SAE 401

Expliquer ou prédire l'âge de décès à partir de plusieurs facteurs

27/03/2025

Sasha Le Potier – Romane Lequeux – Soline Thomas

SOMMAIRE

- Présentation des données
- Méthodes utilisées
- Résultats
- Cas concret

PROBLÉMATIQUE

Dans quelle mesure les facteurs socio-démographiques, les habitudes de vie et les antécédents médicaux influencent-ils l'âge de décès des individus ?

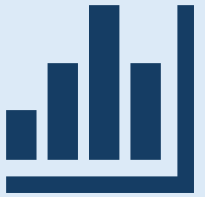
PRÉSENTATION DES DONNÉES

Description des sources de données

Données de Santé

Expliquer l'âge de décès

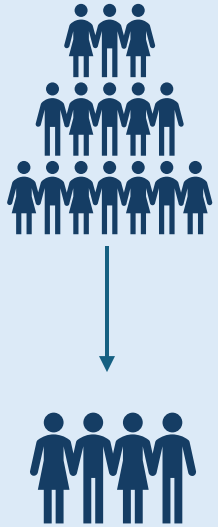
9221 individus



24 variables explicatives

PRÉSENTATION DES DONNÉES

Prétraitements et nettoyage des données



Échantillon
1 000 individus

Changements de labels
Conversion du poids et l'âge

IMC

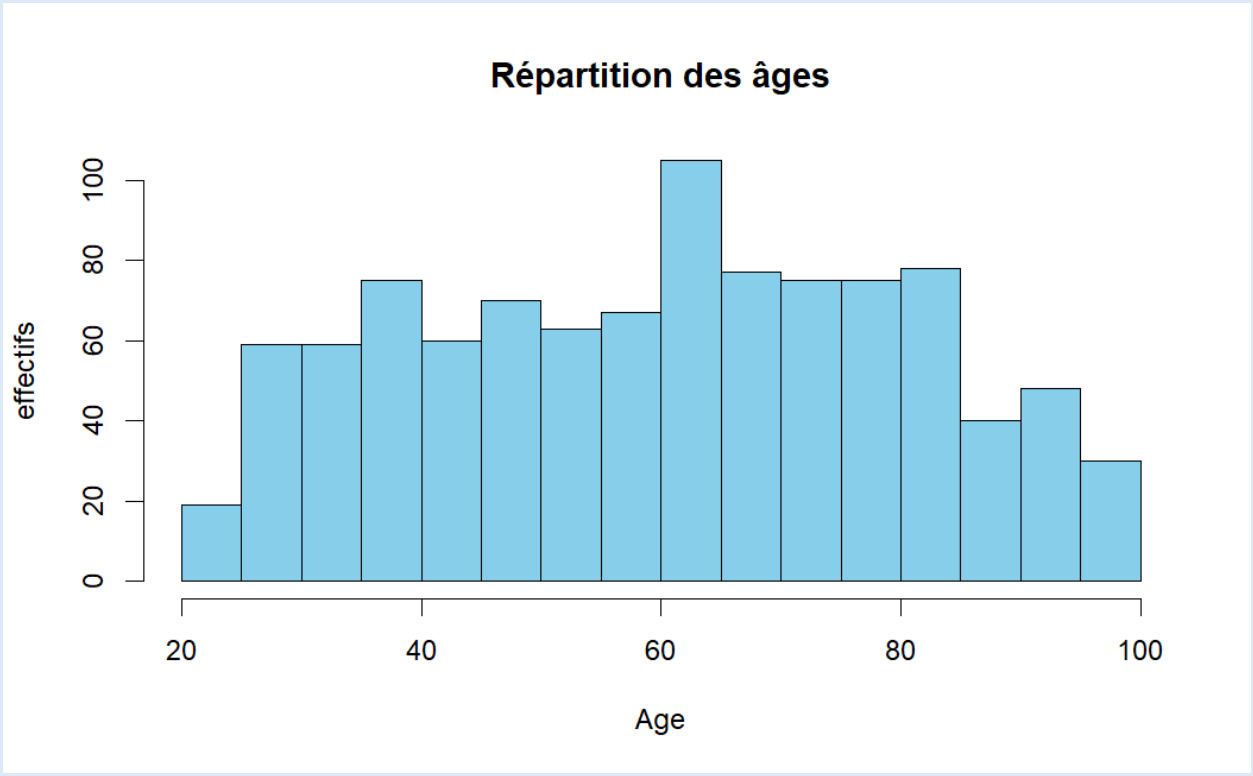
nouvelle variable

ANALYSE EXPLORATOIRE

variable d'intérêt

Statistiques	Valeurs
Moyenne	60.63500
Variance	401.46724
Écart-type	20.03665
Minimum	25.00000
Maximum	99.00000
premier Quartiles	44.00000
Médiane	62.00000
Troisième quartile	77.00000

Analyse de la variable de l'âge



Histogramme de la répartition de l'âge

MÉTHODOLOGIE

Choix de la méthode

Méthodes différentes :

Méthode ascendante descendante

- Combine 2 méthodes et ajoute ou enlève des variables à chaque itération
- 16 variables explicatives

Méthode exhaustive

- Test toutes les combinaisons possibles de variables explicatives
- 14 variables explicatives

MÉTHODOLOGIE

Validité du modèle **ascendant descendant**

Absence de multicolinéarité (VIF)

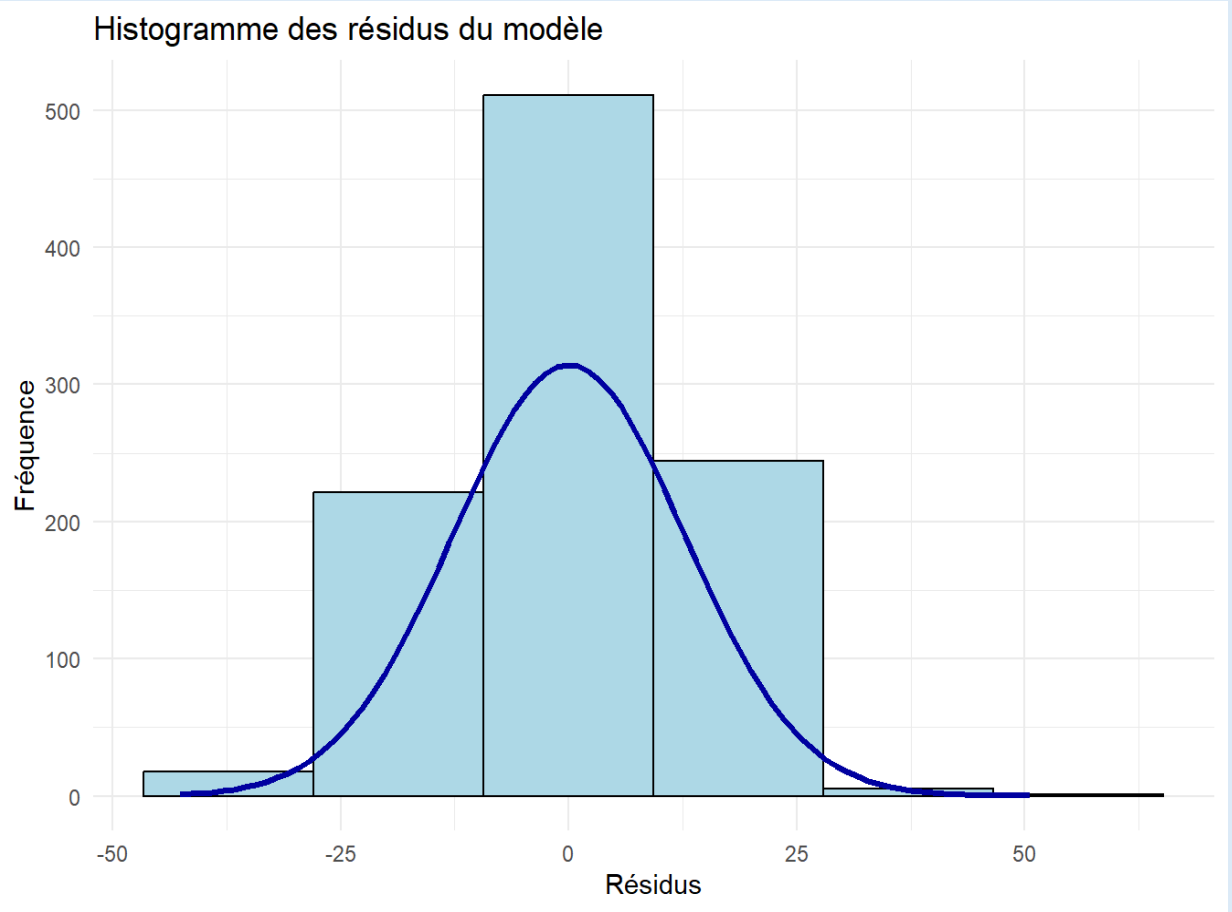
=> Toutes les variables ont un VIF < 2

	GVIF	Df	GVIF^(1/(2*Df))
addiction	1.213637	1	1.101652
cholesterol	1.576593	1	1.255625
num_meds	1.580694	1	1.257257
hds	1.345669	1	1.160030
drinks_aweek	1.352783	1	1.163092
immune_defic	1.027045	1	1.013432
opioids	1.093431	1	1.045672
sex	1.428611	1	1.195245
major_surgery_num	1.699354	1	1.303593
family_cancer	1.014802	1	1.007374
diabetes	1.027652	1	1.013732
other_drugs	1.028665	1	1.014231
smoker	1.057398	1	1.028299
ls_danger_label	1.179559	2	1.042149
family_heart_disease	1.032985	1	1.016359
family_cholesterol	1.009117	1	1.004548
occup_danger_label	1.176245	2	1.041417
sys_bp	1.351846	1	1.162689

MÉTHODOLOGIE

Validité du modèle **ascendant descendant**

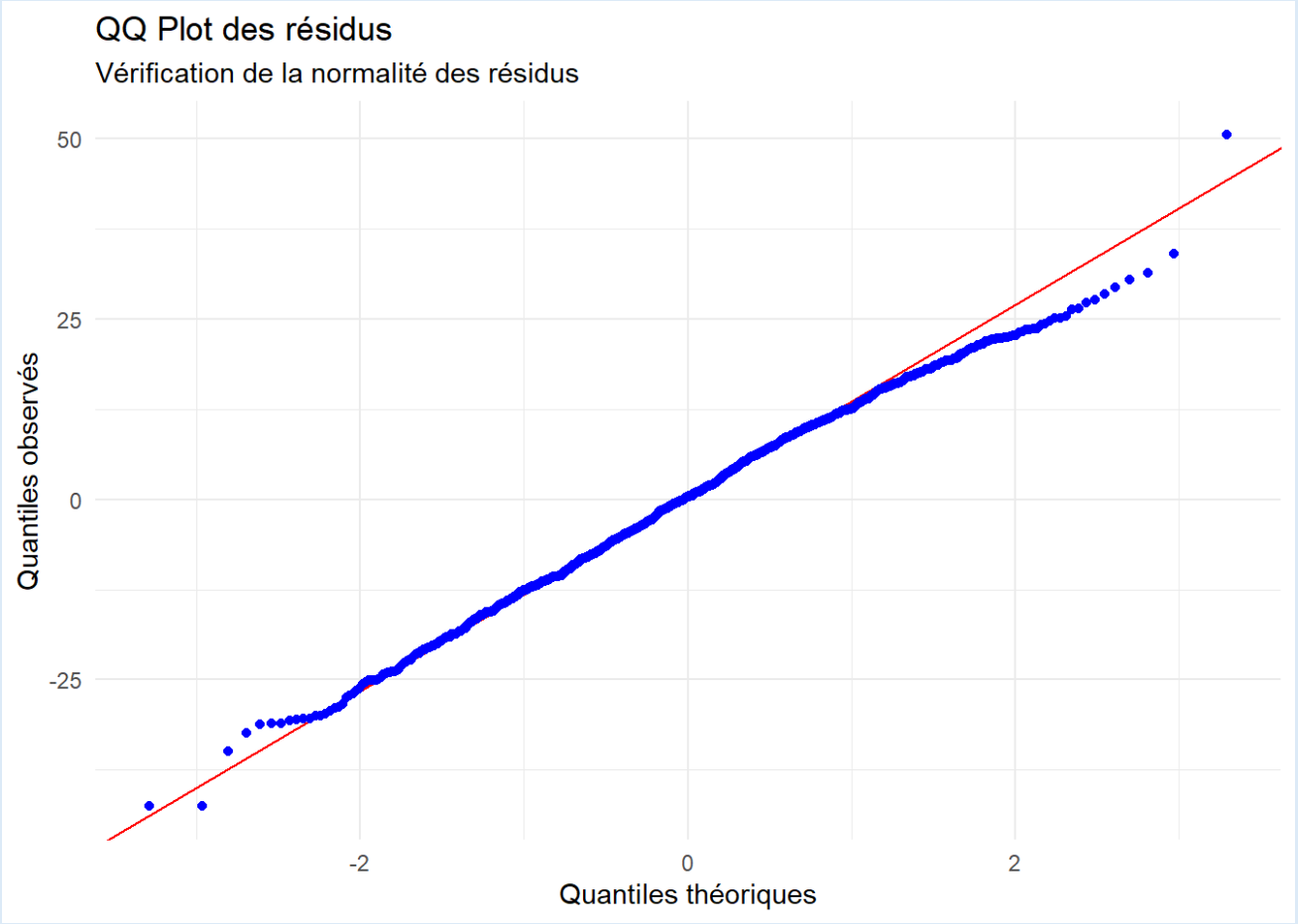
Erreurs centrées autour de zéro



MÉTHODOLOGIE

Validité du modèle **ascendant descendant**

Erreurs normalement distribuées

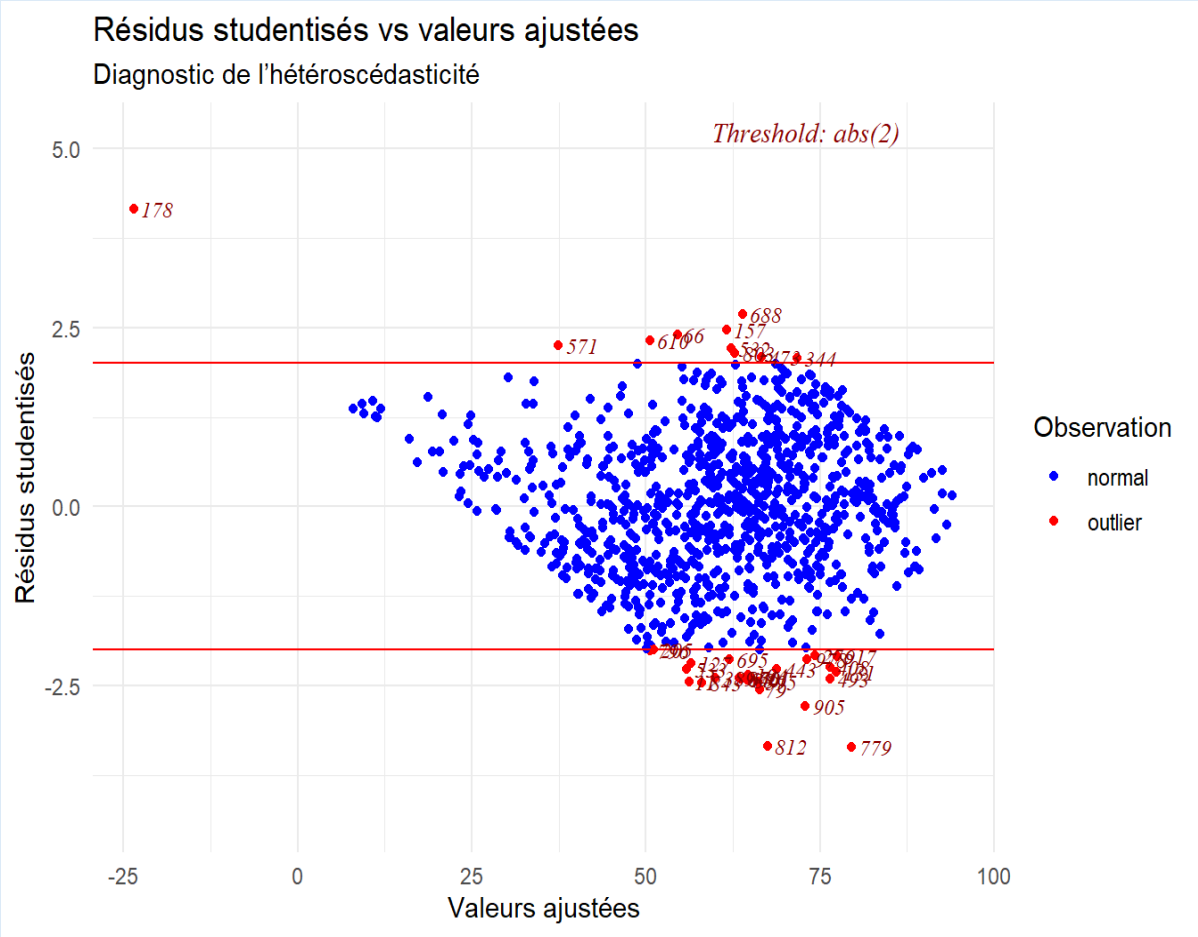


MÉTHODOLOGIE

Validité du modèle **ascendant descendant**

Erreurs de même variance

```
 studentized Breusch-Pagan test

data:  modele_asc_desc
BP = 16.593, df = 20, p-value = 0.6793
```



MÉTHODOLOGIE

Validité du modèle **ascendant descendant**

Erreurs indépendantes

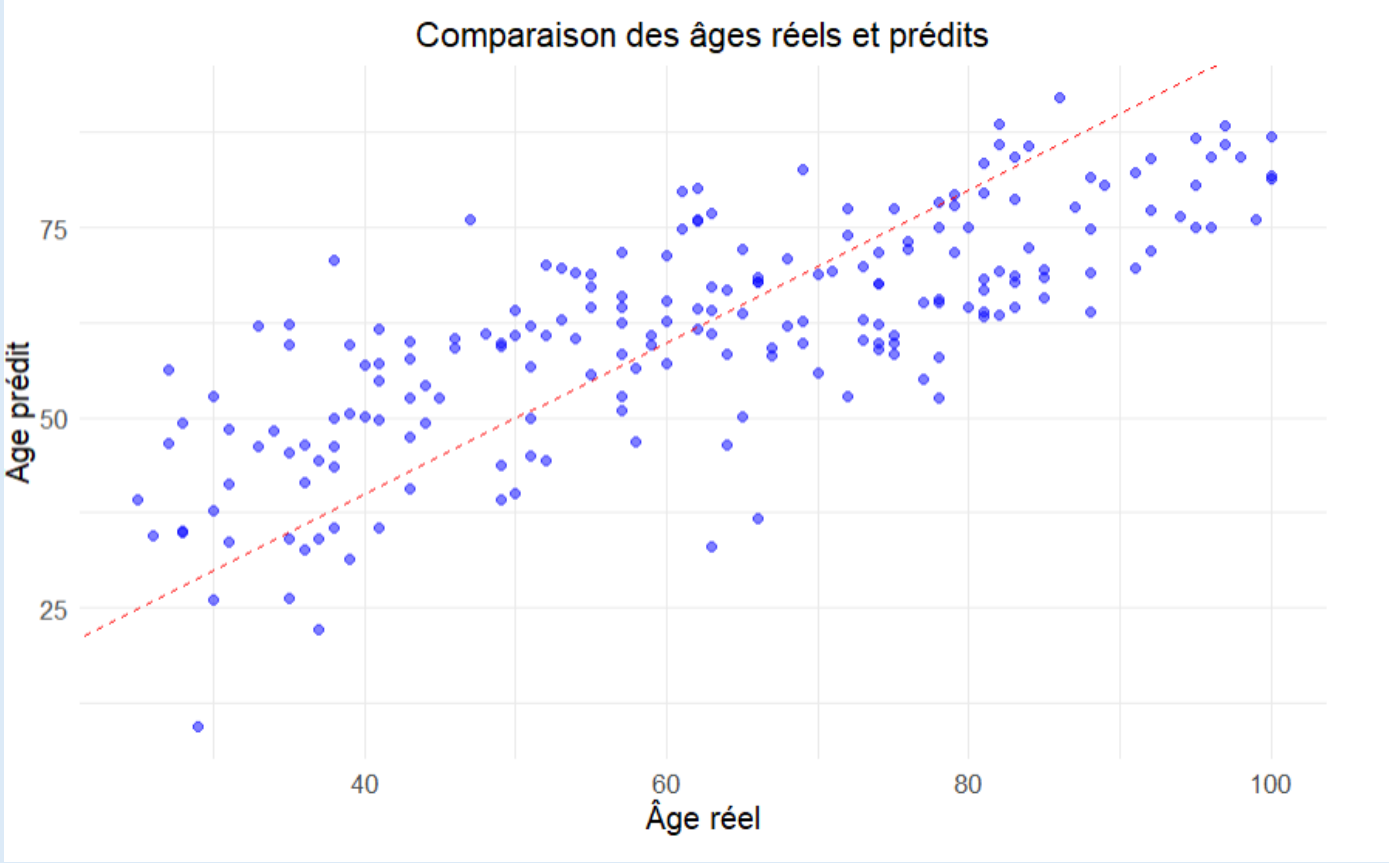
```
Durbin-Watson test

data:  modele_asc_desc
DW = 2.0749, p-value = 0.8819
alternative hypothesis: true autocorrelation is greater than 0
```

DW proche de 2 : Pas d'autocorrélation des erreurs
DW < 2 : Autocorrélation positive des erreurs
DW > 2 : Autocorrélation négative des erreurs

MÉTHODOLOGIE

Qualité du modèle

- Pour vérifier la qualité des prédictions du modèle on fait une validation croisée
- On a un RMSE inferieur à l'écart type de l'âge



RÉSULTATS

Modèle final de prédiction de l'âge de décès

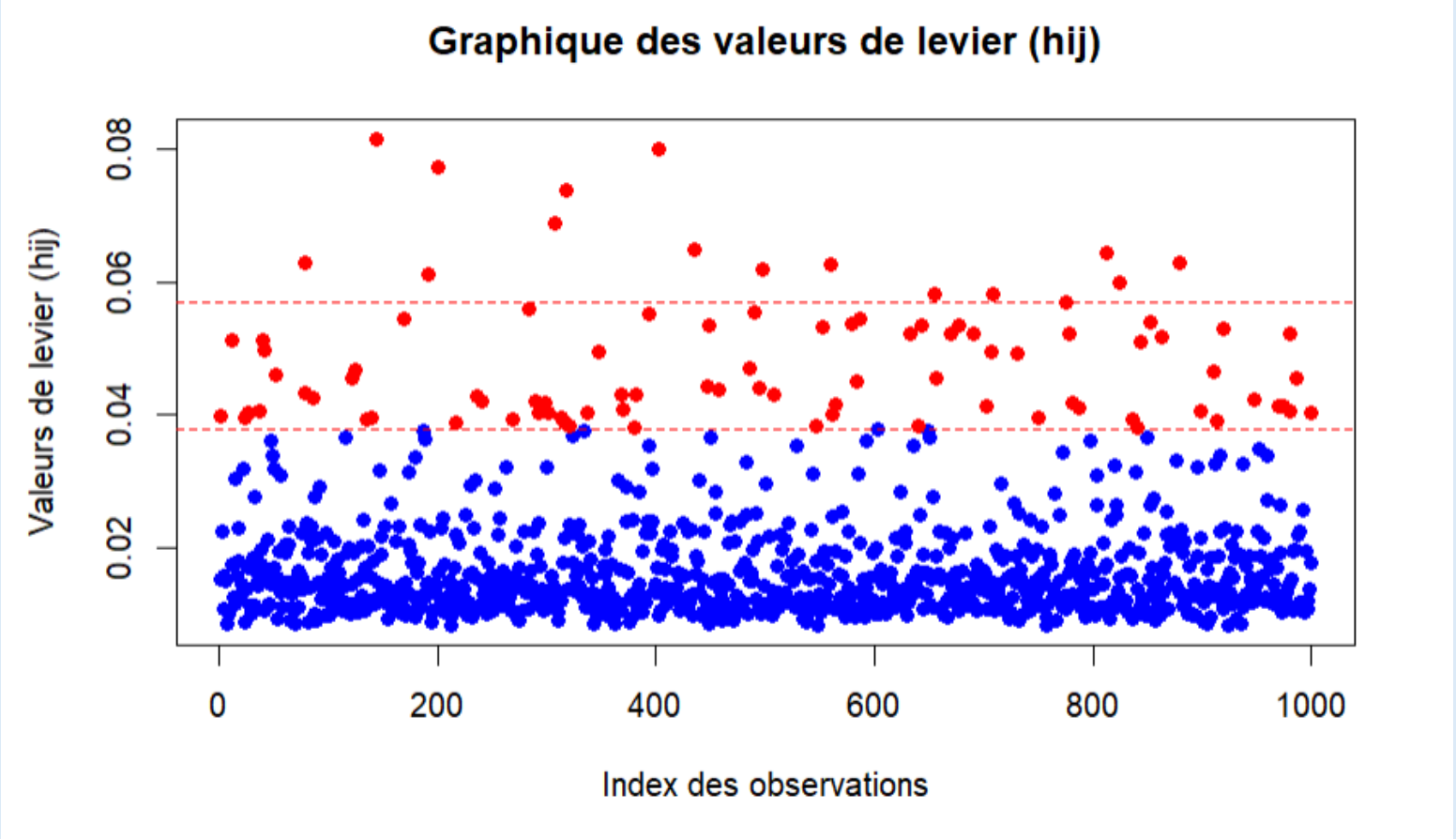
- Ce modèle explique environ 60% de la variance de l'Âge de décès

$$age_i = \beta^0 + \beta^1 \cdot num_{meds_i} + \beta^2 \cdot cholesterol_i + \beta^3 \cdot drinks_{aweek_i} + \beta^4 \cdot immune_{deficy_i} + \beta^5 \cdot opioids_i + \beta^6 \cdot addiction_i + \beta^7 \cdot hds_i + \beta^8 \cdot sex_i + \beta^9 \cdot major_{surgery_{num_i}} + \beta^{10} \cdot family_{cancer_i} + \beta^{11} \cdot diabetes_i + \beta^{12} \cdot other_{drugs_i} + \beta^{13} \cdot smoking_i + \beta^{14} \cdot ls_{danger_{label_{faible_i}}} + \beta^{15} \cdot ls_{danger_{label_{moyen_i}}} + \beta^{16} \cdot family_{heart_{disease_i}} + \beta^{17} \cdot family_{cholesterol_i} + \beta_{18} \cdot occup\_danger\_label\_faible\_i + \beta_{19} \cdot occup\_danger\_label\_moyen\_i + \beta_{20} \cdot sys\_bp\_i + \varepsilon\_i$$

RÉSULTATS

Individus leviers

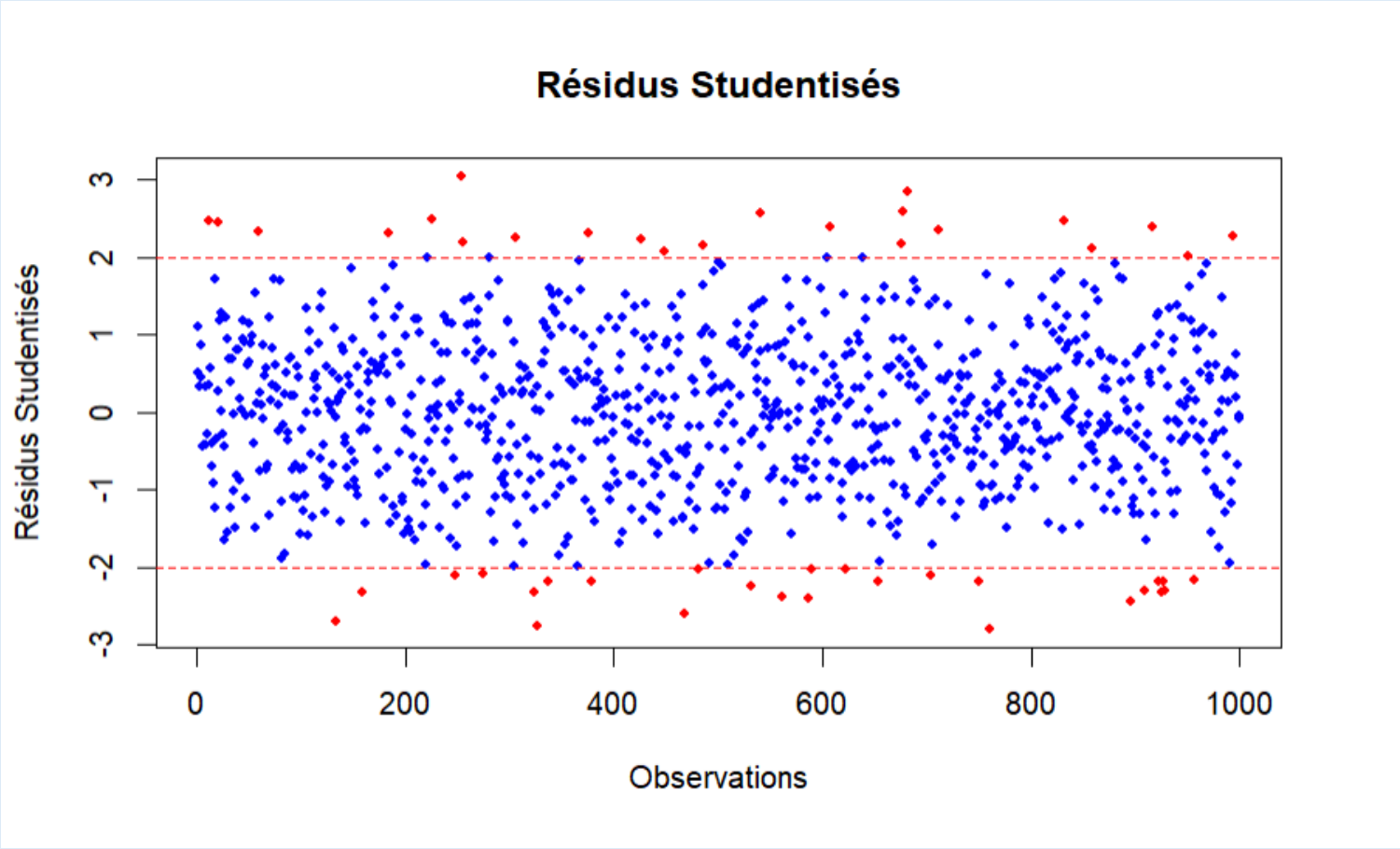
- Les individus au dessus du deuxième seuil doivent faire l'objet de surveillance (biais)



RÉSULTATS

Individus atypiques

- Identification des individus mal expliqués par notre modèle
- Moins de 5% d'individus atypiques au total



Estimation de l'âge de décès de

Marilyn Monroe

- Âge de décès : 36 ans (overdose de médicaments)
- Âge de décès prédit : avant 39 ans
- Intervalle de confiance : [13 .48 ; 65.710]

MERCI DE VOTRE ÉCOUTE !