Prédiction du taux de mortalité par cancer

Projet SAS en Modèle linéaire et généralisation
Auteurs : ***Mengru CHEN, Chuyao LU, Sina MBAYE, Edimah SONGO***
Professeur : Bérangère BERTHO

Introduction

L'objectif de ce projet sera d'expliquer la mortalité liée au cancer à partir des caractéristiques de la population de chaque région.

La base de données ***cancer_reg***, datant de 2016, détaille à partir de plusieurs sources américaines, 3047 enregistrements de taux de mortalité par cancer, et 34 variables agrégées explicatives. Il y a une ligne par région géographique, la variable **Geography** est donc la clef primaire.

Les informations sur les facteurs de mortalité portent pour la plupart sur le revenu, l'âge, la couverture médicale, le nombre de défunts diagnostiqués au cancer etc.

Afin d'expliquer le taux de mortalité lié au cancer compte tenu des caractéristiques de la population de chaque région, nous commencerons par étudier et faire un audit des données du fichier ensuite nous apporterons quelques modifications sur la base de données avant de tester plusieurs modèles.

1. Audit de la base de données

Types de variables

On remarque que la plupart des variables sont numériques, mises à part « **binnedInc** », représentant le revenu médian par habitant sous forme d'intervalle, et « **Geography** », qui donne le nom de chaque région.

Il n'y a pas de doublons, et on peut déjà relever que la clef primaire est la variable « **Geography** », car elle donne un indicateur unique pour chaque observation.

Créations éventuelles de variables

On remarque cependant que les fréquences par « counties » sont parfois négligeables. On crée donc une variable **« Region »** regroupant tous les « counties » d'un même état sous une seule variable.

Certaines régions n'ont qu'un seul county et ne répondent pas correctement à notre première correction : on remarque que les noms des états apparaissent plusieurs fois avec des lettres manquantes ou mal orthographiées, ce qui crée des doublons. On utilise donc des commandes conditionnelles pour y remédier.

Suppressions éventuelles de variables

En analysant de plus près les variables quantitatives, plusieurs d'entre elles semblent redondantes ou peu utiles à notre analyse :

- **« PctSomeColl18_24 »**

Il y a trop peu d'observations (75% de données manquantes), et une redondance avec la variable PctBachDeg18_24. De plus, il n'y a pas variables équivalentes pour les autres tranches d'âge.

- **« PctEmployed16_Over »**

Ici aussi il y a trop peu d'observations (5% de données manquantes).

- **« PctPrivateCoverageAlone »**

Idem, 19% des données sont manquantes. De plus la variable « PctPrivateCoverage » offre des informations similaires.

- **« PctPublicCoverageAlone »**

Cette variable fournit des informations très similaires avec les autres variables de type PctXXXCoverageXXX avec une corrélation supérieure à 89%.

On ignorera donc ces quatre variables dans notre étude.

Modifications de variables

Il y a des erreurs de saisie dans la variable **MedianAge**, à cause d'oublis de point. On corrige donc les lignes fautives en divisant leur valeur par 10 (en particulier lorsque la médiane d'âge est supérieur à 100 ans).

On remarque de plus que la variable **studyPerCap** est souvent nulle : la médiane du nombre d'essais clinique vaut 0. La moitié des régions n'en effectuent donc pas.

2. Représentation des données

Datavisualisation géographique

On génère deux cartes représentant les taux de mortalité les plus faibles et les plus élevés.
https://fr.batchgeo.com/map/2e877c9de37fb0eaf1e5ec3fc62f33d5 (taux les plus faibles)
https://fr.batchgeo.com/map/917e7b691990f8b2925f1d4ee4eabb38 (taux les plus élevés)

Bien que les taux de mortalité les plus faibles semblent a priori assez bien répartis géographiquement, il y a une forte concentration des taux les plus élevés au sud-est du pays.

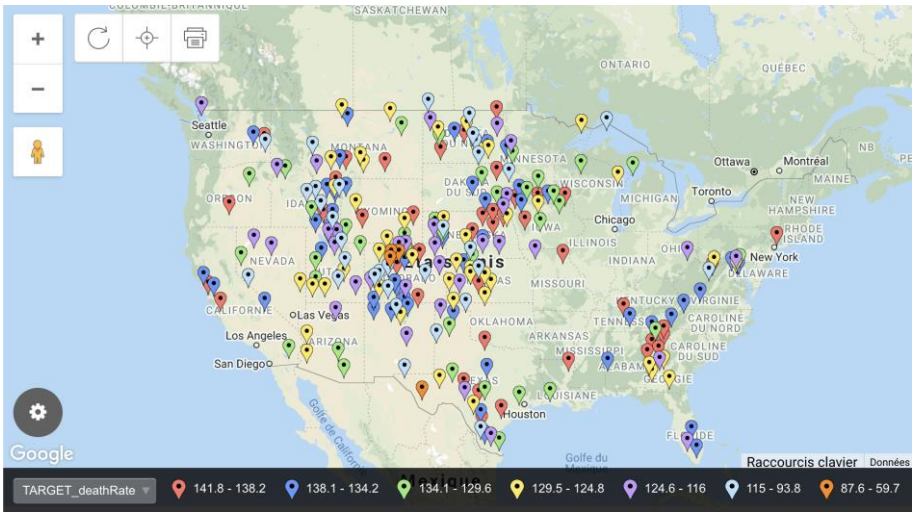

Fig. 1 - Comtés aux taux de mortalité pour 100 000 habitants les plus faibles

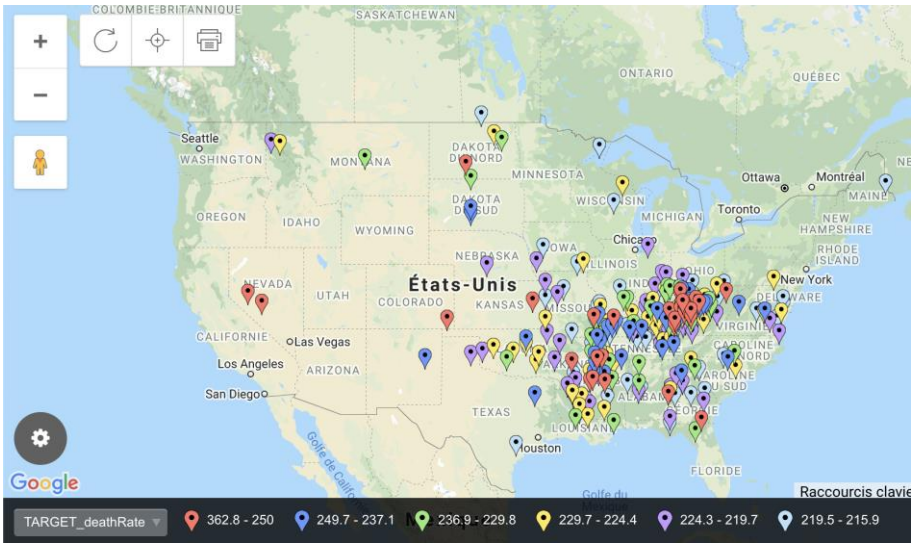

Fig. 2 - Comtés aux taux de mortalité pour 100 000 habitants les plus élevés

Générer une carte des revenus médians les plus faibles (https://fr.batchgeo.com/map/cd65f319dba19b24e1de3df9f88c9100) nous donne une piste sur la pertinence de cette variable explicative :

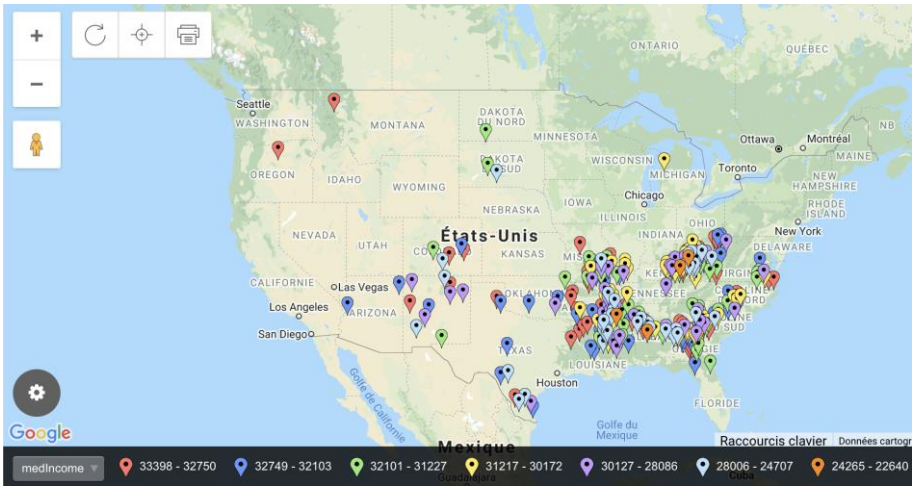

Fig. 3 - Comtés aux revenus annuels les plus faibles (en dollars américains)

Il semblerait que le revenu ait un impact sur le taux de mortalité. On poussera cependant l'analyse pour s'en assurer.

Distribution du taux de mortalité

Les taux de mortalité semblent répartis selon une loi normale, et semblent en moyenne proches de 180 pour 100 000 habitants.

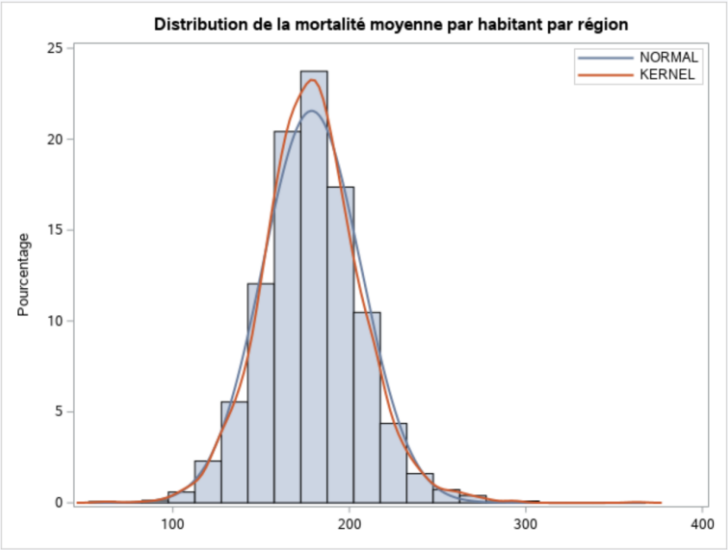

Fig. 4 - Histogramme de la mortalité moyenne et courbe gaussienne s'en rapprochant le plus.)

La gaussianité de la variable à expliquer est avérée, il n'y a donc pas de transformation nécessaire.

Distribution des variables explicatives

En analysant les distributions des variables explicatives, on remarque que certaines variables présentent des asymétries. Si elles ne sont pas supprimées par d’autres outils, il faudra y appliquer une transformation.

3. Observation des corrélations

Nous venons d'observer les données de notre base, on analyse maintenant les corrélations entre les variables explicatives et le taux de mortalité.

Pour ce faire, on va dédier 70% de la base de données à l'apprentissage, et 30% aux tests, afin de valider le modèle avec des jeux de données indépendants.

On étudiera les corrélations sur la table d'apprentissage.

Analyse de la corrélation avec les variables qualitatives

Les deux variables qualitatives sont binnedInc et Region.

En observant les barplots, on remarque qu'un niveau de revenu bas semble être corrélé à un taux de mortalité plus élevé. Ce qui semble rejoindre la datavisualisation géographique vue plus haut (Fig. 2 et 3).

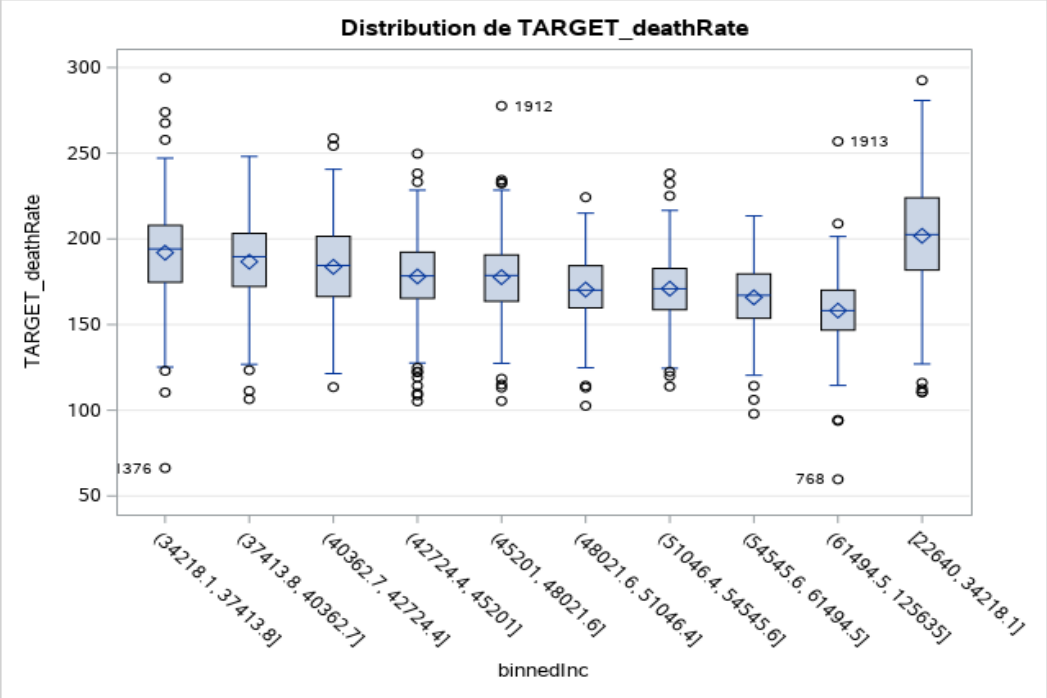

Fig. 5 - Procédure ANOVA entre taux de mortalité et revenus médians divisés par décile.

De même pour les zones géographiques, des différences de taux de mortalité moyens existent en fonction de l'état, comme on l'avait suspecté (Fig. 2).

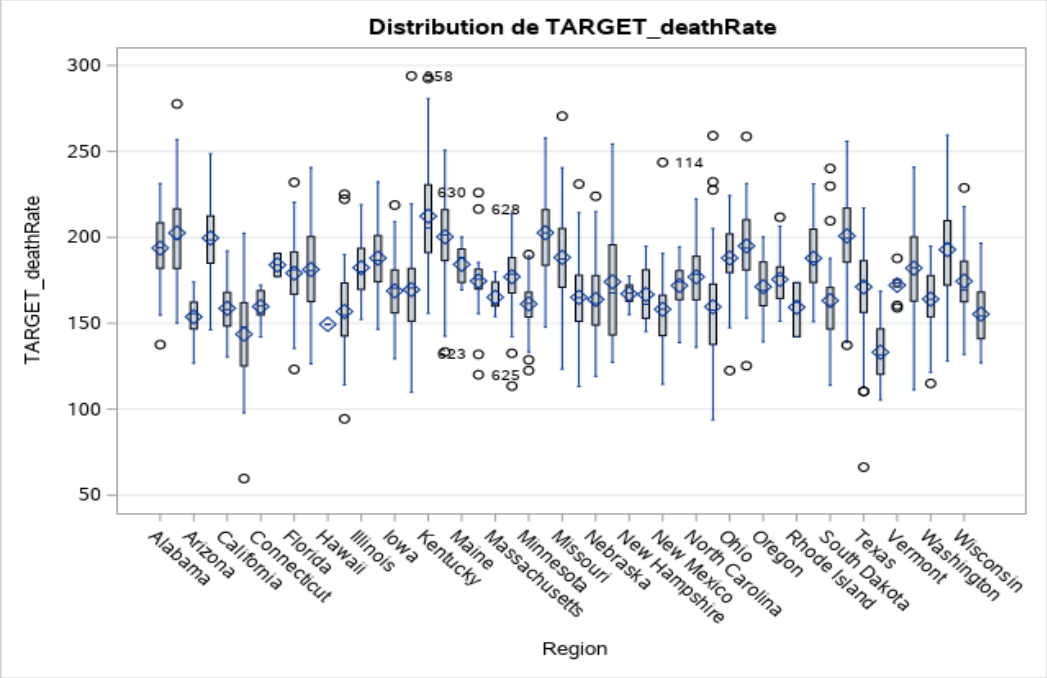

Fig. 5 - Procédure ANOVA entre taux de mortalité et regions/états.

Analyse de la corrélation avec les variables quantitatives

Toutes les variables sont plus ou moins corrélées à la variable cible. Certaines peuvent en revanche être trop corrélées entre elles, ce qui peut créer des problèmes de colinéarité.

On peut prendre l'exemple des variables avgDeathsPerYear, avgAnnCount et popEst2015 semblent très corrélées (Fig. 6). Ce qui est logique, car plus la population augmente, plus le nombre de morts est élevé. De même si on ajoute de facteurs potentiellement mortels (i.e le cancer).

	TARGET_deathRate	avgAnnCount	avgDeathsPerYear	incidenceRate	medIncome	popEst2015	povertyPercent
TARGET_deathRate	1.00000	-0.14366 <.0001	-0.09084 <.0001	0.44944 <.0001	-0.42895 <.0001	-0.12021 <.0001	0.42938 <.0001
avgAnnCount	-0.14366 <.0001	1.00000	0.93936 <.0001	0.07325 <.0001	0.26841 <.0001	0.92683 <.0001	-0.13588 <.0001
avgDeathsPerYear	-0.09084 <.0001	0.93936 <.0001	1.00000	0.06235 0.0006	0.22234 <.0001	0.97761 <.0001	-0.06709 0.0002
incidenceRate	0.44944 <.0001	0.07325 <.0001	0.06235 0.0006	1.00000	-0.00144 0.9365	0.02656 0.1428	0.00900 0.6197
medIncome	-0.42895 <.0001	0.26841 <.0001	0.22234 <.0001	-0.00144 0.9365	1.00000	0.23470 <.0001	-0.78959 <.0001
popEst2015	-0.12021 <.0001	0.92683 <.0001	0.97761 <.0001	0.02656 0.1428	0.23470 <.0001	1.00000	-0.06547 0.0003
povertyPercent	0.42938 <.0001	-0.13588 <.0001	-0.06709 0.0002	0.00900 0.6197	-0.78959 <.0001	-0.06547 0.0003	1.00000

Fig. 6 - Tableau partiel de corrélations : Variables explicatives très corrélées entre elles (en rouge) ayant des corrélations faibles avec la variable cible.

Ces variables sont également peu corrélées avec la variable cible, donc pas forcément les plus pertinentes. On peut choisir de s'en défaire. Il se trouve qu'elles n'avaient pas forcément une distribution gaussienne.

En revanche, d'autres variables — comme le taux de pauvreté — semblent davantage corrélées au taux de mortalité.

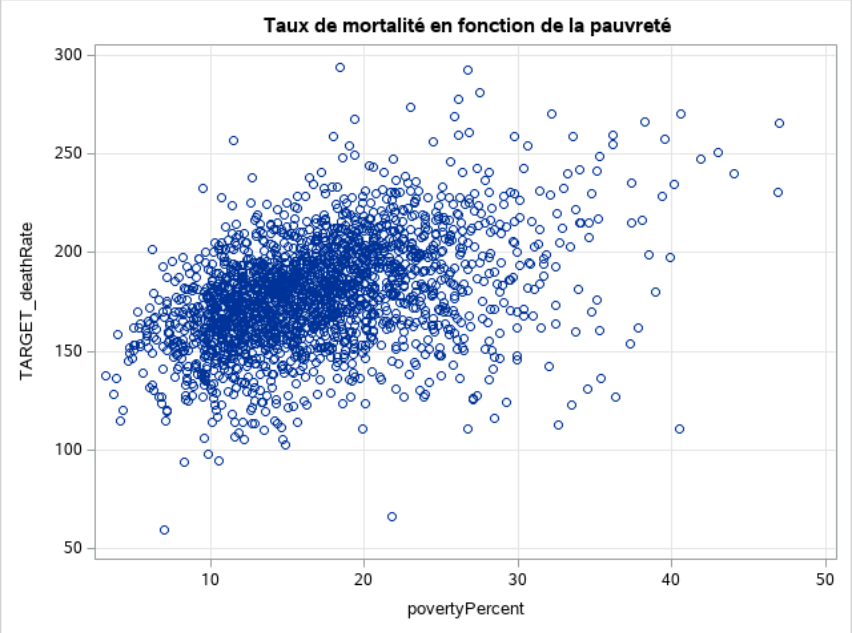

Fig. 7 - Autre type de graphique montrant la corrélation entre le taux de mortalité et la pauvreté

4. Modélisation

La variable d'intérêt étant quantitative et continue, nous effectuons une régression linéaire.

Réduction du modèle

Par la procédure "Stepwise" de SAS, on isole les variables les plus significatives. Après suppression des variables redondantes analysées en première partie, on a comme modèle final :

TARGET_deathRate = 127.17517 *(intercept)*
+ 0.19168 * incidenceRate
+ 0.45535 * povertyPercent
- 0.49875 * MedianAgeMale
+ 0.58154 * PercentMarried
+ 0.30447 * PctHS18_24
+ 0.45610 * PctHS25_Over
- 1.06097 * PctBachDeg25_Over
+ 0.53524 * PctUnemployed16_Over
- 0.50387 * PctPrivateCoverage
+ 0.23791 * PctEmpPrivCoverage
- 0.97990 * PctOtherRace
- 0.79399 * PctMarriedHouseholds
- 0.67198 * BirthRate

Le modèle explique alors environ 51% des variations du taux de mortalité lié au cancer.

On remarquera qu'avec les 13 variables retenues, on explique aussi 51% des variations du taux de mortalité. Étant donné ce faible pourcentage, on les conservera toutes afin de maximiser la pertinence du modèle. Même si on notera que les R^2 partiels sont déjà très faibles à partir de la huitième variable ajoutée par la procédure stepwise. Donc conserver les 8 première variable expliquerait quand même environ 50% des variations du taux de mortalité, avec un autre modèle.

Toutes les variables étant significatives, on passe à la validation du modèle.

On peut aussi noter qu'avec la régression, nous sommes passées de 34 variables explicatives à 13 variables.

Validation du modèle

Les résidus sont bien centrés, normalement distribués et de variable constante (Fig. 8 : centrage en 0 et répartition vraisemblablement uniforme des valeurs).

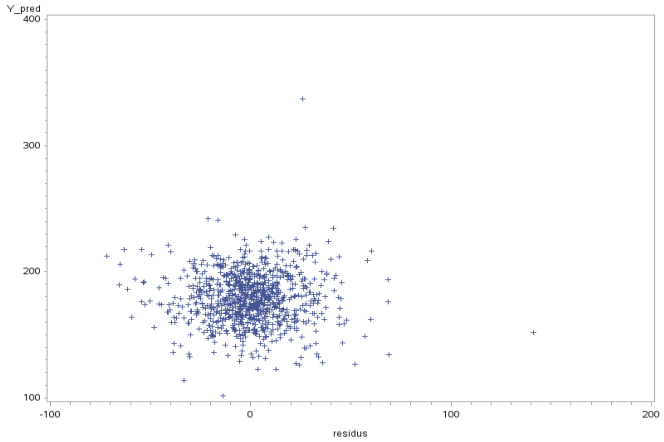

Fig. 8 - Taux de mortalité prédits en fonction des résidus

On valide finalement le modèle avec une comparaison graphique du taux de mortalité avec les valeurs prédites du taux (Fig. 9).

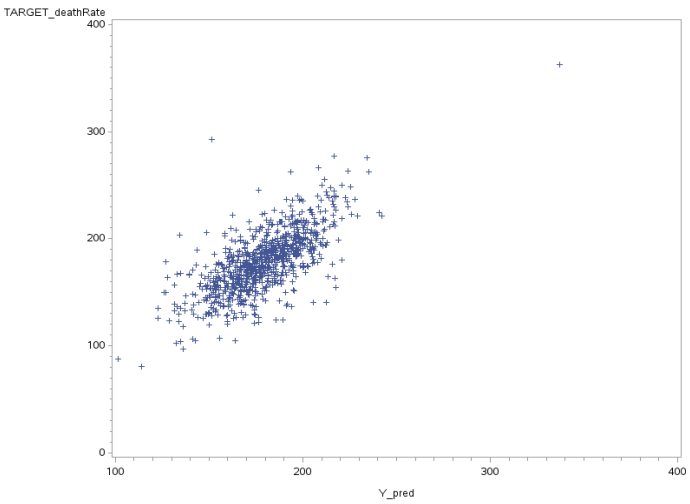

Fig. 9 - Taux de mortalité en fonction des taux prédits

On peut parfois relever une surestimation du taux de mortalité, avec par exemple un taux à 200 pour 100 000 habitants, pour une valeur prédite autour de 150.

Conclusion

Ainsi, le modèle explique environ la moitié des variations du taux de mortalité. Ce faible résultat nous laisse penser qu'il y aurait éventuellement des facteurs inconnus qui n'ont pas été pris en compte dans la base de données.

Par ailleurs, la variable cible peut s'expliquer par treize variables plus ou moins corrélées les unes des autres notamment le taux de pauvreté qui semble très lié au taux de mortalité: l'accès aux besoins primaires est un aspect déterminant quant à la longévité. De même le Pct Private Coverage stipulerait que la mutuelle privée donnerait accès à une meilleure santé. Et enfin le soutien émotionnel au sein d'un foyer d'un couple marié aurait une incidence positive sur la diminution du taux de mortalité.