Marketing et analyse

BOUSSENGUI François

Table des matières

1 Présentation de la base de données

1.1 Description des variables

- `Age` : Age de l'assuré ;
- `Hispanique` : Vaut 1 si l'individu est hispanique et 0 sinon ;
- `Blanc` : Vaut 1 si l'individu est blanc et 0 sinon ;
- `Femme`: Vaut 1 si l'individu est une femme et 0 sinon ;
- `Education` : Nombre d'années d'étude ;
- `Statut matrimonial`: Vaut 1 si l'individu est marié et 0 sinon ;
- `Maladies chroniques`: Nombre de maladies chroniques ;
- `Adl` : Nombre de limitations sur les activités quotidiennes. Ce sont les activités quotidiennes que l'on effectue tous les jours sans besoins d'assistance. Ces activités sont : manger, se doucher, s'habiller, se lever et se coucher, marcher ou encore faire sa toilette.
- `Retraité` : Vaut 1 si l'individu est retaité et 0 sinon ;
- `Statut du conjoint` : Vaut 1 si le conjoint de l'individu est retraité et 0 sinon ;
- `Revenu` : Revenu du ménage ;
- `Assurance` : Vaut 1 si l'indiviu à une complémentaire et 0 sinon ;
- `Etat de sante` : Variable indiquant si l'etat de santé est excellent, très bon, bon, moyen ou mauvais ;

1.2 Contexte et problématique

Medicare est un programme d'assurance maladie qui aide les personnes âgées à payer les services de santé. Il est financé par le gouvernement fédéral des États-Unis.En 2016, environ 57 millions de personnes étaient couvertes par Medicare. Parmi celles-ci, 48 millions avaient 65 ans ou plus, tandis que 9 millions étaient plus jeunes mais atteintes d'un handicap permanent.

Medicare est divisé en plusieurs parties qui couvrent les coûts de différents services :

- Partie A : Soins hospitaliers non ambulatoires, soins palliatifs, établissements de soins infirmiers spécialisés (sous certaines conditions);
- Partie B : Services fournis par les médecins et autres prestataires de soins de santé, soins ambulatoires, y compris certains équipements médicaux.
- Partie C : Régime Medicare Advantage (un régime de santé Medicare proposé par une entreprise privée sous contrat avec Medicare pour gérer les remboursements).
- Partie D : Médicaments sur ordonnance

Bien que le programme soit financé et réglementé par le gouvernement, les interventions quotidiennes de Medicare sont gérées par des entreprises privées. Les prestataires administratifs de Medicare gèrent les Parties A et B, et les compagnies d'assurance privées gèrent Medicare Advantage (Partie C).

Pour être admissible à Medicare il faut avoir les caractéristiques suivantes :

- Être âgé(e) de plus de 65 ans
- Être sous dialyse rénale ou avoir reçu une transplantation rénale
- Être âgé(e) de moins de 65 ans et atteint de certains handicaps
- Être atteint(e) d'une sclérose latérale amyotrophique (maladie de Charcot).

Medicare prend en charge uniquement les services considérés comme appropriés (appelés services couverts). Pour chaque service couvert, Medicare comporte ce qu'on appelle un coût admissible. Le coût admissible est le montant maximal autorisé par Medicare et facturé aux personnes pour un service par le prestataire de soins de santé. Cependant, Medicare ne paie pas la totalité des coûts admissibles pour les services couverts. La première fois que les personnes ont besoin d'un service, elles doivent généralement payer une petite somme fixe (appelée franchise) avant tout paiement de Medicare.

Certaines personnes souscrivent une assurance complémentaire (Medigap) qui les aide à payer le reste à charge de Medicare et les autres dépenses médicales qui ne sont pas couvertes par Medicare. Cette assurance est parfois fournie par le dernier employeur de la personne dans le cadre d'une prestation de retraite. D'autres personnes souscrivent une assurance maladie complémentaire auprès de sociétés d'assurances privées.

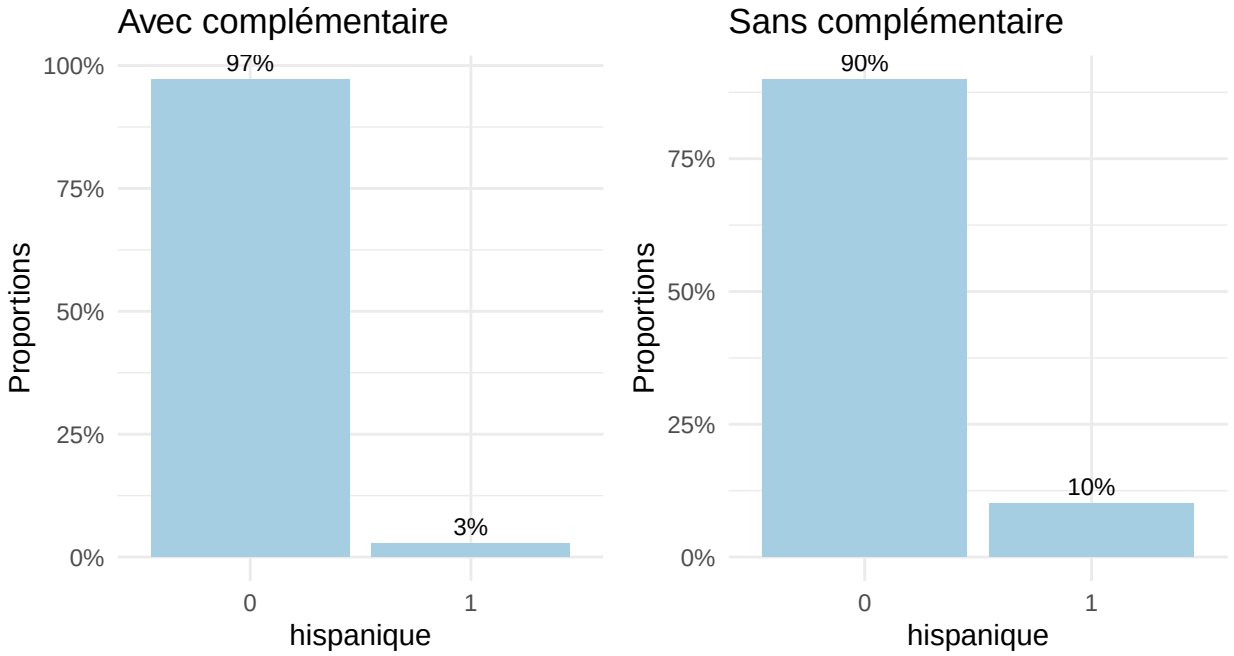
Comme Medicare et Medigap ne couvrent pas les soins de longue durée, certaines personnes souscrivent une autre assurance pour financer les soins de longue durée. La décision de souscrire une assurance de soins de longue durée doit tenir compte entre autres des deux questions suivantes : les personnes s'attendent-elles à avoir besoin d'une aide pour payer les soins de longue durée, ou bien ont-elles les moyens de payer une assurance de soins de longue durée ?

Les personnes qui ont un faible revenu et peu de ressources peuvent avoir droit à une couverture complémentaire par le régime Medicaid financé par le gouvernement.

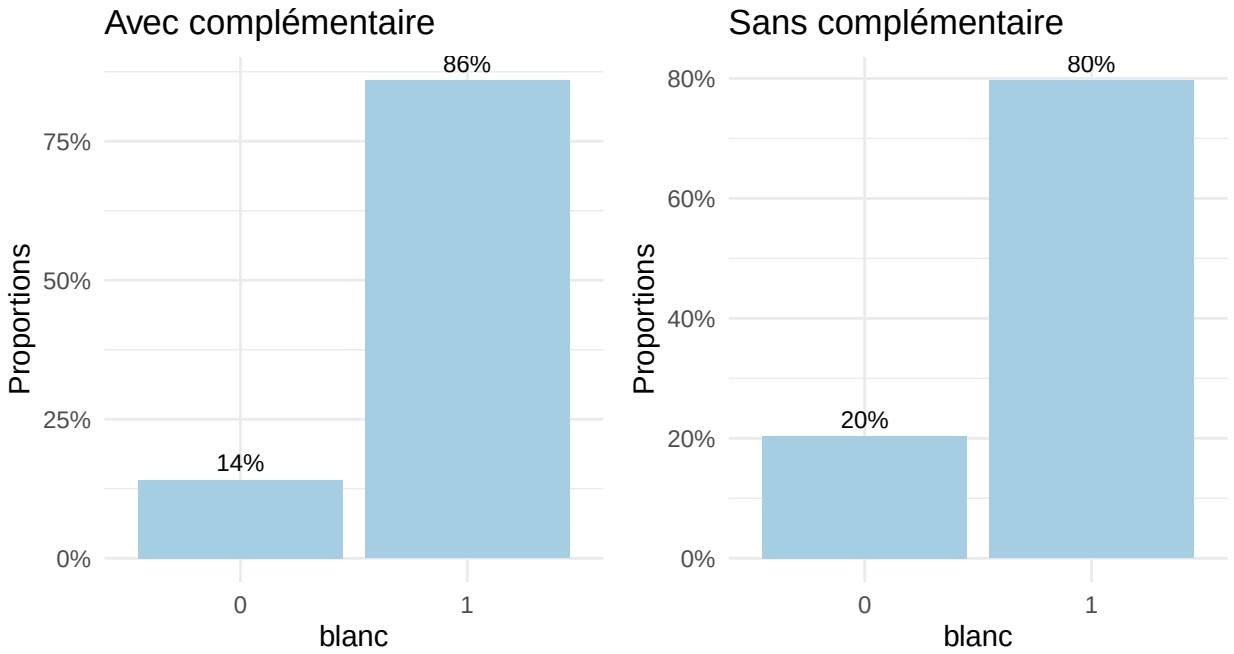
Notre base de données contient donc des informations sur des individus benéficiant des services de Medicare. L'objectif sera donc d'effectuer un travail d'analyse afin de comprendre pourquoi certains individus, en plus d'être assuré chez Medicare, ont en plus une complémentaire. Pour ce faire, nous allons tout d'abord effectuer une analyse descriptive de notre base de données puis mettre en place des régressions à variable à expliquer binaire pour répondre à notre problématique.

2 Statistiques descriptives

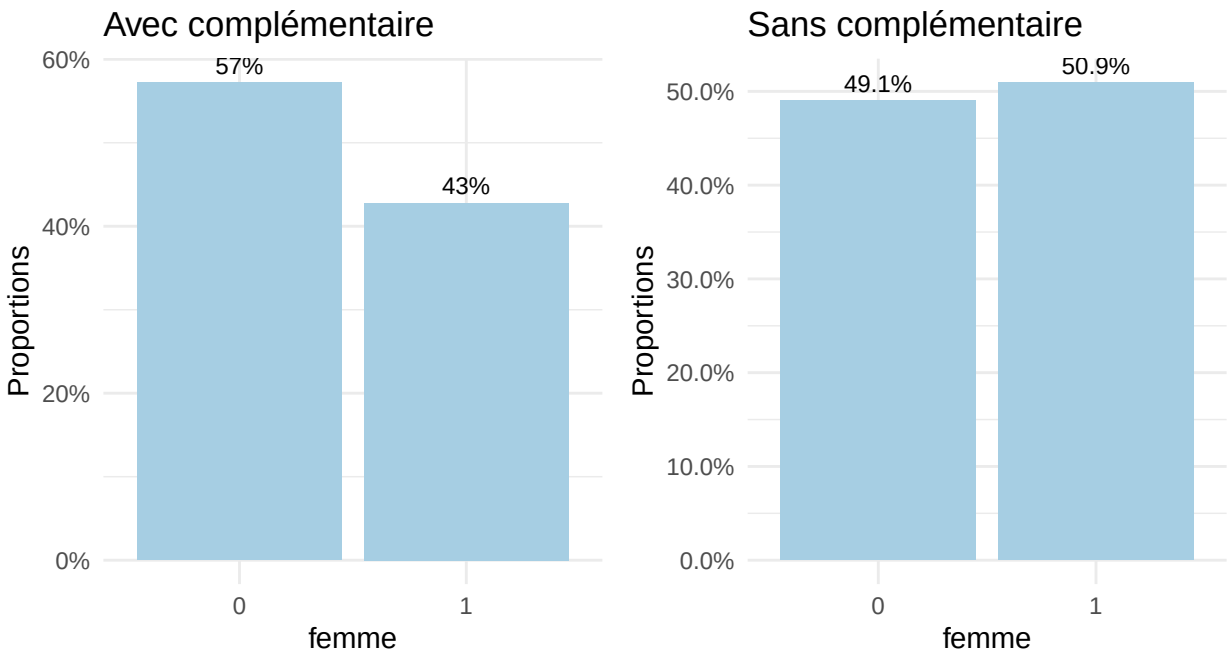
Dans le cadre des statistiques descriptives, la population a été répartie en 2 sous-groupes à savoir : Les bénéficiaires de Medicare ayant une complémentaire et ceux n'en ayant pas. Notons que les individus ayant une complémentaire représentent 39 % de la population totale.



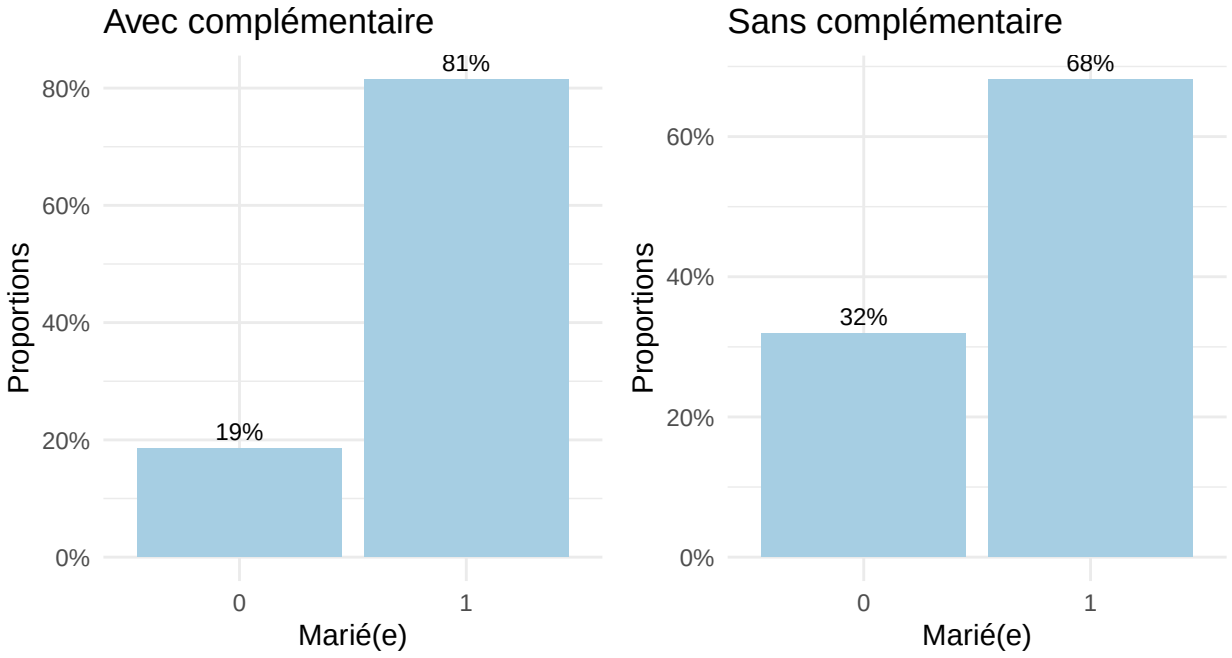
De manière générale, les individus de race hispanique sont très faiblement représentés dans la population. En effet, chez les individus ayant une complémentaire, ils représentent 3 %. Par contre, ils représentent 10 % des individus n'ayant pas de complémentaire.



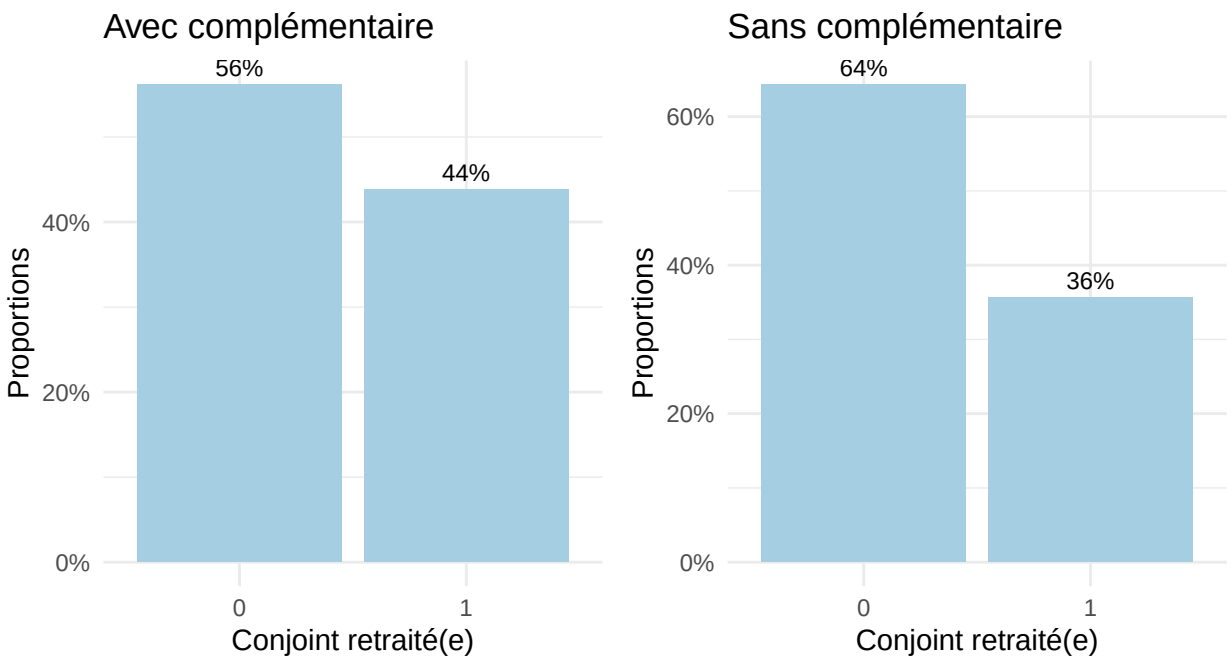
Les individus de race blanche sont majoritairement representés. En effet, Chez les individus ayant une complémentaire, ils représentent 86 % de la population et chez ceux n'ayant pas de complémentaire ils représentent 80 %.



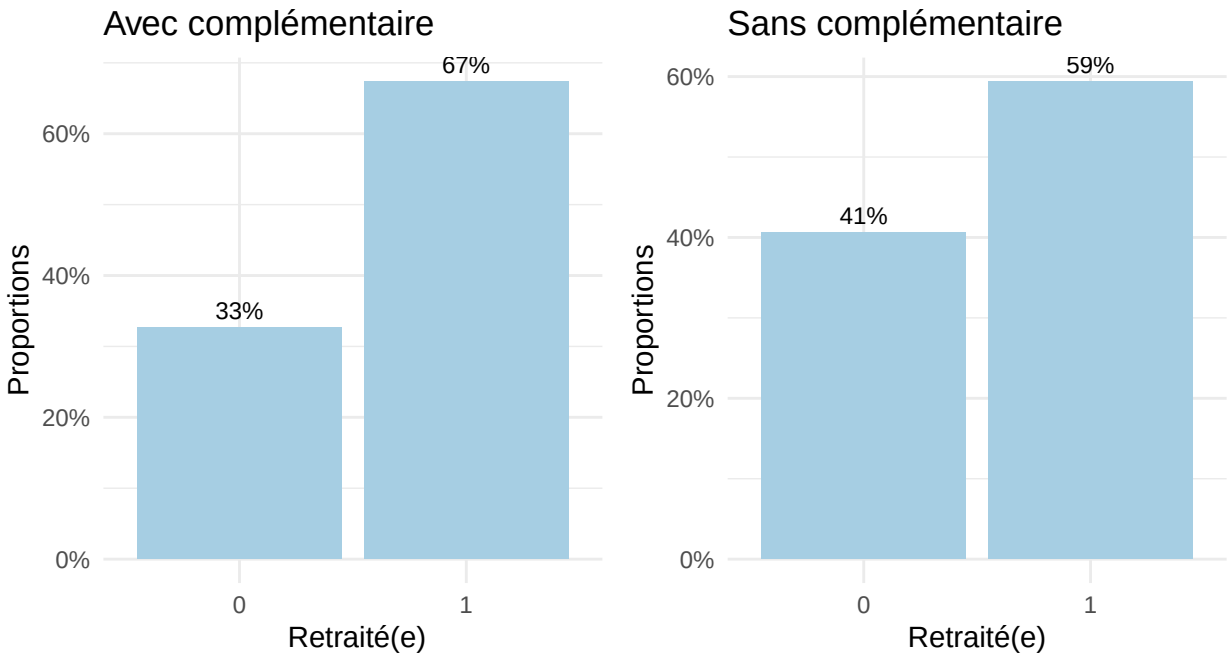
Chez les individus ayant une complémentaire, les femmes sont représentées à hauteur de 43 % tandis que dans la population des individus n'ayant pas de complémentaire, elles représentent 50.9 %.



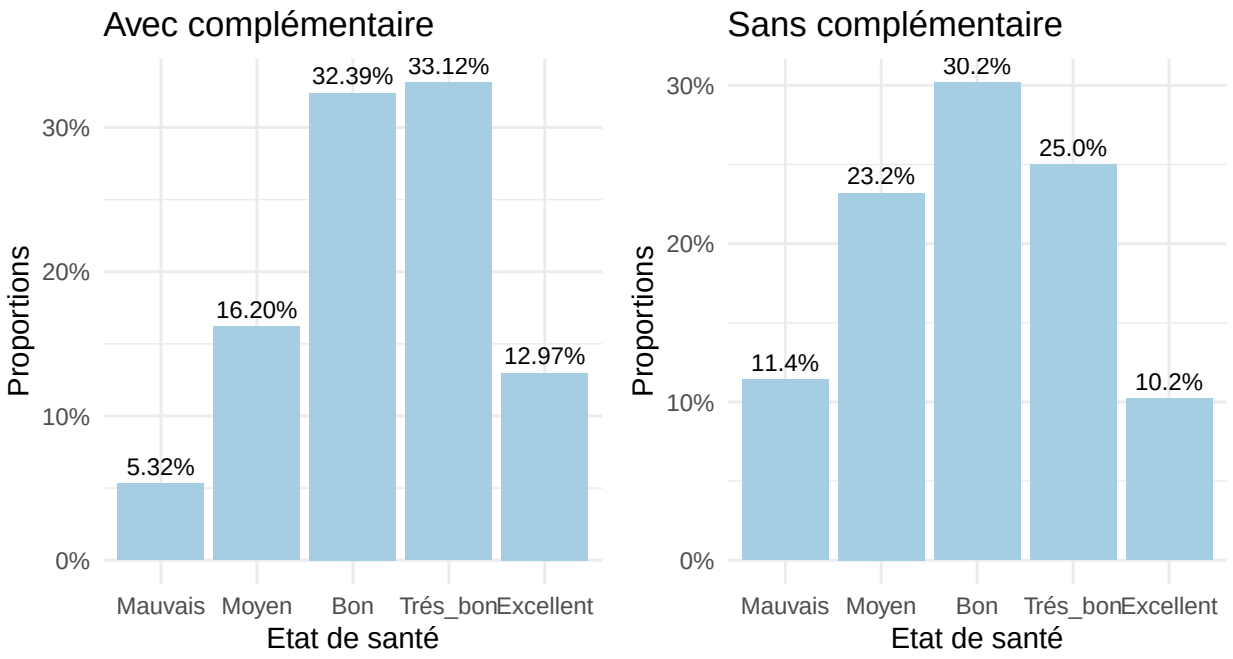
Globalement, les individus de cette base de données sont mariés. En effet, chez les individus ayant une complémentaire, 81 % sonr mariés. Chez les individus n'ayant pas de complémentaire, 68 % sont mariés.



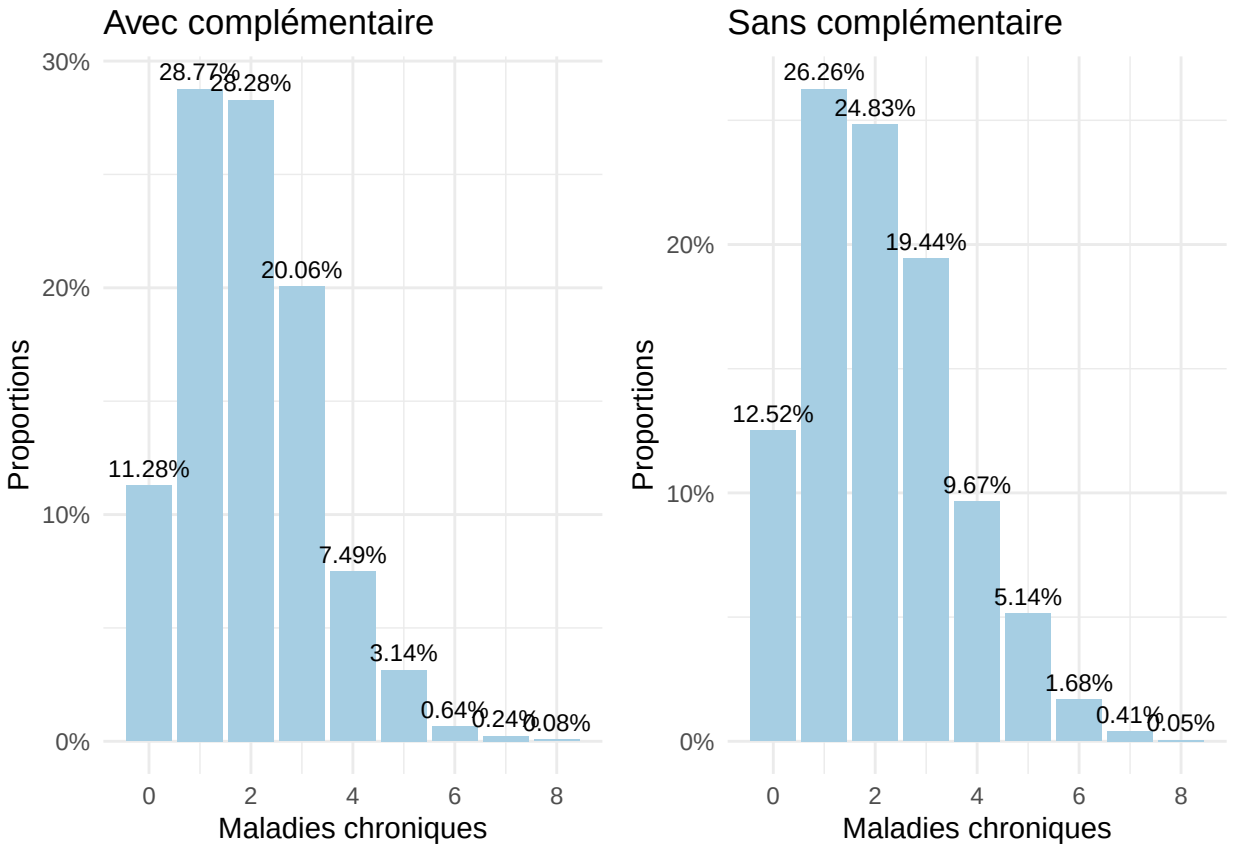
On observe très bien que les individus dont le conjoint n'est pas retraité sont les plus représentés. En effet, chez les individus ayant une complémentaire, 56 % ont un conjoint non retraité contre 64 % chez les individus n'ayant pas de complémentaire.



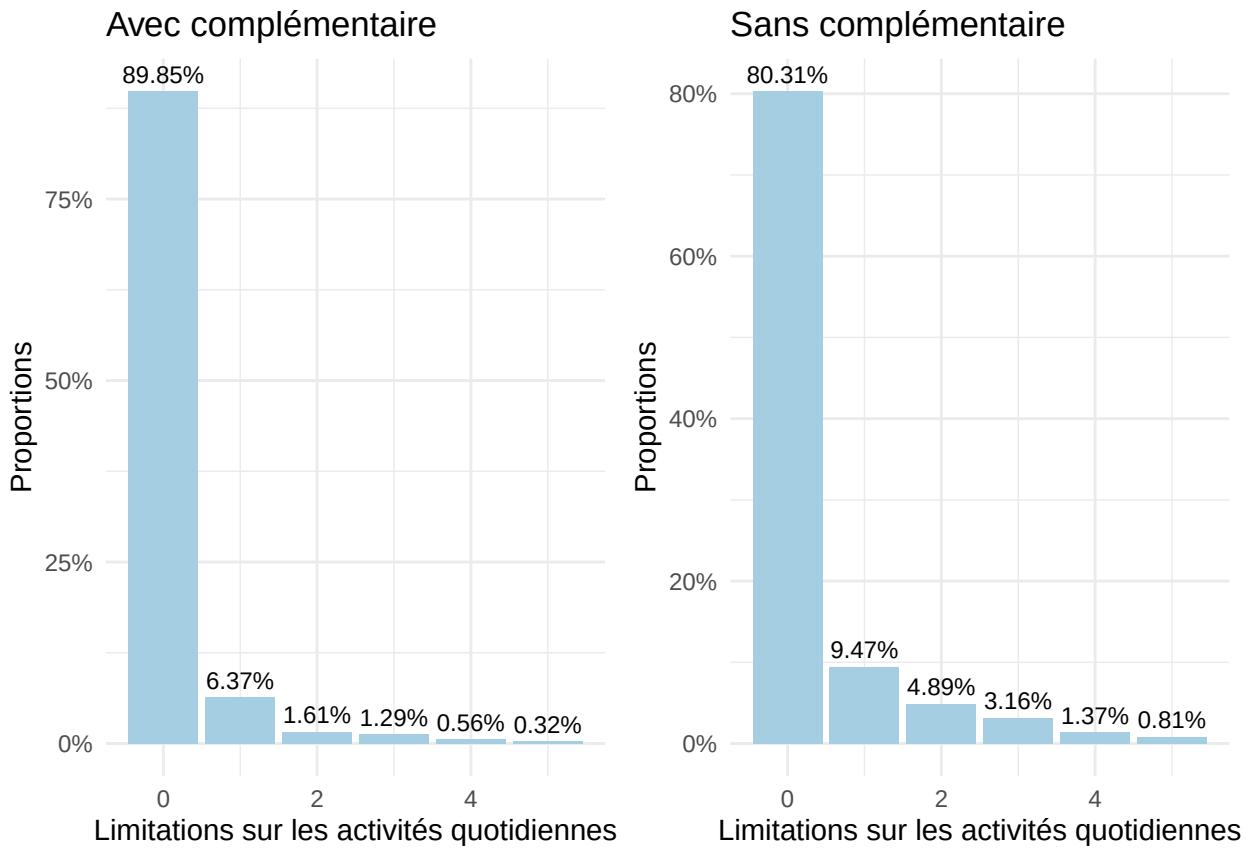
Que ce soit chez les individus ayant une complémentaire ou ceux n'en n'ayant pas, les individus retraités sont les plus présents. En effet, ils représentent 67 % des personnes ayant une complémentaire et 59 % de ceux n'en ayant pas.



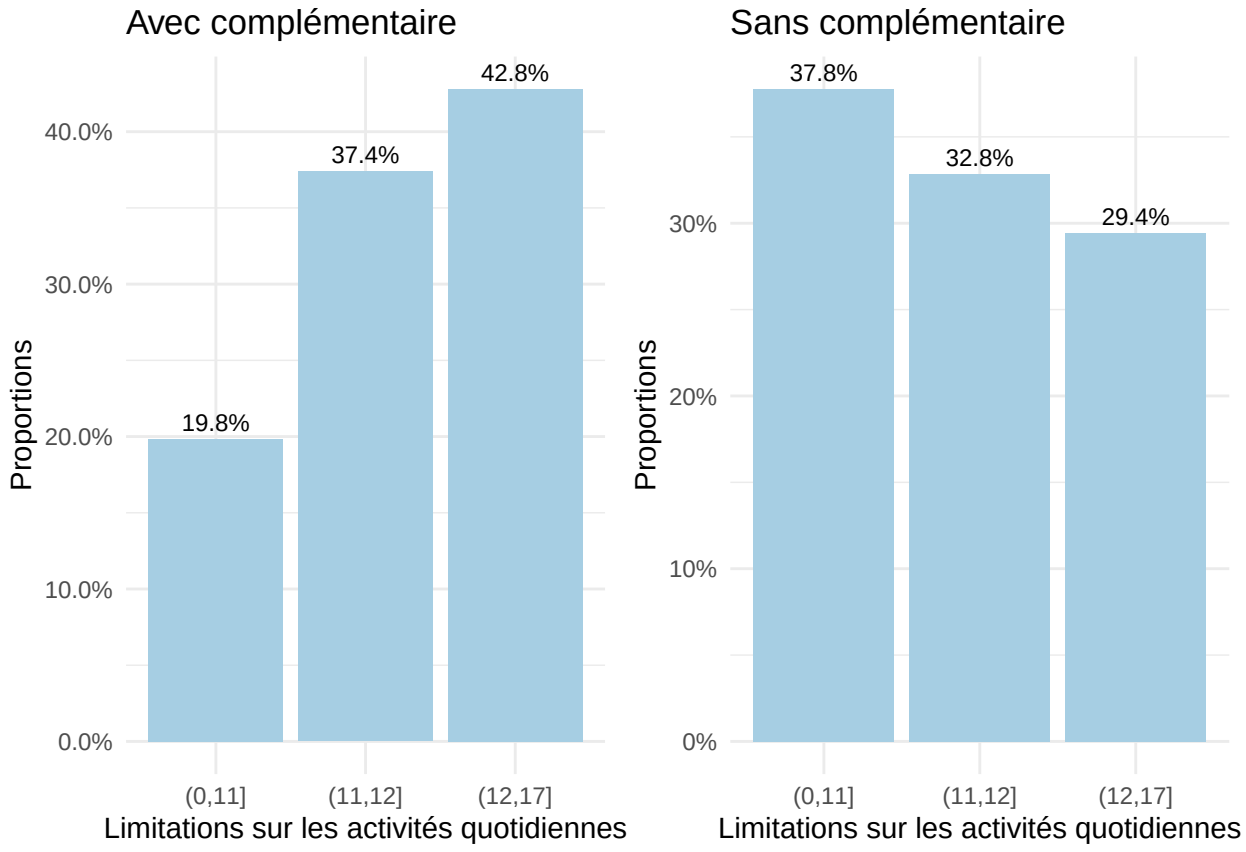
Les individus ayant une complémentaire ont globalement un bon etat de santé. Cepenent, ont voit que chez ceux n'ayant pas de complémentaire, il y a beaucoup plus d'individus ayant un etat de santé mauvais et moyen. De même, on voit que les proportions de personnes ayant un bon, très bon ou excellent etat de santé sont infèrieures par rapport aux mêmes modalités concernant les individus ayant une complémentaire.



Chez les individus ayant une complémentaire, on voit qu'il y a 27 % qui ont 1 maladie chronique, 28 % qui en ont 2 et 20.06 % qui en ont 3. 11 % d'entre eux n'ont pas de maladie chronique. Chez les individus n'ayant pas de complémentaire, 26 % ont 1 maladie chronique, 24 % qui en ont 2 et 19.44 % qui en ont 3. 12.52 % d'entre eux n'ont pas de de maladie chronqiues.

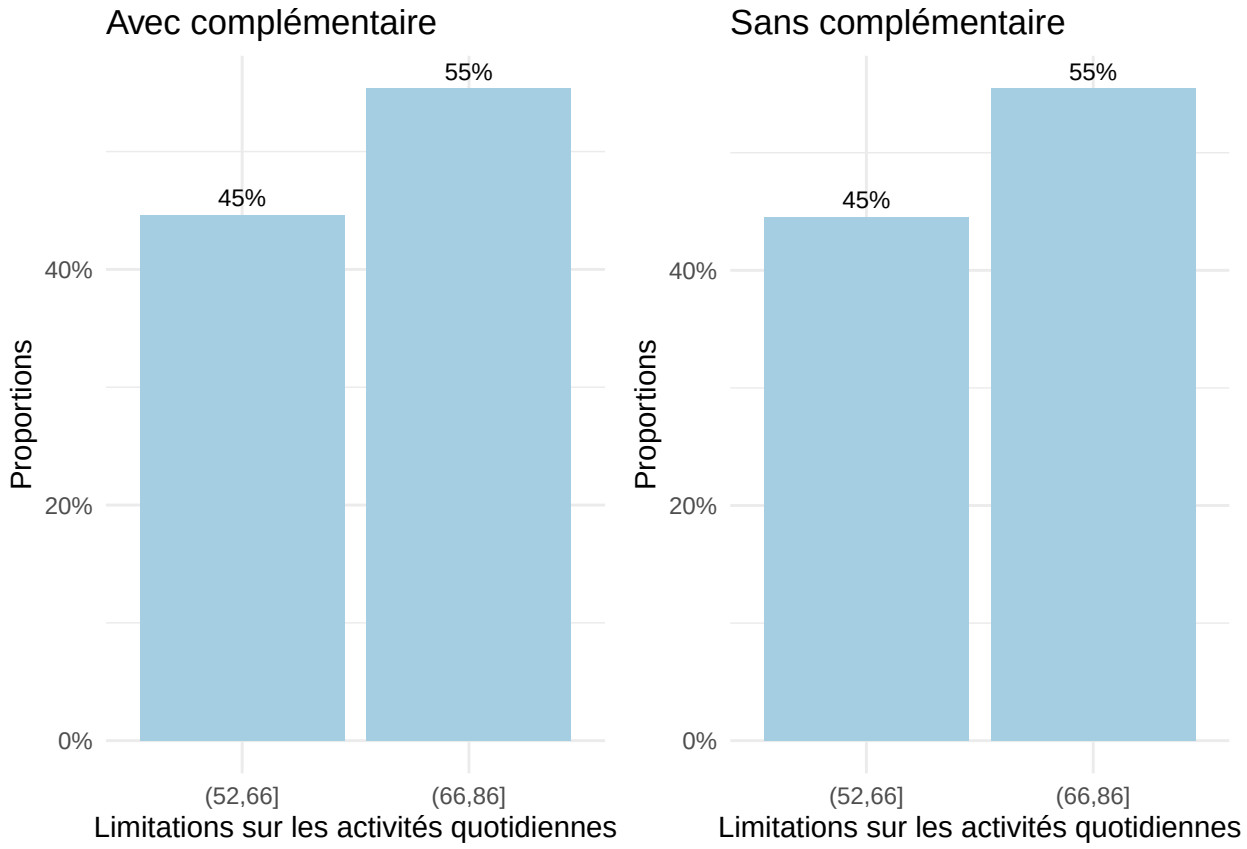


Que ce soit chez les individus ayant une complémentaire ou ceux n'en ayant pas, la majorité des individus n'a pas de limitation sur les activités quotidiennes.

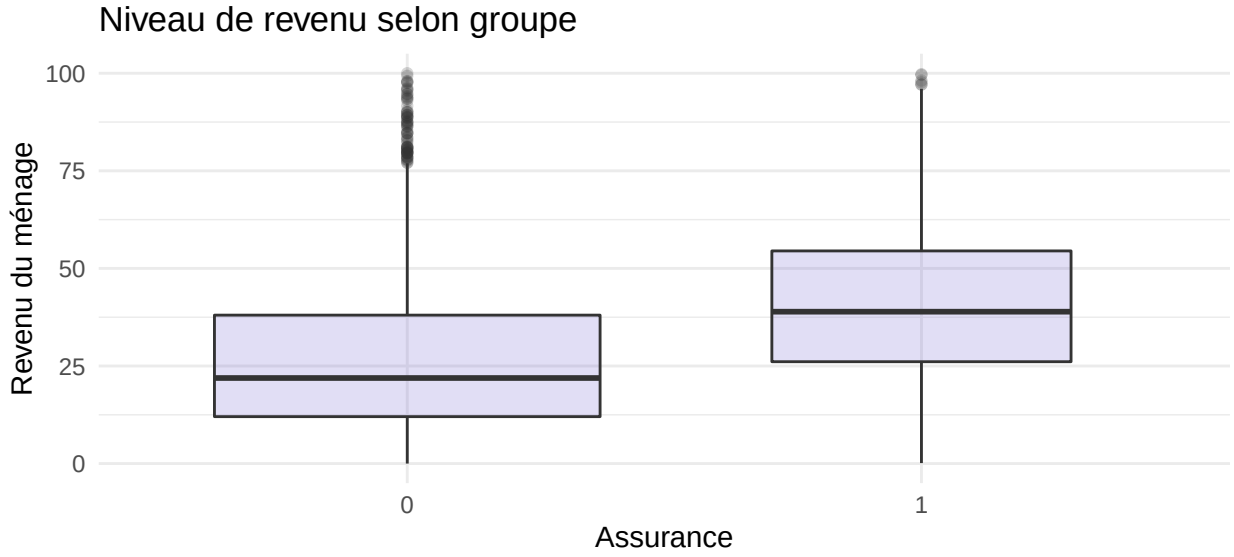


Chez les individus ayant une complémentaire 42.8 % ont effectué 13 à 17 années d'etudes contre 37.8 % chez ceux n'ayant pas de complémentaire. 37.4 % des individus ayant une complémentaire ont effectué 12 années d'études contre 32.8 % chez

ceux n'ayant pas de complémentaire. Enfin, 19.8 % des individus ayant une complémentaire ont effectué 11 années d'etude contre 29.4 %.

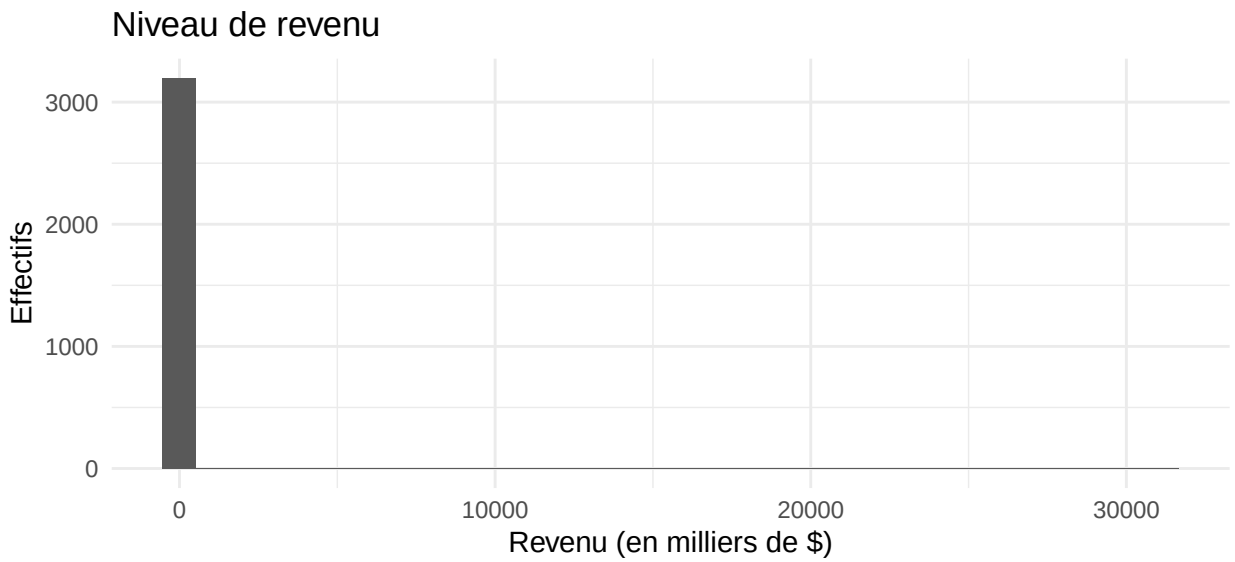


Etant donné que Medicare est attribué aux personnes ayant plus de 65, nous avons décidé de découper cette varibles en 2 classes. On observe que les chez les individus n'ayant pas complémentaire, 55 % ont entre 67 ans et 86 ans. et 45 % ont un age compris entre 52 et 66 ans.

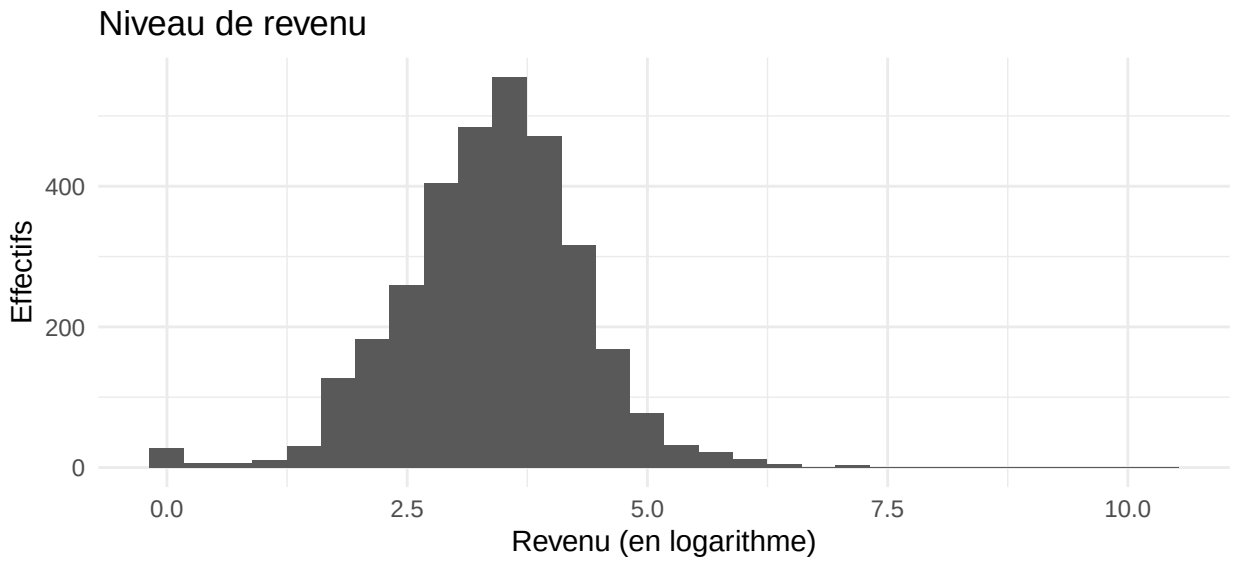


On voit très bien qu'il y a une différence de revenus entre les individus ayant une complémentaire et ceux n'en ayant pas. Globalement, les individus ayant une complémentaire ont un revenu plus élevé que ceux n'en ayant pas.

Une autre représentation graphique de la variable revenu nous permet d'observer qu'il y a une dissymétrique de la distribution de cette variable.



En appliquant le logarithme, la variable revenu suit à présent une distribution normale.



Les caractéristiques des individus ayant une complémentaire sont les suivantes :

- Non - hispanique ;
- Blanc ;
- Homme ;
- Marié ;
- Conjoint non-retraité ;
- Retraité ;
- Etat de santé : Bon, Très bon ;
- Maladies chroniques : 1 à 4 ;
- Nombre de limiations sur les activités du quotidien : Majoritaires aucunes ;
- Ont entre 66 et 86 ans ;
- Ont un niveau d'etude supérieur à celui de ceux n'ayant de complémentaire ;
- Ont un niveau de revenu plus élevé que ceux n'ayant pas de complémentaire ;

Les caractéristiques des individus n'ayant de complémentaire sont les suivantes :

- Non - hispanique ;
- Blanc ;
- femme ;
- Marié(e) ;
- Conjoint non-retraité ;
- Retraitée ;
- Etat de santé : Moyen, Mauvais ;
- Maladies chroniques : 1 à 4 ;
- Nombre de limitations sur les activtés du quotidien : Majoritairment aucunes ;
- Ont un niveau d'etude infèrieur à celui de ceux ayant de complémentaire ;
- Ont entre 66 et 86 ans ;
- Ont un niveau de revenu inférieur à ceux ayant une de complémentaire ;

3 Modèle : Modèle général et fonction de vraisemblance

3.1 Modèle général

Les modèles à choix binaires ont pour objectif de modéliser la probabilité de succès d'un évènement conditionnellement aux caractéristiques des individus. Dans notre cas, la variable binaire à expliquer que nous allons considérer est la variable assurance. Nous allons tenter d'expliquer la probabilté qu'un individu ait une complémentaire.

$$\forall i \in [1, N] \ , \ Y_i = \begin{cases} 1 \ si \ l'individu \ a \ une \ complémentaire \\ 0 \ sinon \end{cases}$$

Pour un modèle général à choix binaire, nous avons :

$$\begin{aligned} \mathbb{P}_i = \mathbb{P}(Y_i = 1 \mid X_i) = \mathbb{P}(Y_i^* > c) = \mathbb{P}(X_i\beta + \epsilon_i > c) \\ = \mathbb{P}(\frac{\epsilon_i}{\sigma} < -\frac{c - X_i\beta}{\sigma}) = \mathbb{P}_i = F(\frac{-c - X_i\beta}{\sigma}) \\ \mathbb{P}(Y_i = 1) = 1 - F(\frac{-c - X_i\beta}{\sigma}) \end{aligned}$$

Pour des raison d'identification du modèle on pose $c = 0$. Dans ce cas :

$$P(Y_i = 1) = 1 - F(\frac{-X_i\beta}{\sigma})$$

$$\begin{aligned} \mathbb{P}(Y_i = 1) = 1 - F(\frac{-X_i\beta}{\sigma}) \\ = 1 - (1 - \mathbb{F}(\frac{-X_i\beta}{\sigma})) \\ \mathbb{P}(Y_i = 1) = \mathbb{F}(\frac{X_i\beta}{\sigma}) \end{aligned}$$

Encore pour des raisons d'identification de notre modèle, on pose $\sigma = 1$ dans le cadre d'une estimation Probit et $\sigma = \frac{\pi^2}{3}$ dans le cadre d'une estimation logit. Ainsi, nous obtenons :

$$\mathbb{P}_i \ = \mathbb{P}(Y_i = 1) = \mathbb{F}(X_i\beta)$$

La fonction de densité de Y_i est donc :

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

avec Y_i = 0,1, et $P_i = F(X_i\beta)$.

3.2 Maximum de vraimsemblance

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

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

Pour tout N :

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

La maximisation de la quantité $ln \ L(Y_i, \beta)$ par le biais de l'algorithme de Newton-Raphson nous permet de converger vers des solutions plausibles à notre problème.

4 Estimations (logit, probit)

Tout d'abord, nous avons séparer notre base de données en un échantillon d'apprentissage, sur lequel nous allons effectuer nos estimations et en un échantillon test sur lequel nous allons attester de la qualité des estimations effectuées sur les données d'apprentissage. Au vue du travail réalisé en amont lors de l'analyse descriptive, nous réalisons des premières estimations dans lesquelles nous incluons toutes les variables. De plus, nous décidons de considérer la variable lnrevenu plutôt que la variable revenu. Par ailleurs, nous avons décidér de découper en 2 classes la variable maladie chronique. 1 regrouprant les individus n'ayant pas de maladie chronique et une autre regroupant ceux ayant des maladies chroniques. De même concernant la variable concernant le nombre de limitation sur les activités du quotidien. Elle est à présent découpée en une classe regroupant les individus n'ayant aucune limitation sur les activités du quotidien et ceux ayant des limitations.

$$\mathbb{P}(Y_i = 1) = \mathbb{F}(\beta_0 + \beta_1 agecl_i + \beta_2 hispanique_i + \beta_3 blanc_i + \beta_4 femme_i + \beta_5 education_i + \beta_6 statut\_matrimonial_i + \beta_7 maladies\_chroniques_i + \beta_8 adl_i + \beta_9 retraité_i + \beta_{10} statut\_du\_conjoint_i + \beta_{11} lnrevenu_i + \beta_{12} etat\_de\_sante_i)$$

La fonction de distribution F dans le cadre d'un logit est :

$$F(X_i\beta) = \Lambda(X_i\beta) = \frac{exp(X_i\beta)}{1 + exp(X_i\beta)}$$

La fonction de distribution dans le cadre d'un probit est

$$F(X_i\beta) = \Phi(X_i\beta) = \int_{-\infty}^{X_i\beta} \frac{2}{\sqrt{2\pi}} exp(-z^2/2)dz$$

Table 1 –

	Dependent variable:	
	assurance	
	logistic	*probit*
	(1)	(2)
age(66,86]	−0.220**	−0.134**
	(0.101)	(0.061)
hisp1	−0.747***	−0.452***
	(0.243)	(0.137)
blanc1	−0.045	−0.022
	(0.137)	(0.082)
femme1	−0.104	−0.067
	(0.108)	(0.066)
educ(11,12]	0.457***	0.283***
	(0.128)	(0.077)
educ(12,17]	0.293**	0.183**
	(0.136)	(0.082)
stat_mat1	0.234	0.144
	(0.152)	(0.091)
maladiec[1,8]	0.407***	0.246***
	(0.156)	(0.094)
adl[1,5]	−0.289*	−0.178*
	(0.160)	(0.096)
retraite1	0.221**	0.132**
	(0.107)	(0.065)
stat_conj1	−0.089	−0.053
	(0.113)	(0.069)
etat_santeMoyen	0.026	0.024
	(0.221)	(0.131)
etat_santeBon	0.039	0.034
	(0.223)	(0.133)
etat_santeTrés_bon	0.076	0.058
	(0.231)	(0.138)
etat_santeExcellent	0.120	0.085
	(0.260)	(0.156)
lnrevenu	0.688***	0.418***
	(0.075)	(0.044)
Constant	−3.512***	−2.154***
	(0.355)	(0.210)
Observations	2,138	2,138
Log Likelihood	−1,283.424	−1,281.033
Akaike Inf. Crit.	2,600.848	2,596.066
Note:	*p<0.1; **p<0.05; ***p<0.01	

5 Tests de spécification et hypothèses

5.1 Méthode pas à pas descendante

Nous décidons à présent d'utiliser un critère de sélection des variables explicatives pour nos modèles dont le but est de minimiser le Critère d'Information d'Akaike. En régression logistique et probit, on cherche à maximiser le logarithme de la vraisemblance L. Lorqu'on ajoute des variables au modèle, on peut augmenter la vraisemblance. L'AIC est pénalisé par le nombre de variables explicatives inclues dans le modèle, satisfaisant le critère de parcimonie. Il se calcule, pour k variables explicatives :

$$AIC = 2k - 2ln(L)$$

Nous utilisons la fonction **stepAIC** qui permet d'appliquer la méthode de pas à pas descendante qui prend en compte l'AIC. Le déroulement de cette procédure se trouve en annexe.

Le modèle contenant des variables pertinentes selon le critère de l'AIC pour les estimations logit et probit est le suivant :

$$\begin{aligned}\mathbb{P}(Y_i = 1) = \mathbb{F}(\beta_0 + \ \beta_1 age(66, 86] \ + \beta_2 hispanique_i + \beta_3 education(11, 12]_i + \beta_4 education(12, 17] \\ + \beta_5 statut\_matrimonial + \beta_6 maladie\_chroniques[1, 8]_i + \\ \beta_7 adl[1, 5]_i + \beta_8 retraite_i + \beta_9 lnrevenu_i)\end{aligned}$$

Les estimations Logit et Probit de ces modèles sont présentées ci-dessous :

TABLE 2 –

	Dependent variable:	
	assurance	
	logistic	*probit*
	(1)	(2)
age(66,86]	−0.209**	−0.127**
	(0.099)	(0.060)
hisp1	−0.735***	−0.443***
	(0.242)	(0.137)
educ(11,12]	0.441***	0.274***
	(0.124)	(0.075)
educ(12,17]	0.285**	0.181**
	(0.133)	(0.081)
stat_mat1	0.210	0.132*
	(0.130)	(0.078)
maladiec[1,8]	0.387**	0.232**
	(0.151)	(0.092)
adl[1,5]	−0.312**	−0.195**
	(0.148)	(0.088)
retraite1	0.236**	0.143**
	(0.104)	(0.063)
lnrevenu	0.697***	0.424***
	(0.073)	(0.043)
Constant	−3.581***	−2.188***
	(0.288)	(0.169)
Observations	2,138	2,138
Log Likelihood	−1,284.693	−1,282.376
Akaike Inf. Crit.	2,589.387	2,584.752
Note:	*p<0.1; **p<0.05; ***p<0.01	

5.2 Tests de significativité globale

Vérifions à présent la significativé globale de nos modèles obtenus grâce à la méthode de pas à pas descendante. Pour cela, on réalise un test du rapport de vraisemblance.

5.2.1 Modèle logit :

Considérons le modèle suivant :

$$\begin{aligned}\mathbb{P}(Y_i = 1) = \Lambda(\beta_0 + \ \beta_1 age(66, 86] \ + \beta_2 hispanique_i + \beta_3 education(11, 12]_i + \beta_4 education(12, 17] \\ + \beta_5 statut\_matrimonial + \beta_6 maladie\_chroniques[1, 8]_i + \\ \beta_7 adl[1, 5]_i + \beta_8 retraite_i + \beta_9 lnrevenu_i)\end{aligned}$$

$$\begin{aligned}H_0 &= \exists\, j\{1, 2, 3, 4, 5, 6, 7, 8, 9\} \in \ \beta j \ = \ 0 \\ H_1 &= \exists\, j\{1, 2, 3, 4, 5, 6, 7, 8, 9\} \in \ \beta j \ \neq \ 0\end{aligned}$$

Soient L_{nc} la vraisemblance du modèle complet et L_c la vraisemblance du modèle contraint (avec tous les coefficents égaux à 0). La statistique de test est :

$$LR \ = \ 2[ln(L_{nc}) - ln(L_c)] \ \rightarrow \chi^2_9$$

La $p - value$ associée à cette statistique est $ 6.471563e^{-56}$. Elle est inférieure au seuil de 5%. Par conséquent, nous pouvons affirmer que le modèle est globalement significatif.

5.2.2 Modèle Probit :

Considérons le modèle probit suivant :

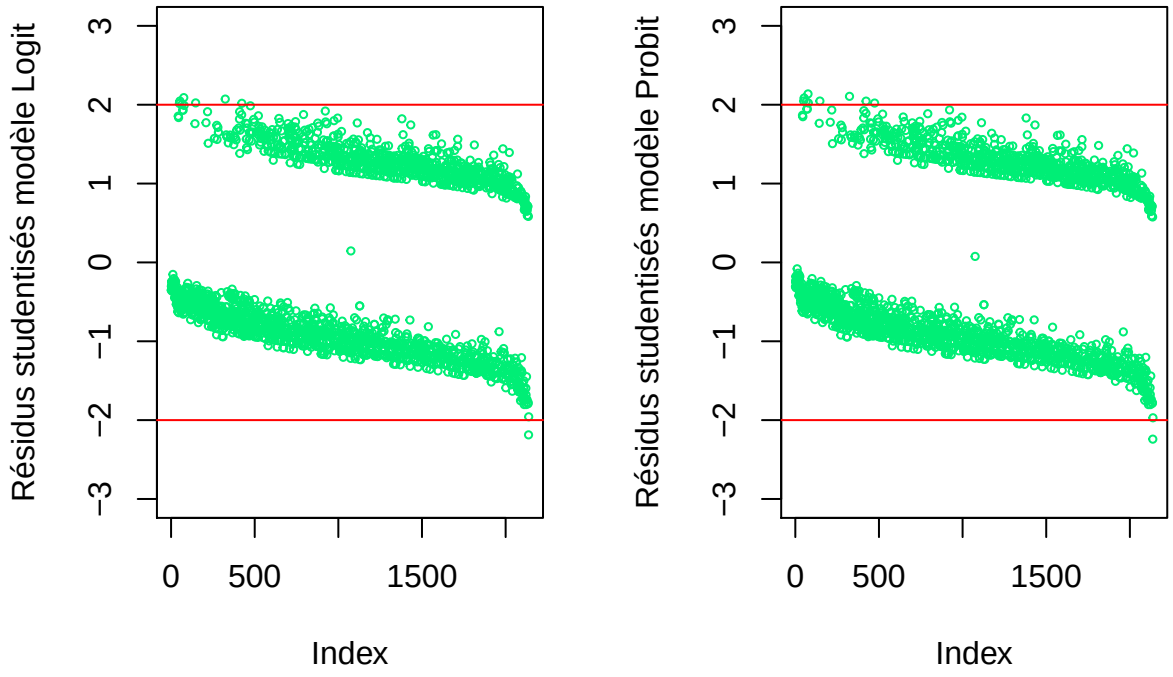
$$\begin{aligned}\mathbb{P}(Y_i = 1) = \Phi(\beta_0 + \ \beta_1 age(66, 86] \ + \beta_2 hispanique_i \ + \beta_3 education(11, 12]_i + \beta_4 education(12, 17] \\ + \beta_5 statut\_matrimonial + \beta_6 maladie\_chroniques[1, 8]_i + \\ \beta_7 adl[1, 5]_i + \beta_8 retraite_i + \beta_9 lnrevenu_i)\end{aligned}$$

$$\begin{aligned}H_0 &= \exists\, j\{1, 2, 3, 4, 5, 6, 7, 8, 9\} \in \ \beta j \ = \ 0 \\ H_1 &= \exists\, j\{1, 2, 3, 4, 5, 6, 7, 8, 9\} \in \ \beta j \ \neq \ 0\end{aligned}$$

$$LR \ = \ 2[ln(L_{nc}) - ln(L_c)] \ \rightarrow \chi^2_9$$

La $p - value$ associée à cette statistique est $6.746039e^{-57}$. Elle est inférieure au seuil de 5%. Par conséquent, nous pouvons affirmer que le modèle est globalement significatif.

On peut également regarder les résidus standardisés de nos modèles :



Il est important de faire diagnostique la régression afin de valider ou non le modèle. L'analyse des résidus est de ce point de vue très importante. En régression logistique, on s'intéresse la plupart du temps aux résidus de déviance. Ils prennent généralement les valeurs qui oscillent entre -2 et 2. On voit très bien qu'il n'y presque pas de valeurs aberrantes, c'est-à-dire au delà des limites.

6 Comparaison des modèles :

Lorsqu'on inclue toutes les variables, on voit que pour les estimations logit (1), et probit (2) un grand nombre des coefficients associés aux variables ne sont pas significatifs. L'intérêt est donc de pouvoir trouver un modèle contenant des variables avec un pouvoir explicatif. Nous avons donc raffiné ce modèle grâce à la méthode de pas à pas descendante. Les modèles obtenus grâce à cette méthode sont nettement mieux. L'ensemble des coefficients associés aux variables du modèle sont significatifs, le critère d'AIC baisse. Cependant, comme le montre les estimations logit (3) et probit (4), il semble que le coefficient associé à la variable statut matrimonial n'est pas significatif dans l'estimation logit (3). Nous avons alors effectué une autre régression logistique [logit (5)] dans laquelle nous avons enlever la variable statut matrimonial et on voit que cela entraine une augmentation de l'AIC du modèle. De même, on voit que le coefficient associé à la variable educ(12,17] baisse en significativité. Il semble donc que l'estimation logit (3) soit à préférer.

Ainsi, si l'on considère les estimations logit (3) et probit (4), il semble que les coefficients associés aux variables explicatives aient les mêmes signes ainsi que les mêmes niveaux de significativité. Sauf pour statut matrimonial. On constate que la valeur des coefficients associés aux variables explicatives du modèle probit est légèrement inférieure à celle des coefficients du modèle logit. De plus, si l'on considère le Critère d'Information d'Akaike, on voit très bien que celui du modèle probit est à préferer par rapport à celui du logit. Néanmoins, pour des raisons d'interprétabilité des résultats nous allons conserver uniquement le modèle logit (3).

Ainsi, le modèle que nous allons considérer est le suivant :

$$\begin{aligned}\mathbb{P}(Y_i = 1) = \Lambda(\beta_0 + \ \beta_1 age(66,86] \ + \beta_2 hispanique_i + \beta_3 education(11,12]_i + \beta_4 education(12,17] \\ + \beta_5 statut\_matrimonial + \beta_6 maladie\_chroniques[1,8]_i + \\ \beta_7 adl[1,5]_i + \beta_8 retraite_i + \beta_9 lnrevenu_i)\end{aligned}$$

TABLE 3 –

	Dependent variable:				
	assurance				
	logistic	*probit*	*logistic*	*probit*	*logistic*
	(1)	(2)	(3)	(4)	(5)
age(66,86]	−0.220**	−0.134**	−0.209**	−0.127**	−0.198**
	(0.101)	(0.061)	(0.099)	(0.060)	(0.099)
hisp1	−0.747***	−0.452***	−0.735***	−0.443***	−0.721***
	(0.243)	(0.137)	(0.242)	(0.137)	(0.242)
blanc1	−0.045	−0.022			
	(0.137)	(0.082)			
femme1	−0.104	−0.067			
	(0.108)	(0.066)			
educ(11,12]	0.457***	0.283***	0.441***	0.274***	0.428***
	(0.128)	(0.077)	(0.124)	(0.075)	(0.124)
educ(12,17]	0.293**	0.183**	0.285**	0.181**	0.258*
	(0.136)	(0.082)	(0.133)	(0.081)	(0.132)
stat_mat1	0.234	0.144	0.210	0.132*	
	(0.152)	(0.091)	(0.130)	(0.078)	
maladiec[1,8]	0.407***	0.246***	0.387**	0.232**	0.398***
	(0.156)	(0.094)	(0.151)	(0.092)	(0.151)
adl[1,5]	−0.289*	−0.178*	−0.312**	−0.195**	−0.299**
	(0.160)	(0.096)	(0.148)	(0.088)	(0.147)
retraite1	0.221**	0.132**	0.236**	0.143**	0.253**
	(0.107)	(0.065)	(0.104)	(0.063)	(0.103)
stat_conj1	−0.089	−0.053			
	(0.113)	(0.069)			
etat_santeMoyen	0.026	0.024			
	(0.221)	(0.131)			
etat_santeBon	0.039	0.034			
	(0.223)	(0.133)			
etat_santeTrés_bon	0.076	0.058			
	(0.231)	(0.138)			
etat_santeExcellent	0.120	0.085			
	(0.260)	(0.156)			
lnrevenu	0.688***	0.418***	0.697***	0.424***	0.746***
	(0.075)	(0.044)	(0.073)	(0.043)	(0.067)
Constant	−3.512***	−2.154***	−3.581***	−2.188***	−3.601***
	(0.355)	(0.210)	(0.288)	(0.169)	(0.288)
Observations	2,138	2,138	2,138	2,138	2,138
Log Likelihood	−1,283.424	−1,281.033	−1,284.693	−1,282.376	−1,286.008
Akaike Inf. Crit.	2,600.848	2,596.066	2,589.387	2,584.752	2,590.016

Note: *p<0.1; **p<0.05; ***p<0.01

7 Qualité de l'ajustement et prédiction

7.1 Matrice de confusion

Une manière d'évaluer la qualité du modèle serait de confronter les vrais valeurs prédites avec les vraies valeurs prises par Y. C'est pourquoi nous décidons d'effectuer une matrice de confusion. A partir de celle-ci, nous allons déduire des indicateurs qui nous permettront effectivement d'évaluer la qualité de notre modèle.

Les prédictions correctes sont telles que :

$$Y_i \ = 1 \ si \ \mathbb{F}(X_i\beta) > 0.5 \ ; \ 0 \ sinon.$$

7.1.1 Matrices de confusion

TABLE 4 – Matrice de confusion sur les données d'entrainement

	Réalité		
	0	1	Sum
0	1046	487	1533
1	265	340	605
Sum	1311	827	2138

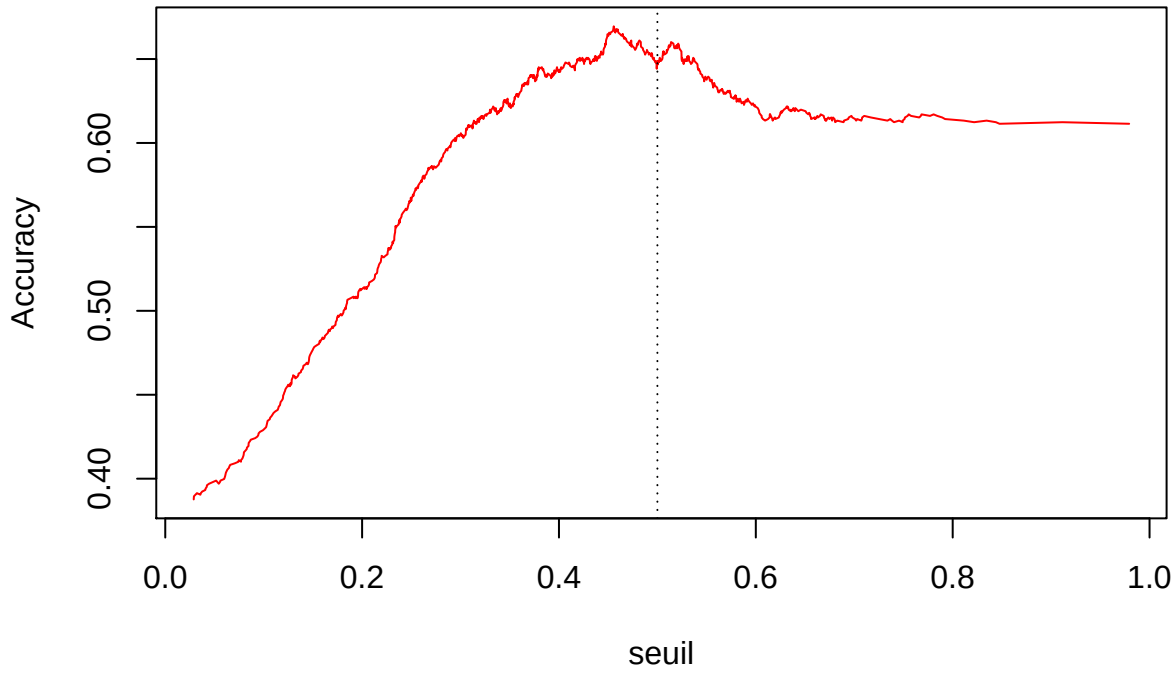
TABLE 5 – Matrice de confusion sue les données tests

	Réalité		
	0	1	Sum
0	514	237	751
1	140	177	317
Sum	654	414	1068

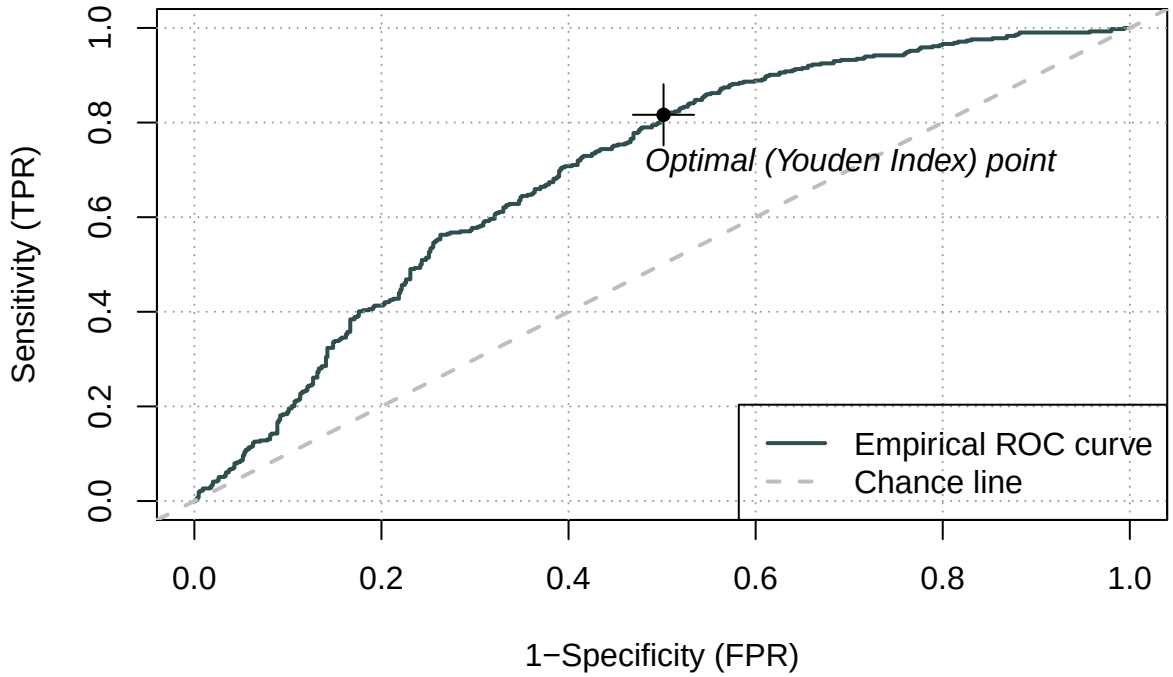
7.2 Indicateurs de performance

- R^2 = 0.54 (entrainement) | 0.54 (test)
- Sensivité (Vrais positifs) = 41 % (entrainement) | 42 % (test)
- Spécificité (Vrais négatifs) = 79 % (entrainement) | 78 % (test)
- Taux de faux positifs = 37 % (entrainement) | 36 % (test)
- Taux de faux négatifs = 32 % (entraiement) | 33 % (test)
- Taux d'erreur globale = 35 % (entraiement) | 35 % (test)

Au vue des résultats que nous avons obtenus, on peut tout d'abord noter que les résultats en apprentissage et en test ne sont pas très éloignés. Le R^2 est de 0.54. Donc notre modèle n'est pas parfait mais il fait mieux que le classifieur par défaut. Le taux d'erreur global n'est pas élevé. Il est de 35 %, que ce soit en test ou en apprentissage. Donc le taux de succès de notre modèle est de 65 %. De plus, le taux de prédiction positives est de 42 % en test et de 41 % en apprentissage. Cependant, il prédit à 78 % les vrais négatifs sur les données tests et à 79 % sur les données d'apprentissage. Donc il prédit, mieux les individus qui n'ont pas de complémentaire.



On voit qu'au seuil de 0.5, la précision de notre modèle est de 65 %.



L'AUC est de 0.7 donc notre classifieur fait mieux que le hasard.

8 Effets marginaux

Rappelons les résultats resultats du modèle :

Table 6 –

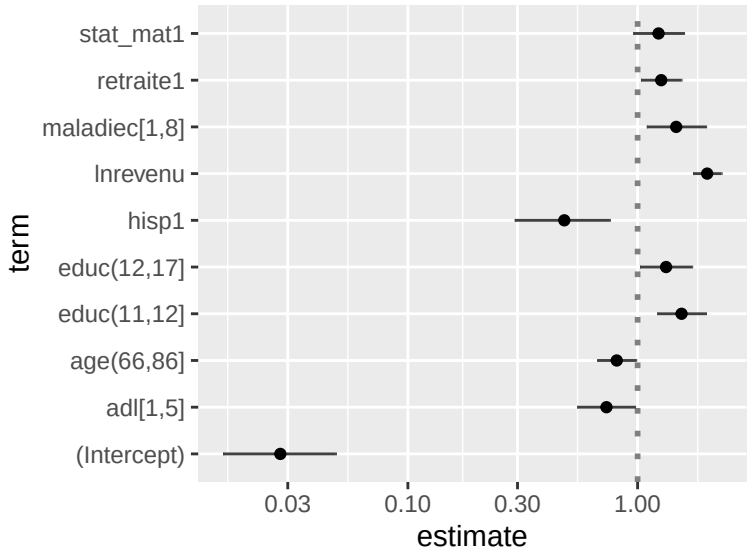
	Dependent variable:
	assurance
age(66,86]	-0.209^{**}
	(0.099)
hisp1	-0.735^{***}
	(0.242)
educ(11,12]	0.441^{***}
	(0.124)
educ(12,17]	0.285^{**}
	(0.133)
stat_mat1	0.210
	(0.130)
maladiec[1,8]	0.387^{**}
	(0.151)
adl[1,5]	-0.312^{**}
	(0.148)
retraite1	0.236^{**}
	(0.104)
lnrevenu	0.697^{***}
	(0.073)
Constant	-3.581^{***}
	(0.288)
Observations	2,138
Log Likelihood	−1,284.693
Akaike Inf. Crit.	2,589.387
Note:	*p<0.1; **p<0.05; ***p<0.01

8.1 Odds Ratio :

Table 7 –

	OR	2.5 %	97.5 %
(Intercept)	0.028	0.016	0.049
age(66,86]	0.811	0.667	0.985
hisp1	0.480	0.292	0.757
educ(11,12]	1.554	1.218	1.984
educ(12,17]	1.330	1.025	1.727
stat_mat1	1.234	0.957	1.593
maladiec[1,8]	1.472	1.097	1.987
adl[1,5]	0.732	0.546	0.975
retraite1	1.267	1.034	1.553
lnrevenu	2.008	1.743	2.321

On peut observer graphiqument les odds ratio de nos coefficients.



Les variables qui indiquent qu'il y a plus de chances que l'individu ait une complémentaire sont les suivantes :

- retraité,
- Maladiec[1,8]
- educ(11,12]
- educ(12,17]
- lnrevenu

Les variables qui indiquent que l'individu ait moins de chance d'avoir une complémentaire sont les suivantes :

- Hispanique
- retraité,
- Maladiec[1,8]
- age(66,86]
- adl[1,5]

8.2 Effets marginaux moyens

Les effets marginaux se calculent comme suit :

$$\frac{\Delta P_i}{\Delta x_{ik}} = \frac{\Delta F(X_i\beta)}{\Delta x_{ik}} = f(X_i\beta)\beta_k$$

Dans le cadre du modèle logit, ils se calculent comme suit :

$$\frac{\Delta P_i}{\Delta x_{ik}} = \frac{exp(X_i\beta)}{[1 + exp(X_i\beta)]^2}\beta_k$$

A partir de ces formules nous pouvons en déduire celle des effets marginaux moyens :

$$AME_k = \beta_k \frac{1}{n}\sum_{i=1}^{n} f(X_i\beta)$$

TABLE 8 – Effets marginaux moyens

factor	AME	SE	z	p	lower	upper
adl[1,5]	-0.0639223	0.0295399	-2.163935	0.0304693	-0.1218194	-0.0060253
age(66,86]	-0.0435317	0.0206082	-2.112350	0.0346564	-0.0839230	-0.0031404
educ(11,12]	0.0922134	0.0259708	3.550651	0.0003843	0.0413115	0.1431153
educ(12,17]	0.0588834	0.0276203	2.131887	0.0330162	0.0047486	0.1130183
hisp1	-0.1426496	0.0423028	-3.372106	0.0007460	-0.2255616	-0.0597376
lnrevenu	0.1449280	0.0139549	10.385471	0.0000000	0.1175769	0.1722790
maladiec[1,8]	0.0780693	0.0294121	2.654320	0.0079468	0.0204225	0.1357160
retraite1	0.0490218	0.0213566	2.295391	0.0217107	0.0071636	0.0908801
stat_mat1	0.0435611	0.0268114	1.624720	0.1042221	-0.0089884	0.0961105

9 Interprétations

— En moyenne, la probabilité que les personnes ayant entre 1 et 5 limitations sur les activités du quotidien ait une complémentaire est de 6.3 points de pourcentage infèrieure à ceux qui n'ont en pas.
— En moyenne, la probabilité que les individus qui ont entre 66 et 86 ans ait une complémentaire est de 4.3 points de pourcentage infèrieure à ceux qui on un âge compris entre 52 et 65 ans.
— En moyenne, la probabilité que les individus ayant effectué 12 années d'etude ait une complémentaire est de 9.2 points de pourcentage supérieure à ceux qui en ont fait moins.
— En moyenne, la probabilité que les individus ayant effectué entre 13 à 17 ans années d'etude ait une complémentaire est de 5.8 points de pourcentage supérieure à ceux qui en ont fait moins.
— En moyenne, la probabilité que les individus hispaniques ait une complémentaire est de 14.26 points de pourcentage infèrieure à ceux qui ne sont pas hispaniques.
— En moyenne, la probabilité d'avoir une complémentaire augmente de 1.5 points de pourcentage compte tenu d'une augmentage du revenu de 10 %.
— En moyenne, la probabilité que les individus qui ont entre 1 à 8 maladies chroniques ait une complémentaire est de presque 8 points de pourcentage supérieure à ceux qui n'en ont pas.
— En moyenne, la probabilité que les retraité(e)s ait une complémentaire est de presque 5 points de pourcentage supérieure à ceux qui ne le sont pas pas.

10 Discussion

Rappeleons que les individus peuvent recevoir les aides de Medicare s'ils ont (eux ou leurs conjoints) cotisé au moins dix ans, sont âgés de plus de 65 ans et habitent de manière permanente aux États-Unis. De plus, les individus ayant moins de 65 ans peuvent aussi être éligibles à condition d'être handicapés ou d'être au stade final d'une maladie rénale.

Au vue de l'intérprétation des effets marginaux moyens de notre modèle nous pouvons dire plusieurs choses. Tout d'abord, on peut voir que pour les individus d'origine hispaniques, il est en moyenne peu probable d'avoir une complémentaire. Cela s'explique peut-être par ce que nous avons observé lors de notre phase de statistiques descriptives. En effet, on a vu qu'en majorité les bénéficiaires des services Medicare sont des individus, blancs et non-hispaniques. En moyenne, les individus ayant effectué au moins 12 années d'étude ont de forte chance d'avoir une complémentaire. En effet, cela sous entend que ce sont des individus instruits. Par conséquent, ils appréhendent sans doute mieux ce qu'est une complémentaire et ses prérogatives. Les personnes retraitées ont en moyenne de forte chance d'avoir une complémentaire santé. Cela paraît logique étant donné qu'à ce moment de la vie il est possible d'avoir certaines pathologies dont les soins ne sont pas pris en charge par Medicare tels que les appareils auditifs ou encore les soins à domicile. Par ailleurs, les individus ayant entre 1 à 8 maladies chroniques ont en moyenne de fortes chance d'avoir une complémentaire. Cela peut s'expliquer par le fait qu'ils aient besoin de services qui ne sont pas couverts par Medicare. De plus, on voit que le revenu du ménage joue un rôle dans la souscription à une assurance complémantaire. En effet, le coût de souscription à une complémentaire aux les Etats-Unis n'est pas non-négligeable. Donc les personnes à faibles revenus peuvent être réticentes à y souscrire. Cependant, les individus ayant 1 à 5 limitations sur les activités quoitidiennes ont de faibles chance d'avoir une complémenaire. Cela peut s'expliquer par le fait que lorsque l'on est assuré à Medicare, on peut être éligibile au progamme dual de Medicaid. En effet, Medicare ne couvrent les soins liés aux limitations sur les activités quotidiennes mais proposent de rattacher les individus qui en ont besoin au régime Medicaid, une institution de santé gérée par le gouvernement américaine. De plus, les limitations sur les activtés du quotidien apparaissent surtout chez les personnes âgées (70 ans).

11 Limitations

Lors de l'interprétation de nos résultats, nous avons noté qu'il ait en moyenne peu probable que les individus agés de 67 ans à 86 ans aient une complémentaire par rapport à ceux qui sont plus plus jeunes. Or, dans la partie discussion, nous avons dit qu'il était fortement probable que les individus à la retraite ait une complémentaire. Or, au vue de la legislation américaine qui prévoit que l'âge de départ à la retraite est à 65 ans, on peut penser qu'ils y aient des individus à la retraite parmi ceux qui sont dans la tranche d'âge de 67 ans à 86 ans. Donc au vue de l'interprétation que nous avons faite tantôt concernant les individus retraités, l'effet marginal moyen associé à la variable concernant les personnes agées de 67 à 86 ans devrait aller dans le même sens. Ce qui n'est pas le cas. Cela prend peut-être en compte le fait que les personnes de cette tranche d'âge n'ont pas forcément de revenu suffisants afin de souscrire à une complémentaire.

12 Annexe

12.1 Méthode pas à pas descendante modèle logit (1)

```
## Start:  AIC=2600.85
## assurance ~ age + hisp + blanc + femme + educ + stat_mat + maladiec +
##     adl + retraite + stat_conj + etat_sante + lnrevenu
##
##              Df Deviance    AIC
## - etat_sante  4   2567.2 2593.2
## - blanc       1   2567.0 2599.0
## - stat_conj   1   2567.5 2599.5
## - femme       1   2567.8 2599.8
## <none>            2566.8 2600.8
## - stat_mat    1   2569.2 2601.2
## - adl         1   2570.1 2602.1
## - retraite    1   2571.1 2603.1
## - age         1   2571.6 2603.6
## - maladiec    1   2573.8 2605.8
## - hisp        1   2577.4 2609.4
## - educ        2   2579.8 2609.8
## - lnrevenu    1   2661.1 2693.1
##
## Step:  AIC=2593.21
## assurance ~ age + hisp + blanc + femme + educ + stat_mat + maladiec +
##     adl + retraite + stat_conj + lnrevenu
##
##             Df Deviance    AIC
## - blanc      1   2567.3 2591.3
## - stat_conj  1   2567.8 2591.8
## - femme      1   2568.1 2592.1
## <none>           2567.2 2593.2
## - stat_mat   1   2569.5 2593.5
## - retraite   1   2571.7 2595.7
## - age        1   2571.8 2595.8
## - adl        1   2571.8 2595.8
## - maladiec   1   2573.8 2597.8
## - hisp       1   2577.9 2601.9
## - educ       2   2580.9 2602.9
## - lnrevenu   1   2665.7 2689.7
##
## Step:  AIC=2591.3
## assurance ~ age + hisp + femme + educ + stat_mat + maladiec +
##     adl + retraite + stat_conj + lnrevenu
##
##             Df Deviance    AIC
## - stat_conj  1   2567.9 2589.9
## - femme      1   2568.2 2590.2
## <none>           2567.3 2591.3
## - stat_mat   1   2569.5 2591.5
## - retraite   1   2571.7 2593.7
## - adl        1   2571.8 2593.8
## - age        1   2571.9 2593.9
## - maladiec   1   2573.9 2595.9
## - hisp       1   2578.1 2600.1
## - educ       2   2580.9 2600.9
## - lnrevenu   1   2667.7 2689.7
##
## Step:  AIC=2589.87
## assurance ~ age + hisp + femme + educ + stat_mat + maladiec +
```

```
##     adl + retraite + lnrevenu
##
##            Df Deviance    AIC
## - femme     1   2569.4 2589.4
## - stat_mat  1   2569.5 2589.5
## <none>          2567.9 2589.9
## - retraite  1   2571.9 2591.9
## - adl       1   2572.3 2592.3
## - age       1   2572.8 2592.8
## - maladiec  1   2574.4 2594.4
## - hisp      1   2578.4 2598.4
## - educ      2   2581.3 2599.3
## - lnrevenu  1   2668.3 2688.3
##
## Step:  AIC=2589.39
## assurance ~ age + hisp + educ + stat_mat + maladiec + adl + retraite +
##     lnrevenu
##
##            Df Deviance    AIC
## <none>          2569.4 2589.4
## - stat_mat  1   2572.0 2590.0
## - age       1   2573.8 2591.8
## - adl       1   2574.0 2592.0
## - retraite  1   2574.6 2592.6
## - maladiec  1   2576.1 2594.1
## - hisp      1   2579.7 2597.7
## - educ      2   2582.0 2598.0
## - lnrevenu  1   2672.1 2690.1

##
## Call:  glm(formula = assurance ~ age + hisp + educ + stat_mat + maladiec +
##     adl + retraite + lnrevenu, family = binomial(link = "logit"),
##     data = train_data)
##
## Coefficients:
##   (Intercept)     age(66,86]          hisp1    educ(11,12]    educ(12,17]
##       -3.5810        -0.2093        -0.7350         0.4405         0.2853
##     stat_mat1  maladiec[1,8]       adl[1,5]      retraite1       lnrevenu
##        0.2099         0.3867        -0.3123         0.2365         0.6970
##
## Degrees of Freedom: 2137 Total (i.e. Null);  2128 Residual
## Null Deviance:       2853
## Residual Deviance: 2569  AIC: 2589
```

12.2 Méthode pas à pas descendante modèle probit (1)

```
## Start:  AIC=2596.07
## assurance ~ age + hisp + blanc + femme + educ + stat_mat + maladiec +
##     adl + retraite + stat_conj + etat_sante + lnrevenu
##
##              Df Deviance    AIC
## - etat_sante  4   2562.5 2588.5
## - blanc       1   2562.1 2594.1
## - stat_conj   1   2562.7 2594.7
## - femme       1   2563.1 2595.1
## <none>            2562.1 2596.1
## - stat_mat    1   2564.6 2596.6
## - adl         1   2565.5 2597.5
## - retraite    1   2566.2 2598.2
## - age         1   2566.9 2598.9
```

```
## - maladiec    1   2569.0 2601.0
## - hisp        1   2573.3 2605.3
## - educ        2   2575.6 2605.6
## - lnrevenu    1   2659.0 2691.0
##
## Step:  AIC=2588.51
## assurance ~ age + hisp + blanc + femme + educ + stat_mat + maladiec +
##     adl + retraite + stat_conj + lnrevenu
##
##             Df Deviance    AIC
## - blanc      1   2562.6 2586.6
## - stat_conj  1   2563.0 2587.0
## - femme      1   2563.5 2587.5
## <none>           2562.5 2588.5
## - stat_mat   1   2564.9 2588.9
## - retraite   1   2566.9 2590.9
## - age        1   2567.1 2591.1
## - adl        1   2567.4 2591.4
## - maladiec   1   2569.1 2593.1
## - hisp       1   2573.8 2597.8
## - educ       2   2576.9 2598.9
## - lnrevenu   1   2664.1 2688.1
##
## Step:  AIC=2586.56
## assurance ~ age + hisp + femme + educ + stat_mat + maladiec +
##     adl + retraite + stat_conj + lnrevenu
##
##             Df Deviance    AIC
## - stat_conj  1   2563.1 2585.1
## - femme      1   2563.6 2585.6
## <none>           2562.6 2586.6
## - stat_mat   1   2564.9 2586.9
## - retraite   1   2566.9 2588.9
## - age        1   2567.2 2589.2
## - adl        1   2567.5 2589.5
## - maladiec   1   2569.2 2591.2
## - hisp       1   2574.0 2596.0
## - educ       2   2576.9 2596.9
## - lnrevenu   1   2666.2 2688.2
##
## Step:  AIC=2585.1
## assurance ~ age + hisp + femme + educ + stat_mat + maladiec +
##     adl + retraite + lnrevenu
##
##            Df Deviance    AIC
## - femme     1   2564.8 2584.8
## - stat_mat  1   2564.9 2584.9
## <none>          2563.1 2585.1
## - retraite  1   2567.1 2587.1
## - adl       1   2567.9 2587.9
## - age       1   2568.0 2588.0
## - maladiec  1   2569.6 2589.6
## - hisp      1   2574.3 2594.3
## - educ      2   2577.3 2595.3
## - lnrevenu  1   2666.9 2686.9
##
## Step:  AIC=2584.75
## assurance ~ age + hisp + educ + stat_mat + maladiec + adl + retraite +
##     lnrevenu
##
##            Df Deviance    AIC
```

```
## <none>          2564.8 2584.8
## - stat_mat  1   2567.7 2585.7
## - age       1   2569.2 2587.2
## - adl       1   2569.7 2587.7
## - retraite  1   2570.0 2588.0
## - maladiec  1   2571.3 2589.3
## - hisp      1   2575.7 2593.7
## - educ      2   2578.1 2594.1
## - lnrevenu  1   2670.8 2688.8
```

```
##
## Call:  glm(formula = assurance ~ age + hisp + educ + stat_mat + maladiec +
##     adl + retraite + lnrevenu, family = binomial(link = "probit"),
##     data = train_data)
##
## Coefficients:
##   (Intercept)     age(66,86]          hisp1    educ(11,12]    educ(12,17]
##       -2.1884        -0.1266        -0.4431         0.2740         0.1807
##     stat_mat1  maladiec[1,8]       adl[1,5]      retraite1       lnrevenu
##        0.1325         0.2321        -0.1951         0.1430         0.4239
##
## Degrees of Freedom: 2137 Total (i.e. Null);  2128 Residual
## Null Deviance:       2853
## Residual Deviance: 2565  AIC: 2585
```