

# Predictive Modeling of Health Insurance Costs: From Linear Regression to Gamma GLM

---



Github:

<https://github.com/chaimae-elyaouti>



LinkedIn:

<https://www.linkedin.com/in/chaimae-el-yaouti-95897b320/>

# Key Findings (Aperçu des Résultats)

- Primary Factor: Smoking multiplies medical costs by 4.58x
- Best Model: Gamma GLM outperformed OLS Regression (AIC = 20,947 vs. 21,267)
- Data Integrity: 87 influential points were removed to ensure unbiased predictions.)

# Plan

- (01) Introduction
- (02) Préparation des données
- (03) Régression linéaire multiple
- (04) Méthode progressive
- (05) Méthode pas à pas

- (06) ANOVA : partiethéorique et pratiques sur les variables dum m
- (07) Limites de la regression linéair
- (08) Méthode alterntives
- (09) Comparaison des modèles
- (10) Conclusion

# Introduction



Dans un contexte de hausse constante des dépenses de santé, la modélisation des coûts médicaux est un enjeu crucial pour anticiper et optimiser la gestion des ressources. Cette étude utilise la régression linéaire pour prédire les dépenses médicales annuelles à partir de données patients (âge, sexe, tabagisme, etc.)

02

# Préparation des données

---

## Objectif:

Télécharger l'ensemble de données "MedicalCost PersonalDataset" (`insurance.csv`) depuis Kaggle afin de préparer le dataset pour une régression linéaire robuste.

## Nettoyage:

détection et gestion des outliers sur charges  
a - (boîte à moustaches) et des points influents  
b - (distance de Cook).

## Transformation:

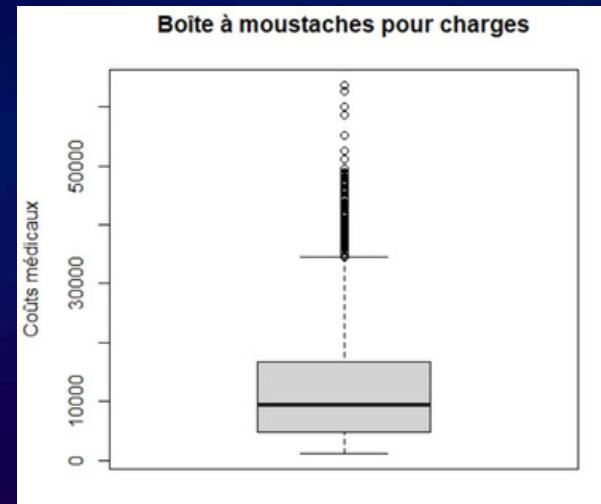
conversion des variables catégoriques en facteurs (`dummyvariables`)

# Nettoyage -Détection des outliers avec boîte à moustaches

```
> data <- read.csv("C:\\\\Users\\\\hp EliteBook\\\\Downloads\\\\PROJET_RL\\\\insurance.csv")
> # Résumé des données
> summary(data)
   age          sex      bmi    children
Min. :18.00  Length:1338  Min. :15.96  Min. :0.000
1st Qu.:27.00 Class :character  1st Qu.:26.30  1st Qu.:0.000
Median :39.00 Mode  :character  Median :30.40  Median :1.000
Mean   :39.21                   Mean   :30.66  Mean   :1.095
3rd Qu.:51.00                   3rd Qu.:34.69  3rd Qu.:2.000
Max.   :64.00                   Max.   :53.13  Max.   :5.000
   smoker        region      charges
Length:1338  Length:1338  Min.   : 1122
Class :character  Class :character  1st Qu.: 4740
Mode  :character  Mode  :character  Median : 9382
                  Mode  :character  Mean   :13270
                  Mode  :character  3rd Qu.:16640
                  Mode  :character  Max.   :63770
```

```
> # Calculer Q1, Q3 et IQR pour 'charges'
> Q1 <- quantile(data$charges, 0.25, na.rm = TRUE)
> Q3 <- quantile(data$charges, 0.75, na.rm = TRUE)
> IQR <- Q3 - Q1
> lower_bound <- Q1 - 1.5 * IQR
> upper_bound <- Q3 + 1.5 * IQR
>
> # Identifier les outliers
> outliers_boxplot <- data$charges < lower_bound | data$charges > upper_bound
> cat("Nombre d'outliers détectés :", sum(outliers_boxplot, na.rm = TRUE), "\n")
Nombre d'outliers détectés : 139
```

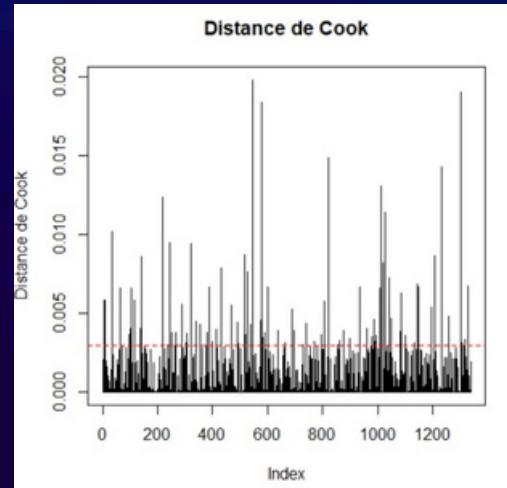
```
> boxplot(data$charges, main = "Boite à moustaches pour charges", ylab = "Coûts médicaux")
```



# Nettoyage -Détection des points influents avec distance de Cook

```
> # Ajuster le modèle
> model <- lm(charges ~ age + sex + bmi + children + smoker + region, data = data)
>
> # Calculer la distance de Cook
> cooks_d <- cooks.distance(model)
> n <- nrow(data)
> threshold <- 4 / n
> influential_points <- cooks_d > threshold
> cat("Nombre de points influents détectés :", sum(influential_points, na.rm = TRUE), "\n")
Nombre de points influents détectés : 87
```

- Commentaire : Seuil de  $4/n$  pour identifier les points influents (observations qui impactent fortement le modèle).



# Transformation des variables

```
> str(data)
'data.frame': 1338 obs. of 7 variables:
 $ age      : int 19 18 28 33 32 31 46 37 37 60 ...
 $ sex       : chr "female" "male" "male" "male" ...
 $ bmi       : num 27.9 33.8 33 22.7 28.9 ...
 $ children: int 0 1 3 0 0 0 1 3 2 0 ...
 $ smoker    : chr "yes" "no" "no" "no" ...
 $ region   : chr "southwest" "southeast" "southeast" "northwest" ...
 $ charges  : num 16885 1726 4449 21984 3867 ...
```

- Objectif : Convertir les variables catégoriques (sex, smoker, region) en facteurs pour la régression.

```
> # Convertir les variables catégoriques en facteurs
> data$sex <- as.factor(data$sex)
> data$smoker <- as.factor(data$smoker)
> data$region <- as.factor(data$region)
> str(data)
'data.frame': 1338 obs. of 7 variables:
 $ age      : int 19 18 28 33 32 31 46 37 37 60 ...
 $ sex       : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
 $ bmi       : num 27.9 33.8 33 22.7 28.9 ...
 $ children: int 0 1 3 0 0 0 1 3 2 0 ...
 $ smoker    : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 ...
 $ region   : Factor w/ 4 levels "northeast","northwest",...: 4 3 3 2 2 3 3 2 1 2 ...
 $ charges  : num 16885 1726 4449 21984 3867 ...
```

# Résumé et datasetfinal

```
> # Supprimer les outliers et points influents
> data_cleaned <- data[!outliers_boxplot & !influential_points, ]
>
> # Vérification finale
> cat("Nombre de lignes avant nettoyage :", nrow(data), "\n")
Nombre de lignes avant nettoyage : 1338
> cat("Nombre de lignes après nettoyage :", nrow(data_cleaned), "\n")
Nombre de lignes après nettoyage : 1127
```

Résultat : Dataset prêt pour la régression linéaire!!

03

# Régression linéaire multiple

---

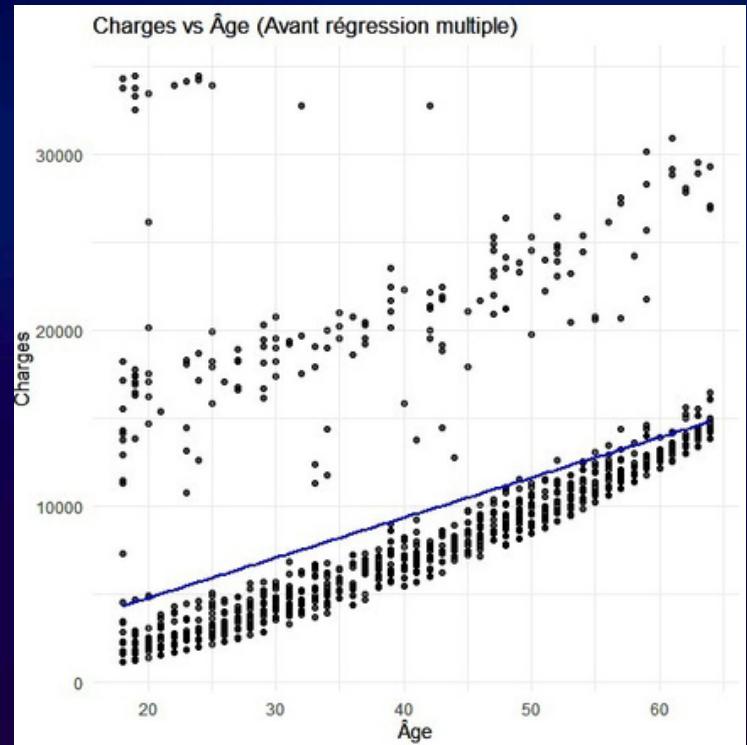
# Lien Statistique entre les Variables Clés et les Dépenses de Santé:

## Nuage de points

```
> library(corrplot)
corrplot 0.95 loaded
> cor_matrix <- cor(data_cleaned[c("age", "bmi", "children", "charges")])
> print(cor_matrix)
      age         bmi   children   charges
age 1.00000000 0.12216952 0.05178231 0.48936154
bmi 0.12216952 1.00000000 0.02117866 -0.02267430
children 0.05178231 0.02117866 1.00000000 0.05466755
charges 0.48936154 -0.02267430 0.05466755 1.00000000
```

# Nuage de points entre age et les charges

Interprétation : Tendance linéaire positive entre charges et age (charges augmentent avec l'âge).  
Forte dispersion, suggérant l'influence d'autres variables (ex. smoker,bmi).



# Théorie:

Modèle : Prédire  $Y$  [charges] à partir de  $k$  variables explicatives ( $X_1$ : âge,  $X_2$ : sexe,  $X_3$ : BMI,  $X_4$ : enfants,  $X_5$ : fumeur,  $X_6$ : région)

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \dots + a_6 X_6 + \varepsilon$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \quad \text{avec } \boldsymbol{\alpha} = \mathbf{X}'\mathbf{X}^{-1}\mathbf{X}'\mathbf{Y}$$

Hypothèses :

$E[\varepsilon] = 0$ ,  $\text{Var}(\varepsilon\varepsilon') = \sigma^2 \mathbf{I}_{nn}$  (indépendance, homoscédasticité).

$\varepsilon \sim N(0, \sigma^2 \mathbf{I}_{nn})$  (normalité des résidus).

$\text{rg}(X) = k$  (linéarité, matrice  $X$  de rang plein)

# Résultat de la régression linéaire :

```
> new_model <- lm(charges ~ age + sex + bmi + children + smoker + region, data = data_cleaned)
> summary(new_model)

Call:
lm(formula = charges ~ age + sex + bmi + children + smoker +
region, data = data_cleaned)

Residuals:
    Min      1Q  Median      3Q     Max 
-4708.0 -1125.5  -601.4   55.4 15964.1 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4157.228    556.198  -7.474 1.56e-13 ***
age          243.569     6.467   37.664 < 2e-16 ***
sexmale      -341.270   179.945  -1.897 0.058148 .  
bmi          84.029    16.294    5.157 2.97e-07 ***
children     367.727    74.741    4.920 9.95e-07 ***
smokeryes    16263.713   304.166   53.470 < 2e-16 ***
regionnorthwest -598.258  255.529  -2.341 0.019394 *  
regionsoutheast -1002.652  263.048  -3.812 0.000146 *** 
regionsouthwest -1016.862  257.414  -3.950 8.29e-05 *** 
---
Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 ' ' 1 

Residual standard error: 3015 on 1118 degrees of freedom
Multiple R-squared:  0.7895,    Adjusted R-squared:  0.788 
F-statistic: 524.2 on 8 and 1118 DF,  p-value: < 2.2e-16
```

# Reamrque: comparaison entre le dataset avant et après nettoyage:

```
> new_model <- lm(charges ~ age + sex + bmi + children + smoker + region, data = data_cleaned)
> summary(new_model)

Call:
lm(formula = charges ~ age + sex + bmi + children + smoker +
region, data = data_cleaned)

Residuals:
    Min      1Q  Median      3Q     Max 
-4708.0 -1125.5 -601.4   55.4 15964.1 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4157.228   556.198 -7.474 1.56e-13 ***
age          243.569    6.467 37.664 < 2e-16 ***
sexmale     -341.270  179.945 -1.897 0.058148 .  
bmi           84.029   16.294  5.157 2.97e-07 ***
children     367.727   74.741  4.920 9.95e-07 ***
smokeryes   16263.713  304.166 53.470 < 2e-16 ***
regionnorthwest -598.258  255.529 -2.341 0.019394 *  
regionsoutheast -1002.652  263.048 -3.812 0.000146 *** 
regionsouthwest -1016.862  257.414 -3.950 8.29e-05 *** 
...
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3015 on 1118 degrees of freedom
Multiple R-squared:  0.7895   Adjusted R-squared:  0.788 
F-statistic: 524.2 on 8 and 1118 DF,  p-value: < 2.2e-16
```

```
> model <- lm(charges ~ age + sex + bmi + children + smoker + region, data = data)
> summary(model)

Call:
lm(formula = charges ~ age + sex + bmi + children + smoker +
region, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-11304.9 -2848.1 - 982.1   1393.9 29992.8 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -11938.5    987.8 -12.086 < 2e-16 ***
age          256.9     11.9  21.587 < 2e-16 ***
sexmale     -131.3    332.9 -0.394 0.693348  
bmi           339.2     28.6 11.860 < 2e-16 ***
children     475.5    137.8  3.451 0.000577 *** 
smokeryes   23848.5   413.1  57.723 < 2e-16 ***
regionnorthwest -353.0    476.3 -0.741 0.458769  
regionsoutheast -1035.0   478.7 -2.162 0.030782 *  
regionsouthwest -960.0    477.9 -2.009 0.044765 * 
...
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,  Adjusted R-squared:  0.7494 
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

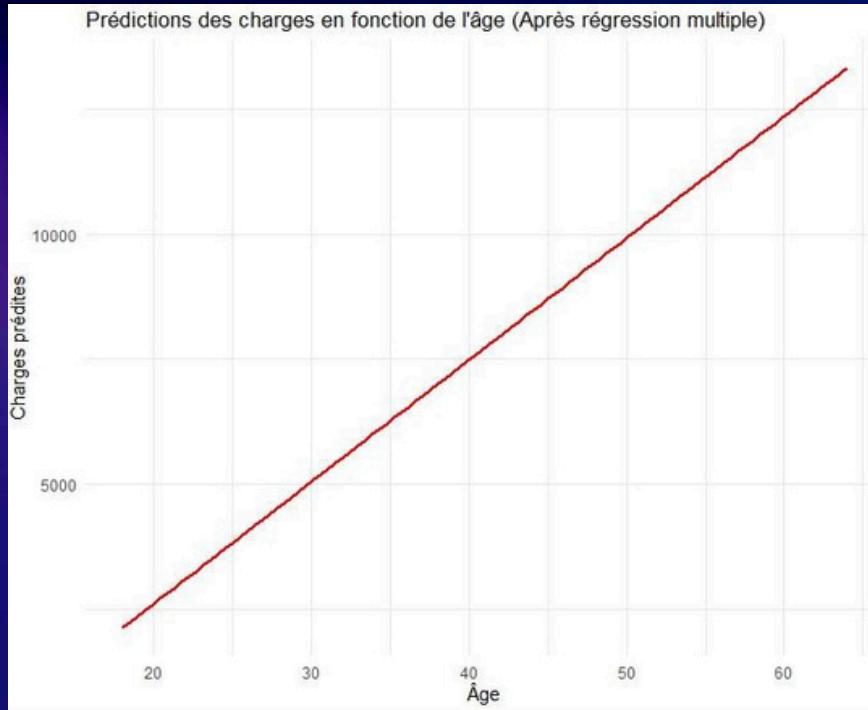
## Interprétation des résultats :

- **R<sup>2</sup>ajusté:** Le R<sup>2</sup>a augmenté après le nettoyage, ce qui indique que le modèle explique une plus grande proportion de la variabilité de charges (de 74,9 % à 78,7 %). Cela suggère que les outliers et points influents introduisaient du bruit, et leur suppression a amélioré la qualité du modèle.
- **Erreur standard résiduelle:** Avant : 6062.

Après : 3015.

Interprétation : L'erreur standard résiduelle a diminué de manière significative (presque divisée par 2). Cela signifie que les prédictions du modèle sont plus précises après nettoyage, car les résidus (écart entre les valeurs observées et prédites) sont plus petits.

## Exemple de la régression multiple : Prédiction des charges en fonction de l'âge:



# Test des paramètres individuellement:

Test :  $H_0: a_i = 0$   
H1:  $a_i \neq 0$

Statistique :  $T = \frac{\hat{a}_i}{\hat{\sigma}_i}$  avec  $i \in \{1, \dots, 6\}$

Décision: on prend  $\alpha = 0.05$ . D'après les résultats précédents on a : p-value de Sexmale est inférieure à

sexmale	-341.270	179.945	-1.897	0.058148
---------	----------	---------	--------	----------

Mais on ne peut pas décider!! : Vu que la variable sex est catégorielle et Sexmale représente une de ses deux variables indicatrices (dummies) ; (R a pris Sexfemale comme référence) et donc on va décider la significativité de cette variable Sex dans les prochaines slides (dans la partie anova). Même scénario avec la variable Région.  
D'où l'importance de recourir à un test global tel que l'ANOVA, qui permet de tester l'utilité de la variable catégorielle dans son ensemble

# Test de significativité globale :

$$H_0: a_1 = a_2 = \dots = a_6 = 0$$

$$H_1: \exists j \in \{1, 6\} a_j \neq 0$$

Test :

$$\{ \quad \}$$

Statistique :  $F = \frac{\text{Variance expliquée par le modèle}_K}{\text{Variance résiduelle}_{(n-k-1)}}$

Décision: on prend  $\alpha = 0.05$ . D'après les résultats précédents on a :

F-statistic: 524.2 on 8 and 1118 DF, p-value: < 2.2e-16

Donc on rejette  $H_0$  : Donc la régression est utile , et on peut procéder aux étapes suivantes!!

## Interprétation des résultats :

- **R<sup>2</sup>ajusté:** Le R<sup>2</sup>a augmenté après le nettoyage, ce qui indique que le modèle explique une plus grande proportion de la variabilité de charges (de 74,9 % à 78,7 %). Cela suggère que les outliers et points influents introduisaient du bruit, et leur suppression a amélioré la qualité du modèle.
- **Erreur standard résiduelle:** Avant : 6062.

Après : 3015.

Interprétation : L'erreur standard résiduelle a diminué de manière significative (presque divisée par 2). Cela signifie que les prédictions du modèle sont plus précises après nettoyage, car les résidus (écart entre les valeurs observées et prédites) sont plus petits.

04

# Méthode progressive

---

# Méthode progressive

On commence avec un modèle vide :  $Y = a_0 + \varepsilon$  (interceptseul).

Ajoute une variable à la fois ( $X_j$ : âge, sexe, IMC, enfants, fumeur, région) en maximisant l'amélioration du modèle.

Critère d'ajout : coefficient de détermination (on choisit le max) :

tester la contribution marginale:

Si p-value <  $\alpha$  (typiquement 0.05), la variable est ajoutée.

Critère d'arrêt : Arrêt si aucune variable restante n'améliore significativement le modèle (p-value  $> \alpha$ )

$$\begin{aligned} r^2_{Y, X_{k+1}/X_1, \dots, X_k} &= \frac{SCR(X_1, \dots, X_k) - SCR(X_1, \dots, X_k, X_{k+1})}{SCR(X_1, \dots, X_k)} \\ &= \frac{SCE(X_1, \dots, X_k, X_{k+1}) - SCE(X_1, \dots, X_k)}{SCR(X_1, \dots, X_k)} \end{aligned}$$

$$\begin{aligned} F &= \frac{CME(X_{j_2}/X_{j_1})}{CMR(X_{j_1}, X_{j_2})} \\ &= \frac{SCE(X_{j_1}, X_{j_2}) - SCE(X_{j_1})}{\frac{SCR(X_{j_1}, X_{j_2})}{n-2}} \end{aligned}$$

# AIC (AkaikeInformation Criterion)

Critère AIC (AkaikeInformation Criterion)

**Théorie :**Formule :  $AIC = -2\log(L) + 2k \cdot \log(L)$

Log-vraisemblance (qualité d'ajustement).

k : Nombre de paramètres (pénalité pour la complexité)

**Objectif :**Minimiser l'AIC pour choisir le meilleur modèle (équilibre ajustement et simplicité). Utilisé par défaut dans step() (R) pour la méthode progressive / pas à pas .

# Méthode progressive

```
# Modèle initial vide (intercept seul)
model_null <- lm(charges ~ 1, data = data_cleaned)

# Modèle complet avec toutes les variables
model_full <- lm(charges ~ age + sex + bmi + children + smoker + region, data = data_cleaned)

# Sélection progressive (forward selection)
step_model <- step(model_null, direction = "forward",
  scope = formula(model_full), trace = 1)
```

Start: AIC=19007.07

charges ~ 1

	Df	Sum of Sq	RSS	AIC
+ smoker	1	2.3736e+10	2.4558e+10	19047
+ age	1	1.1565e+10	3.6729e+10	19501
+ children	1	1.4433e+08	4.8150e+10	19806
<none>		4.8294e+10		19807
+ region	3	2.2510e+08	4.8069e+10	19808
+ bmi	1	2.4829e+07	4.8269e+10	19809
+ sex	1	1.8186e+07	4.8276e+10	19809

Step: AIC=19046.91

charges ~ smoker

Étape 1

	Df	Sum of Sq	RSS	AIC
+ age	1	1.3791e+10	1.0767e+10	18120
+ bmi	1	7.0934e+08	2.3849e+10	19016
+ children	1	4.0571e+08	2.4152e+10	19030
+ sex	1	5.4763e+07	2.4503e+10	19046
<none>		2.4558e+10		19047
+ region	3	1.2289e+08	2.4435e+10	19047

Step: AIC=18119.62

charges ~ smoker + age

	Df	Sum of Sq	RSS	AIC
+ children	1	208695092	1.0558e+10	18100
+ bmi	1	183912178	1.0583e+10	18102
+ region	3	113743452	1.0653e+10	18114
+ sex	1	21500855	1.0745e+10	18119
<none>			1.0767e+10	18120

Step: AIC=18099.56

charges ~ smoker + age + children

	Df	Sum of Sq	RSS	AIC
+ bmi	1	181806830	1.0376e+10	18082
+ region	3	123962166	1.0434e+10	18080
+ sex	1	24913628	1.0533e+10	18099
<none>			1.0558e+10	18100

Step: AIC=18081.98

charges ~ smoker + age + children + bmi

	Df	Sum of Sq	RSS	AIC
+ sex	1	32703607	1.0165e+10	18067
<none>			1.0198e+10	18068

Step: AIC=18066.83

charges ~ smoker + age + children + bmi + region + sex

Étape 5

## Modèle finale par la Méthode progressive:

```
>
> # Afficher le modèle final
> summary(step_model)

Call:
lm(formula = charges ~ smoker + age + children + bmi + region +
    sex, data = data_cleaned)
```

# Analyse de la MulticolinéaritéVIF : (facteurs d'inflation de la variance)

VIF (Variance Inflation Factor) : Mesure la colinéarité entre variables explicatives.

Formule:  $VIF_j = \frac{1}{1-R_j^2}$  Avec  $j \in \{1, \dots, 6\}$

\* VIF <5 : Pas de colinéarité problématique.

\* VIF  $\geq 5$  : Colinéarité potentielle, peut affecter la stabilité des coefficients.

```
>
> # Charger la bibliothèque car
> library(car)
Le chargement a nécessité le package : carData
>
> # Ajuster le modèle final (basé sur les résultats de la méthode progressive)
> final_model <- lm(charges ~ smoker + age + bmi + children + region, data = data_cleaned)
>
> # Calculer le VIF pour chaque variable
> vif_values <- vif(final_model)
>
> # Afficher les résultats
> print(vif_values)
      GVIF Df GVIF^(1/(2*Df))
smoker   1.050037  1        1.024713
age       1.022309  1        1.011093
bmi       1.149764  1        1.072270
children  1.006526  1        1.003258
region    1.093747  3        1.015047
>
> # Interprétation des VIF : Ajouter un message pour signaler les valeurs élevées
> threshold <- 5 # Seuil couramment utilisé pour le VIF
> if (any(vif_values > threshold)) {
+   cat("Attention : Colinéarité détectée ! Les variables suivantes ont un VIF >", threshold, ":\n")
+   print(names(vif_values[vif_values > threshold]))
+   print(vif_values[vif_values > threshold])
+ } else {
+   cat("Aucune colinéarité problématique détectée (tous les VIF <", threshold, ".)\n")
+ }
Aucune colinéarité problématique détectée (tous les VIF < 5 ).
```

# Résultats de la Sélection des Variables et Analyse de la Multicolinéarité

**Toutes les variables explicatives** (age, sex, bmi, children, smoker, region) ont été retenues.

Chaque variable améliore significativement le modèle (basé sur l'AIC)

Analyse de la multicolinéarité(VIF):

- VIF calculé pour chaque variable : tous les VIF < 5.
- Interprétation : Aucune multicolinéaritéproblématique détectée.

Conclusion:

- Le modèle final est robuste, avec des variables pertinentes et indépendantes.
- Les résultats confirmant la validité du modèle pour prédire charges.

05

# Méthode pas à pas

---

# Méthode pas à pas (stepwiseselection)

On commence avec un modèle vide:

Étape 1: Régression simple pour chaque

$X_j$  (âge, sexe, BMI, enfants, fumeur, région) en calculant  $F_j$ :

$$F_j = \frac{CME(X_j)}{CMR(X_j)} \quad \text{Avec } j \in \{1, \dots, 6\}$$

Choisir  $X_{j1}$  avec  $F_{j1}$  maximale, tester

$F_{j1} > F_{\alpha=0.05}$ . Si non, arrêt (pas de lien linéaire significatif).

Étape 2: Sélectionner une deuxième variable

$X_{j2}$ , en calculant:

$$F_{j/j_1} = \frac{CME(X_j/X_{j1})}{CMR(X_{j1}, X_j)}, \quad \text{Avec } j \in \{1, \dots, 6\}$$

Vérifier  $X_j$ : Calculer  $F_{j1/j2}$  si  $F_{j1+j2} > F_{\alpha=0.05}$ , garder  $X_{j1}$ , sinon retirer

Étape 3: Continuer pour les autres variables (ajout et suppression).

Critère d'arrêt:  $\hat{e}$  Aucun ajout ( $F_j < F_{\alpha}$ ) ni suppression ( $F_j < F_r$ ) possible.

# Méthode pas à pas

```

> # Modèle initial vide (intercept seul)
> model_null <- lm(charges ~ 1, data = data_cleaned)
>
> # Modèle complet avec toutes les variables
> model_full <- lm(charges ~ age + sex + bmi + children + smoker + region, data = data_cleaned)
>
> # Méthode pas-à-pas (stepwise selection) avec critère AIC
> step_model <- step(model_null,
+                     direction = "both",
+                     scope = list(lower = model_null, upper = model_full),
+                     trace = 1)

```

Start: AIC=19807.07  
charges ~ 1

	Df	Sum of Sq	RSS	AIC
+ smoker	1	2.3736e+10	2.4558e+10	19047
+ age	1	1.1565e+10	3.6729e+10	19501
+ children	1	1.4433e+08	4.8150e+10	19806
<none>		4.8294e+10		19807
+ region	3	2.2510e+08	4.8069e+10	19808
+ bmi	1	2.4829e+07	4.8269e+10	19809
+ sex	1	1.8186e+07	4.8276e+10	19809

Step: AIC=19046.91  
charges ~ smoker

	Df	Sum of Sq	RSS	AIC
+ age	1	1.3791e+10	1.0767e+10	18120
+ bmi	1	7.0934e+08	2.3849e+10	19016
+ children	1	4.0571e+08	2.4152e+10	19030
+ sex	1	5.4763e+07	2.4503e+10	19046
<none>		2.4558e+10		19047
+ region	3	1.2289e+08	2.4435e+10	19047
- smoker	1	2.3736e+10	4.8294e+10	19807

Step: AIC=18119.62  
charges ~ smoker + age

Étape 2

	Df	Sum of Sq	RSS	AIC
+ children	1	2.0870e+08	1.0558e+10	18100
+ bmi	1	1.8391e+08	1.0583e+10	18102
+ region	3	1.1374e+08	1.0653e+10	18114
+ sex	1	2.1501e+07	1.0745e+10	18119
<none>			1.0767e+10	18120
- age	1	1.3791e+10	2.4558e+10	19047
- smoker	1	2.5962e+10	3.6729e+10	19501

Step: AIC=18099.56  
charges ~ smoker + age + children

Étape 3

	Df	Sum of Sq	RSS	AIC
+ bmi	1	1.8181e+08	1.0376e+10	18082
+ region	3	1.2396e+08	1.0434e+10	18092
+ sex	1	2.4914e+07	1.0533e+10	18099
<none>			1.0558e+10	18100
- children	1	2.0870e+08	1.0767e+10	18120
- age	1	1.3594e+10	2.4152e+10	19030
- smoker	1	2.6129e+10	3.6687e+10	19501

Step: AIC=18081.98  
charges ~ smoker + age + children + bmi

	Df	Sum of Sq	RSS	AIC
+ sex	1	3.2704e+07	1.0165e+10	18067
<none>			1.0198e+10	18068
- region	3	1.7830e+08	1.0376e+10	18082
- children	1	2.1605e+08	1.0414e+10	18090
- bmi	1	2.3614e+08	1.0434e+10	18092
- age	1	1.2947e+10	2.3145e+10	18990
- smoker	1	2.5965e+10	3.6163e+10	19493

Step: AIC=18066.83  
charges ~ smoker + age + children + bmi + region + sex

Étape 5

## Modèle finale par la Méthode pas à pas:

```
>  
> # Afficher les variables sélectionnées  
> selected_vars <- names(coef(step_model))[-1] # Exclure l'intercept  
> cat("Variables sélectionnées : ", paste(selected_vars, collapse = ", "), "\n")  
Variables sélectionnées : smokeryes, age, children, bmi, regionnorthwest, regionsoutheast, regionsouthwest, sexmale  
> |
```

Remarque : La variable sexa été ajoutée à la fin, car bien qu'ellene soit significative seule (p-value=0.058), elle améliorait légèrement le critère AIC global du modèle.

# Vérification de l'importance relative des variables du modèle finale données par les méthodes progressive et pas à pas :

Les coefficients standardisés permettent de comparer l'importance relative des variables sur une échelle commune (en écarts-types) :

$$\beta_j^{\text{std}} = \beta_j \cdot \frac{\text{écart-type}(X_j)}{\text{écart-type}(Y)}$$

Les coefficients standardisés confirment les résultats des méthodes progressive et pas à pas :

L'ordre d'importance (smoker > age > bmi > children > region > sex) reflète les attentes théoriques (par exemple, le tabagisme et l'âge sont des facteurs majeurs des coûts médicaux).

```
>
> # Charger le package
> library(lm.beta)
>
> # Calculer le modèle avec coefficients standardisés
> standardized_model <- lm.beta(step_model)
>
> # Afficher les coefficients standardisés
> summary(standardized_model)
```

```
Call:
lm(formula = charges ~ smoker + age + children + bmi + region +
   sex, data = data_cleaned)

Residuals:
    Min      1Q  Median      3Q     Max 
-4708.0 -1125.5  -601.4   55.4 15964.1 

Coefficients:
            Estimate Standardized   Std. Error t value Pr(>|t|)    
(Intercept) -4.157e+03        NA 5.562e+02 -7.474 1.56e-13 ***
smokeryes   1.626e+04        7.520e-01 3.042e+02 53.470 < 2e-16 ***
age          2.436e+02        5.228e-01 6.467e+00 37.664 < 2e-16 ***
children    3.677e+02        6.775e-02 7.474e+01 4.920 9.95e-07 ***
bmi          8.403e+01        7.592e-02 1.629e+01 5.157 2.97e-07 ***
regionnorthwest -5.983e+02   -3.977e-02 2.555e+02 -2.341 0.019394 *  
regionsoutheast -1.003e+03   -6.681e-02 2.630e+02 -3.812 0.000146 ***
regionsouthwest -1.017e+03   -6.720e-02 2.574e+02 -3.950 8.29e-05 ***
sexmale      -3.413e+02   -2.606e-02 1.799e+02 -1.897 0.058148 .  
---
Signif. codes:  0 '*****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 ' ' 1

Residual standard error: 3015 on 1118 degrees of freedom
Multiple R-squared:  0.7895,    Adjusted R-squared:  0.788
F-statistic: 524.2 on 8 and 1118 DF,  p-value: < 2.2e-16
```

# Résumé des coefficients et de l'importance relative des variables dans le modèle de prédiction des charges médicales

Variable	Coefficient (non std)	Interprétation (non std)	Coefficient (std)	Importance relative
smokeryes	26263.71	+26263.71 par rapport à non-fumeur	2.171	Impact le plus important
age	243.57	+243.57 par année	0.282	Deuxième impact le plus important
bmi	84.03	+84.03 par point d'IMC	0.042	Impact modéré
children	367.73	+367.73 par enfant	0.036	Impact modéré
regionnnorthwest	-598.25	-598.25 par rapport à southeast	-0.049	Impact faible
regionsouthwest	-1002.65	-1002.65 par rapport à southeast	-0.083	Impact faible
regionnortheast	-1016.86	-1016.86 par rapport à southeast	-0.084	Impact faible
sexmale	-341.27	-341.27 par rapport à femme (non significatif)	-0.028	Impact faible

## Remarque:

Les coefficients bruts (non standardisés) mesurent l'effet d'une unité de chaque variable sur charges , par exemple :

- Une unité de changement pour smoker(0 à 1) est un changement majeur (non-fumeur à fumeur).
- Une unité de changement pour age(1 an) est relativement petite (1 an sur une plage de 20 à 60 ans).
- Une unité de changement pour bmi(1 point) est également petite par rapport à sa plage.

06

# ANOVA: partiethéorie et pratique

---

## Contexte et utilité de ANOVA

Utilité : Évaluer si les variables explicatives (dummy) : smoker et sex contribuent significativement à expliquer la variable dépendante (coûts médicaux).

# 1) a-ANOVA Théorique -Une Variable Dummy

Contexte : Modèle avec une variable dummy(ex. : smoker avec valeurs yes/no).

Modèle :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

avec  $X_1$ : âge et  $X_2$ : smoker = 1 si fumeur, 0 sinon.

Remarque : L'ANOVA est utilisée pour évaluer la contribution de chaque variable au modèle. Inclure âge permet de comparer la significativité des variables dummy par rapport à une variable continue importante.

# 1) a-ANOVA Théorique -Une Variable Dummy

Test ANOVA:

H<sub>0</sub>:  $\beta_2 = 0$  (la variable smoker n'a pas d'effet significatif).  
H<sub>1</sub>:  $\beta_2 \neq 0$  (la variable smoker a un effet significatif).

Statistique :

$$F = \frac{\frac{SCE}{k-1}}{\frac{SCR}{n-k}}$$

Décision :

Une valeur F élevée avec une p-value < 0,05 rejette (H<sub>0</sub>), indiquant que smoker est significatif.

Source de variance	Somme des carrés	ddl	Carres moyens SCE	Fobs
Régression	SCE	K-1=2	$\frac{2}{SCR}$	
Résiduelle	SCR	n-k = n-3 = 1127 3 = 1124	1124	$\frac{SCE}{\frac{k-1}{n-k}} = \frac{1124}{2} \times \frac{SCE}{SCR}$
Total	SCT	n-1 = 1126		

AVEC :

n : nombre d'observations = 1127  
K : nombre de paramètres du modèle = 3

## 1) b -ANOVA Pratique -Une Variable Dummy

```
> model <- lm(charges ~ age + smoker, data = data_cleaned)
> anova(model)
Analysis of Variance Table

Response: charges
            Df    Sum Sq   Mean Sq F value    Pr(>F)
age          1 1.1565e+10 1.1565e+10 1207.4 < 2.2e-16 ***
smoker       1 2.5962e+10 2.5962e+10 2710.3 < 2.2e-16 ***
Residuals 1124 1.0767e+10 9.5790e+06
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### Interprétation :

- smokera une p-valeur < 0,05, donc significatif.
- Contribution desmokerà lavarianceexpliquéeest importante (voir Sum Sq).

## 2) a-ANOVA Théorique –Deux Variables Dummy

Contexte : Modèle avec deux variables dummy(ex.: smokeret sex).

Modèle :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

avec  $X_1$ : âge et  $X_2$  smoker = 1 si fumeur, 0 sinon  
et  $X_3$  sex=1 si homme, 0 si femme.

## 2) b-ANOVA Théorique -Deux Variable Dummy

### Test ANOVA:

H<sub>0</sub>:  $\beta_2 = \beta_3 = 0$  (la variable smoker n'a pas d'effet significatif).

H<sub>1</sub>: Au moins une variable  $\in \{1, 2\}$  (au moins une variable a un effet significatif)

### Statistique :

$$F = \frac{\frac{SCE}{k-1}}{\frac{SCR}{n-k}}$$

Source de variance	Somme des carrés	ddl	Carrés moyens SCE	Fobs
Régression	SCE	K-1=3	$\frac{3}{SCR}$	
Résiduelle	SCR	$n-k = n-4 = 1127 - 4 = 1123$	$\frac{1123}{3}$	$\frac{SCE}{k-1} = \frac{1123}{3} \times \frac{SCE}{SCR}$
Total	SCT	$n-1 = 1126$		

### Décision:

Une valeur F élevée avec une p-valeur < 0,05 rejette (H<sub>0</sub>), indiquant que smoker est significatif.

### AVEC :

n : nombre d'observations = 1127

K : nombre de paramètres du modèle = 3

## 2) b -ANOVA Pratique -Deux Variables Dummy

```
> model <- lm(charges ~ age + smoker + sex, data = data_cleaned)
> anova(model)
Analysis of Variance Table

Response: charges
            Df    Sum Sq   Mean Sq   F value Pr(>F)
age          1 1.1565e+10 1.1565e+10 1208.6955 <2e-16 ***
smoker       1 2.5962e+10 2.5962e+10 2713.3354 <2e-16 ***
sex          1 2.1501e+07 2.1501e+07   2.2471 0.1341
Residuals 1123 1.0745e+10 9.5684e+06
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### Interprétation :

- `smoker` a une p-valeur < 0,05, donc significatif.
- `sex` n'est pas significatif (p-valeur > 0,05), suggérant un effet faible.

07

# Limites de la régression

---

# L'ANOVA:

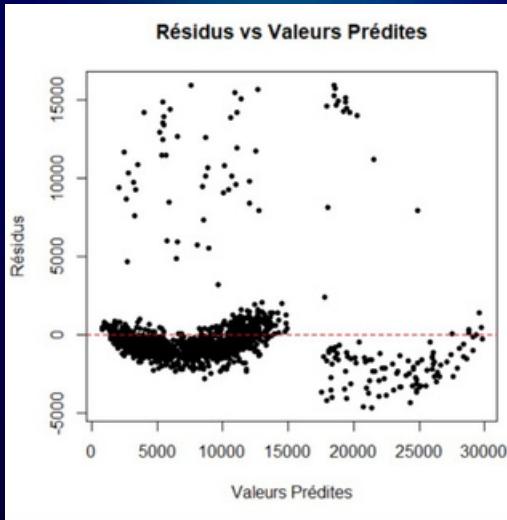
Étant donné que l'ANOVA repose sur les hypothèses de linéarité, d'indépendance et de normalité des résidus, nous vérifierons dans les diapositives suivantes si ces conditions sont remplies. Si ce n'est pas le cas, nous identifierons les limites de la régression et explorerons d'autres alternatives à la régression.

# Vérification des hypothèses d'homoscédacité et de normalité :



## homoscédacité

```
> # Graphique des résidus vs valeurs prédictes
> plot(fitted(step_model), residuals(step_model),
+       main = "Résidus vs Valeurs Prédictes",
+       xlab = "Valeurs Prédictes", ylab = "Résidus",
+       pch = 20)
> abline(h = 0, col = "red", lty = 2)
```



NUAGE DE POINTS DISPERSÉ DE MANIÈRE UNIFORME DE PART ET D'AUTRE DE LA LIGNE  $y = 0$ .

AUCUNE TENDANCE OU MOTIF VISIBLE (PAS D'ÉLARGISSEMENT, DE RÉTRÉCISSEMENT, OU DE FORME EN U).

FORME EN ENTONNOIR : La dispersion des résidus augmente clairement à mesure que les valeurs prédictes augmentent : Cette forme en entonnoir (variance croissante) est un signe classique d'hétérosécédasticité.

ASYMÉTRIE : Les résidus positifs (au-dessus de la ligne  $y = 0$ ) semblent plus nombreux et plus dispersés que les résidus négatifs, surtout pour les grandes valeurs prédictes. Cela suggère une possible asymétrie dans les erreurs, qui peut être liée à une non-normalité des résidus

tre

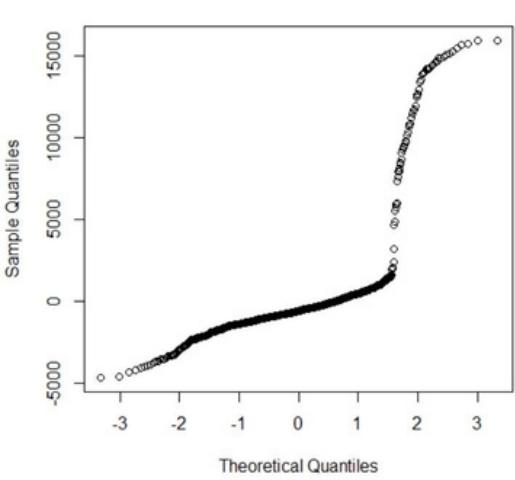
```
> # Q-Q Plot pour vérifier la normalité
> qqnorm(residuals(step_model), main = "Q-Q Plot des Résidus")
> qqline(residuals(step_model), col = "red")
> # Test de Shapiro-Wilk
> shapiro.test(residuals(step_model))

Shapiro-Wilk normality test

data: residuals(step_model)
W = 0.5347, p-value < 2.2e-16
```

**H0 :** Les résidus sont normaux.  
**H1 :** Les résidus ne sont pas normaux.  
 Si la p-valeur < 0,05, rejetez H0 → les résidus ne sont pas normaux.

Q-Q Plot des Résidus



Forme générale : Le Q-Q plot montre une courbure en S (les extrémités s'écartent de la droite dans des directions opposées), ce qui est un signe classique de non-normalité

```
library(rms)

# Ajoutez le modèle final (méthode pas-à-pas avec step())
final_model <- lm(charges ~ smoker + age + children + bmi + region + sex,
                   data = data_cleaned)

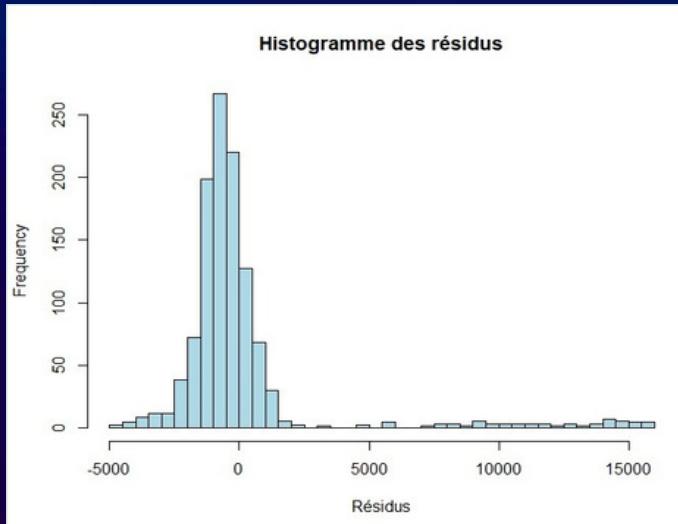
# --- Diagnostic 1 : Normalité des résidus ---
# Extraire les résidus
residuale <- residuals(final_model)

# Vérifier que les résidus ont été correctement extraits
if (length(residuale) == 0) {
  stop("Erreur : Les résidus n'ont pas été correctement extraits. Vérifiez le modèle.")
}

# --- Histogramme des résidus avec couche de densité ---
# Avez-vous que la fenêtre graphique est active (facultatif, mais utile pour certains environnements)
dev.new() # Ouvre une nouvelle fenêtre graphique (optionnel dans RStudio)

hist(residuale,
     breaks = 50,
     main = "Histogramme des résidus",
     xlab = "Résidus",
     ylab = "Fréquence",
     col = "lightblue",
     border = "black")
```

Histogramme des résidus



## Correction de la non-normalité : La transformation logarithmique de Y : charges

```
> # Ajuster le modèle avec log(charges)
> step_model_log <- lm log(charges) ~ smoker + age + children + bmi + region + sex,
+                         data = data_cleaned)
>
> # Test de Shapiro-Wilk sur les résidus
> shapiro.test(residuals(step_model_log))
```

Shapiro-Wilk normality test

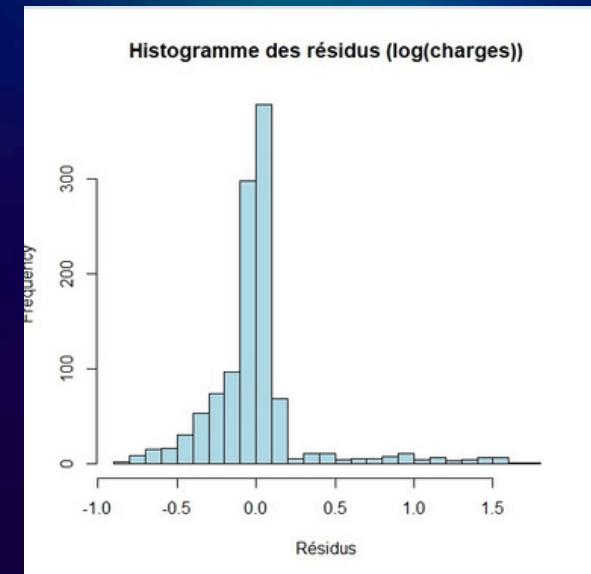
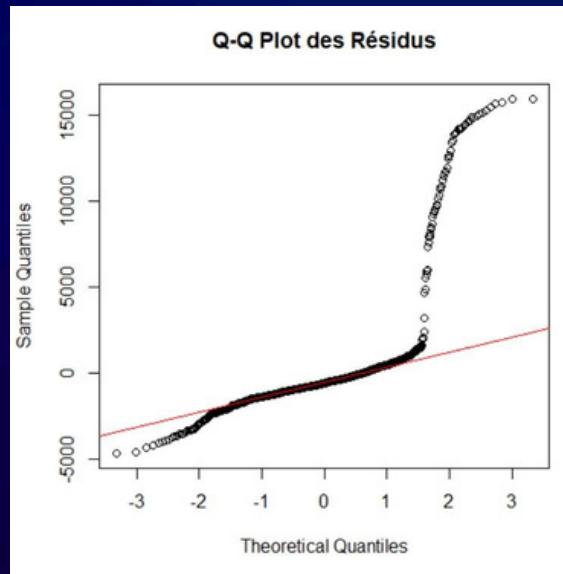
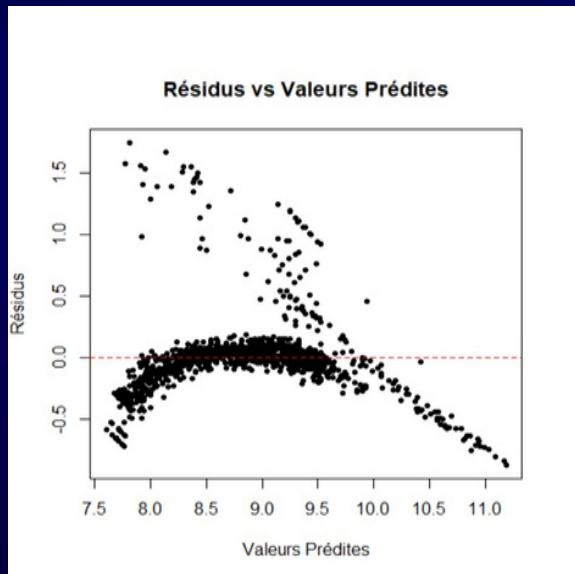
```
data: residuals(step_model_log)
W = 0.75767, p-value < 2.2e-16
```

# Résultats de La transformation logarithmique

```
> # Réajuster le modèle avec log(charges)
> step_model_log <- lm(log(charges) ~ smoker + age + children + bmi + region + sex,
+                         data = data_cleaned)
>
> # Test de Shapiro-Wilk sur les résidus
> shapiro.test(residuals(step_model_log))

Shapiro-Wilk normality test

data: residuals(step_model_log)
W = 0.75767, p-value < 2.2e-16
```



# Conclusion : nature des données (charges) posent un problème pour la normalité des erreurs !!

La transformation logarithmique a partiellement corrigé la non-normalité, mais les résidus ne sont pas encore complètement normaux en raison des déviations résiduelles dans les queues: Cela est dû à la nature asymétrique des dépenses médicales.

## Que faire si la normalité n'est pas complètement atteinte?

a) Accepter la non-normalité (car l'échantillon est grand):

L'échantillon est grand ( $>100$ ), la non-normalité résiduelle a un impact limité sur les tests de significativité grâce au théorème central limite (Les estimateurs des coefficients sont asymptotiquement normaux)

b) Utiliser un modèle alternatif :

- Les **GLM** nous permettent de contourner ce problème en utilisant une **distribution Gamma**, qui correspond mieux à la forme des dépenses médicales. Cette flexibilité en fait un outil puissant pour notre étude.

08

# Méthode alternative : GLM

---

# Introduction aux Modèles Linéaires Généralisés (GLM)

**Contexte:** La régression linéaire a montré des limites pour modéliser les dépenses médicales ("charges") :

Problème : Non-normalité des résidus (test de Shapiro-Wilk: p-valeur < 2.2e-16).

Conséquence : Estimations biaisées et tests de significativité non fiables.

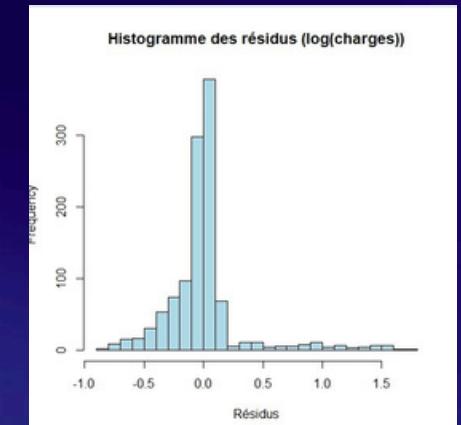
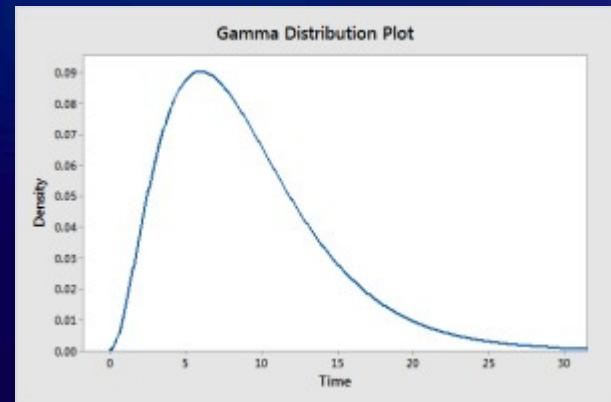
**Solution :** Les Modèles Linéaires Généralisés (GLM) offrent une alternative flexible :

Adaptés aux données non normales, comme les "charges" (positives et asymétriques).

Exemple : GLM avec distribution Gamma pour modéliser les dépenses médicales.

**Objectif :** Améliorer la précision des prédictions en respectant la nature des données (non-normales)

« Pour nos données, la distribution Gamma est idéale car elle gère les valeurs positives et asymétriques, contrairement à la distribution normale utilisée dans la régression linéaire. »



## 2) Théorie : Modèle Linéaire Généralisé (GLM) :

Modèle :

- Prédire Y (charges) à partir des variables explicatives (X<sub>1</sub>, X<sub>2</sub> : sexe, X<sub>3</sub> : BMI, X<sub>4</sub> : enfants, X<sub>5</sub> : fumeur, X<sub>6</sub> : région)

GLM Gamma avec lien logarithmique:

$$\text{Log}(E(Y|X)) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \dots + \alpha_6 X_6 + \varepsilon$$

- Y ~ Gamma (adapté aux données positives et asymétriques).

Hypothèses:

- Distribution Gamma : Y ~ Gamma,  $\text{Var}(Y) \propto [E(Y | X)]^2$
- Résidus : Résidus de déviance centrés autour de 0, sans motifs systématiques.
- Indépendance : Observations indépendantes.

### 3) vérification des hypothèses du GLM Gamma:

#### a-Résultat de GLM :

```
> # Charger les bibliothèques nécessaires
> library(stats)
>
>
> glm_model <- glm(charges ~ smoker + age + children + bmi + region + sex,
+                     family = Gamma(link = "log"),
+                     data = data_cleaned)
> summary(glm_model)
```

```
Call:
glm(formula = charges ~ smoker + age + children + bmi + region +
    sex, family = Gamma(link = "log"), data = data_cleaned)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.054216  0.091702 76.925 < 2e-16 ***
smokeryes   1.521683  0.050149 30.343 < 2e-16 ***
age          0.035615  0.001066 33.403 < 2e-16 ***
children     0.093245  0.012323  7.567 7.97e-14 ***
bmi          0.012406  0.002687  4.618 4.32e-06 ***
regionnorthwest -0.116134  0.042130 -2.757 0.00594 **
regionsoutheast -0.185531  0.043370 -4.278 2.05e-05 ***
regionsouthwest -0.169817  0.042441 -4.001 6.71e-05 ***
sexmale      -0.070972  0.029668 -2.392 0.01691 *
---
Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1

(Dispersion parameter for Gamma family taken to be 0.2471607)

Null deviance: 598.08 on 1126 degrees of freedom
Residual deviance: 152.98 on 1118 degrees of freedom
AIC: 20947

Number of Fisher Scoring iterations: 7
```

Le Fisher Scoring est une méthode (un algorithme) utilisée pour trouver les meilleurs coefficients.

Le modèle a convergé après 7 itérations de Fisher Scoring, indiquant un ajustement rapide et stable.

### 3) vérification des hypothèses du GLM Gamma:

#### b- Pseudo-R<sup>2</sup>:

```
> # 1. Vérifier l'ajustement global (Pseudo-R2)
> pseudo_r2 <- 1 - (glm_model$deviance / glm_model>null.deviance)
> cat("Pseudo-R2 : ", pseudo_r2, "\n")
Pseudo-R2 : 0.7442208
>
> # 2. Extraire les résidus de déviance
> residuals_dev <- residuals(glm_model, type = "deviance")
> summary(residuals_dev)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.81387 -0.19463 -0.05471 -0.04746  0.02393  2.11790
> |
```

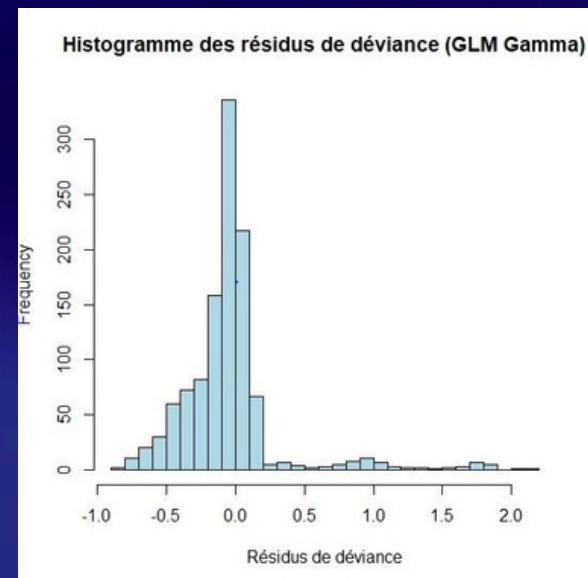
Pseudo-R<sup>2</sup>: de 0.744 (74.4% de la déviance expliquée), ce qui indique un bon ajustement. Un élevé montre que le GLM Gamma capture bien la variabilité de charges.

Pseudo-R<sup>2</sup>

### 3) vérification des hypothèses du GLM Gamma : c-Histogrammes des résidus de deviance :

```
> # 2.1 Histogramme des résidus de déviance
> dev.new()
> hist(residuals_dev,
+       breaks = 30,
+       main = "Histogramme des résidus de déviance (GLM Gamma)",
+       xlab = "Résidus de déviance",
+       col = "lightblue",
+       border = "black")
```

Conclusion : Les résidus de déviance sont globalement bien centrés autour de 0, avec une dispersion raisonnable, ce qui indique que le modèle GLM Gamma ajuste bien les données.



### 3) vérification des hypothèses du GLM Gamma : c-Histogrammes résidus de deviance :

#### Analyse des Résidus :

Hétérosécédasticité contrôlée : Pas de forme de cone: Dispersion homogène des résidus → La structure Gamma + lien log est adaptée.

Biais systématique : Légère sous-estimation des faibles charges (résidus

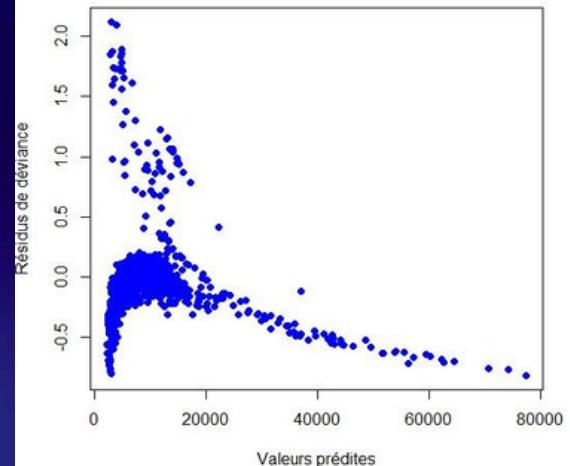
> 0). Légère surestimation des charges élevées (résidus < 0);

Nous

allons immédiatement confirmer que ce biais systématique n'est pas problématique en examinant l'impact des points influents.

```
> # 2.1 Histogramme des résidus de déviance
> dev.new()
> hist(residuals_dev,
+       breaks = 30,
+       main = "Histogramme des résidus de déviance (GLM Gamma)",
+       xlab = "Résidus de déviance",
+       col = "lightblue",
+       border = "black")
> residuals_dev <- residuals(glm_model, type = "deviance")
> plot(fitted(glm_model), residuals_dev,
+       main = "Résidus de déviance vs. Valeurs prédictes (GLM Gamma)",
+       xlab = "Valeurs prédictes",
+       ylab = "Résidus de déviance",
+       pch = 19,
+       col = "blue")
> abline(h = 0, col = "red", lty = 2)
> |
```

Résidus de déviance vs. Valeurs prédictes (GLM Gamma)



### 3) vérification des hypothèses du GLM Gamma :

#### d-Impact des points influents :

#### Détection des points influents :

```
> # Détection des points influents (distance de Cook)
> cooks_d <- cooks.distance(glm_model)
> n <- nrow(data_cleaned)
> influential <- which(cooks_d > 4/n)
> cat("Nombre de points influents : ", length(influential), "\n")
Nombre de points influents : 49
```

Résultat : 49 points influents sur 1127 observations. Proportion:  $49/1127 \approx 0.0435$ , soit 4.35% des observations.

Seuil classique : Un modèle est généralement considéré comme robuste si le nombre de points influents est inférieur à 5% des observations. Ici, **4.35% est juste en dessous de ce seuil, ce qui est acceptable.**

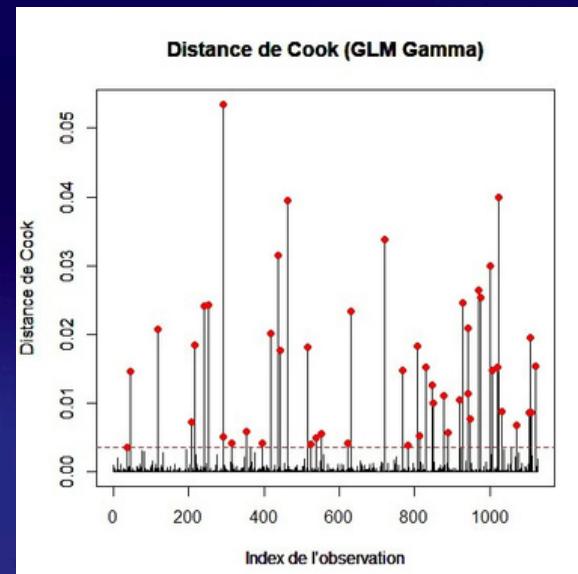
### 3) vérification des hypothèses du GLM Gamma :

#### d-**Impact des points influents :**

##### Distance de cook:

La majorité des distances de Cook sont proches de 0, ce qui est normal. Les 49 points influents (marqués en rouge) dépassent le seuil  $4/n$ , qui est d'environ  $4/1127 \approx 0.00355$ . Quelques points ont des distances de Cook élevées (autour de 0.05), indiquant qu'ils influencent davantage le modèle.

```
> # Graphique des distances de Cook
> dev.new()
> plot(cooks_d, type = "h", main = "Distance de Cook (GLM Gamma)",
+       xlab = "Index de l'observation",
+       ylab = "Distance de Cook")
> abline(h = 4/n, col = "red", lty = 2)
> points(influential, cooks_d[influential], col = "red", pch = 19)
> |
```



## 4) Conclusion: Interprétation des coefficients directs via $\exp(\beta_j)$

Variable	Coefficient	Interprétation (GLM Gamma)	Effet multiplicatif	Significativité
smokeryes	1.521683	Être fumeur multiplie les charges par 4.58	$e^{1.521} \approx 4.58$ (x4.58)	p < 2e-16 ***
age	0.035615	Chaque année multiplie les charges par 1.036	$e^{0.0356} \approx 1.036$ (+3.6%)	p < 2e-16 ***
children	0.083245	Chaque enfant multiplie les charges par 1.087	$e^{0.0832} \approx 1.087$ (+8.7%)	p = 2.37e-11 ***
bmi	0.012406	Chaque point d'IMC multiplie les charges par 1.012	$e^{0.0124} \approx 1.012$ (+1.2%)	p = 4.32e-06 ***
regionnorthwest	-0.116134	Northwest divise les charges par 0.890	$e^{-0.116} \approx 0.890$ (-11%)	p = 0.00595 **
regionsouthwest	-0.185531	Southwest divise les charges par 0.831	$e^{-0.185} \approx 0.831$ (-16.9%)	p = 2.05e-05 ***
regionnortheast	-0.169817	Northeast divise les charges par 0.844	$e^{-0.170} \approx 0.844$ (-15.6%)	p = 6.71e-05 ***
sexmale	-0.070972	Être homme divise les charges par 0.931	$e^{-0.071} \approx 0.931$ (-6.9%)	p = 0.01691 *

Tous les variables sont significatives ( p-value < 0.05 )

## 4) Test de rapport de vraisemblance(modèles emboîtés):

objectif : Tester si un groupe de variables améliore significativement l'ajustement du modèle.

Test :  $H_0:$  Modèle complet: Inclut toutes les variables.

$H_1:$  Modèle réduit : Exclut un sous - ensemble de variables (par exemple, region ou sex ).

Statistique :

$$LR = -2 \times (\log(\text{vraisemblance}_{\text{réduit}}) - \log(\text{vraisemblance}_{\text{complet}}))$$

LR suit une distribution du Chi-deux ( $\chi^2$ ) avec des degrés de liberté égaux à la différence du nombre de paramètres entre les deux modèles.

Décision :

Si p-value < 0.05, le modèle complet est significativement meilleur, donc les variables exclues (par exemple, region) sont significatives dans leur ensemble.

Si La déviance du modèle complet est plus faible que celle du modèle réduit, alors on accepte  $H_0$  (La déviance mesure l'écart entre les données observées et les prédictions du modèle. Une déviance plus faible indique un meilleur ajustement.)

## 4) Test de rapport de vraisemblance(modèles emboîtés) :

### Test de vraisemblance pour région :

```
> # Test de rapport de vraisemblance pour région
> anova(glm_model, glm_model_reduced_region, test = "Chisq")
Analysis of Deviance Table

Model 1: charges ~ smoker + age + children + bmi + region + sex
Model 2: charges ~ smoker + age + children + bmi + sex
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     1118   152.98
2     1121   158.75 -3  -5.7777 3.371e-05 ***
---
Signif. codes:  '*****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



La déviance du modèle complet (152.98) est plus faible que celle du modèle réduit (158.75), ce qui montre que l'ajout de region améliore l'ajustement du modèle.

La p-valeur (3.371e-05) est bien inférieure à 0.05, donc region est significative au niveau global.

## 4) Test de rapport de vraisemblance(modèles emboîtés) :

### Test de vraisemblance pour sex:

```
>
> # Test de rapport de vraisemblance pour sex
> anova(glm_model, glm_model_reduced_sex, test = "Chisq")
Analysis of Deviance Table

Model 1: charges ~ smoker + age + children + bmi + region + sex
Model 2: charges ~ smoker + age + children + bmi + region
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      1118   152.98
2      1119   154.38 -1    -1.4082  0.01699 *
---
Signif. codes:  **** 0.001 *** 0.01 ** 0.05 . 0.1 ' ' 1
>
```

La déviance du modèle complet (152.98) est plus faible que celle du modèle réduit (154.38), ce qui montre que l'ajout de sex améliore l'ajustement du modèle.

la p-value (0.01699) est < 0.05, la variable sex est significative dans son ensemble dans le modèle GLM Gamma.

## 5) Conclusion : interprétation des résultats du modèle GLM :

Le GLM Gamma prédit efficacement les coûts médicaux, identifiant `smoker`, `age`, `children`, et `bmi` comme principaux facteurs d'augmentation, avec des coûts plus élevés pour les fumeurs, dans `northeast` et pour les femmes, confirmés par des tests globaux significatifs pour `region`( $p\text{-value} = 3.371\text{e-}05$ ) et `sex`( $p\text{-value}=0.01699$ ), offrant ainsi des insights clés pour optimiser la gestion des dépenses de santé.

09

# Comparaison des modèles: GML & RL

---

# 1) Introduction à : AIC (AkaikeInformation Criterion) et BIC (Bayesian Information Criterion)

Definition :

AIC et BIC sont des critères utilisés pour comparer des modèles statistiques. Ils évaluent deux aspects :

L'ajustement du modèle : À quel point le modèle explique bien les données (mesuré par la vraisemblance, ou "likelihood").

La complexité du modèle : Le nombre de paramètres du modèle (plus un modèle a de paramètres, plus il est complexe). Ces critères cherchent un compromis : un modèle qui ajuste bien les données sans être inutilement complexe ; C'est à dire plus AIC et BIC sont petits cela indique que le modèle ajuste bien les données :

$$\text{AIC} = -2 \times \log(\text{vraisemblance}) + 2 \times k$$

$$\text{BIC} = -2 \times \log(\text{vraisemblance}) + k \times \log(n)$$

## 2) Comparaison des modèles: GLM : Modèle Linéaire Généralisé et RL : regression lineaire :

```
> # Calculer AIC et BIC pour les deux modèles
> model_comparison <- data.frame(
+   Model = c("GLM Gamma", "RL"),
+   AIC = c(AIC(glm_model), AIC(rl_model)),
+   BIC = c(BIC(glm_model), BIC(rl_model)))
+ )
>
> # Trier selon AIC et BIC et afficher
> cat("Trié selon AIC (plus petit = meilleur):\n")
Trié selon AIC (plus petit = meilleur):
> cat("\nTrié selon BIC (plus petit = meilleur):\n")

Trié selon BIC (plus petit = meilleur):
> print(model_comparison[order(model_comparison$AIC), ])
      Model     AIC     BIC
1 GLM Gamma 20947.42 20997.70
2          RL 21267.11 21317.39
> |
```

### 3) Conclusion : Comparaison des résultats

Comparaison RLvs. GLM GammaRégression

- Linéaire (RL) :

Hypothèses violées :

Résidus non normaux ( $p\text{-value} < 2.2\text{e-}16$ )

hétéroscédasticité( $p\text{-value} = 3.14\text{e-}07$ ).

AIC = 21267.11, BIC = 21317.39.GLM

• Gamma :

Hypothèses respectées :

Distribution Gamma adaptée  
, hétéroscédasticitécorrigée.

Résidus : Centrés autour de 0 (moyenne = -0.047), tendance systématique mineure.

Pseudo- $R^2$  = 0.744 (74.4% de la déviance expliquée).

AIC = 20947.42, BIC = 20987.70 .

Conclusion : **GLM Gamma plus adapté** (meilleur ajustement, AIC et BICplusfaibles).

10

# Conclusion

---

## Conclusion finale :

Cette étude a visé à modéliser les coûts médicaux annuels à l'aide de techniques de régression pour prédire les dépenses en fonction des caractéristiques des patients (âge, sexe, IMC, tabagisme, nombre d'enfants, région). Grâce à une préparation rigoureuse des données et à une analyse approfondie, nous avons obtenu un ensemble de données fiable et comparé deux approches de modélisation : la régression linéaire (RL) et le modèle linéaire généralisé (GLM) avec distribution Gamma.

## Conclusion finale :

En adoptant le modèle GLM Gamma, nous pouvons prédire avec précision les dépenses médicales, offrant des perspectives concrètes pour une gestion efficace du système de santé. Des travaux futurs pourraient explorer d'autres modèles ou intégrer des variables supplémentaires pour améliorer encore la précision des prédictions.