Devoir_marketing

Sabaye Fried-Junior

Table des matières

Préambule

Contexte et problématique

Medicare est un programme d'assurance maladie qui aide les personnes âgées à payer les services et les soins liés à leurs santés. Il est financé en partie par le gouvernement americain. Pour être admissible à **Medicare** il faut remplir certaines conditions : être âgé(e) de plus de 65 ans si on a pas de handicap, être sous dialyse ou avoir reçu une transplantation, être atteint de la maladie de Charcot etc. . .

Medicare couvre une multitude de services mais sous certaines conditions, telles que : les soins hospitaliers non ambulatoires, les soins palliatifs, les établissements de soins infirmiers spécialisés, les médicaments sur ordonnance et les divers soins à domicile.

Cependant **Medicare** ne paie pas la totalité des coûts. Un montant limite est fixé soit à l'ensemble des services, soit à des services en particulier. Au-dessus de ce montant les concernés payent eux même la différence.

C'est pourquoi certains souscrivent à une assurance complémentaire, soit dans la même structure : **Medigap**, soit dans des sociétés privées et ce, afin de payer les services qui ne sont pas couverts par **Medicare** et de les aider à couvrir les divers coûts qui ne sont pas couverts. Aussi pour les aider à payer les soins de longues durées car ces derniers ne sont pas couverts par **Medicare**.

Notons que cette assurance complémentaire est parfois fournie par le dernier employeur dans le cadre d'une prestation de retraite. Les personnes qui ont un faible revenu et qui remplissent certains critères peuvent avoir droit à une *pseudo* couverture complémentaire par le régime **Medicaid**, lui aussi financé par le gouvernement.

Tous les individus de notre base de données bénéficient du programme **Medicare**, certains d'entre eux ont souscrit à la complémentaire santé : **Medigap** en plus de leur assurance et d'autres non.

L'objectif de notre étude sera d'essayer de déterminer les variables et/ou les combinaisons de variables qui expliquent ces faits.

Pour ce faire nous allons commencer par une analyse descriptive, puis nous construiront divers modèles basés sur la régression logistique, puis nous choisirons le meilleur d'entre eux et enfin, nous discuterons des résultats ainsi obtenus.

1 Analyse descriptive

1.1 Visualisation des données

private	age	hisp	white	female	educyear	married	excel	vegood	good	fair	poor	chronic	adl	retire	sretire	hhincome	ins	hstatusg
0	62	0	0	1	12	0	0	0	0	1	0	3	0	0	0	0.000	0	0
0	59	0	1	1	12	0	0	0	0	1	0	1	3	0	0	0.000	0	0
0	60	0	0	0	13	0	0	1	0	0	0	2	0	1	0	0.000	0	1
0	62	0	1	1	10	0	0	0	0	1	0	4	3	0	0	0.000	0	0
0	54	0	1	1	9	0	0	0	0	0	1	6	0	0	0	0.000	0	0
0	62	0	1	1	12	1	0	1	0	0	0	0	0	1	1	0.000	0	1
0	59	0	0	0	5	1	0	0	0	0	1	4	0	0	0	0.000	0	0
0	59	0	1	1	11	0	0	0	0	0	1	2	2	0	0	0.000	0	0
0	65	0	0	0	14	0	0	0	0	0	1	2	3	0	0	0.000	0	0
0	58	0	1	1	12	0	0	0	0	0	1	3	1	0	0	0.101	0	0

Ce tableau ne contient que les 10 première lignes de notre base de données.

1.2 Description des variables

La base de données renseigne sur 20 informations corncernant 3206 individus.
Ces dernières peuvent etre séparé comme suit :

- **Variables renseignant sur l'état de santé** :
  - *Excellent* : prend la valeur 1 si l'individu est en exxcellente santé et 0 sinon;
  - *vegood* : prend la valeur 1 si l'individu est en très bonne santé et 0 sinon;
  - *good* : prend la valeur 1 si l'individu est en bonne santé et 0 sinon;
  - *fair* : prend la valeur 1 si l'individu est plutot en bonne santé et 0 sinon;
  - *hstatusg* : variable qui renseigne sur l'autoévaluation de son état de santé, elleprend la valeur 1 si l'individu se pense en bonne santé et 0 sinon;
- **Les autres renseignant sur la signalitique des individus**
  - *hisp* : renseigne sur la race de la personne concerné, prend la valeur 1 si la personne est hispanique et à sinon;
  - *white* : renseigne sur la race de la personne concerné, prend la valeur 1 si la personne est blanche et à sinon;
  - *married* : prend la valeur 1 si la personne est marrié et à sinon;
  - *female* : prend la valeur 1 si l'individu est une femme et à si il s'agit d'un homme;
  - *retire* : prend la valeur 1 si l'individu est retraité et 0 sinon;
  - *sretire* : variable qui renseigne sur le statut de retraite du conjoint, prend la valeur 1 si le conjoint est assuré et sinon;
  - *poor* : prend la valeur 1 si la personne est pauvre et 0 sinon;
- **Les variables quantitatives et qualitatives ordonées**
  - *age* : renseigne sur l'âge des personnes, qu'elles soient assurés ou non;
  - *educyear* : renseignent sur le nombre d'années d'éducation des personnes;
  - *hhincome* : renseignent sur le revenus des ménages;
  - *chronic* : renseigne sur le nombre total de maladies chroniques; il s'agit d'une variable quantitatives

1.3 Analyse des variables

• Commencons par étudier les variables relatives à la **signalitique des individus**: *female, white, hisp, married, retire, sretire* et *poor* .

— **Le sexe** :

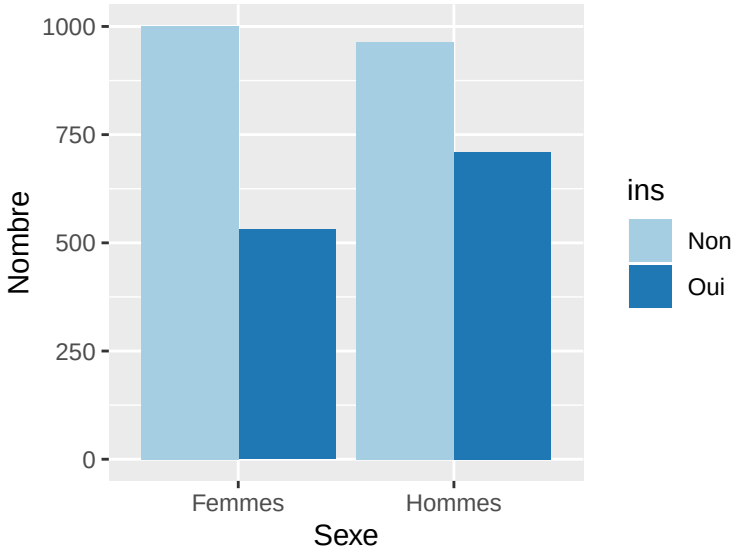


Table 1 – Statistiques concernant la race

Situation	Genre	
	Femmes	Hommes
Ayant une assurance complémentaire	531	710
Sans assurance complémentaire	1001	964

Le nombre d'hommes et de femmes participant à l'étude n'est pas significativement different. Il y'a au total : 1532 femmes contre 1674 hommes. Qu'il s'agissent des hommes ou des femmes, le nombre d'individu n'ayant pas d'assurance complémentaire est supérieur à celui qui en a une : 65,3% des femmes et 57,6% des hommes n'ont pas de complémentaire santé.

— **La race**:

Pour une meilleure compréhension visuelle et statistique, on fusionne les variables *hisp* et *white* pour mieux les representer.

Notons que certains individus sont à la fois Blanc et hispanic. On les appellera "Blanc et Hisp". Les "autres" sont les individus qui ne sont ni Blanc ni hispanic.

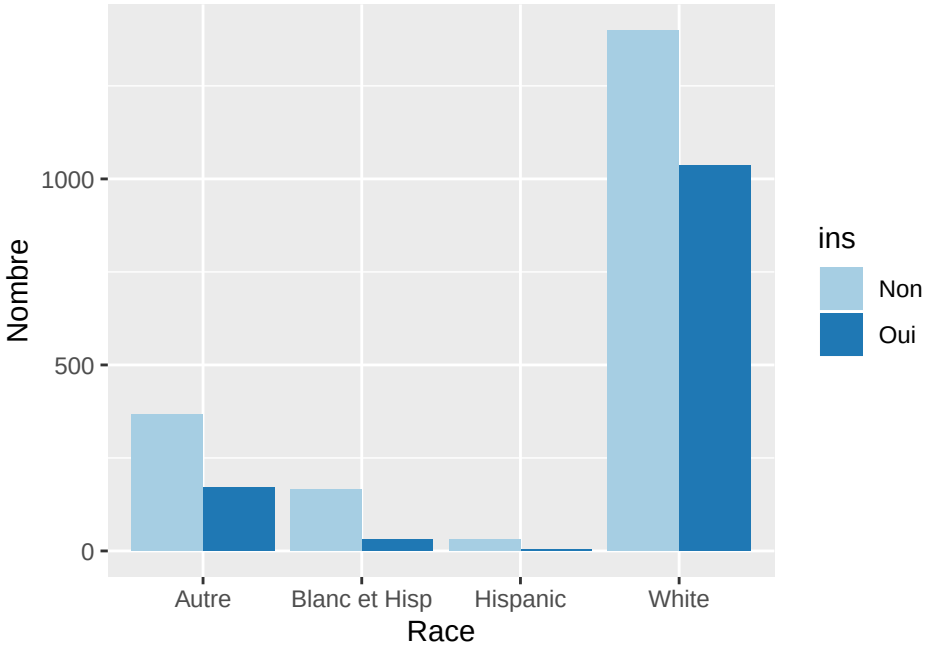


TABLE 2 – Statistiques concernant le sexe

Situation	Race			
	White	Autre	Hispanic	Blanc_et_Hisp
Ayant une assurance complémentaire	1036	170	5	30
Sans assurance complémentaire	1399	368	32	166

On constate que la majorité écrasante des individus enquétés sont de race blanche. Soit 75% du total des individus enquété. 57% d'entre eux n'ont pas de complémentaire santés.

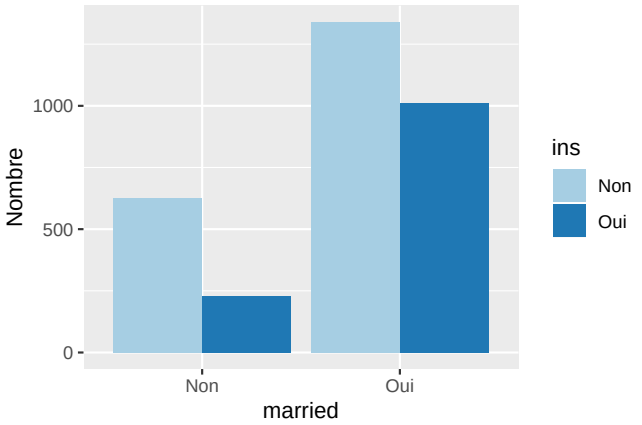
On note aussi, a l'inverse, que l' effectif des hispanic participant à l'enquète est très faible. Ils ne représente que 1,15% du total des individus enquétés et l'immense majorité d'entre eux n'a pas de complémentaire santé.

Cependant on note qu'un nombre plus important est à la fois hispanic et blanc. Eux aussi, pour la plupart, n'ont pas de complémentaire santé

Enfin, les autres, sont eux eux aussi, pour la plupart, n'ont pas de complémentaire santé.

• Intéressons nous maintenant aux variables qui renseignent sur les situations dites **matrimoniale et socio-professionnelle** :

On s'interesse aux variables : *married et retire*. Qui pour la premiere renseigne sur la situation matrimoniale de l'individu : si il (elle) est marrié(e) ou non, la seconde sur la situation socio-professionnelle de l'individu : si il (elle) est acitif(ve) ou à la retraite.



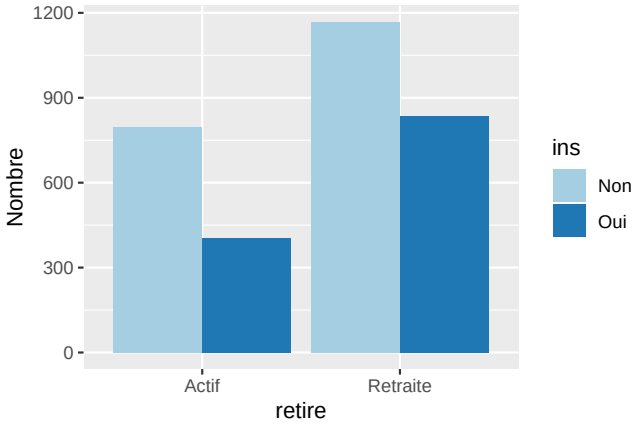


TABLE 3 – Statistiques concernant les divers situations

	Situation			
	Matrimoniale		Socio-professionnelle	
Situation	Marié	Célibataire	Retraité	Actif
Ayant une assurance complémentaire	1011	230	836	405
Sans assurance complémentaire	1339	626	1167	798

Grossièrement, on constate que la majorité des individus enquétés sont mariés et à la retraite.

En ce qui concerne la situation matrimoniale : 73% des individus enquétés sont marriés et parmi ces derniers, 57% n'ont pas de complémentaire santé.

Pour ce qui est de la situation socio-profésionnelle : 62,5% des personnes enquétés sont à la retraite et 58% d'entre eux n'ont pas de complémentaire santé.

• Passons aux variables qui renseigne sur **l'état de santé** :

Les variables concernés étant : *excellent, verygood, good, fair et poor.*

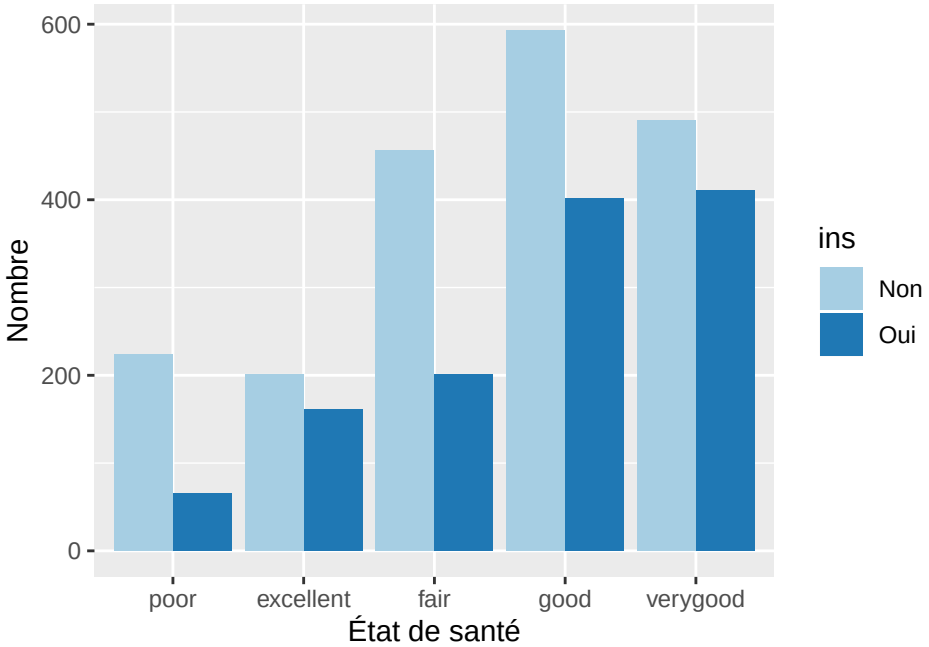


TABLE 4 – Statistiques concernant l'état de santé

Situation	État de santé				
	Excellent	Very_good	Good	Fair	poor
Ayant une assurance complémentaire	161	411	402	201	66
Sans assurance complémentaire	201	491	593	456	224

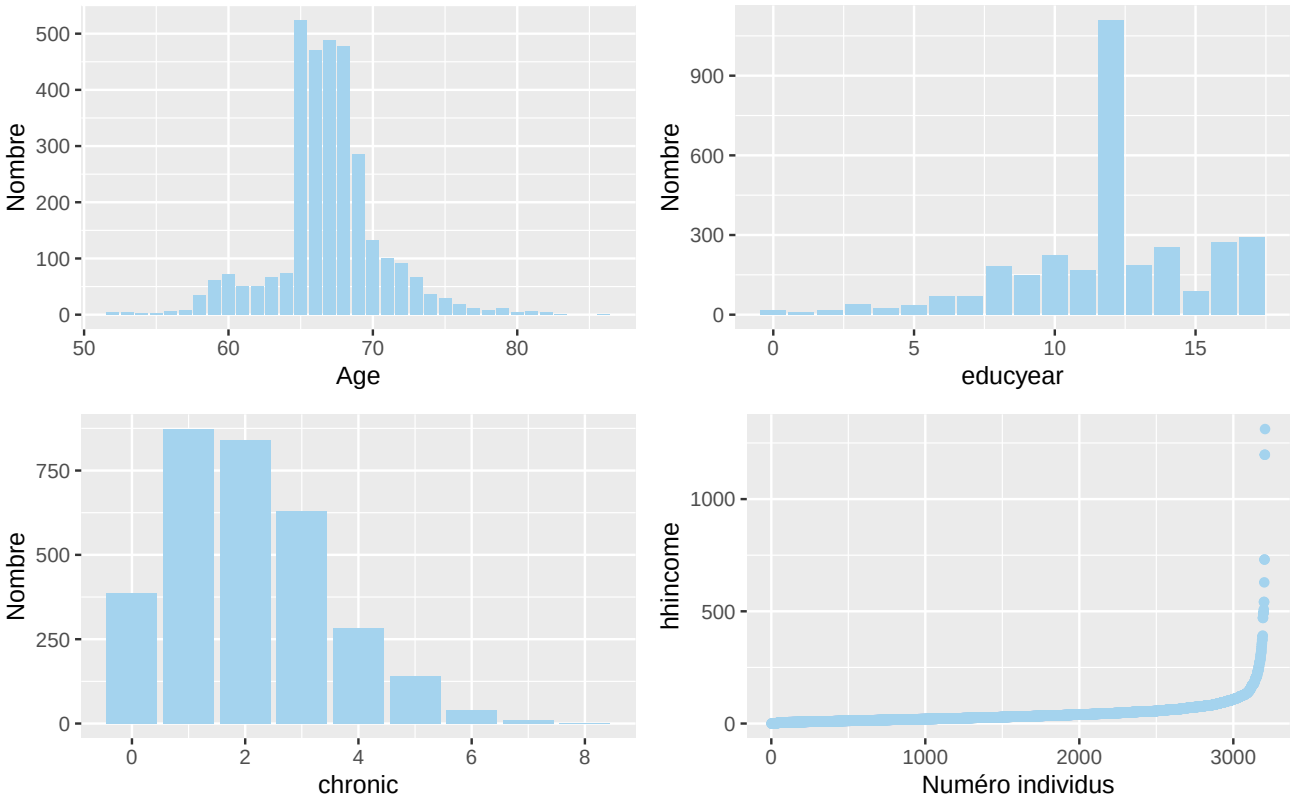
La plupart des individus se considèrent comme étant en bonne santé et n'ont pas de complémentaire santé

Seul 9% des individus enquétés ne se considèrent pas comme étant en bonne santé.

Il serait intéressant de comparer ces auto-évaluations au nombre de maladies chroniques ou simplement à l'évaluation faites par un médecin. Cette comparaison sera faite dans la suite de notre étude.

• Intéressons nous maintenant aux **variables quantitatives et qualitatives ordinales** :

Ces variables sont : *age* (l'âge des individus), *educyear*(nombre d'années d'éducation), *hhincome*(le revenu des individus) et *chronic*(nombre de maladies chroniques).



La moyenne d'âge est de 66 ans, la plupart des personnes enquétées ont un âge compris entre 65 et 69 ans. L'écart type est de 3,6, il est relativement faible.

	N	Mean	Sd.	Var.	Min	Q1	Q3	Max
Age	3206	66.913911	3.675794	13.511463	52	65.000	69.0	86.000
educyear	3206	11.898628	3.304611	10.920454	0	10.000	14.0	17.000
hhincome	3206	45.263914	64.339364	4139.553762	0	17.001	52.8	1312.124
chronic	3206	2.063319	1.416434	2.006286	0	1.000	3.0	8.000

Les personnes enquetées ont, pour la plupart, un niveau d'étude superieur à 10 ans, avec un pic significatif à 12 ans.

Il est difficile d'appréhender la variable *hhincome*. En effet on ne sait pas si il s'agit d'un indice, d'un nombre de point ou même d'une valeur financière. On note cependant une très grande variabilité. La moyenne n'a de ce fait, aucun sens.

Enfin, en ce qui concerne le nombre de maladie chroniques, les personnes enquetées ont pour la plupart des maladies chroniques. La majeure partie d'entre eux en ont entre une et trois. La moyenne étant de 2 Seul un nombre très réduit de personnes ont plus de 6 maladies chroniques.

2 Modèle général et maximum de vraissemblance

2.1 Généralités

Les modèles dichotomiques probit et logit admettent pour variable expliquée, non pas un codage quantitatif associé à la réalisation d'un évènement comme dans le cas de la spécification linéaire, mais la probabilité d'apparition de cet évènement, conditionnellement aux variables exogènes. notre variable à expliquer étant *ins* on à:

$$ins = \begin{cases} 1 & \text{avec la probabilité P}i \\ 0 & \text{avec la probabilité 1-P}i \end{cases}$$

Ainsi, on considère le modèle suivant :

$$p_i = Prob(y_i = 1|x_i) = F(x_i\beta)$$

dont la fonction de densité est :

$$f(Y_i|X_i) = P^{Y_i}(1-P)^{1-Y_i}$$

Avec : $Y_i = 0, 1$ et $P_i = F(x_i\beta)$

Maximum de vraissemblance:

Pour un individu i sachant ses caractéristiques, la vraissemblance associée est :

$$lnL(Y_i, \beta) = Y_i ln[F(X\beta)] + (1 - Y_i) ln[1 - F(X\beta)]$$

Pour tout N :

$$lnL(Y_i, \beta) = \sum_{i=1}^{N} \{Y_i ln[F(X\beta)] + (1 - Y_i) ln[1 - F(X\beta)]\}$$

La maximisation de la quantité $lnL(Y_i, \beta)$ par le biais de Newton-Raphson nous permet de converger vers es solutions plausibles à notre problème.

Estimations:

Le modèle logit général s'écrit :

$$Y = \mathbf{P}(Y = \frac{1}{\{X_j\}}) + \epsilon = \pi(\{X_i\}) + \epsilon = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}} + \epsilon$$

Le modèle probit général s'écrit :

$$F(X_i\theta) = \Phi(X_i\theta) = \int_{-\infty}^{X_i\theta} \frac{e^{-t^2/2}}{\sqrt{2\pi}} \, \mathrm{d}x$$

où F est la fonction de répartition d'un gaussienne centrée réduite, usuellement notée Φ.

avec

$$\epsilon = 1 - \pi(X) \quad si \quad Y = 1 \tag{1}$$
$$\epsilon = -\pi(X) \quad si \quad Y = 0 \tag{2}$$

2.2 Echantillonnage : Apprentissage vs test

Dans le but d'évaluer les differents modèles que nous allons estimer, nous allons séparer la base de données de deux.

On effectue un tirage aléatoire sans remise et on décide que 90% des données, soit 2885 individus, seront affectées à l'échantillon d'apprentissage, à partir duquel seront construit les modèles.

Les 10% restant, soit 321 individus, seront affectés à l'échantillon test. Ce dernier sera utilisé pour tester les modèles.

2.3 Modèle général

Le modèle général est estimé sur toutes les variables explicatives de notre base de données sans aucune transformation préalable. Il s'ecrtit :

$$Ln(\frac{P_i}{1-P_i}) = \beta_1\ private_i + \beta_2\ age_i + \beta_3\ hisp_i + \beta_4\ white_i + \beta_5\ educyear_i + \beta_6\ married_i + \beta_7\ excel_i + \beta_8\ vegood_i + \beta_9\ good_i + \beta_{10}\ fair_i + \beta_{11}\ poor_i + \beta_{12}\ chronic_i + \beta_{13}\ adl_i + \beta_{14}\ retire_i + \beta_{15}\ sretire_i + \beta_{16}\ hhincome_i + \beta_{17}\ hstatusg_i$$

Remarque: Une première régréssion nous a permise d'observer que : parmi les variables présentent dans la base de données certaines sont colinéaires et renseignent sur les mêmes informations. C'est le cas notamment de la variable *private* (assurance complémentaire privé) qui renseigne sur les mêmes informations que la variable à expliquée *ins*. Les variables renseignant sur l'état de santé : *excel, good, vegood, fair et poor* sont résumé dans la variable *hstatug* et sont donc colinéaires avec cette dernière. Elles ne seront donc pas incluses, elles non plus, dans la régression du modèle général.

L'équation du modèle général qui sera estimée est donc :

$$Ln(\frac{P_i}{1-P_i}) = \beta_1\ age_i + \beta_2\ hisp_i + \beta_3\ white_i + \beta_4\ educyear_i + \beta_5\ married_i + \beta_6\ chronic_i + \beta_7\ adl_i + \beta_8\ retire_i + \beta_9\ sretire_i + \beta_{10}\ hhincome_i + \beta_{11}\ hstatusg_i$$

3 Estimation

Les résultats de ces régressions sont renseignés dans le tableau suivant :

TABLE 5 – Résultats modèles générale

	Dependent variable:	
	ins	
	logistic	*probit*
	(1)	(2)
age	−0.020	−0.012
	(0.013)	(0.008)
hisp1	−0.715***	−0.432***
	(0.217)	(0.123)
white1	−0.048	−0.025
	(0.118)	(0.071)
female1	−0.081	−0.054
	(0.095)	(0.058)
educyear	0.077***	0.048***
	(0.016)	(0.010)
married1	0.168	0.102
	(0.129)	(0.078)
chronic1	0.265*	0.161*
	(0.142)	(0.086)
chronic2	0.386***	0.236***
	(0.144)	(0.088)
chronic3	0.567***	0.342***
	(0.156)	(0.095)
chronic4	0.429**	0.258**
	(0.196)	(0.119)
chronic5	0.440*	0.265*
	(0.254)	(0.152)
chronic6	0.163	0.126
	(0.444)	(0.257)
chronic7	0.617	0.363
	(0.776)	(0.464)
chronic8	1.204	0.738
	(1.587)	(0.947)
adl1	−0.213	−0.133
	(0.163)	(0.098)
adl2	−0.679**	−0.405**
	(0.283)	(0.162)
adl3	−0.324	−0.194
	(0.322)	(0.188)
adl4	−0.733	−0.448
	(0.562)	(0.315)
adl5	−0.197	−0.133
	(0.589)	(0.344)
retire1	0.210**	0.124**
	(0.094)	(0.057)
sretire1	−0.025	−0.014
	(0.100)	(0.061)
hhincome	0.628***	0.377***
	(0.066)	(0.039)
hstatusg1	0.155	0.100
	(0.113)	(0.068)
Constant	−2.783***	−1.695***
	(0.874)	(0.525)
Observations	2,885	2,885
Log Likelihood	−1,721.060	−1,718.456
Akaike Inf. Crit.	3,490.119	3,484.913
Note:	*p<0.1; **p<0.05; ***p<0.01	

4 Autres modèles

Nous allons maintenant estimer différents modèles et réaliser diverses régréssions. Pour chacun d'eux, les résultats des régréssions probit et logit seront présentées.

La comparaion des modèles logit et probit, ainsi que les interprétations des differents coefficients ne seront faites que sur le modèle final retenu.

4.1 Second modèle

Nous allons, en nous basant sur le modèle général, effectuer une sélection des variables et ce pour plusieurs raisons, la principale étant que : un modèle avec peu de variables sera plus facilement généralisable en termes de robustesse, c'est le *Principe du rasoir d'Occam.*

Pour ce faire nous allons faire une sélection automatique des variables du modèle général, sur le critère d'Akaïke (AIC). Ce dernier s'écrit comme suit: $AIC = 2k - 2\ln(L)$; où k est le nombre de paramètres à estimer du modèle et L est le maximum de la fonction de vraisemblance du modèle.

Si l'on considère un ensemble de modèles candidats, le modèle choisi est celui qui aura la plus faible valeur d'AIC. On à :

Table 6 – Résultats modèle AIC sur modèle général

	Dependent variable:	
	ins	
	logistic	*probit*
	(1)	(2)
hisp1	−0.749***	−0.442***
	(0.215)	(0.121)
educyear	0.077***	0.048***
	(0.016)	(0.009)
married1	0.161	0.103
	(0.107)	(0.065)
retire1	0.228***	0.136***
	(0.087)	(0.053)
hhincome	0.646***	0.389***
	(0.064)	(0.038)
Constant	−3.898***	−2.382***
	(0.231)	(0.134)
Observations	2,885	2,885
Log Likelihood	−1,735.179	−1,732.879
Akaike Inf. Crit.	3,482.358	3,477.757
Note:	*p<0.1; **p<0.05; ***p<0.01	

Les régression obtenu après l'AIC sont donc :

$$Logit(ins) = \text{-}0.74891\ hisp_1 + 0.07706\ educyear + 0.16105\ married_1 + 0.22802\ retire_1 + 0.64627\ hhincome + \varepsilon$$

$$Probit(ins) = \text{-}0.442\ hisp_1 + 0.04842\ educyear + 0.10284\ married_1 + 0.13612\ retire_1 + 0.38863\ hhincome + \varepsilon$$

A l'exception du coefficient de la variable *married* tous les coefficients sont significatifs au seuil de 0.01.

4.2 Le troisième modèle : modèle général transformé

Pour construire ce modèle nous allons revenir au modèle général et transformer diverses variables pour tenter de les rendre plus pertinentes. Nous nous baserons essentiellement sur notre intuiton et sur les différentes informations que nous avons pu receuillir sur la base de données et le systême américain en général.

Commençons par rappeler les effectifs des modalités de la variable à expliquer :

TABLE 7 – Variable à expliquer : ins

	Effectif
0	1965
1	1241

Remarque : Toutes les valeurs des effectifs ainsi que les autres résultats qui suivront seront ceux correspondant à notre échantillon d'apprentissage, sur lequel nos modèles sont construits.

4.2.1 Transformation des variables

Passons à la transformation des variables.

chronic

Les effectifs des différentes modalités de la variable *chronic* en fonction de la variable à expliquer *ins* sont renseignés dans le tableau suivant :

TABLE 8 – Tableau croisé des effectifs : ins et chronic

	chronic								
	0	1	2	3	4	5	6	7	8
0	222	471	448	337	170	92	30	7	1
1	125	316	309	224	85	36	8	3	1

Les catégories d'individus ayant plus de 3 maladies chroniques ne sont pas du tout représentatives de l'échantillon. On aurait pu supposer que les individus ayant plus de 3 maladies chroniques seraient plus susceptible d'avoir une assurance complémentaire par rapport à ceux qui ont moins de trois maladies chroniques étant donné que **Médicare** ne couvre pas tous les frais liés à leur santé . Ce n'est pas le cas.

L'immense majorité des individus ayant plus de 3 maladies chroniques n'ont pas de complémentaire santé.

On va donc simplement séparer l'échantillon en trois catégories : la première catégorie sera constituée des individus n'ayant aucunes maladies chroniques, la seconde sera constituée des individus ayant une à deux maladies chroniques et enfin la dernière sera constituée des individus ayant plus de deux maladies chroniques, et ce afin de tenter de palier à ces incohérences observées.

On aura donc :

TABLE 9 – Tableau croisé après tranformation

	chronic		
ins	0	[1:2]	[3:8]
0	191	760	527
1	104	525	297

adl

Concernant la variable *adl*, les effectifs des modalités de cette variable par rapport à la variable à expliquer sont renseignées dans le tableau suivant :

TABLE 10 – Répartition adl :: ins

ins	adl					
	0	1	2	3	4	5
0	1423	170	89	57	24	15
1	994	72	18	15	4	4

Comme le montre la **Table 10** : la répartition des effectifs des différentes modalités par rapport à la variable à expliquer est très inégalement répartit.

La variable *adl* renseigne sur le nombre de limitations de la vie quotidienne. Nous allons transformer cette variable et la séparer en deux modalités. Ceux qui ont des contraintes et ceux qui n'en ont pas.

On aura donc :

TABLE 11 – Table modifié adl :: ins

	adl	
ins	0	[1:5]
0	1182	296
1	826	100

hhincome

La **Figure 1** montre la distribution initiale de la variable *hhincome*. Cette dernière est très mal distribuée sur l'ensemble de l'échantillon. On décide donc de la transformer en logarithme.

Après cette transformation, on constate que la variable suit une distribution normale, comme le montre la **Figure 2**.

Figure 1

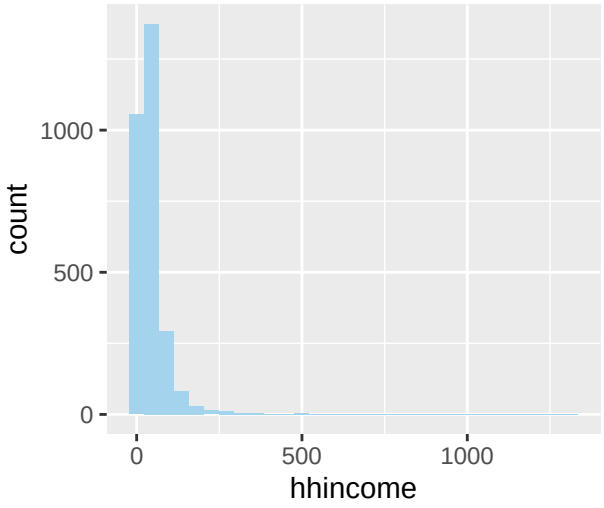
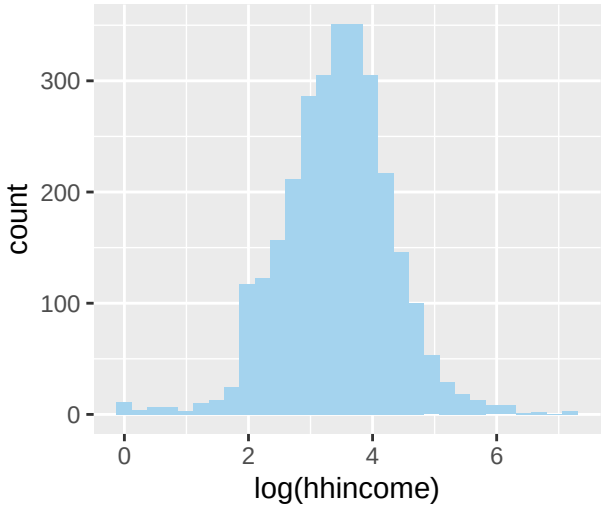


Figure 2



En plus de cette transformation en logarithme, nous allons transformer la variable *hhincome* et la subdiviser en deux classes séparées par la moyenne, tel que :

TABLE 12 – Répartition hhincome

ins	hhincome	
	[0,3.59]	(3.59,7.18]
0	1197	581
1	434	673

age

Des recherche sur le **Medicare** nous ont permis d'apprendre qu'a partir de 65 ans, ils n'y avait plus de condition pour bénéficier de l'assurance santé, d'ou le pic des effectifs.

On va donc séparé la variable *age* en deux modalités : avant et après 65 ans.

Les effectifs des nouvelles modalités de la variable ainsi transformée par rapport à la variable à expliquer sont:

TABLE 13 – Table modifié ins - age

ins	age	
	[52,64]	(64,86]
0	281	1497
1	112	995

educyear

Le parcours classique dans le systême américain s'éffectue en 12 ans.

On decide donc de séparer la variable *educyear* en trois modalités, la première : constituée des individus ayant fait moins de 12 ans d'études, la seconde : constituée des individus ayant fait 12 ans d'études et la dernière : de ceux ayant fait plus de 12 ans d'études.

Les effectifs des modalités de la variable *educyear* ainsi transformée par rapport à la variable à expliquer sont donc :

TABLE 14 – Table modifié ins - age

ins	educyear		
	0-11	12	13-17
0	695	584	499
1	215	416	476

Remarque : Diverses transformations plus ou moins pertinentes ont été effectués sur les divers variables. Seules les transformations les plus pertinentes et permettant d'obtenir le meilleur des modèles ont été susmentionnées et seront utilisées dans la suite de notre étude. Les autres transformations et résultats seront présentés en **Annexe**.

4.2.2 Régressions

Les résultats des régressions du modèle général transformé sont renseignés dans le tableau suivant :

TABLE 15 – Résultats modèles générale

	Dependent variable:	
	ins	
	logistic	*probit*
	(1)	(2)
age.d(64,86]	0.020	0.012
	(0.138)	(0.082)
hisp1	−0.924***	−0.546***
	(0.213)	(0.119)
white1	0.049	0.035
	(0.116)	(0.069)
female1	−0.104	−0.065
	(0.093)	(0.057)
educyear12	0.513***	0.312***
	(0.110)	(0.066)
educyear13-17	0.574***	0.353***
	(0.115)	(0.070)
married1	0.327***	0.195***
	(0.123)	(0.074)
chronic1-2	0.300**	0.181**
	(0.130)	(0.079)
chronic3-8	0.457***	0.277***
	(0.146)	(0.088)
adl1-5	−0.393***	−0.241***
	(0.134)	(0.079)
retire1	0.164*	0.099*
	(0.092)	(0.056)
sretire1	−0.036	−0.018
	(0.099)	(0.061)
hhincome.d(3.59,7.18]	0.779***	0.481***
	(0.093)	(0.057)
hstatusg1	0.216*	0.134**
	(0.111)	(0.067)
Constant	−1.937***	−1.188***
	(0.231)	(0.138)
Observations	2,885	2,885
Log Likelihood	−1,751.009	−1,749.173
Akaike Inf. Crit.	3,532.019	3,528.345
Note:	*p<0.1; **p<0.05; ***p<0.01	

Ce modèle transformé semble beaucoup plus pertinent que le modèle initial.

4.3 Quatrième modèle

Le **quatrième modèle** à été obtenu, comme le second, en faisant une sélection automatique des variables du modèle 3, sur le critère d'Akaïke (AIC).

Les résultats des régressions du modèle retenu sont renseignés dans le tableau suivant :

Table 16 – Résultats régréssion AIC

	Dependent variable:	
	ins	
	logistic	*probit*
	(1)	(2)
hisp1	−0.908***	−0.537***
	(0.213)	(0.118)
educyear12	0.504***	0.306***
	(0.109)	(0.065)
educyear13-17	0.573***	0.353***
	(0.115)	(0.069)
married1	0.346***	0.210***
	(0.102)	(0.061)
chronic1-2	0.297**	0.179**
	(0.130)	(0.079)
chronic3-8	0.451***	0.273***
	(0.145)	(0.088)
adl1-5	−0.406***	−0.248***
	(0.133)	(0.079)
retire1	0.188**	0.115**
	(0.087)	(0.052)
hhincome.d(3.59,7.18]	0.789***	0.487***
	(0.092)	(0.057)
hstatusg1	0.214**	0.134**
	(0.109)	(0.066)
Constant	−1.967***	−1.206***
	(0.186)	(0.111)
Observations	2,885	2,885
Log Likelihood	−1,752.012	−1,750.220
Akaike Inf. Crit.	3,526.024	3,522.440
Note:	*p<0.1; **p<0.05; ***p<0.01	

Les modèles obtenu après l'AIC sont donc :

$$Logit(ins) = \text{- } 0.9078\ 1\ hisp_1 + 0.5036\ educyear_{12} + 0.5732\ educyear_{13-17} + 0.3460\ married_1 + 0.2971\ chronic_{1-2} + 0.4507\ chronic_{3-8} \text{ - } 0.4059\ adl_{1-5} + 0.1885\ retire_1 + 0.7893\ hhincome_{(3.59,7.18]} + 0.2141\ hstatusg1 + \varepsilon$$

$$Probit(ins) = \text{-}0.5373\ 1\ hisp_1 + 0.3065\ educyear_{12} + 0.3528\ educyear_{13-17} + 0.2102\ married_1 + 0.1792\ chronic_{1-2} + 0.2731\ chronic_{3-8} \text{ - } 0.2478\ adl_{1-5} + 0.1153\ retire_1 + 0.4875\ hhincome_{(3.59,7.18]} + 0.1343\ hstatusg1 + \varepsilon$$

5 Comparaison des modèles

Dans cette partie nous allons comparer les résultats des modèles 2 et 4, qui sont respectivement les modèles obtenus apres l'AIC effectué sur le modèle général et le modèle général transformé.

Notons que nous allons içi comparer les modèles Logit. Nous allons d'abord effectuer des tests afin de déterminer si les modèles sont pertinents. Nous allons ensuite appliquer ces modèles à l'échantillon test afin de déterminer lequel est le meilleur.

Modèle 2

$$Logit(ins) = \text{-}0.74891\ hisp_1 + 0.07706\ educyear + 0.16105\ married_1 + 0.22802\ retire_1 + 0.64627\ hhincome + \varepsilon$$

Modèle 4

$$Logit(ins) = \text{-}0.9078\ 1\ hisp_1 + 0.5036\ educyear_{12} + 0.5732\ educyear_{13-17} + 0.3460\ married_1 + 0.2971\ chronic_{1-2} + 0.4507\ chronic_{3-8} \text{ - } 0.4059\ adl_{1-5} + 0.1885\ retire_1 + 0.7893\ hhincome_{(3.59,7.18]} + 0.2141\ hstatusg1 + \varepsilon$$

5.0.1 Test du rapport de vraisemblance

La statistique de test est basée sur la différence des rapports de vraisemblance entre le modèle complet et le modèle sous H_0. La statistique de test est :

$$2[\zeta_n(\widehat{\beta}) - \zeta_n(\widehat{\beta_{H_0}})] \xrightarrow{\zeta} {\chi^2}_q$$

On à :

$$H_0 : \beta_0 = \beta_1 = ... = \beta_{q-1} = 0$$
$$\text{contre}$$
$$H_1 : \exists k \in \{0, ..., q-1\} : \beta_k \neq 0$$

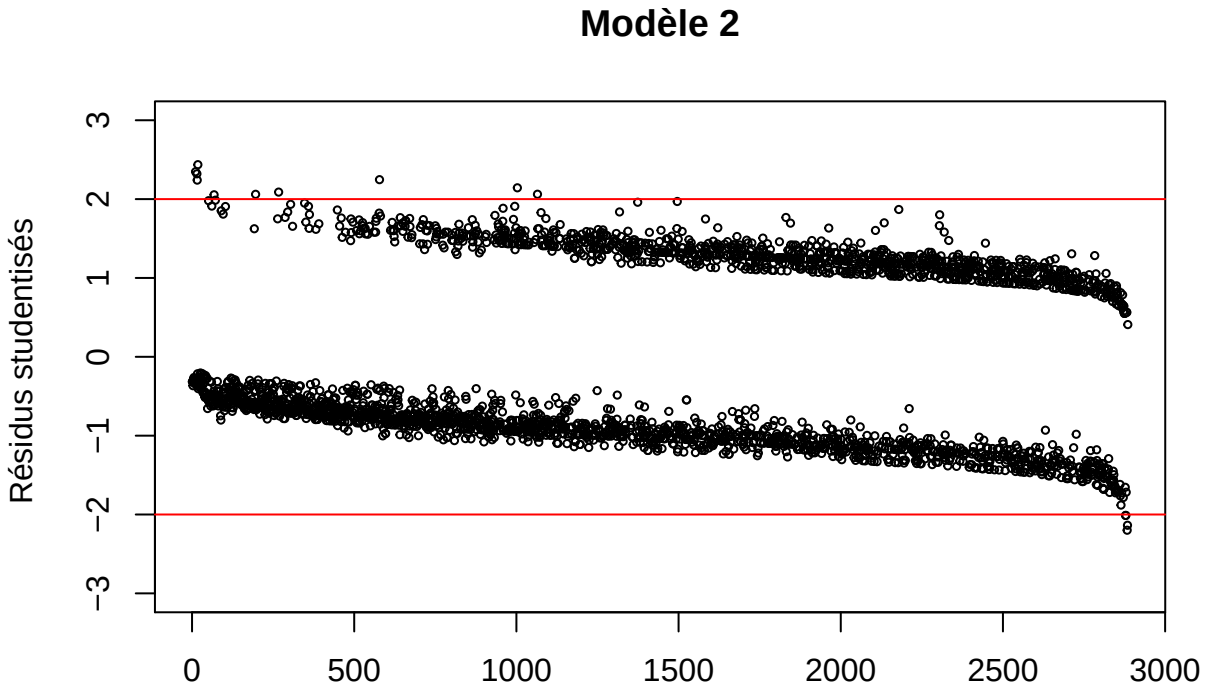
Table 17 – Résultats des tests

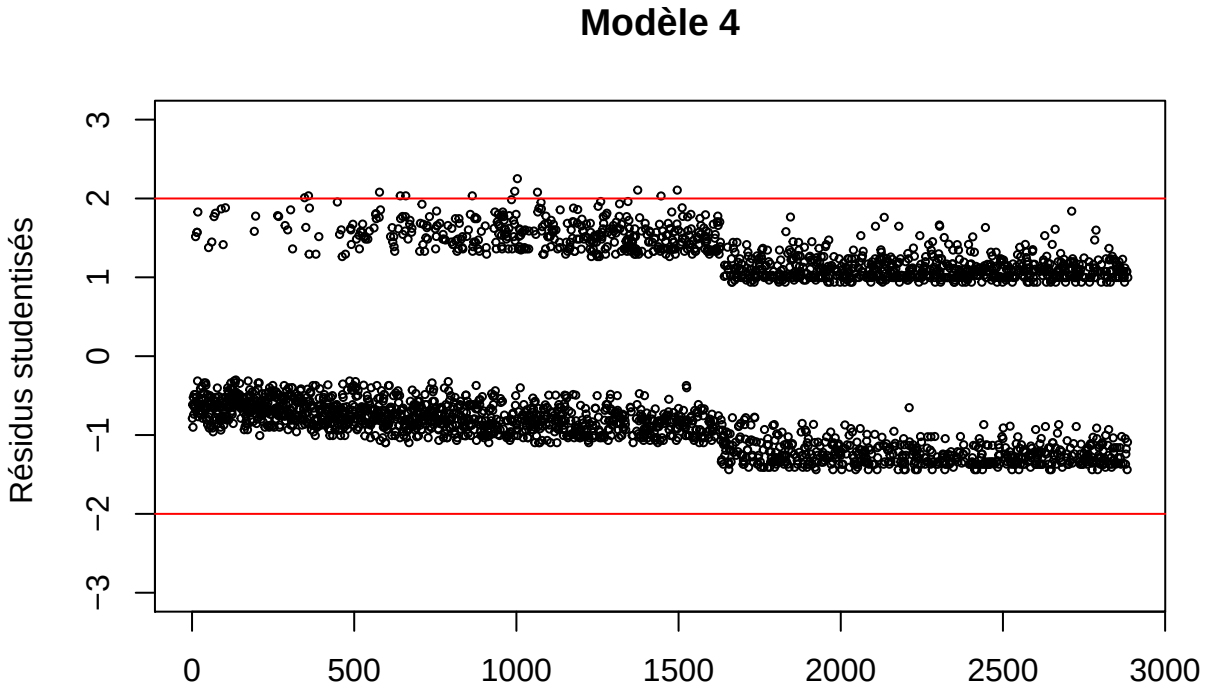
	Modèle 2	Modèle 4
chi2	371.5999	337.9344
ddl	5.0000	10.0000
pvalue	0.0000	0.0000

La *p-value* associée à ces statistiques de test sont inferieur au seuil de 5% et ce pour les deux modèles. Par conséquent nous pouvons donc conclure que les deux modèles sont globalement significatifs.

5.0.2 Résidus de déviances

Pour les régressions logistiques, on s'intéresse la plupart du temps aux résidus de déviance. Ils prennent généralement les valeurs qui oscillent entre -2 et 2. Construisons un index plot pour détecter les valeurs aberrantes.





Il semblerait que le nombre de valeurs pouvant être considérées comme aberrantes soit très minime.

5.1 Application

5.1.1 Application à l'échantillon apprentissage

TABLE 18 – Matrice de confusion apprentissage

	Modèle 2			Modèle 4	
	0	1		0	1
0	1375	600		1447	706
1	403	507		331	401

TABLE 19 – Taux d'erreur

Modèle 2	Modèle 4
0.3594454	0.3476603

Comparons les courbes ROC obtenues sur l'échantillon d'apprentissage

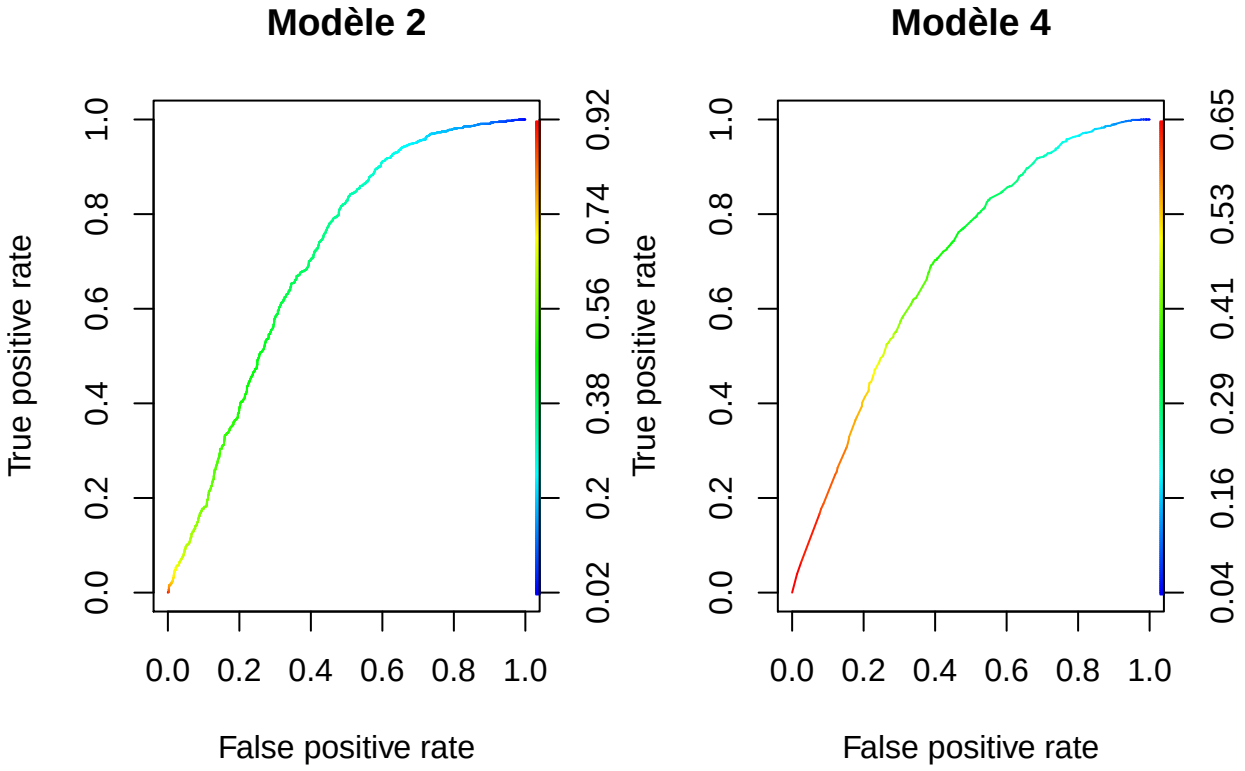


TABLE 20 – Aire sous les courbes

Modèle 2	Modèle 4
0.7043116	0.6922882

Dans la théorie de la détection du signal, l'AUC ou "aire sous la coube" fournit une mesure agrégée des performances pour tous les seuils de classification possibles. On peut interpréter l'AUC comme une mesure de la probabilité pour que le modèle classe un exemple positif aléatoire au-dessus d'un exemple négatif aléatoire. Nos courbes s'écartent de la ligne du classificateur aléatoire (modèle dans lequel l'esperance est égale à 0.5) et se rapproche du coude du classificateur idéal (qui passe de (0, 0) à (0, 1) à (1, 1)).

Ce qui signifie simplement qu'utiliser nos modèles afin de déterminer si une personne est assurée ou non est bien plus pertinent que de faire un simple pile ou face.

5.1.2 Application à l'échantillon test

TABLE 21 – Matrice de confusion test

	Modèle 2		Modèle 4	
	0	1	0	1
0	149	88	144	71
1	38	46	43	63

TABLE 22 – Taux d'erreur test

Modèle 2	Modèle 4
0.3925234	0.3551402

Le taux d'erreur de prédiction du modèle 4 est plus faible que celui du modèle 2. Il sera donc plus fiable que ce dernier. Les transformations apportées afin de construire le modèle 4 ont donc été utiles.

Comparons les courbes ROC obtenues sur l'échantillon test

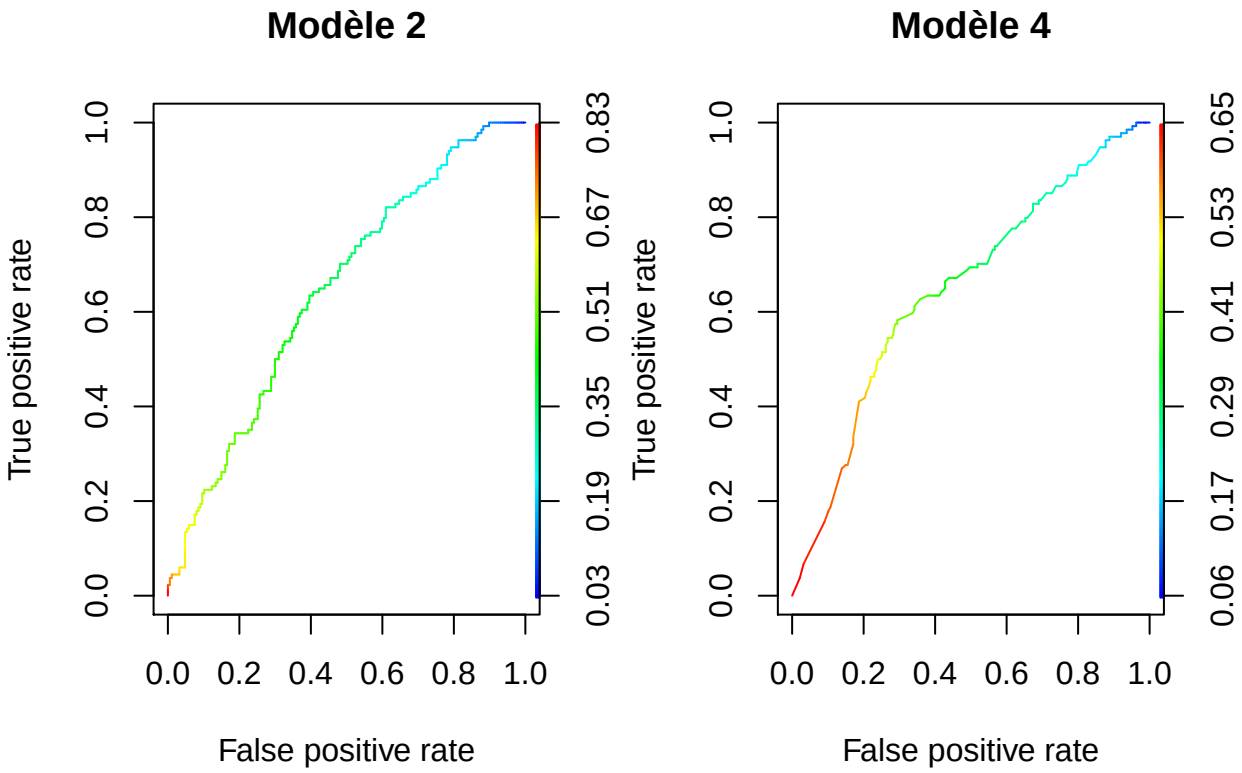


TABLE 23 – Aire sous les courbes

Modèle 2	Modèle 4
0.6443052	0.6475178

5.1.3 Modèle à choisir :

TABLE 24 – Comparaison des modèles 2 et 4

	Taux d'erreur test	sensibilité	spécificité	précision	AUC test
Modèle 2	0.3925234	0.7967914	0.3432836	0.6286920	0.6443052
Modèle 4	0.3551402	0.7700535	0.4701493	0.6697674	0.6475178

La **Table 24** résume les probabilités innhérantes aux deux modèles.

Nous savons que : la sensibilité indique la probabilité qu'un individu appartienne à la bonne catégorie sachant qu'il doit y appartenir et qu'à l'inverse la spécificité indique la probabilité qu'un individu n'appartienne pas à une catégorie sachant qu'il ne doit pas y appartenir.

Notre objectif étant de déterminer si un individu et assuré ou non, nous allons donc sélectionner le modèle ayant à la fois la sensibilité et la spécificité la plus élevée. Aussi, nous voulons que notre modèle soit le plus précis possible, nous allons donc tenir compte de la précision.

Considérant les faits susmentionnés, le modèle 4 est à tout point de vu meilleur que le modèle 2. C'est donc bien le modèle 4 qui sera retenu et étudié dans la suite de notre étude.

6 Comparaison modèles logit et probit

Nous allons içi comparer les modèles logit et probit de notre régression finale : le modèle 4.

Rappellons que :

$$Logit(ins) = \text{-}0.9078\ 1\ hisp_1 + 0.5036\ educyear_{12} + 0.5732\ educyear_{13-17} + 0.3460\ married_1 + 0.2971\ chronic_{1-2} + 0.4507\ chronic_{3-8} \text{ - } 0.4059\ adl_{1-5} + 0.1885\ retire_1 + 0.7893\ hhincome_{(3.59,7.18]} + 0.2141\ hstatusg1 + \varepsilon$$

$$Probit(ins) = \text{-}0.5373\ 1\ hisp_1 + 0.3065\ educyear_{12} + 0.3528\ educyear_{13-17} + 0.2102\ married_1 + 0.1792\ chronic_{1-2} + 0.2731\ chronic_{3-8} \text{ - } 0.2478\ adl_{1-5} + 0.1153\ retire_1 + 0.4875\ hhincome_{(3.59,7.18]} + 0.1343\ hstatusg1 + \varepsilon$$

Les valeurs des coefficients des régréssions logit et probit sont de mêmes signes mais sont différents car les spécifications ne sont pas les mêmes. Cependant, nous pouvons retrouver, approximativement, les valeurs estimées du modèle Logit en multipliant chacun des coefficients des variables explicatives du modèle Probit par la constante $\frac{\pi}{\sqrt{3}} = 1.81288$.

Table 25 – Comparaison Logit et Probit transformé

	hisp1	educyear12	educyear13-17	married1	chronic1-2
Logit	-0.9077592	0.5036353	0.5732427	0.3460205	0.2970560
Probit * 1.81288	-0.9739899	0.5556259	0.6395984	0.3809770	0.3248183

Table 26 – Comparaison Logit et Probit transformé

	chronic3-8	adl1-5	retire1	hhincome.d(3.59,7.18]	hstatusg1
Logit	0.4506626	-0.4059149	0.1884811	0.7893258	0.2140812
Probit * 1.81288	0.4951646	-0.4492702	0.2090767	0.8837142	0.2433852

Il apparaît que les résultats des modèles probit et logit sont généralement similaires que ce soit en termes de probabilité ou en termes d'estimation des coefficients β si l'on prend en compte les problèmes de normalisation. En raison de l'étroite similitude des deux distributions, il est difficile de les distinguer statistiquement à moins que l'on ne dispose d'un nombre extrêmement élevé d'observations. Ainsi, peu importe que l'on utilise le modèle probit ou le modèle logit.

Cependant, pour faciliter les interprétations nous allons, dans la suite de notre étude, nous focaliser sur le modèle Logit.

7 Interpretation

Rappellons que le modèle s'écrit :

$$Logit(ins) = \text{-}0.9078\ 1\ hisp_1 + 0.5036\ educyear_{12} + 0.5732\ educyear_{13-17} + 0.3460\ married_1 + 0.2971\ chronic_{1-2} + 0.4507\ chronic_{3-8} \text{ - } 0.4059\ adl_{1-5} + 0.1885\ retire_1 + 0.7893\ hhincome_{(3.59,7.18]} + 0.2141\ hstatusg1 + \varepsilon$$

L'estimation d'un coefficient positif est associée à une augmentation de la probabilité d'avoir une complémentaire santé.

Nous pouvons donc déduire, grâce aux signes de nos coéfficients :

- Le fait qu'un individu soit hispanique diminue la probabilité qu'il ait souscrit une assurance complémentaire ;
- Le nombre d'années d'éducation agit positivement sur la probabilité d'avoir souscrit une assurance complémentaire, plus les individus sont éduqués plus la probabilité qu'ils aient souscris une assurance complémentaire augmente ;
- Être marié agit positivement sur la probabilité de souscrire à une assurance complémentaire;
- Le nombre de maladies chroniques agit lui aussi positivement, plus les individus ont des maladies chroniques plus la probabilité qu'ils souscrivent une assurance complémentaire augmente ;
- Avoir des contraintes liés à la vie quotidienne semble agir négativement sur la probabilité de souscrire une assurance complémentaire ;
- Être retraité agit positivement sur la probabilité d'avoir une complémentaire santé ;
- Le revenu semble lui aussi agir positivement, plus les individus ont des revenus élevés plus la probabilité qu'ils souscrivent une assurance complémentaire est grande;
- La fait d'être en bonne santé est un facteur positif au fait de souscrire une assurance complémentaire ;

Intéréssons nous maintenant à l'effet de ses différentes variables. Pour ce faire nous allons calculer les **oods ratios** : "rapport de côtes".

Table 27 – OODS-RATIO

	OR	2.5 %	97.5 %	p
(Intercept)	0.1398369	0.0968252	0.2007010	0.0000000
hisp1	0.4034272	0.2612559	0.6031149	0.0000197
educyear12	1.6547258	1.3382249	2.0486203	0.0000035
educyear13-17	1.7740103	1.4179882	2.2217090	0.0000006
married1	1.4134316	1.1572646	1.7291646	0.0007274
chronic1-2	1.3458906	1.0437911	1.7409973	0.0227519
chronic3-8	1.5693516	1.1816241	2.0905469	0.0019494
adl1-5	0.6663669	0.5126413	0.8624109	0.0022034
retire1	1.2074143	1.0191511	1.4311795	0.0295120
hhincome.d(3.59,7.18]	2.2019113	1.8382485	2.6396147	0.0000000
hstatusg1	1.2387232	1.0012707	1.5338088	0.0490125

Un oods-ratio de 1 indique l'absence d'effets. On note qu'aucuns de nos intervalles de confiance ne contient la valeur 1 et que toutes les *p-value* sont inferieures à 0.05, ce qui signifie que tous les coefficients ont un effet.

hisp

Un **OR** à 0.4034272 signifie que la probabilité d'avoir une complémentaire santé est 0.4034272 fois plus faible chez les individus qui sont hispanique. Ce qui signifie simplement que les hispanique ont 60% de chance en moins davoir une complémentaire santé par rapport aux individus des autres races.

Remarque : En réalité il ne s'agit pas vraiment de la probabilité, mais plutôt de la chance ou de l'espérance de gain. Si l'étude portait sur le tirage du loto par exemple on aurait plutôt dit : l'esperance de gain est 0.4034272 fois plus faible chez les individus qui sont hispanique ou dans ce cas précis on aurait simplement pût dire: les hispaniques ont 0.4034272 fois moins de chance d'avoir une complémentaire santé ... Cependant parler de *"chance d'avoir une complémentaire santé"* me semble incohérent étant donné que l'on ne raisonne pas en terme de gain et de perte. C'est donc le terme *probabilité* qui sera utilisé.

educyear

La classe de référence de la variable *educyear* étant $educyear_{0-11}$; Le **OR** à 1.6547258 de la classe $educyear_{12}$ signifie que la probabilité d'avoir une complémentaire santé est 1.6547258 fois plus élevé chez les individus ayant 12 ans d'études que chez les individus ayant moins de 12 ans d'études, soit 65,4% de plus.

Aussi, le **OR** à 1.7740103 de la classe $educyear_{13-17}$ signifie que la probabilité d'avoir une complémentaire santé est 1.7740103 fois plus élevé chez les individus ayant entre 13 et 17 ans d'études que chez les individus ayant moins de 12 ans d'études, soit 77.4% e plus.

married

Un **OR** à 1.4134316 signifie que la probabilité d'avoir une complémentaire santé est 1.4134316 fois plus forte chez les individus qui sont marriés par rapport aux individus qui ne le sont pas. Ce qui signifie simplement que les personnes marriés ont 41,3% chance en plus d'avoir une complémentaire santé par rapport aux individus qui ne le sont pas.

chronic

La classe de référence de la variable *chronic* étant $chronic_0$; Le **OR** à 1.3458906 de la classe $chronic_{1-2}$ signifie que la probabilité d'avoir une complémentaire santé est 1.3458906 fois plus élevé chez les individus ayant entre une et deux maladies chroniques, que chez les individus n'ayant aucunes maladies chroniques. Pour le dire simplement, les individus ayant entre une et deux maladies chroniques ont 34,5% de chance en plus d'avoir une complémentaire santé par rapport aux individus n'ayant pas de maladies chroniques.

Aussi, le **OR** à 1.5693516 de la classe $chronic_{3-8}$ signifie que la probabilité d'avoir une complémentaire santé est 1.5693516 fois plus élevé chez les individus ayant entre trois et huit maladies chroniques, que chez les individus n'ayant aucunes maladies chroniques, soit 57% de chance en plus d'avoir une complémentaire santé que ceux qui n'ont pas de maladies chroniques.

adl

La classe de référence de la variable *adl* étant adl_0 ; Le **OR** à 0.6663669 de la classe adl_{1-5} signifie que la probabilité d'avoir une complémentaire santé est 0.6663669 fois plus faible chez les individus ayant des limitations par rapport aux individus n'ayant aucunes limitations. Un individu ayant des limitations à donc 33.36% chance en moins d'avoir une complémentaire santé.

retire

Le **OR** à 1.2074143 de la variable *retire* signifie que la probabilité d'avoir une complémentaire santé est 1.2074143 fois plus élevé chez les individus qui sont retraités par rapport à ceux qui sont actifs. Un individu à la retraite aura 20% de chance en plus d'avoir une complémentaire santé, par rapport à un individu actif.

hhincome

La classe de référence de la variable *hhincome* étant $hhincome_{0-3,59}$; Le **OR** à 2.2019113 de la classe $hhincome_{3.59-7.18}$ signifie que la probabilité d'avoir une complémentaire santé est 2.2019113 fois plus élevé chez les individus ayant un revenu en logaritme compris entre 3.59 et 7.18, que chez les individus ayant un revenu plus faible.

Pour le dire simplement, les individus ayant un revenu superieur à la moyenne ont 120% de chance en plus d'avoir une complémentaire santé par rapport à ceux qui ont un revenu inférieur à la moyenne.

hstatug

Le **OR** à 1.2387232 de la variable *retire* signifie que la probabilité d'avoir une complémentaire santé est 1.2387232 fois plus élevé chez les individus qui sont en bonne santé par rapport à

ceux qui ne le sont pas. Un individu en bonne santé aura 24% de chance en plus d'avoir une complémentaire santé, par rapport à un individu en mauvaise santé.

8 Marginal effects

Un des principaux problème avec le rapport de côtes est que de nombreuses paires de résultats donnent exactement le même rapport de côtes. Au lieu des rapports de côtes, il serait intérréssant de calculer les effets marginaux. Ces derniers sont plus simples à interpréter et à comprendre car ils donnent une mesure directe : la différence moyenne de probabilité en termes de points de pourcentage du les classes. Ainsi, les effets marginaux fournissent une statistique bien meilleure et plus informative par rapport aux rapports de côtes.

Table 28 – Effets marginaux

hisp1	educyear12	educyear13-17	married	chronic1-2
-0.1731724	0.1057364	0.1212139	0.0724989	0.0601876

hisp

L'effet marginale de la classe *hisp1* est de -0.1731724 ce qui signifie que la différence moyenne de probabilité entre les personnes assurés qui ne sont pas hispaniques et les personnes assuré qui le sont est de -17.3 points de pourcentage.

educyear

L'effet marginale de la classe $educyear_{12}$ est de 0.1057364 ce qui signifie que la différence moyenne de probabilité entre les personnes assurés qui ont de 12 ans d'études et les personnes assurés qui ont moins 12 ans d'études est de 10.5 points de pourcentage.

L'effet marginale de la classe $educyear_{13-17}$ est de 0.1212139 ce qui signifie que la différence moyenne de probabilité entre les personnes assurés qui ont plus de 12 ans d'étude et les personnes assurés qui ont moins de 12 ans d'étude est de 12 points de pourcentage.

married

L'effet marginale de la classe *married1* est de 0.0724989 ce qui signifie que la différence moyenne de probabilité entre les personnes assuré qui sont marriés et les personnes assuré qui ne le sont pas est de 7.2 points de pourcentage.

Table 29 – Effets marginaux

chronic3-8	adl1-5	retire	hhincome.d(3.59,7.18]	hstatusg1
0.0926718	-0.0835464	0.039587	0.1757659	0.0449592

chronic

L'effet marginale de la classe $chronic_{1-2}$ est de 0.0601876 ce qui signifie que la différence moyenne de probabilité entre les personnes assurés qui ont 1 à 2 maladie chroniques et les personnes assurés qui n'en ont aucune est de 6 points de pourcentage .

L'effet marginale de la classe $chronic_{3-8}$ est de 0.0926718 ce qui signifie que la différence moyenne de probabilité entre les personnes assurés qui ont 3 à 8 maladies chroniques et les personnes assurés qui n'en ont aucune est de 9.2 points de pourcentage.

adl

L'effet marginale de la classe adl_{1-5} est de -0.0835464 ce qui signifie que la différence moyenne de probabilité entre les personnes assurés qui ont des limitations et ceux qui n'en ont pas chronique est de -8.3 points de pourcentage.

retire

L'effet marginale de la classe *retire1* est de 0.039587 ce qui signifie que la différence moyenne de probabilité entre les personnes assurés qui n'ont sont retraités avec ceuxx qui ne le sont pas est de 4 points de pourcentage.

hhincome

L'effet marginale de la classe $hhincome_{(3.59,7.18]}$ est de 0.1757659 ce qui signifie que la différence moyenne de probabilité entre les personnes assurés qui ont un revenu en logarithme superieur à la moyenne et les personnes assurés qui ont un revenu inferieur à la moyenne est de 17.6 points de pourcentage.

hstatug

L'effet marginale de la classe *hstatug* est de 0.0449592 ce qui signifie que la différence moyenne de probabilité entre les personnes assurés qui sont en bonne santé et les personnes assurés qui ne le sont pas est de 4.4 points de pourcentage.

9 Discussion

Commençons par la variable **hisp**: s'agissant d'une variable de type signalétique, il n'y a pas grand chose à dire. Ils ressort simplement que les hispaniques ont moins tendance à souscrire une assurance que les autres races. Sur l'ensemble de notre base de données seuls 233 individus sont hispaniques ce qui représente à peine 6 % des individus et parmi eux seuls 35 individus ont une complémentaire santé. Étant donné le faible nombre d'individus concernés on aurait pu pensé qu'il faudrait être prudent et éviter de faire des généralités mais ce faible effectif s'explique simplement par le fait que **Medicare** est une institution dont les services sont principalement destinées aux américains.

Concernant la variable **educyear** : la séparer en 3 modalités était une très bonne idée. Cela nous a permis d'observer l'influence de ses diverses modalités sur le fait de souscrire à une complémenatire santé. Les différents résultats nous ont permis de conclure que le nombre d'années d'éducation avait un impact significatif sur la probabilité de souscrire à une complémenatire santé. On peut supposé sans trop s'avancer que le nombre d'années d'éducation est corrélé

d'une manière ou d'une autre avec la variable salaire : **hhincome**, qui elle aussi contribue de manière significative et augmente la probabilité d'avoir une complémentaire santé à mesure qu'elle augmente. Cette *pseudo-corrélation* pourrait expliquer ces effets marginaux très élevés.

En ce qui concerne les variables **married** et **retire**: les résultats obtenus précédemment nous ont permis de déduire que le fait d'etre marrié et d'être retraité augmentait la probabilité de souscire à une complémentaire santé.

En ce qui concerne les maladies chroniques et le nombre de limitations : **Medicare** permet de couvrir seulement certaines maladies chroniques, ce qui implique trivialement qu'un grand nombre de maladies chroniques augmentent la probabilité de recourir à une complémentaire santé. Sachant que **Medicare** ne s'occupe pas des soins de longues durées, cette relation est plus qu'évidente.
Avoir des limitations liés à la vie quotidienne semble diminuer la probabilité de recourir à une complémentaire. Ceçi s'explique certainement par le fait que lorsqu'on à une assurance santé **Medicare** on peut etre éligible au programme appelée **Medicaid** : il s'agit d'une autre assurance santé conçu spécialement pour les personnes en situation de handicap et qui ne nécessite pas le recours à une assurance complémenataire.

10 Limitations

Les résultats obtenus précedemment nous permettent de conclure que certaines des caractéristiques qui induisent une faible probabilité d'avoir une complémentaire santé sont les suivantes : avoir beaucoup de handicap, ne pas avoir fait de longues études, être pauvre (avoir des revenus en dessous de la moyenne), être célibataire, être en mauvaise santé.

Ces quelques caractéristiques suffisent pour montrer les limites de la complémentaire santé **Medigap** et plus généralement du systême de santé américain. Ce sont ceux qui en ont le plus besoin qui ont la probabilité la plus faible d'y souscrire. C'est un systême qui est censé aider à couvrir la différence de coût qui n'est pas couverte par l'assurance maladie et donc aider les personnes dans le besoin mais on constate que ce systême bénéficie aux plus riches alors que : à maladies et limitations égales, ce sont les pauvres qui en ont le plus besoin.

11 Annexe

11.1 AIC modèle général :

```
## Start:  AIC=3490.12
## ins ~ age + hisp + white + female + educyear + married + chronic +
##     adl + retire + sretire + hhincome + hstatusg
##
##            Df Deviance    AIC
## - sretire   1   3442.2 3488.2
## - white     1   3442.3 3488.3
## - female    1   3442.8 3488.8
## - adl       5   3451.3 3489.3
## - chronic   8   3457.4 3489.4
## - married   1   3443.8 3489.8
## - hstatusg  1   3444.0 3490.0
## <none>          3442.1 3490.1
## - age       1   3444.8 3490.8
## - retire    1   3447.1 3493.1
## - hisp      1   3454.2 3500.2
## - educyear  1   3465.7 3511.7
## - hhincome  1   3540.1 3586.1
##
## Step:  AIC=3488.18
## ins ~ age + hisp + white + female + educyear + married + chronic +
##     adl + retire + hhincome + hstatusg
##
##            Df Deviance    AIC
## - white     1   3442.3 3486.3
## - female    1   3443.2 3487.2
## - adl       5   3451.4 3487.4
## - chronic   8   3457.4 3487.4
## - married   1   3444.0 3488.0
## - hstatusg  1   3444.1 3488.1
## <none>          3442.2 3488.2
## - age       1   3444.9 3488.9
## - retire    1   3447.1 3491.1
## - hisp      1   3454.2 3498.2
## - educyear  1   3465.7 3509.7
## - hhincome  1   3540.4 3584.4
##
## Step:  AIC=3486.35
## ins ~ age + hisp + female + educyear + married + chronic + adl +
##     retire + hhincome + hstatusg
##
##            Df Deviance    AIC
## - female    1   3443.3 3485.3
## - adl       5   3451.5 3485.5
## - chronic   8   3457.6 3485.6
## - married   1   3444.1 3486.1
## - hstatusg  1   3444.2 3486.2
```

```
## <none>          3442.3 3486.3
## - age       1   3445.2 3487.2
## - retire    1   3447.3 3489.3
## - hisp      1   3454.6 3496.6
## - educyear  1   3465.7 3507.7
## - hhincome  1   3540.9 3582.9
##
## Step:  AIC=3485.32
## ins ~ age + hisp + educyear + married + chronic + adl + retire +
##     hhincome + hstatusg
##
##            Df Deviance    AIC
## - chronic   8   3458.6 3484.6
## - adl       5   3452.7 3484.7
## - hstatusg  1   3445.0 3485.0
## <none>          3443.3 3485.3
## - age       1   3445.7 3485.7
## - married   1   3445.9 3485.9
## - retire    1   3449.3 3489.3
## - hisp      1   3455.3 3495.3
## - educyear  1   3466.3 3506.3
## - hhincome  1   3543.5 3583.5
##
## Step:  AIC=3484.58
## ins ~ age + hisp + educyear + married + adl + retire + hhincome +
##     hstatusg
##
##            Df Deviance    AIC
## - adl       5   3466.8 3482.8
## - hstatusg  1   3458.8 3482.8
## <none>          3458.6 3484.6
## - age       1   3460.7 3484.7
## - married   1   3461.5 3485.5
## - retire    1   3465.3 3489.3
## - hisp      1   3471.6 3495.6
## - educyear  1   3480.7 3504.7
## - hhincome  1   3556.2 3580.2
##
## Step:  AIC=3482.77
## ins ~ age + hisp + educyear + married + retire + hhincome + hstatusg
##
##            Df Deviance    AIC
## - age       1   3468.6 3482.6
## <none>          3466.8 3482.8
## - hstatusg  1   3469.1 3483.1
## - married   1   3469.5 3483.5
## - retire    1   3474.5 3488.5
## - hisp      1   3480.2 3494.2
## - educyear  1   3488.6 3502.6
## - hhincome  1   3571.8 3585.8
```

```
##
## Step:  AIC=3482.64
## ins ~ hisp + educyear + married + retire + hhincome + hstatusg
##
##            Df Deviance    AIC
## - hstatusg  1   3470.4 3482.4
## <none>          3468.6 3482.6
## - married   1   3471.0 3483.0
## - retire    1   3475.0 3487.0
## - hisp      1   3482.3 3494.3
## - educyear  1   3491.0 3503.0
## - hhincome  1   3572.5 3584.5
##
## Step:  AIC=3482.36
## ins ~ hisp + educyear + married + retire + hhincome
##
##            Df Deviance    AIC
## <none>          3470.4 3482.4
## - married   1   3472.6 3482.6
## - retire    1   3477.3 3487.3
## - hisp      1   3484.0 3494.0
## - educyear  1   3495.5 3505.5
## - hhincome  1   3582.9 3592.9

##
## Call:  glm(formula = ins ~ hisp + educyear + married + retire + hhincome,
##     family = binomial(link = logit), data = appren1)
##
## Coefficients:
## (Intercept)        hisp1     educyear     married1      retire1
##    -3.89777     -0.74891      0.07706      0.16105      0.22802
##    hhincome
##     0.64627
##
## Degrees of Freedom: 2884 Total (i.e. Null);  2879 Residual
## Null Deviance:       3842
## Residual Deviance: 3470  AIC: 3482
```

11.2 AIC Modèle général transformé :

```
## Start:  AIC=3532.02
## ins ~ age.d + hisp + white + female + educyear + married + chronic +
##     adl + retire + sretire + hhincome.d + hstatusg
##
##              Df Deviance    AIC
## - age.d       1   3502.0 3530.0
## - sretire     1   3502.1 3530.1
## - white       1   3502.2 3530.2
## - female      1   3503.3 3531.3
## <none>            3502.0 3532.0
## - retire      1   3505.2 3533.2
```

```
## - hstatusg    1   3505.8 3533.8
## - married     1   3509.1 3537.1
## - chronic     2   3512.0 3538.0
## - adl         1   3510.9 3538.9
## - hisp        1   3523.9 3551.9
## - educyear    2   3531.1 3557.1
## - hhincome.d  1   3573.4 3601.4
##
## Step:  AIC=3530.04
## ins ~ hisp + white + female + educyear + married + chronic +
##     adl + retire + sretire + hhincome.d + hstatusg
##
##              Df Deviance    AIC
## - sretire     1   3502.2 3528.2
## - white       1   3502.2 3528.2
## - female      1   3503.3 3529.3
## <none>            3502.0 3530.0
## - retire      1   3505.5 3531.5
## - hstatusg    1   3506.1 3532.1
## - married     1   3509.1 3535.1
## - chronic     2   3512.0 3536.0
## - adl         1   3511.1 3537.1
## - hisp        1   3523.9 3549.9
## - educyear    2   3531.1 3555.1
## - hhincome.d  1   3573.6 3599.6
##
## Step:  AIC=3528.16
## ins ~ hisp + white + female + educyear + married + chronic +
##     adl + retire + hhincome.d + hstatusg
##
##              Df Deviance    AIC
## - white       1   3502.4 3526.4
## - female      1   3503.8 3527.8
## <none>            3502.2 3528.2
## - retire      1   3505.5 3529.5
## - hstatusg    1   3506.1 3530.1
## - chronic     2   3512.1 3534.1
## - married     1   3510.3 3534.3
## - adl         1   3511.2 3535.2
## - hisp        1   3523.9 3547.9
## - educyear    2   3531.1 3553.1
## - hhincome.d  1   3573.9 3597.9
##
## Step:  AIC=3526.35
## ins ~ hisp + female + educyear + married + chronic + adl + retire +
##     hhincome.d + hstatusg
##
##              Df Deviance    AIC
## - female      1   3504.0 3526.0
## <none>            3502.4 3526.4
```

```
## - retire      1   3505.7 3527.7
## - hstatusg    1   3506.5 3528.5
## - chronic     2   3512.3 3532.3
## - married     1   3510.9 3532.9
## - adl         1   3511.5 3533.5
## - hisp        1   3523.9 3545.9
## - educyear    2   3532.5 3552.5
## - hhincome.d  1   3574.8 3596.8
##
## Step:  AIC=3526.02
## ins ~ hisp + educyear + married + chronic + adl + retire + hhincome.d +
##     hstatusg
##
##              Df Deviance    AIC
## <none>            3504.0 3526.0
## - hstatusg    1   3507.9 3527.9
## - retire      1   3508.8 3528.8
## - chronic     2   3513.8 3531.8
## - adl         1   3513.6 3533.6
## - married     1   3515.6 3535.6
## - hisp        1   3525.3 3545.3
## - educyear    2   3533.3 3551.3
## - hhincome.d  1   3577.9 3597.9

##
## Call:  glm(formula = ins ~ hisp + educyear + married + chronic + adl +
##     retire + hhincome.d + hstatusg, family = binomial(link = logit),
##     data = appren)
##
## Coefficients:
##           (Intercept)                  hisp1             educyear12
##               -1.9673                -0.9078                 0.5036
##         educyear13-17               married1             chronic1-2
##                0.5732                 0.3460                 0.2971
##            chronic3-8                  adl1-5                retire1
##                0.4507                -0.4059                 0.1885
## hhincome.d(3.59,7.18]              hstatusg1
##                0.7893                 0.2141
##
## Degrees of Freedom: 2884 Total (i.e. Null);  2874 Residual
## Null Deviance:       3842
## Residual Deviance: 3504  AIC: 3526
```

11.3 Les autres transformations

On sépare hhincome en trois classes.

Age en classe de 10 ans.

Régression obtenu

```
##
```

TABLE 30 – Répartition hhincome

ins	hhincome		
	[0,2.39]	(2.39,4.79]	(4.79,7.18]
0	360	1518	87
1	22	1141	78

TABLE 31 – Répartition hhincome

ins	hhincome		
	[52,60]	(60,70]	(70,86]
0	134	1427	217
1	47	924	136

```
## Call:
## glm(formula = ins ~ agee.d + hisp + white + female + educyear +
##     married + chronic + adl + retire + sretire + hhincome.d +
##     hstatusg, family = binomial(link = logit), data = appren)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4794  -0.9471  -0.6408   1.1019   2.2376
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -2.054411   0.272001  -7.553 4.25e-14 ***
## agee.d(60,70]           0.165256   0.195742   0.844  0.39853
## agee.d(70,86]           0.007589   0.227746   0.033  0.97342
## hisp1                  -0.917786   0.213392  -4.301 1.70e-05 ***
## white1                  0.048691   0.115663   0.421  0.67378
## female1                -0.128058   0.094992  -1.348  0.17763
## educyear12              0.512632   0.110333   4.646 3.38e-06 ***
## educyear13-17           0.574554   0.115525   4.973 6.58e-07 ***
## married1                0.334970   0.123485   2.713  0.00668 **
## chronic1-2              0.304686   0.130529   2.334  0.01958 *
## chronic3-8              0.468314   0.145891   3.210  0.00133 **
## adl1-5                 -0.378012   0.133762  -2.826  0.00471 **
## retire1                 0.160200   0.091768   1.746  0.08086 .
## sretire1               -0.036169   0.099178  -0.365  0.71535
## hhincome.d(3.59,7.18]  0.774344   0.092724   8.351  < 2e-16 ***
## hstatusg1               0.214695   0.110065   1.951  0.05110 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3842.0  on 2884  degrees of freedom
## Residual deviance: 3499.9  on 2869  degrees of freedom
## AIC: 3531.9
##
## Number of Fisher Scoring iterations: 4
```