Université Sidi Mohammed ben Abdellah
Faculté des Sciences Dhar El Mahraz
Département Informatique
Web Intelligence et Science des Données
Année Universitaire 2021/2022

Master WISD

Introduction à la Modélisation et à la Simulation

Prédiction du Taux de Mortalité par Cancer

Régression Linéaire Multiple

Réalisé par:
Youness HOURRI
Mohammed BATI
Ibrahim CHERRATE

Encadré par:
Pr. Abdelkamel ALJ

1 Juin 2022

Résumé

Une estimation de 8,2. millions de personnes sont mortes du cancer en 2012 dans le monde. La proportion de décès attribués au cancer varie considérablement entre les régions du monde et au sein de celles-ci, les taux ayant tendance à être plus élevés dans les régions les plus développées du monde, bien que les taux de mortalité de plusieurs formes de cancer soient élevés dans de nombreuses économies à revenu faible ou intermédiaire. Dans le monde, la cause la plus fréquente de décès par cancer est le poumon, qui représente près d'un cinquième de la charge globale de mortalité par cancer. Le cancer du foie se classe au deuxième rang (environ 9,1 % de tous les décès par cancer), suivi de près par le cancer de l'estomac (8,8 %). Dans cet article, nous discutons des caractéristiques, des sources de données et de la disponibilité de la mortalité par cancer dans le monde et décrivons certaines des variations entre les pays et heures supplémentaires. Nous examinons tous les néoplasmes combinés, ainsi que les trois causes les plus fréquentes de décès par cancer : les néoplasmes du poumon, du foie et de l'estomac. Nous considérons également les tendances de la mortalité par cancer comme une mesure des progrès dans la lutte contre le cancer et fournissons des estimations du futur fardeau de la mortalité par cancer pour l'année 2025.

Table des matières

1 Modèle linéaire multiple

1.1 Définition

En statistiques, en économétrie et en apprentissage automatique, un modèle de régression linéaire est un modèle de régression qui cherche à établir une relation linéaire entre une variable, dite expliquée, et une ou plusieurs variables, dites explicatives.

Dans la régression linéaire, les relations sont modélisées à l'aide de fonctions prédictives linéaires dont les paramètres de modèle inconnus sont estimés à partir des données. Ces modèles sont appelés modèles linéaires. Le plus souvent, l'espérance conditionnelle[1] de la réponse donnée aux valeurs des variables explicatives (ou prédicteurs) est supposée être une fonction affine[2] de ces valeurs; moins fréquemment, la médiane conditionnelle ou un autre quantile est utilisé. Comme toutes les formes d'analyse de régression, la régression linéaire se concentre sur la distribution de probabilité conditionnelle de la réponse compte tenu des valeurs des prédicteurs, plutôt que sur la distribution de probabilité conjointe de toutes ces variables, qui est le domaine de l'analyse multivariée.

La régression linéaire a été le premier type d'analyse de régression à être étudié rigoureusement et à être largement utilisé dans des applications pratiques. En effet, les modèles qui dépendent linéairement de leurs paramètres inconnus sont plus faciles à ajuster que les modèles qui ne sont pas liés de manière linéaire à leurs paramètres et parce que les propriétés statistiques des estimateurs[3] résultants sont plus faciles à déterminer.

La régression linéaire a de nombreuses utilisations pratiques. La plupart des applications appartiennent à l'une des deux grandes catégories suivantes :

- Si l'objectif est la prédiction, la prévision ou la réduction des erreurs, [clarification nécessaire] la régression linéaire peut être utilisée pour ajuster un modèle prédictif à un ensemble de données observées de valeurs de la réponse et de variables explicatives. Après avoir

[1] l'espérance conditionnelle d'une variable aléatoire réelle est, selon les cas, un nombre ou une nouvelle variable aléatoire.

[2] une fonction affine est une fonction obtenue par addition et multiplication de la variable par des constantes.

[3] un estimateur est une fonction permettant d'évaluer un paramètre inconnu relatif à une loi de probabilité

développé un tel modèle, si des valeurs supplémentaires des variables explicatives sont collectées sans valeur de réponse d'accompagnement, le modèle ajusté peut être utilisé pour faire une prédiction de la réponse.

- Si le but est d'expliquer la variation de la variable de réponse qui peut être attribuée à la variation des variables explicatives, une analyse de régression linéaire peut être appliquée pour quantifier la force de la relation entre la réponse et les variables explicatives, et en particulier pour déterminer si certaines les variables explicatives peuvent n'avoir aucune relation linéaire avec la réponse, ou pour identifier quels sous-ensembles de variables explicatives peuvent contenir des informations redondantes sur la réponse.

1.2 Applications

La régression linéaire est largement utilisée dans les sciences biologiques, comportementales et sociales pour décrire les relations possibles entre les variables. Il se classe comme l'un des outils les plus importants utilisés dans ces disciplines.

1.2.1 Machine Learning

Dans l'apprentissage automatique, le but de la régression est d'estimer une valeur (numérique) de sortie à partir des valeurs d'un ensemble de caractéristiques en entrée. Par exemple, estimer le prix d'une maison en se basant sur sa surface, nombre des étages, son emplacement, etc. Donc, le problème revient à estimer une fonction de calcul en se basant sur des données d'entrainement.

1.2.2 Economique

La régression linéaire est l'outil empirique prédominant en économie. Par exemple, il est utilisé pour prédire les dépenses de consommation, les dépenses d'investissement fixe, les investissements en stocks, les achats d'exportations d'un pays, les dépenses d'importation, la demande de détention d'actifs liquides, la demande de main-d'œuvre et l'offre de main-d'œuvre.

1.2.3 Finance

Le modèle d'évaluation des actifs financiers utilise la régression linéaire ainsi que le concept de bêta pour analyser et quantifier le risque systématique d'un investissement. Cela provient directement du coefficient bêta du modèle de régression linéaire qui relie le rendement de l'investissement au rendement de tous les actifs risqués.

1.2.4 Ligne de tendance

Une ligne de tendance représente une tendance, le mouvement à long terme des données de séries chronologiques après la prise en compte d'autres composants. Il indique si un ensemble de données particulier (par exemple, le PIB, les prix du pétrole ou les cours des actions) a augmenté ou diminué au cours de la période. Une ligne de tendance pourrait simplement être tracée à l'œil nu à travers un ensemble de points de données, mais plus précisément, leur position et leur pente sont calculées à l'aide de techniques statistiques telles que la régression linéaire. Les lignes de tendance sont généralement des lignes droites, bien que certaines variantes utilisent des polynômes de degré supérieur en fonction du degré de courbure souhaité dans la ligne. Les lignes de tendance sont parfois utilisées dans l'analyse commerciale pour montrer les changements dans les données au fil du temps. Cela a l'avantage d'être simple. Les lignes de tendance sont souvent utilisées pour affirmer qu'une action ou un événement particulier (tel qu'une formation ou une campagne publicitaire) a provoqué des changements observés à un moment donné. Il s'agit d'une technique simple qui ne nécessite pas de groupe témoin, de conception expérimentale ou de technique d'analyse sophistiquée. Cependant, il souffre d'un manque de validité scientifique dans les cas où d'autres changements potentiels peuvent affecter les données.

1.3 Modèle

1.3.1 Présentation formelle

Étant donné un ensemble de données $\{y_i,\, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$ de n unités statistiques, un modèle de régression linéaire suppose que la relation entre la variable dépendante y et le vecteur p des régresseurs x est linéaire. Cette relation est modélisée par un terme de perturbation ou une variable d'erreur ϵ - une variable aléatoire non observée qui ajoute du "bruit" à la relation linéaire entre la variable dépendante et les régresseurs. Ainsi le modèle prend la forme :

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^\mathsf{T}\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

Où T désigne la transposition, de sorte que $x_i^\mathsf{T}\boldsymbol{\beta}$ est le produit scalaire entre les vecteurs x_i et $\boldsymbol{\beta}$.
Souvent, ces n équations sont empilées et écrites en notation matricielle comme :

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

Où

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$$X = \begin{pmatrix} \mathbf{x}_1^\mathsf{T} \\ \mathbf{x}_2^\mathsf{T} \\ \vdots \\ \mathbf{x}_n^\mathsf{T} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

1.3.2 Notation et terminologie

- **$\mathbf{y}$:** est un vecteur de valeurs observées y_i $(i = 1, \ldots, n)$ de la variable appelée régressand, variable critère ou variable dépendante.
- **X :** peut être vu comme une matrice de vecteurs-lignes $\mathbf{x}_i$ ou de vecteurs-colonnes à n dimensions X_j, appelés régresseurs, variables explicatives ou variables indépendantes
- **$\boldsymbol{\beta}$:** est un vecteur de paramètre de $(p + 1)$ dimension, où β_0 est le terme d'interception (s'il y en a un dans le modèle, sinon $\boldsymbol{\beta}$ est de dimension p). Ses éléments sont appelés effets ou coefficients de régression
- **$\boldsymbol{\varepsilon}$:** est un vecteur de valeurs ε_i. Cette partie du modèle est appelée terme d'erreur, terme de perturbation ou parfois bruit

1.4 Estimations des paramètres

1.4.1 Hypothèses

Comme en régression simple, les hypothèses permettent de déterminer : les propriétés des estimateurs (biais, convergence) ; et leurs distributions (pour les estimations d'intervalle et les tests d'hypothèses).

- **H1 :** Les X_j sont déterminées sans erreurs, j = 1, . . . , p.
- **H2 :** $E(\varepsilon_i) = 0$ Le modèle est bien spécifié en moyenne.
- **H3 :** $\text{Var}(\varepsilon_i) = \sigma^2 \ \forall i$ Homoscédasticité des erreurs (variance constante).
- **H4 :** $\text{cov}(\varepsilon_i, \varepsilon_j) = 0 \ \forall i \neq j$ Pas d'autocorrélation des erreurs.
- **H5 :** $\text{cov}(X_i, \varepsilon_j) = 0 \ \forall i \neq j$ Les erreurs sont linéairement indépendantes des variables exogènes.
- **H6 :** $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$ Les erreurs suivent une loi normale multidimensionnelle.

1.4.2 Estimateur des moindres carrés ordinaires

Du modèle complet :

$$y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \epsilon_i$$

On va estimer les paramètres et obtenir:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_p x_{i,p}$$

Les résidus estimés sont la différence entre la valeur de y observée et estimée, Soit:

$$\hat{\epsilon}_i \equiv y_i - \hat{y}_i$$

Le principe des moindres carrés consiste à rechercher les valeurs des paramètres qui minimisent la somme des carrés des résidus

$$\min \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \min_{\hat{\beta}_0,.,\hat{\beta}_p} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \cdots - \hat{\beta}_p x_{i,p})^2$$

Ce qui revient à rechercher les solutions de $\frac{\partial(\sum \hat{\epsilon}_i^2)}{\partial \hat{\beta}_j} = 0$. Nous avons j=p+1 équations, dites équations normales, à résoudre. La solution obtenue est l'estimateur des moindres carrés ordinaires, il s'écrit :

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Propriétés des estimateurs: Si les hypothèses initiales sont respectées, l'estimateur des MCO possède d'excellentes propriétés

2 Analyse de données

2.1 Objectifs de projet

L'objectif de ce projet est de prédire les résultats des taux de mortalité par cancer, afin de construire un modèle de régression linéaire.

Source des données : ces données ont été récuperer à partir d'un certain nombre de sources, notamment l'American Community Survey [census.gov], National Library of Medicine [clinicaltrials.gov] et National Cancer Institute [cancer.gov]. La plupart du processus de préparation des données peut être visualisé ici [https://data.world/nrippner/cancer-trials].

Travail à faire

(i) La création d'un modèle de régression avec la méthode des moindres carrés ordinaire pour prédire "TARGET-deathRate"

(ii) La sortie du logiciel statistique, y compris le R-carré (ajusté) et l'erreur quadratique moyenne (RMSE)

(iii) L'étude sous R

(iv) Diagnostics du modèle comprenant des statistiques et des visualisations:

- Évaluer la linéarité du modèle (paramètres)
- Évaluer l'indépendance sérielle des erreurs
- Évaluer l'hétéroscédasticité
- Évaluer la normalité de la distribution résiduelle
- Évaluer la multicolinéarité

(v) L'interprétation du modèle

(vi) Autres facteurs à considérer:

- Y a-t-il des valeurs aberrantes ?
- Y a-t-il des valeurs manquantes ?
- Comment allez-vous gérer les variables catégorielles ?

2.2 Description du jeu de données

Le tableau contient les 34 caractéristiques suivantes :

- **TARGET-deathRate :** variable dépendante. Mortalités moyennes par habitant (100 000) par cancer(a)
- **avgAnnCount :** nombre moyen de cas déclarés de cancer diagnostiqués chaque année(a)
- **avgDeathsPerYear :** nombre moyen de décès déclarés dus au cancer(a)
- **taux d'incidence :** nombre moyen par habitant (100 000) de diagnostics de cancer(a)
- **medianIncome :** revenu médian par comté (b)
- **popEst2015 :** Population du comté (b)
- **PovertyPercent :** Pourcentage de la population vivant dans la pauvreté (b)
- **studyPerCap :** nombre par habitant d'essais cliniques liés au cancer par comté (a)
- **binnedInc :** Revenu médian par habitant regroupé par décile (b)
- **MedianAge :** âge médian des résidents du comté (b)
- **MedianAgeMale :** âge médian des résidents masculins du comté (b)
- **MedianAgeFemale :** âge médian des femmes résidant dans le comté (b)
- **Geography :** Nom du comté (b)
- **AvgHouseholdSize :** taille moyenne des ménages du comté (b)
- **PercentMarried :** Pourcentage de résidents du comté qui sont mariés (b)
- **PctNoHS18-24 :** Pourcentage de résidents du comté âgés de 18 à 24 ans ayant atteint le niveau d'études le plus élevé : inférieur au lycée (b)

- **PctHS18-24 :** Pourcentage de résidents du comté âgés de 18 à 24 ans ayant atteint le niveau d'études le plus élevé : diplôme d'études secondaires (b)
- **PctSomeCol18-24 :** Pourcentage de résidents du comté âgés de 18 à 24 ans ayant atteint le niveau d'études le plus élevé : certaines études collégiales (b)
- **PctBachDeg18-24 :** pourcentage de résidents du comté âgés de 18 à 24 ans ayant atteint le niveau d'études le plus élevé : baccalauréat (b)
- **PctHS25-Over :** pourcentage de résidents du comté âgés de 25 ans et plus ayant atteint le niveau d'études le plus élevé : diplôme d'études secondaires (b)
- **PctBachDeg25-Over :** Pourcentage de résidents du comté âgés de 25 ans et plus ayant atteint le niveau d'études le plus élevé : licence (b)
- **PctEmployed16-Over :** Pourcentage de résidents du comté âgés de 16 ans et plus employés (b)
- **PctUnemployed16-Over :** pourcentage de résidents du comté âgés de 16 ans et plus sans emploi (b)
- **PctPrivateCoverage :** pourcentage de résidents du comté bénéficiant d'une couverture médicale privée (b)
- **PctPrivateCoverageAlone :** pourcentage de résidents du comté bénéficiant uniquement d'une couverture médicale privée (pas d'assistance publique) (b)
- **PctEmpPrivCoverage :** pourcentage de résidents du comté bénéficiant d'une couverture médicale privée fournie par l'employé (b)
- **PctPublicCoverage :** Pourcentage de résidents du comté bénéficiant d'une couverture maladie fournie par le gouvernement (b)
- **PctPubliceCoverageAlone :** pourcentage de résidents du comté bénéficiant uniquement d'une couverture maladie fournie par le gouvernement (b)
- **PctWhite :** Pourcentage de résidents du comté qui s'identifient comme blancs (b)

- **PctBlack :** Pourcentage de résidents du comté qui s'identifient comme noirs (b)
- **PctAsian :** pourcentage de résidents du comté qui s'identifient comme asiatiques (b)
- **PctOtherRace :** pourcentage de résidents du comté qui s'identifient dans une catégorie qui n'est pas blanche, noire ou asiatique (b)
- **PctMarriedHouseholds :** Pourcentage de ménages mariés (b)
- **BirthRate :** nombre de naissances vivantes par rapport au nombre de femmes dans le comté (b)

(a) : années 2010-2016

(b) : Estimations du recensement de 2013

2.3 Préparation de données sous R

Importation des données :

```
install.packages("data.table")
library("data.table")

data <- fread("data/cancer_reg.csv", select =
                c("TARGET_deathRate",
                "incidenceRate",
                "PctBachDeg25_Over",
                "PctOtherRace"))
View(data)
```

Affichage des données :

```
   TARGET_deathRate incidenceRate PctBachDeg25_Over PctOtherRace
1:            164.9      489.8000              19.6    1.8434785
2:            161.3      411.6000              22.7    3.7413515
3:            174.7      349.7000              16.0    2.7473583
4:            194.8      430.4000               9.3    1.3626432
```

```
   5:          144.4      350.1000          15.0     0.4921355
  ---          ---        ---               ---      ---
3043:          149.6      453.5494          15.2     1.7004680
3044:          150.1      453.5494          12.4    14.1302884
3045:          153.9      453.5494          12.8     5.6807052
3046:          175.0      453.5494          14.4     2.1317905
3047:          213.6      453.5494          13.7     1.3564574
```

3 Mise en œuvre

3.1 Division de données

Avant de créer le modèle, nous devons diviser les données en un ensemble de données de formation et un ensemble de données de test. Mais d'abord, pour des raisons de simplicité, nous allons éliminer tous les facteurs catégoriels de notre ensemble de données et n'utiliser que les variables numériques ainsi que les variables telles que ID et CarName qui sont inutiles.

```
canser_dt <- subset(canser_dt, select = c('avgAnnCount', 'avgDeathsPerYear', 'AvgHouseholdSize', 'BirthRate',
                                          'incidenceRate', 'MedianAge', 'MedianAgeFemale', 'MedianAgeMale', 'medIncome', 'PctAsian',
                                            'PctBachDeg18_24', 'PctBachDeg25_Over', 'PctBlack', 'PctEmployed16_Over', 'PctEmpPrivCoverage',
                                            'PctHS18_24', 'PctHS25_Over', 'PctMarriedHouseholds', 'PctNoHS18_24', 'PctOtherRace',
                                            'PctPrivateCoverage', 'PctPrivateCoverageAlone', 'PctPublicCoverage', 'PctPublicCoverageAlone',
                                            'PctSomeCol18_24', 'PctUnemployed16_Over','PctWhite', 'PercentMarried', 'popEst2015', 'povertyPercent',
                                            'studyPerCap', 'TARGET_deathRate'))
```

Nous utiliserons l'ensemble de données de formation pour former le modèle de régression linéaire. L'ensemble de données de test sera utilisé comme comparaison pour voir si le modèle est trop ajusté et ne peut pas prédire de nouvelles données qui n'ont pas été vues pendant la phase
de formation. Nous utiliserons 80formation et le reste comme données de test.

```
# split to test and training dataset
set.seed(123)
samplesize <- round(0.8 * nrow(canser_dt), 0)
index <- sample(seq_len(nrow(canser_dt)), size = samplesize)

data_train <- canser_dt[index, ]
data_test <- canser_dt[-index, ]
```

3.2 Construction du modèle de régression

3.2.1 Le modèle initial

Dans tous les prochains tests de ce rapport, nous utiliserons la fonction "lm" qui est utilisée pour ajuster des modèles linéaires. Elle peut être utilisée pour effectuer une régression, une analyse de variance à strate unique et une analyse de covariance. Et maintenant, nous allons construire le modèle initial pour notre ensemble de données en utilisant toutes les variables de l'ensemble de données.

```
# calculer le modele de regression
reg <- lm(TARGET_deathRate ~ ., data = data_train)
summary(reg)
```

```
Call:
lm(formula = TARGET_deathRate ~ ., data = data_train)

Residuals:
    Min      1Q  Median      3Q     Max
-71.036 -10.709   0.109  10.348 109.720

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.738e+03  1.774e+03   0.979 0.327882
avgAnnCount            -1.610e-03  1.984e-03  -0.811 0.417670
avgDeathsPerYear        8.404e-03  9.844e-03   0.854 0.393738
AvgHouseholdSize        4.105e-01  2.550e+00   0.161 0.872198
BirthRate              -4.738e-01  5.309e-01  -0.893 0.372604
incidenceRate           1.519e-01  1.924e-02   7.895 2.29e-14 ***
MedianAge              -2.154e-02  1.827e-02  -1.179 0.239018
MedianAgeFemale         4.633e-01  5.868e-01   0.789 0.430241
MedianAgeMale          -9.069e-01  5.773e-01  -1.571 0.116948
medIncome               2.013e-04  2.397e-04   0.840 0.401452
PctAsian                2.700e-01  7.146e-01   0.378 0.705689
PctBachDeg18_24        -1.655e+01  1.772e+01  -0.934 0.350636
PctBachDeg25_Over      -1.317e+00  4.459e-01  -2.954 0.003302 **
PctBlack                1.572e-01  1.563e-01   1.006 0.314962
PctEmployed16_Over     -6.037e-01  2.987e-01  -2.021 0.043854 *
PctEmpPrivCoverage      5.396e-01  3.440e-01   1.569 0.117410
PctHS18_24             -1.580e+01  1.772e+01  -0.891 0.373233
PctHS25_Over            3.271e-01  2.830e-01   1.156 0.248310
PctMarriedHouseholds   -2.080e+00  4.363e-01  -4.766 2.55e-06 ***
PctNoHS18_24           -1.611e+01  1.771e+01  -0.910 0.363495
PctOtherRace           -1.464e+00  4.281e-01  -3.421 0.000681 ***
PctPrivateCoverage     -1.308e-01  7.853e-01  -0.167 0.867788
PctPrivateCoverageAlone -1.947e-01  9.219e-01  -0.211 0.832827
PctPublicCoverage      -7.003e-01  9.477e-01  -0.739 0.460359
PctPublicCoverageAlone  1.322e+00  1.082e+00   1.222 0.222410
PctSomeCol18_24        -1.581e+01  1.772e+01  -0.892 0.372845
PctUnemployed16_Over    1.051e-01  4.806e-01   0.219 0.826936
PctWhite                2.097e-02  1.646e-01   0.127 0.898640
PercentMarried          2.049e+00  4.608e-01   4.447 1.10e-05 ***
popEst2015             -4.812e-06  1.222e-05  -0.394 0.693950
povertyPercent          3.223e-01  4.506e-01   0.715 0.474790
studyPerCap            -1.422e-04  2.745e-03  -0.052 0.958700
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.66 on 445 degrees of freedom
  (1961 observations effacées parce que manquantes)
Multiple R-squared:  0.4828,    Adjusted R-squared:  0.4467
F-statistic:  13.4 on 31 and 445 DF,  p-value: < 2.2e-16
```

Le résumé du modèle reg montre beaucoup d'informations. Mais pour l'instant, il est préférable de se concentrer sur le $Pr(>|\ t\ |)$. Cette colonne montre le niveau de signification de la variable par rapport au modèle. Si la valeur est inférieure à 0,05, nous pouvons supposer sans risque que la variable a un effet significatif sur le modèle (ce qui signifie que le coefficient estimé n'est pas différent de 0), et vice versa. Ainsi, nous pouvons créer un modèle plus simple en supprimant les variables dont la valeur p est inférieure ou égale à 0,05, car elles n'ont pas d'effet significatif sur notre modèle. La valeur estimée indique le coefficient de chaque variable.

3.2.2 Le facteur d'inflation de la variance

A) **Qu'est-ce qu'un facteur d'inflation de la variance ?**

Un facteur d'inflation de la variance (VIF) détecte la multicollinéarité dans l'analyse de régression. On parle de multicollinéarité lorsqu'il existe une corrélation entre les prédicteurs (c'est-à-dire les variables indépendantes) d'un modèle ; sa présence peut avoir une incidence négative sur les résultats de la régression. Le VIF estime dans quelle mesure la variance d'un coefficient de régression est gonflée en raison de la multicollinéarité du modèle.

Les VIF sont généralement calculés par un logiciel, dans le cadre d'une analyse de régression. Vous verrez une colonne VIF dans le résultat. Les VIF sont calculés en prenant un prédicteur et en le régressant par rapport à tous les autres prédicteurs du modèle. Cela vous donne les valeurs de R-carré, qui peuvent ensuite être insérées dans la formule VIF. "i" est le prédicteur que vous examinez (par exemple, x1 ou x2) :

$$VIF = \frac{1}{R_i^2}$$

B) Interprétation du facteur d'inflation de la variance

Les facteurs d'inflation de la variance vont de 1 à plus. La valeur numérique du VIF vous indique (sous forme décimale) quel pourcentage de la variance (c'est-à-dire l'erreur standard au carré) est gonflé pour chaque coefficient. Par exemple, un VIF de 1,9 indique que la variance d'un coefficient particulier est 90% plus élevée que ce à quoi on pourrait s'attendre s'il n'y avait pas de multicollinéarité

- s'il n'y avait pas de corrélation avec les autres prédicteurs. Une règle empirique pour interpréter le facteur d'inflation de la variance :
  (a) 1 = non corrélé.
  (b) entre 1 et 5 = corrélation modérée.
  (c) Supérieur à 5 = forte corrélation.

La taille exacte d'un VIF avant qu'il ne pose problème est un sujet de débat. Ce que l'on sait, c'est que plus votre VIF augmente, moins vos résultats de régression seront fiables. En général, un VIF supérieur à 10 indique une corrélation élevée et est source d'inquiétude. Certains auteurs suggèrent un niveau plus prudent de 2,5 ou plus.

C) Calcul d'un nouveau modèle en éliminant les facteurs avec un VIF élevé

```
> vif(reg)
            avgAnnCount         avgDeathsPerYear         AvgHouseholdSize               BirthRate           incidenceRate
              20.699153                65.855987                 1.446316                 1.206919                 1.247784
              MedianAge           MedianAgeFemale             MedianAgeMale                medIncome                PctAsian
               1.044413                10.371024                 9.960623                 8.567281                 2.511958
        PctBachDeg18_24          PctBachDeg25_Over                PctBlack      PctEmployed16_Over      PctEmpPrivCoverage
            6511.937547                 6.136502                 5.165190                 6.102442                11.920201
             PctHS18_24              PctHS25_Over     PctMarriedHouseholds            PctNoHS18_24             PctOtherRace
           31168.225405                 4.304914                 8.311341             21958.335317                 1.881875
    PctPrivateCoverage PctPrivateCoverageAlone       PctPublicCoverage  PctPublicCoverageAlone        PctSomeCol18_24
             75.127521                92.707107                57.726627                45.985792             44525.820114
  PctUnemployed16_Over               PctWhite          PercentMarried               popEst2015            povertyPercent
             2.914539                 7.288456                10.132561                51.800336                 9.075286
           studyPerCap
             1.107258
```

En appliquant la fonction vif au modèle précédent, nous décidons de supprimer la variable *PctSomeCol*18_24 ayant la valeur vif la plus élevée et de construire un nouveau modèle.

```
# remove PctSomeCol18_24
canser_dt1 <- subset(canser_dt, select = c('avgAnnCount', 'avgDeathsPerYear', 'AvgHouseholdSize', 'BirthRate',
                                           'incidenceRate', 'MedianAge', 'MedianAgeFemale', 'MedianAgeMale', 'medIncome', 'PctAsian',
                                           'PctBachDeg18_24', 'PctBachDeg25_Over', 'PctBlack', 'PctEmployed16_Over', 'PctEmpPrivCoverage',
                                           'PctHS18_24', 'PctHS25_Over', 'PctMarriedHouseholds', 'PctNoHS18_24', 'PctOtherRace',
                                           'PctPrivateCoverage', 'PctPrivateCoverageAlone', 'PctPublicCoverage', 'PctPublicCoverageAlone',
                                           'PctUnemployed16_Over','PctWhite', 'PercentMarried', 'popEst2015', 'povertyPercent',
                                           'studyPerCap', 'TARGET_deathRate'))

set.seed(123)
samplesize <- round(0.8 * nrow(canser_dt1), 0)
index <- sample(seq_len(nrow(canser_dt1)), size = samplesize)

data_train1 <- canser_dt1[index, ]
data_test1 <- canser_dt1[-index, ]

reg1 = lm(TARGET_deathRate~ . , data=data_train1)
summary(reg1)
```

```
Call:
lm(formula = TARGET_deathRate ~ ., data = data_train1)

Residuals:
    Min      1Q  Median      3Q     Max
-94.313 -10.509  -0.471  10.615 133.236

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.803e+02  2.054e+01   8.775  < 2e-16 ***
avgAnnCount             -2.953e-03  1.013e-03  -2.915  0.00360 **
avgDeathsPerYear         1.625e-02  5.311e-03   3.059  0.00225 **
AvgHouseholdSize         1.146e+00  1.178e+00   0.973  0.33060
BirthRate               -1.002e+00  2.372e-01  -4.223 2.53e-05 ***
incidenceRate            1.790e-01  9.747e-03  18.363  < 2e-16 ***
MedianAge                5.220e-03  1.035e-02   0.504  0.61423
MedianAgeFemale         -3.228e-01  2.826e-01  -1.142  0.25350
MedianAgeMale           -3.221e-01  2.662e-01  -1.210  0.22640
medIncome                2.028e-04  1.032e-04   1.964  0.04966 *
PctAsian                -1.069e-01  2.367e-01  -0.452  0.65144
PctBachDeg18_24         -1.576e-01  1.373e-01  -1.148  0.25131
PctBachDeg25_Over       -1.563e+00  2.050e-01  -7.624 3.94e-14 ***
PctBlack                -2.133e-02  7.251e-02  -0.294  0.76866
PctEmployed16_Over      -6.321e-01  1.447e-01  -4.370 1.32e-05 ***
PctEmpPrivCoverage       4.427e-01  1.635e-01   2.707  0.00685 **
PctHS18_24               1.690e-01  6.491e-02   2.604  0.00928 **
PctHS25_Over             2.666e-01  1.277e-01   2.088  0.03693 *
PctMarriedHouseholds    -1.601e+00  2.160e-01  -7.410 1.92e-13 ***
PctNoHS18_24            -1.974e-01  7.297e-02  -2.705  0.00690 **
PctOtherRace            -9.639e-01  1.659e-01  -5.810 7.35e-09 ***
PctPrivateCoverage      -4.260e-01  3.349e-01  -1.272  0.20357
PctPrivateCoverageAlone  3.564e-04  4.037e-01   0.001  0.99930
PctPublicCoverage       -1.961e-01  4.029e-01  -0.487  0.62660
PctPublicCoverageAlone   3.066e-01  4.624e-01   0.663  0.50731
PctUnemployed16_Over     3.224e-02  2.193e-01   0.147  0.88314
PctWhite                -1.419e-01  7.532e-02  -1.885  0.05966 .
PercentMarried           1.599e+00  2.227e-01   7.182 1.00e-12 ***
```

```
popEst2015              -1.333e-05  7.136e-06  -1.868  0.06193 .
povertyPercent           3.227e-01  2.100e-01   1.536  0.12461
studyPerCap             -3.903e-04  7.703e-04  -0.507  0.61247
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.74 on 1821 degrees of freedom
  (586 observations effacées parce que manquantes)
Multiple R-squared:  0.5013,    Adjusted R-squared:  0.4931
F-statistic: 61.03 on 30 and 1821 DF,  p-value: < 2.2e-16
```

Une fois de plus, nous appliquons la fonction "vif" à ce nouveau modèle, et nous continuons à faire la même chose jusqu'à ce que nous éliminions toutes les variables avec une valeur VIF élevée.

```
> vif(reg1)
             avgAnnCount           avgDeathsPerYear           AvgHouseholdSize                 BirthRate             incidenceRate                  MedianAge 
               11.188394                 38.223224                  1.312255                  1.206976                  1.291345                  1.026763 
       MedianAgeFemale             MedianAgeMale                 medIncome                   PctAsian          PctBachDeg18_24          PctBachDeg25_Over 
               10.424704                  9.045158                  7.553138                  1.955342                  1.973999                  5.778507 
             PctBlack          PctEmployed16_Over          PctEmpPrivCoverage                PctHS18_24                PctHS25_Over      PctMarriedHouseholds 
                5.246309                  6.768629                 11.484075                  1.637070                  3.828123                  9.478231 
           PctNoHS18_24                PctOtherRace       PctPrivateCoverage  PctPrivateCoverageAlone       PctPublicCoverage  PctPublicCoverageAlone 
                1.698117                  1.577499                 61.955642                 79.754590                 47.733797                 39.205532 
   PctUnemployed16_Over                   PctWhite             PercentMarried                popEst2015           povertyPercent             studyPerCap 
                2.776785                  7.414875                 11.353659                 32.076986                  8.781123                  1.048268 
```

La prochaine variable que nous devons supprimer de notre modèle est *PctPrivateCoverageAlone*

```
# remove PctPrivateCoverageAlone
canser_dt2 <- subset(canser_dt, select = c('avgAnnCount', 'avgDeathsPerYear', 'AvgHouseholdSize', 'BirthRate',
                                           'incidenceRate', 'MedianAge', 'MedianAgeFemale', 'MedianAgeMale', 'medIncome', 'PctAsian',
                                           'PctBachDeg18_24', 'PctBachDeg25_Over', 'PctBlack', 'PctEmployed16_Over', 'PctEmpPrivCoverage',
                                           'PctHS18_24', 'PctHS25_Over', 'PctMarriedHouseholds', 'PctNoHS18_24', 'PctOtherRace',
                                           'PctPrivateCoverage','PctPublicCoverage', 'PctPublicCoverageAlone',
                                           'PctUnemployed16_Over','PctWhite', 'PercentMarried', 'popEst2015', 'povertyPercent',
                                           'studyPerCap', 'TARGET_deathRate'))

set.seed(123)
samplesize <- round(0.8 * nrow(canser_dt2), 0)
index <- sample(seq_len(nrow(canser_dt2)), size = samplesize)

data_train2 <- canser_dt2[index, ]
data_test2 <- canser_dt2[-index, ]

reg2 = lm(TARGET_deathRate~ . , data=data_train2)
summary(reg2)
```

```
Call:
lm(formula = TARGET_deathRate ~ ., data = data_train2)

Residuals:
    Min      1Q  Median      3Q     Max 
-96.818 -10.489  -0.644  10.935 133.129 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)             1.733e+02  1.836e+01   9.436  < 2e-16 ***
avgAnnCount            -3.341e-03  9.109e-04  -3.667 0.000251 ***
avgDeathsPerYear        1.789e-02  4.475e-03   3.997 6.61e-05 ***
AvgHouseholdSize        6.033e-01  1.079e+00   0.559 0.576150    
BirthRate              -9.419e-01  2.157e-01  -4.368 1.31e-05 ***
incidenceRate           1.846e-01  8.670e-03  21.295  < 2e-16 ***
MedianAge               4.169e-03  8.826e-03   0.472 0.636706    
MedianAgeFemale        -2.966e-01  2.498e-01  -1.187 0.235231    
MedianAgeMale          -3.737e-01  2.372e-01  -1.575 0.115317    
medIncome               1.475e-04  9.254e-05   1.593 0.111201    
PctAsian               -2.147e-02  2.050e-01  -0.105 0.916588    
PctBachDeg18_24        -9.953e-02  1.204e-01  -0.826 0.408682    
PctBachDeg25_Over      -1.381e+00  1.825e-01  -7.569 5.42e-14 ***
PctBlack               -2.980e-02  6.372e-02  -0.468 0.639997    
PctEmployed16_Over     -6.634e-01  1.269e-01  -5.227 1.88e-07 ***
PctEmpPrivCoverage      4.055e-01  1.196e-01   3.390 0.000710 ***
PctHS18_24              2.354e-01  5.671e-02   4.151 3.43e-05 ***
PctHS25_Over            2.368e-01  1.122e-01   2.110 0.034934 *  
PctMarriedHouseholds   -1.457e+00  1.902e-01  -7.657 2.79e-14 ***
PctNoHS18_24           -1.426e-01  6.422e-02  -2.220 0.026522 *  
PctOtherRace           -9.692e-01  1.516e-01  -6.392 1.98e-10 ***
PctPrivateCoverage     -3.740e-01  1.503e-01  -2.489 0.012875 *  
PctPublicCoverage      -2.942e-01  2.613e-01  -1.126 0.260290    
PctPublicCoverageAlone  5.240e-01  3.241e-01   1.617 0.106037    
PctUnemployed16_Over   -4.123e-02  1.947e-01  -0.212 0.832325    
PctWhite               -1.109e-01  6.530e-02  -1.699 0.089472 .  
PercentMarried          1.478e+00  1.961e-01   7.537 6.91e-14 ***
```

```
popEst2015             -1.467e-05  6.123e-06  -2.397 0.016612 *
povertyPercent          3.317e-01  1.847e-01   1.797 0.072539 .
studyPerCap             9.351e-05  7.241e-04   0.129 0.897252
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.5 on 2284 degrees of freedom
  (124 observations effacées parce que manquantes)
Multiple R-squared:  0.5115,    Adjusted R-squared:  0.5053
F-statistic: 82.47 on 29 and 2284 DF,  p-value: < 2.2e-16
```

```
> vif(reg2)
          avgAnnCount         avgDeathsPerYear       AvgHouseholdSize             BirthRate          incidenceRate             MedianAge
            11.650792              35.431560               1.350423               1.205009               1.296563               1.023131
      MedianAgeFemale          MedianAgeMale              medIncome               PctAsian       PctBachDeg18_24      PctBachDeg25_Over
            10.744238               9.484154               7.632935               1.943280               1.960384               5.889163
            PctBlack     PctEmployed16_Over     PctEmpPrivCoverage            PctHS18_24          PctHS25_Over   PctMarriedHouseholds
             5.129329               6.933103               7.884971               1.629118               3.833022               9.318769
         PctNoHS18_24          PctOtherRace     PctPrivateCoverage     PctPublicCoverage PctPublicCoverageAlone   PctUnemployed16_Over
             1.686583               1.585699              16.069680              25.817157              24.420808               2.786770
            PctWhite        PercentMarried             popEst2015        povertyPercent          studyPerCap
             7.003957              11.237947              29.129541               8.717650               1.053017
> |
```

Sur la base de la dernière observation, la prochaine variable à supprimer est *avgDeathsPerYear*

```
# remove avgDeathsPerYear
canser_dt3 <- subset(canser_dt, select = c('avgAnnCount','AvgHouseholdSize', 'BirthRate',
                                           'incidenceRate', 'MedianAge', 'MedianAgeFemale', 'MedianAgeMale', 'medIncome', 'PctAsian',
                                           'PctBachDeg18_24', 'PctBachDeg25_Over', 'PctBlack', 'PctEmployed16_Over', 'PctEmpPrivCoverage',
                                           'PctHS18_24', 'PctHS25_Over', 'PctMarriedHouseholds', 'PctNoHS18_24', 'PctOtherRace',
                                           'PctPrivateCoverage','PctPublicCoverage', 'PctPublicCoverageAlone',
                                           'PctUnemployed16_Over','PctWhite', 'PercentMarried', 'popEst2015', 'povertyPercent',
                                           'studyPerCap', 'TARGET_deathRate'))

set.seed(123)
samplesize <- round(0.8 * nrow(canser_d3), 0)
index <- sample(seq_len(nrow(canser_dt3)), size = samplesize)

data_train3 <- canser_dt3[index, ]
data_test3 <- canser_dt3[-index, ]

reg3 = lm(TARGET_deathRate~ . , data=data_train3)
summary(reg3)
```

```
Call:
lm(formula = TARGET_deathRate ~ ., data = data_train3)

Residuals:
    Min      1Q  Median      3Q     Max
-99.084 -10.403  -0.379  10.794 135.113

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.765e+02  1.841e+01   9.588  < 2e-16 ***
avgAnnCount            -1.751e-03  8.221e-04  -2.130  0.03332 *
AvgHouseholdSize        6.496e-01  1.083e+00   0.600  0.54855
BirthRate              -9.637e-01  2.163e-01  -4.456 8.77e-06 ***
incidenceRate           1.866e-01  8.684e-03  21.489  < 2e-16 ***
MedianAge               3.409e-03  8.852e-03   0.385  0.70017
MedianAgeFemale        -2.747e-01  2.505e-01  -1.096  0.27306
MedianAgeMale          -3.870e-01  2.380e-01  -1.626  0.10409
medIncome               1.347e-04  9.279e-05   1.451  0.14688
PctAsian               -6.564e-02  2.053e-01  -0.320  0.74927
PctBachDeg18_24        -7.572e-02  1.207e-01  -0.627  0.53046
PctBachDeg25_Over      -1.335e+00  1.827e-01  -7.308 3.74e-13 ***
PctBlack               -2.355e-02  6.390e-02  -0.369  0.71247
PctEmployed16_Over     -6.965e-01  1.271e-01  -5.481 4.68e-08 ***
PctEmpPrivCoverage      4.428e-01  1.196e-01   3.701  0.00022 ***
PctHS18_24              2.378e-01  5.689e-02   4.179 3.04e-05 ***
PctHS25_Over            2.529e-01  1.125e-01   2.249  0.02464 *
PctMarriedHouseholds   -1.492e+00  1.907e-01  -7.824 7.79e-15 ***
PctNoHS18_24           -1.474e-01  6.442e-02  -2.288  0.02225 *
PctOtherRace           -9.827e-01  1.521e-01  -6.461 1.26e-10 ***
PctPrivateCoverage     -4.040e-01  1.506e-01  -2.683  0.00734 **
PctPublicCoverage      -2.725e-01  2.621e-01  -1.040  0.29856
PctPublicCoverageAlone  5.349e-01  3.251e-01   1.645  0.10006
PctUnemployed16_Over    2.819e-02  1.946e-01   0.145  0.88483
PctWhite               -1.042e-01  6.549e-02  -1.591  0.11176
PercentMarried          1.452e+00  1.966e-01   7.384 2.14e-13 ***
```

```
popEst2015              5.364e-06  3.527e-06   1.521  0.12841
povertyPercent          2.357e-01  1.837e-01   1.283  0.19948
studyPerCap             5.007e-05  7.263e-04   0.069  0.94504
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.56 on 2285 degrees of freedom
  (124 observations effacées parce que manquantes)
Multiple R-squared:  0.5081,    Adjusted R-squared:  0.5021
F-statistic:  84.3 on 28 and 2285 DF,  p-value: < 2.2e-16
```

```
> vif(reg3)
          avgAnnCount       AvgHouseholdSize             BirthRate          incidenceRate              MedianAge        MedianAgeFemale
             9.429879              1.350268              1.204237              1.292320              1.022657             10.739068
       MedianAgeMale              medIncome              PctAsian       PctBachDeg18_24      PctBachDeg25_Over              PctBlack
             9.482304              7.623782              1.937633              1.955590              5.865796              5.126241
   PctEmployed16_Over     PctEmpPrivCoverage            PctHS18_24            PctHS25_Over   PctMarriedHouseholds          PctNoHS18_24
             6.903576              7.837179              1.628942              3.828045              9.299030              1.685993
         PctOtherRace     PctPrivateCoverage PctPublicCoverage PctPublicCoverageAlone   PctUnemployed16_Over               PctWhite
             1.584911             16.029635             25.806017             24.419079              2.764605              6.999281
      PercentMarried            popEst2015       povertyPercent           studyPerCap
            11.225604              9.601440              8.570195              1.052780
```

Sur la base de la dernière observation, la prochaine variable à supprimer est *PctPublicCoverage*

```
#remove PctPublicCoverage
canser_dt4 <- subset(canser_dt, select = c('avgAnnCount','AvgHouseholdSize', 'BirthRate',
                                           'incidenceRate', 'MedianAge', 'MedianAgeFemale', 'MedianAgeMale', 'medIncome', 'PctAsian',
                                           'PctBachDeg18_24', 'PctBachDeg25_Over', 'PctBlack', 'PctEmployed16_Over', 'PctEmpPrivCoverage',
                                           'PctHS18_24', 'PctHS25_Over', 'PctMarriedHouseholds', 'PctNoHS18_24', 'PctOtherRace',
                                           'PctPrivateCoverage','PctPublicCoverageAlone',
                                           'PctUnemployed16_Over','PctWhite', 'PercentMarried', 'popEst2015', 'povertyPercent',
                                           'studyPerCap', 'TARGET_deathRate'))

set.seed(123)
samplesize <- round(0.8 * nrow(canser_dt4), 0)
index <- sample(seq_len(nrow(canser_dt4)), size = samplesize)

data_train4 <- canser_dt4[index, ]
data_test4 <- canser_dt4[-index, ]

reg4 = lm(TARGET_deathRate~ . , data=data_train4)
summary(reg4)
```

```
Call:
lm(formula = TARGET_deathRate ~ ., data = data_train4)

Residuals:
     Min       1Q   Median       3Q      Max
-99.171 -10.351  -0.464  10.776 134.810

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.753e+02  1.837e+01   9.543  < 2e-16 ***
avgAnnCount            -1.784e-03  8.215e-04  -2.171 0.030015 *
AvgHouseholdSize        6.599e-01  1.083e+00   0.610 0.542223
BirthRate              -9.622e-01  2.163e-01  -4.449 9.05e-06 ***
incidenceRate           1.867e-01  8.683e-03  21.501  < 2e-16 ***
MedianAge               3.482e-03  8.852e-03   0.393 0.694109
MedianAgeFemale        -3.528e-01  2.390e-01  -1.476 0.140115
MedianAgeMale          -4.250e-01  2.352e-01  -1.807 0.070851 .
medIncome               1.499e-04  9.163e-05   1.636 0.102062
PctAsian               -6.322e-02  2.053e-01  -0.308 0.758209
PctBachDeg18_24        -6.854e-02  1.205e-01  -0.569 0.569507
PctBachDeg25_Over      -1.317e+00  1.819e-01  -7.241 6.06e-13 ***
PctBlack               -1.556e-02  6.344e-02  -0.245 0.806303
PctEmployed16_Over     -6.624e-01  1.228e-01  -5.396 7.53e-08 ***
PctEmpPrivCoverage      4.916e-01  1.100e-01   4.468 8.30e-06 ***
PctHS18_24              2.416e-01  5.677e-02   4.256 2.16e-05 ***
PctHS25_Over            2.533e-01  1.125e-01   2.252 0.024414 *
PctMarriedHouseholds   -1.460e+00  1.881e-01  -7.757 1.30e-14 ***
PctNoHS18_24           -1.477e-01  6.442e-02  -2.293 0.021965 *
PctOtherRace           -9.809e-01  1.521e-01  -6.449 1.37e-10 ***
PctPrivateCoverage     -4.752e-01  1.341e-01  -3.542 0.000404 ***
PctPublicCoverageAlone  2.483e-01  1.724e-01   1.440 0.149970
PctUnemployed16_Over    7.329e-03  1.935e-01   0.038 0.969795
PctWhite               -1.039e-01  6.549e-02  -1.587 0.112639
PercentMarried          1.439e+00  1.962e-01   7.332 3.13e-13 ***


popEst2015              5.552e-06  3.522e-06   1.576 0.115092
povertyPercent          2.454e-01  1.834e-01   1.338 0.181113
studyPerCap             2.234e-05  7.259e-04   0.031 0.975448
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.56 on 2286 degrees of freedom
  (124 observations effacées parce que manquantes)
Multiple R-squared:  0.5079,    Adjusted R-squared:  0.5021
F-statistic: 87.37 on 27 and 2286 DF,  p-value: < 2.2e-16
```

```
> vif(reg4)
          avgAnnCount       AvgHouseholdSize             BirthRate          incidenceRate              MedianAge        MedianAgeFemale 
             9.415834              1.350155              1.204184              1.292184              1.022594              9.773877 
       MedianAgeMale             medIncome               PctAsian       PctBachDeg18_24       PctBachDeg25_Over             PctBlack 
             9.258257              7.434076              1.937384              1.949196              5.810346              5.051989 
   PctEmployed16_Over     PctEmpPrivCoverage             PctHS18_24             PctHS25_Over   PctMarriedHouseholds           PctNoHS18_24 
             6.444075              6.629333              1.621957              3.828000              9.055152              1.685955 
         PctOtherRace PctPrivateCoverage PctPublicCoverageAlone PctUnemployed16_Over            PctWhite        PercentMarried 
             1.584691             12.720324              6.866269              2.735218              6.999184             11.178967 
           popEst2015       povertyPercent          studyPerCap 
             9.576189              8.548215              1.051360 
```

Sur la base de la dernière observation, la prochaine variable à supprimer est *PctPrivateCoverage*

```
#remove PctPrivateCoverage
canser_dt5 <- subset(canser_dt, select = c('avgAnnCount','AvgHouseholdSize', 'BirthRate',
                                           'incidenceRate', 'MedianAge', 'MedianAgeFemale', 'MedianAgeMale', 'medIncome', 'PctAsian',
                                           'PctBachDeg18_24', 'PctBachDeg25_Over', 'PctBlack', 'PctEmployed16_Over', 'PctEmpPrivCoverage',
                                          'PctHS18_24', 'PctHS25_Over', 'PctMarriedHouseholds', 'PctNoHS18_24', 'PctOtherRace',
                                        'PctPublicCoverageAlone','PctUnemployed16_Over','PctWhite', 'PercentMarried', 'popEst2015', 'povertyPercent',
                                           'studyPerCap', 'TARGET_deathRate'))

set.seed(123)
samplesize <- round(0.8 * nrow(canser_dt5), 0)
index <- sample(seq_len(nrow(canser_dt5)), size = samplesize)

data_train5 <- canser_dt5[index, ]
data_test5 <- canser_dt5[-index, ]

reg5 = lm(TARGET_deathRate~ . , data=data_train5)
summary(reg5)
```

```
Call:
lm(formula = TARGET_deathRate ~ ., data = data_train5)

Residuals:
    Min      1Q  Median      3Q     Max 
-99.680 -10.567  -0.419  10.831 134.493 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)             1.473e+02  1.663e+01   8.861  < 2e-16 ***
avgAnnCount            -2.104e-03  8.186e-04  -2.570   0.0102 *  
AvgHouseholdSize        8.637e-01  1.084e+00   0.797   0.4256    
BirthRate              -1.057e+00  2.152e-01  -4.911 9.70e-07 ***
incidenceRate           1.824e-01  8.621e-03  21.161  < 2e-16 ***
MedianAge               3.849e-03  8.874e-03   0.434   0.6645    
MedianAgeFemale        -4.668e-01  2.374e-01  -1.966   0.0494 *  
MedianAgeMale          -3.935e-01  2.356e-01  -1.670   0.0950 .  
medIncome               1.627e-04  9.179e-05   1.773   0.0764 .  
PctAsian               -8.264e-02  2.058e-01  -0.402   0.6880    
PctBachDeg18_24        -9.199e-02  1.206e-01  -0.763   0.4457    
PctBachDeg25_Over      -1.422e+00  1.799e-01  -7.907 4.07e-15 ***
PctBlack               -3.023e-02  6.347e-02  -0.476   0.6339    
PctEmployed16_Over     -5.966e-01  1.217e-01  -4.904 1.00e-06 ***
PctEmpPrivCoverage      2.945e-01  9.517e-02   3.094   0.0020 ** 
PctHS18_24              2.789e-01  5.593e-02   4.986 6.62e-07 ***
PctHS25_Over            2.114e-01  1.121e-01   1.885   0.0596 .  
PctMarriedHouseholds   -1.431e+00  1.884e-01  -7.593 4.52e-14 ***
PctNoHS18_24           -8.466e-02  6.207e-02  -1.364   0.1727    
PctOtherRace           -9.547e-01  1.523e-01  -6.269 4.33e-10 ***
PctPublicCoverageAlone  6.019e-01  1.409e-01   4.271 2.03e-05 ***
PctUnemployed16_Over    7.315e-02  1.931e-01   0.379   0.7049    
PctWhite               -1.185e-01  6.552e-02  -1.809   0.0707 .  
PercentMarried          1.394e+00  1.963e-01   7.100 1.66e-12 ***


popEst2015              7.173e-06  3.501e-06   2.049   0.0406 *  
povertyPercent          3.837e-01  1.797e-01   2.135   0.0329 *  
studyPerCap            -5.114e-05  7.274e-04  -0.070   0.9440    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.61 on 2287 degrees of freedom
  (124 observations effacées parce que manquantes)
Multiple R-squared:  0.5052,    Adjusted R-squared:  0.4995 
F-statistic:  89.8 on 26 and 2287 DF,  p-value: < 2.2e-16
```

```
> vif(reg5)
          avgAnnCount       AvgHouseholdSize             BirthRate          incidenceRate              MedianAge        MedianAgeFemale 
             9.302045              1.346343              1.185854              1.267393              1.022454              9.596502 
       MedianAgeMale             medIncome               PctAsian       PctBachDeg18_24       PctBachDeg25_Over             PctBlack 
             9.245045              7.422432              1.936003              1.943315              5.654976              5.030459 
   PctEmployed16_Over     PctEmpPrivCoverage             PctHS18_24             PctHS25_Over   PctMarriedHouseholds           PctNoHS18_24 
             6.296622              4.934548              1.566322              3.785606              9.038470              1.557364 
         PctOtherRace PctPublicCoverageAlone PctUnemployed16_Over            PctWhite        PercentMarried           popEst2015 
             1.580955              4.564628              2.710007              6.971578             11.132022              9.414389 
       povertyPercent          studyPerCap 
             8.161326              1.050502 
```

Sur la base de la dernière observation, la prochaine variable à supprimer est *PercentMarried*

```
#remove PercentMarried
canser_dt6 <- subset(canser_dt, select = c('avgAnnCount','AvgHouseholdSize', 'BirthRate',
                                           'incidenceRate', 'MedianAge', 'MedianAgeFemale', 'MedianAgeMale', 'medIncome', 'PctAsian',
                                           'PctBachDeg18_24', 'PctBachDeg25_Over', 'PctBlack', 'PctEmployed16_Over', 'PctEmpPrivCoverage',
                                           'PctHS18_24', 'PctHS25_Over', 'PctMarriedHouseholds', 'PctNoHS18_24', 'PctOtherRace',
                                           'PctPublicCoverageAlone','PctUnemployed16_Over','PctWhite','popEst2015', 'povertyPercent',
                                           'studyPerCap', 'TARGET_deathRate'))

canser_dt6<-na.omit(canser_dt6)
set.seed(123)
samplesize <- round(0.8 * nrow(canser_dt6), 0)
index <- sample(seq_len(nrow(canser_dt6)), size = samplesize)

data_train6 <- canser_dt6[index, ]
data_test6 <- canser_dt6[-index, ]

reg6 = lm(TARGET_deathRate~ . , data=data_train6)
summary(reg6)
```

```
Call:
lm(formula = TARGET_deathRate ~ ., data = data_train6)

Residuals:
     Min       1Q   Median       3Q      Max
-109.342  -11.018   -0.129   10.883  139.655

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.142e+02  1.633e+01   6.992 3.54e-12 ***
avgAnnCount            -2.078e-03  8.406e-04  -2.472  0.01352 *
AvgHouseholdSize        9.133e-02  1.059e+00   0.086  0.93129
BirthRate              -6.229e-01  2.212e-01  -2.816  0.00491 **
incidenceRate           1.939e-01  8.664e-03  22.386  < 2e-16 ***
MedianAge               1.509e-03  8.646e-03   0.175  0.86145
MedianAgeFemale        -1.573e-01  2.390e-01  -0.658  0.51051
MedianAgeMale          -2.415e-01  2.339e-01  -1.032  0.30214
medIncome               1.010e-05  8.967e-05   0.113  0.91034
PctAsian                1.590e-02  2.145e-01   0.074  0.94090
PctBachDeg18_24         1.174e-01  1.247e-01   0.941  0.34681
PctBachDeg25_Over      -1.389e+00  1.765e-01  -7.865 5.62e-15 ***
PctBlack               -3.624e-02  6.601e-02  -0.549  0.58304
PctEmployed16_Over     -1.811e-01  1.027e-01  -1.764  0.07783 .
PctEmpPrivCoverage      1.791e-01  9.289e-02   1.928  0.05399 .
PctHS18_24              3.045e-01  5.544e-02   5.493 4.38e-08 ***
PctHS25_Over            2.835e-01  1.114e-01   2.544  0.01102 *
PctMarriedHouseholds   -2.496e-01  1.071e-01  -2.329  0.01992 *
PctNoHS18_24           -6.686e-02  6.358e-02  -1.051  0.29316
PctOtherRace           -8.593e-01  1.413e-01  -6.081 1.39e-09 ***
PctPublicCoverageAlone  5.556e-01  1.396e-01   3.981 7.08e-05 ***
PctUnemployed16_Over    1.939e-01  1.918e-01   1.011  0.31215
PctWhite               -8.624e-02  6.817e-02  -1.265  0.20598
popEst2015              6.981e-06  3.632e-06   1.922  0.05470 .
povertyPercent          4.393e-01  1.805e-01   2.433  0.01503 *
studyPerCap            -2.821e-04  7.522e-04  -0.375  0.70762
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.62 on 2290 degrees of freedom
Multiple R-squared:  0.5067,    Adjusted R-squared:  0.5013
F-statistic: 94.09 on 25 and 2290 DF,  p-value: < 2.2e-16
```

```
> vif(reg6)
          avgAnnCount        AvgHouseholdSize               BirthRate          incidenceRate               MedianAge         MedianAgeFemale
             8.194903                1.314799                1.174385                1.271671                1.023457                9.607596
       MedianAgeMale               medIncome                PctAsian       PctBachDeg18_24      PctBachDeg25_Over               PctBlack
             9.112420                7.282154                1.965262                1.979464                5.536721                5.330767
   PctEmployed16_Over     PctEmpPrivCoverage             PctHS18_24             PctHS25_Over   PctMarriedHouseholds           PctNoHS18_24
             4.484901                4.735009                1.526256                3.709118                2.955915                1.560593
        PctOtherRace PctPublicCoverageAlone   PctUnemployed16_Over               PctWhite            popEst2015         povertyPercent
             1.614789                4.466977                2.706414                7.318504                8.228485                8.324235
         studyPerCap
             1.048804
```

Et enfin, il n'y a plus de variables dans notre modèle avec un facteur d'inflation de variance élevé qui pourrait éventuellement affecter la fiabilité de notre modèle. Nous notons que le modèle final que nous avons trouvé est globalement significatif, mais il est toujours nécessaire de corriger ce modèle, même s'il est globalement significatif, car il y a toujours des erreurs puisque certains paramètres ne sont pas significatifs dans le modèle.

3.2.3 Sélection des variables

De nombreux critères de choix de modèle sont présentés dans la littérature sur la régression linéaire multiple, dans notre cas nous utiliserons le critère d'information $d'Akaike(AIC)$.

A) **Introduction à ł'AIC**

Le critère d'information $d'Akaike(AIC)$ vous permet de tester dans quelle mesure votre modèle s'adapte à l'ensemble des données sans les surdimensionner.
Le score AIC récompense les modèles qui obtiennent un score élevé de qualité d'ajustement et les pénalise s'ils deviennent trop complexes.
En soi, le score AIC n'est pas d'une grande utilité s'il n'est pas comparé au score AIC d'un modèle concurrent.
Le modèle ayant le score AIC le plus bas est censé trouver un meilleur équilibre entre sa capacité à s'ajuster à l'ensemble des données et sa capacité à éviter un ajustement excessif de l'ensemble des données.

B) Comment ça marche ?

$L'AIC$ fonctionne en évaluant l'ajustement du modèle sur les données d'apprentissage, et en ajoutant un terme de pénalité pour la complexité du modèle (principes fondamentaux similaires à la régularisation). Le résultat souhaité est de trouver l'AIC le plus bas possible, qui indique le meilleur équilibre entre l'ajustement du modèle et sa généralisation. Cela sert l'objectif final de maximiser l'ajustement sur les données hors échantillon.

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

Le score AIC n'est utile que lorsqu'il est utilisé pour comparer deux modèles. Disons que nous avons deux modèles de ce type avec un nombre k1 et k2 de paramètres, et des scores AIC_1 et AIC_2.
Supposons que AIC_1 ¡ AIC_2, c'est-à-dire que le modèle 1 est meilleur que le modèle 2.
Dans quelle mesure le modèle 2 est-il plus mauvais que le modèle 1 ?
On peut répondre à cette question en utilisant la formule suivante :

$$Probabilite\ relative = e^{\frac{AIC_1 - AIC_2}{2}}$$

Pourquoi utiliser la fonction exp() pour calculer la vraisemblance relative ? Pourquoi ne pas simplement soustraire $l'AIC_2$ de $l'AIC_1$?

D'une part, la fonction exp() garantit que la vraisemblance relative est toujours un nombre positif et donc plus facile à interpréter.

C) Méthodes de sélection

la méthode de sélection vous permet de spécifier comment les variables indépendantes sont entrées dans l'analyse. En utilisant différentes méthodes, vous pouvez construire une variété de modèles de régression à partir du même ensemble de variables. Nous utiliserons dans ce rapport deux méthodes différentes qui se basent sur le critère d'information d'Akaike (AIC) que nous avons expliqué plus haut pour le processus d'élimination.

-Élimination à rebours :
Procédure de sélection des variables dans laquelle toutes les variables sont introduites dans l'équation, puis éliminées de manière séquentielle. La variable présentant la plus faible corrélation partielle avec la variable dépendante est considérée en premier lieu pour être éliminée. Si elle répond au critère d'élimination, elle est supprimée. Une fois la première variable éliminée, la variable restante dans l'équation avec la plus petite corrélation partielle est considérée comme la suivante. La procédure s'arrête lorsqu'il n'y a plus de variables dans l'équation qui satisfont aux critères d'élimination.

-Sélection avant :
Une procédure de sélection des variables par étapes dans laquelle les variables sont introduites séquentiellement dans le modèle. La première variable prise en compte pour l'entrée dans l'équation est celle qui présente la plus grande corrélation positive ou négative avec la variable dépendante. Cette variable n'est introduite dans l'équation que si elle satisfait au critère d'introduction. Si la première variable est entrée, la variable indépendante qui n'est pas dans l'équation et qui présente la plus grande corrélation partielle est considérée comme la suivante. La procédure s'arrête lorsqu'il n'y a plus de variables qui répondent au critère d'entrée.

3.2.4 La rétro-élimination

Cette méthode est très similaire à la méthode d'élimination à rebours, mais nous partons de la régression constante jusqu'à ce que nous traitions toutes les autres variables que nous avons spécifiées, jusqu'à ce que nous trouvions

les meilleures variables pour notre modèle. Nous appliquons la sélection directe à partir de la régression constante en utilisant la fonction step.

```
# Forward Tracing With constant regression
step(lm(TARGET_deathRate~1, data=data_train6), scope=~.+avgAnnCount + AvgHouseholdSize + BirthRate + incidenceRate + MedianAge + MedianAgeFemale +
     MedianAgeMale + medIncome + PctAsian + PctBachDeg18_24 + PctBachDeg25_Over + PctBlack +
     PctEmployed16_Over + PctEmpPrivCoverage + PctHS18_24 + PctHS25_Over + PctMarriedHouseholds +
     PctNoHS18_24 + PctOtherRace + PctPublicCoverageAlone + PctUnemployed16_Over + PctWhite +
     popEst2015 + povertyPercent + studyPerCap, direction="forward", trace = TRUE)
```

Et après de nombreuses étapes de sélection, le meilleur modèle que nous avons pu trouver en utilisant cette méthode est :

```
Call:
lm(formula = TARGET_deathRate ~ PctBachDeg25_Over + incidenceRate +
    povertyPercent + PctHS18_24 + PctOtherRace + PctMarriedHouseholds +
    MedianAgeFemale + BirthRate + PctPublicCoverageAlone + PctHS25_Over +
    PctEmployed16_Over + PctEmpPrivCoverage + PctWhite + avgAnnCount +
    popEst2015, data = data_train6)

Coefficients:
          (Intercept)       PctBachDeg25_Over          incidenceRate         povertyPercent              PctHS18_24             PctOtherRace
            1.156e+02            -1.326e+00             1.955e-01             4.185e-01             3.066e-01            -8.579e-01
  PctMarriedHouseholds       MedianAgeFemale             BirthRate  PctPublicCoverageAlone          PctHS25_Over      PctEmployed16_Over
           -2.780e-01            -4.014e-01            -6.336e-01             5.922e-01             2.649e-01            -2.164e-01
    PctEmpPrivCoverage             PctWhite           avgAnnCount            popEst2015
            2.128e-01            -6.859e-02            -2.050e-03             7.236e-06
```

Les variables restantes après la sélection sont les suivantes :
$PctBachDeg25_Over$, $incidenceRate$, $povertyPercent$, $PctHS18_24$, $PctOtherRace$, $PctMarriedHouseholds$
$MedianAgeFemale, BirthRate, PctPublicCoverageAlone$,
$PctHS25_Over, PctEmployed16_Over, PctEmpPrivCoverage$,
$PctWhite, avgAnnCount$, $popEst2015$.
Nous pouvons maintenant construire un nouveau modèle linéaire en utilisant ces variables.

Si nous comparons le modèle que nous avons réussi à trouver en utilisant la méthode de sélection avant, et le modèle que nous trouvons à partir de la méthode d'élimination arrière, nous concluons que ce sont les mêmes modèles avec les mêmes variables.

3.2.5 Élimination des variables non significatives

Après de nombreux tests et étapes d'élimination et de sélection, nous avons réussi à construire un modèle formé par les variables
Mais nous pouvons remarquer, d'après le résumé du modèle, que le variables $PctWhite$ n'est pas significative pour le modèle, et que nous devrions probablement les éliminer. Pour voir si cette action affecte notre modèle, nous pouvons vérifier la valeur du R-carré ajusté du modèle précédent.

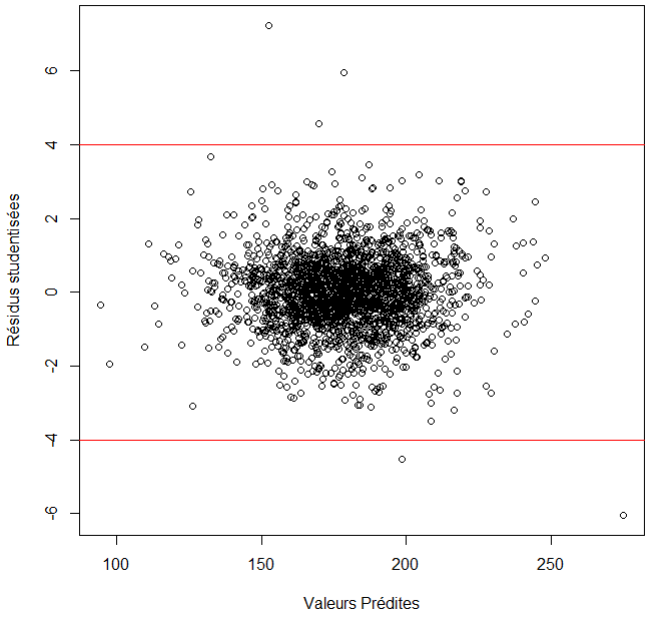
```
reg9 = lm(TARGET_deathRate ~ PctBachDeg25_Over + incidenceRate +
            povertyPercent + PctHS18_24 + PctOtherRace + PctMarriedHouseholds +
            MedianAgeFemale + BirthRate + PctPublicCoverageAlone + PctHS25_Over +
            PctEmployed16_Over + PctEmpPrivCoverage + avgAnnCount +
            popEst2015, data = data_train6)
summary(reg9)
```

```
Call:
lm(formula = TARGET_deathRate ~ PctBachDeg25_Over + incidenceRate +
    povertyPercent + PctHS18_24 + PctOtherRace + PctMarriedHouseholds +
    MedianAgeFemale + BirthRate + PctPublicCoverageAlone + PctHS25_Over +
    PctEmployed16_Over + PctEmpPrivCoverage + avgAnnCount + popEst2015,
    data = data_train6)

Residuals:
     Min       1Q   Median       3Q      Max
-112.931  -11.044   -0.106   10.974  140.079

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.161e+02  1.424e+01   8.150 5.89e-16 ***
PctBachDeg25_Over      -1.309e+00  1.621e-01  -8.075 1.08e-15 ***
incidenceRate           1.951e-01  8.484e-03  22.992  < 2e-16 ***
povertyPercent          4.500e-01  1.578e-01   2.851  0.00440 **
PctHS18_24              3.194e-01  5.243e-02   6.091 1.31e-09 ***
PctOtherRace           -8.163e-01  1.301e-01  -6.273 4.22e-10 ***
PctMarriedHouseholds   -3.506e-01  8.897e-02  -3.940 8.38e-05 ***
MedianAgeFemale        -4.454e-01  1.109e-01  -4.017 6.08e-05 ***
BirthRate              -6.043e-01  2.178e-01  -2.775  0.00557 **
PctPublicCoverageAlone  5.793e-01  1.321e-01   4.384 1.22e-05 ***
PctHS25_Over            2.514e-01  1.072e-01   2.345  0.01914 *
PctEmployed16_Over     -2.356e-01  9.325e-02  -2.527  0.01159 *
PctEmpPrivCoverage      2.095e-01  8.048e-02   2.603  0.00931 **
avgAnnCount            -2.055e-03  8.347e-04  -2.462  0.01389 *
popEst2015              7.489e-06  3.582e-06   2.091  0.03666 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.61 on 2301 degrees of freedom
Multiple R-squared:  0.5049,    Adjusted R-squared:  0.5019
F-statistic: 167.6 on 14 and 2301 DF,  p-value: < 2.2e-16
```

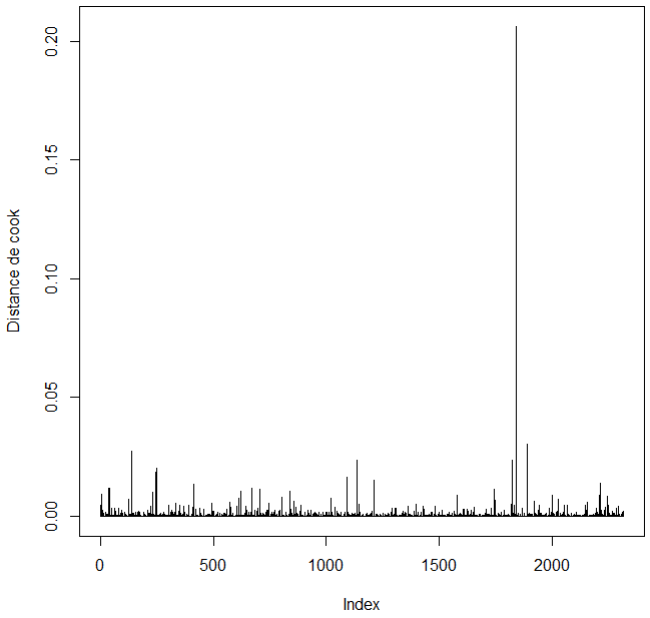
Le premier modèle avec des variables complètes a un R-carré ajusté de 0,5024, ce qui signifie que le modèle peut expliquer 50,24% de la variance de la variable cible (taux de mortalité). En revanche, notre modèle plus simple a un R-carré ajusté de 50,19 %, ce qui ne présente pas de grande différence avec notre premier modèle. Cela montre qu'il est possible de supprimer les variables dont les coefficients ne sont pas significatifs. En outre, la suppression de ces deux variables a augmenté l'importance des variables empattement et taux de compression dans le modèle. Ce qui prouve que c'est le bon modèle.

3.2.6 Vérification des valeurs aberrantes

Une valeur aberrante est une observation qui se situe à une distance anormale des autres valeurs dans un échantillon aléatoire d'une population. L'examen des données permet de repérer des observations inhabituelles, éloignées de la masse des données. Ces points sont souvent qualifiés de valeurs aberrantes.



D'après le graphique ci-dessus, nous pouvons remarquer qu'il y a quelques valeurs aberrantes dans nos données, et nous devrions travailler à les éliminer pour améliorer la performance de notre modèle.
Pour trouver ces valeurs aberrantes, nous utilisons la distance de Cook, afin de détecter les observations qui influencent fortement les valeurs ajustées du modèle. La distance de Cook a été introduite par le statisticien américain R Dennis Cook en 1977. Elle est utilisée pour identifier les points de données influents. Elle dépend à la fois du résidu et de l'effet de levier, elle prend en compte à la fois la valeur x et la valeur y de l'observation.



Nous pouvons maintenant remarquer qu'il y a une valeur aberrante qui affecte beaucoup notre modèle et qui doit être supprimée, à savoir la ligne '1875' dans l'ensemble de données d'entraînement.

```
#removing outliers
data_train6 = data_train6[-c(1875),]

reg10 = lm(TARGET_deathRate ~ PctBachDeg25_Over + incidenceRate +
             povertyPercent + PctHS18_24 + PctOtherRace + PctMarriedHouseholds +
             MedianAgeFemale + BirthRate + PctPublicCoverageAlone + PctHS25_Over +
             PctEmployed16_Over + PctEmpPrivCoverage + avgAnnCount +
             popEst2015, data = data_train6)
summary(reg10)
```

```
Call:
lm(formula = TARGET_deathRate ~ PctBachDeg25_Over + incidenceRate +
    povertyPercent + PctHS18_24 + PctOtherRace + PctMarriedHouseholds +
    MedianAgeFemale + BirthRate + PctPublicCoverageAlone + PctHS25_Over +
    PctEmployed16_Over + PctEmpPrivCoverage + avgAnnCount + popEst2015,
    data = data_train6)

Residuals:
     Min       1Q   Median       3Q      Max
-113.028  -11.020   -0.109   10.979  140.089

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.160e+02  1.425e+01   8.145 6.14e-16 ***
PctBachDeg25_Over      -1.312e+00  1.622e-01  -8.091 9.47e-16 ***
incidenceRate           1.952e-01  8.487e-03  22.999  < 2e-16 ***
povertyPercent          4.515e-01  1.579e-01   2.860  0.00428 **
PctHS18_24              3.218e-01  5.250e-02   6.129 1.04e-09 ***
PctOtherRace           -8.154e-01  1.302e-01  -6.265 4.45e-10 ***
PctMarriedHouseholds   -3.495e-01  8.900e-02  -3.927 8.85e-05 ***
MedianAgeFemale        -4.462e-01  1.109e-01  -4.023 5.92e-05 ***
BirthRate              -6.134e-01  2.181e-01  -2.812  0.00497 **
PctPublicCoverageAlone  5.785e-01  1.322e-01   4.377 1.26e-05 ***
PctHS25_Over            2.486e-01  1.073e-01   2.317  0.02062 *
PctEmployed16_Over     -2.368e-01  9.332e-02  -2.538  0.01123 *
PctEmpPrivCoverage      2.114e-01  8.052e-02   2.626  0.00870 **
avgAnnCount            -2.054e-03  8.356e-04  -2.458  0.01403 *
popEst2015              7.473e-06  3.584e-06   2.085  0.03716 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.61 on 2299 degrees of freedom
Multiple R-squared:  0.5051,    Adjusted R-squared:  0.502
F-statistic: 167.6 on 14 and 2299 DF,  p-value: < 2.2e-16
```

Après avoir supprimé la valeur aberrante, nous pouvons remarquer qu'il y a un changement notable par rapport au modèle précédent, par exemple le R-carré ajusté a augmenté.
Bien sûr, nous pouvons continuer à éliminer toutes les valeurs aberrantes une par une de notre ensemble de données jusqu'à ce que nous atteignions une solution satisfaisante, mais cela prendra trop de temps d'éliminer point par point et nous nous contenterons donc de ce modèle pour travailler.

4 Évaluation du modèle

4.1 Performance du modele

RMSE est un moyen utile de voir dans quelle mesure un modèle de régression est capable de s'adapter à un jeu de données.Il indique à quelle distance nos valeurs prédites sont éloignées de nos valeurs observées dans une analyse de régression, en moyenne. Il est calculé comme suit :

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(P_i - O_i)^2}$$

- Pi : est la valeur prédite de la ième observation dans l'ensemble de données
- Oi : est la valeur observée pour la ième observation dans l'ensemble de données
- n : est la taille de l'échantillon

Nous pouvons utilisé la fonction RMSE du paquet caret.

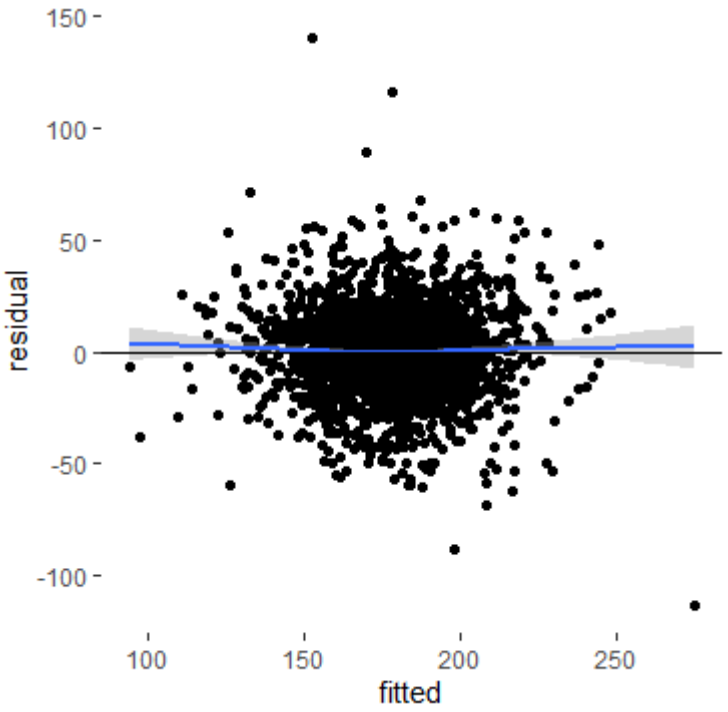
```
> # RMSE of train dataset
> RMSE(pred = reg10$fitted.values, obs = data_train6$TARGET_deathRate)
[1] 19.54371
> # RMSE of test dataset
> RMSE(pred = lm_pred, obs = data_test6$TARGET_deathRate)
[1] 19.68974
>
```

D'après nos observations, nous avons remarqué que la valeur de RMSE augmente lorsque nous testons le modèle sur l'ensemble de données de test, par rapport à l'ensemble de données d'entraînement, mais ce n'est pas une augmentation importante compte tenu de la taille de l'ensemble de données que nous utilisons, et de la distribution des données entre les données d'entraînement et de test.

4.2 Test de linéarité

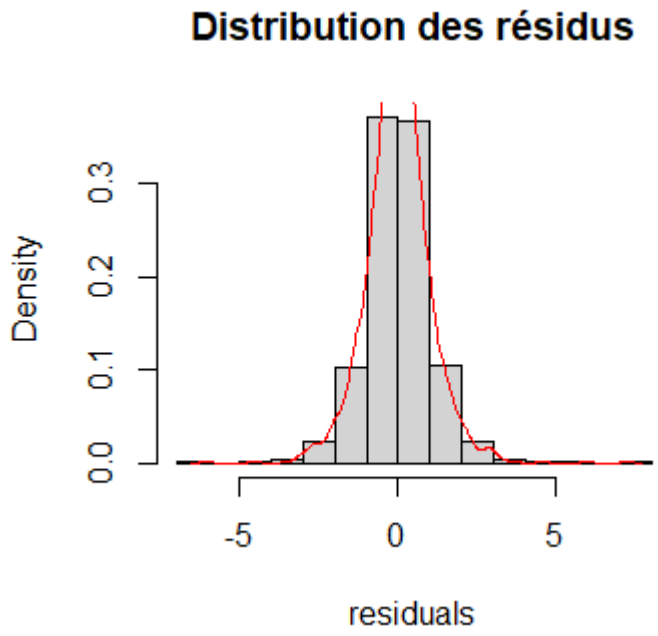
Le modèle de régression linéaire suppose qu'il existe une relation linéaire entre les prédicteurs et la réponse. Si la relation réelle est loin d'être linéaire, pratiquement toutes les conclusions que nous tirons de l'ajustement sont suspectes. En outre, la précision de la prédiction du modèle peut être considérablement réduite. Les tracés résiduels sont un outil graphique utile pour identifier la non-linéarité. Si le tracé des résidus

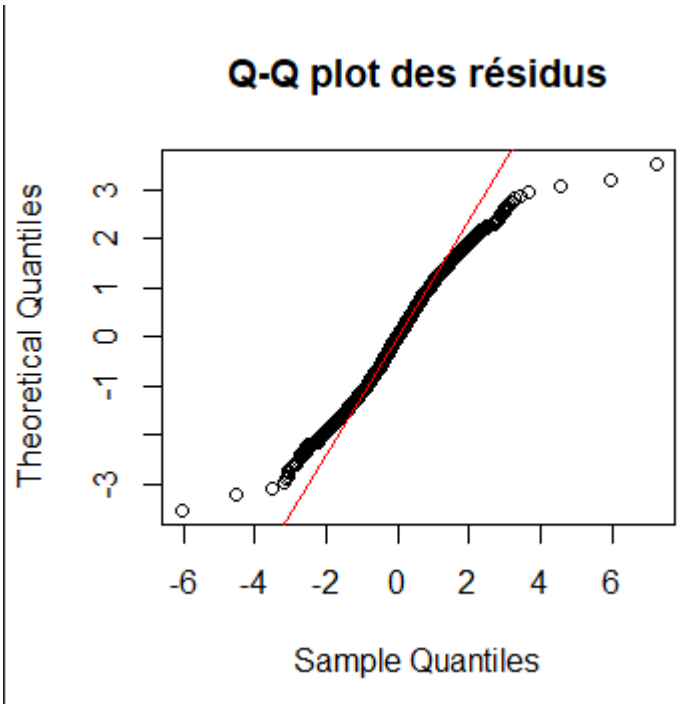
présente une tendance, cela signifie que le modèle peut être encore amélioré ou qu'il ne répond pas à l'hypothèse de linéarité. Le graphique montre la relation entre les résidus/erreurs et les valeurs prédites/adaptées.



4.3 Test de normalité

L'hypothèse de la régression linéaire est que les résidus suivent une distribution normale. Nous pouvons facilement vérifier cela en plaçant la distribution des résidus.





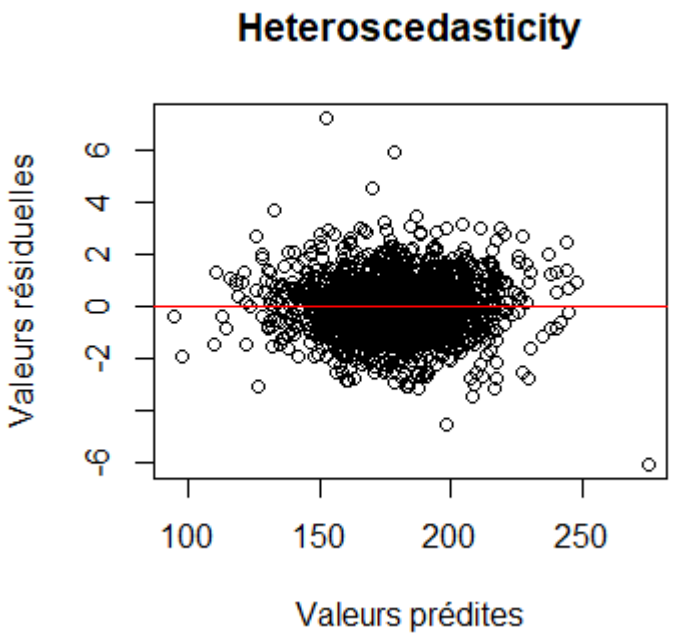
A partir de ces graphiques, nous pouvons conclure que notre modèle ne suit pas la distribution normale. Pour prouver notre conclusion, nous pouvons également utiliser le test de normalité de Saphiro-Wilk.

```
> shapiro.test(rstudent(reg10))

        Shapiro-Wilk normality test

data:  rstudent(reg10)
W = 0.97667, p-value < 2.2e-16
```

L'hypothèse nulle est que les résidus suivent une distribution normale. Avec une valeur p inférieur a 0,05, nous pouvons conclure que notre hypothèse est rejetée, et que nos résidus ne suivent pas la distribution normale.

4.4 Test d'hétéroscédasticité

L'hétérodasticité signifie que les variances des termes d'erreur ne sont pas constantes. On peut identifier des variances non constantes dans les erreurs par la présence d'une forme d'entonnoir dans le graphique des résidus, comme dans le cas de la linéarité. En plaçant les valeurs prédites sur la base des valeurs résidentielles :



pour dire qu'il ya une hétérodasticité on va effectuer le test de Breusch-Pagan avec la fonction bptest().

```
> bptest(reg10)

        studentized Breusch-Pagan test

data:  reg10
BP = 89.146, df = 14, p-value = 5.501e-13
```

La statistique de test est 89,146 et la valeur de p correspondante est 0,0000000005501. Puisque la valeur de p est inférieure à 0,05, nous rejetons l'hypothèse(les résidus sont distribués avec une variance égale) et on conclut à la présence d'hétéroscédasticité(les résidus ne sont pas distribués avec une variance égale).

4.5 Test d'autocorrélation

Les erreurs types qui sont calculées pour les coefficients de régression estimés ou les valeurs ajustées sont basées sur l'hypothèse de termes d'erreur non corrélés (pas d'auto-corrélation). Si, en fait, il existe une corrélation entre les termes d'erreur, les erreurs standard estimées auront tendance à sous-estimer les véritables erreurs standard. Par conséquent, les intervalles de confiance et de prédiction seront plus étroits qu'ils ne devraient l'être. Par exemple, un intervalle de confiance de 95% peut en réalité avoir une probabilité bien plus faible que 0,95 de contenir la vraie valeur du paramètre. En outre, les valeurs p associées au modèle seront inférieures à ce qu'elles devraient être ; cela pourrait nous amener à conclure à tort qu'un paramètre est statistiquement significatif. En bref, si les termes d'erreur sont corrélés, nous pouvons avoir un sentiment de confiance injustifié dans notre modèle. L'auto-corrélation peut être détectée à l'aide du test de durbin watson, avec l'hypothèse nulle qu'il n'y a pas d'auto-corrélation.

```
> ############
> # autocorrelation test
> dwtest(reg10)

        Durbin-Watson test

data:  reg10
DW = 2.0399, p-value = 0.8316
alternative hypothesis: true autocorrelation is greater than 0

>
```

Le résultat montre que l'hypothèse nulle est vraie, ce qui signifie qu'il n'y a pas d'auto-corrélation dans notre modèle final.c-a-d qu'il n ya pas d'indépendance entre les résiduelle de notre modèle.

5 Conclusion

Notre modèle final a satisfait aux certains hypothèses classiques. Le R-carré du modèle est moyen. La précision du modèle dans la prédiction du Taux de mortalité cible est mesurée avec RMSE, avec les données de formation a RMSE de 19.54371 et les données de test a RMSE de 19.68974, ce qui signifie que notre modèle peut adapter l'ensemble de données de formation. Après nos tests sur le modèle, nous avons conclu ce qui suit :

— Le modèle est linéaire.

— Les résidus du modèle ne suivent pas la distribution normale.

— Les variances des termes d'erreur ne sont pas constantes dans notre modèle.

— Il n'y a pas d'autocorrélation dans notre modèle.

Bien sûr, notre modèle n'est pas le modèle parfait, pour de nombreuses raisons, l'une d'entre elles étant la taille de notre ensemble de données, puisqu'il n'y a pas beaucoup d'observations à partir desquelles notre modèle peut apprendre, ce qui entraîne une faible précision. Il est toujours possible d'améliorer les modèles de régression, nous essayons également d' améliorer notre modèle, par exemple nous avons effectuer les transformations suivantes:

-avec la fonction *lm_robust*() :

```
reg11<-lm_robust(TARGET_deathRate ~ PctBachDeg25_Over + incidenceRate +
                 povertyPercent + PctHS18_24 + PctOtherRace + PctMarriedHouseholds +
                 MedianAgeFemale + BirthRate + PctPublicCoverageAlone + PctHS25_Over +
                 PctEmployed16_Over + PctEmpPrivCoverage + avgAnnCount +
                 popEst2015, data = data_train6)

summary(reg11)
```

-avec la transformation log:

```
reg12<-lm(log(TARGET_deathRate) ~ PctBachDeg25_Over + incidenceRate +
                  povertyPercent + PctHS18_24 + PctOtherRace + PctMarriedHouseholds +
                  MedianAgeFemale + BirthRate + PctPublicCoverageAlone + PctHS25_Over +
                  PctEmployed16_Over + PctEmpPrivCoverage + avgAnnCount +
                  popEst2015, data = data_train6)
############
```

mais il n'y a pas de changement surtout au niveau de R-carré du modèle et la valeur du p ,dont le R-carré du modèle = 0.505

```
Multiple R-squared:  0.505 ,    Adjusted R-squared:  0.502
F-statistic: 139.2 on 14 and 2300 DF,  p-value: < 2.2e-16
```

et pour la valeur du p reste = 0,0000000005501