MÉTHODES D'ÉCHANTILLONAGE

Les détermianants du taux de mortalité dans les pays de l'OCDE

Réalisé par :
Redouane ESSAMMAA
Anas OUBIDA
Ibrahima SAGNO

Sous l'encadrement du:
Pr. AMOR KEZIOU

Année universitaire : 2019-2020

Sommaire :

Introduction :

Depuis 1960, les taux de mortalité dans les pays de l'OCDE n'ont cessé de baisser passant ainsi de 225.164 morts sur 1 000 personnes vivantes en 1960 à 115.161 en 2015[1] ; soit une baisse de près de 49%.

La banque mondiale définie le taux de mortalité comme le nombre de décès au cours de l'année pour 1 000 personnes. L'OCDE rapporte son nombre de décès à 100 000 habitants.

Dans les pays de l'OCDE, les principales causes de mortalité sont les maladies cardiovasculaires et le cancer : Plus d'un décès sur : Plus d'un décès sur trois est lié à une maladie de l'appareil circulatoire et un décès sur quatre est lié au cancer. Dans l'ensemble des pays membres le dénominateur commun à l'origine de la mortalité est le vieillissement de la population. Par ailleurs, d'autres facteurs comme le régime alimentaire le tabagisme, la consommation d'alcool ainsi que l'accès aux soins de santé jouent un rôle dans ces maladies.

L'objectif de cette étude est d'identifier les déterminants de la santé qui ont un effet sur le taux de mortalité. Pour ce faire, nous allons dans une première partie (réalisée par I.Sagno) présenter nos données puis dans une deuxième partie (réalisée par R.Essammaâ) nous allons tenter de proposer un modèle qui explique le taux de mortalité à travers les déterminants de la santé, et dans la troisième partie (réalisée par A.Oubida) on va utiliser des méthodes d'estimation en grande dimension.

[1] D'après les données de la banque mondiale disponible à l'adresse : https://donnees.banquemondiale.org/indicator/SP.DYN.AMRT.MA?locations=OE

Partie I : Présentation et description des données

Notre étude porte sur 28 pays européens de l'OCDE. Les données présentées dans cette étude proviennent d'une part de la base de données statistique de l'OCDE et d'autre part de l'Eurostat.

Pour chaque pays, nous avons recueilli le taux de mortalité puis les déterminants non médicaux de la santé comme le nombre de fumeurs quotidien, le litre d'alcool consommé par an, les déterminants sociaux de la santé comme les dépenses de santé, le revenu par habitant net des dépenses de santé, le chômage de longue durée mais aussi les causes de décès comme le taux de suicide et le nombre de cas de cancer, le tout pour l'année 2014.

Tableau 1 : Extraire de la base de données

Pays	Taux_mort	besoin	Fumeurs	alcool	dep_sante	chômage	pibhbtnet	cancer	suicide
Allemagne	1017,07	1,6	15,9	11,6	5142,4	44	31007,6	253,23	11,94
Autriche	957,15	0,1	24,3	12,4	4858,6	27,2	34131,4	249,28	15,26
Belgique	970,63	2,5	18,9	10,6	4477,8	49,9	31472,2	199,4	16,1
Bulgarie	1646,5	5,6	28,2	15	1550,4	60,4	4389,6	242,41	11,5
Danemark	1028,33	1,4	13,8	9,5	4536,2	25,2	42553,8	300,61	11,91
Espagne	837,46	0,6	23	8,7	2852,7	52,8	19367,3	232,7	8,17
Estonie	1269,27	11,3	22,1	11,1	1752,2	45,3	13587,8	299,41	18,31
Finlande	994,74	3,3	15,4	8,8	3812,8	22,1	34067,2	218,57	14,55
France	829,86	2,8	22,4	12	4641,5	43,9	27778,5	245,41	14,13

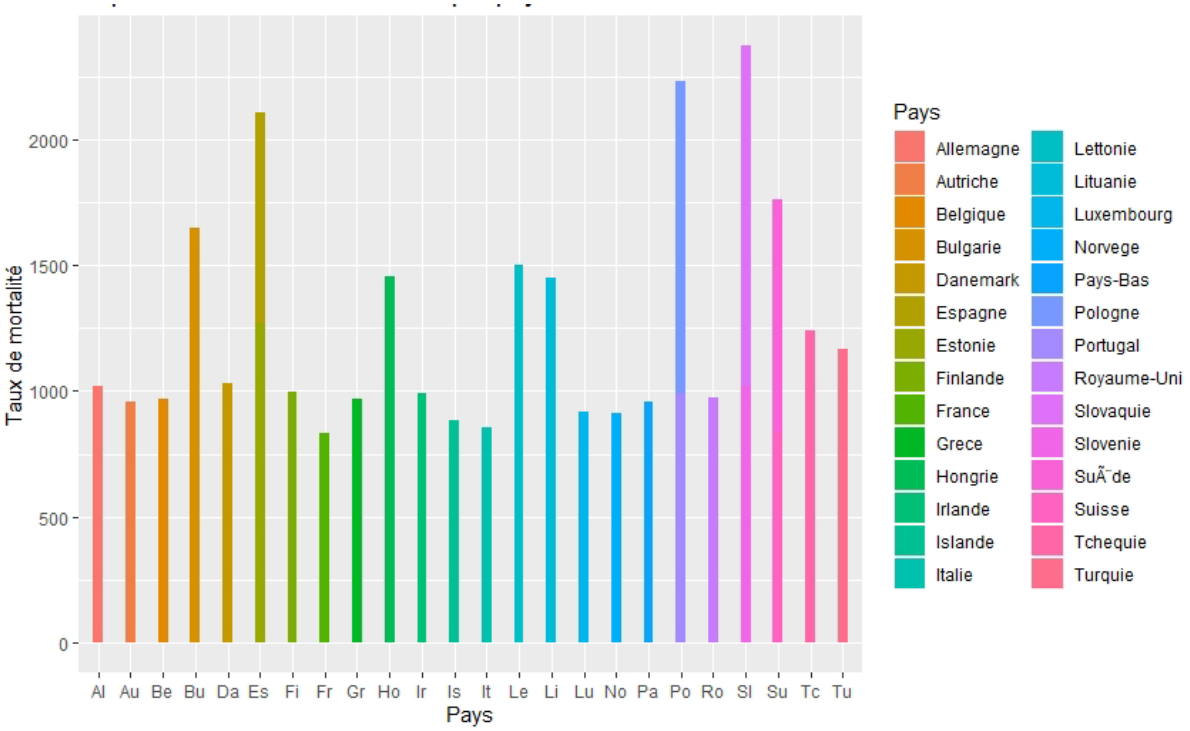
Source : les données de l'OCDE et de l'Eurostat.

Note : En Allemagne, le taux de mortalité est de 1017,07 personnes sur 100 000 personnes vivantes ; 1.6 personnes déclarent avoir des besoins non satisfaites en matière de santé ; 15,9% des + 15 ans sont fumeurs ; Ils consomment en moyenne 11,6 litres d'alcool par an ; les dépenses de santé annuelles s'élèvent 5142,4 dollars USD ; le revenu net par habitant s'élève à 310007,6 dollars USD ; 253,23 personnes sur 100 000 meurent du cancer et 11,94 sur 100 000 se suicident. [2014].

En effet, toutes ces paramètres ont un lien avec l'état de santé d'un individu. Certains sont des facteurs de risques pour la santé comme le tabagisme et l'alcoolisme, d'autres ont des effets bénéfiques sur la santé comme les dépenses de santé et d'autres comme le chômage ont des effets mitigés sur l'état de santé. (Panorama de la santé de l'OCDE).

En moyenne 1078 personnes sur 100 000 habitants dans ces pays en 2014. 9 des 28 pays ont un taux de mortalité supérieur à la moyenne. Il est beaucoup plus élevé pour la Bulgarie (1646.5), la Lettonie (1502.96), la Hongrie (1455.45), la Lituanie (1449.22), et atteint son niveau faible en France (829.86).

***Figure 1**: Répartition des taux de mortalité par pays*



Source : réalisé par les auteurs.

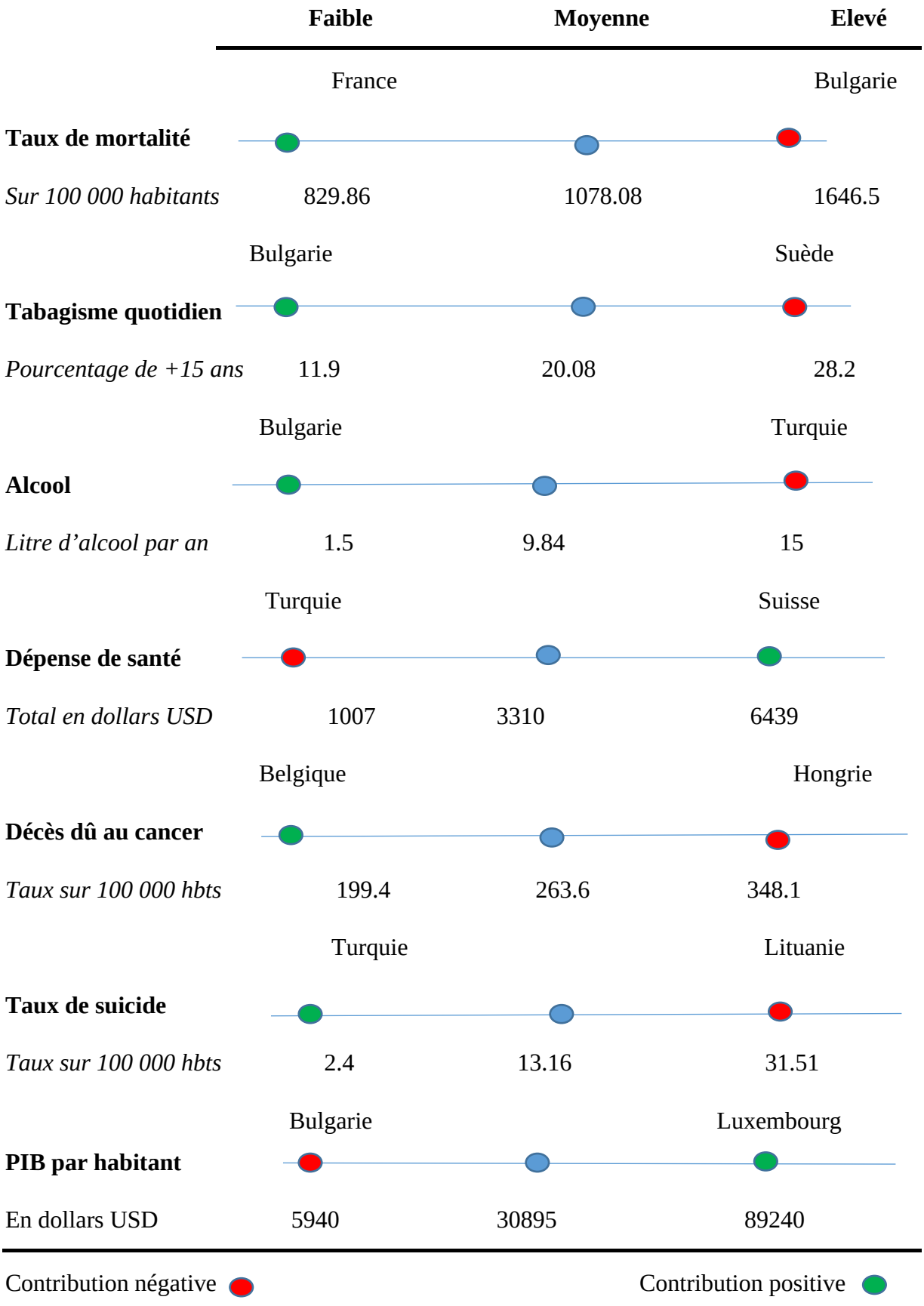
Note : le taux de mortalité en Allemagne est environ de 1000 habitants sur 100 000 personnes vivantes.

Dans cette étude, nous nous interrogeons également sur les effets entrainés par le taux de suicide et le cancer sur les pertes de vies. Dans les pays de l'OCDE, un décès sur 4 est causé par le cancer (panorama de l'OCDE) mais ce taux est en recul depuis 1990.

En parallèle, le taux de tabagisme diminue mais 18% des adultes fument encore quotidiennement. Les plus gros fumeurs sont observés en Turquie, en Hongrie et en Grèce. La consommation d'alcool quant à elle est en hausse pour 13 pays principalement en Lettonie, Belgique, Islande et Pologne. En moyenne dans les pays de l'OCDE, on consomme 9 litres d'alcool pur par an par personne.

Dans les pays de l'OCDE, les dépenses de santé s'élèvent en moyenne à 4000 dollars par personne soit 12% du PIB par habitant dans une fourchette comprise entre 1000,6 en Turquie et 6439 pour la Suisse. Dans la totalité de ces pays, l'assurance maladie obligatoire reste la principale source de financement des dépenses de santé.

***Figure 2**: Vue d'ensemble des variables*



Source : réalisé par les auteurs, d'après les données de l'OCDE.Note : une baisse du tabagisme contribue à la baisse du taux de mortalité.

Partie II : Sélection et choix de modèle

Cette partie est consacrée au choix du modèle optimal. Le fait d'avoir un modèle riche en variables explicatives ne signifie pas toujours que le modèle est parfait en termes d'explication et de prévision. Ainsi, la présence de plusieurs variables exogènes peut poser un problème de multi colinéarité entre ces variables, ce qui remet en cause l'une des hypothèses de la régression, notamment celle de l'indépendance des variables explicatives.

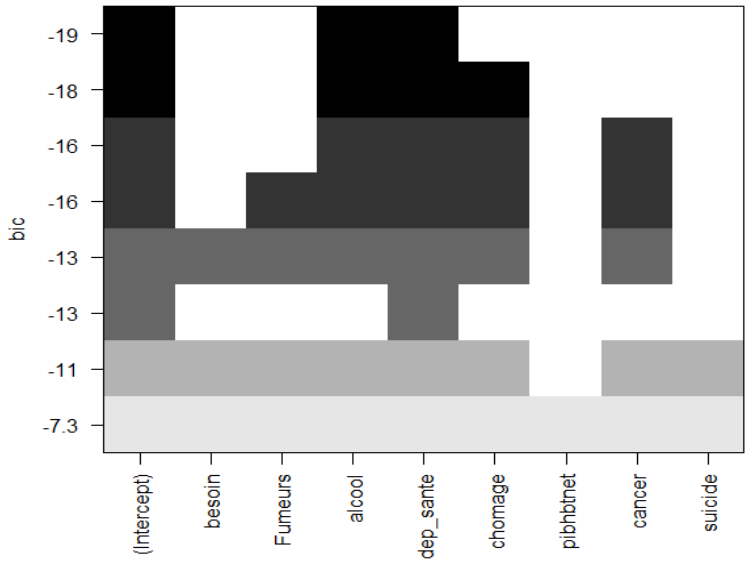
Pour prendre le nombre optimal de variables exogène, on suivre une procédure nous permettant de détecter le meilleur modèle, qui va nous permettre de mieux expliquer notre variable d'intérêt. Parmi ces méthodes, on trouve la méthode exhaustive qui consiste à construire $2^p - 1$ modèles et puis choisir celui qui optimise un critère donné (AIC, BIC, R^2 ajusté ou Cp de Mallow). Ensuite, il y a la méthode ascendante (forward selection) qui consiste à construire un modèle trivial (avec uniquement la constante), et on l'alimente variable par variable jusqu'à temps d'avoir le modèle optimal. En revanche, il y a la méthode descendante (backward selection) qui permet de construire un modèle complet (avec toutes les variables) et on commence à supprimer les variables l'une après l'autre jusqu'à ce qu'on arrive au modèle parfait. Tous ces méthodes prennent beaucoup de temps afin de choisir le modèle optimal. On peut remédier à ce problème en utilisant la méthode de sélection par algorithme génétique. L'utilisation de cette méthode nous permet de gagner du temps, car elle est supposée trouver le meilleur modèle sans passer par le calcul des critères exigés dans la recherche exhaustive.

Dans notre projet on va se baser sur la méthode exhaustive, plus particulièrement en utilisant les résultats du critère BIC, et puis on va utiliser l'algorithme génétique pour vérifier si on tombe sur le même modèle. Cela revient au fait que la dernière méthode s'est avérée la plus recommandée en cas de présence de plusieurs de variables explicatives.

1. *Choix de modèle par la méthode exhaustive (critère BIC) :*

Comme nous l'avons mentionné, cette méthode nous permet de construire 2^p - 1 modèles et grâce au calcul sur R on obtient les résultats du critère BIC pour chaque modèle, et on choisit le modèle qui minimise le plus ce critère. D'après notre application, on a obtenu les résultats suivants :

***Figure 3** : Résultats de BIC pour toutes les variables de chaque modèle*



D'après cette figure on peut déduire que le meilleur modèle est celui présenté par l'équation suivante :

$$Taux\_mort_i = \beta_0 + \beta_1 alcool + \beta_2 dep\_sante + \varepsilon_i \qquad (1)$$

Cette méthode exhaustive qui consiste à comparer entre toutes les combinaisons possibles, nous a permis de déduire le meilleur modèle avec 2 variables explicatives. Pour valider ces résultats, on va utiliser également l'algorithme génétique qui sert à sélectionner le modèle optimal, et vu que nous avons plusieurs variables explicative, cette méthode est la plus efficace. Ainsi, l'utilisation de la fonction glmulti nous a donné le même modèle optimal que la méthode exhaustive.

2. *L'estimation de l'erreur théorique de prévision*

Lors de cette étape on va essayer d'estimer correctement l'erreur théorique de prévision de chaque modèle, et choisir le modèle ayant l'erreur estimé la plus faible.

Pour estimer l'erreur de prévision nous allons utiliser la méthode LOOCV qui consiste à diviser l'ensemble des observations en deux parties. Cette méthode se caractérise par le fait qu'elle prend uniquement une seule observation (X_1, Y_1) pour la validation et tout le reste sera utilisée pour l'apprentissage (estimation du modèle).

Le modèle qui minimise l'erreur théorique de prévision peut s'écrire sous la forme suivante :

$$Taux\_mort_i = \beta_0 + \beta_1 alcool + \beta_2 alcool^2 + \beta_3 dep\_sante + \beta_4 dep\_dante^2 + \varepsilon_i$$

Après la vérification des hypothèses de la régression linéaire (la normalité des erreurs, l'indépendance des erreurs, l'homoscédasticité, etc.) *cf. Annexe I*, on a pu faire l'estimation et obtenir les résultats suivants :

***Tableau 2** : Estimation des résultats*

Variables	***Estimation***	***P-value***
Alcool	-7,42	0,85679
Alcool²	2,36	0,31081
Dep_sante	$-2{,}765^{e-01}$**	0,00485
Dep_sante²	$2{,}694^{e-05}$*	0,03977

**** le coefficient est significatif à 1%, ** le coefficient est significatif à 5%, * le coefficient est significatif à 10%*

D'après ces résultats, on s'aperçoit que la variable alcool ainsi que celle-ci au carré ne sont pas significatives ce qui nous incite à reprendre le modèle choisi par l'algorithme génétique pour gagner plus en termes d'explication et de prédiction.

Cependant, vu que nous disposons maintenant d'un modèle linéaire, il convient de tester les hypothèses de la régression linéaire afin de donner une conclusion sur la significativité des variables.

- *L'hypothèse de l'autocorrélation des erreurs :*

Cette hypothèse peut se vérifier par l'application du test de Durbin&Watson :

Statistique du test	***P-value***
2,1279	0,807

Les résultats de ce test nous permettent de conclure que les erreurs sont indépendantes. Vu que p-value > 5% on accepte l'hypothèse nulle qui stipule que les erreurs sont corrélées. Ce résultat peut se conformer avec le graphique figurant dans l'annexe.

- *Test d'homoscédasticité ou test de Breusch-Pagan :*

Ce test consiste à tester l'hypothèse nulle, selon laquelle, les erreurs sont homoscédastiques contre l'alternative où il y a une hétéroscédasticité.

Statistique du test	*P-value*
2,4339	0,2961

D'après les résultats, on accepte l'hypothèse d'homoscédasticité des erreurs car la p-value > 5%.

- *Test de normalité des résidus :*

On peut tester la normalité des erreurs en utilisant le test de Shapiro-Wilk, qui consiste à tester l'hypothèse nulle qui stipule que les erreurs suivent une loi normale contre l'alternative où les erreurs ne sont pas normales.

Statistique du test	*P-value*
0,95727	0,2998

Selon les résultats obtenus, on accepte l'hypothèse nulle parce que la p-value > 5% et on peut déduire que les erreurs suivent une loi normale.

- *Les valeurs influentes :*

Il s'agit des observations qui peuvent influencer le résultat de la prédiction du modèle, donc il est généralement préférable de supprimer ces valeurs avant de passer à l'étape de l'estimation des résultats. La détection de ces valeurs peut se faire de différentes manières. Dance cette étude nous allons se baser sur la distance de Cook (présentée en annexe). D'après les résultats on va supprimer l'observation (Turquie) qui s'est avérée influente.

- *Estimation des résultats :*

Dans cette phase on a fait l'estimation par la méthode des moindres carrés ordinaires. On note qu'on a introduit le logarithme sur nos variables et on a obtenu les résultats suivants :

	Estimation	**t_value**	**P-value**
Intercept	8,376727***	16,113	$2,26^{e-14}$
Alcool	0,29076**	2,669	0,0134
Dep_sante	-0,25996***	-5,563	$1,01^{e-05}$

**** le coefficient est significatif à 1%, ** le coefficient est significatif à 5%, * le coefficient est significatif à 10%*

D’après ces résultats on peut conclure que la consommation d’alcool impacte positivement le taux de mortalité des personnes. Toute augmentation de la consommation d’alcool d’une unité augmente le taux de mortalité de 0,29%. Cela peut se traduire par le fait que les gens alcooliques sont souvent risqués d’avoir des accidents de la route ou bien être victimes d’homicides.

Par contre la variable de la dépense en santé a un effet négatif sur le taux de mortalité des gens. Pour ce, toute augmentation des dépenses en santé réduit le taux de mortalité de la population de 0,26%. Cet impact s’explique par le fait que les dépenses en santé sont destinées à améliorer le système sanitaire afin d’équiper tous les établissements santé par le matériel nécessaire qui joue en faveur de la facilité de détection de maladies, ainsi que les investissements dans la recherche et développement des médicaments qui permettent aux individus de faire face à toute sorte de maladie et donc ça va permettre de réduire relativement leur taux de mortalité.

Partie III : régression en grande dimension

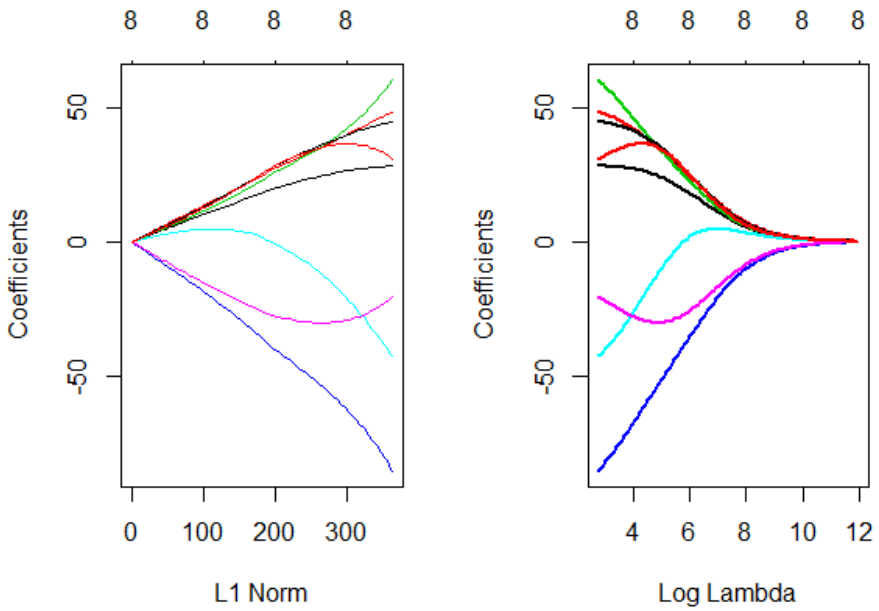
Dans notre étude on dispose d'un nombre assez important de variables explicatives, ce qui implique que les estimateurs de MCO possèdent une variance très élevée, cela peut entrainer probablement une non-significativité des coefficients. Donc la solution la plus adaptée à ce type de problème est de recourir à des régressions en grande dimension, qui permettent de contraindre les paramètres du modèle. A cet égard on va utiliser deux types d'estimation de notre modèle de régression : l'estimateur Ridge, et Lasso.

1. *Régression Ridge :*

Cette méthode consiste à pénaliser l'estimateur de MC par la norme L_2 du vecteur (w1,... ,wp).

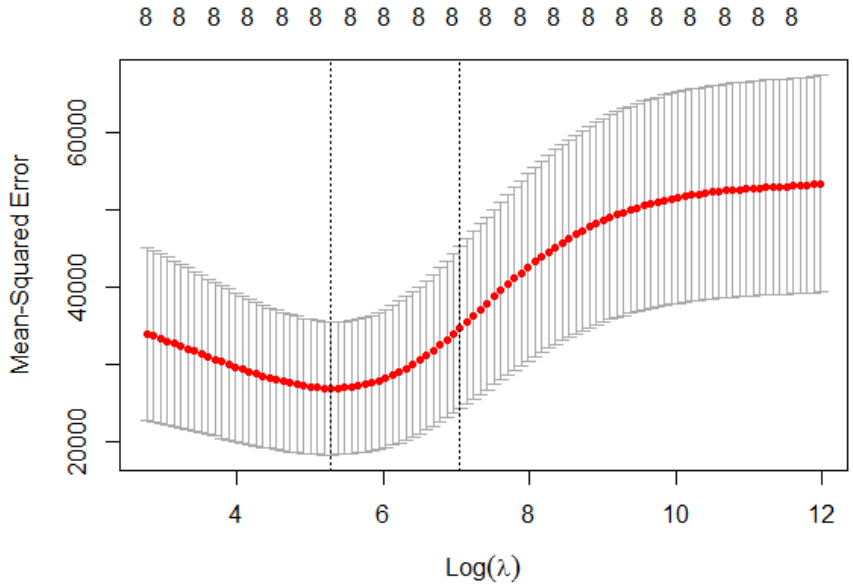
Présentant maintenant les résultats de cette régression.

***Figure 4** : Variation des coefficients en fonction de lambda et la norme L1*



Il s'agit de noter que si lambda=0, on tombera sur l'estimateur de MC (qui est sans biais), et plus lambda augmente plus le biais des estimateurs augmente et leurs variances diminuent, l'inverse est vrai aussi. Donc le problème qui se présente à cet égard, est le choix de lambda, pour déterminer cette valeur on va utiliser la méthode de validation croisée et on choisira lambda qui minimise l'erreur de prévision estimé.

Figure 5 : Evolution de l'erreur de prévision en fonction de la valeur de lambda



Dans le graphique ci-dessus on observe que l'erreur de prévision minimale est atteinte lorsque log(lambda) = 5.28, et donc la valeur optimale de lambda est de $\lambda = 196.5107$.

Après avoir déterminer la valeur de λ, les coefficients du modèle estimé par le Ridge sont les suivants :

Variables	***Estimation***	**Variables**	**Estimation**
Besoin	10.81	Chômage	4.80
Fumeurs	14.54	Pibhbtnet	-16.23
Alcool	12.36	Cancer	14.134885
Dep_sante	-20.04	Suicide	13.610309

On constate que les signes des coefficients sont très logiques, et par conséquent leurs impacts sont plausibles aussi. Les valeurs des paramètres de ce modèle sont différentes de ceux du modèle RLM complet, que vous trouverez en annexes, ainsi que le coefficient relatif à la variable qui désigne le chômage a changé de signe par rapport au modèle de RLM complet, sans prendre en considération la significativité de ces paramètres.

En effet, l'erreur de prévision estimée du modèle de régression Ridge optimal, qui est égale à 26875.07, est inférieure à celle du modèle RLM complet qui vaut 35771.92[2], cela confirme la qualité de l'estimateur Ridge face à celui de MC, lorsque le nombre de variables explicatives est important. Cependant cette erreur de prévision du modèle de régression Ridge est supérieur

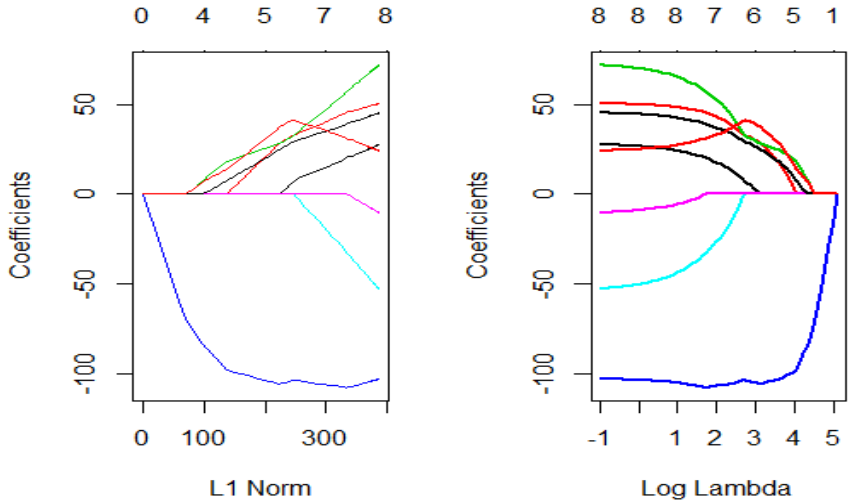
[2] On a estimé cette valeur par la méthode LOOCV.

à celle du modèle optimal choisi par l'algorithme génétique, donc ce dernier reste plus puissant en termes de qualité d'estimation et de prévision. Raison pour laquelle on va faire une régression Lasso qui permet à la fois, au contraire du Ridge, à estimer et à choisir le modèle optimal.

2. *Régression Lasso :*

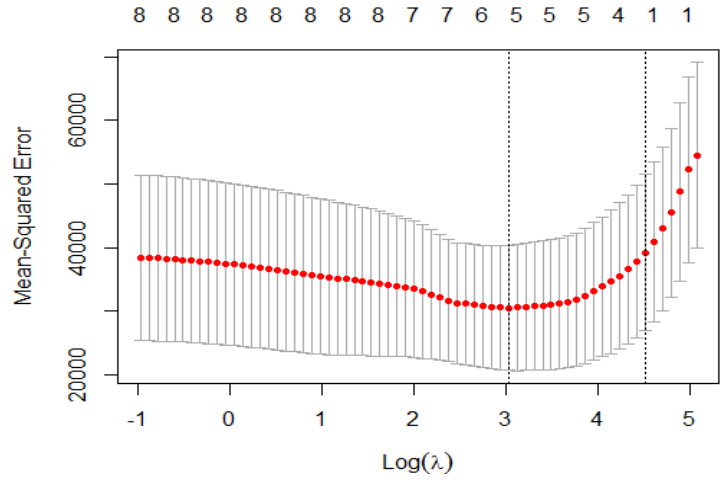
L'estimateur Lasso consiste à pénaliser l'estimateur de MC par la norme L_1 du vecteur (w1,... ,wp), et comme on l'avait cité, cette méthode permet de sélectionner les variables pertinentes et aussi à estimer les paramètres du modèle.

***Figure 6** : variation des coefficients en fonction de lambda et la norme L1.*



Comme c'était le cas pour la régression Ridge, plus le lambda augmente plus le biais augmente et la variance des estimateurs diminue, et plus lambda diminue, le biais diminue et la variance croît, mais contrairement à l'estimateur Ridge, lorsque lambda augmente le nombre de variables explicative diminue, c'est bien clair la procédure de choix de modèle. Passant maintenant à la détermination de la valeur de lambda.

***Figure 7**: Evolution de l'erreur de prévision en fonction de la valeur de lambda.*



On constate d'après la figure ci-dessus que l'erreur de prévision minimale est atteinte lorsque log(lambda) =3.024, c'est ce qui est relatif à une valeur de $\lambda = 20.59$.

Pour cette valeur de λ, une seule variable est choisie, c'est celle relative aux dépenses de santé, et par conséquent le modèle de régression lasso s'écrit comme suit :

$$Taux\_mort_i = 1078.08000 + -69.43 * dep\_sante + \varepsilon_i$$

En effet, le fait que la régression Lasso a choisi une seule variable explicative, son erreur de prévision sera plus élevée que celles des modèles estimés auparavant (RLM complet, optimal d'après l'algorithme génétique, et Ridge), cette valeur vaut 30508.81.

Donc on peut tirer comme conclusion que le modèle choisi par l'algorithme génétique et la méthode exhaustive, est le meilleur en termes d'explication et de prévision, en se basant sur l'erreur théorique de prévision.

Conclusion :

A l'issue de notre étude qui porte sur l'explication du taux de mortalité au sein des pays de l'OCDE, on a conclu que parmi huit variables exogènes, deux ont été choisies dans le modèle optimal par la méthode exhaustive et l'algorithme génétique, cela est dû à une multi-colinéarité et une dépendance entre ces variables explicatives, autrement dit on trouve que certaines variables expliquent le taux de mortalité de la même manière, c'est pour cette raison que plusieurs entre elles avaient un impact non significatif. La non-significativité des coefficients de la régression peut être relative aussi à la hausse des variances des estimateurs de MC, donc on a utilisé une régression Lasso et Ridge afin de régler ce problème, en revanche aucun de ces modèles ne présente une bonne alternative à l'estimation par MCO en termes d'explication des variables, et de prédiction. Donc on peut conclure que le problème principal de la non-significativité de certaine variable est le fait qu'elle sont dépendantes les unes des autres. Cependant toute augmentation de la consommation de l'alcool par unité implique un accroissement du taux de mortalité de 0.29%, et toute baisse de ce dernier par 0.26% est due à un accroissement des dépenses de santé par un dollars.

En effet, Il ressort de cette étude qu'il n'est pas facile de déterminé les facteurs permettant d'expliquer les variations du taux de mortalité dans les pays de l'OCDE. Ainsi, plus des variables liées au système de santé et la consommation de l'alcool, il faut prendre en considération un certain nombre de facteurs, principalement environnementaux, non observables.

Liste des figures :

Listes des annexes :

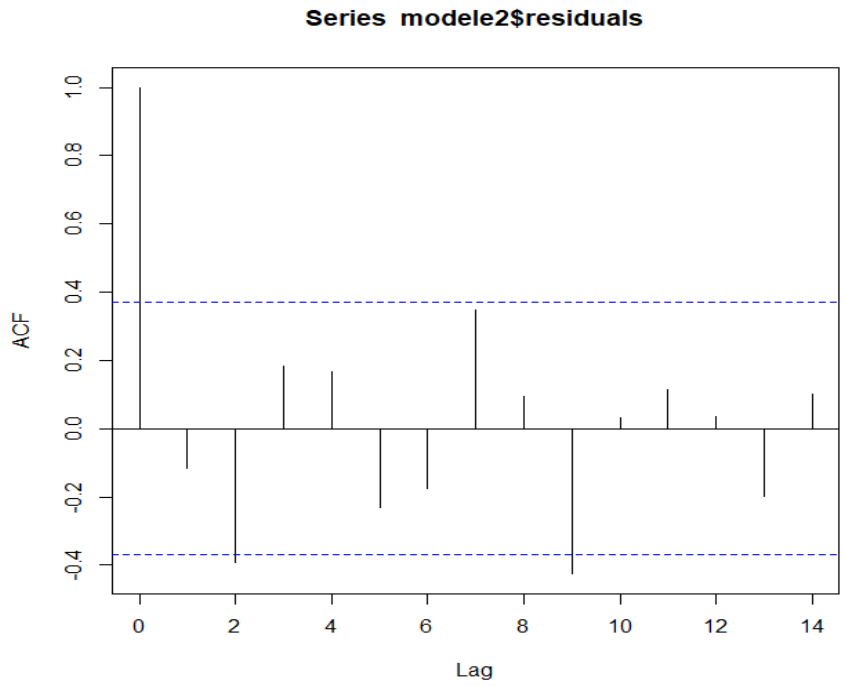
Annexes :

Annexe I :

- *L'hypothèse de non corrélation des erreurs :*

Statistique du test	***P-value***
2,2253	0,6942

Puisque la p-value > 5%, on a ccepte H_0 qui stipule l'indépendance des erreurs.



- *L'hypothèse d'homoscédasticité : Test de Breush-Pagan*

Statistique du test	***P-value***
4,0108	0,4045

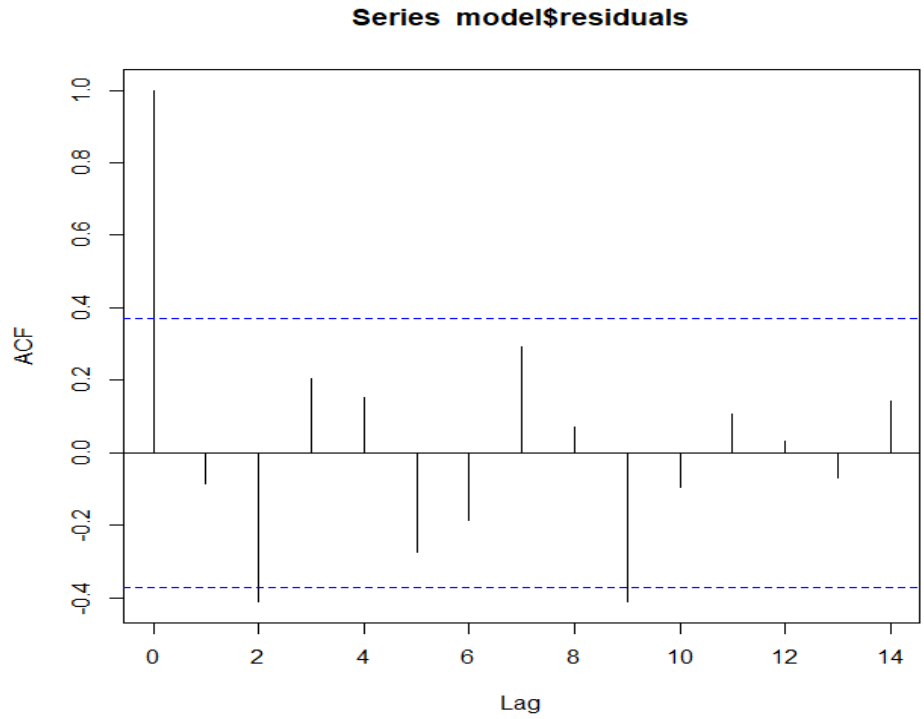
On accepte l'hypothèse, selon laquelle, les erreurs sont homoscédastiques car la p-value > 5%.

- *L'hypothèse de normalité des erreurs : Test de Shapiro-Wilk*

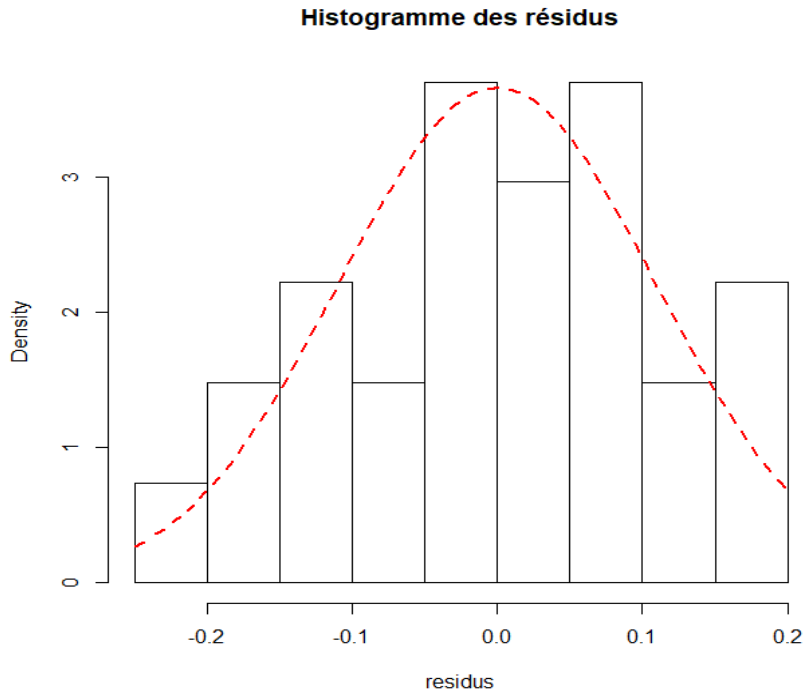
Statistique du test	***P-value***
0,97491	0,7162

On peut déduire que les erreurs suivent une loi normale, vu que la p-value > 5%.

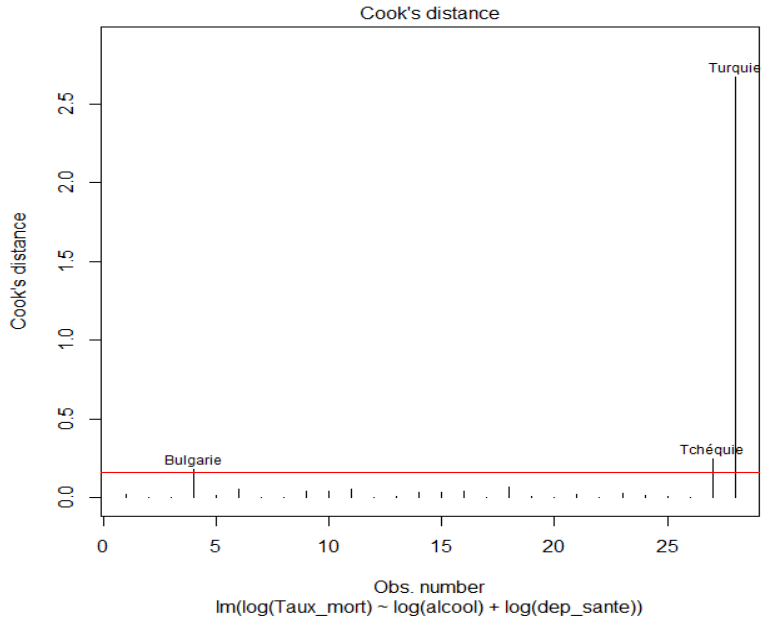
Annexe II : L'hypothèse de corrélation des erreurs :



Annexe III : Histogramme et normalité des erreurs



IV : Distance de Cook



Annexe V : Classement des variables selon les résultats du test de Student et celui de Fisher

- *Classement selon le test de Student :*

Variables	***T-Student***
Dep_sante	$5{,}190052^{e\text{-}06}$
Alcool	$4{,}855766^{e\text{-}03}$

- *Classement selon le test de Fisher :*

Variables	***T-Student***
Dep_sante	$5{,}190052^{e\text{-}06}$
Alcool	$1{,}738091e^{-03}$

Les deux tests nous ont donné le même classement, mais d'une manière générale il est très recommandé de retenir le classement selon les résultats du test de Fisher.

Annexe VI : Les coefficients de la RLM du modèle complet :

Variables	***Estimation***	**Variables**	**Estimation**
Besoin	7.881	Chômage	-3.407
Fumeurs	11.06	Pibhbtnet	-5.854e-04
Alcool	27.43	Cancer	1.281
Dep_sante	-6.537e-02	Suicide	4.280