

Wave 5 of Health and Retirement Study

Grégoire AMATO

5 avril 2020

Sommaire

1	Introduction	1
2	Statistiques descriptives	2
2.1	Variables qualitatives	2
2.2	Variables quantitatives	2
3	Modèle général et fonction de log-vraisemblance.	3
3.1	Modèle général	3
3.2	Log-vraisemblance	3
4	Estimation	4
4.1	Logit	4
4.1.1	Détermination des variables influentes	4
4.1.2	Premier modèle candidat et diagnostic	8
4.1.3	Variables en interaction	14
4.1.4	Comparaison des modèles	17
4.1.5	Modèle final	20
4.2	Probit	25
4.2.1	Premier modèle et diagnostic	28
4.2.2	Variables en interaction	31
4.2.3	Comparaison des modèles	34
4.2.4	Modèle final	36
5	Conclusion	39
6	Annexe	40
6.1	Répartition des variables qualitatives	40
6.2	Répartition des variables quantitatives	43
6.3	Matrices de corrélation et p-values détaillées	49
6.4	Test du rapport des vraisemblances	51
6.5	Test de Wald	51

1 Introduction

Médicare est le système d'assurance-santé géré par le gouvernement des Etats-Unis. Pour y avoir accès il faut impérativement :

- Avoir plus de 65 ans.
- Avoir cotisé au moins 10 ans.
- Habiter de manière permanente aux Etats-Unis.

Néanmoins, il existe des exceptions pour les personnes de moins de 65 ans si :

- Elles sont handicapées.
- Elles sont au stade final d'une maladie rénale.

Dans ce cas, pour être éligibles, elles doivent recevoir des aides de la sécurité sociale ou du financement de la retraite des cheminots. Le but de cette étude est d'analyser les achats d'assurances.

Ce jeu de données contient 3206 observations, chacune représentant un bénéficiaire de medicare.

Le jeu de données contient 18 variables :

Variable à expliquer :

- **ins, private** : achat d'assurances privées de toutes sources, incluant les marchés privées, les assurances allouées par l'employeur, Medica et autres.

Variables socioéconomiques :

- **age** : variable quantitative discrète (continue théoriquement) , l'âge de l'individu.
- **female** : variable qualitative binaire, le sexe de l'individu, prend pour valeur 1 si l'individu est une femme, 0 sinon.
- **white** : variable qualitative binaire.
- **hisp** : variable qualitative binaire.
- **married** : variable qualitative binaire, prend pour valeur 1 si l'individu est marié, 0 sinon.
- **educyear** : variable quantitative discrète, nombre d'années d'études.
- **retire** : variable qualitative binaire, prend pour valeur 1 si l'individu a pris sa retraite, 0 sinon.

Variables relatives à l'état de santé :

- **hstatusg** : variable qualitative binaire indiquant si l'état de santé est excellent, très bon ou bon.
- **variables dummies** : 5 variables qualitatives binaires relatives à l'état de santé, excellent, très bon, bon, juste ou faible.
- **chronic** : variables quantitatives, nombre de maladies chroniques.
- **adl** : variable qualitative ordinale, activités de la vie quotidienne, rend compte des difficultés qu'ont les personnes âgées à effectuer des tâches de la vie quotidienne.

Variables relative à l'époux/l'épouse :

- **sretire** : variable qualitative, présence d'un(e) époux/épouse à la retraite.
- **hhincome** : variable quantitative, revenu du ménage.

2 Statistiques descriptives

Tous les graphiques associés aux statistiques descriptives sont disponibles en Annexe, il s'agit ici de faire un court résumé de l'étude des variables.

2.1 Variables qualitatives

Les graphiques sont disponibles ici.

Variable	Moyenne	Ecart.type	Min	Max
hisp	7.27	-	0	1
white	82.06	-	0	1
female	47.79	-	0	1
married	73.30	-	0	1
retire	62.48	-	0	1
sretire	38.83	-	0	1
ins	38.71	-	0	1
hstatusg	70.46	-	0	1

TABLE 1

- La majorité des personnes sont en bonne santé (environ 70%)
- 30% des personnes sont encore en activité.

2.2 Variables quantitatives

Les graphiques sont disponibles ici.

Variable	Moyenne	Ecart.type	Min	Max
age	66.91	3.68	52.00	86.00
educyear	11.90	3.30	0.00	17.00
chronic	2.06	1.42	0.00	8.00
adl	0.30	0.83	0.00	5.00
hhincome	45.26	64.34	0.00	1312.12

TABLE 2

- Il y a un pic d'arrêt d'études après le Baccalauréat¹ (environ 50% des bacheliers n'ont pas continué dans le supérieur), on remarque également une hausse de souscription d'assurances avec la durée de scolarisation.
- La répartition de la variable *hhincome* est très inégale, l'écart-type est très élevé, pour palier à ce problème, il vaut mieux prendre son logarithme. La variable est donc remplacée par *lnincome*.
- 86% des personnes possèdent tous leurs droits de retraite²

1. Aux Etats-Unis, l'université étant très chère, beaucoup d'étudiants préfèrent arrêter leurs études pour aller travailler.

2. Calculé d'après le site de la Social Security.

3 Modèle général et fonction de log-vraisemblance.

3.1 Modèle général

Notre étude porte sur un modèle dichotomique univarié, cherchant à expliquer la souscriptions d'assurances privées. Exprimons notre modèle en terme de variable latente :

$$y_i = \begin{cases} 1 & \text{si } y_i^* > \gamma \\ 0 & \text{sinon} \end{cases}$$

avec

$$y_i^* = x_i\beta + \varepsilon_i$$

y_i^* étant la variable latente du modèle, c'est-à-dire la variable théorique, estimée par la variable observée y_i .

Nous effectuons une normalisation du seuil γ car dans notre cas, nous n'avons aucune information à-priori sur la valeur du seuil.

3.2 Log-vraisemblance

La log-vraisemblance d'un modèle général logistique est :

$$\text{loglik}_{\text{logit}}(\hat{\beta}) = \sum_{i=1}^n (y_i - \Lambda(x_i\hat{\beta}))x_i'$$

avec $\Lambda()$ la fonction de densité de la loi logistique.

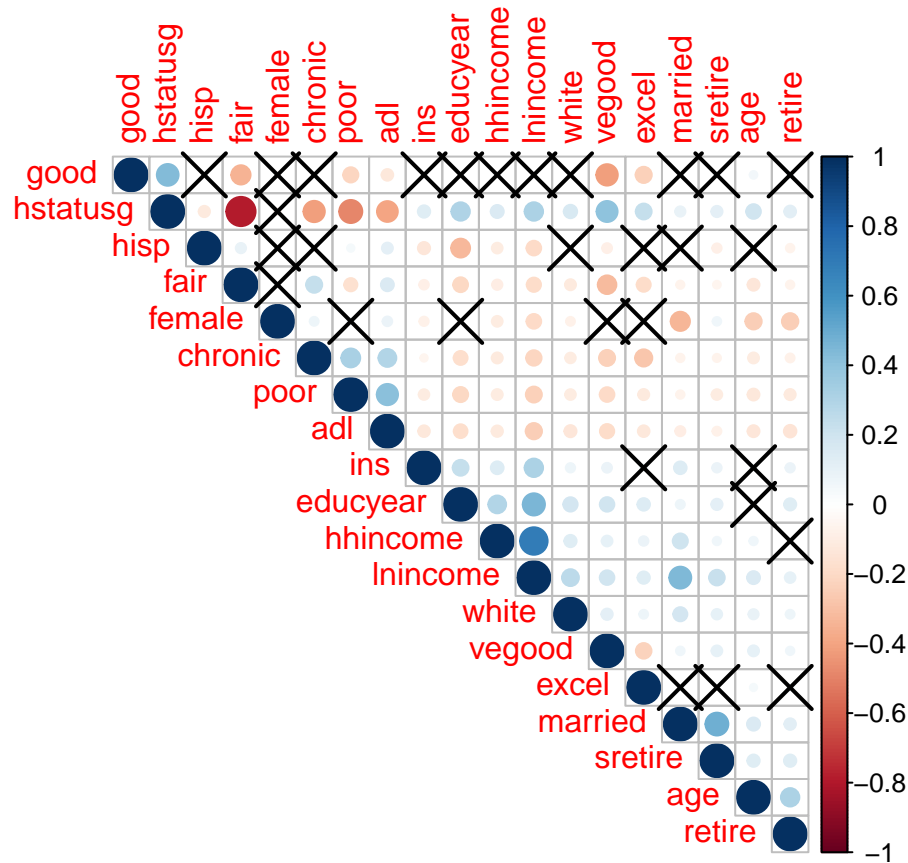
La log-vraisemblance d'un modèle probit est :

$$\text{loglik}_{\text{probit}}(\hat{\beta}) = \sum_{i=1}^n \frac{[y_i - \Phi(x_i\hat{\beta})]\phi(x_i\hat{\beta})}{\Phi(x_i\hat{\beta})[1 - \phi(x_i\hat{\beta})]} x_i'$$

avec $\Phi()$ la fonction de densité de la loi normale.

4 Estimation

Avant de commencer les estimations, regardons la matrice de corrélation des variables :



Les corrélations non significativement différentes de 0 sont barrées, la matrice des p-values est disponible en annexe. Les corrélations positives les plus importantes se situent entre la variable *hstatusg* et les variables dummies associées (*good*, *fair*...), il faudra faire attention à ne pas les mettre dans le même modèle si elles sont significatives. Par ailleurs, *adl* est également assez fortement corrélé négativement à *hstatusg*, ce qui est plutôt normal car *adl* dépend directement de l'état de santé. On observe une dernière corrélation négative entre les variables *married* et *sretire*.

4.1 Logit

4.1.1 Détermination des variables influentes

Débutons notre étude par l'estimation d'un modèle logit. Dans un premier temps, nous mettons en place une validation croisée, à raison d'un échantillon d'entraînement comprenant deux-tiers des données, et un échantillon test comprenant un tiers des données. Puis, il faut trier les variables explicatives ayant le plus d'influence, nous utiliserons un algorithme stepwise bidirectionnel, cherchant à minimiser l'AIC³ afin de créer un modèle logit optimal. Le modèle préconisé par l'algorithme est présenté dans la table 3.

3. Le critère d'information d'Akaike est une mesure de la qualité d'un modèle statistique, permettant de pénaliser les modèles en fonction du nombre de paramètres, on choisit alors le modèle avec l'AIC le plus faible.

TABLE 3: Modèle stepwise

<i>Dependent variable :</i>	
ins	
lnincome	0.687*** (0.067)
educyear	0.060*** (0.018)
retire1	0.216** (0.099)
hisp1	-0.491** (0.237)
adl	-0.139** (0.071)
white1	-0.208 (0.129)
Constant	-3.489*** (0.276)
Observations	2,149
Log Likelihood	-1,306.826
Akaike Inf. Crit.	2,627.652
<i>Note :</i> *p<0.1 ; **p<0.05 ; ***p<0.01	

Nous pouvons observer immédiatement que la variable *white* n'est pas significativement différente de zéro. Confirmons cela par un test de Wald} sur chaque variable⁴ afin d'étudier de plus près la significativité des variables et ainsi écarter de notre modèle les variables non significatives.

	Df	Chisq	Pr(>Chisq)
hisp	1	4.29	0.0383
educyear	1	11.51	0.0007
retire	1	4.74	0.0294
lnincome	1	105.55	0.0000
adl	1	3.89	0.0485
white	1	2.63	0.1052

TABLE 4: Test de Wald

4. Pour tous les tests de l'étude, nous appliquerons un seuil de significativité critique équivalent à $\alpha = 0.05$

La p-value de *white* est de 0.1052, qui est supérieure à la valeur critique de 0.05, donc nous conservons H_0 , le coefficient associé à la variable *white* n'est pas significativement différent de zéro. Mais nous pouvons également remarquer que la p-value de la variable *adl* est très proche du seuil critique, il va falloir surveiller la réaction de cette variable lors de la modification du modèle.

Pour tester notre modèle, nous effectuerons maintenant un test du rapport des vraisemblances afin de confronter notre modèle simplifié (sans *white*) au modèle complexe (comprenant *white*). Les résultats sont compilés dans la table 5.

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
1	5	-1310.01			
2	6	-1308.89	1	2.25	0.1332

TABLE 5: Rapport des vraisemblances

La première ligne représente le modèle simplifié, la seconde le modèle complexe. Les degrés de liberté (#Df) représentent le nombre de variable explicative du modèle, LogLik la log-vraisemblance du modèle et Pr(>Chisq) la p-value. Le test du rapport des vraisemblances conserve le modèle simplifié, en effet, la p-value est supérieure à 0.05, nous allons donc supprimer la variable *white*. La table 6 contient la nouvelle estimation du modèle :

TABLE 6

<i>Dependent variable :</i>	
ins	
hisp1	−0.507** (0.237)
educyear	0.057*** (0.018)
retire1	0.210** (0.099)
lnincome	0.670*** (0.066)
adl	−0.133* (0.071)
Constant	−3.568*** (0.272)
Observations	2,149
Log Likelihood	−1,308.131
Akaike Inf. Crit.	2,628.263
<i>Note :</i> *p<0.1 ; **p<0.05 ; ***p<0.01	

Comme nous le craignons, le fait de retirer la variable *white* a affaibli la significativité de la variable *adl*, il vaudrait mieux effectuer un test de Wald, ainsi qu'un test du rapport des vraisemblances pour décider de la conserver ou non, ces résultats sont disponibles dans la table 7 :

	Df	Chisq	Pr(>Chisq)
hisp	1	4.58	0.0323
educyear	1	10.60	0.0011
retire	1	4.52	0.0335
lnincome	1	103.32	0.0000
adl	1	3.56	0.0591

TABLE 7: Test de Wald

On remarque en effet que la variable n'est pas significative à un seuil de 5%.

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
1	5	-1310.01			
2	6	-1308.13	1	3.76	0.0524

TABLE 8: Rapport des vraisemblances

Le test du rapport de vraisemblance (table 8) conserve le modèle simplifié, on retire donc *adl* du modèle. Effectuons tout de même un test de Wald sur le modèle simplifié pour vérifier que cela n'a pas entraîné de changement sur les autres variables :

	Df	Chisq	Pr(>Chisq)
hisp	1	4.82	0.0281
educyear	1	11.35	0.0008
retire	1	5.36	0.0206
lnincome	1	111.32	0.0000

TABLE 9: Test de Wald

4.1.2 Premier modèle candidat et diagnostic

Voici donc le modèle que nous conservons pour la suite :

TABLE 10

<i>Dependent variable :</i>	
	ins
hisp1	−0.519** (0.236)
educyear	0.059*** (0.017)
retire1	0.228** (0.098)
lnincome	0.688*** (0.065)
Constant	−3.696*** (0.265)
Observations	2,149
Log Likelihood	−1,310.013
Akaike Inf. Crit.	2,630.025
<i>Note :</i> *p<0.1 ; **p<0.05 ; ***p<0.01	

La déviance⁵ du modèle est de 2620.0250171, avec un degré de liberté associé de 2144.

5. La déviance permet de comparer la vraisemblance du modèle à celle d'un modèle parfait en terme d'adequation aux données : le modèle saturé.

Voici les valeurs du pseudo- R^2 par trois modes de calculs différents⁶.

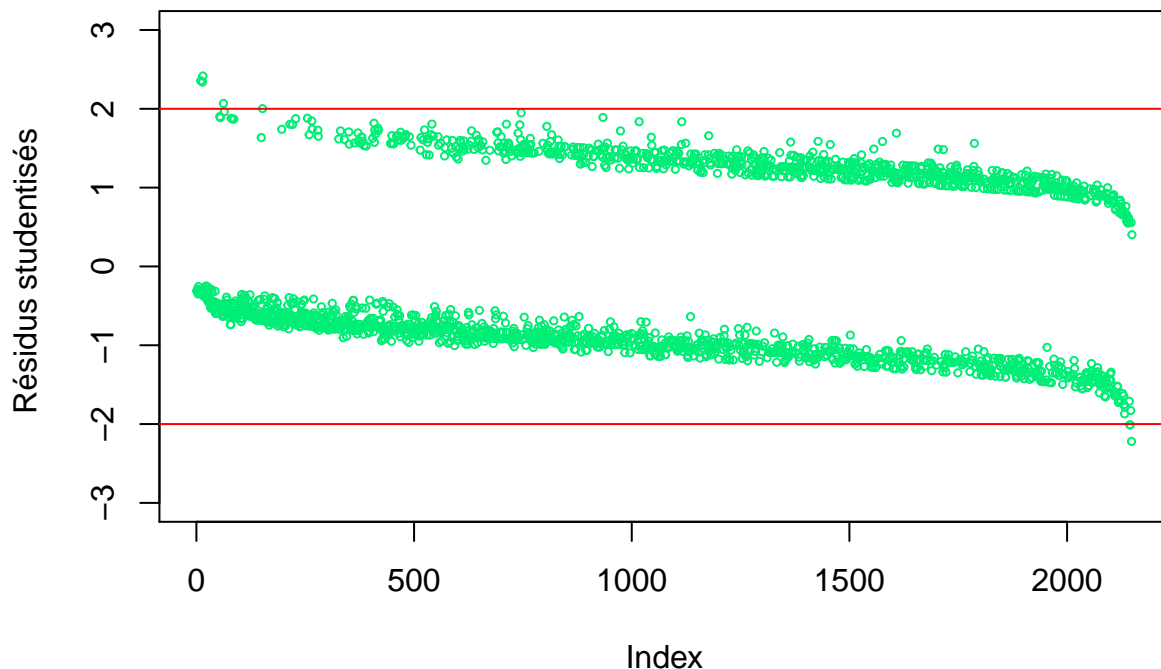
	Pseudo.R.squared
McFadden	0.09
Cox and Snell (ML)	0.11
Nagelkerke (Cragg and Uhler)	0.15

TABLE 11: Simple

Les valeurs du pseudo- R^2 ne sont pas très grandes, il va falloir poursuivre notre diagnostic pour pouvoir conclure sur ce modèle et également s'attarder sur la capacité prédictive du modèle.

On vérifie qu'il n'y a pas de problème au niveau des résidus, en effet, si trop de valeurs se situent hors de l'intervall $[-2,2]$, cela signifie qu'il y a beaucoup de valeurs extrêmes, ce qui peut nuire à la qualité du modèle.

Etude des résidus



Globalement, les résidus sont bien compris entre -2 et 2, ce qui signifie qu'il n'y a pas trop de variables extrêmes, nous n'auront pas de problème avec ce modèle. Il faut être vigilant car la loi logistique tend à attribuer aux valeurs extrêmes une probabilité plus forte que probit.

Vérifions maintenant qu'il n'y a pas de multicolinéarité dans le modèle en calculant les VIF⁷ du modèle :

6. Il faut rester prudent avec ces indicateurs, en effet tout comme les modèles linéaires avec le R^2 , il ne reste qu'un indicateur et ne doit pas être pris à la lettre comme nous allons le voir, notre modèle est viable même avec un faible pseudo- R^2 , il ne peut pas se substituer à une étude complète.








7. Variance Inflation Factor, quotient de la variance d'un modèle avec toutes les variables sur la variance d'un modèle ne possédant que l'une des variables, ainsi, plus le VIF est proche de 1 et moins il y a de risque de multicolinéarité de la variable.

Variable	VIF
retire	1.00932009415571
hisp	1.05320698024992
lnincome	1.17804711319185
educyear	1.20884350245951

TABLE 12

Nous pouvons remarquer que les VIF de toutes les variables sont très proches de 1, il n'y a donc pas de problème de multicolinéarité dans notre modèle.

Continuons avec l'interprétation du modèle, en effet, un modèle logistique possède une caractéristique spéciale que l'on ne retrouve pas chez Probit : les coefficients ne sont pas interprétables (tout du moins facilement), et donc pour se faciliter la tâche, nous calculons les Odds-Ratios des coefficients des variables explicatives du modèle. Les Odds-Ratios de notre modèles sont résumés dans la table ci-dessous :

Variable	N	Odds ratio	p
hisp	0 2001		Reference
	1 148		0.60 (0.37, 0.93) 0.03
educyear	2149		1.06 (1.03, 1.10) <0.001
retire	0 835		Reference
	1 1314		1.26 (1.04, 1.52) 0.02
lnincome	2149		1.99 (1.75, 2.27) <0.001
(Intercept)			0.02 (0.01, 0.04) <0.001

0.02 0.05 0.1 0.2 0.5 1 2

Vérifications :

- Tous les intervalles de confiance ne contiennent pas 1
- Toutes les p-values sont bien inférieures à 0.05

Si l'une de ces condition n'est pas validée, l'Odd-Ratio n'est pas interprétable. Puisque ce n'est pas le cas ici, tous les Odds-Ratios sont interprétables.

Ainsi :

- **retire** : Toutes choses égales par ailleurs, une personne retraitée a 1.26 fois plus de chance de souscrire à une assurance supplémentaire.
- **lnincome** : Toutes choses égales par ailleurs, une personne ayant un revenu élevé a 2 fois plus de chances de souscrire à une assurance supplémentaire.
- **educyear** : Toutes choses égales par ailleurs, une personne ayant un cursus scolaire long a 1.06 fois plus de chances de souscrire à une assurance supplémentaire.
- **hisp** : Toutes choses égales par ailleurs, une personne hispanique a 0.6 fois moins de chances de souscrire à une assurance supplémentaire.

Intéressons-nous maintenant aux prédictions de notre modèle :

		Réalité	
		Achat	Achat
Prédictions	Achat	910	342
	Achat	407	490

TABLE 13: Matrice de confusion

A première vue, le modèle détecte bien les vrai négatifs et moins bien les vrai positifs, nous allons donc approfondir notre étude avec les données de la table 13 :

	Modèle entraînement
Accuracy	0.6514658
95% IC	[0.6163, 0.6574]
No Information Rate IC	0.6128
P-value [Acc>NIR]	0.01106
Kappa	0.1834
Sensitivity	0.6909643
Specificity	0.5889423

TABLE 14

Dans un premier temps, il faut vérifier si l'Accuracy est significativement différente du No Information Rate. Si ce n'est pas le cas, le modèle sera moins bon qu'un modèle basé sur de l'aléatoire. Dans ce cas, la P-value du test est bien inférieur à 0.05, ainsi notre modèle est bien significativement meilleur qu'un modèle "aléatoire". On remarque une statistique de Kappa de 0.18, ce qui est faible mais acceptable.

Il est très important de tester notre modèle sur l'échantillon de test prévu à cet effet, ainsi cela nous permet de détecter plusieurs problèmes, tels que le sur-apprentissage ou le sous-apprentissage⁸.

		Réalité	
		Achat	Achat
Prédictions	Achat	371	91
	Achat	277	318

8. Le sur-apprentissage survient dans un modèle logit ou probit lorsque le modèle est inutilement complexe, la capacité prédictive du modèle en est fortement affecté, il se traduit par une très bonne prédiction sur l'échantillon d'entraînement, mais une très mauvaise prédiction sur l'échantillon de test. A l'inverse, le sous-apprentissage survient lorsque le modèle est excessivement simple, la prédiction sera alors très mauvaise sur les deux échantillons car le modèle ne permet pas de discriminer correctement.

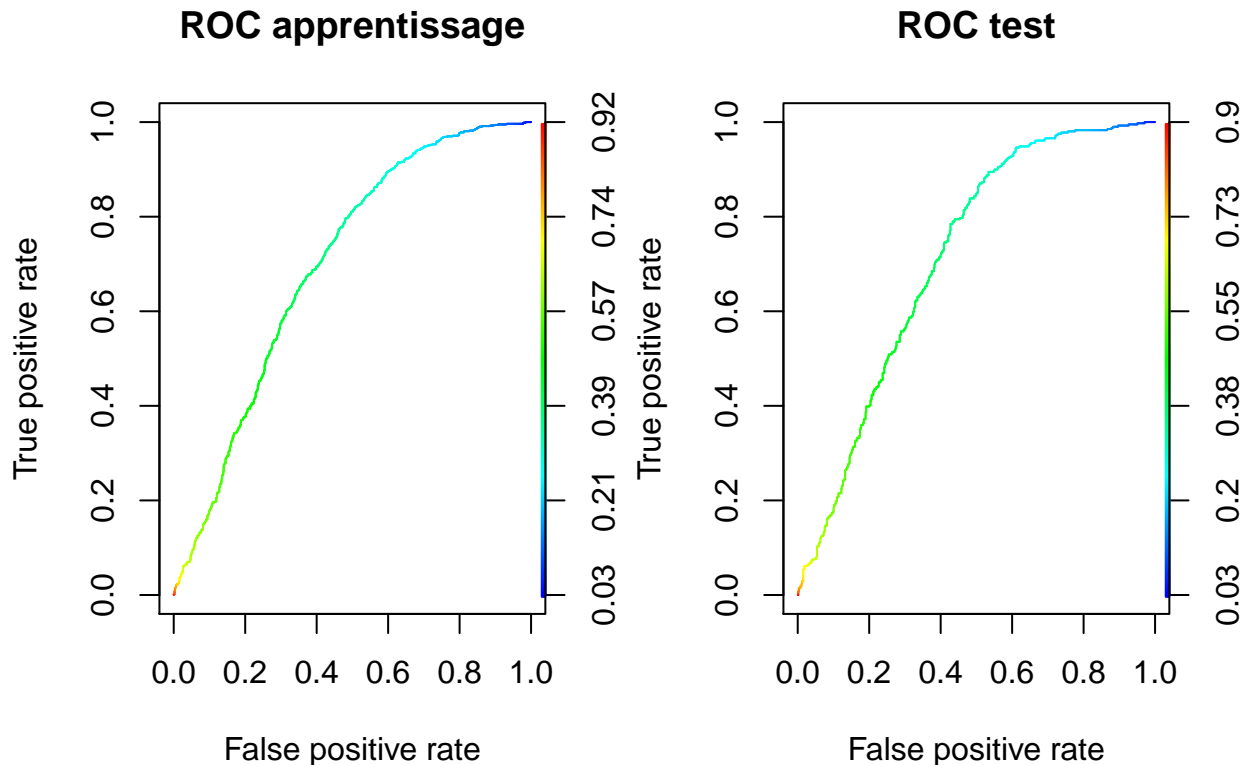
TABLE 15: Matrice de confusion

Dan un premier temps, la matrice de confusion nous montre que le modèle détecte très bien les vrai positifs, ce qui est encourageant, nous allons pouvoir continuer l'étude avec les informations contenues dans la table 15) :

	Modèle de test
Accuracy	0.6518448
95% IC	[0.6079, 0.6667]
No Information Rate	0.6131
P-value [Acc>NIR]	0.05322
Kappa	0.1757
Sensitivity	0.5725309
Specificity	0.7775061

TABLE 16

Comme sur les données d'entraînement, la précision (accuracy) du modèle reste bonne et significativement différente du No Information Rate. La sensibilité⁹ et la spécificité¹⁰ sont également bonnes, le modèle est donc utilisable. Terminons avec l'étude des courbes ROC du modèle :



L'aire sous la courbe ROC sur les données d'apprentissage est égale à 0.6958377, en test elle est à 0.711699, ce qui est globalement correct. Nous n'avons pas de problème de sur-apprentissage puisque le résultat est bon

9. Capacité du modèle à détecter les négatifs.

10. Capacité du modèle à détecter les positifs.

sur les données de test et nous n'avons pas de sous-apprentissage puisque les deux résultats sont globalement bons.

Maintenant que nous avons estimé notre premier modèle, essayons de l'améliorer en ajoutant des variables en interaction.

4.1.3 Variables en interaction

Les variables sélectionnées ont une influence sur la variable à expliquer, mais certaines variables ont des influences entre elles, et elles peuvent être également liées à la variable à expliquer. Intuitivement, nous pouvons penser à une dépendance entre les variables de revenu et de temps passé à l'école.

TABLE 17

	<i>Dependent variable :</i>
	ins
retire1	0.459 (0.550)
hisp1	-5.916*** (1.916)
educyear	0.381*** (0.083)
lnincome	1.559*** (0.289)
educyear :lnincome	-0.079*** (0.021)
hisp1 :lnincome	1.717*** (0.575)
retire1 :educyear	-0.068* (0.039)
retire1 :lnincome	0.163 (0.130)
Constant	-7.220*** (1.051)
Observations	2,149
Log Likelihood	-1,292.093
Akaike Inf. Crit.	2,602.186
<i>Note :</i> *p<0.1 ; **p<0.05 ; ***p<0.01	

Comme attendu, l'algorithme a sélectionné une interaction significative entre *lnincome :educyear*, mais il en a sélectionné d'autres que nous allons devoir trier.

On remarque que les coefficients associés aux variables *retire*, *retire :educyear* et *retire :lnincome* ne sont pas significativement différents de zéro, faisons un test de Wald pour affiner nos premières impressions.

	Df	Chisq	Pr(>Chisq)
retire	1	4.53	0.0333
hisp	1	0.88	0.3491
educyear	1	8.57	0.0034
lnincome	1	100.68	0.0000
educyear :lnincome	1	13.53	0.0002
hisp :lnincome	1	8.91	0.0028
retire :educyear	1	3.03	0.0817
retire :lnincome	1	1.59	0.2080

TABLE 18: Test de Wald

Le test de Wald confirme la non-significativité de la variable *retire :educyear* ainsi que *retire :lnincome*, mais également celle de *hisp*. Nous n'allons pas la supprimer tout de suite, attendons d'observer l'effet qu'aura la suppression des variables en interaction non significatives avant de nous prononcer. Nous allons confirmer notre choix par un test de rapport de vraisemblances.

Le modèle complexe est le modèle complet de la table 16), le modèle simplifié est le modèle complet mais sans les variables *retire :educyear* et *retire :lnincome*.

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
1	7	-1293.86			
2	9	-1292.09	2	3.53	0.1709

TABLE 19

Le test conserve comme attendu le modèle simplifié, nous retirons les deux variables en interactions.

Relançons un test de Wald sur le modèle simplifié :

	Df	Chisq	Pr(>Chisq)
retire	1	4.58	0.0324
hisp	1	1.10	0.2937
educyear	1	8.96	0.0028
lnincome	1	102.89	0.0000
educyear :lnincome	1	13.42	0.0002
hisp :lnincome	1	9.16	0.0025

TABLE 20

Le test de Wald confirme la non-significativité de la variable *hisp*, essayons de relancer un test de Wald sans la variable en interaction :

	Df	Chisq	Pr(>Chisq)
retire	1	4.24	0.0395
hisp	1	2.95	0.0861
educyear	1	9.38	0.0022
lnincome	1	104.61	0.0000
educyear :lnincome	1	18.77	0.0000

TABLE 21

La variable *hisp* n'est toujours pas significative, effectuons un test de rapport de vraisemblances pour sélectionner le modèle à conserver :

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
1	6	-1300.03			
2	9	-1292.09	3	15.87	0.0012

TABLE 22

Le test conserve le modèle complexe dans ce ca-là. Continuons avec un dernier test de rapport de vraisemblance entre le modèle sélectionné et le modèle ne comprenant pas la variable *hisp* :

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
1	5	-1301.58			
2	7	-1293.86	2	15.44	0.0004

TABLE 23

Le modèle complexe est retenu, néanmoins, la mauvaise significativité du coefficient *hisp* laisse à présager un risque au niveau des prédictions, c'est pourquoi nous allons confronter ces modèles au niveau de leurs prédictions. Terminons par effectuer un test de Wald sur le modèle simplifié :

	Df	Chisq	Pr(>Chisq)
retire	1	4.47	0.0344
educyear	1	11.90	0.0006
lnincome	1	108.82	0.0000
educyear :lnincome	1	20.61	0.0000

TABLE 24

Les deux modèles retenus sont donc :

TABLE 25

	<i>Dependent variable :</i>	
	ins	
	(1)	(2)
retire1	0.211** (0.099)	0.209** (0.099)
hisp1	-6.024*** (1.916)	
educyear	0.332*** (0.078)	0.392*** (0.075)
lnincome	1.639*** (0.282)	1.885*** (0.274)
educyear :lnincome	-0.078*** (0.021)	-0.094*** (0.021)
hisp1 :lnincome	1.742*** (0.576)	
Constant	-6.953*** (0.997)	-7.884*** (0.958)
Observations	2,149	2,149
Log Likelihood	-1,293.860	-1,301.581
Akaike Inf. Crit.	2,601.720	2,613.162
<i>Note :</i>	*p<0.1 ; **p<0.05 ; ***p<0.01	

4.1.4 Comparaison des modèles

Ajoutons notre modèle retenu sans variable en interaction :

TABLE 26

	<i>Dependent variable :</i>		
	ins		
	(1)	(2)	(3)
retire1	0.211** (0.099)	0.209** (0.099)	0.228** (0.098)
hisp1	-6.024*** (1.916)		-0.519** (0.236)
educyear	0.332*** (0.078)	0.392*** (0.075)	0.059*** (0.017)
lnincome	1.639*** (0.282)	1.885*** (0.274)	0.688*** (0.065)
educyear :lnincome	-0.078*** (0.021)	-0.094*** (0.021)	
hisp1 :lnincome	1.742*** (0.576)		
Constant	-6.953*** (0.997)	-7.884*** (0.958)	-3.696*** (0.265)
Observations	2,149	2,149	2,149
Log Likelihood	-1,293.860	-1,301.581	-1,310.013
Akaike Inf. Crit.	2,601.720	2,613.162	2,630.025
<i>Note :</i> *p<0.1 ; **p<0.05 ; ***p<0.01			

On remarque tout de suite que nos modèles en interaction sont meilleurs que le modèle précédemment retenu.

Pour départager nos 2 modèles restants, puisqu'ils ont un AIC comparable, nous allons nous baser sur les prédictions.

TABLE 27

	<i>Dependent variable :</i>	
	ins	
	(1)	(2)
retire1	0.211** (0.099)	0.209** (0.099)
hisp1	-6.024*** (1.916)	
educyear	0.332*** (0.078)	0.392*** (0.075)
lnincome	1.639*** (0.282)	1.885*** (0.274)
educyear :lnincome	-0.078*** (0.021)	-0.094*** (0.021)
hisp1 :lnincome	1.742*** (0.576)	
Constant	-6.953*** (0.997)	-7.884*** (0.958)
Observations	2,149	2,149
Log Likelihood	-1,293.860	-1,301.581
Akaike Inf. Crit.	2,601.720	2,613.162
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01	

Voici les prédictions des deux modèles :

	Modèle 1	Modèle 2
Accuracy	0.6471145	0.6499527
95% IC	[0.6175, 0.676]	[0.6203, 0.6787]
No Information Rate	0.6131	0.6131
P-value [Acc>NIR]	0.01215	0.007258
Kappa	0.2273	0.231
Sensitivity	0.7762346	0.7839506
Specificity	0.4425428	0.4376528

TABLE 28

Le modèle 2 possède une précision légèrement plus importante que le modèle 1, il a donc une capacité prédictive plus intéressante, même si elle est légère, cela doit être pris en compte lors du choix du modèle. De plus, la p-value du modèle 2 étant plus faible que celle du modèle 1 cela montre que la précision est plus fiable que celle du modèle 2. C'est donc pourquoi nous allons sélectionner le modèle 2 comme modèle final.

4.1.5 Modèle final

TABLE 29

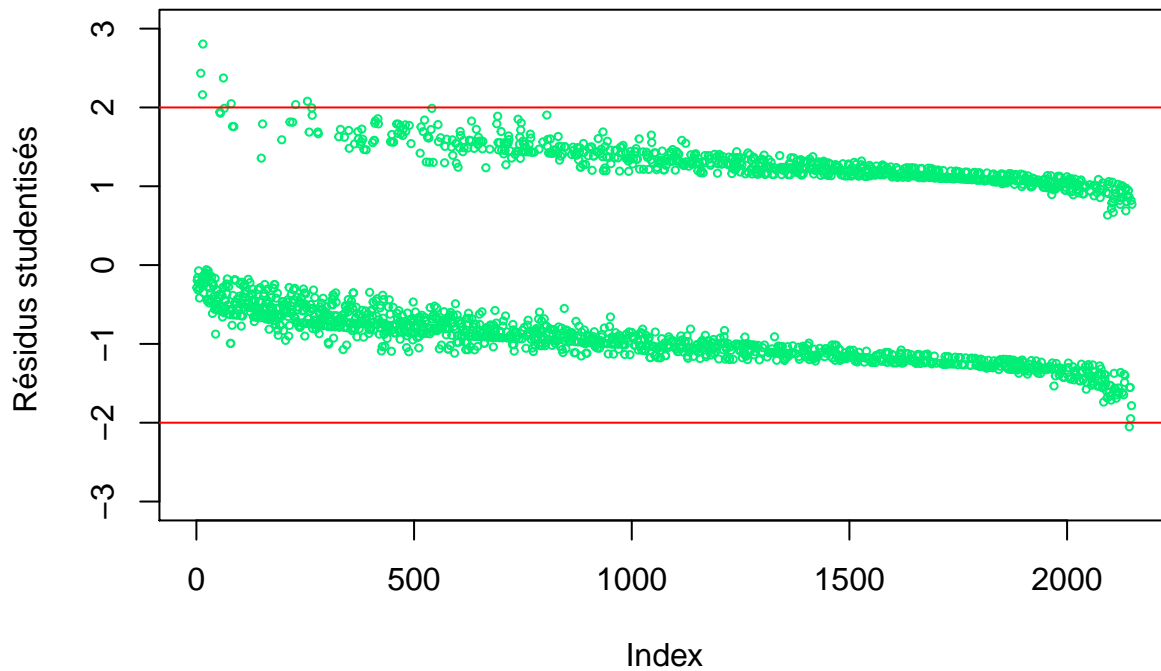
	<i>Dependent variable :</i>
	ins
retire1	0.209** (0.099)
educyear	0.392*** (0.075)
lnincome	1.885*** (0.274)
educyear :lnincome	-0.094*** (0.021)
Constant	-7.884*** (0.958)
Observations	2,149
Log Likelihood	-1,301.581
Akaike Inf. Crit.	2,613.162
<i>Note :</i> *p<0.1 ; **p<0.05 ; ***p<0.01	

Voici les valeurs du pseudo- R^2 par trois modes de calculs différents.

	Pseudo.R.squared
McFadden	0.09
Cox and Snell (ML)	0.12
Nagelkerke (Cragg and Uhler)	0.16

TABLE 30: fin







Le pseudo- R^2 de McFadden n'a pas changé, par contre celui des autres a légèrement augmenté.



Variable	VIF
retire	1.00965021336758
lnincome	20.2888344259488
educyear	20.6916130983677
educyear :lnincome	53.311562789151

TABLE 31

Au niveau de la répartition des résidus on ne voit pas réellement de problème. Au niveau des VIF par contre on remarque que les valeurs sont très largement différentes de 1, mais c'est tout à fait normal. En effet, la variable en interaction dépend totalement de deux variables du modèle, ainsi il est normal de retrouver des valeurs aussi élevées, il ne faut pas s'en inquiéter puisque nous avons déjà diagnostiqué le modèle sans ajout de variable en interaction et il n'avait aucun problème de multicollinéarité. Maintenant que nous avons sélectionné notre modèle, il est temps d'étudier ses coefficients, plus précisément les Odds-Ratios des coefficients.

Variable	N	Odds ratio	p
retire	0 835		Reference
	1 1314		1.23 (1.02, 1.50) 0.03
educyear	2149		1.48 (1.28, 1.72) <0.001
lnincome	2149		6.58 (3.89, 11.38) <0.001
(Intercept)			0.00 (0.00, 0.00) <0.001
educyear:lnincome			0.91 (0.87, 0.95) <0.001

0.0001 0.01 1

Interprétation des Odds-ratio :

L'interprétation des variables seules est la même que précédemment, mais nous pouvons tout de même remarquer un changement par rapport à l'autre modèle, cette fois c'est le revenu qui compte le plus, c'est-à-dire qu'une personne ayant un revenu plus important a quasiment 7 fois plus de chances de souscrire à une assurance supplémentaire !

Nous n'interpréterons pas l'odd-ratio de la variable en interaction, en effet, celui-ci est assez compliqué à interpréter. Par contre son effet marginal sur la souscription d'assurances est, lui, beaucoup plus abordable.

Prédictions modèle entraînement :

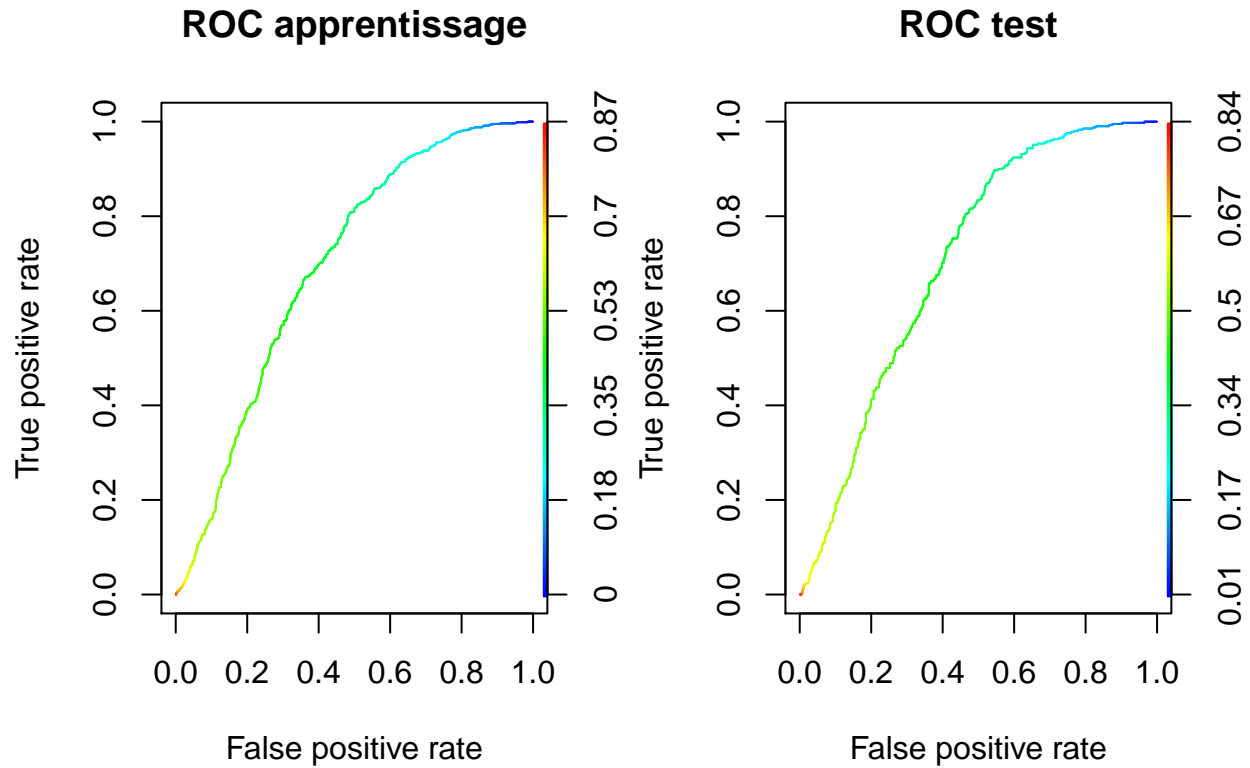
		Réalité	
		Achat	Achat
Prédictions	Achat	883	316
	Achat	434	516

TABLE 32: Matrice de confusion entraînement

Prédictions modèle test :

		Réalité	
		Achat	Achat
Prédictions	Achat	508	230
	Achat	140	179

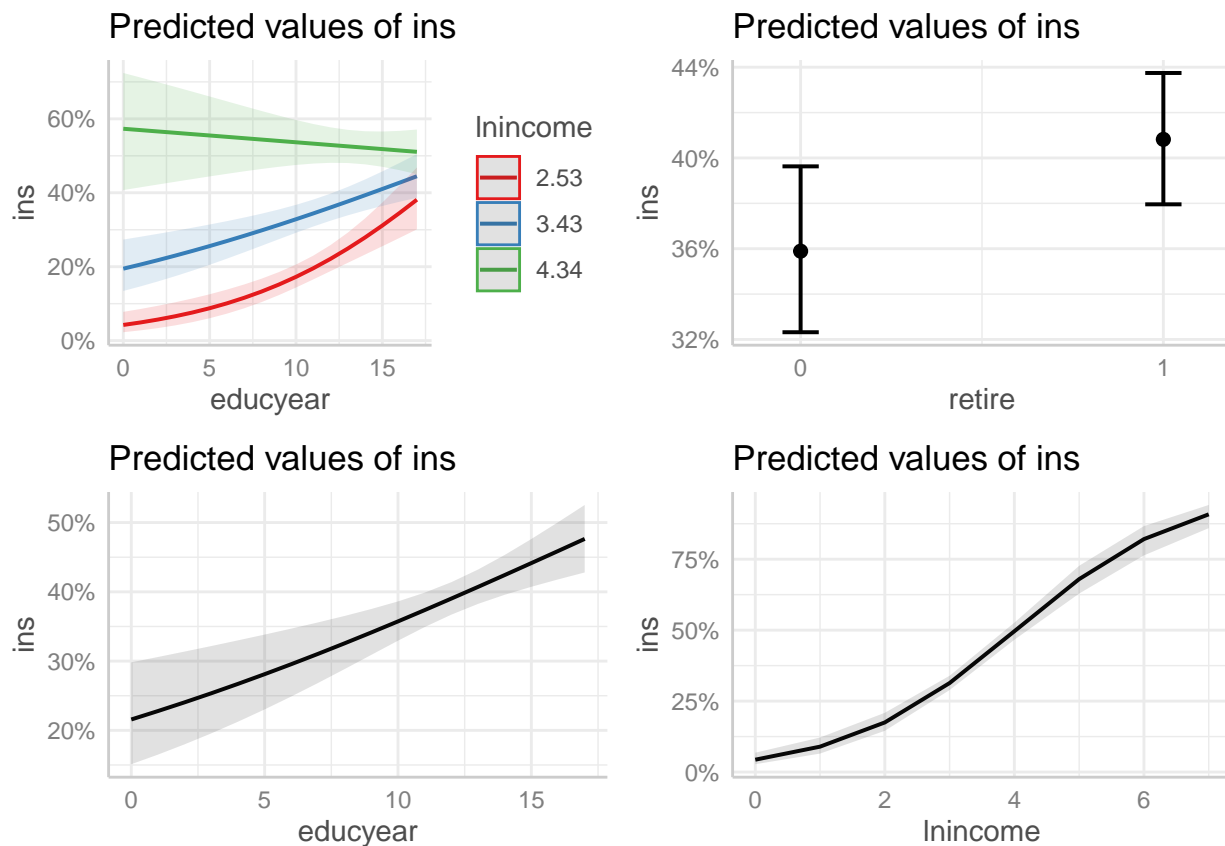
TABLE 33: Matrice de confusion test



L'aire sous la courbe ROC en apprentissage est de 0.6942958, l'aire sous la courbe ROC en test est de 0.7041772, ce qui est correct.

Il n'y a pas de sur-apprentissage ni de sous-apprentissage car les prédictions des deux échantillons sont correctes.

Pour terminer l'estimation logistique, il nous reste à étudier les effets marginaux du modèle :



	dF/dx	Std. Err.	p-value
retire1	0.0477143	0.0223566	0.03282
educyear	0.0903156	0.0169663	1.019e-07
lnincome	0.434257	0.0615469	1.717e-12
educyear :lnincome	-0.0215963	0.0046854	4042e-06

TABLE 34

Commençons par étudier les effets marginaux des variables seules :

- **retire** : L'effet marginal est positif, c'est-à-dire qu'une personne retraitée souscrit plus souvent à des assurances supplémentaires qu'une personne encore active.
- **educyear** : L'effet marginal est positif, ainsi, plus une personne a fait d'études et plus elle va être encline à souscrire à des assurances supplémentaires. Par cette interprétation, on peut commencer à ressentir le lien entre éducation et salaire.
- **lnincome** : L'effet marginal est positif, une personne plus aisée financièrement sera plus apte à souscrire à des assurances supplémentaires. Si on observe bien la courbe, on se rend compte qu'elle n'est pas aussi linéaire que celle du nombre d'années d'études, en effet, entre 0 et 2 puis 6+, la pente est légèrement plus faible qu'entre 2 et 6. Nous pouvons peut-être envisager une modification de notre réflexion, cela serait la classe moyenne qui aurait la plus grande demande pour des produits d'assurance complémentaires, cela peut-être intéressant d'envisager un produit destiné aux ménages les plus modestes. Mais nous allons pouvoir confirmer notre théorie grâce aux effets marginaux de la variable en interaction.

Il faut faire attention à la valeur de l'effet marginal de la variable *educyear :lnincome* se trouvant dans la table 32 car elle est donnée automatiquement par beaucoup de fonctions précompilées mais elle n'est pas interprétable. On le comprend lorsqu'on regarde le graphique, cette fois il n'y a pas une courbe mais trois. La variable *lnincome* a été séparée en trois classes, par tranche de revenu. On s'aperçoit alors que même si dans la table 32 le signe de l'effet marginal est négatif, il n'en est rien pour les tranches de revenus faibles et moyens, seule la tranche de revenus élevés est décroissante avec le nombre d'années d'études. Ainsi, les personnes à haut revenu sont bien celles qui s'assurent le plus, sauf que la tendance baisse avec le nombre d'années d'études. En revanche, pour les classes de revenus inférieurs, la tendance est clairement à la hausse avec le nombre d'années d'études, ce qui confirme nos prévisions faites lors de l'étude des variables seules.

4.2 Probit

Continuons notre étude avec une estimation Probit. Les modèles Logits et Probits sont toutefois très proches car le modèle Logit est lui-même un dérivé du modèle Probit. Ainsi, les commentaires des tests ne seront pas autant détaillés que dans la partie Logit car ils sont complètement similaires. Bien évidemment, s'il existe à un moment une différence, celle-ci sera développée.

TABLE 35: Modèle stepwise

<i>Dependent variable :</i>	
	ins
lnincome	0.413*** (0.039)
educyear	0.038*** (0.011)
hisp1	-0.304** (0.136)
retire1	0.128** (0.060)
adl	-0.087** (0.041)
white1	-0.120 (0.078)
Constant	-2.133*** (0.161)
Observations	2,149
Log Likelihood	-1,305.392
Akaike Inf. Crit.	2,624.784
<i>Note :</i> *p<0.1 ; **p<0.05 ; ***p<0.01	

Comme pour l'estimation Logit, la variable *white* ne semble pas significative, nous répétons la même procédure que précédemment, en effectuant un test de Wald :

	Df	Chisq	Pr(>Chisq)
lnincome	1	109.77	0.0000
educyear	1	12.80	0.0003
hisp	1	4.99	0.0255
retire	1	4.52	0.0335
adl	1	4.45	0.0350
white	1	2.39	0.1222

TABLE 36

Terminons avec un test du rapport des vraisemblances :

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
1	5	-1308.63			
2	6	-1307.65	1	1.96	0.1613

TABLE 37

Nous pouvons supprimer la variable *white*. Voici la nouvelle estimation du modèle :

TABLE 38

<i>Dependent variable :</i>	
ins	
hisp1	-0.313** (0.136)
educyear	0.036*** (0.011)
retire1	0.125** (0.060)
lnincome	0.404*** (0.039)
adl	-0.083** (0.041)
Constant	-2.180*** (0.159)
Observations	2,149
Log Likelihood	-1,306.564
Akaike Inf. Crit.	2,625.127
<i>Note :</i> *p<0.1 ; **p<0.05 ; ***p<0.01	

Le fait de retirer la variable *white* n'a pas affaibli la significativité de la variable *adl* comme dans le modèle logit, il vaudrait mieux effectuer un test de Wald, ainsi qu'un test du rapport des vraisemblances pour décider

de la conserver ou non selon la p-value, car celle-ci était déjà proche de la probabilité critique dans l'estimation Logit. Si cette fois-ci le modèle probit décide de la conserver, nous prenons le risque d'inclure dans notre modèle une variable n'ayant que peu de différence avec une variable non significative qui pourrait influencer sur les prédictions. Ainsi, si la p-value est trop proche du seuil critique, il faudra réfléchir quant au fait de conserver la variable.

	Df	Chisq	Pr(>Chisq)
hisp	1	5.32	0.0211
educyear	1	11.88	0.0006
retire	1	4.36	0.0369
lnincome	1	107.91	0.0000
adl	1	4.07	0.0436

TABLE 39

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
1	5	-1308.63			
2	6	-1306.56	1	4.13	0.0421

TABLE 40

La variable est significative à un seuil de 0.05 pour le test de Wald (mais très proche de la probabilité critique), le test du rapport de vraisemblance conserve le modèle complexe cette fois-ci. Mais à la vue de la p-value, on ne peut pas réellement trancher. Ce choix est donc personnel, tout comme le choix de la probabilité critique, mais à la vue de la p-value des différents tests effectués et à la vue du modèle logit produit, il me semble plus rationnel d'écarter la variable *adl* et de conserver le même modèle simplifié. Effectuons tout de même un test de Wald sur le modèle simplifié pour confirmer notre choix :

	Df	Chisq	Pr(>Chisq)
hisp	1	5.59	0.0180
educyear	1	12.72	0.0004
retire	1	5.18	0.0229
lnincome	1	116.64	0.0000

TABLE 41

4.2.1 Premier modèle et diagnostic

Voici donc le modèle que nous conservons pour la suite (le même modèle que pour Logit) :

TABLE 42

<i>Dependent variable :</i>	
	ins
hisp1	−0.320** (0.135)
educyear	0.037*** (0.010)
retire1	0.136** (0.060)
lnincome	0.415*** (0.038)
Constant	−2.258*** (0.154)
Observations	2,149
Log Likelihood	−1,308.630
Akaike Inf. Crit.	2,627.260
<i>Note :</i> *p<0.1 ; **p<0.05 ; ***p<0.01	

La déviance du modèle est de 2617.2595399, avec un degré de liberté associé de 2144.

Voici les valeurs du pseudo- R^2 par trois modes de calculs différents.

	Pseudo.R.squared
McFadden	0.09
Cox and Snell (ML)	0.11
Nagelkerke (Cragg and Uhler)	0.15

TABLE 43: Simple

Comme pour le modèle logit, la valeur de cet indicateur n'est pas très élevée, malgré tout, nous avons tout de même des résultats corrects.

Regardons si nous avons des problèmes de multicolinéarité :

Variables	VIF
retire	1.00932009415571
hisp	1.05320698024992
lnincome	1.17804711319185
educyear	1.20884350245951

TABLE 44

Les VIF sont proches de 1, nous n'avons donc à priori pas de problèmes de multicolinéarité.

Attardons-nous maintenant sur les prédictions du modèle :

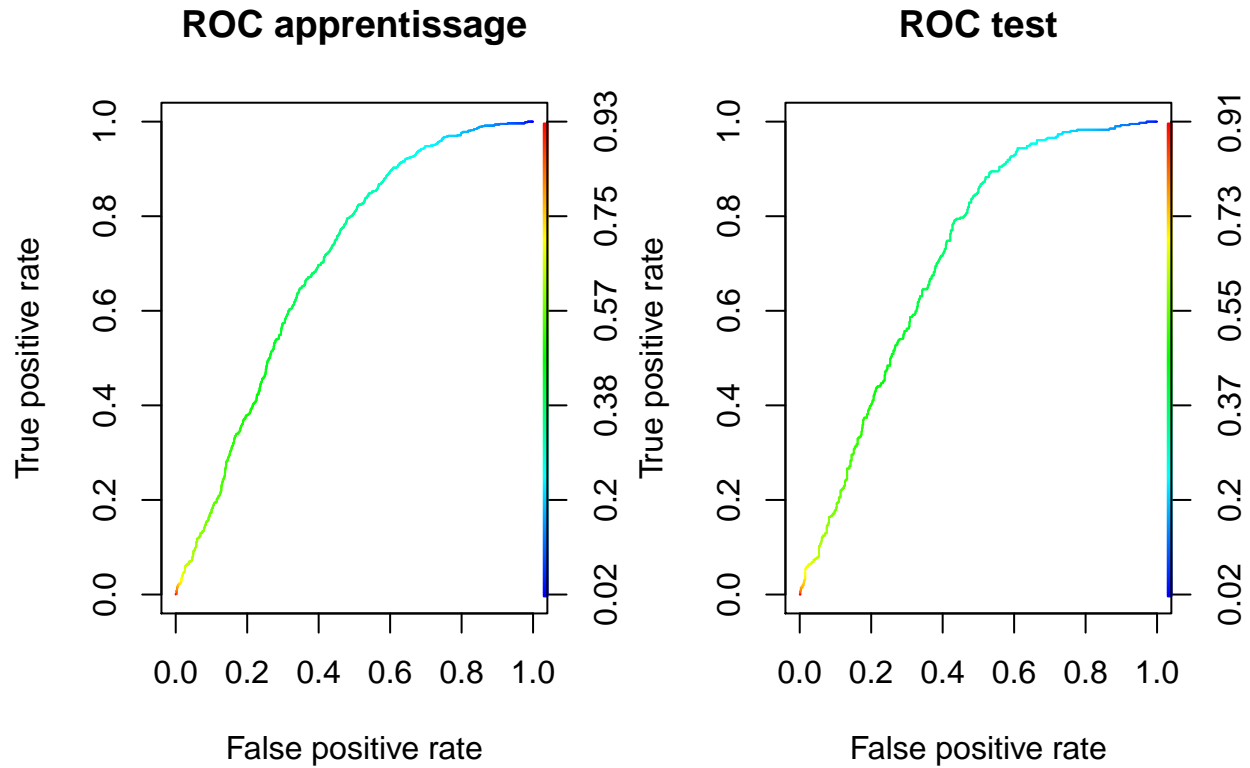
		Réalité	
		Achat	Achat
Prédictions	Achat	911	345
	Achat	406	487

TABLE 45: Matrice de confusion

	Modèle d'entraînement
Accuracy	0.6499527
95% IC	[0.6299, 0.6707]
No Information Rate	0.6128
P-value [Acc>NIR]	0.0001676
Kappa	0.2734
Sensitivity	0.7839506
Specificity	0.4376528

TABLE 46

Le modèle possède une bonne précision, bien différent du No Information Rate, et possède une bonne Kappa value, ce qui en fait un modèle assez intéressant.



L'aire sous la courbe ROC en apprentissage est de 0.6955785, et l'aire sous la courbe ROC en test est de 0.7118612 ce qui est correct.

		Réalité	
		Achat	Achat
Prédictions	Achat	368	87
	Achat	280	322

TABLE 47: Matrice de confusion

Modèle de test	
Accuracy	0.6499527
95% IC	[0.6203, 0.6787]
No Information Rate	0.6131
P-value [Acc>NIR]	0.007258
Kappa	0.231
Sensitivity	0.7839506
Specificity	0.4376528

TABLE 48

Tout comme pour le modèle Logit, nous allons essayer d'ajouter des variables en interaction :

4.2.2 Variables en interaction

En appliquant la même procédure que précédemment, adaptée au modèle probit, on obtient le modèle suivant :

TABLE 49

	<i>Dependent variable :</i>
	ins
retire1	0.238 (0.320)
hisp1	-3.073*** (0.997)
educyear	0.222*** (0.048)
lnincome	0.919*** (0.167)
educyear :lnincome	-0.046*** (0.013)
hisp1 :lnincome	0.894*** (0.306)
retire1 :educyear	-0.038 (0.023)
retire1 :lnincome	0.099 (0.077)
Constant	-4.268*** (0.596)
Observations	2,149
Log Likelihood	-1,291.641
Akaike Inf. Crit.	2,601.282
<i>Note :</i> *p<0.1 ; **p<0.05 ; ***p<0.01	

On remarque que les coefficients associés aux variables *retire*, *retire :educyear* et *retire :lnincome* ne sont pas significativement différents de zéro, faisons tout de même un test de Wald pour affiner nos premières impressions.

	Df	Chisq	Pr(>Chisq)
retire	1	4.43	0.0354
hisp	1	1.49	0.2217
educyear	1	10.51	0.0012
lnincome	1	107.09	0.0000
educyear :lnincome	1	13.45	0.0002
hisp :lnincome	1	8.57	0.0034
retire :educyear	1	2.67	0.1023
retire :lnincome	1	1.64	0.2006

TABLE 50

Appliquons un test de rapport des vraisemblances pour validation :

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
1	7	-1293.28			
2	9	-1291.64	2	3.28	0.1942

TABLE 51

Comme pour le modèle Logit, on remarque que la p-value de la variable *hisp* est très importante, ainsi, il nous faut poursuivre les tests jusqu'à obtenir le modèle optimal.

	Df	Chisq	Pr(>Chisq)
retire	1	4.56	0.0327
hisp	1	1.89	0.1696
educyear	1	10.98	0.0009
lnincome	1	108.93	0.0000
educyear :lnincome	1	13.34	0.0003
hisp :lnincome	1	8.84	0.0029

TABLE 52

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
1	5	-1300.71			
2	7	-1293.28	2	14.87	0.0006

TABLE 53

Comme précédemment, le modèle retenu est le modèle complet. Pour être en accord avec la méthodologie précédente, nous conserverons le modèle complet pour le confronter avec le modèle simplifié au niveau des prédictions. Effectuons un dernier test de Wald sur le modèle simplifié :

	Df	Chisq	Pr(>Chisq)
retire	1	4.36	0.0368
educyear	1	15.03	0.0001
lnincome	1	114.58	0.0000
educyear :lnincome	1	20.22	0.0000

TABLE 54

Les deux modèles retenus sont donc :

TABLE 55

	<i>Dependent variable :</i>	
	ins	
	(1)	(2)
retire1	0.128** (0.060)	0.125** (0.060)
hisp1	-3.138*** (0.996)	
educyear	0.196*** (0.045)	0.229*** (0.043)
lnincome	0.967*** (0.164)	1.099*** (0.158)
educyear :lnincome	-0.045*** (0.012)	-0.054*** (0.012)
hisp1 :lnincome	0.908*** (0.305)	
Constant	-4.138*** (0.570)	-4.647*** (0.542)
Observations	2,149	2,149
Log Likelihood	-1,293.280	-1,300.713
Akaike Inf. Crit.	2,600.560	2,611.427
<i>Note :</i>	*p<0.1 ; **p<0.05 ; ***p<0.01	

4.2.3 Comparaison des modèles

Pour effectuer une comparaison plus globale, ajoutons notre modèle retenu sans variable en interaction :

TABLE 56

	<i>Dependent variable :</i>		
		ins	
	(1)	(2)	(3)
retire1	0.128** (0.060)	0.125** (0.060)	0.136** (0.060)
hisp1	-3.138*** (0.996)		-0.320** (0.135)
educyear	0.196*** (0.045)	0.229*** (0.043)	0.037*** (0.010)
lnincome	0.967*** (0.164)	1.099*** (0.158)	0.415*** (0.038)
educyear :lnincome	-0.045*** (0.012)	-0.054*** (0.012)	
hisp1 :lnincome	0.908*** (0.305)		
Constant	-4.138*** (0.570)	-4.647*** (0.542)	-2.258*** (0.154)
Observations	2,149	2,149	2,149
Log Likelihood	-1,293.280	-1,300.713	-1,308.630
Akaike Inf. Crit.	2,600.560	2,611.427	2,627.260
<i>Note :</i> *p<0.1 ; **p<0.05 ; ***p<0.01			

On remarque tout de suite que nos modèles en interaction sont meilleurs que le modèle précédemment retenu.

Pour départager nos 2 modèles restants, puisqu'ils ont un AIC comparable, nous allons nous baser sur les prédictions.

TABLE 57

	<i>Dependent variable :</i>	
	ins	
	(1)	(2)
retire1	0.128** (0.060)	0.125** (0.060)
hisp1	-3.138*** (0.996)	
educyear	0.196*** (0.045)	0.229*** (0.043)
lnincome	0.967*** (0.164)	1.099*** (0.158)
educyear :lnincome	-0.045*** (0.012)	-0.054*** (0.012)
hisp1 :lnincome	0.908*** (0.305)	
Constant	-4.138*** (0.570)	-4.647*** (0.542)
Observations	2,149	2,149
Log Likelihood	-1,293.280	-1,300.713
Akaike Inf. Crit.	2,600.560	2,611.427
<i>Note :</i>	*p<0.1 ; **p<0.05 ; ***p<0.01	

Voici les prédictions des deux modèles :

	Modèle 1	Modèle 2
Accuracy	0.6461684	0.6508988
95% IC	[0.6175, 0.676]	[0.6203, 0.6787]
No Information Rate IC	0.6131	0.6131
P-value [Acc>NIR]	0.01215	0.007258
Kappa	0.2273	0.231
Sensitivity	0.757716	0.7638889
Specificity	0.757716	0.7638889

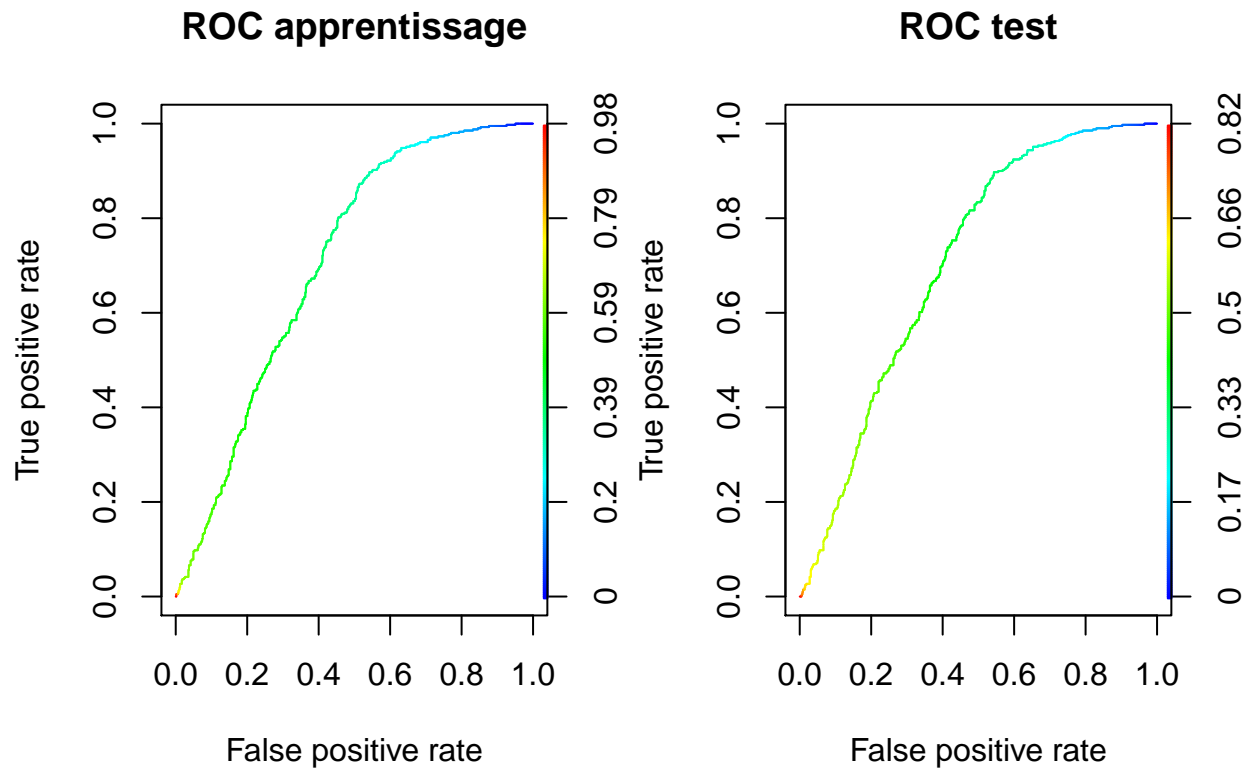
TABLE 58

Encore une fois, le modèle 2 a une précision plus importante que le modèle 1, et la p-value est encore une fois plus faible. Sachant que ce modèle produit de meilleures prédictions avec moins de variables, il est alors plus intéressant à conserver. On retient donc le même modèle que précédemment avec le modèle Logit.

4.2.4 Modèle final

TABLE 59

	<i>Dependent variable :</i>
	ins
retire1	0.125** (0.060)
educyear	0.229*** (0.043)
lnincome	1.099*** (0.158)
educyear :lnincome	-0.054*** (0.012)
Constant	-4.647*** (0.542)
Observations	2,149
Log Likelihood	-1,300.713
Akaike Inf. Crit.	2,611.427
<i>Note :</i> *p<0.1 ; **p<0.05 ; ***p<0.01	



L'aire sous la courbe ROC en apprentissage est de 0.7039263, l'aire sous la courbe ROC en test est de 0.7044263

Voici les prédictions du modèle :

		Réalité	
		Achat	Achat
Prédictions	Achat	883	316
	Achat	434	516

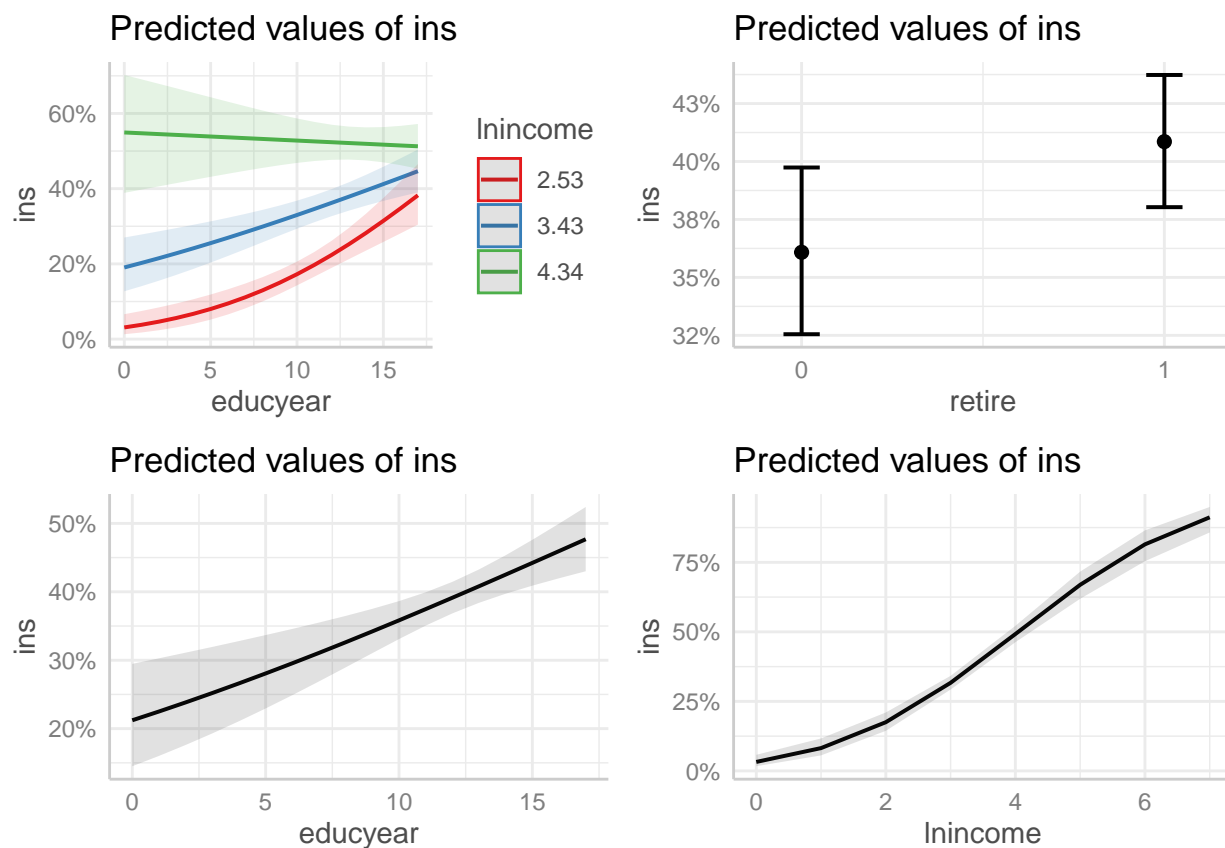
TABLE 60: Matrice de confusion entrainement

Les prédictions sont correctes.

Passons maintenant à l'étude des effets marginaux du modèle :

		Réalité	
		Achat	Achat
Prédictions	Achat	508	230
	Achat	140	179

TABLE 61: Matrice de confusion test



	dF/dx	Std. Err.	p-value
retire1	0.0466450	0.0221979	0.03561
educyear	0.0857509	0.0159647	7.817e-08
lnincome	0.4124981	0.0583299	1.529e-12
educyear :lnincome	-0.0202284	0.0044601	5.748e-06

TABLE 62

On remarque immédiatement que les effets marginaux sont exactement les mêmes que ceux du modèle Logit, les interprétations sont donc équivalentes.

5 Conclusion

Pour conclure cette étude, nous allons confronter les modèles Logit et Probit retenus :

TABLE 63

	<i>Dependent variable :</i>	
	ins	
	<i>probit</i>	<i>logistic</i>
	(1)	(2)
retire1	0.125** (0.060)	0.209** (0.099)
educyear	0.229*** (0.043)	0.392*** (0.075)
lnincome	1.099*** (0.158)	1.885*** (0.274)
educyear :lnincome	-0.054*** (0.012)	-0.094*** (0.021)
Constant	-4.647*** (0.542)	-7.884*** (0.958)
Observations	2,149	2,149
Log Likelihood	-1,300.713	-1,301.581
Akaike Inf. Crit.	2,611.427	2,613.162
<i>Note :</i>	*p<0.1 ; **p<0.05 ; ***p<0.01	

On remarque un lien assez intéressant entre les modèles, d'une part, les coefficients possèdent le même signe, de plus, nous pouvons également constater, selon l'approximation d'Amemiya, que les coefficients du modèle Logit sont équivalents à 1.6 fois ceux du modèle Probit ¹¹.

Finalement, si nous devons choisir entre ces deux modèles, à résultats équivalents, le modèle Logit est tout de même beaucoup plus simple à interpréter, et propose une interprétation plus complète avec les Odds-Ratios. Il faut tout de même rester vigilant avec le modèle Logit en cas de présence de variables extrêmes car cela peut le faire mal réagir.

Axes d'étude pour améliorer la vente de produits d'assurance :

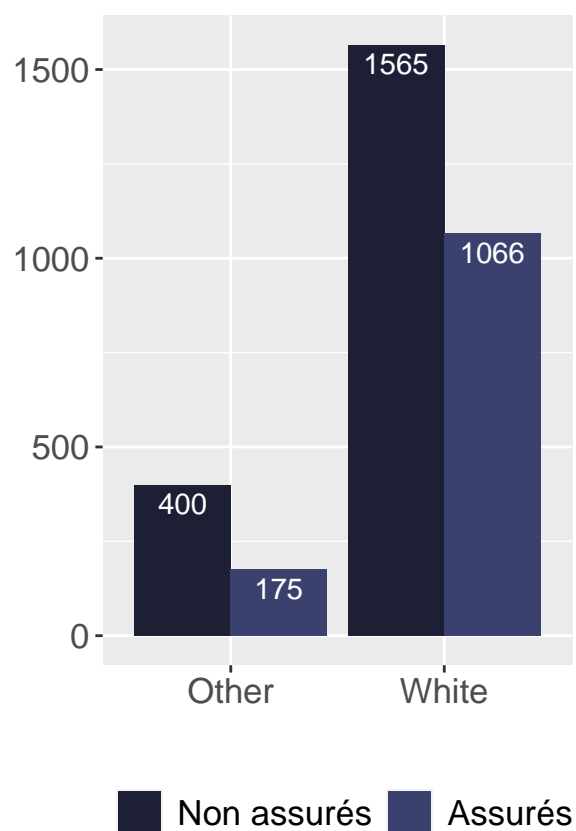
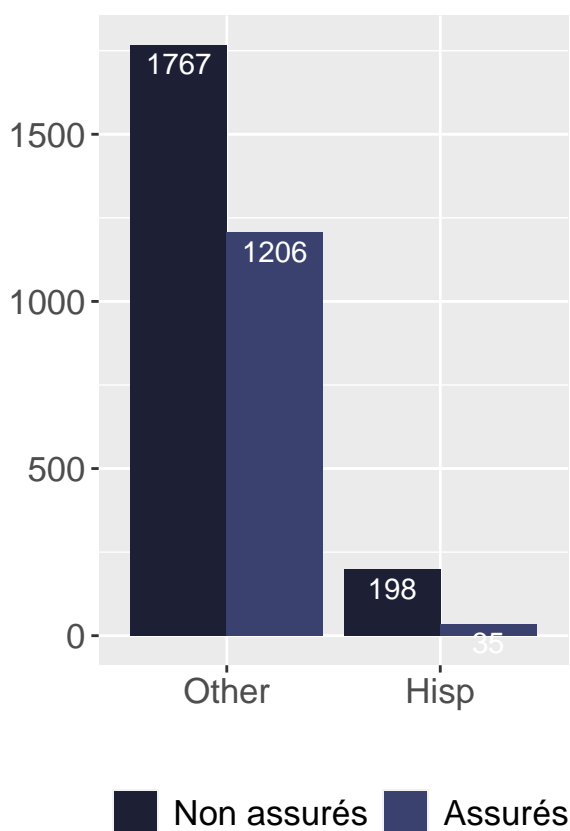
- Proposer une offre plus accessible financièrement et mettre en place des outils d'éducation pour permettre une meilleure compréhension des produits proposés, car une grande partie des non-acheteurs sont des personnes avec un faible niveau d'études et/ou un faible salaire.
- Proposer une offre pour les travailleurs handicapés et/ou malades (ceux pouvant bénéficier de Medicare) car ils sont très peu à souscrire à des assurances supplémentaire.
- Conserver et développer l'accompagnement proposé aux senior car ils représentent la majeure partie de la clientèle.

11. Voir les interprétations ici.

6 Annexe

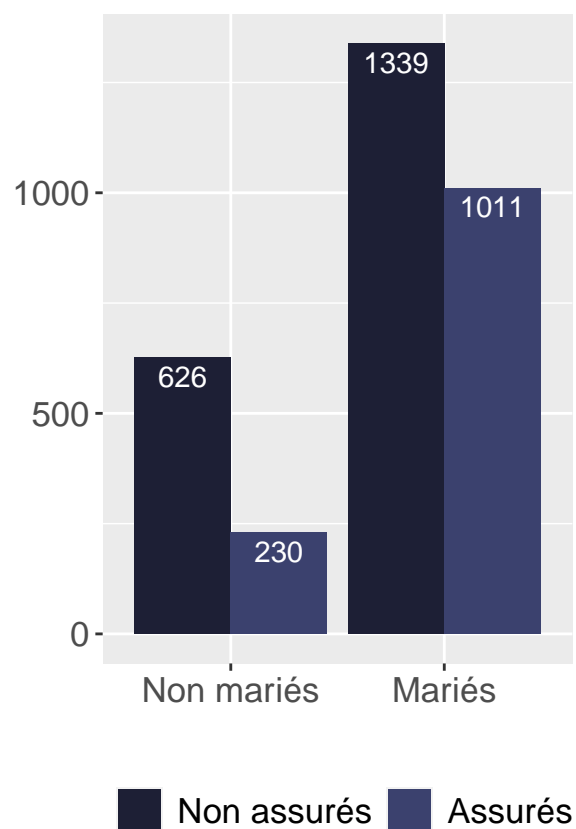
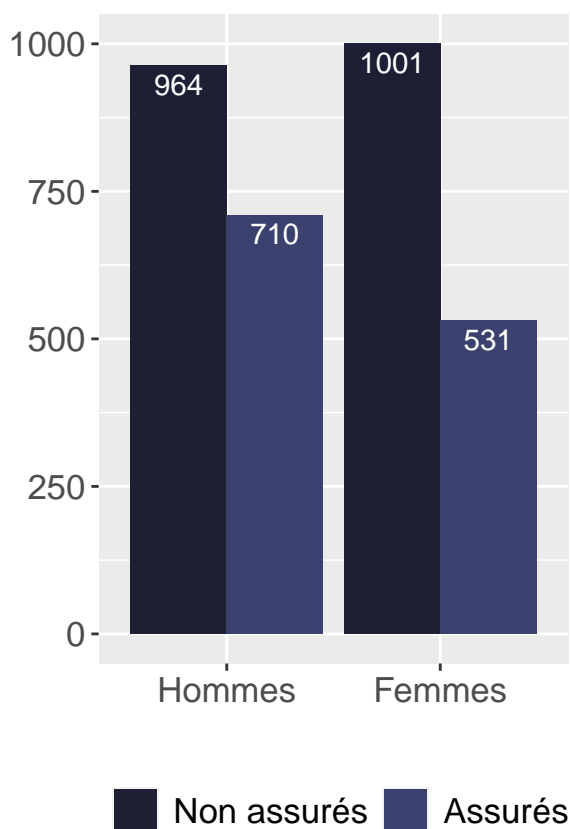
6.1 Répartition des variables qualitatives

Répartition des variables *hisp* et *white*¹² :

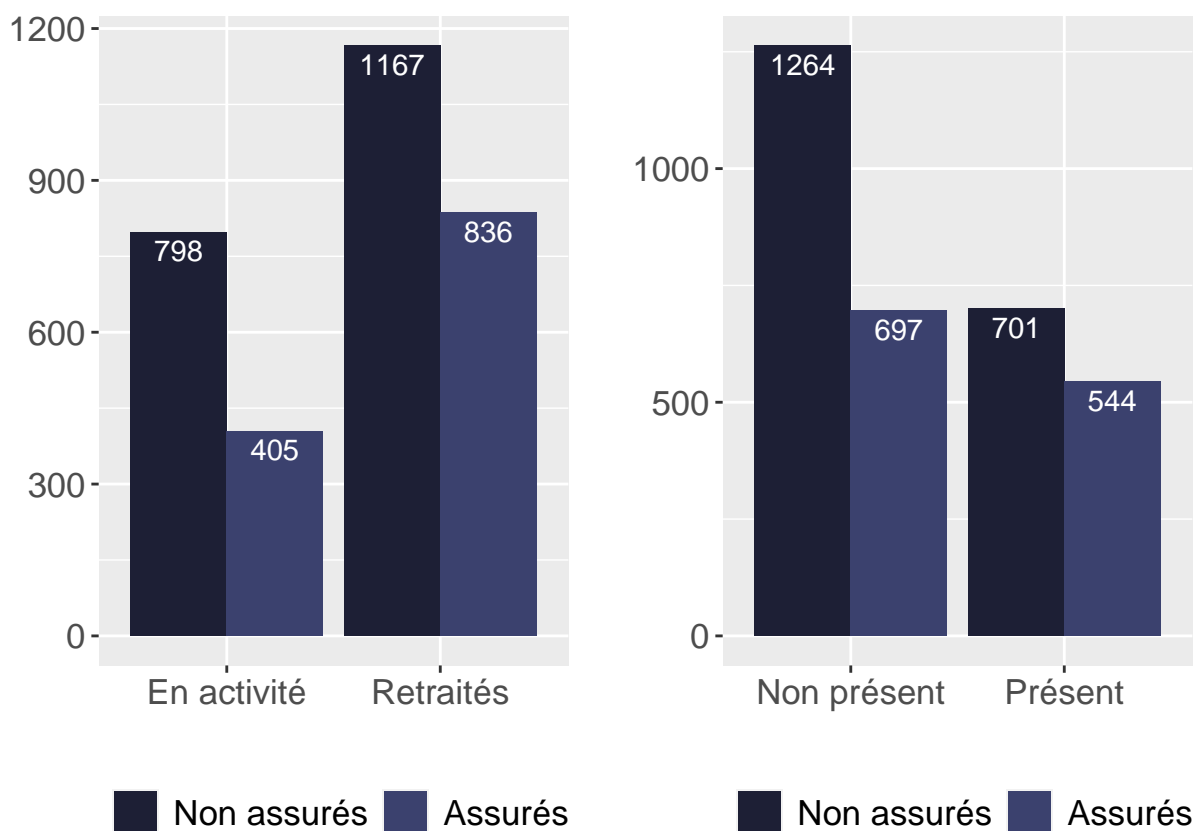


12. Retour à l'introduction ici.

Répartition des variables *female* et *married* :

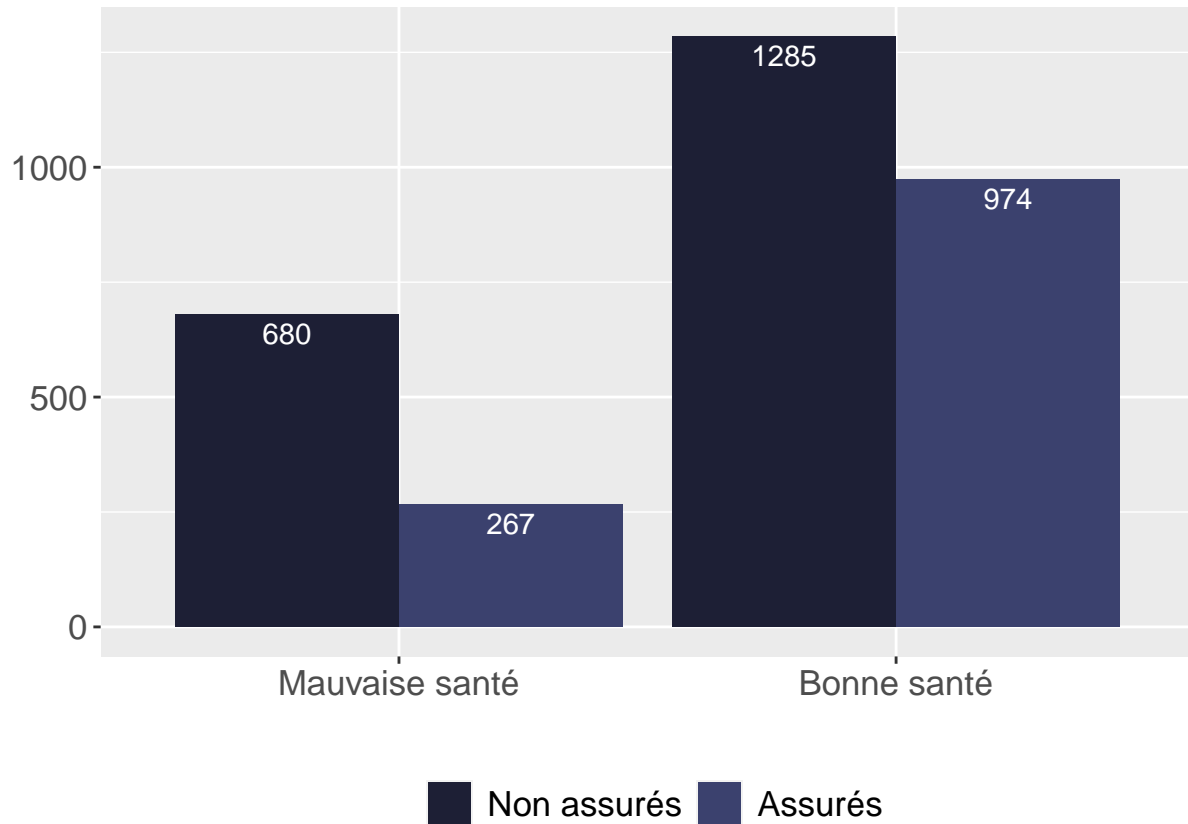


Répartition des variables *retire* et *sretire* :



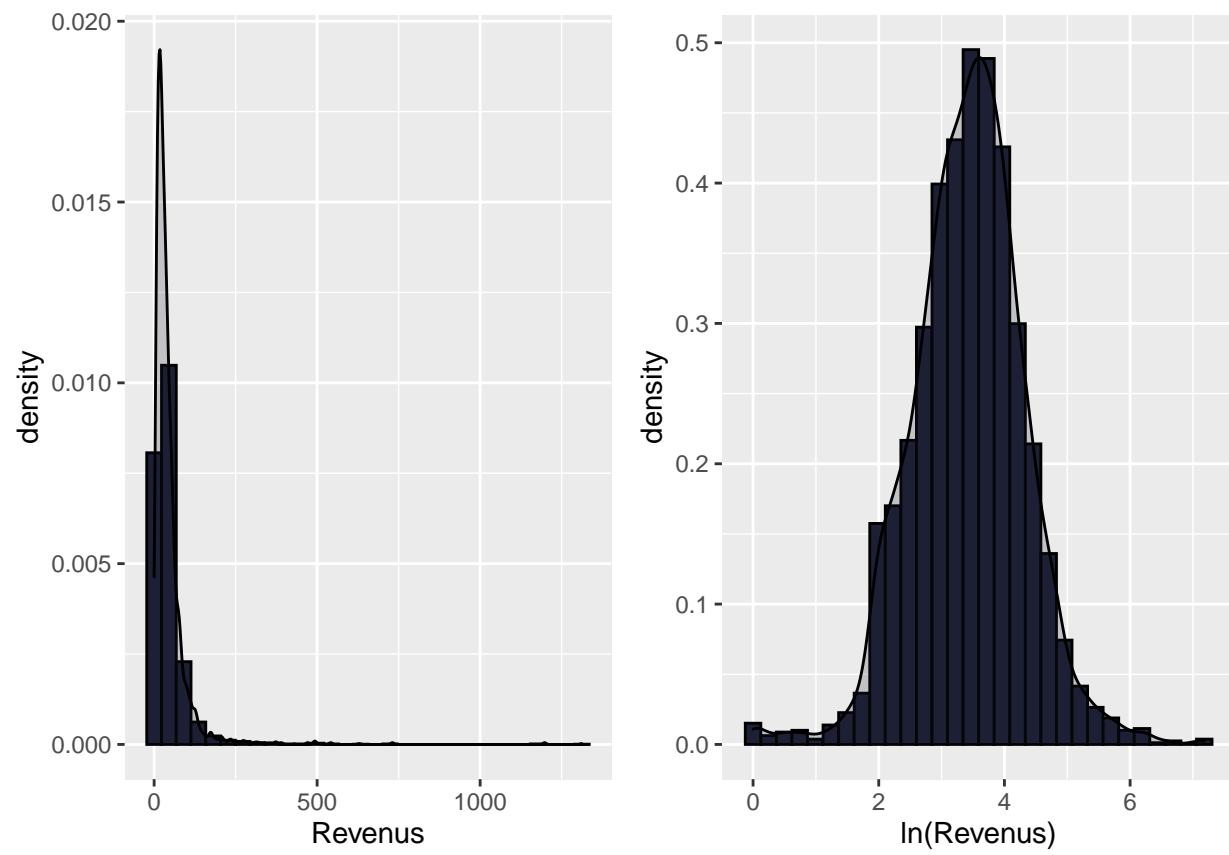
Par présent et non présent, nous entendons présence ou non d'un(e) époux-se à la retraite.

Répartition de *hstatusg* :

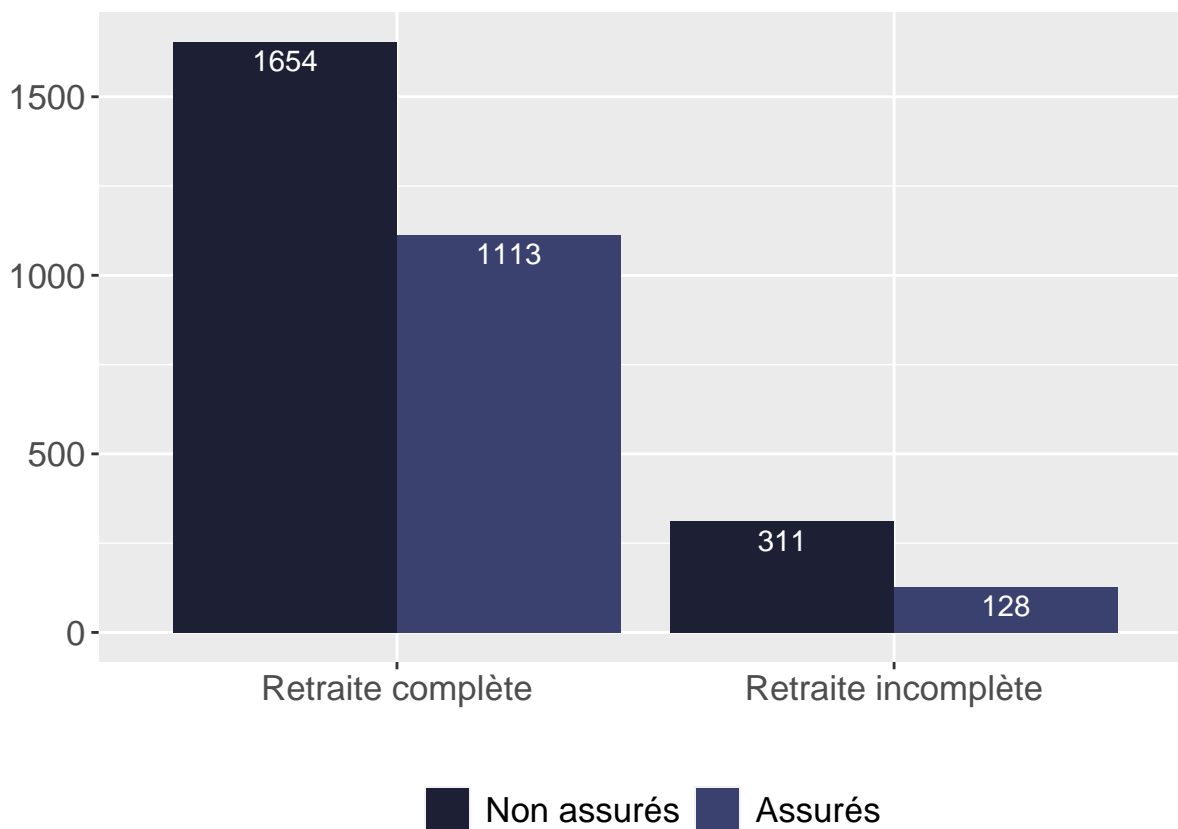


6.2 Répartition des variables quantitatives

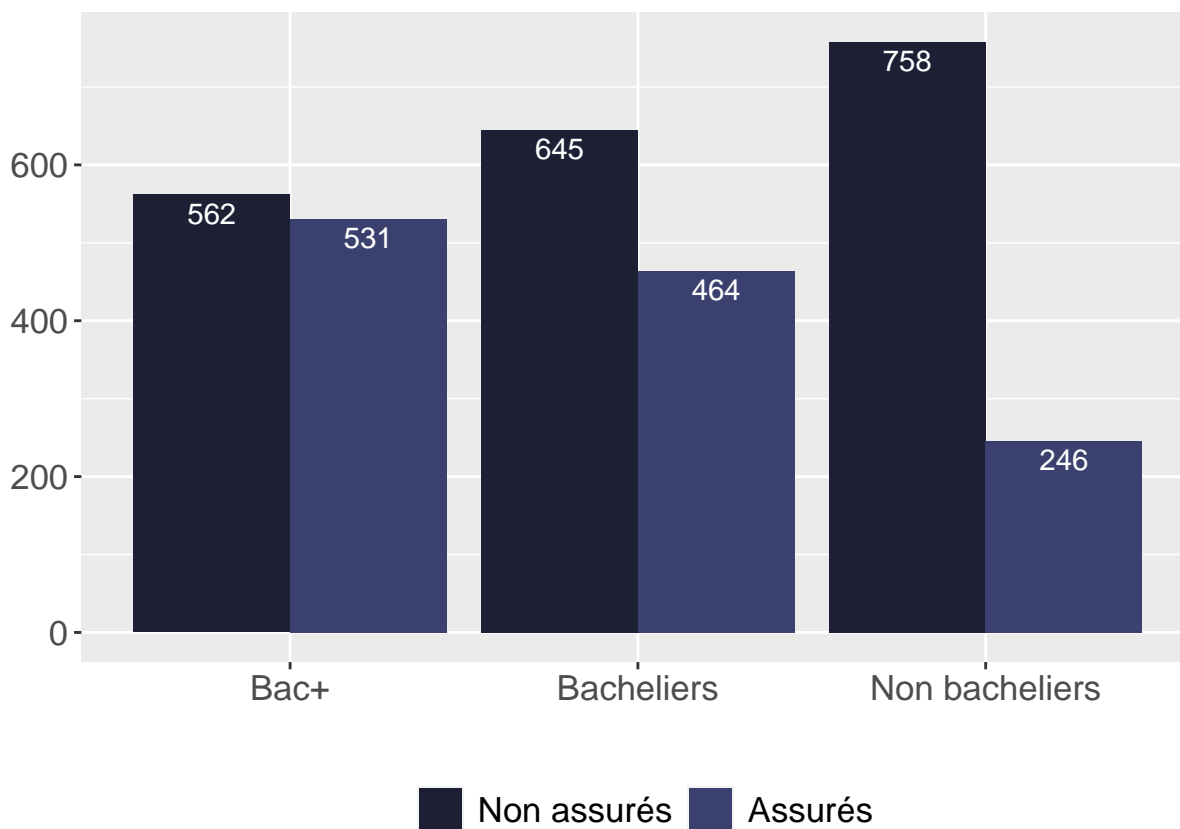
Répartition des variables *hhincome* et *lnincome* :



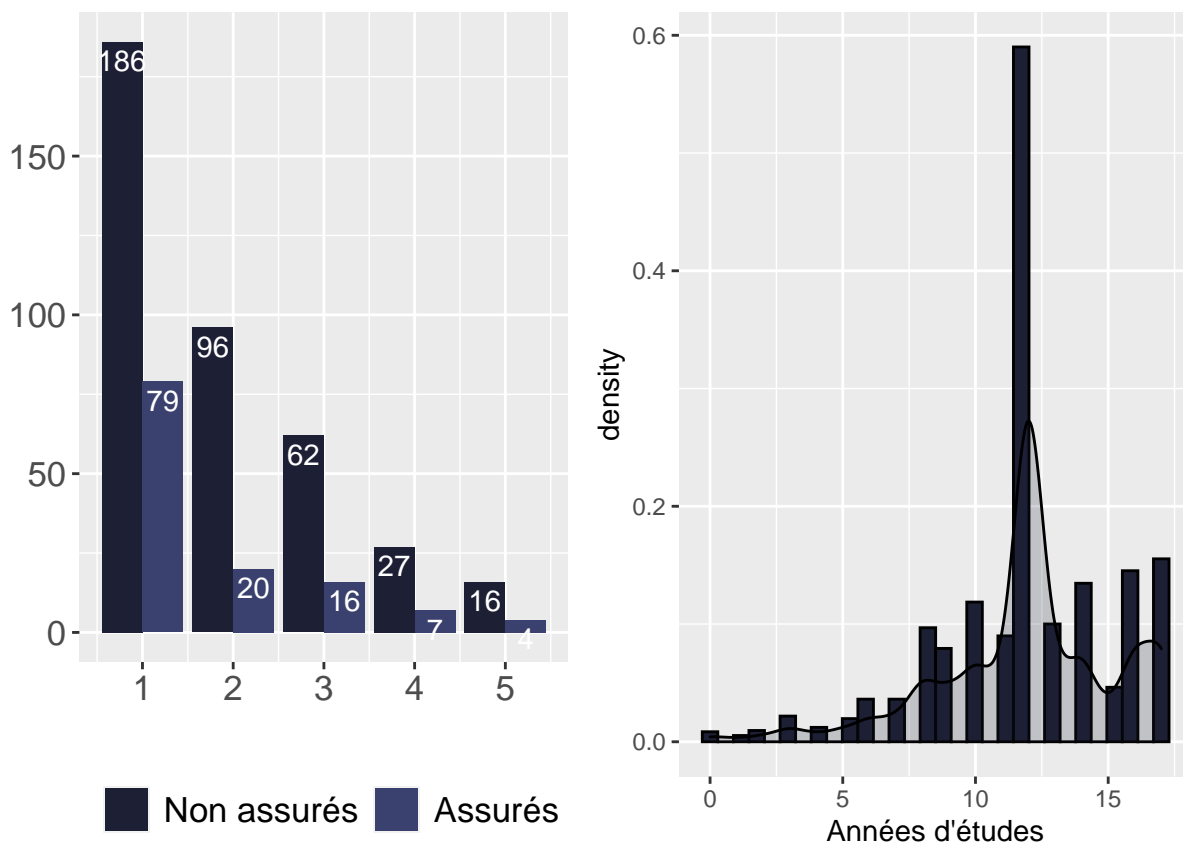
Fusion des variables *retire* et *age* afin de déterminer si la personne possède tous ses droit à la retraite ou non :



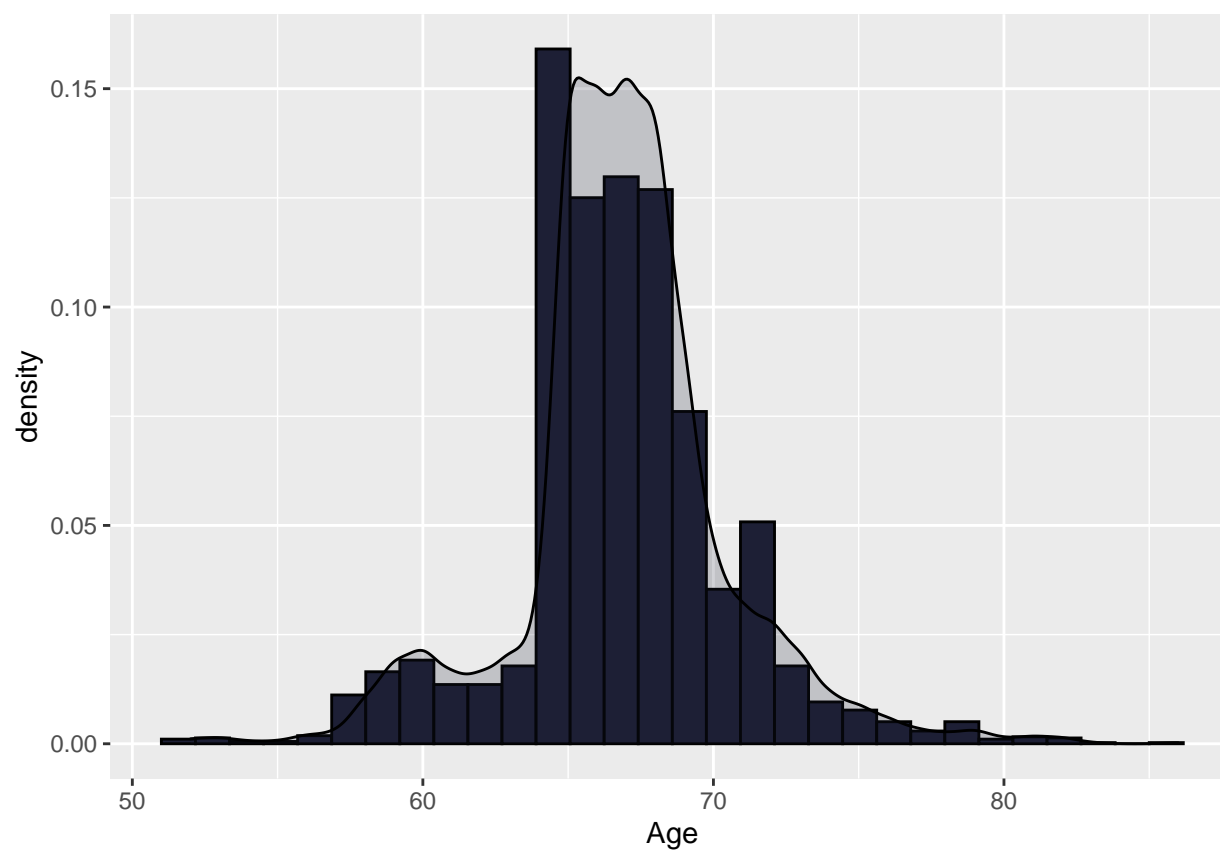
Simplification de la variable *educyear*, selon l'obtention de l'équivalent du Baccalauréat aux Etats-Unis :



Représentation “brute” de la variable *educyear* ainsi que de la variable *adl* :



Répartition de la variable *age* :



	50-60	60-65	65-70	70-80	80+	Total
Non assurés	90.00	221.00	1336.00	307.00	11.00	1965.00
Assurés	34.00	94.00	909.00	198.00	6.00	1241.00

TABLE 64

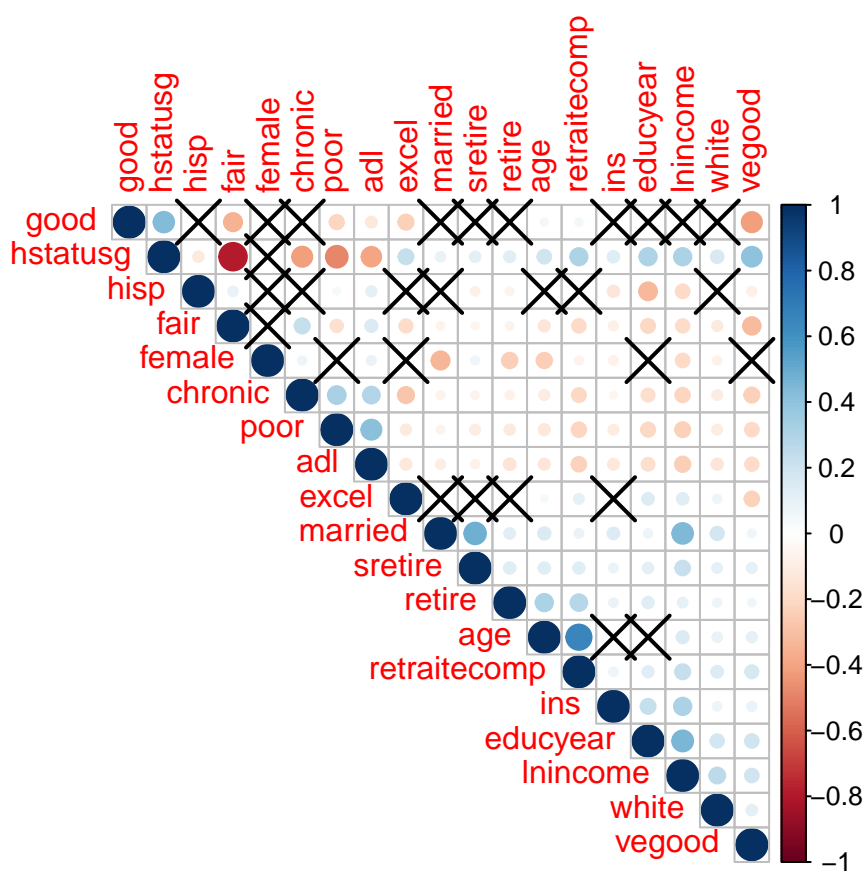
6.3 Matrices de corrélation et p-values détaillées

	age	hisp	white	female	educyear	married	excel	vegood	good	fair	poor
age		0.97	0.00	0.00	0.02	0.00	0.01	0.00	0.00	0.00	0.00
hisp	0.97		0.40	0.65	0.00	0.13	0.12	0.00	0.63	0.00	0.00
white	0.00	0.40		0.00	0.00	0.00	0.00	0.00	0.96	0.00	0.00
female	0.00	0.65	0.00		0.24	0.00	0.22	0.88	0.10	0.29	0.02
educyear	0.02	0.00	0.00	0.24		0.00	0.00	0.00	0.29	0.00	0.00
married	0.00	0.13	0.00	0.00	0.00		0.56	0.00	0.24	0.00	0.00
excel	0.01	0.12	0.00	0.22	0.00	0.56		0.00	0.00	0.00	0.00
vegood	0.00	0.00	0.00	0.88	0.00	0.00	0.00		0.00	0.00	0.00
good	0.00	0.63	0.96	0.10	0.29	0.24	0.00	0.00		0.00	0.00
fair	0.00	0.00	0.00	0.29	0.00	0.00	0.00	0.00	0.00		0.00
poor	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	
chronic	0.00	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.81	0.00	0.00
adl	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
retire	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.04	0.00	0.00
sretire	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.00	0.62	0.00	0.00
ins	0.08	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.19	0.00	0.00
hstatusg	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
lnincome	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.00	0.00

TABLE 65

	chronic	adl	retire	sretire	ins	hstatusg	lnincome
age	0.00	0.00	0.00	0.00	0.08	0.00	0.00
hisp	0.95	0.00	0.00	0.00	0.00	0.00	0.00
white	0.00	0.00	0.00	0.00	0.00	0.00	0.00
female	0.00	0.00	0.00	0.00	0.00	0.02	0.00
educyear	0.00	0.00	0.00	0.00	0.00	0.00	0.00
married	0.00	0.00	0.00	0.00	0.00	0.00	0.00
excel	0.00	0.00	0.03	0.12	0.02	0.00	0.00
vegood	0.00	0.00	0.00	0.00	0.00	0.00	0.00
good	0.81	0.00	0.04	0.62	0.19	0.00	0.19
fair	0.00	0.00	0.00	0.00	0.00	0.00	0.00
poor	0.00	0.00	0.00	0.00	0.00	0.00	0.00
chronic		0.00	0.00	0.00	0.00	0.00	0.00
adl	0.00		0.00	0.00	0.00	0.00	0.00
retire	0.00	0.00		0.00	0.00	0.00	0.00
sretire	0.00	0.00	0.00		0.00	0.00	0.00
ins	0.00	0.00	0.00	0.00		0.00	0.00
hstatusg	0.00	0.00	0.00	0.00	0.00		0.00
lnincome	0.00	0.00	0.00	0.00	0.00	0.00	

TABLE 66



6.4 Test du rapport des vraisemblances

Présentation du test ¹³ :

Posons $L(m_i)$ la vraisemblance du modèle i , K le degré de liberté de la loi du $\tilde{\chi}^2$ suivie par la statistique de test LR , $\loglik(m_i)$ la log-vraisemblance du modèle i , et α la probabilité critique.

Hypothèses du test :

$$\begin{cases} H_0 : \text{Modèle contraint} \\ H_1 : \text{Modèle non contraint} \end{cases}$$

Statistique de test :

$$LR = -2\ln\left(\frac{L(m_1)}{L(m_2)}\right) = 2(\loglik(m_1) - \loglik(m_2))$$

Avec :

$$LR \sim \tilde{\chi}^2(K)$$

Règle de décision :

- Si la p-value, $(Pr(> Chisq) < \alpha)$, alors on rejette H_0
- Si la p-value, $(Pr(> Chisq) > \alpha)$, alors on conserve H_0

6.5 Test de Wald

Présentation du test :

Hypothèse du test :

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

Statistique du test :

$$W = \frac{(\hat{\theta} - \theta_0)^2}{V(\hat{\theta})}$$

Avec :

$$W \sim \tilde{\chi}^2(K)$$

Règle de décision :

- Si la p-value, $(Pr(> Chisq) < \alpha)$, alors on rejette H_0
- Si la p-value, $(Pr(> Chisq) > \alpha)$, alors on conserve H_0

13. Retour à 4.1.1, retour à 4.2.1