projet marketing

KASHALA ILUNGA Caleb

03/04/2020

Table des matières

1 Nos variable étudiées

-La base de données que nous allons étudier est une base de données qui contient des informations sur diverses utilisations des services médicaux. Les personnes âgées peuvent obtenir une assurance complémentaire soit en l'achetant elles-mêmes ou en adhérant à des régimes parrainés par l'employeur.Le but de l'étude que nous allons ici réaliser,est de savoir si oui ou non les personnes âgées ont souscrit une assurance complémentaire,la variable *assuré* sera donc notre variable à expliquer.Pour ce faire nous allons commencer par une études globale sur la base de données puis ensuite nous allons procéder à une estimation d'un modéle logit et probit et définir le meilleur modéle de "prédiction".
-Notre base de données contient 3206 observations et 20 variables.
Nous observons que notre base de données a une varibale appelé "X" qui correspond à la numérotation des lignes de chaque individu, nous l'avons donc retiré cette variable car elle apporte aucune information importante dans l'étude que nous aurons à mener ici.
Il nous reste donc 19 variables sur notre base de données dont la variable à expliquer *assuré* avec lesquelles nous allons faire notre étude. - Nos variables explicatives sont :
Pour toute les variables la valeur 1 veut dire OUI et 0 veut dire NON sauf celle dont nous allons le préciser.

assurance_privé : Si oui ou non l'individu a souscrit une assurance privée,prend la valeur 1 ou 0 .

age :Qui correspond à l'âge de l'individu.

hispanique :Si oui ou non l'individu est hispanique,prend la valeur 1 et 0.

blanc :si oui ou non l'individu est blanc.

femme :Si oui ou non l'individu est une femmme,prend la valeur 1 et 0.

année_d_éducation: Qui correspond au nombre d'année d'étude éffectuer par l'individus.

marié:Si oui ou non l'individu est marié.

excellente_santé:Si la personne a oui ou non une excellente santé,prend la valeur 1 et 0.

trés_bonne_santé:Si la personne a oui ou non une trés bonne santé,prend la valeur 1 et 0.

bonne_santé:Si la personne a oui ou non une bonne santé,prend la valeur 1 et 0.

santé_passable:Si la personne a oui ou non une santépassable ,prend la valeur 1 et 0.

mauvaise_santé:Si la personne a oui ou non une mauvaise santé,prend la valeur 1 et 0.

maladie_chronique:Nombre de maladie chronique chez l'individu.

adl:le nombre de limitations (jusqu'à cinq) sur les activités de la vie quotidienne.

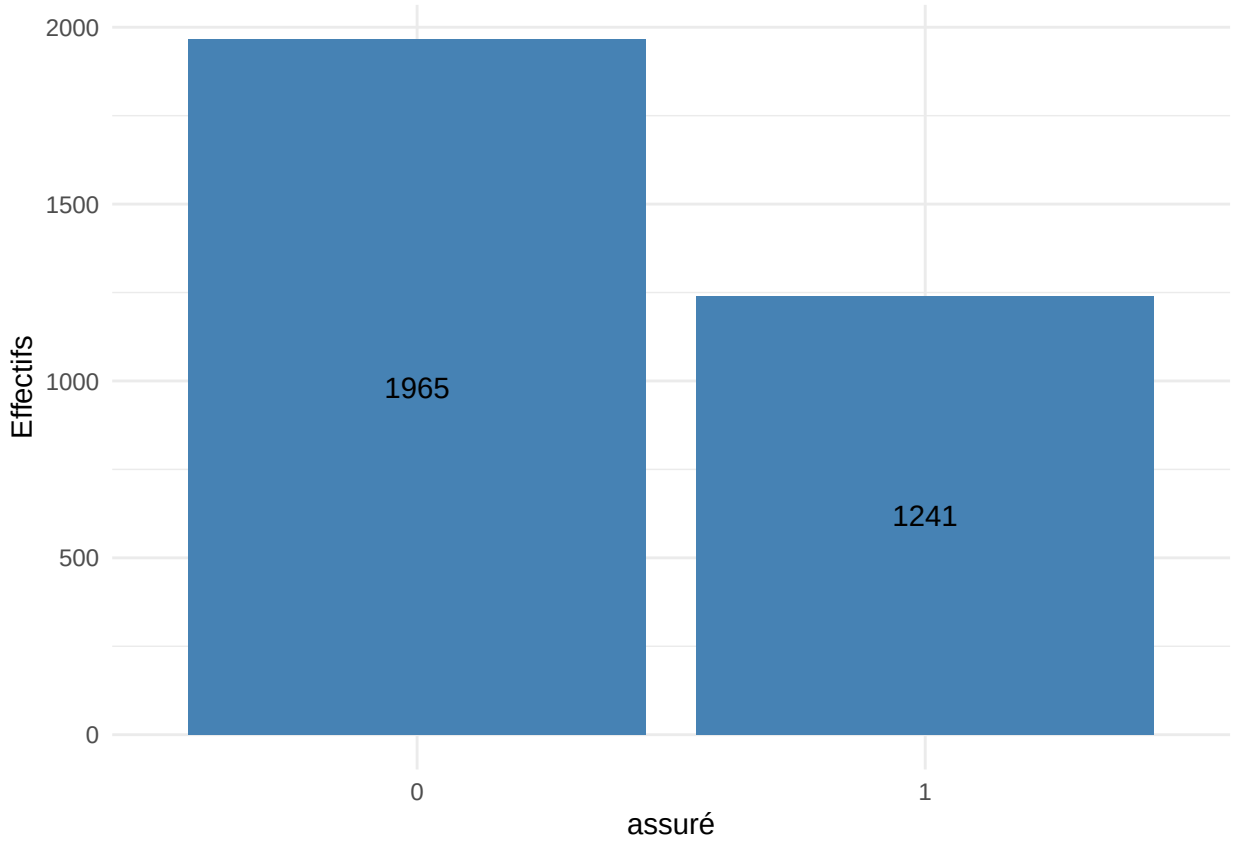
retraité:Si l'individu est retaité.

conjoint_retraité:Si le conjoint de l'individu est retraité.

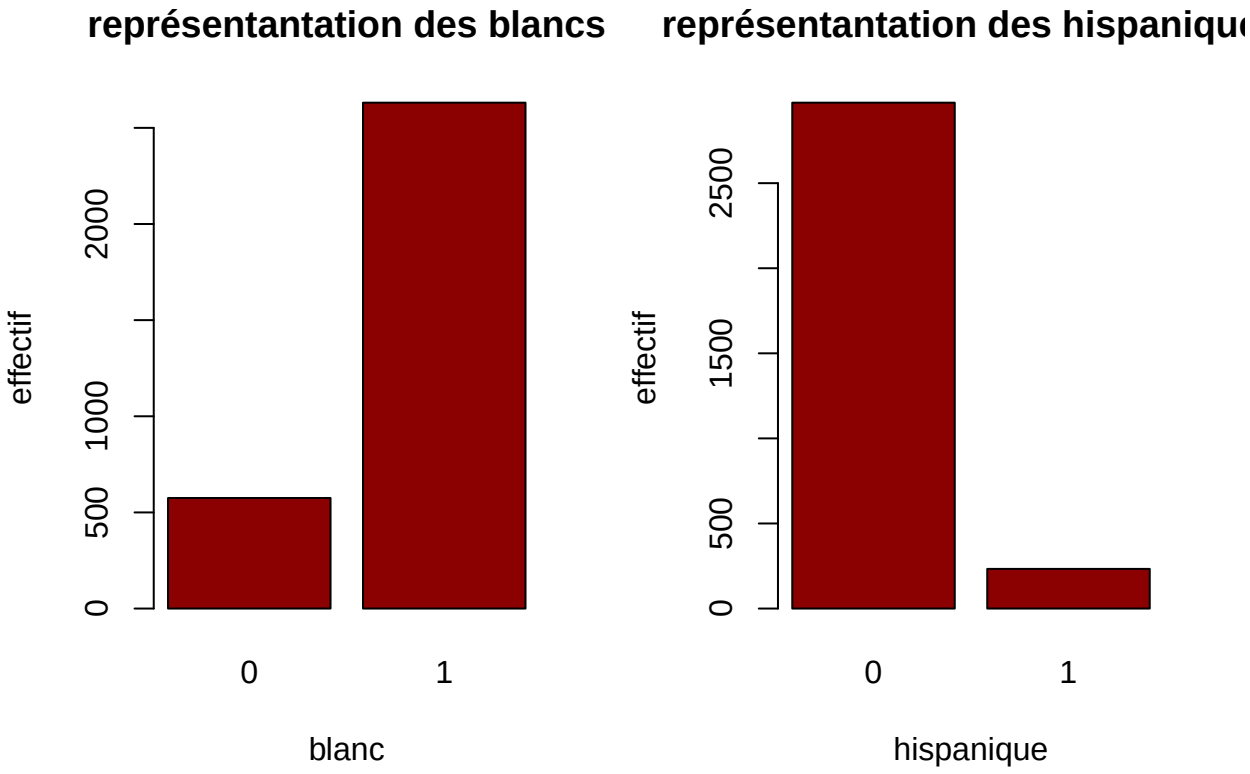
revenu:revenu de l'individu.

statut_santé:statut santé de l'individu.

— Notre variable à expliquer *assuré*:

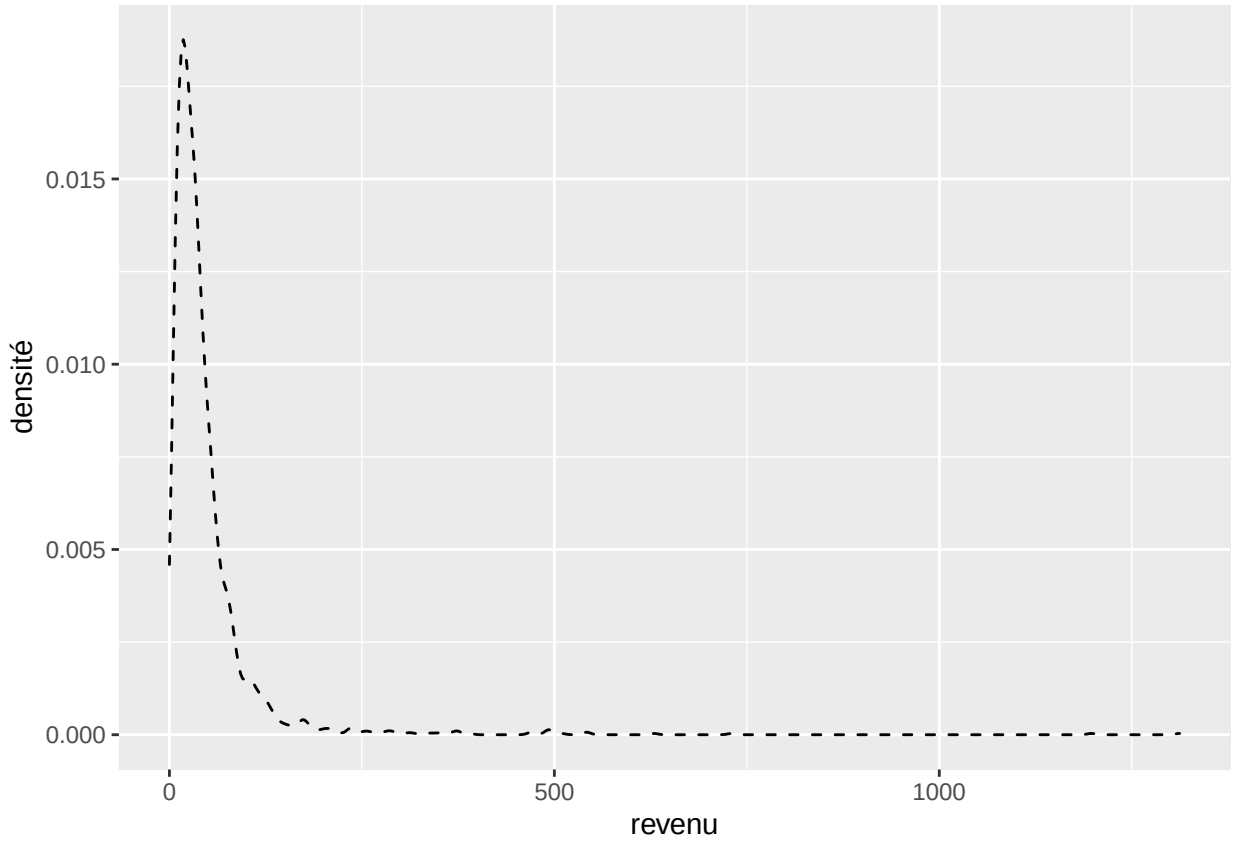


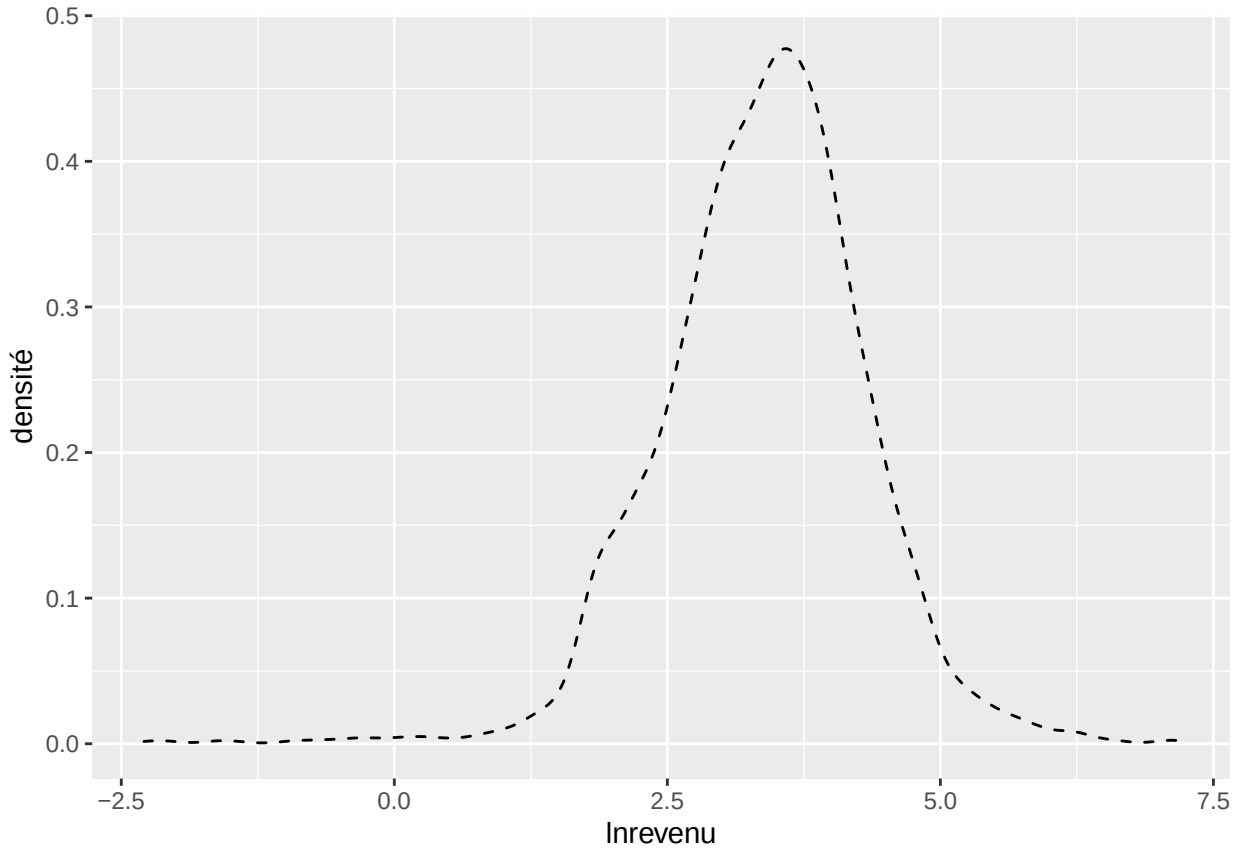
On constate que sur notre base de données il y'a environs 2000 indivus qui ne sont pas souscrit à une assurance complémentaire et environs 1250 qui sont souscrit à une assurance complémentaire.



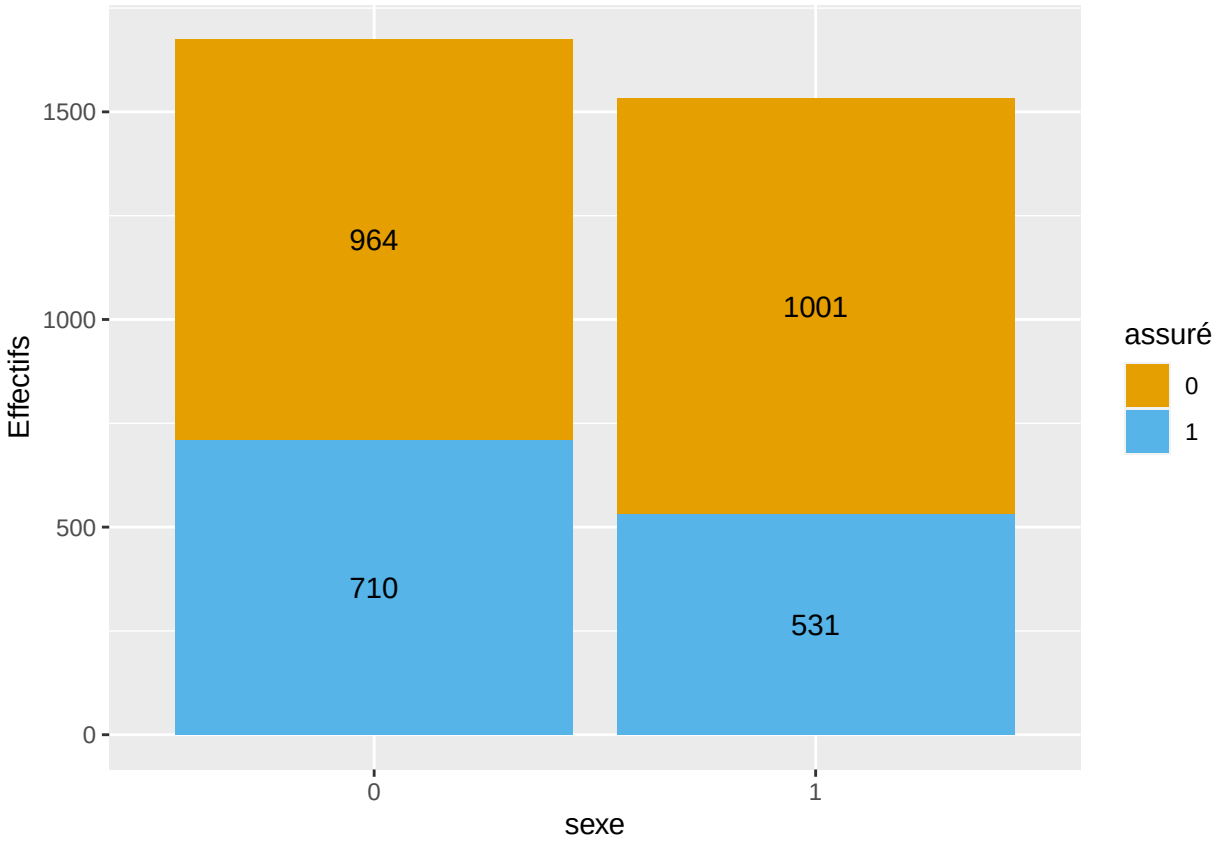
-

On constate qu'il y'a bien plus de blanc que d'hispanique dans notre base de données.

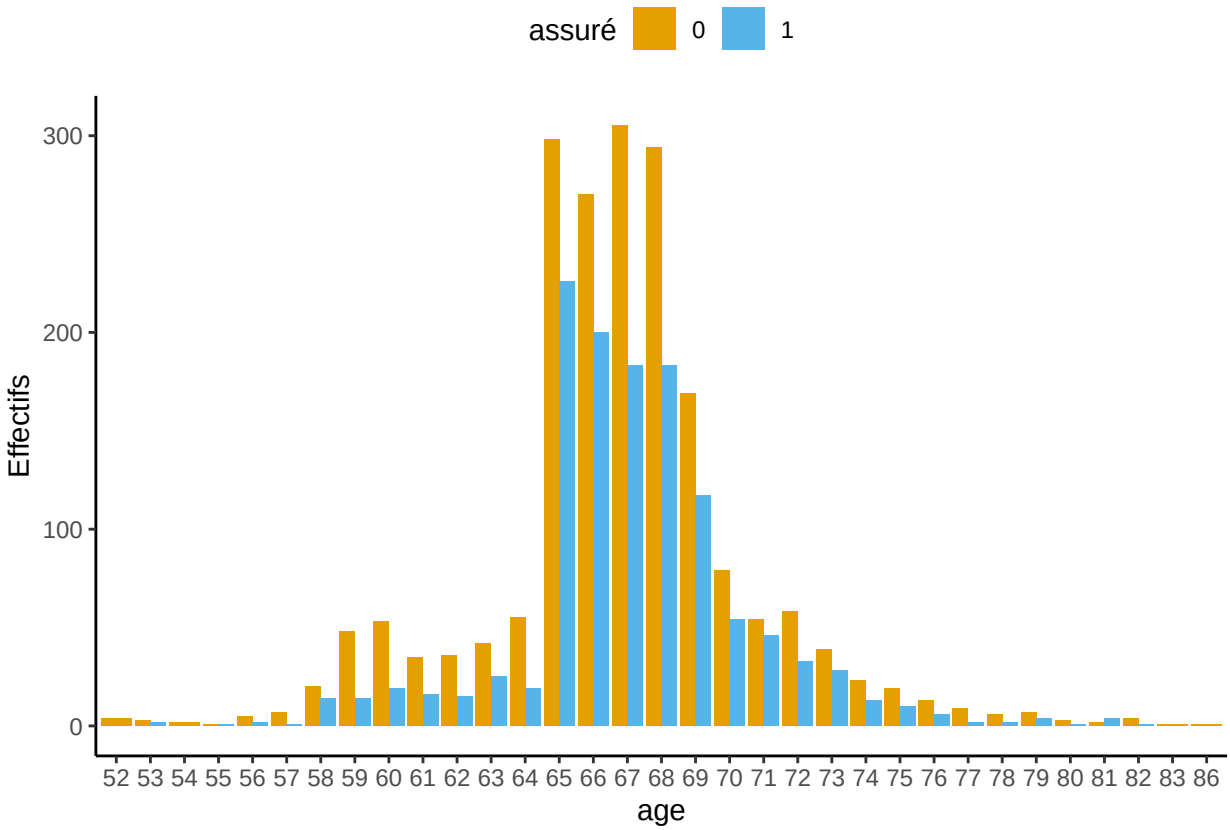




On contate que la majorité des individus ont des revenu reparti entre 0 et 250.Nous avons pas l'information pour savoir si les revenus sont donnés dans quel ordre d'échelle. Nous avons mis une representaion du revenu au logarithme pour montrer qu'il y'a un effet de lissage ce quin permet de mieux capter les changements.



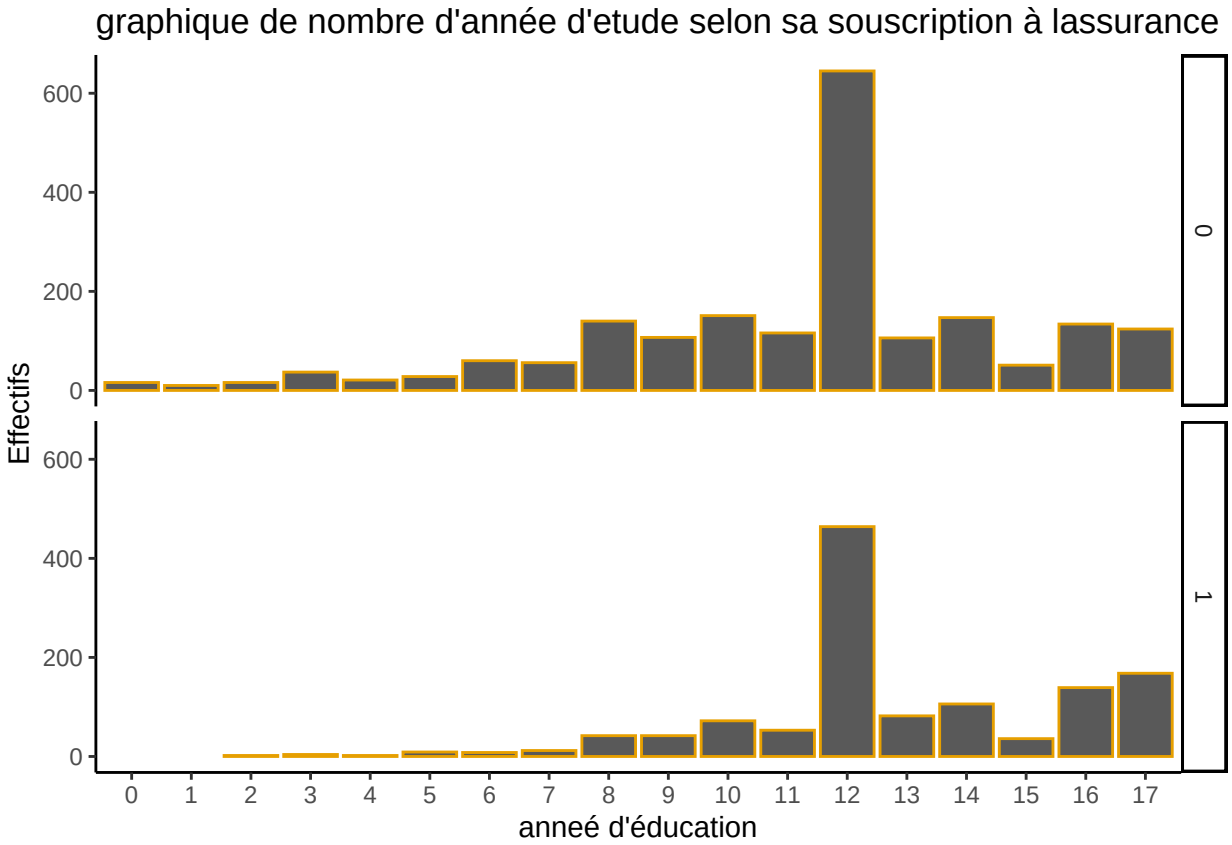
-À travers ce graphique on peut voir deux imformations.La premiere est qu'il y'a plus d'homme que de femme dans notre base de données et la deuxiéme et que chez les femmes comme chez les hommes il y' a plus d'individu qu n'ont pas de d'assurance complémentaire.



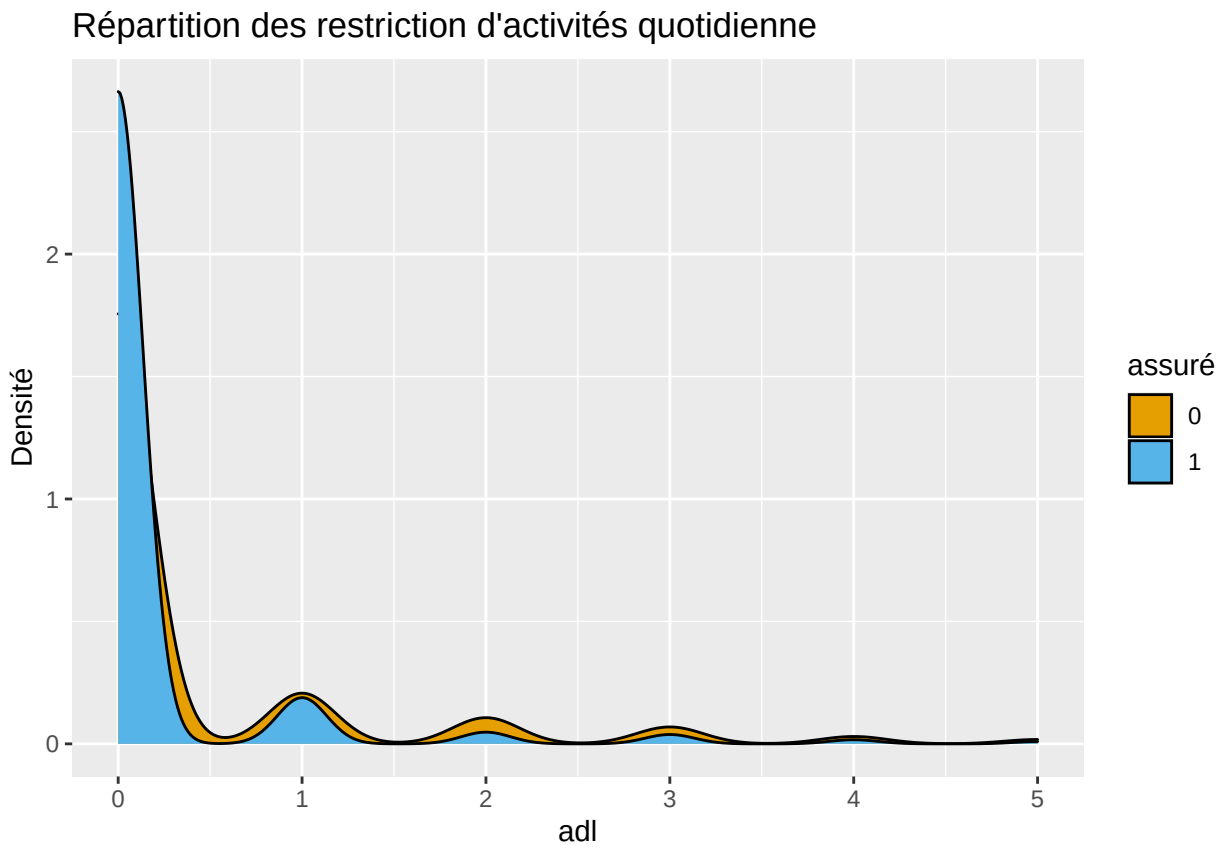
-

Comme on peut le voir sur ce graphique la tranche d'âge la plus representé dans notre base de données

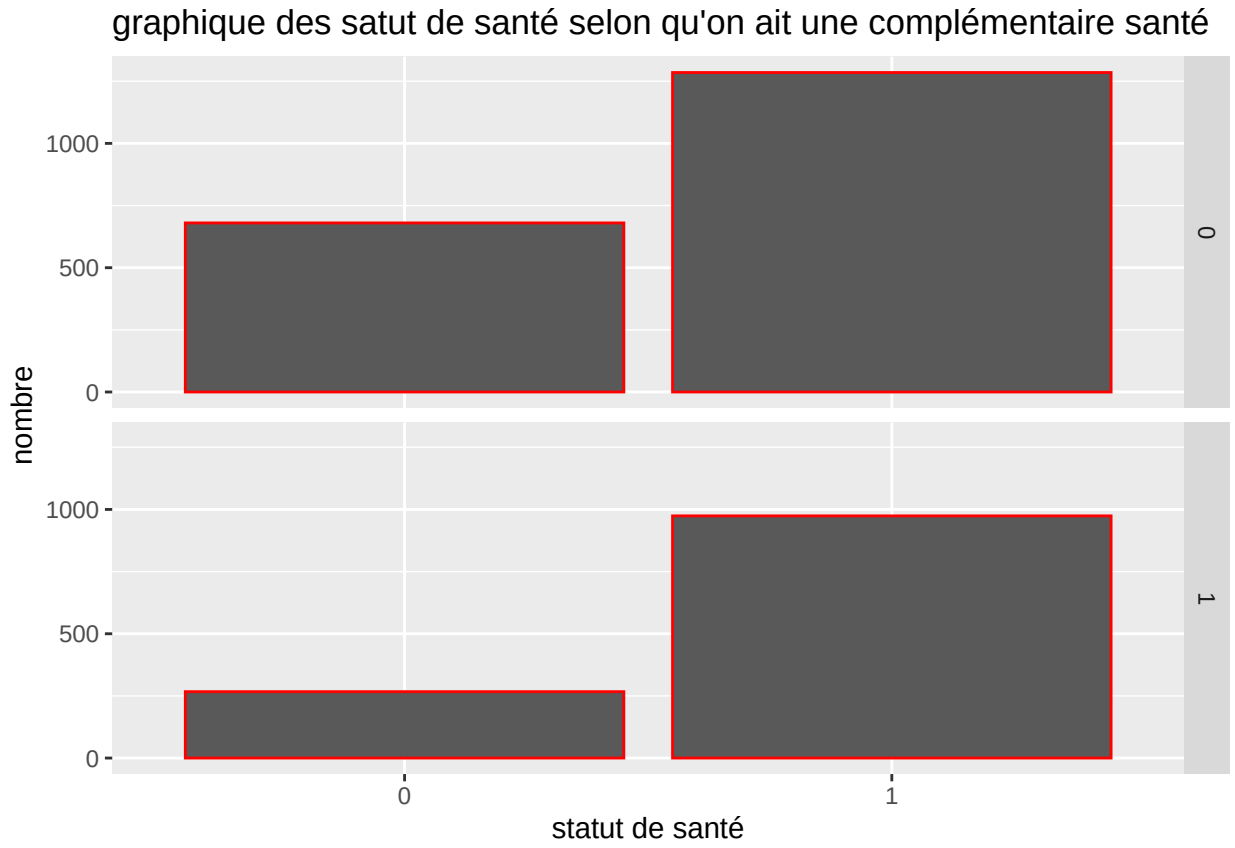
et la tranche d'age 65 72 ans environs.Et en majorité il y'a toujours plus de non souscrit à une assurance complémentaire que des souscrit



-Ce graphique nous montre que la majorité des individus non souscrit et souscrit ont effectué 12 ans d'étude.Et que la repartions d'individus souscrit à l'assurance commence à partir de 8 ans études.



-on voit assez bien que la plus part d'individus souscrit ou pas à une assurance complémentaire n'ont pas de restriction dans pour la limite d'activités dans la vie quotidenne et que plus il y'a de restricton moins la densité est grande.



-On remarque sur ce graphique que la majorité des personnes qui se souscrivent à une assurance complémentaire on un bonne santé et que c'est la même chose chez les non assuré.Donc on d'autre termes que la santé

n'a pas forcément une grande influence dans le fait d'avoir ou pas une assurance complémentaire,il faudra le vérifier.

2 Modéle général , Vraisemblance et Log-vraisemblance.

Le modéle probit et logit sont des modéle dichotomique par modéle dichotomique, on entend un modéle statistique dans lequel la variable expliquée ne peut prendre que deux modalités (variable dichotomique). Il s'agit alors généralement d'expliquer la survenue ou non d'un événement, ou d'un choix. On considére un échantillon de n individus d'indices i = 1, .., n. Pour chaque individu, on observe si un certain évènement s'est réalisé et l'on pose:

$Y_i = \begin{cases} 1 \ si \ l'evenement \ se \ réalise \ , Y^* \geq 0 \\ 0 \ sinon \ , Y^* \leq 0 \end{cases}$ ~~avec~~Yˆ{*}=$\beta X_i + \varepsilon_i$

— On utilise la méthode du maximum de vraisemblance pour estimer nos paramétres qui s'ecrit de façon suivante:

$L(\theta) = \prod_{i=1}^{N} F(X_i\theta)^{Y_i}(1 - F(X_i\theta))^{1-Y_i}$

— Et sa log-vraisemblance qui s'ecrit de façon suivante

$log(L(\theta)) = \sum_{i:Y_i=1} logF(X_i\theta) + \sum_{i:Y_i=1} log(1 - F(X_i\theta))$

— Elle permet d'obtenir les différents coefficients associés à nos modéles .

3 Estimation (probit et logit)

Nous allons ici vous présenter les Modéles probit et logit.

3.1 Probit

— Pour le modéle Probit on pause F ,qui est la fonction de répartition d'une gaussienne centrée réduite,usuellement notée :

F(X_{i} θ) = $\Phi(X\_\{i\}\theta) = \int$ {i=1}ˆ{X{i}θ}$\frac{\exp(-t\check{s}/2)}{\sqrt{2\pi}}dt$

— Et sa densité correspondante, usuellement notée , est :
f(X_{i} θ) = $\phi(X\_\{i\}\theta) = \frac{\exp -((X_i\theta)\check{s}/2)}{\sqrt{2\pi}}dt$
Logit

-Pour le modéle Logit on pause F ,qui est la fonction de répartition introduite spécialement pour ce type de modéle,usuellement notée :

F(X_{i} θ) = $\Lambda(X\_\{i\}\theta) = \frac{\exp(X_i\theta)}{1+\exp(X_i\theta)}dt = \frac{1}{1+\exp(-X_i\theta)}dt$

— Et sa densité correspondante, usuellement notée , est :

f(X_{i} θ) = $\lambda(X\_\{i\}\theta) = \frac{\exp(-X_i\theta)}{(1+\exp(-X_i\theta))\check{s}}dt = \Lambda(X\_\{i\}\theta)(1 - \Lambda(X\_\{i\}\theta))$

— Il n'y a pratiquement pas de différence entre ces deux lois, l'introduction de la loi logistique étant simplement plus simple en terme de calcul en général. La seule différence notable entre les deux modéles probit et logit vient tout simplement de la spécification de la fonction de répartition F.

4 Hypothése et spécification de test

Test de wald

Le test de Wald est un test paramétrique économétrique qui permet de tester la "vraie" valeur du paramètre basé sur l'estimation de l'échantillon.À savoir si nos paramétres sont siginificatifs ou pas.

le test s'écrit de façon suivante:

H0:$\beta_k = 0$,alors le paramétre est n'est pas significativement différent de 0
H1:$\beta_k \neq$,0 alors le paramétre est significativement différent de 0

la statistique:

$$W = \frac{\beta\check{s}_k}{S\check{s}_k}$$

Décision

On rejette H0 au risque α si W $\geq \chi\check{s}_{1-\alpha}(1)$

Test de spécification Hosmer-Lemeshow

Le test de spécification Hosmer-Lemeshow est un test statistique pour la qualité d'ajustement pour la régression logistique modèles, permet de tester si les proportions observées et attendues diffèrent de manière significative . soit G le nombre de groupe,les observations sont regroupées par probabilité attendue. les observations avec une probabilité attendue similaire sont regroupées dans le même groupe, pour créer 10 groupes. soit p_g la probabilité moyenne prédite dans le groupe g soit y_g la fréquence d'échantillonnage moyenne dans le groupe g.

la statistique:

$HL = \sum_{g=1}^{G} \frac{(p_g - y_g)\check{s}}{y_g(1-y_g)}$

Sous la valeur nulle de spécification correcte, la statistique est distribuée comme $\chi\check{s}(G-2)$.

Décision

Si la p-value est inférieur à l'alpha choisi alors l'hypothése nulle selon laquelle les proportions observées et attendues sont les mêmes est rejetée.

5 Comparaison des modéles

— Nous avons éffectué une premiére régression probit et logit avec tout nos variables pour avoir une vue d'ensemble sur nos variables et données et pour essayer de voir la significativité de chaque varible et on s'aperçoit que toutes les variables sauf assurance privée et la "constante" n'ont pas de coéfficient β associé ce qui suggére une certaine corrélation forte entre l'un de nos variables explicatives et notre variable à expliquer.

le modéle est le suivant:

$$assuré_i = \beta_0 + \beta_1\ assuranceprivé + \beta_2\ age + \beta_3\ hispanique + \beta_4\ blanc + \beta_5\ femme + \beta_6\ annéed'éducation + \beta_7\ marié + \beta_8\ excellentesanté + \beta_9\ trésbonnesanté + \beta_10\ bonnesanté + \beta_11\ santépassable + \beta_12\ mauvaisesanté + \beta_13\ maladiechronique + \beta_{14}\ adl + \beta_{15}\ retraité + \beta_{16}\ conjointretraité + \beta_{17}\ revenu + \beta_{18}\ statutsanté + \varepsilon_i$$

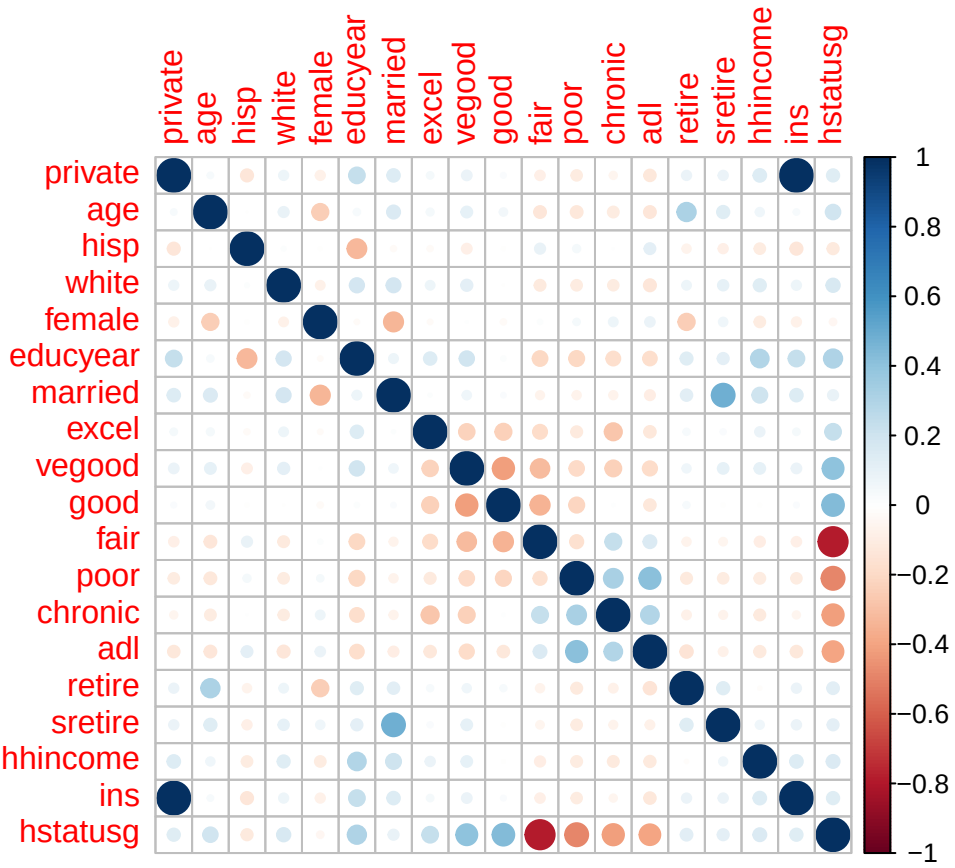
Et on s'aperçoit rapidement à travers ce graphique de matrice de corrélation que ceci est dû à la forte corrélation que nous avons entre la variable assurance privée et assuré.

De même nous constatons aussi des corrélation plus ou moins important entre d'autres variables comme:

— statut stanté et bonne santé
— statut stanté et mauvaise santé
— statut stanté et santé passable
— statut stanté et maladie chronique

TABLE 1 – regression logit et probit avec toute les variables

	Dependent variable:	
	assuré	
	logistic	*probit*
	(1)	(2)
assurance_privé	53.13213	13.98244
	(16,522.58000)	(3,432.36700)
age	−0.00000	0.00000
	(2,306.54900)	(479.15820)
hispanique	0.00000	0.00000
	(32,311.46000)	(6,712.32800)
blanc1	−0.00000	−0.00000
	(21,544.94000)	(4,475.70500)
femme	0.00000	−0.00000
	(17,951.37000)	(3,729.18400)
année_d_édu_ation	−0.00000	0.00000
	(2,774.17800)	(576.30260)
marié	−0.00000	−0.00000
	(22,660.49000)	(4,707.44900)
excellente_santé	−0.00000	−0.00000
	(42,117.78000)	(8,749.47100)
trés_bonne_santé	−0.00000	−0.00000
	(36,576.05000)	(7,598.24400)
bonne_santé	0.00000	0.00000
	(34,416.78000)	(7,149.68200)
santé_passable	−0.00000	−0.00000
	(33,561.38000)	(6,971.98200)
mauvaise_santé		
maladie_chronique	0.00000	0.00000
	(6,374.88200)	(1,324.30600)
adl	0.00000	0.00000
	(11,320.09000)	(2,351.61600)
retraité	0.00000	0.00000
	(17,451.86000)	(3,625.41600)
conjoint_retraité	−0.00000	−0.00000
	(19,145.87000)	(3,977.32600)
revenu	0.00000	0.00000
	(123.85110)	(25.72858)
statut_santé		
Constant	−26.56607	−6.99122
	(159,990.50000)	(33,236.13000)
Observations	2,138	2,138
Log Likelihood	−0.00000	−0.00000
Akaike Inf. Crit.	34.00000	34.00000
Note:		*p<0.1; **p<0.05; ***p<0.01



Pour la suite de notre étude, toute les estimations de nos modéles qui seront réalisées sans la variable la variable *"assurance privée"*.

5.1 Mise en place des différent modéle

— Le modéle 0 est le modéle suivant:

Dans ce modéle la variable "statut santé" a été retiré dû à sa colinéarité de plus nous avons remarquer que cette même variable est une variable qui traduit de façon global les informations liées aux cinq variables associées à l'état de santé.

— le modéle 0 que nous proposant est donc le suivant:

$$assuré_i = \beta_0 + \beta_1\ age_i + \beta_2\ hispanique_i + \beta_3\ blanc_i + \beta_4\ femme_i + \beta_5\ annéed'éducation_i + \beta_6\ marié_i + \beta_7\ excellentesanté_i + \beta_8\ trésbonnesanté_i + \beta_9\ bonnesanté_i + \beta_10\ santépassable_i + \beta_11\ mauvaisesanté_i + \beta_12\ maladiechronique_i + \beta_{13}\ adl_i + \beta_{14}\ retraité_i + \beta_{15}\ conjointretraité_i + \beta_{16}\ revenu_i + \varepsilon_i$$

— Pour l'éstimation de ce premier modéle nous avons enlever "excellente_santé","trés_bonne_santé","bonne_santé","santé et "mauvaise santé" dû à leur faible et moyenne corrélation avec la variable statut_santé,dont nous avons également pu rémarquer qu'elle traduisait les informations de ces cinq variables liées à la santé.

— Le modéle 1 est le modéle suivant:

$$assuré_i = \beta_0 + \beta_1\ age_i + \beta_2\ hispanique_i + \beta_3\ blanc_i + \beta_4\ femme_i + \beta_5\ annéed'éducation_i + \beta_6\ marié_i + \beta_7\ maladiechronique_i + \beta_8\ adl_i + \beta_9\ retraité_i + \beta_{10}\ conjointretraité_i + \beta_{11}\ revenu_i + \beta_{12}\ statutsanté_i + \varepsilon_i$$

— Le modéle 0 et 1 nous permettra de jauger et savoir s'il est plus pertinent de garder les 5 variables associé à l'état de santé ou simplement de garder la variable "statut sante" et de retirer les cinq autres.

— Pour l'éstimation de ce deuxiéme modéle nous avons éffectué un changement sur la variable age,que nous avons divisé en 3 classe d'age,une premiére classe de 52 à 64 ans une deuxiéme classe de 64 à 76 ans et une troisiéme clase de 76 à 86 ans.

— Le modéle 2 est le suivant:

$$assuré_i = \beta_0 + \beta_1\ age_i + \beta_2\ hispanique_i + \beta_3\ blanc_i + \beta_4\ femme_i + \beta_5\ annéed'éducation_i + \beta_6\ marié_i + \beta_7\ maladiechronique_i + \beta_8\ adl_i + \beta_9\ retraité_i + \beta_{10}\ conjointretraité_i + \beta_{11}\ revenu_i + \beta_{12}\ statutsanté_i + \varepsilon_i$$

```
## [1] 52
```

— Pour l'éstimation de ce troisiéme modéle nous avons appliquer le logarithme à la variable revenu pour baissé son effet d'echelle et ainsi mieux le comparer aux autres variables et ainsi pouvoir mieux capter les variations liéés à cette variable en la lissant.
— Le modéle 3 est le modéle suivant:

$$assuré_i = \beta_0 + \beta_1\ age_i + \beta_2\ hispanique_i + \beta_3\ blanc_i + \beta_4\ femme_i + \beta_5\ annéed'éducation_i + \beta_6\ marié_i + \beta_7\ maladiechronique_i + \beta_8\ adl_i + \beta_9\ retraité_i + \beta_{10}\ conjointretraité_i + \beta_{11}\ lnrevenu_i + \beta_{12}\ statutsanté_i + \varepsilon_i$$

Pour l'éstimation de ce quatriéme modele nous avons tous simplement mis au carré la variable age dans le but d'éssayer de capter au mieux les changement liés à cette variable.

— Le modéle 4 est le modéle suivant:

$$assuré_i = \beta_0 + \beta_1\ Ageš_i + \beta_2\ hispanique_i + \beta_3\ blanc_i + \beta_4\ femme_i + \beta_5\ annéed'éducation_i + \beta_6\ marié_i + \beta_7\ maladiechronique_i + \beta_8\ adl_i + \beta_9\ retraité_i + \beta_{10}\ conjointretraité_i + \beta_{11}\ revenu_i + \beta_{12}\ statutsanté_i + \varepsilon_i$$

Pour l'éstimation de ce cinquiéme modéle nous avons tout simplement combiné l"age au carré et le logarithme sur le revenu .

— Le modéle 5 est le modéle suivant:

$$assuré_i = \beta_0 + \beta_1\ Ageš_i + \beta_2\ hispanique_i + \beta_3\ blanc_i + \beta_4\ femme_i + \beta_5\ annéed'éducation_i + \beta_6\ marié_i + \beta_7\ maladiechronique_i + \beta_8\ adl_i + \beta_9\ retraité_i + \beta_{10}\ conjointretraité_i + \beta_{11}\ lnrevenu_i + \beta_{12}\ statutsanté_i + \varepsilon_i$$

— Pour l'estimation de sixiéme modéle nous avons effectué un changement sur la variable statut_santé,ainsi elle prend la valeur de:
1 si l'individu a une excellente santé
2 si l'individu a une trés bonne santé
3 si l'individu a une bonne santé
4 si l'individu a une santé passable
5 si l'individu a une mauvaise santé
— Le modéle 6 est le modéle suivant:

$$assuré_i = \beta_0 + \beta_1\ Ageš_i + \beta_2\ hispanique_i + \beta_3\ blanc_i + \beta_4\ femme_i + \beta_5\ annéed'éducation_i + \beta_6\ marié_i + \beta_7\ maladiechronique_i + \beta_8\ adl_i + \beta_9\ retraité_i + \beta_{10}\ conjointretraité_i + \beta_{11}\ lnrevenu_i + \beta_{12}\ statutsanté_2 + \beta_{13}\ statutsanté_3 + \beta_{14}\ statutsanté_4 + \beta_{15}\ statutsanté_5 + \varepsilon_i$$

— Pour l'estimation de septiéme modéle nous avons rajouté la variable age² et lnrevenu en plus des age et revenu que nous avions au départ.
— Le modéle 7 est le modéle suivant:

$$assuré_i = \beta_0 + \beta_1\ age_i + \beta_2\ hispanique_i + \beta_3\ blanc_i + \beta_4\ femme_i + \beta_5\ annéed'éducation_i + \beta_6\ marié_i + \beta_7\ maladiechronique_i + \beta_8\ adl_i + \beta_9\ retraité_i + \beta_{10}\ conjointretraité_i + \beta_{11}\ revenu_i + \beta_{12}\ Ageš_i + \beta_{13}\ lnrevenu_i + \beta_{12}\ statutsanté_2 + \beta_{13}\ statutsanté_3 + \beta_{14}\ statutsanté_4 + \beta_{15}\ statutsanté_5 + \varepsilon_i$$

— Aprés avoir mise en place nos sept modéles différent,nous allons sélectionner les modéles probits et logits qui présente les meilleurs critéres d'information et estimations sur les données d'entrainement.Ensuite nous allons essayer d'ameliorer nos différents modéles à travers plusieur procéder pour en tirer le meilleur modéle possible.

— Aprés avoir pu étudier le critère d'information d'Akaike (AIC,Akaike information criterion) qui est une mesure de la qualité d'un modèle statistique ainsi que le critère d'information bayésien (BIC,bayesian information criterion) dérivé du critère d'information d'Akaike.Nous avons sélectionné les modéles qui présente les plus faibles mesure d'AIC et de BIC combinés. Les modéles 5 et 7 sont les modéles que nous avons séléctionné pour la suite de l'étude que nous allons mener. À travers ces deux modéles nous allons comparer les modéles probit et logit. -Notre modéle 5 est le suivant

$$assuré_i = \beta_0 + \beta_1\ Ageš_i + \beta_2\ hispanique_i + \beta_3\ blanc_i + \beta_4\ femme_i + \beta_5\ annéed'éducation_i + \beta_6\ marié_i + \beta_7\ maladiechronique_i + \beta_8\ adl_i + \beta_9\ retraité_i + \beta_{10}\ conjointretraité_i + \beta_{11}\ lnrevenu_i + \beta_{12}\ statutsanté_i + \varepsilon_i$$

-Notre modéle 7 est le suivant:

$$assuré_i = \beta_0 + \beta_1\ age_i + \beta_2\ hispanique_i + \beta_3\ blanc_i + \beta_4\ femme_i + \beta_5\ annéed'éducation_i + \beta_6\ marié_i + \beta_7\ maladiechronique_i + \beta_8\ adl_i + \beta_9\ retraité_i + \beta_{10}\ conjointretraité_i + \beta_{11}\ revenu_i + \beta_{12}\ Ageš_i + \beta_{13}\ lnrevenu_i + \beta_{12}\ statutsanté_2 + \beta_{13}\ statutsanté_3 + \beta_{14}\ statutsanté_4 + \beta_{15}\ statutsanté_5 + \varepsilon_i$$

TABLE 2 – Comparaison des AIC et BIC modèles

	AIC	BIC
Modéle logit 0	2685.494	2776.176
Modéle probit 0	2684.614	2775.296
Modéle logit 1	2683.492	2757.172
Modéle probit 1	2682.314	2755.993
Modéle logit 2	2684.144	2780.478
Modéle probit 2	2683.200	2779.534
Modéle logit 3	2617.711	2680.055
Modéle probit 3	2614.359	2676.703
Modéle logit 4	2683.627	2751.639
Modéle probit 4	2682.368	2750.380
Modéle logit 5	2617.647	2685.658
Modéle probit 5	2614.251	2682.262
Modéle logit 6	2685.494	2776.176
Modéle probit 6	2684.614	2775.296
Modéle logit 7	2600.871	2702.888
Modéle probit 7	2598.469	2700.487

TABLE 3 – regression logit et probit modéle 5

	Dependent variable:	
	assuré	
	logistic	*probit*
	(1)	(2)
Age^2	−0.00010	−0.00007
	(0.00011)	(0.00006)
hispanique	−0.72104***	−0.42336***
	(0.25360)	(0.14262)
blanc	−0.18087	−0.10826
	(0.13610)	(0.08212)
femme	−0.02093	−0.01639
	(0.10910)	(0.06653)
année_d_édu_ation	0.09560***	0.05872***
	(0.01847)	(0.01106)
marié	0.09670	0.06192
	(0.15041)	(0.09087)
retraité	0.15061	0.09228
	(0.10684)	(0.06489)
conjoint_retraité	0.01514	0.00650
	(0.11474)	(0.07021)
lnrevenu	0.62068***	0.37752***
	(0.07495)	(0.04462)
statut_santé	−0.02650	−0.01050
	(0.12408)	(0.07505)
adl	−0.13845*	−0.08416*
	(0.07572)	(0.04428)
Constant	−3.22173***	−1.96702***
	(0.54526)	(0.32671)
Observations	2,138	2,138
Log Likelihood	−1,296.82300	−1,295.12500
Akaike Inf. Crit.	2,617.64700	2,614.25100
Note:	*p<0.1; **p<0.05; ***p<0.01	

La différence que nous remarquons en premier entre les deux modéles est que les estimation β associées aux variables explicatives du modéle probit sont toujours plus faibles que les estimation β associées aux variables explicatives du modéle logit.Cette différence découle probablement de la différence de leur fonction de répartition F et de densité f.

De plus nous avons aussi remarqué que les AIC et BIC du modéle probit sont toujours un peu plus faible que celle du logit en général.

— Les deux tableau qui vont suivre nous donnes la valeurs des odds ration associés à nosvariables explicatives ainsi que leurs intervalles de confiance qui peut nous renseigner en amont surla significativité d'un

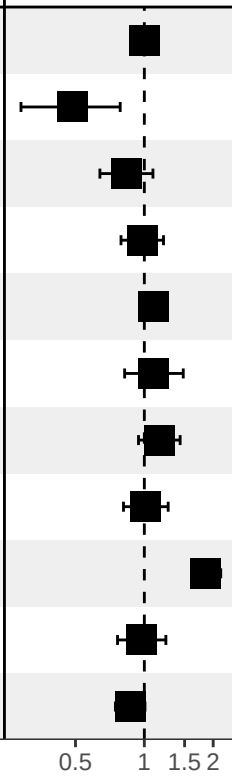
TABLE 4 – regression logit et probit modéle 7

	Dependent variable:	
	assuré	
	logistic	*probit*
	(1)	(2)
age	0.20674	0.12900
	(0.25113)	(0.14949)
blanc	−0.21510	−0.13011
	(0.13863)	(0.08326)
hispanique	−0.65206**	−0.38452***
	(0.25638)	(0.14470)
femme	−0.03859	−0.03085
	(0.11025)	(0.06714)
année_d_édu_ation	0.09531***	0.05908***
	(0.01881)	(0.01125)
marié	−0.01276	−0.00259
	(0.15447)	(0.09325)
maladie_chronique	0.08897**	0.05342**
	(0.04022)	(0.02431)
adl	−0.12234	−0.07535
	(0.08002)	(0.04691)
retraité	0.10647	0.06558
	(0.10815)	(0.06566)
conjoint_retraité	0.00092	0.00027
	(0.11557)	(0.07065)
revenu	−0.00499***	−0.00275***
	(0.00124)	(0.00069)
statut_santé2	0.19908	0.11887
	(0.16316)	(0.10002)
statut_santé3	0.11026	0.06950
	(0.16884)	(0.10317)
statut_santé4	0.17574	0.10032
	(0.20041)	(0.12170)
statut_santé5	−0.09505	−0.05467
	(0.27723)	(0.16571)
age^2	−0.00162	−0.00101
	(0.00185)	(0.00110)
lnrevenu	0.96857***	0.56906***
	(0.11395)	(0.06559)
Constant	−11.40848	−6.98772
	(8.51336)	(5.06577)
Observations	2,138	2,138
Log Likelihood	−1,282.43500	−1,281.23500
Akaike Inf. Crit.	2,600.87000	2,598.46900
Note:	*p<0.1; **p<0.05; ***p<0.01	

coefficient.Un peut plus tard dans l'etude nous expliquerons le rôle des odds ration et qu'elle imformations nous pouvons en tirer de ces odds ratio qui sont généralement compris entre 0 et $+\infty$

\begin{center} **Tableau des odds ratio et intervalles de confiance pour le modéle logit 5**

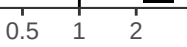
Variable	N	Odds ratio	p
Age²	2138	1.00 (1.00, 1.00)	0.329
hispanique	2138	0.49 (0.29, 0.78)	0.004
blanc	2138	0.83 (0.64, 1.09)	0.184
femme	2138	0.98 (0.79, 1.21)	0.848
année_d_édu_ation	2138	1.10 (1.06, 1.14)	<0.001
marié	2138	1.10 (0.82, 1.48)	0.520
retraité	2138	1.16 (0.94, 1.43)	0.159
conjoint_retraité	2138	1.02 (0.81, 1.27)	0.895
lnrevenu	2138	1.86 (1.61, 2.16)	<0.001
statut_santé	2138	0.97 (0.76, 1.24)	0.831
adl	2138	0.87 (0.75, 1.01)	0.067



\end{center}

\begin{center} **Tableau des odds ratio et intervalles de confiance pour le modéle logit 7**

Variable		N	Odds ratio		p
age		2138		1.23 (0.76, 2.04)	0.41
blanc		2138		0.81 (0.61, 1.06)	0.12
hispanique		2138		0.52 (0.31, 0.85)	0.01
femme		2138		0.96 (0.78, 1.19)	0.73
année_d_édu_ation		2138		1.10 (1.06, 1.14)	<0.001
marié		2138		0.99 (0.73, 1.34)	0.93
maladie_chronique		2138		1.09 (1.01, 1.18)	0.03
adl		2138		0.88 (0.75, 1.03)	0.13
retraité		2138		1.11 (0.90, 1.38)	0.32
conjoint_retraité		2138		1.00 (0.80, 1.26)	0.99
revenu		2138		1.00 (0.99, 1.00)	<0.001
statut_santé	1	241		Reference	
	2	621		1.22 (0.89, 1.68)	0.22
	3	678		1.12 (0.80, 1.56)	0.51
	4	420		1.19 (0.81, 1.77)	0.38
	5	178		0.91 (0.53, 1.56)	0.73
age²		2138		1.00 (0.99, 1.00)	0.38
lnrevenu		2138		2.63 (2.12, 3.31)	<0.001



\end{center}

6 Prédiction et fitted

— Pour la suite de l'étude ,aprés avoir sélectionner nos modéles 3 et 5 nous avons effectués une selection du meilleur modéle à partir de ces deux modéle de départ à travers une sélection par AIC pour déterminer le modéle 5 et 7 "finaux" puis nous allons les comparer.

Le modéle 5 "final" est le suivant:

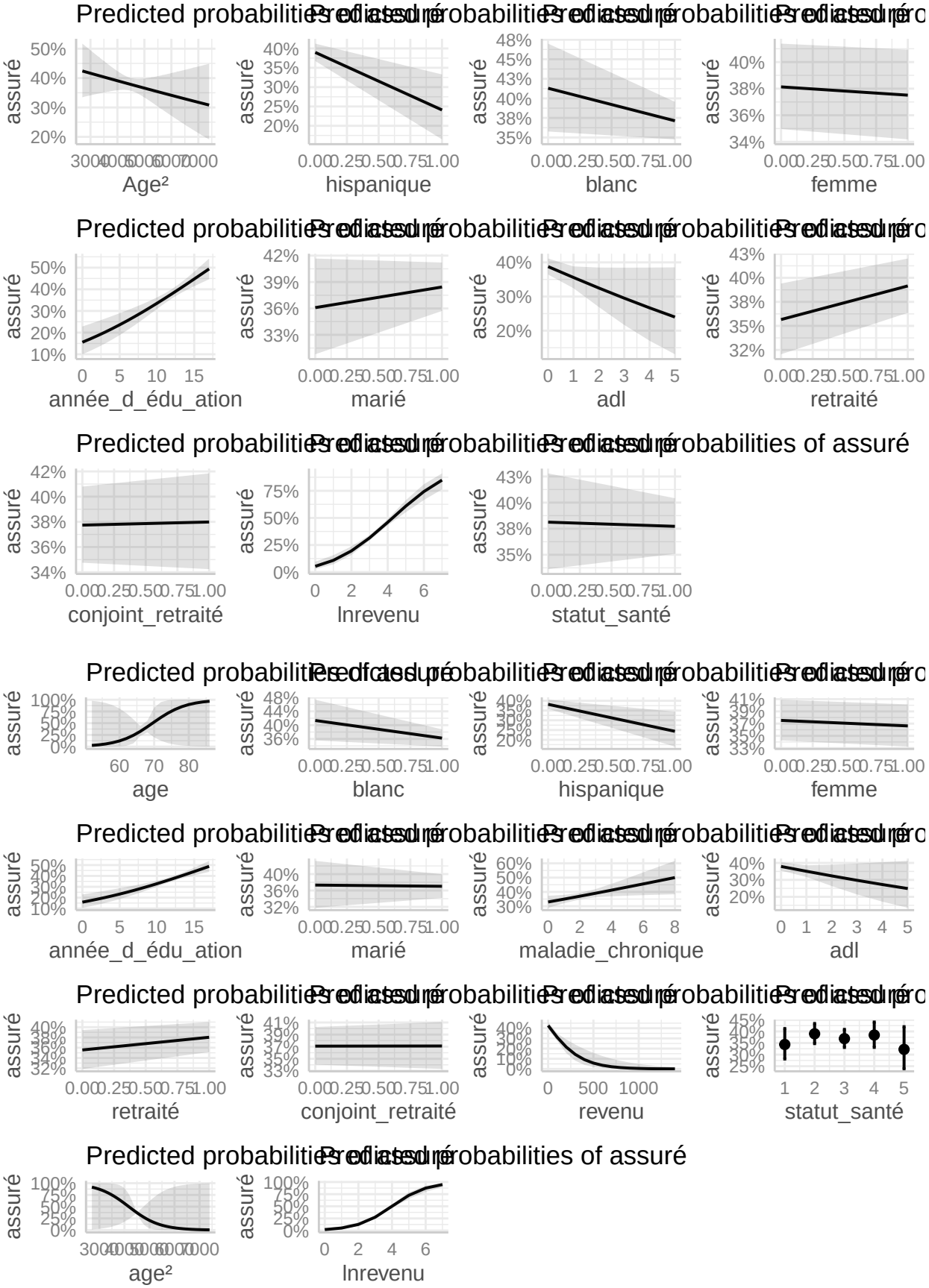
$$assuré_i = \beta_0 + \beta_1\ hispanique_i + \beta_2\ annéed'éducation_i + \beta_3\ adl_i + \beta_4\ lnrevenu_i + \varepsilon_i$$

Le modéle 7 "final" est le suivant:

$$assuré_i = \beta_0 + \beta_1\ hispanique_i + \beta_2\ blanc_i + \beta_3\ annéed'éducation_i + \beta_4\ maladiechronique_i + \beta_5\ adl_i + \beta_6\ revenu_i + \beta_7\ lnrevenu_i + \varepsilon_i$$

— Ces graphique ci-dessous nous montre visuellement la relation qu'à notre variables à expliquer "assuré" avec nos différent variables explicatives,à savoir si elle ont réellement une influence sur celle-ci.

GRAPHIQUE SUR LES LIENS ENTRE LA VARIABLE À EXPLIQUER ET LES VARIABLES EXPLICATIVES



6.1 training

TABLEAU DE CONTINGENCE SUR LES DONNÉE D'APPRENTISSAGE

	Modéle logit 5	
	0	1
0	1023	502
1	268	345

	Modéle logit 5 final	
	0	1
0	1025	514
1	266	333

	Modéle probit 5	
	0	1
0	1025	508
1	266	339

	Modéle probit 5 final	
	0	1
0	1027	518
1	264	329

	Modéle logit 7	
	0	1
0	994	452
1	297	395

	Modéle logit 7 final	
	0	1
0	998	453
1	293	394

	Modéle probit 7	
	0	1
0	1000	456
1	291	391

	Modéle probit 7 final	
	0	1
0	1004	458
1	287	389

TABLE 5 – Taux d'erreur sur les données d'apprentissage

	Taux d'érreur
Modéle logit 5	0.3503274
Modéle logit 5 final	0.3489242
Modéle probit 5	0.3493920
Modéle probit 5 final	0.3484565
Modéle logit 7	0.3601497
Modéle logit 7 final	0.3648269
Modéle probit 7	0.3620206
Modéle probit 7 final	0.3657624

6.2 TEST

TABLEAU DE CONTINGENCE SUR LES DONNÉE TEST

	Modéle logit 5	
	0	1
0	541	255
1	133	139

	Modéle logit 5 final	
	0	1
0	540	254
1	134	140

	Modéle probit 5	
	0	1
0	542	258
1	132	136

	Modéle probit 5 final	
	0	1
0	543	255
1	131	139

	Modéle logit 7	
	0	1
0	523	227
1	151	167

	Modéle logit 7 final	
	0	1
0	532	225
1	142	169

	Modéle probit 7	
	0	1
0	526	227
1	148	167

	Modéle probit 7 final	
	0	1
0	534	229
1	140	165

Table 6 – Taux d'erreur sur les données test

	Taux d'érreur
Modéle logit 5	0.3632959
Modéle logit 5 final	0.3632959
Modéle probit 5	0.3651685
Modéle probit 5 final	0.3614232
Modéle logit 7	0.3539326
Modéle logit 7 final	0.3436330
Modéle probit 7	0.3511236
Modéle probit 7 final	0.3455056

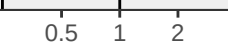
D'aprés le calcul des taux d'erreur les estimation du modéle logit 7 "final" et les estimations du modéle probit 7 "final" semble être les meilleurs tout en notant que l'erreur sur le modéle logit 7 final est faiblement plus petite comparer à l'erreur sur le modéle 7 probit final.Pour la suite nous nous focaliserons sur le modéle logit 7 final.

7 Odds ratio ,test et effet marginal.

7.1 ODDS-RATIO

— Bien que l'odds-ratio est d'une valeur qui ne s'interprete pas aisément néanmoins elle donne quelques informations globales.L'odds ratio s'interprète comme le risque relatif à la réalisation d'un évenement.Si l'odds ration est inférieur à un alors on a moins de "chance" à la survenue d'un évenement par rapport à une autre.Si l'odds ration est supérieur à un alors on a plus de "chance" à la survenue d'un évenement par rapport à une autre.Si elle est egale à zéro alors les "chances" des deux évenement sont les même.Pour faire simple dans notre cas l'odds ratio nous montre si la variable explicative à un effet positif (de chance) ou négatif(malchance) sur le fait d'être assuré.

Variable	N	Odds ratio	p
blanc	2138	0.81 (0.62, 1.06)	0.122
hispanique	2138	0.51 (0.30, 0.82)	0.008
année_d_édu_ation	2138	1.10 (1.06, 1.14)	<0.001
maladie_chronique	2138	1.09 (1.02, 1.17)	0.017
adl	2138	0.85 (0.73, 0.98)	0.031
revenu	2138	0.99 (0.99, 1.00)	<0.001
lnrevenu	2138	2.64 (2.17, 3.25)	<0.001



Ce tableau nous donne dans un premier temps la valeur de l'odds ratio associée à chaque variables explcatives.On peut aussi voir l'intervalle de confiance associé à chaque odds ration,ce qui peut nous donner une information sur la significatvité de l'odds ratio.

-D'aprés ce tableau ,on peut déja observer que les odds ratio associés aux variables "blanc","hispanique","adl" et "revenu" sont inférieur à 1,ce qui nous suggére que ces variables là ont un effet dit"négatif" sur le fait d'être assuré. Et que les variables "année d'éducation", "maladie chronique" et "lnrevenue" ont un effet positif sur l'évenement être assuré. De plus on remarque la valeur assez élevée du odds ration lié à la variable "lnrevenu"

7.2 Test

7.2.1 Test de wald

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 68.694, df = 6, P(> X2) = 7.5728e-13
```

Pour notre test on rejette l'hypothèse H0 si la p-value est inférieure à 5%. Ici notre p.value vaut 7.5728e-13 donc au rique de 5% on rejette HO et on accepte l'hypothése que nos coéfficient sont différent de 0.

7.2.2 Test de spécification Hosmer-Lemeshow

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  regl7_l$data$assuré, fitted(regl7_l)
## X-squared = 28.976, df = 8, p-value = 0.0003202
```

```
##
##  Hosmer and Lemeshow test (binary model)
##
## data:  regl7_l$data$assuré, fitted(regl7_l)
## X-squared = 28.976, df = 8, p-value = 0.0003202
```

Notre modéle à une p-value inférieur à 5% ce qui suggére que nous avons une erreur d'étalonnage globale au risque de 5%.

7.3 Effet marginal

```
## Call:
## logitmfx(formula = regl7_l, data = test_data7, atmean = TRUE)
##
## Marginal Effects:
##                          dF/dx   Std. Err.       z     P>|z|
## blanc                0.01994744  0.04161988  0.4793   0.63174
## hispanique          -0.10991305  0.05913847 -1.8586   0.06309 .
## année_d_édu_ation    0.00261051  0.00564049  0.4628   0.64350
## maladie_chronique    0.00999343  0.01142885  0.8744   0.38190
## adl                 -0.04253592  0.02228103 -1.9091   0.05625 .
## revenu              -0.00338054  0.00072055 -4.6916 2.711e-06 ***
## lnrevenu             0.34308394  0.04210969  8.1474 3.719e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## dF/dx is for discrete change for the following variables:
##
## [1] "blanc"      "hispanique"
```

-L'effet marginal de la variable blanc est de 0.02 , cela equivaut à dire que si un individu est blanc alors la probabilité qu'il ait une assurance complémentaire augmente de 2 point. -L'effet marginal correspondant à la variable hispanique vaut -0.11,ceci nous dit que si une personne est d'hispanique alors la probabilité qu'il ait une assurance complémentaire baisse de 10 point.

-L'effet marginal associé à la variable année d'éducation vaut 0.002,on peut donc dire qu'en moyenneune variation positif de 1% sur l'année d'étude augmente la probababilité d'être souscrit à une complémentaire santé augmente de 0.2 point.

-L'effet marginal associé à la variable maladie chronique vaut 0.010,on peut donc dire qu'en moyenne une variation positif de 1% sur la variable maladie chronique augmente la probabilité d'avoir une souscrition à une assurance complémentaire de 1 point.

-L'effet marginal correspondant à la variable adl vaut -0.04 ce qui indique qu'en moyenne une variation positif de 1% sur la variable adl augmente la probabilité d'avoir une souscription à une assurance complémentaire de 4 point.

-L'effet marginal correspondant à la variable revenu vaut -0.003 ce qui nous emméne à dire qu'en moyenne une variation positif de 1% sur le revenu baisse la probabilité d'avoir une souscription à l'assurance complémentaire de 0.3 point.

-L'effet marginal correspondant à notre variable lnrevenu vaut 0.34,ce qui veut dire qu'en une variation positif de 1% sur le revenu au logarithme augmente la probabilité d'avoir une souscription à l'assurance complémentaire de 34 point.

8 Interprétation

— On peut interpreter l'effet marginaux positf sur les blancs et négatif sur les hispaniques en disant que l'étude a été faite aux états-unis ,où il y'a une différence entre les blancs et les hispaniques qui viennent de l'amérique du sud genéralement clandestinements .Les blancs trés majoritaire occupent des bons postes tandis que les immigrés hispaniques ont des postes moins bien remunérés donc n'ont pas forcément le moyen de se payer une souscription à une assurance complémentaire contrairement aux indivudus blancs qui dans la moyenne le peuvent.

-Les variables année d'éducation et maladie chronique ont tout les deux des éffets marginaux positifs qui peut s'expliquer par le fait que plus on étudie plus on a un revenu et donc se permettre une complémentaire santé.Et généralement quand on est sujet à revenr continuellement dans un hopital comme quand on a une maladie chronique on opte généralement pour une complémentaire santé.

-Pour les variables revenu et lnrevenu on a effet marginaux qui s'oppose,on va tenter de l'expliquer. D'abord il est nécécessaire de dire que la majorité des individus avait des revenu pas trés élevé ,donc de ce fait on peut dire que n'ayant pas forcément des revenus illimités les personnes à faible revenu préferent tout simplement se souscrire à une complémentaire santé et ainsi en cas de problémes médicaux il ne débourseront pas leurs argents.Alors qu'une personne aisé elle n'a pas forcement ce besoin là elle peut elle même se payer ses séjours à l'hôpital et ne pas opter pour une assurance complémentaire .Donc si le revenu augmente ils vont de moins en moins souscrire une assurance complémentaire.

Et d'un autre coté on peut se dire que justement une personne qui n'est pas assez riche et/ou qui as souvent des problémes médicaux va préférer garder son argent pour ses loisirs,son alimentation ou autre probléme alors qu'une personne moyennement riche ou riche va elle se prevenir de tout risque d'autant plus qu'on a vu que les années d'education(plus on étudie mieux est le salaire) avait un effet positif, donc cela vont opter pour une assurance santé,d'autant plus qu'on sait qu'aux états-uns les assurances santé coûtent particuliérement chére.Et ces donc ce sont ces individus là qu'arrive à capter la variables lnrevenu même s'ils ne sont pas nombreux.

9 Discussion et Limite

Dans notre études ,je pense que l'environnement de l'étude est déjà un peu affiné car ici on observe des individus bien précis de plus de 52 ans qui sont déjà dans des hopitaux et ayant des suivis.Parce que ce qui nous frappe en premier c'est le fait que la variable statut_santé ou encore les variables associées à la santé ne soient aucunement significatif alors que le but d'une assurance complémentaire c'est justement d'alléger les coût liés à la santé.Pour moi d'un point de vue personnel ça me semble un peu bizarre et je n'ai su trouver le pourquoi et c'est peut-là la limite de l'étude.

Les variables mis en avance ne sont pour la plus part non liées à la santé.On pourrait alors se dire que la santé n'a pas vraiment une influence sur le fait qu'une personne opte ou pas pour une souscription à une assurance complémentaire. Pour se faire,il faudrait aussi peut-être faire une étude sur l'assurance santé auquelle les individus ce sont souscrire et ce qu'elle couvre.