MASTER 1 ECONOMÉTRIE - STATISTIQUES

Facteurs explicatifs des coûts médicaux

Abdoul Aziz BERRADA, Baptiste GOUMAIN, Hugo HAMON

ECONOMÉTRIE APPLIQUÉE DES MODÈLES DE SANTÉ

MME HÉLÈNE HUBER

2020-2021

Contents

1 Introduction

L'Organisation mondiale de la santé (OMS) mesure tous les ans le Fardeau Global de la Maladie (FGM) dans le monde. En France, le vieillissement de la population prévu devrait amener d'ici 2025, une baisse relative des pathologies intervenant aux âges jeunes comme le VIH/sida, la dépression et les malformations congénitales. En revanche le fardeau des pathologies intervenant aux âges élevés devrait augmenter et notamment les cancers, les infractus et les crises cardiaques. Dans plusieurs pays, et en France en particulier, la croissance de la dépense de santé est supérieure à celle du PIB auxquelles les recettes sont mécaniquement liées. Cela conduit à des déficits récurrents de l'assurance maladie auxquels des réponses conjoncturelles ont été apportées par tous les gouvernements sous la forme de plans de redressement ou de réforme. Toutefois, et en dépit d'une volonté politique de maîtrise des dépenses, l'Etat et l'assurance maladie ont échoué dans la réalisation de cet objectif. Dans le même temps, il a été constaté que les performances sanitaires de la France n'étaient pas supérieures à la moyenne des pays développés.
C'est pourquoi, comprendre la raison de l'augmentation des frais médicaux demeure un enjeu crucial. Il est primordial de bien contrôler les dépenses en frais médicaux notamment pour les plus démunis ou même pour une meilleure application des politiques publiques et sociales. Toutefois, il convient tout d'abord de définir ce que l'on entend par frais médicaux; dans notre analyse ce sera au sens de coûts médicaux individuels facturés par l'assurance maladie.

Le présent travail a pour objet d'estimer l'influence des caractéristiques des individus ou des ménages basés aux États-Unis sur leurs dépenses de santé.
Pour cela nous allons dans un premier temps présenter nos données ainsi que les statistiques descriptives, puis nous analyserons différents modèles afin d'expliquer au mieux les déterminants des frais médicaux.

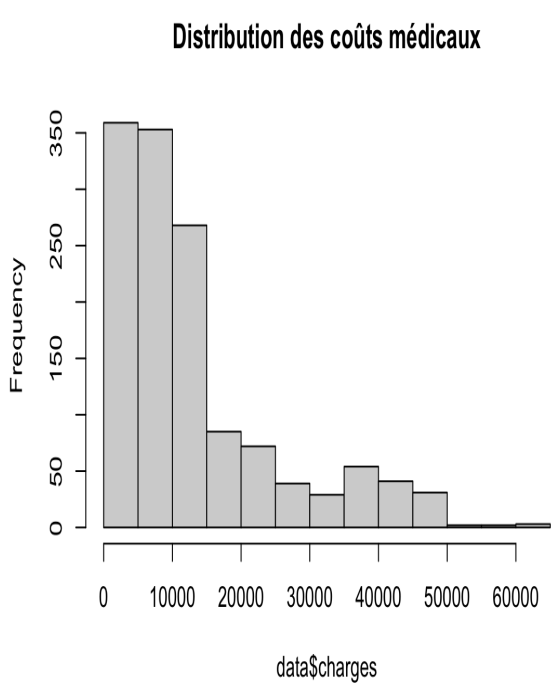
2 Présentation des données et statistiques descriptives

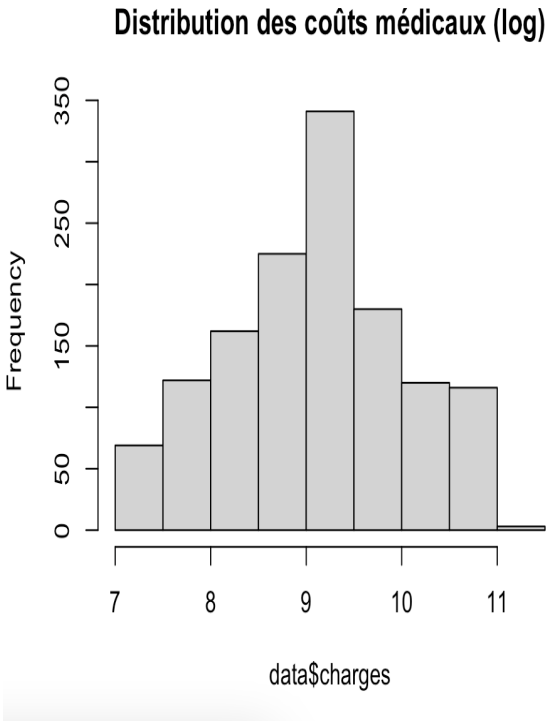
Nous disposons d'une base de données issue de l'ouvrage *Machine Learning with R* de Brett Lantz, portant sur les coûts médicaux et comportant des données simulées à l'aide de statistiques démographiques aux Etats-Unis. Notre base de données comporte 1338 observations et initialement 7 variables :

Age: âge du bénéficiaire principal.
Sex: sexe de l'individu assuré (femme, homme).
BMI: indice de masse corporelle, permettant de caractériser la corpulence, et est calculé en fonction du poids et de la taille, IMC (kg / m*2) idéalement entre 18,5 à 24,9.
Children: nombre d'enfants couverts par l'assurance maladie / nombre de personnes à charge.
Smoker: l'individu est-il fumeur ou non (Yes, No).
Region: la zone résidentielle du bénéficiaire aux États-Unis (nord-est, sud-est, sud-ouest, nord-ouest).
Charges: Frais médicaux individuels facturés par l'assurance maladie.

Ainsi on essaiera de déterminer les paramètres qui caractérisent le mieux les frais médicaux (charges) parmi l'âge, le sexe, le nombre d'enfants, l'indice de masse corporelle, le fait d'être fumeur ou

non. Nous aborderons dans cette partie des statistiques descriptives ainsi que des descriptions graphiques de nos données.
Commençons avec notre variable expliquée : les coûts médicaux (charges).
Celle-ci n'est pas centrée, sa distribution est concentrée vers la gauche et la skewness est de 1,51. Comme la variable ne contient aucune valeur nulle ou négative, nous lui appliquons donc le logarithme pour la centrer.



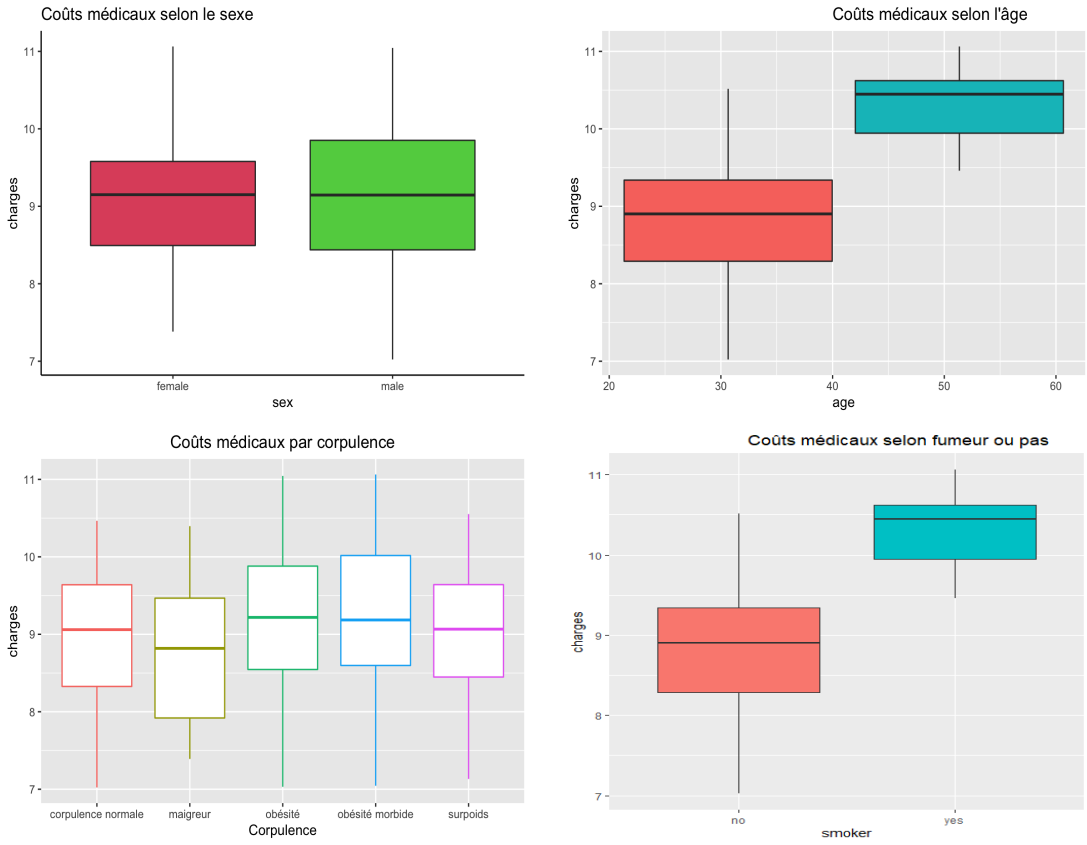


Procédons maintenant à un descriptif global de notre base de données.

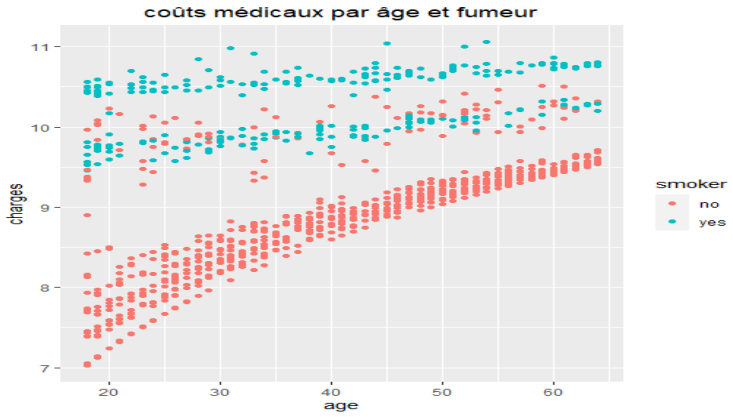
Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	1,338	39.207	14.050	18	27	51	64
bmi	1,338	30.663	6.098	15.960	26.296	34.694	53.130
children	1,338	1.095	1.205	0	0	2	5
charges	1,338	13,270.420	12,110.010	1,121.874	4,740.287	16,639.910	63,770.430

Nous observons que les individus présents dans les données sont âgés de 18 à 64 ans, que leur âge moyen est de 39 ans, qu'ils ont au maximum 5 enfants, que leur indice de masse corporelle (bmi) oscille entre 15.96 et 53.13, et qu'ils sont en moyenne au bord de l'obésité.

Analysons les boxplots de quelques variables explicatives croisées aux frais médicaux.



A travers ces 4 graphiques nous constatons premièrement que les hommes et les femmes, quelle que soit leur corpulence, ne dépensent pas des montants très éloignés en termes de frais de médicaux. Ensuite que les fumeurs dépensent plus que les non fumeurs ainsi que les personnes âgées par rapport aux moins âgées.
Ainsi, en croisant alors l'âge et le fait de fumer ou non nous obtenons le graphique suivant :



Le constat est sans appel : les dépenses de santé augmentent avec l'âge et le fait de fumer, les fumeurs avec un âge avancé sont ceux qui sont le plus exposés à d'importants frais médicaux.
Nos statistiques descriptives nous montrent que nos données n'étaient pas centrées ainsi qu'une prédominance d'importants frais médicaux chez les individus fumeurs avec un âge avancé. Maintenant, essayons de montrer comment les frais médicaux varient effectivement lorque les caractéristiques associées aux individus et aux ménages varient.

3 Modèles statistiques

L'objectif de cette partie est d'estimer la valeur des différents coefficients associés à chaque variable à travers différents modèles, de sélectionner le meilleur, le corriger au mieux, pour enfin déterminer quels sont les facteurs principaux des coûts médicaux.
Pour commencer, nous estimons un premier modèle très simple, dans lequel les frais médicaux dépendent uniquement de l'âge et du fait d'être fumeur ou non conformément à la conclusion faite lors des statistiques descriptives.

Le modèle (M1) est le suivant :

$$charges_i = \beta_0 + \beta_1 age_i + \beta_2 smoker_i + \epsilon_i \tag{1}$$

Voici les premiers résultats :

	Dependent variable:
	charges
age	0.036***
	(0.001)
smokeryes	1.539***
	(0.032)
Constant	7.387***
	(0.039)
Observations	1,329
R^2	0.738
Adjusted R^2	0.738
Residual Std. Error	0.469 (df = 1326)
F Statistic	1,869.342*** (df = 2; 1326)
Note:	*p<0.1; **p<0.05; ***p<0.01

Nous obtenons déjà un R^2 assez élevé de l'ordre de 73,8%. Donc ce modèle, assez simple explique déjà une grande partie de la variance. L'âge et le fait de fumer ou non sont significatives au seuil 1% ainsi que le modèle globalement. Les coefficients associés à nos deux variables sont tous les deux positifs, ce qui confirme l'intuition issue de nos statistiques descriptives. Cependant, le modèle semble trop simple, il convient donc de l'amender.
Cette fois-ci nous ajoutons le statut de corpulence des individus ainsi que le nombre d'enfants des

ménages. Pour ce faire nous créons une nouvelle variable "obèse" qui sera indexée sur la valeur de l'IMC. Comme son nom l'indique, la variable nous dit si l'individu est obèse ou non, elle sera égale à "yes" si l'indice de masse corporelle est supérieur à 30 et sera "no" sinon.
Nous supprimons enfin les outliers compris dans la variable "bmi" afin d'éviter de fausser nos estimations par la suite.

Nous régressons alors le modèle (M2) suivant :

$$charges_i = \beta_0 + \beta_1 age_i + \beta_2 sex_i + \beta_3 children_i + \beta_4 smoker_i + \beta_5 obese_i + \epsilon_i \quad (2)$$

Nos résultats sont les suivants :

	Dependent variable:
	charges
age	0.035^{***}
	(0.001)
sexmale	-0.074^{***}
	(0.025)
children	0.101^{***}
	(0.010)
smokeryes	1.544^{***}
	(0.031)
obeseyes	0.140^{***}
	(0.025)
Constant	7.272^{***}
	(0.041)
Observations	1,329
R^2	0.763
Adjusted R^2	0.762
Residual Std. Error	0.447 (df = 1323)
F Statistic	851.928^{***} (df = 5; 1323)
Note:	*p<0.1; **p<0.05; ***p<0.01

Nous obtenons un modèle légèrement plus performant. Le R^2 est de 76,3% (73,8% pour (M1)) et tous les coefficients sont très significatifs. Le modèle est également globalement significatif et explique bien les frais médicaux. Les variables "age", "smoker" et "obese" haussent significativement les frais médicaux alors que les hommes eux voient leurs frais médicaux baisser par rapport aux femmes.

Toutefois nous pouvons considérer que même si l'âge est un facteur explicatif important, il n'est pas une fonction linéaire des coûts médicaux. On introduit alors la variable "age*2" qui sera le carré de la variable "age". On peut penser que plus on vieillit plus les frais médicaux augmentent mais passée un certain seuil cette augmentation marginale devient décroissante.

De plus, nous pensons que croiser les variables "obese" et "smoker" permettra d'améliorer encore le modèle.

Donc notre nouveau modèle (M3) est le suivant :

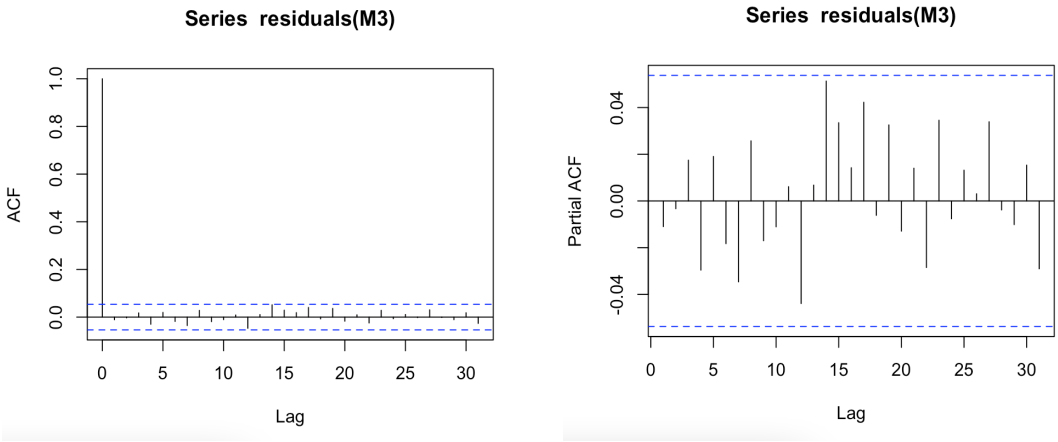
$$charges_i = \beta_0 + \beta_1 age_i + \beta_2 age2_i + \beta_3 children_i + \beta_4 sex_i + \beta_5 obese * smoker_i + \epsilon_i \quad (3)$$

Nous obtenons enfin les résultats ci-après :

	Dependent variable:
	charges
age	0.054^{***}
	(0.006)
age2	-0.0002^{***}
	(0.0001)
children	0.093^{***}
	(0.010)
obeseyes	0.016
	(0.026)
smokeryes	1.212^{***}
	(0.042)
sexmale	-0.085^{***}
	(0.024)
obeseyes:smokeryes	0.634^{***}
	(0.058)
Constant	7.020^{***}
	(0.104)
Observations	1,329
R^2	0.784
Adjusted R^2	0.783
Residual Std. Error	0.427 (df = 1321)
F Statistic	685.667^{***} (df = 7; 1321)
Note:	*p<0.1; **p<0.05; ***p<0.01

Le modèle est globalement significatif et explique 78,4% (76,3% pour (M2)) de la variance. Toutes les variables sont significatives. L'âge a un coefficient toujours positif et comme nous pouvions nous y attendre, le coefficient associé à "age*2" vient en négatif. Le croisement de "obese" et "smoker" est très concluant. En effet il confirme encore plus notre idée de base. Nous garderons ce modèle pour la suite de notre analyse.

Cependant il convient encore de traiter l'autocorrélation et l'hétéroscédasticité des résidus. Commençons par l'autocorrélation en menant un test ACF (Autocorrelation function) et puis un autre test PACF (PartielACF).



Nous pouvons conclure qu'il n'y a pas d'autocorrélation des résidus.

Testons à présent l'hétéroscédasticité avec le test de Breush-Pragan :

Studentized Breusch-Pagan test :
BP = 100.08, df = 7, p-value < 2.2e-16

On rejette H_0 donc on a bel et bien de l'hétéroscédasticité. En effet nous pouvons considérer que ce problème d'hétéscédasticité des perturbations apparait car on a des données en coupe et ceci cumulé à un "effet de taille". En effet tous les déterminants exhaustifs des frais médicaux ne sont pas pris en compte dans notre modèle, cette spécification conduit donc à une incertitude sur le modèle à travers la perturbation. Il est fort vraisemblable que cette incertitude corresponde à une erreur relative et non absolue sur l'explication des frais médicaux. Ainsi un individu jeune et non fumeur ne dépense pas le même montant qu'un individu âgé et fumeur. Ainsi les erreurs absolues sont de tailles très différentes et par la même occasion nos estimateurs $\hat{\beta}_{mco}$ ne sont plus efficaces. Nous traitons le problème en utilisant les moindres carrés quasi généralisés.
Nous pouvons écrire le modèle sous la forme suivante :

$$\underbrace{\hat{\Omega}^{-1/2}y}_{y^*} = \underbrace{\hat{\Omega}^{-1/2}X}_{X^*}\beta + \underbrace{\hat{\Omega}^{-1/2}\epsilon}_{\epsilon^*}$$

Avec $\hat{\Omega} \xrightarrow{P} \Omega$ et Ω est la matrice variance-covariance des résidus.

$$\hookrightarrow y^* = X^*\beta + \epsilon^*$$

Avec $\hat{\beta}_{mcqg} \xrightarrow{P} \hat{\beta}_{mcg}$, $\hat{\beta}_{mcg} \xrightarrow{P} \hat{\beta}_{mco}$ et $\hat{\beta}_{mcqg}$ l'estimateur des moindres carrés quasi généralisés.

Notre nouveau modèle (M4) est alors le suivant :

$$\hookrightarrow charges_i^* = \beta_0 + \beta_1 age_i^* + \beta_2 age2_i* + \beta_2 children_i^* + \beta_4 sex^* + \beta_5 obese * smoker_i^* + \epsilon_i^* \qquad (4)$$

La première étape de l'estimation consiste à définir un estimateur convergent de Ω. On considère que $\sigma_t^2 = exp(Z_t'\phi)$. Dans ce cas, nous pouvons appliquer les moindres carrés ordnaires au modèle :

$$ln(\hat{\epsilon}_t^2) = Z_t'\Phi + \xi_t \tag{5}$$

où les $ln(\hat{\epsilon}_t^2)$ désignent les logarithmes des carrés des résidus estimés par l'application des MCO au modèle initial y=Xβ + ϵ. Ainsi lorsque les MCO sur ce modèle sont convergents, nous savons alors que l'estimateur $\hat{\phi}$ peut donc être utilisé pour contruire l'estimateur convergent des MCQG $\hat{\beta}_{mcqg}$. On utilise alors $\hat{\sigma}_t^2 = exp(Z_t'\hat{\phi})$.

On obtient les résultats ci-dessous :

	Dependent variable:
	charges
age	0.058***
	(0.006)
age2	−0.0003***
	(0.0001)
children	0.097***
	(0.010)
obeseyes	0.012
	(0.026)
smokeryes	1.233***
	(0.042)
sexmale	−0.092***
	(0.023)
obeseyes:smokeryes	0.638***
	(0.059)
Constant	6.894***
	(0.101)
Observations	1,329
R^2	0.795
Adjusted R^2	0.794
Residual Std. Error	1.159 (df = 1321)
F Statistic	730.685*** (df = 7; 1321)
Note:	*p<0.1; **p<0.05; ***p<0.01

Ainsi, toutes nos variables sont significatives au seuil de 5%. Le modèle est aussi globalement significatif et a un pouvoir significatif très élevé de l'ordre de 79,5%. Il a un R^2 sensiblement égal à celui du modèle précèdent mais les coefficients ne sont plus biaisés. On remarque céteris paribus qu'être fumeur augmente les frais médicaux de [exp(1,233)-1]*100=243% (on l'interprête de cette façon car c'est une dummy dans un modèle semi-log), être obèse de [exp(0.012)-1]*100=1,2% et qu'un invidu fumeur et obèse voit ses frais médicaux augmenter de [exp(0.638)-1]*100=89,2%.

Nous observons aussi que les hommes payent moins de frais médicaux que les femmes. De plus, une année de vie supplémentaire augmente les coûts médicaux de 5,74% (0,058-2*0,0003).

4 Conclusion

Nous venons donc d'analyser l'évolution des frais médicaux en fonction de différentes variables, le fait de fumer ou non, le genre, l'âge, l'indice de masse corporelle (bmi) et le nombre d'enfants. Il en ressort que pour les patients non fumeurs, la correlation entre une situation d'obésité (bmi>30) et ses frais médicaux n'est pas claire (elle augmente très peu les coûts médicaux). Ceci est un résultat contre intuitif car la relation obésité et frais médicaux parait évidente alors qu'il s'agit d'un paralogisme. En effet ce n'est pas le fait d'être obèse qui augmente les frais médicaux mais d'être obèse et d'avoir en plus d'autres mauvaises habitudes, ici en l'occurence celle de fumer. Un second résultat contre intuitif provient du fait qu'être obèse et fumeur augmente les coûts de seulement 89% alors qu'être fumeur augmente les coûts de 243%. Ce résultat peut provenir d'une limite de notre modèle, en effet il peut être mal spécifié, ou lié à l'importance des variables, puisque logiquement, être fumeur et obèse devrait davantage faire exploser les coûts qu'être simplement fumeur. Or ici, la différence est énorme. Les frais médicaux des hommes sont un tout petit plus élevés que ceux des femmes mais rien de vraiment interprétable et significatif. C'est donc surtout le fait de fumer ou non qui a le plus d'impact sur le niveau des frais médicaux, et d'autant plus si la personne est obèse, comme nous l'avons signifié plus haut, et que la personne est âgée. Or il s'agit d'une des problématiques majeure de notre siècle quand on sait qu'aux Etats-unis 39,6% des adultes étaient obèses en 2016, d'après la National Health and Nutrition Examination Survey. Notre analyse pourrait être utilisée entre autres dans la lutte contre le tabac. En effet, en plus des dangers sur la santé, nous avons montré ici que fumer augmentait en plus les frais médicaux. Nous sommes toutefois conscients que notre modèle ne contient pas l'ensemble des variables explicatives capables d'expliquer les frais médicaux, en effet nous sommes convaincus que le revenu, le pays d'origine, la catégorie socio-professionnelle de même que le type de maladie auraient aussi un pouvoir explicatif non négligeable.