèle précédent c'est l'utilisation de la variable à expliquer en dummy qui nous aide à identifier les individus qui ont un salaire supérieur à 2000 € 3 ans après la fin des études et les autres individus qui ont un salaire inférieur à 2000€.

➢ *Résultat de l'estimation*

La variable expliquée du modèle présenté ci dessous est la variable dummy salaire Y , cela va nous aider à identifier le positionnement des individus par rapport à un salaire de 2000€ 3 ans après la fin de leurs études.

variable à expliquer : Y (Tranche de salaire)

Variable	DDL	Valeur estimée des paramètres	Pr > |t|
Intercept	1	0.17543	<.0001
âge inférieur à 20 ans	1	-0.03368	0.0402
âge entre 26 ans et 30 ans	1	0.02522	0.1885
âge entre 31 ans et 35 ans	1	-0.03294	0.2982
Femme	1	-0.09345	<.0001
Non Apprentissage	1	-0.04011	0.0059
Père Ouvrier	1	-0.05735	0.0029
Père Employé	1	-0.04082	0.0349
Père Cadre ingénieur	1	-0.00772	0.6755
Père Artisan chef d'entreprise	1	-0.02364	0.2190
non diplômé	1	-0.02157	0.5376
Cap Bep	1	-0.03244	0.1091
Bac plus 2/Bts/Dut	1	0.04324	0.0288
Bac Plus 2/3 santé	1	0.10854	<.0001
Bac plus 3/4 hors santé	1	0.07867	<.0001
Bac plus 5/M2	1	0.42620	<.0001
Doctorat	1	0.63892	<.0001
R carré		**0.3281**	
R car. ajust		**0.3261**	

➢ *interprétation des résultats de la régression du modèle*

- le fait d'être un homme âgé entre 20 ans et 25 ans qui a le bac sortant d'une formation apprentissage et la position professionnelle de son père est de catégorie "technicien/agent de maîtrise/vendeur" a une probabilité 17% d'avoir un salaire supérieur à 2000€

- [0.42620 +0.17543-0.09345] = 0,50818 = 50% de chance d'avoir un salaire supérieur à 2000€ pour un profil d'une femme âgé entre 20 ans et 25 ans qui a un diplome de bac plus 5/M2 sortant d'une formation apprentissage et la position professionnelle de son père est de catégorie "technicien/agent de maîtrise/vendeur".

Afin de tester la qualité de notre modèle et de s'assurer que notre modèle est significatif, on vérifie les hypothèses des moindres carrés ordinaires, nous allons effectuer plusieurs tests statistiques.

➢ *Test statistique de la pertinence du modèle*

❖ *Test de fisher:*

Afin de tester la significativité globale du modèle nous allons effectuer un test de Fisher

On pose deux hypothèses :

H0 : tous les paramètres sont nuls

H1 : au moins un paramètre est non nul

On pose le seuil de significativité égale à α= 5% On pose la règle de décision basée sur la p-value est donc de refuser H0 si le niveau de la p-value est inférieure à un seuil α donnée

On remarque qu'on a une P-value <.0001 donc on rejette H0.

On conclut que notre modèle est significatif.

❖ *Test de student:*

Afin de tester la significativité partielle c'est à dire de tester la significativité de chaque variable nous allons établir le test de Student

Test de student t=(b^/σ^)

On pose deux hypothèses :

- H0: coefficient estimé est non significative
- H1: coefficient estimé est significativement différent de 0

On pose le seuil de significativité égale à α= 5% On pose la règle de décision basée sur la p-value est donc de refuser H0 si le niveau de la p-value est inférieure à un seuil α donnée

Test de student pour la variable Sexe la p-value < 0,001 , moins de 1% de chance de se tromper en rejetant H0 -> on rejette H0

➔ Nous pouvons conclure d'après ce tableau et les résultats du test de Student que toutes les modalités nos variables sont significatives à l'exception de la modalité aucun diplôme de la variable plus haut diplôme obtenu et la modalité cadre/ingénieur pour la variable position professionnelle du père.

❖ *Test Durbin and watson*

Durbin-Watson D	1.902
Nombre d'observations	5478
Autocorrélation de 1er ordre	0.049

Afin de vérifier la qualité du modèle on a utilisé le test de Durbin-Watson pour détecter s'il y a l'autocorrélation entre les résidus* d'une régression linéaire qui peut entraîner une mauvaise estimation des paramètres (variables dont dépendent les coefficients de l'équation).

* : Un résidu est dans une régression le terme qui n'est pas expliqué par les autres variables. En effet, c'est l'ensemble des facteurs variables qui ne s'inscrit pas dans la formule estimée.

On pose deux hypothèses et on établit un test statistique qui va nous permettre de choisir la bonne conclusion entre ces deux hypothèses :

- H0 : absence d'autocorrélation des résidus
- H1 : présence d'autocorrélation des résidus

La statistique DW prend ses valeurs entre 0 (auto-corrélation linéaire positive) et 4 (auto-corrélation linéaire négative). L'hypothèse nulle est retenue si la statistique a une valeur proche de 2 (pas d'auto-corrélation linéaire).

Nous avons obtenu une valeur du test de DW proche de 2 (1.9) pour notre modèle donc on accepte l'hypothèse HO.

Conclusion : nous concluions qu'il y a une absence d'autocorrélation des résidus ce qui signifie que l'augmentation observée dans un intervalle de temps n'entraîne pas une augmentation proportionnelle de l'intervalle de temps décalé ce qui implique que la valeur d'une variable à un instant donné n'est pas liée à sa valeur à un instant antérieur.

❖ *Test d'hétéroscédasticité:*

Nous avons décidé de réaliser le test de Gleijer pour tester l'existence ou l'absence de l'hétéroscédasticité dans notre modèle.

Test de Gleijer Le test de Gleijer se décline entre deux hypothèses :

- ∀ i (εi) = k*Ti (l'écart-type du résidu est proportionnel à la taille Ti de l'unité i).

- ∀ i (εi) = k*Ti (la variance du résidu est proportionnelle à la taille Ti de l'unité i).

Le test porte donc sur k.

- H0: k = 0 (homoscédasticité).
- H1: k ≠ 0 (hétéroscédasticité).

On pose la règle de décision suivante: si k significativement différent de zéro on conclura à l'Hétéroscédasticité du résidu, si en revanche k n'est pas significativement différent de 0, on conclura à l'homoscédasticité du résidu.

Première régression avec la variable absresid = la valeur absolue de résidu

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	0.17760	0.00313	56.80	<.0001
Y	1	0.32367	0.00621	52.16	<.0001

Deuxième régression avec la variable carresid = le résidu au carré

Variable	DDL	Valeur estimée des paramètres	Erreur	Valeur du test t	Pr > |t|
Intercept	1	0.06587	0.00278	23.73	<.0001
Y	1	0.24186	0.00551	43.91	<.0001

Les tableaux représentent les résultats des deux régressions avec un indicateur de plausibilité H0 inférieur à 0,005.

Nous constatons l'existence de l'hétéroscédasticité .

afin de palier à cela voici quelques solutions:

- estimer la variance de l'erreur et utiliser les MCO sans autocorrélation .
- MCG moindre carré généralisé .

Mesures de la qualité d'ajustement du modèle

➢ *Coefficient de détermination du modèle*

Une fois un modèle élaboré, la question de sa solidité et de sa qualité reste primordiale. Plusieurs indicateurs peuvent permettre d'apprécier la qualité d'un modèle. Le premier critère est le coefficient de détermination. On peut également regarder la matrice de confusion.

D'après le tableau ci-dessus nous remarquons qu'on a un coefficient de détermination égale à 0.32 donc nous pouvons dire que 32% de la probabilité d'obtention d'un salaire supérieur à 2000.

➢ *la matrice de confusion de modèle*

La matrice de confusion est associée à un seuil de 50% permettant de confronter les valeurs prédites et les valeurs observées. Il permet de voir la qualité de la prévision. Cette matrice va nous permettre de calculer des taux d'erreur. Elle permet de juger de la qualité des prévisions du modèle.

Valeur réelle	Valeur prédite du salaire qualitatif		
	0	1	Total
0	3675	412	4087
1	535	856	1391
Total	4210	1268	5478

- *3675: c'est le nombre d' étudiants que nous avons estimés qu'ils ont un salaire inférieur à 2000€ et qu'ils ont effectivement en réalité un salaire inférieur à 2000€.*
- *535: c'est le nombre d'étudiants que notre modèle prévoit qu'ils ont un salaire inférieur à 2000€ alors qu'en réalité ils ont un salaire supérieur à 2000€.*
- *412: c'est le nombre d'étudiants que notre modèle prévoit qu'ils ont un salaire supérieur à 2000€ alors qu'en réalité ils ont un salaire inférieur à 2000€.*
- *856 : c'est le nombre d'étudiants que notre modèle prévoit qu'ils ont un salaire supérieur à 2000€ et qu'ils ont effectivement un salaire supérieur à 2000€.*

Calcul des taux d'erreur:

taux d'erreur alpha: C'est l'erreur de prévoir l'événement (0) alors que l'individu doit connaître l'événement (1).En effet Dans le cas de notre étude c'est l'erreur de prévoir qu'un étudiant a un salaire inférieur à 2000€ alors qu'en réalité ce dernier a un salaire supérieur à 2000€ .

α=(535/1391)*100=38.46%

Le taux d'erreur α est élevé. Cela veut dire que le modèle a tendance à prévoir l'événement 0 (salaire<2000 euros) pour des individus qui gagnent plus de 2000 €

taux d'erreur beta: C'est l'erreur de prévoir l'événement (1) alors que l'individu doit connaître l'événement (0).En effet Dans le cas de notre étude c'est l'erreur de prévoir qu'un

étudiant a un salaire supérieur à 2000€ alors qu'en réalité ce dernier a un salaire inférieur à 2000€ .

β=(412/4087)*100=10.08%

Le taux d'erreur β est faible donc notre modèle a tendance à bien estimer les gens qui gagnent plus de 2000 €.

taux d'erreur moyen: c'est la proportion de mal classés, la probabilité de mal classer un individu pris au hasard dans notre échantillon.En effet, ce taux est une synthèse entre les deux taux d'erreur. Il donne l'erreur moyen de prédiction du modèle.

taux moyen= ((535+412)/5478)*100=17.29%

De manière globale, les erreurs de prévision sont de 17.29%. La qualité du modèle est relativement bonne.

le tableau ci-dessous est un récapitulatif des taux d'erreurs

Taux d'erreur moyen	17,29%
Taux d'erreur alpha	38,46%
Taux d'erreur Beta	10,08%

➢ *le score du modèle*

Le score normalisé va nous permettre de déterminer le profil le plus favorable ou défavorable pour l'obtention d'un salaire supérieur à 2000€.Les variables utilisées dans l'établissement du score sont les variables significatives pour tous les modèles.

Score du modèle linéaire variable latente						
Paramètre	Estimation	Variables	Minimum	Maximum	Ecart	Score
Formation apprentissage	0	Sortant de formation par voie apprentissage (CFAD)	-0,04011	0	0,04011	4,64332847
Formation initial	-0,04011					0
Homme	0	Sexe	-0,09345	0	0,09345	10,818226

Femme	-0,09345					0
Age entre 20 ans et 25 ans	0	Âge	-0,03368	0,02522	0,0589	3,89896043
âge inférieur à 20 ans	-0,03368					0
âge entre 26 ans et 30 ans	0,02522					6,81855016
âge entre 31 ans et 35 ans	-0,03294					0,085666
BAC	0	Plus haut diplôme obtenu (PHD)	-0,03244	0,63892	0,67136	3,75541201
non diplômé	-0,02157					1,25836401
CAP_BEP_MC	-0,03244					0
Bac plus 2/Bts/Dut	0,04324					8,76108449
Bac Plus 2/3 santé	0,10854					16,3205297
Bac plus 3/4 hors santé	0,07867					12,8626334
Bac plus 5/M2	0,4262					53,0943947
Doctorat	0,63892					77,7198953
La somme des écarts		0,86382				

Les résultats de notre scoring nous permettent ainsi de définir deux profils :

- le profil le plus favorable à l'obtention d'un haut salaire :
  - ➔ Homme
  - ➔ Sortant d'une formation par voie d'apprentissage
  - ➔ Age entre 26 ans et 30 ans
  - ➔ plus haut diplôme obtenu est le Doctorat

- le profil le moins favorable à l'obtention d'un haut salaire :
  - ➔ Femme
  - ➔ Non Sortant d'une formation par voie d'apprentissage
  - ➔ Âge inférieur à 20 ans
  - ➔ plus haut diplôme obtenu est CAP/BEP/MC

➢ *Limites du modèle et solutions*

- ➔ Nous cherchons une probabilité donc, elle doit être comprise entre 0 et 1. Or nous trouvons des probabilités n'appartenant pas à cet intervalle.
- ➔ Le second problème pour un tel modèle est le problème d'hétéroscédasticité. À partir du test de Gleisjer, nous notons une normalité des résidus et de leur variance qui n'est pas constante ce qui va engendrer une inefficacité des estimateurs et l'estimation de la variance des estimateurs qui est biaisée.

Solutions :

- ➔ L'estimation de la variance de l'erreur et on utilise les MCO sans autocorrélation
- ➔ L' utilisation de la méthode des moindre carré généralisé MCG

III. Modèle Probit de détermination d'un salaire supérieur à 2000€

Nous allons maintenant utiliser un troisième modèle, le modèle probit est un modèle probabiliste basé sur une loi normale. Les coefficients d'un modèle probit se lisent en faisant une correspondance dans la table de la loi normale. Le modèle probit présente l'avantage d'avoir une justification théorique solide et d'être un modèle cohérent. La seule limite qui lui est reprochée est la difficulté d'interprétation des coefficients estimés.

➢ *Résultat de l'estimation*

variable à expliquer : Y (Tranche de salaire)

Paramètre	DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > khi-2
Intercept	1	-0.7255	0.2403	9.1170	0.0025
âge inférieur à 20 ans	1	-0.1190	0.0465	6.5486	0.0105
âge entre 26 ans et 30 ans	1	0.0435	0.0383	1.2865	0.2567
âge entre 31 ans et 35 ans	1	-0.0518	0.0645	0.6448	0.4220
Femme	1	-0.2052	0.0235	76.4793	<.0001
Non Apprentissage	1	-0.1157	0.0342	11.4289	0.0007
Père Ouvrier	1	0.2631	0.0852	9.5353	0.0020
Père Employé	1	0.1795	0.0840	4.5659	0.0326
Père Cadre ingénieur	1	0.0298	0.0763	0.1528	0.6959
Père Artisan chef d'entreprise	1	0.0902	0.0819	1.2146	0.2704
non diplômé	1	-0.1048	0.1185	0.7818	0.3766
Cap Bep	1	-0.1804	0.0698	6.6885	0.0097
Bac plus 2/Bts/Dut	1	0.1495	0.0502	8.8890	0.0029
Bac Plus 2/3 santé	1	0.3063	0.0500	37.5350	<.0001
Bac plus 3/4 hors santé	1	0.2310	0.0491	22.1284	<.0001
Bac plus 5/M2	1	0.7555	0.0441	293.3570	<.0001
Doctorat	1	1.0380	0.0589	310.2776	<.0001

➢ *interprétation des coefficients:*

intercept : -0.7255 fu(-0.7255)=23%

- En faisant ainsi la correspondance dans la table de la loi normale, le faite d'être un homme âgé entre 20 ans et 25 ans qui a le bac sortant d'une formation apprentissage et la position professionnelle de son père est de catégorie "technicien/agent de maîtrise/vendeur" a une probabilité 23% d'avoir un salaire supérieur à 2000€.

fu (-0.7255-0.2052)=18%

- C'est la variation marginale d'avoir un salaire supérieur à 2000€ impliqué par le fait d'être une femme 23%-18%= 5% Parmi les hommes ayant le bac âgé entre 20 ans et 25 ans et une position professionnelle du père est de catégorie "technicien/agent de maîtrise/vendeur"

➢ *Test de significativité globale : le test du maximum de vraisemblance*

Afin de tester la significativité globale du modèle nous allons établir le test du maximum de vraisemblance sur la base du critère de l'efficacité statistique puisqu'il est plus contraignant que le test de wald

on pose deux hypothèses:

- Ho : les 5 restrictions posées sont vraies
- H1 : les 5 restrictions posées sont fausses

la statistique du test :$X^2 = 2(LLC - LLR)$ avec LLC modèle complet et LLR modèle restreint

- H0 : $X^2 <= X^2alpha$
- H1 : $X^2 > X^2alpha$

On pose le seuil de significativité égale à α=5% On pose la règle de décision basée sur la p-value est donc de refuser H0 .si le niveau de la p-value est inférieure à un seuil α donnée.

D'après le tableau ci-dessus nous pouvons remarquer qu'on a une p-value*<.0001* donc nous allons rejeter H0

Test de l'hypothèse nulle globale			
Test	*khi-2*	*DDL*	*Pr > khi-2*
Rapport de vraisemblance	*1823.0603*	*16*	*<.0001*

Conclusion : notre modèle est globalement significatif.

➢ *Test de significativité partielle : Test de Wald*

Afin de tester la significativité partielle du modèle nous allons établir un test de Wald, nous allons alors tester nos variables une à une pour voir s'ils ont une action significative sur notre variable d'intérêt c'est-à-dire la probabilité d'avoir un salaire supérieur à 2000€.

Libellé	Khi-2 de Wald	DDL	Pr > khi-2
Formation apprentissage	11.4289	1	0.0007
Le sexe	76.4793	1	<.0001
Position professionnel du père	17.0277	4	0.0019
l'âge	9.9787	3	0.0187
Plus haut diplôme obtenu	628.7555	7	<.0001

On pose deux hypothèses :

- H0 : AJ=0 (variable n'agit pas sur Y)
- H1: AJ #0 (variable agit sur Y)

On pose le seuil de significativité égale à α=10% On pose la règle de décision basée sur la p-value est donc de refuser H0 .si le niveau de la p-value est inférieure à un seuil α donnée.

➔ on accepte HO si khi2 >khi2 de wald
➔ et on accepte H1 si khi2 alfa < khi2 wald

D'après les résultats du tableau ci-dessus nous remarquons qu'on a une p-value <10% pour toutes nos variables donc nous pouvons dire que toutes nos variables ont une action significative sur notre variable d'intérêt.

➢ *Mesure de la qualité de l'ajustement par le R^2*

Pour mesurer la qualité de l'ajustement nous allons regarder le pseudo R carré de Mcfadden puisque ce dernier est liée directement à la statistique du test (test de maximum du vraisemblance) comme le R carré du linéaire avec le test de fisher.En effet le Pseudo R carré de mcfadden présente les mêmes propriétés que le coefficient de détermination classique, c'est-à-dire compris entre 0 et 1 et croissant avec la qualité de l'ajustement. Il permet aussi d'effectuer une comparaison avec le coefficient de détermination du modèle linéaire.

R^2 de McFadden	0.2937

Le pseudo-R2 de McFadden sera utilisé pour ce modèle qui explique 29,37% les différences de salaire .

➢ *La matrice de confusion du modèle*

La matrice de confusion est associée à un seuil de 50% permettant de confronter les valeurs prédites et les valeurs observées. Il permet de voir la qualité de la prévision. Cette matrice va nous permettre de calculer des taux d'erreur. Elle permet de juger de la qualité des prévisions du modèle.

Valeur réelle	Valeur prédite du salaire qualitatif		
	0	1	Total
0	3695	392	4087
1	558	833	1391
Total	4253	1225	5478

- *3695: c'est le nombre d' étudiants que nous avons estimés qu'ils ont un salaire inférieur à 2000€ et qu'ils ont effectivement en réalité un salaire inférieur à 2000€.*
- *558: c'est le nombre d'étudiants que notre modèle prévoit qu'ils ont un salaire inférieur à 2000€ alors qu'en réalité ils ont un salaire supérieur à 2000€.*
- 392*: c'est le nombre d'étudiants que notre modèle prévoit qu'ils ont un salaire supérieur à 2000€ alors qu'en réalité ils ont un salaire inférieur à 2000€.*
- *833 : c'est le nombre d'étudiants que notre modèle prévoit qu'ils ont un salaire supérieur à 2000€ et qu'ils ont effectivement un salaire supérieur à 2000€.*

Calcul des taux d'erreur:

taux d'erreur alpha: C'est l'erreur de prévoir l'événement (0) alors que l'individu doit connaître l'événement (1).En effet Dans le cas de notre étude c'est l'erreur de prévoir qu'un étudiant a un salaire inférieur à 2000€ alors qu'en réalité ce dernier a un salaire supérieur à 2000€ .

α=(558/1391)*100=40,11%

Le taux d'erreur α est élevé. Cela veut dire que le modèle a tendance à prévoir l'événement 0 (salaire<2000 euros) pour des individus qui gagnent plus de 2000 €

taux d'erreur beta: C'est l'erreur de prévoir l'événement (1) alors que l'individu doit connaître l'événement (0).En effet Dans le cas de notre étude c'est l'erreur de prévoir qu'un étudiant a un salaire supérieur à 2000€ alors qu'en réalité ce dernier a un salaire inférieur à 2000€ .

β=(392/4087)*100=9,59%

Le taux d'erreur β est faible donc notre modèle a tendance à bien estimer les gens qui gagnent plus de 2000 €.

taux d'erreur moyen: c'est la proportion de mal classés, la probabilité de mal classer un individu pris au hasard dans notre échantillon.En effet, ce taux est une synthèse entre les deux taux d'erreur. Il donne l'erreur moyen de prédiction du modèle.

taux moyen= ((558+392)/5478)*100=17,34%

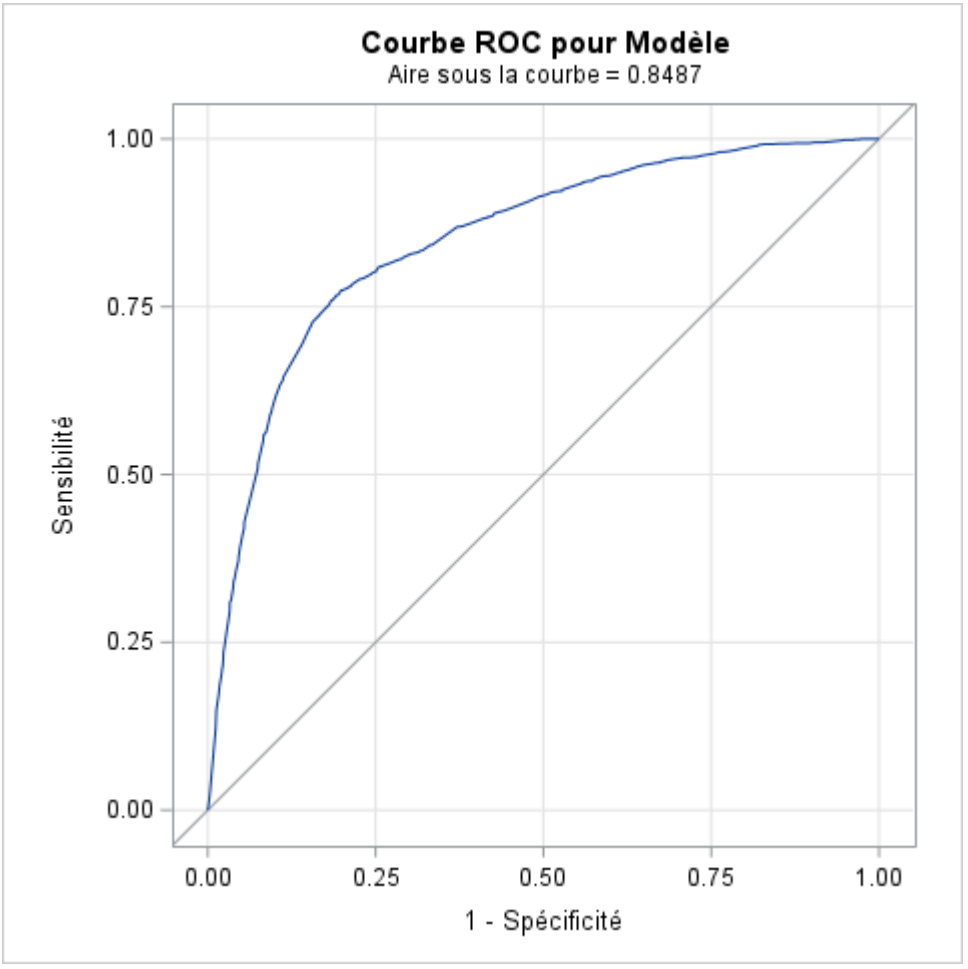
De manière globale, les erreurs de prévision sont de 17.29%. La qualité du modèle est relativement bonne.

le tableau ci-dessous est un récapitulatif des taux d'erreurs

Taux d'erreur moyen	17,34%
Taux d'erreur alpha	40,11%
Taux d'erreur Beta	9,59%

➢ *Courbe de ROC du modèle PROBIT*

La courbe de ROC est une mesure de la performance d'un classificateur binaire, c'est-à-dire d'un système qui a pour objectif de catégoriser des éléments en deux groupes distincts sur la base d'une ou plusieurs des caractéristiques de chacun de ces éléments. Cette courbe est très utilisée en médecine pour prévoir les personnes malades et les personnes non malades. Dans le cas de notre étude, nous nous intéresserons à la capacité de prédiction de notre modèle . La sensibilité c'est la capacité à prévoir la maladie alors que la spécificité, c'est la capacité à prévoir la non maladie. L'objectif est de maximiser à la fois sensibilité et spécificité. La méthode s'adapte donc parfaitement à notre étude qui cherche à prédire la capacité d'obtention d'un haut salaire. La courbe permet de mesurer le pouvoir discriminant de notre modèle à partir de l'indice de L'AUC.



➢ *Indice de l'AUC*

L'indice de l'Auc, c'est la surface sous la courbe. Elle donne une idée du pouvoir discriminant de notre modèle. Dans ce cas il est de 0,8487 ; c'est la probabilité estimée d'un individu ayant un salaire supérieur à 2000€ pris au hasard soit supérieur ou égale à la probabilité estimé d'un individu ayant un salaire inférieur à 2000€ pris au hasard. Notre indice de l'AUC est compris entre 0,8 et 0,9 donc le pouvoir de discrimination de ce modèle est excellent.

➢ *score du modèle probit*

Score modèle probit						
Paramètre	Estimation	Variables	Minimum	Maximum	Ecart	Score
Formation Apprentissage	0	Sortant de formation par voie	-0,12	0	0,12	7
Formation initial	-0,12	apprentissage (CFAD)				0
Homme	0	Sexe	-0,21	0	0,21	12
Femme	-0,21					0
Age entre 20 ans et 25 ans	0	Âge	-0,12	0,04	0,16	7
âge inférieur à 20 ans	-0,12					0
âge entre 26 ans et 30 ans	0,04					10
âge entre 31 ans et 35 ans	-0,05					4
BAC	0	Plus haut diplôme obtenu	-0,18	1,04	1,22	11
non diplômé	0,105	(PHD)				17
CAP_BEP_MC	-0,18					0
Bac plus 2/Bts/Dut	0,15					19
Bac Plus 2/3 santé	0,31					29
Bac plus 3/4	0,23					24

hors santé						
Bac plus 5/M2	0,76					55
Doctorat	1,04					72
La somme des écarts		1,7				

D'après les résultats du score normalisé -dessus nous pouvons dire que la variable qui influence le plus la probabilité d'obtention d'un salaire supérieur à 2000€ est le plus haut diplôme obtenu .En effet, la modalité doctorat de cette variable est le facteur le plus favorables pour l'obtention d'un haut salaire avec un score de 72 qui est le plus score le plus élévé du tableau .l'âge est le deuxième facteur en terme d'influence et le sexe est le troisième.En ce qui concerne la variable qui a le plus faible effet on trouve la variable apprentissage.

Les résultats de notre scoring nous permettent ainsi de définir deux profils :

- le profil le plus favorable à l'obtention d'un haut salaire :
  - ➔ Homme
  - ➔ Sortant d'une formation par voie d'apprentissage
  - ➔ Age entre 26 ans et 30 ans
  - ➔ plus haut diplôme obtenu est le Doctorat

- le profil le moins favorable à l'obtention d'un haut salaire :
  - ➔ Femme
  - ➔ Non Sortant d'une formation par voie d'apprentissage
  - ➔ Âge inférieur à 20 ans
  - ➔ plus haut diplôme obtenu est CAP/BEP/MC

IV. Modèle Logit de détermination d'un salaire supérieur à 2000€

La régression logistique ou modèle logit est un modèle de régression binomiale. Il est utilisé lorsque la variable dont on cherche à expliquer se présente sous une forme discrète avec un nombre de modalités restreint. Dans notre étude, nous cherchons à estimer la capacité d'obtention d'un salaire supérieur à 2000 € 3 ans après la fin des études. Le modèle logit est tout à fait adapté à cette problématique,son utilisation donne les résultats présents sur le tableau ci-dessous. En effet l'avantage de ce modèle par rapport au probit c'est la facilité d'interpréter les coefficients estimés. On va pouvoir interpréter les coefficients en termes de rapport de chance.

➢ *Résultat de l'ajustement*

variable à expliquer : Y (Tranche de salaire)

Paramètre	DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > khi-2
Intercept	1	-2.3493	0.2460	91.2334	<.0001
âge inférieur à 20 ans	1	-0.4694	0.1928	5.9239	0.0149
âge entre 26 ans et 30 ans	1	0.1381	0.1309	1.1125	0.2915
âge entre 31 ans et 35 ans	1	0.1003	0.1098	0.8340	0.3611
Femme	1	-0.7139	0.0821	75.7033	<.0001
Apprentissage	1	-0.4713	0.1248	14.2614	0.0002
Père Ouvrier	1	0.2334	0.0765	9.3023	0.0023
Père Employé	1	0.1584	0.0747	4.4955	0.0340
Père Cadre ingénieur	1	0.0285	0.0667	0.1832	0.6687
Père Artisan chef d'entreprise	1	0.0732	0.0722	1.0257	0.3112
non diplômé	1	-0.4542	0.5331	0.7259	0.3942
Cap Bep	1	-0.8037	0.3156	6.4846	0.0109
Bac plus 2/Bts/Dut	1	0.5957	0.1986	8.9936	0.0027
Bac Plus 2/3 santé	1	1.1683	0.1967	35.2798	<.0001
Bac plus 3/4 hors santé	1	0.8979	0.1924	21.7899	<.0001

Bac plus 5/M2	1	2.6725	0.1732	237.9884	<.0001
Doctorat	1	3.6171	0.2202	269.8524	<.0001

➢ *interprétation des coefficients*

Afin d'interpréter les résultats du modèle Logit nous devons passer par l' exponentielle

- ➔ Le coefficient de la constante de notre modèle est de -2.3493. En termes de probabilité cela donne une probabilité de 8.67% pour notre individu de référence. Dans notre cas, un homme âgé entre 20-25 ans qui a le BAC de père de catégorie "technicien/agent de maîtrise/vendeur" a une probabilité de 8.67% d'avoir un salaire supérieur à 2000€.
- ➔ Un doctorant a des chances multipliés par 37 fois d'avoir un salaire supérieur à 2000€ par rapport un individu qui a le bac, toute chose égale par ailleurs
- ➔ une femme a des chances divisés par 2 environ d'avoir un salaire supérieur à 2000€ par rapport à un homme,toutes choses égales par ailleurs.

➢ *Test de significativité globale : le test du maximum de vraisemblance*

Afin de tester la significativité globale du modèle nous allons établir le test du maximum de vraisemblance sur la base du critère de l'efficacité statistique puisqu'il est plus contraignant que le test de wald

on pose deux hypothèses:

- Ho : les 5 restrictions posées sont vraies
- H1 : les 5 restrictions posées sont fausses

la statistique du test :$X^2 = 2(LLC - LLR)$ avec LLC modèle complet et LLR modèle restreint

- H0 : $X^2 <= X^2alpha$
- H1 : $X^2 > X^2alpha$

On pose le seuil de significativité égale à α=5% On pose la règle de décision basée sur la p-value est donc de refuser H0 .si le niveau de la p-value est inférieure à un seuil α donnée.

D'après le tableau ci-dessus nous pouvons remarquer qu'on a une p-value*<.0001* donc nous allons rejeter H0

Test de l'hypothèse nulle globale			
Test	*khi-2*	*DDL*	*Pr > khi-2*
Rapport de vraisemblance	1825.4381	16	<.0001

Conclusion : notre modèle est globalement significatif.

➢ *Test de significativité partielle : Test de Wald*

Afin de tester la significativité partielle du modèle nous allons établir un test de Wald, nous allons alors tester nos variables une à une pour voir s'ils ont une action significative sur notre variable d'intérêt c'est-à-dire la probabilité d'avoir un salaire supérieur à 2000€.

Libellé	Khi-2 de Wald	DDL	Pr > khi-2
Formation apprentissage	14.2614	1	0.0002
Le sexe	75.7033	1	<.0001
Position professionnel du père	16.4467	4	0.0025
l'âge	9.4287	3	0.0241
Plus haut diplôme obtenu	575.1120	7	<.0001

On pose deux hypothèses :

- H0 : AJ=0 (variable n'agit pas sur Y)
- H1: AJ #0 (variable agit sur Y)

On pose le seuil de significativité égale à α=10% On pose la règle de décision basée sur la p-value est donc de refuser H0 .si le niveau de la p-value est inférieure à un seuil α donnée.

➔ on accepte HO si khi2 >khi2 de wald
➔ et on accepte H1 si khi2 alfa < khi2 wald

D'après les résultats du tableau ci-dessus nous remarquons qu'on a une p-value <10% pour toutes nos variables donc nous pouvons dire que toutes nos variables ont une action significative sur notre variable d'intérêt.

➢ *Mesure de la qualité de l'ajustement par le R^2*

Pour mesurer la qualité de l'ajustement nous allons regarder le pseudo R carré de Mcfadden puisque ce dernier est liée directement à la statistique du test (test de maximum du vraisemblance) comme le R carré du linéaire avec le test de fisher.En effet le Pseudo R carré de mcfadden présente les mêmes propriétés que le coefficient de détermination classique, c'est-à-dire compris entre 0 et 1 et croissant avec la qualité de l'ajustement. Il permet aussi d'effectuer une comparaison avec le coefficient de détermination du modèle linéaire.

R^2 de McFadden	0.2941

Le pseudo-R2 de McFadden sera utilisé pour ce modèle .Il correspond à la part expliquée du phénomène. Autrement dit, les différences de salaire sont expliquées à 29.41% par les variables explicatives utilisées dans notre modèle.

➢ *La matrice de confusion du modèle*

La matrice de confusion est associée à un seuil de 50% permettant de confronter les valeurs prédites et les valeurs observées. Il permet de voir la qualité de la prévision. Cette matrice va nous permettre de calculer des taux d'erreur. Elle permet de juger de la qualité des prévisions du modèle.

Valeur réelle	Valeur prédite du salaire qualitatif		
	0	1	Total
0	3693	394	4087
1	557	834	1391
Total	4250	1228	5478

- 3693*: c'est le nombre d' étudiants que nous avons estimés qu'ils ont un salaire inférieur à 2000€ et qu'ils ont effectivement en réalité un salaire inférieur à 2000€.*
- 557*: c'est le nombre d'étudiants que notre modèle prévoit qu'ils ont un salaire inférieur à 2000€ alors qu'en réalité ils ont un salaire supérieur à 2000€.*
- 394*: c'est le nombre d'étudiants que notre modèle prévoit qu'ils ont un salaire supérieur à 2000€ alors qu'en réalité ils ont un salaire inférieur à 2000€.*
- 834*: c'est le nombre d'étudiants que notre modèle prévoit qu'ils ont un salaire supérieur à 2000€ et qu'ils ont effectivement un salaire supérieur à 2000€.*

Calcul des taux d'erreur:

taux d'erreur alpha: C'est l'erreur de prévoir l'événement (0) alors que l'individu doit connaître l'événement (1).En effet Dans le cas de notre étude c'est l'erreur de prévoir qu'un étudiant a un salaire inférieur à 2000€ alors qu'en réalité ce dernier a un salaire supérieur à 2000€ .

α=(557/1391)*100=40,04%

Le taux d'erreur α est élevé. Cela veut dire que le modèle a tendance à prévoir l'événement 0 (salaire<2000 euros) pour des individus qui gagnent plus de 2000 €

taux d'erreur beta: C'est l'erreur de prévoir l'événement (1) alors que l'individu doit connaître l'événement (0).En effet Dans le cas de notre étude c'est l'erreur de prévoir qu'un étudiant a un salaire supérieur à 2000€ alors qu'en réalité ce dernier a un salaire inférieur à 2000€ .

β=(394/4087)*100=9,64%

Le taux d'erreur β est faible donc notre modèle a tendance à bien estimer les gens qui gagnent plus de 2000 €.

taux d'erreur moyen: c'est la proportion de mal classés, la probabilité de mal classer un individu pris au hasard dans notre échantillon.En effet, ce taux est une synthèse entre les deux taux d'erreur. Il donne l'erreur moyen de prédiction du modèle.

taux moyen= ((557+394)/5478)*100=17,36%

De manière globale, les erreurs de prévision sont de 17.29%. La qualité du modèle est relativement bonne.

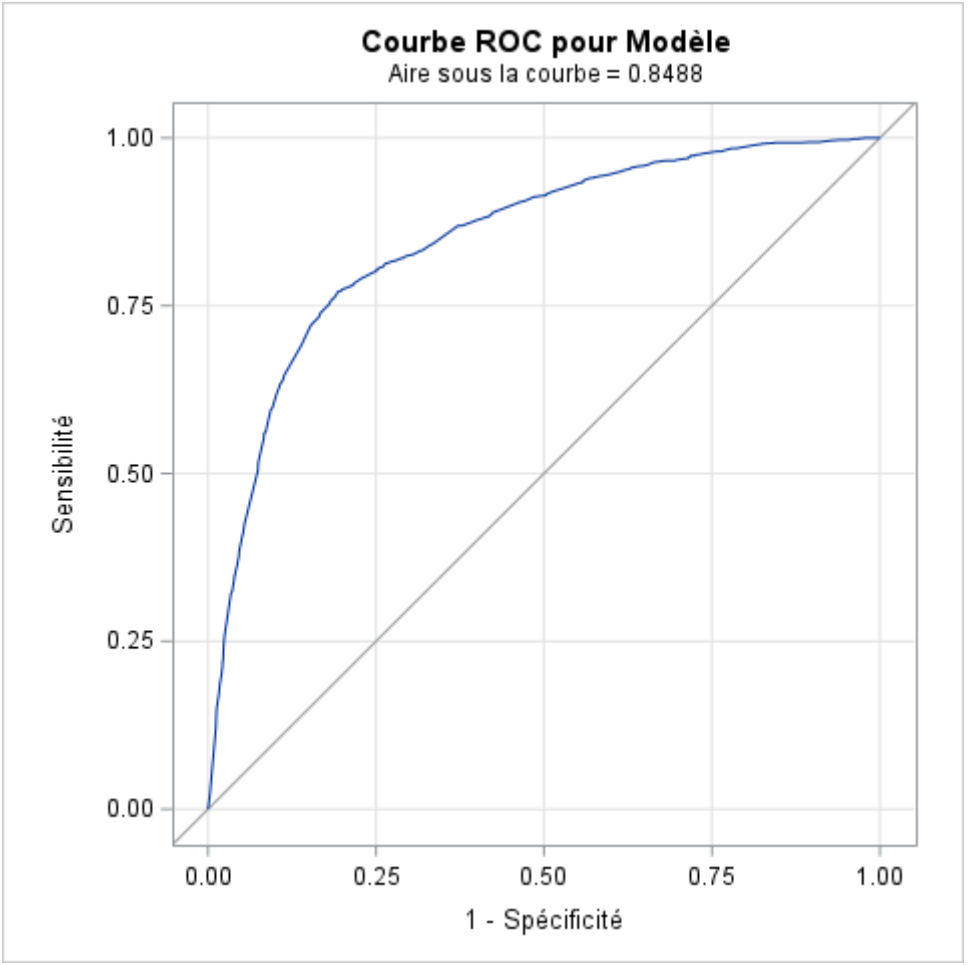
le tableau ci-dessous est un récapitulatif des taux d'erreurs

Taux d'erreur moyen	17,36%
Taux d'erreur alpha	40,04%
Taux d'erreur Beta	9,64%

➢ *Courbe de ROC du modèle LOGIT*

C'est une représentation graphique qui met en ordonné la sensibilité et en abscisse la spécificité.c'est une mesure de la performance d'un classificateur binaire, c'est-à-dire d'un système qui a pour objectif de catégoriser des éléments en deux groupes distincts sur la base d'une ou plusieurs des caractéristiques de chacun de ces éléments. Cette courbe est très utilisée en médecine pour prévoir les personnes malades et les personnes non malades. Dans le cas de notre étude, nous nous intéresserons à la capacité de prédiction de notre modèle . La sensibilité c'est la capacité à prévoir la maladie alors que la spécificité, c'est la capacité à prévoir la non maladie. L'objectif est de maximiser à la fois sensibilité et spécificité. La méthode s'adapte donc parfaitement à notre étude qui cherche à prédire la capacité

d'obtention d'un haut salaire. La courbe permet de mesurer le pouvoir discriminant de notre modèle à partir de l'indice de L'AUC.



➢ *Indice de l'AUC*

L'indice de l'Auc, c'est la surface sous la courbe. Elle donne une idée du pouvoir discriminant de notre modèle. Dans ce cas il est de 0,8488 ; c'est la probabilité estimée d'un individu ayant un salaire supérieur à 2000€ pris au hasard soit supérieur ou égale à la probabilité estimé d'un individu ayant un salaire inférieur à 2000€ pris au hasard. Notre indice de l'AUC est compris entre 0,8 et 0,9 donc le pouvoir de discrimination de ce modèle est excellent.

➢ *score du modèle logit*

Le score normalisé va nous permettre de déterminer le profil le plus favorable ou défavorable pour l'obtention d'un salaire supérieur à 2000€.Les variables utilisées dans l'établissement du score sont les variables significatives pour tous les modèles.

Score modèle Logit						
Paramètre	Estimation	Variables	Minimum	Maximum	Ecart	Score
Formation Apprentissage	0	Sortant de formation par voie	-0,47	0	0,47	8
Formation Initiale	-0,47	apprentissage (CFAD)				0
Homme	0	Sexe	-0,71	0	0,71	11
Femme	-0,71					0
Age entre 20 ans et 25 ans	0	Âge	-0,47	0,14	0,61	8
âge inférieur à 20 ans	-0,47					0
âge entre 26 ans et 30 ans	0,14					10
âge entre 31 ans et 35 ans	-0,1					6
BAC	0	Plus haut diplôme obtenu	-0,8	3,62	4,42	13
non diplômé	-0,45	(PHD)				6
CAP_BEP_MC	-0,8					0
Bac plus 2/Bts/Dut	0,6					22
Bac Plus 2/3	1,17					32

santé						
Bac plus 3/4 hors santé	0,9					27
Bac plus 5/M2	2,67					56
Doctorat	3,62					71
La somme des écarts		6,21				

D'après les résultats du score normalisé -dessus nous pouvons dire que la variable qui influence le plus la probabilité d'obtention d'un salaire supérieur à 2000€ est le plus haut diplôme obtenu .En effet, les modalités doctorat et Bac+5/ M2 de cette dernière sont les facteurs les plus favorables pour l'obtention d'un haut salaire puisque ces dernières ont les plus grands score.l'âge est le deuxième facteur en terme d'influence et le sexe est le troisième.En ce qui concerne la variable qui a le plus faible effet on trouve la variable apprentissage.

Les résultats de notre scoring nous permettent ainsi de définir deux profils :

- le profil le plus favorable à l'obtention d'un haut salaire :
  - ➔ Homme
  - ➔ Sortant d'une formation par voie d'apprentissage
  - ➔ Age entre 26 ans et 30 ans
  - ➔ plus haut diplôme obtenu est le Doctorat

- le profil le moins favorable à l'obtention d'un haut salaire :
  - ➔ Femme
  - ➔ Non Sortant d'une formation par voie d'apprentissage
  - ➔ Âge inférieur à 20 ans
  - ➔ plus haut diplôme obtenu est CAP/BEP/MC

Analyse synthétique et comparative des différents modèles prédictifs du salaire 3 ans après la sortie du système scolaire

Lors de cette étape nous allons élaborer une analyse synthétique et comparative de nos 4 modèles différents en se basant sur les résultats obtenus et les analyses tout au long de cette étude.

1) Comparaison entre les signes des coefficients estimés :

Variable	Valeur estimée des paramètres				
	Modèle linéaire variable quanti	Modèle linéaire variable Dummy	modèle probit	modèle logit	Cohérent/ Différent
Intercept	1667.34396	0.17543	-0.7255	-2.3493	Différent
âge inférieur à 20 ans	-69.80577	-0.03368	-0.1190	-0.4694	Cohérent
âge entre 26 ans et 30 ans	19.83514	0.02522	0.0435	0.1381	Cohérent
âge entre 31 ans et 35 ans	43.09268	-0.03294	-0.0518	0.1003	Diffèrent
Femme	-116.08561	-0.09345	-0.2052	-0.7139	Cohérent
Non Apprentissage	-72.48092	-0.04011	-0.1157	-0.4713	Cohérent
Père Ouvrier	-67.22039	-0.05735	0.2631	0.2334	Différent
Père Employé	-46.76598	-0.04082	0.1795	0.1584	Différent
Père Cadre ingénieur	11.01641	-0.00772	0.0298	0.0285	Différent
Père Artisan chef d'entreprise	-11.20138	-0.02364	0.0902	0.0732	Différent
non diplômé	-35.15227	-0.02157	-0.1048	-0.4542	Cohérent
Cap Bep	-39.79603	-0.03244	-0.1804	-0.8037	Cohérent
Bac plus 2/Bts/Dut	69.83722	0.04324	0.1495	0.5957	Cohérent

Bac Plus 2/3 santé	256.74650	0.10854	0.3063	1.1683	Cohérent
Bac plus 3/4 hors santé	140.58637	0.07867	0.2310	0.8979	cohérent
Bac plus 5/M2	568.38535	0.42620	0.7555	2.6725	Cohérent
Doctorat	850.22953	0.63892	1.0380	3.6171	Cohérent

Le tableau ci-dessus montre une comparaison au niveau des signes des coefficients d'estimation de chaque modèle puisque ces différents coefficients ne sont pas comparable directement.Généralement nous pouvons dire qu'on a une cohérence à l'exception des modalités de la variable position professionnelle du père, nous pouvons voir que même si nous avons pas une cohérence pour les 4 modèles la cohérence est présente entre les modèles logit probit puisque nous avons les mêmes signes et la même chose pour les deux modèles linéaires D'autres part pour le signe de la constante nous remarquons une différence au niveau des deux modèles linéaires avec les modèles Probit et Logit. De manière globale, nous pouvons dire qu'il existe une forte cohérence au niveau des signes de nos différents modèles.

2) Comparaison coefficients de détermination :

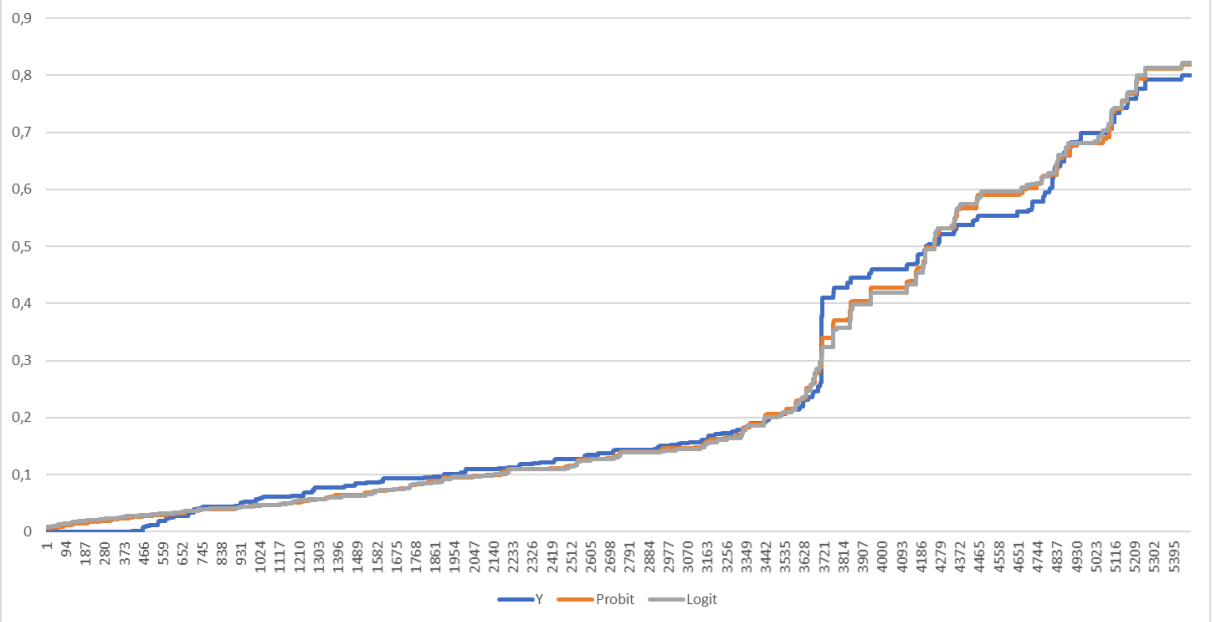
Afin de comparer les 4 modèles nous avons choisi de regarder le pseudo R carré de Mcfadden pour les modèles Logit et Probit puisque ce pseudo R carré a les mêmes propriétés que celui du linéaire

Modèle	Coefficient de détermination	
Modèle linéaire variable quanti	R carré du linéaire	0.3767
Modèle linéaire variable Dummy	R carré du linéaire	0.3281
Modèle Logit	Pseudo R^2 de McFadden	0.2941
Modèle Probit	Pseudo R^2 de McFadden	0.2937

D'après le tableau ci-dessus nous remarquons d'une part que les coefficients de détermination des modèles Logit et probit sont semblables. D'autre part, le coefficient de détermination le plus plus élevé est celui du modèle linéaire avec le salaire quantitatif, nous estimons que cela revient au type de la variable à expliquer .

Globalement, les coefficients de détermination sont assez faibles. Cependant, cela peut être expliqué par le manque de variables explicatives qui peuvent expliquer les différences du salaire.

3) Comparaison des probabilités des trois modèles :Y dummy/Probit/Logit



Nous remarquons du graphique ci-dessus que pour le modèle linéaire les prédictions ne sont pas bornées entre 0,1 et pas symétrique cela nous mène à introduire la fonction de translation (modèle logit et Probit

En général le modèle linéaire est moyennement similaire aux prédictions des trois modèles, Les deux modèles logistique probit et logit sont proches en termes des prédictions.

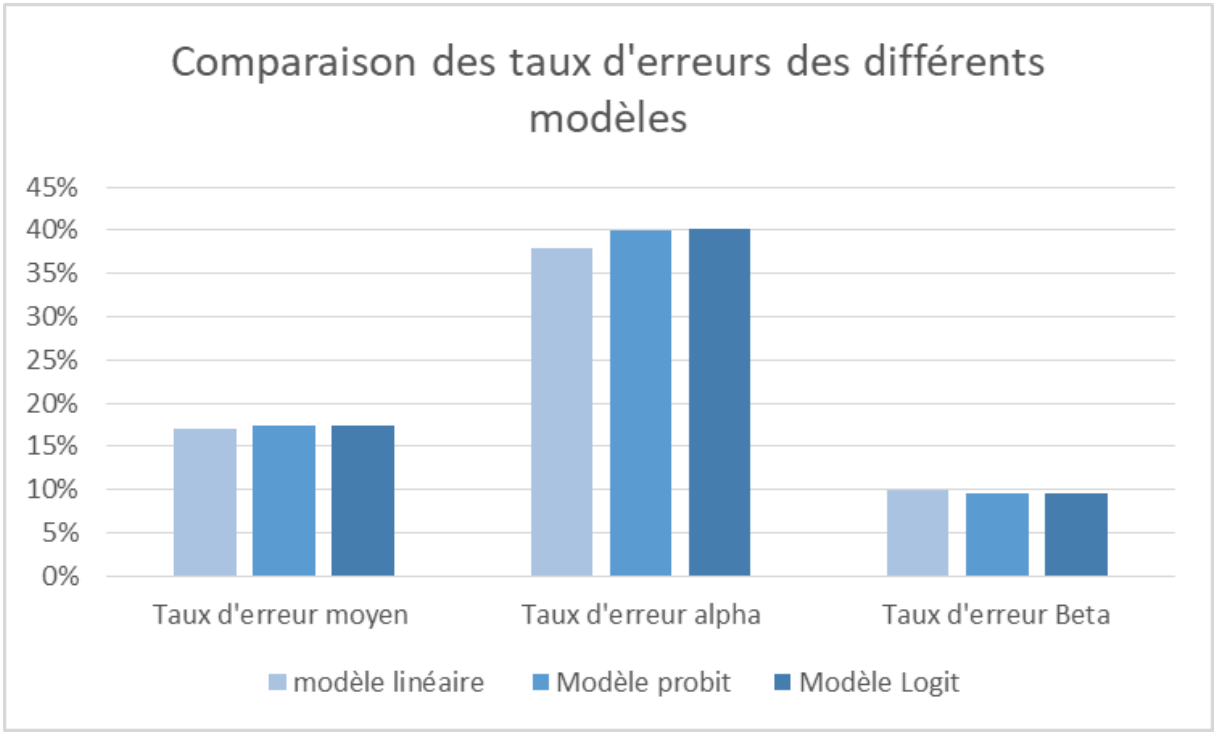
coefficient de corrélation		
Y et Probit	Y et Logit	Probit et Logit
0,99696345	0,9959034	0,999836323

Les résultats du tableau qui illustre les coefficients de corrélations entre les 3 modèles confirment les résultats obtenus dans le graphique précédent. Les modèles sont généralement très proches.

4) Comparaison entre les taux d'erreurs :

A partir du graphique ci-dessous nous remarquons que les taux d'erreurs moyen et beta sont très proches entre les trois modèles. En ce qui concerne le taux d'erreur alpha nous remarquons que les modèles logit et probit sont similaires , nous notons aussi une légère différence avec le modèle linéaire.

- ➔ le taux d'erreur alpha le plus faible est celui du modèle linéaire
- ➔ le taux d'erreur beta le plus élevé est celui du modèle linéaire

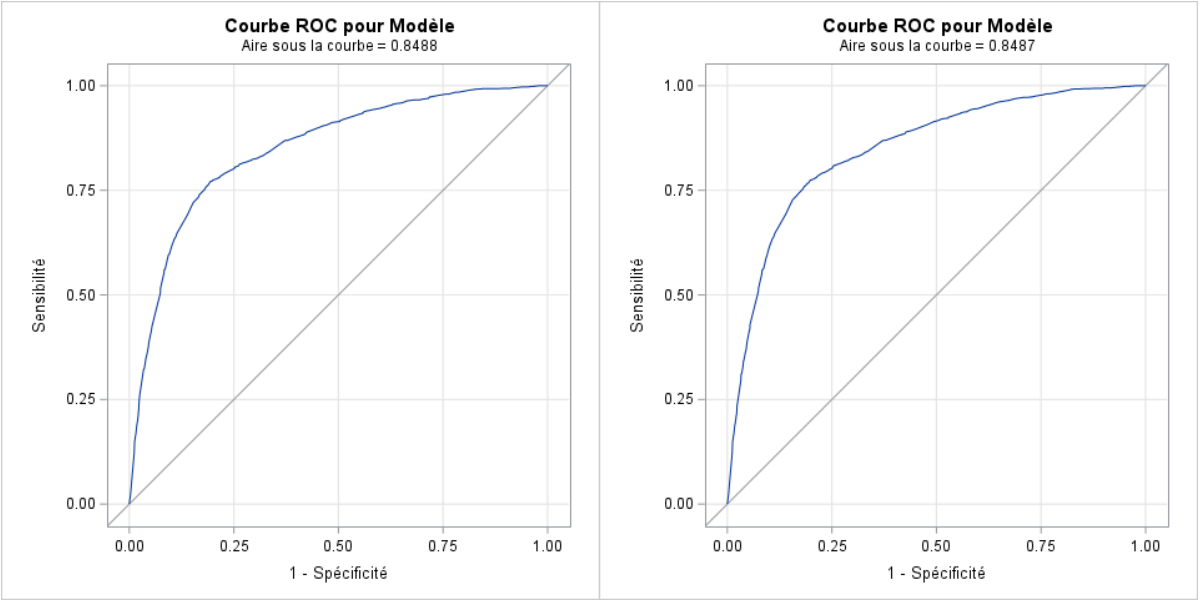


5) Comparaison des différents scores par modèle :

comparaison des scores				
Paramètre	Modèle linéaire variable quanti	Modèle linéaire variable Dummy	modèle probit	modèle logit
Formation apprentissage	6	5	7	8
Formation initial	0	0	0	0
Homme	10	11	12	11
Femme	0	0	0	0
Age entre 20 ans et 25 ans	6	4	7	8
âge inférieur à 20 ans	0	0	0	0
âge entre 26 ans et 30 ans	8	7	10	10
âge entre 31 ans et 35 ans	9	0	4	6
BAC	3	4	11	13
non diplômé	0	1	17	6
CAP_BEP_MC	0	0	0	0
Bac plus 2/Bts/Dut	9	9	19	22
Bac Plus 2/3 santé	25	16	29	32
Bac plus 3/4 hors santé	15	13	24	27
Bac plus 5/M2	51	53	55	56
Doctorat	75	78	72	71

➔ Nous remarquons qu'on a les mêmes profils les plus favorables à l'obtention d'un haut salaire et les mêmes profils les plus défavorables à l'obtention d'un haut salaire pour les 4 modèles.

6) comparaison des courbes des ROC et des AUC :



- Les courbes de ROC pour le modèle Probit et le modèle Logit sont relativement identiques .

- Les indices de l'AUC des deux modèles sont proches. Ils sont compris entre 0.8 et 0.9.
  Le pouvoir discriminant des deux modèles est excellent.

conclusion

En guise de conclusion, la mobilisation des conceptions économétriques nous a mené à une analyse plus concrète sur la manière dont se comporte le salaire en fonction des différents profils observés.

Pour les différents modèles, le profil qui est plus susceptible à avoir un salaire supérieur à 2000 euro est le profil d'un homme âgé entre 26 et 30 ans, qui était en formation d'apprentissage et qui est doctorant et que son père était cadre-ingénieure lors de la sortie de sa formations, nous soulignons que ce constat est valable pour les 4 modèles.

L'analyse réalisé nous permets de dire que les modèles sont en globalité significatifs mais restituent qu'une partie de l'information, ce qui nous permet de remettre en question la qualité et la quantité des variables utilisé; bien évidemment le salaire ne dépend pas que des variables évoqué dans le rapport mais le champ d'étude du salaire est plus vaste et dépend de la conjoncture économique, la situation du pays et les choix de carrière au delà de catégorie socioprofessionnelle.