Modélisation des risques d'AVC avec la GEV

Adama Sall

2025-06-19

```
library(evd)

library(readxl)
library(dplyr)

library(tidyverse)

library(extRemes)

library(ggplot2)
library(pROC)

library(bgeva)

library(lmom)

library(ismev)

library(gridExtra)

df <- read.csv("C:/Users/Hp/Desktop/formation R/Données/base_impropre.csv")
head(df)
```

```
##   Genre Age Hypertension Maladie_Cardiaque Situation_Matrimoniale Type_tra
vail
## 1     1  67            0                 1                       1
2
## 2     0  61            0                 0                       1
3
## 3     1  80            0                 1                       1
2
## 4     0  49            0                 0                       1
2
## 5     0  79            1                 0                       1
3
## 6     1  81            0                 0                       1
2
##   Residence Taux_glucose_moyen      IMC Statut_Fumer AVC
## 1         1              228.69 36.60000            0   1
## 2         0              202.21 28.89324            1   1
## 3         0              105.92 32.50000            1   1
## 4         1              171.23 34.40000            2   1
## 5         0              174.12 24.00000            1   1
## 6         1              186.21 29.00000            0   1
```

```r
attach(df)

summary(df)
```

```
##      Genre             Age         Hypertension     Maladie_Cardiaque
##  Min.   :0.0000   Min.   : 0.08   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:25.00   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.0000   Median :45.00   Median :0.00000   Median :0.00000
##  Mean   :0.4143   Mean   :43.23   Mean   :0.09746   Mean   :0.05401
##  3rd Qu.:1.0000   3rd Qu.:61.00   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :2.0000   Max.   :82.00   Max.   :1.00000   Max.   :1.00000
##  Situation_Matrimoniale  Type_travail     Residence     Taux_glucose_moyen
##  Min.   :0.0000         Min.   :0.000   Min.   :0.000   Min.   : 55.12
##  1st Qu.:0.0000         1st Qu.:2.000   1st Qu.:0.000   1st Qu.: 77.25
##  Median :1.0000         Median :2.000   Median :1.000   Median : 91.89
##  Mean   :0.6562         Mean   :2.168   Mean   :0.508   Mean   :106.15
##  3rd Qu.:1.0000         3rd Qu.:3.000   3rd Qu.:1.000   3rd Qu.:114.09
##  Max.   :1.0000         Max.   :4.000   Max.   :1.000   Max.   :271.74
##       IMC         Statut_Fumer         AVC
##  Min.   :10.30   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:23.80   1st Qu.:1.0000   1st Qu.:0.00000
##  Median :28.40   Median :1.0000   Median :0.00000
##  Mean   :28.89   Mean   :0.9812   Mean   :0.04873
##  3rd Qu.:32.80   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :97.60   Max.   :2.0000   Max.   :1.00000
```

Modele complet

```r
# Définition de la régression GEV
model_gev <- bgeva(AVC ~ Age + Taux_glucose_moyen + IMC + Genre +
                   Hypertension + Maladie_Cardiaque + Situation_Matrimoniale
+
                   Type_travail + Residence + Statut_Fumer,
                 data = df)

# Résumé du modèle estimé
summary(model_gev)
```

```
##
## Family: BGEVA
## Equation: AVC ~ Age + Taux_glucose_moyen + IMC + Genre + Hypertension +
##     Maladie_Cardiaque + Situation_Matrimoniale + Type_travail +
##     Residence + Statut_Fumer
##
## Parametric coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -3.0879428  0.2331480 -13.245  < 2e-16 ***
## Age                     0.0291518  0.0022475  12.971  < 2e-16 ***
## Taux_glucose_moyen      0.0019601  0.0005923   3.310 0.000935 ***
## IMC                    -0.0003419  0.0050096  -0.068 0.945583
## Genre                  -0.0005547  0.0650015  -0.009 0.993192
```

```
## Hypertension            0.2016873  0.0818150   2.465 0.013695 *
## Maladie_Cardiaque       0.1872828  0.1001375   1.870 0.061448 .
## Situation_Matrimoniale -0.1242641  0.0989373  -1.256 0.209120
## Type_travail            0.0072606  0.0337100   0.215 0.829467
## Residence               0.0471372  0.0635211   0.742 0.458044
## Statut_Fumer            0.0026796  0.0500565   0.054 0.957308
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## n = 5110  tau = -0.25  total edf = 11
```

Modele réduit (Backward)

```
Modele_final <- bgeva(AVC ~ Age + Taux_glucose_moyen  +
                    Hypertension + Maladie_Cardiaque ,
                  data = df)

# Résumé du modèle
summary(Modele_final)
```

```
##
## Family: BGEVA
## Equation: AVC ~ Age + Taux_glucose_moyen + Hypertension + Maladie_Cardiaqu
e
##
## Parametric coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -3.1188365  0.1418626 -21.985  < 2e-16 ***
## Age                 0.0284901  0.0021529  13.233  < 2e-16 ***
## Taux_glucose_moyen  0.0018983  0.0005759   3.296 0.000981 ***
## Hypertension        0.2024592  0.0812001   2.493 0.012655 *
## Maladie_Cardiaque   0.1950544  0.0988397   1.973 0.048445 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## n = 5110  tau = -0.25  total edf = 5
```

Chaque année supplémentaire augmente significativement le risque d'AVC. L'effet est fort et très significatif

Une augmentation du taux de glucose est associée à une augmentation du risque d'AVC, avec un effet plus modéré mais significatif.

Les patients hypertendus ont un risque d'AVC plus élevé, significatif au seuil 5%

La présence d'une maladie cardiaque augmente aussi le risque d'AVC, avec une signification statistique juste au seuil 5%.

Calcul de probabilité

```r
# Fonction  GEV
pgev <- function(x, mu = 0, sigma = 1, tau = -0.25) {
  z <- (x - mu) / sigma
  if (tau == 0) {
    p <- exp(-exp(-z))
  } else {
    t <- 1 + tau * z
    if (any(t <= 0)) stop("Argument hors domaine de la GEV")
    p <- exp(-t^(-1/tau))
  }
  return(p)
}


tau <- Modele_final$tau

# coefficients
coef <- c(
  Intercept      = -3.1188365,
  Age = 0.0284901,
  Taux_glucose_moyen  = 0.0018983,
  Hypertension   = 0.2024592,
  Maladie_Cardiaque = 0.1950544
)


# Calcul du prédicteur
eta <- with(df,
  coef["Intercept"] +
  coef["Age"] * Age +
  coef["Taux_glucose_moyen"] * Taux_glucose_moyen +
  coef["Hypertension"] * Hypertension +
  coef["Maladie_Cardiaque"] * Maladie_Cardiaque
)

# Calcul des probabilités :
prob_pred <- 1 - pgev(-eta, mu = 0, sigma = 1, tau = tau)

# les probabilités
head(prob_pred)
```

```
## [1] 0.4136763 0.2721336 0.4647141 0.1639927 0.5057362 0.4593958
```

```r
exp(coef(Modele_final))
```

```
##        (Intercept)                Age Taux_glucose_moyen       Hypertensio
n
##         0.04420858         1.02889981         1.00190009         1.2244101
```

```
3
##  Maladie_Cardiaque
##          1.21537706
```

Chaque année supplémentaire augmente le risque d'AVC d'environ 2,9 %

Chaque unité supplémentaire de glucose augmente le risque d'AVC d'environ 0,19 %

Une personne hypertendue a 1,22 fois plus de risque d'avoir un AVC qu'une personne non hypertendue

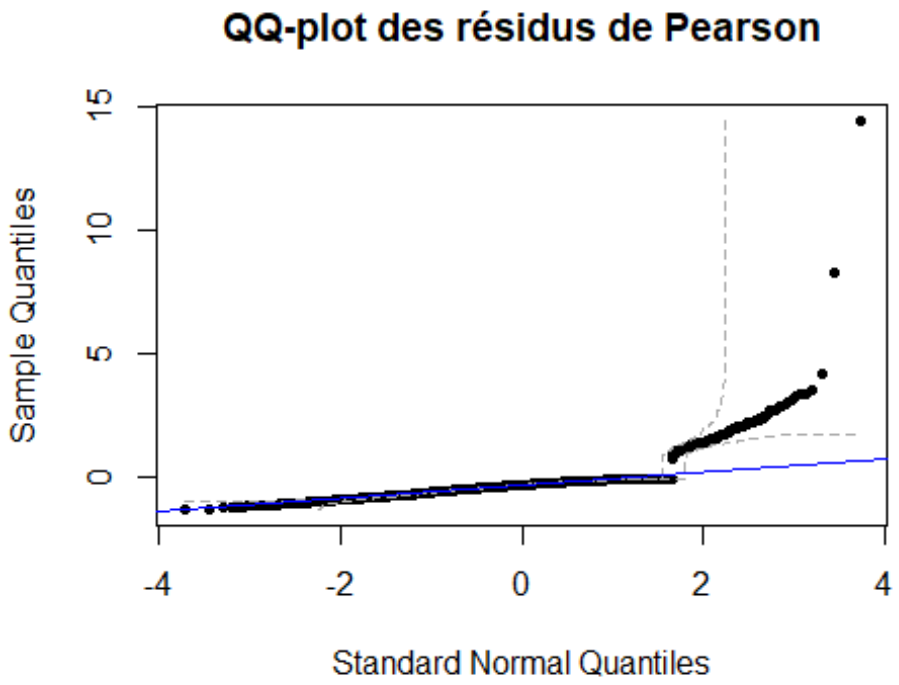
Une personne cardiaque a 1,21 fois plus de risque d'avoir un AVC qu'une personne non cardiaque

```
# Ajouter à ton dataframe
df$prob_pred <- prob_pred
head(df)
```

```
##   Genre Age Hypertension Maladie_Cardiaque Situation_Matrimoniale Type_tra
vail
## 1     1  67            0                 1                       1
2
## 2     0  61            0                 0                       1
3
## 3     1  80            0                 1                       1
2
## 4     0  49            0                 0                       1
2
## 5     0  79            1                 0                       1
3
## 6     1  81            0                 0                       1
2
##   Residence Taux_glucose_moyen      IMC Statut_Fumer AVC prob_pred
## 1         1             228.69 36.60000            0   1 0.4136763
## 2         0             202.21 28.89324            1   1 0.2721336
## 3         0             105.92 32.50000            1   1 0.4647141
## 4         1             171.23 34.40000            2   1 0.1639927
## 5         0             174.12 24.00000            1   1 0.5057362
## 6         1             186.21 29.00000            0   1 0.4593958
```

Qualité d'ajustement

```
# résidus de Pearson
residus_pearson <- (df$AVC - prob_pred) / sqrt(prob_pred * (1 - prob_pred))

# QQ-plot
qqnorm(residus_pearson, main = "QQ-plot des résidus de Pearson")
qqline(residus_pearson, col = "blue")
```



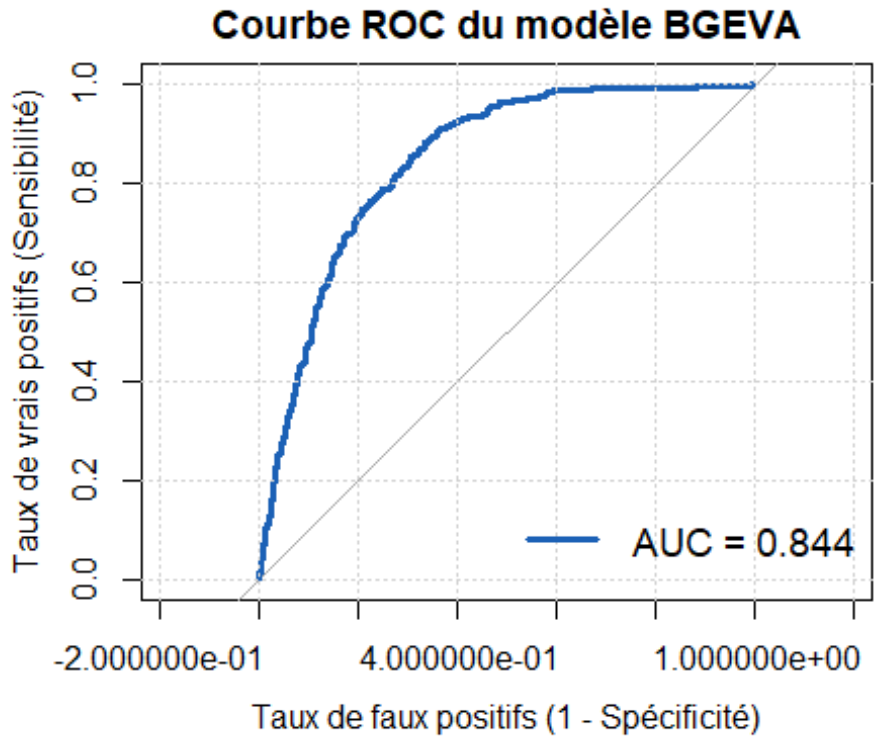
AIC et BIC du modele

```
logLik_val <- sum(AVC * log(prob_pred) + (1 - AVC) * log(1 - prob_pred))

k <- length(coef(Modele_final))
n <- length(AVC)

AIC_val <- -2 * logLik_val + 2 * k
BIC_val <- -2 * logLik_val + log(n) * k

cat("AIC =", AIC_val, "\n")

## AIC = 2272.877

cat("BIC =", BIC_val, "\n")

## BIC = 2305.572
```

la courbe ROC

```
roc <- roc(df$AVC, prob_pred)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

# graphe
plot(roc,
```

```
     col = "#1c61b6",
     lwd = 3,
     main = "Courbe ROC du modèle BGEVA",
     xlab = "Taux de faux positifs (1 - Spécificité)",
     ylab = "Taux de vrais positifs (Sensibilité)",
     legacy.axes = TRUE)

grid()
# Calcul de l'AUC
auc <- auc(roc)
legend("bottomright", legend = paste("AUC =", round(auc, 3)),
       col = "#1c61b6", lwd = 3, bty = "n", cex = 1.2)
```



L'aire sous la courbe AUC = 0.844, le modele a donc une très bonne capacité de discrimination.

Seuil optimal de probabilité qui equilibe la sensibilté et la spécificité

Methode de Youden index

Cette fonction cherche le point optimal sur la courbe ROC qui maximise l'indice de Youden (Youden index)

```r
# Obtenir le seuil optimal selon Youden index
opt <- coords(roc, "best", best.method = "youden", transpose = FALSE)
opt
```

```
##   threshold specificity sensitivity
## 1 0.1884308    0.693273   0.8514056
```

```r
# le seuil optimal
seuil_optimal <- opt["threshold"]

# seuil_optimal
seuil <- seuil_optimal[[1]]

# Affichage
cat("Seuil optimal pour la classification :", seuil, "\n")
```

```
## Seuil optimal pour la classification : 0.1884308
```

Seuil optimal (threshold) : 0.1884

Spécificité au seuil optimal : 0.6933 (69,3 %)

Sensibilité au seuil optimal : 0.8514 (85,1 %)

Interprétation

Le seuil optimal correspond au point sur la courbe ROC qui maximise la somme Sensibilité + Spécificité - 1.

Le seuil de 0.1884 signifie que si la probabilité prédite par le modèle est supérieure à 18,84 %, on classera l'observation comme positive (présence d'AVC). À ce seuil, le modèle détecte correctement environ 85 % des AVC (sensibilité élevée) tout en maintenant une spécificité correcte (69 %).

Taux de bon est de mauvais classement

```r
seuil= 0.188
# Classification binaire selon le seuil
pred_class <- ifelse(prob_pred >= seuil, 1, 0)

# Matrice de confusion
confusion <- table(Observé = AVC, Prédit = pred_class)
print(confusion)
```

```
##         Prédit
## Observé    0    1
##       0 3365 1496
##       1   37  212
```

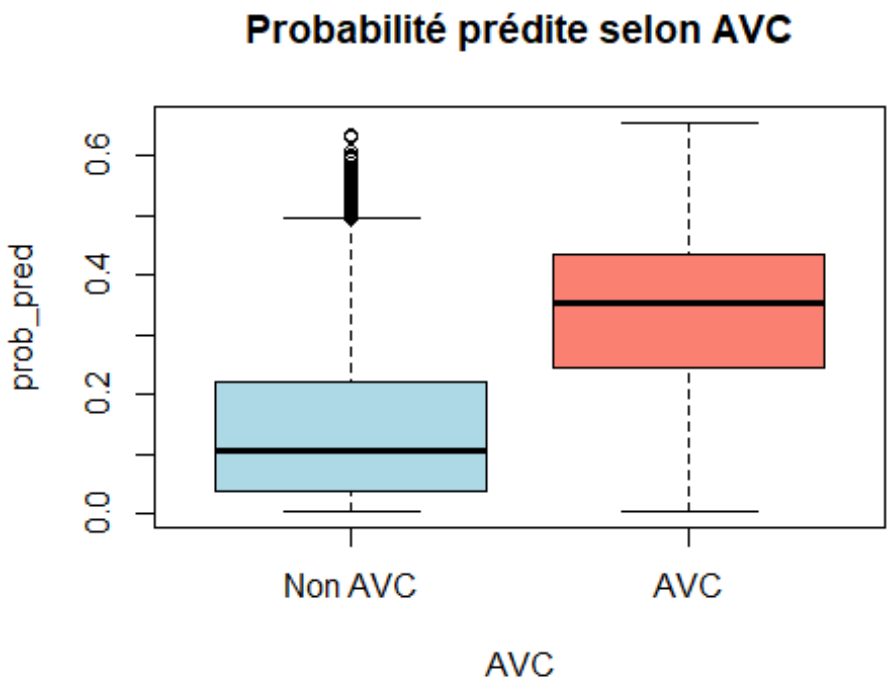
```
# Nombre de bonnes classifications
bon_classement <- sum(diag(confusion))

# Nombre total d'observations
n <- length(AVC)

# Taux de bon classement
taux_bon_classement <- bon_classement / n

# Taux de mauvais classement
taux_mauvais_classement <- 1 - taux_bon_classement

cat("Taux de bon classement :", round(taux_bon_classement * 100, 2), "%\n")
```

```
## Taux de bon classement : 70 %
```

```
cat("Taux de mauvais classement :", round(taux_mauvais_classement * 100, 2),
"%\n")
```

```
## Taux de mauvais classement : 30 %
```

```
boxplot(prob_pred ~ AVC, data = df,
        names = c("Non AVC", "AVC"),
        col = c("lightblue", "salmon"),
        main = "Probabilité prédite selon AVC")
```



Classement des patients

```r
df$classe <- ifelse(prob_pred >= seuil, "risque", "pas de risque")

# Afficher le résumé
table(df$classe)
```

```
##
## pas de risque        risque
##          3402          1708
```

Etude sur les patients à risque

l'idée est d' extraire les patients classés à risque et sélectionner les colonnes numeriques.

```r
# Extraction des individus classés "risque"
patients_a_risque <- subset(df, classe == "risque")
patients_a_risque$prob_patien_risque <- prob_pred[df$classe == "risque"]


# Afficher le nombre de patients à risque
cat("Nombre de patients à risque :", nrow(patients_a_risque), "\n")
```

```
## Nombre de patients à risque : 1708

# Sélection des colonnes IMC, Taux_glucose_moyen et Age
df_selection <- select(patients_a_risque, IMC, Taux_glucose_moyen, Age, AVC,
prob_patien_risque)

head(df_selection)

##        IMC Taux_glucose_moyen Age AVC prob_patien_risque
## 1 36.60000              228.69  67   1          0.4136763
## 2 28.89324              202.21  61   1          0.2721336
## 3 32.50000              105.92  80   1          0.4647141
## 5 24.00000              174.12  79   1          0.5057362
## 6 29.00000              186.21  81   1          0.4593958
## 7 27.40000               70.09  74   1          0.4510198

# dimension
dim(df_selection)

## [1] 1708    5

summary(df_selection)

##       IMC        Taux_glucose_moyen      Age             AVC
##  Min.   :11.30   Min.   : 55.23     Min.   :42.00   Min.   :0.0000
##  1st Qu.:26.70   1st Qu.: 81.51     1st Qu.:61.00   1st Qu.:0.0000
##  Median :29.20   Median :101.39     Median :68.00   Median :0.0000
##  Mean   :30.47   Mean   :127.91     Mean   :67.89   Mean   :0.1241
##  3rd Qu.:33.70   3rd Qu.:191.70     3rd Qu.:76.00   3rd Qu.:0.0000
##  Max.   :60.90   Max.   :271.74     Max.   :82.00   Max.   :1.0000
##  prob_patien_risque
##  Min.   :0.1881
##  1st Qu.:0.2359
##  Median :0.3074
##  Mean   :0.3209
##  3rd Qu.:0.3862
##  Max.   :0.6550
```

Analyse

IMC : La moyenne est autour de 30, ce qui correspond à la limite supérieure du surpoids et le début de l'obésité. L'IMC maximal atteint 60.9, indiquant des cas d'obésité sévère

Taux de glucose : La moyenne élevée (127.96 mg/dL) et le 3e quartile très élevé (191.74) indiquent une hyperglycémie fréquente chez ces patients à risque.

Âge : Les patients à risque sont majoritairement âgés (médiane 68 ans)

Les probabilités prédites à risque, avec une moyenne autour de 32 % et un maximum à 65.5 %.

ajustement de la loi GEV sur l'âge des patients à risque, selon les trois méthodes : MLE, L-moments et Moindres carrés (OLS)

```r
age_risque <- patients_a_risque$Age
n <- length(age_risque)

# # Méthode des maximum de vraisemblance
model_mle <- fgev(age_risque)
params_mle <- model_mle$estimate
loglik_mle <- sum(dgev(age_risque, loc = params_mle["loc"],
                       scale = params_mle["scale"],
                       shape = params_mle["shape"], log = TRUE))
aic_mle <- -2 * loglik_mle + 2 * 3

cat("Méthode : Maximum de vraisemblance (MLE)\n")
```

```
## Méthode : Maximum de vraisemblance (MLE)
```

```r
print(params_mle)
```

```
##        loc      scale      shape
## 65.9557223  9.8738854 -0.5545276
```

```r
cat("AIC (MLE) :", round(aic_mle, 2), "\n\n")
```

```
## AIC (MLE) : 12187.09
```

```r
# Méthode des L-moments

lmom_age <- samlmu(age_risque)
model_lmom <- pelgev(lmom_age)
params_lmom <- c(loc =64.9352581 ,
                 scale =9.2130145 ,
                 shape =0.3266781 )

loglik_lmom <- sum(dgev(age_risque, loc = params_lmom["loc"],
                        scale = params_lmom["scale"],
                        shape = params_lmom["shape"], log = TRUE))
aic_lmom <- -2 * loglik_lmom + 2 * 3

cat("Méthode : L-moments\n")
```

```
## Méthode : L-moments
```

```r
print(params_lmom)
```

```
##        loc      scale      shape
## 64.9352581  9.2130145  0.3266781
```

```r
cat("AIC (L-moments) :", round(aic_lmom, 2), "\n\n")
```

```
## AIC (L-moments) : 13756.3

# Méthode des Moindres carrés

gev_ols <- function(par, data) {
  mu <- par[1]; sigma <- par[2]; xi <- par[3]
  if (sigma <= 0) return(Inf)
  n <- length(data)
  p <- (1:n - 0.5)/n
  y <- sort(data)
  theo <- qgev(p, loc = mu, scale = sigma, shape = xi)
  sum((y - theo)^2)
}
start <- c(mean(age_risque), sd(age_risque), 0.1)
model_ols <- optim(start, gev_ols, data = age_risque, method = "BFGS")
params_ols <- model_ols$par

loglik_ols <- sum(dgev(age_risque, loc = params_ols[1],
                       scale = params_ols[2],
                       shape = params_ols[3], log = TRUE))
aic_ols <- -2 * loglik_ols + 2 * 3

cat("Méthode : Moindres carrés (OLS)\n")

## Méthode : Moindres carrés (OLS)

names(params_ols) <- c("loc", "scale", "shape")
print(params_ols)

##        loc      scale      shape
## 65.0980828  9.1305136 -0.3590795

cat("AIC (OLS) :", round(aic_ols, 2), "\n\n")

## AIC (OLS) : 12238.21
```

Le modèle avec la methode de MLE donne le meilleur ajustement (plus bas AIC). Le parametre de forme varie :

MLE : shape < 0 indique une queue bornée (type Weibull).

L-moments : shape > 0 → queue lourde (type Fréchet).

OLS : shape < 0 → bornée aussi, mais moins prononcée que MLE.

Le paramètre scale σ est assez stable, autour de 9.

Nous avons comparé trois méthodes d'estimation des paramètres appliquée à l'âge des patients à risque d'AVC. Les résultats montrent que la méthode du maximum de vraisemblance fournit l'ajustement le plus performant, avec un AIC plus faible de 12141.55 contre 13391.93 pour les L-moments. Le paramètre de forme est négatif pour MLE et OLS,

ce qui indique que les âges extrêmes sont bornés, c'est-à-dire qu'il existe une limite supérieure au risque lié à l'âge.

```r
# Variable : taux de glucose moyen des patients à risque
glucose_risque <- patients_a_risque$Taux_glucose_moyen
n <- length(glucose_risque)

# Méthode des  MLE
model_mle_glucose <- fgev(glucose_risque)
params_mle_glucose <- model_mle_glucose$estimate
loglik_mle_glucose <- sum(dgev(glucose_risque, loc = params_mle_glucose["loc"
],
                               scale = params_mle_glucose["scale"],
                               shape = params_mle_glucose["shape"], log = TRU
E))
aic_mle_glucose <- -2 * loglik_mle_glucose + 2 * 3

cat("Méthode : Maximum de vraisemblance (MLE)\n")

## Méthode : Maximum de vraisemblance (MLE)

print(params_mle_glucose)

##        loc      scale      shape
## 92.3277201 33.5450822  0.4232177

cat("AIC (MLE) :", round(aic_mle_glucose, 2), "\n\n")

## AIC (MLE) : 18204.66

# Méthode des L-moments
lmom_glucose <- samlmu(glucose_risque)
model_lmom_glucose <- pelgev(lmom_glucose)

# Vérification de la validité de l'estimation
params_lmom_glucose <- c(loc   = 98.96124223,
                         scale = 42.68533986,
                         shape = 0.09414472 )

loglik_lmom_glucose <- sum(dgev(glucose_risque, loc = params_lmom_glucose["lo
c"],
                                scale = params_lmom_glucose["scale"],
                                shape = params_lmom_glucose["shape"], log =
TRUE))
aic_lmom_glucose <- -2 * loglik_lmom_glucose + 2 * 3


cat("Méthode : L-moments\n")

## Méthode : L-moments
```

```r
print(params_lmom_glucose)
```

```
##         loc       scale       shape
## 98.96124223 42.68533986  0.09414472
```

```r
cat("AIC (L-moments) :", round(aic_lmom_glucose, 2), "\n\n")
```

```
## AIC (L-moments) : 18353.16
```

```r
# === 3. MOINDRES CARRES (OLS) ===
gev_ols <- function(par, data) {
  mu <- par[1]; sigma <- par[2]; xi <- par[3]
  if (sigma <= 0) return(Inf)
  n <- length(data)
  p <- (1:n - 0.5)/n
  y <- sort(data)
  theo <- qgev(p, loc = mu, scale = sigma, shape = xi)
  sum((y - theo)^2)
}
start <- c(mean(glucose_risque), sd(glucose_risque), 0.1)
model_ols_glucose <- optim(start, gev_ols, data = glucose_risque, method = "B
FGS")
params_ols_glucose <- model_ols_glucose$par

loglik_ols_glucose <- sum(dgev(glucose_risque, loc = params_ols_glucose[1],
                               scale = params_ols_glucose[2],
                               shape = params_ols_glucose[3], log = TRUE))
aic_ols_glucose <- -2 * loglik_ols_glucose + 2 * 3

cat("Méthode : Moindres carrés (OLS)\n")
```

```
## Méthode : Moindres carrés (OLS)
```

```r
names(params_ols_glucose) <- c("loc", "scale", "shape")
print(params_ols_glucose)
```

```
##        loc      scale      shape
## 104.038514  49.660983  -0.106817
```

```r
cat("AIC (OLS) :", round(aic_ols_glucose, 2), "\n\n")
```

```
## AIC (OLS) : 18533.2
```

Le modèle avec la methode de MLE donne le meilleur ajustement (plus bas AIC).Le parametre de forme varie :

MLE : shape > 0 Indique une distribution de Fréchet avec queue lourde à droite

L-moments : shape > 0 Indique une distribution de Fréchet avec queue lourde à droite

OLS : shape < 0 indique bornée .

L'ajustement de la loi GEV sur le taux de glucose moyen des patients à risque d'AVC a révélé une meilleure performance de la méthode du maximum de vraisemblance (AIC = 18154.42), comparée à celle des L-moments (AIC = 18300.95) et des moindres carrés (AIC = 18480). Le paramètre de forme positif suggère une distribution à queue lourde, confirmant la présence de valeurs extrêmes importantes de glycémie chez les individus à risque.

```
# Variable : IMC des patients à risque
imc_risque <- patients_a_risque$IMC
n <- length(imc_risque)

# MÉTHODE du  Maximum de vraisemblance (MLE)
model_mle_imc <- fgev(imc_risque)
params_mle_imc <- model_mle_imc$estimate
loglik_mle_imc <- sum(dgev(imc_risque,
                      loc = params_mle_imc["loc"],
                      scale = params_mle_imc["scale"],
                      shape = params_mle_imc["shape"],
                      log = TRUE))
aic_mle_imc <- -2 * loglik_mle_imc + 2 * 3

cat("Méthode : Maximum de vraisemblance (MLE)\n")

## Méthode : Maximum de vraisemblance (MLE)

print(params_mle_imc)

##         loc       scale       shape
## 27.87744931  5.30684234 -0.08284068

cat("AIC (MLE) :", round(aic_mle_imc, 2), "\n\n")

## AIC (MLE) : 10898.16

# MÉTHODE des L-moments
lmom_imc <- samlmu(imc_risque)
model_lmom_imc <- pelgev(lmom_imc)
params_lmom_imc <- c(loc = 27.82420228,
                 scale = 4.97604918,
                 shape = 0.04494865
                 )

loglik_lmom_imc <- sum(dgev(imc_risque,
                        loc = params_lmom_imc["loc"],
                        scale = params_lmom_imc["scale"],
                        shape = params_lmom_imc["shape"],
                        log = TRUE))
aic_lmom_imcc <- -2 * loglik_lmom_imc + 2 * 3
```

```r
cat("Méthode : L-moments (valeurs fixées)\n")
```

```
## Méthode : L-moments (valeurs fixées)
```

```r
print(params_lmom_imc)
```

```
##         loc       scale       shape
## 27.82420228  4.97604918  0.04494865
```

```r
cat("AIC (L-moments) :", round(aic_lmom_imcc, 2), "\n\n")
```

```
## AIC (L-moments) : 11026.73
```

```r
# MÉTHODE des  Moindres carrés (OLS)
gev_ols <- function(par, data) {
  mu <- par[1]; sigma <- par[2]; xi <- par[3]
  if (sigma <= 0) return(Inf)
  n <- length(data)
  p <- (1:n - 0.5)/n
  y <- sort(data)
  theo <- qgev(p, loc = mu, scale = sigma, shape = xi)
  sum((y - theo)^2)
}

start <- c(mean(imc_risque), sd(imc_risque), 0.1)
model_ols_imc <- optim(start, gev_ols, data = imc_risque, method = "BFGS")
params_ols_imc <- model_ols_imc$par
names(params_ols_imc) <- c("loc", "scale", "shape")

loglik_ols_imc <- sum(dgev(imc_risque,
                       loc = params_ols_imc["loc"],
                       scale = params_ols_imc["scale"],
                       shape = params_ols_imc["shape"],
                       log = TRUE))
aic_ols_imc <- -2 * loglik_ols_imc + 2 * 3

cat("Méthode : Moindres carrés (OLS)\n")
```

```
## Méthode : Moindres carrés (OLS)
```

```r
print(params_ols_imc)
```

```
##         loc       scale       shape
## 27.73474853  4.93113338 -0.02211008
```

```r
cat("AIC (OLS) :", round(aic_ols_imc, 2), "\n\n")
```

```
## AIC (OLS) : 10937.71
```

Le modèle avec la methode de MLE donne le meilleur ajustement (plus bas AIC).Le parametre de forme varie :

MLE : shape < 0 distribution de type Weibull (queue bornée à droite)

L-moments : shape > 0 indique queue lourde (type Fréchet).

OLS : shape < 0 indique une queue bornée.

L'ajustement de la loi GEV sur l'indice de masse corporelle (IMC) montre que la méthode du maximum de vraisemblance fournit le meilleur ajustement (AIC = 10865.69), légèrement devant celle des moindres carrés (AIC = 10905.15), tandis que la méthode des L-moments présente un ajustement moins optimal (AIC = 10992.62). Le paramètre de forme ξ est proche de zéro pour toutes les méthodes, indiquant une distribution proche de Gumbel, c'est-à-dire à queue exponentielle. Cela suggère que les IMC extrêmes sont possibles, mais pas excessivement fréquents ni illimités.

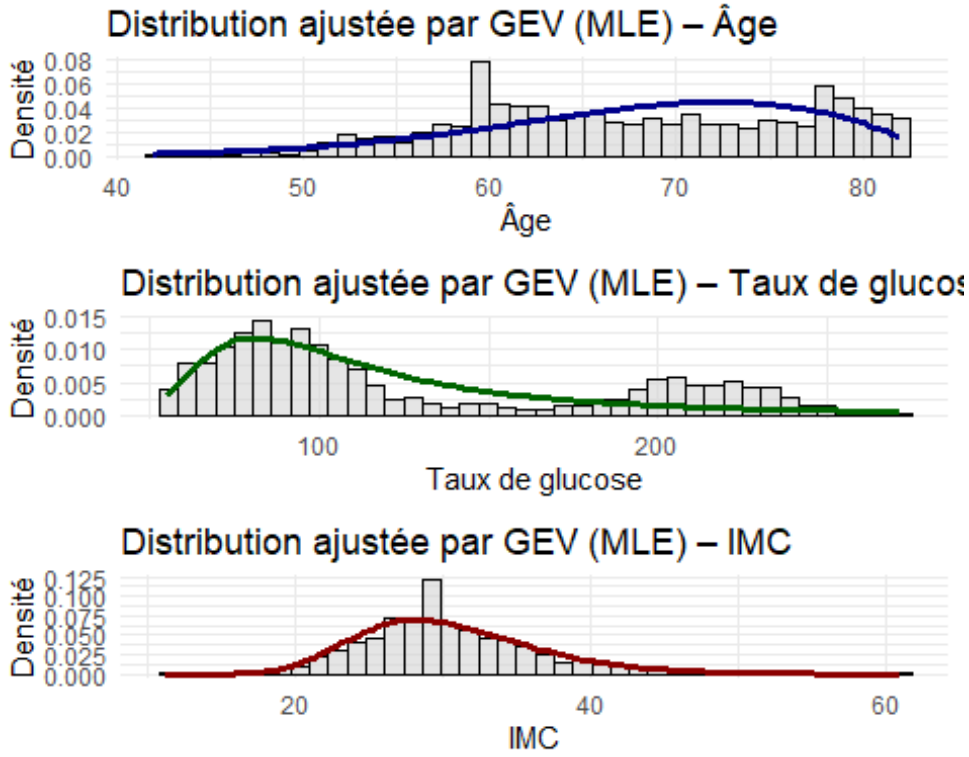
Visualisation des variables dans le cas des patients à risque

```
# Paramètres MLE déjà estimés
params_age     <- c(loc = 66.00, scale = 9.84, shape = -0.55)
params_glucose <- c(loc = 92.38, scale = 33.60, shape = 0.42)
params_imc     <- c(loc = 27.89, scale = 5.30, shape = -0.083)

# Créer une fonction pour tracer une distribution GEV ajustée
plot_gev_fit <- function(data, params, var_name, color = "steelblue") {
  df <- data.frame(x = data)
  ggplot(df, aes(x = x)) +
    geom_histogram(aes(y = ..density..), bins = 40,
                   fill = "lightgray", color = "black", alpha = 0.6) +
    stat_function(fun = function(x) dgev(x, loc = params["loc"],
                                         scale = params["scale"],
                                         shape = params["shape"]),
                  color = color, size = 1.2) +
    labs(title = paste("Distribution ajustée par GEV (MLE) –", var_name),
         x = var_name, y = "Densité") +
    theme_minimal()
}

# Tracés
p1 <- plot_gev_fit(patients_a_risque$Age, params_age, "Âge", color = "darkblu
e")
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

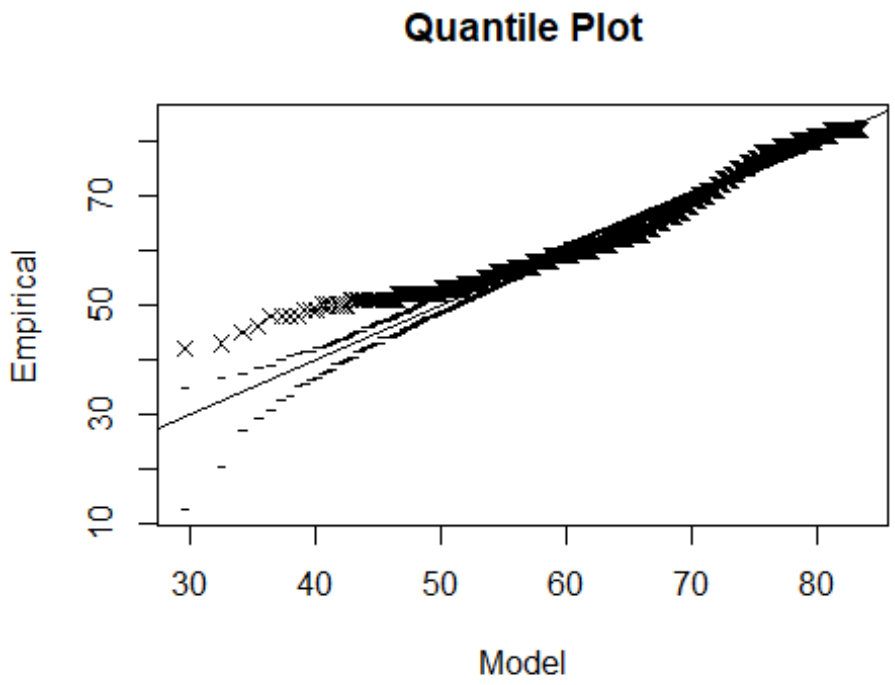
```r
p2 <- plot_gev_fit(patients_a_risque$Taux_glucose_moyen, params_glucose, "Taux de glucose", color = "darkgreen")
p3 <- plot_gev_fit(patients_a_risque$IMC, params_imc, "IMC", color = "darkred")

# Afficher les trois graphiques
grid.arrange(p1, p2, p3, nrow = 3)
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## ℹ Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

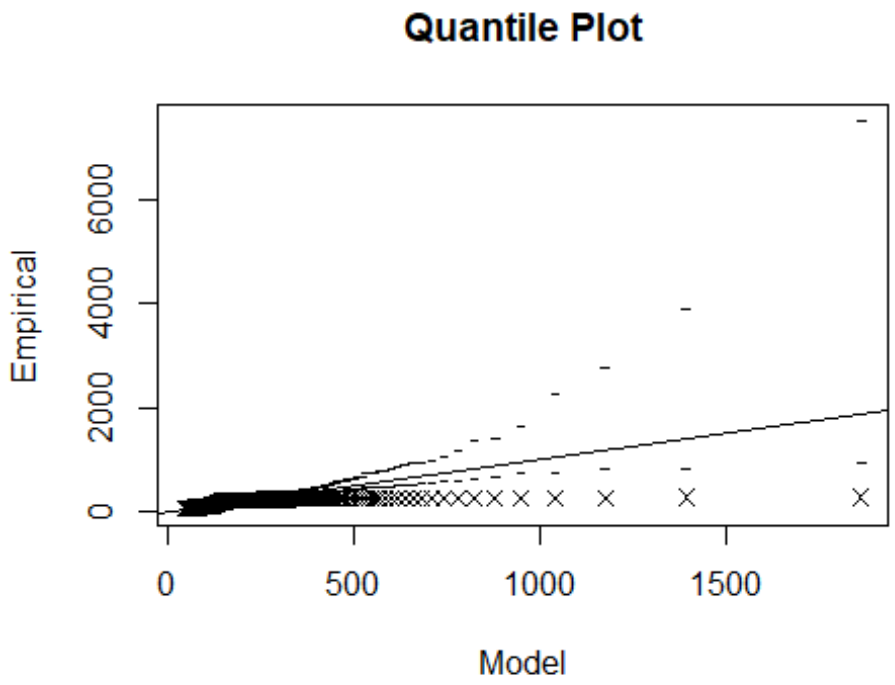


Qualité d'ajustement

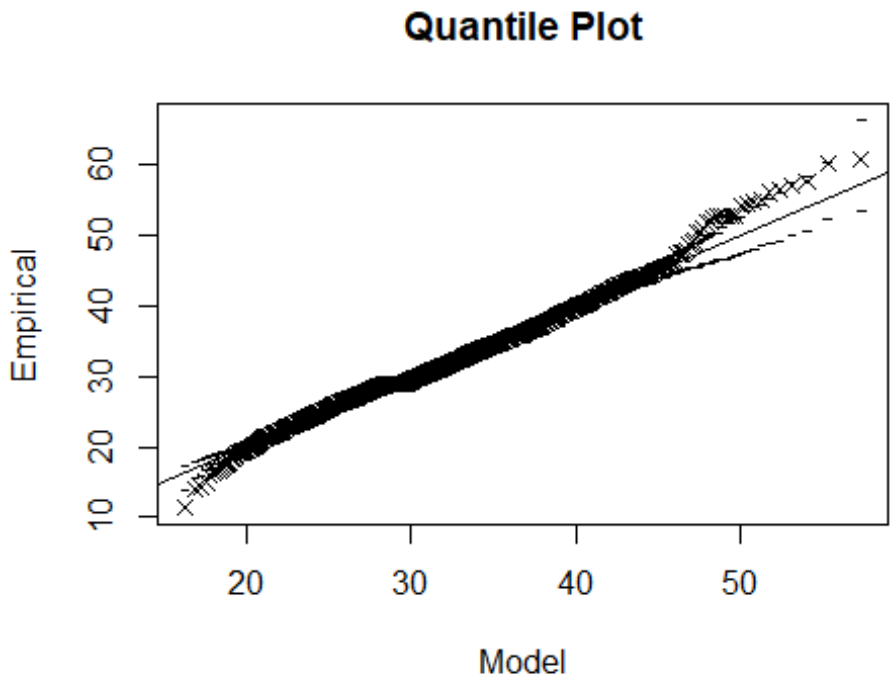
```r
# QQ-plot pour l'âge
plot(model_mle, which = 2)
```




```r
# Même principe pour le glucose
plot(model_mle_glucose, which = 2)
```



```r
# Et pour l'IMC

plot(model_mle_imc, which = 2)
```



Calcul des quantiles pour identifier les valeurs extrêmes dans chaque variable,(zones de risque)

```r
# Paramètres MLE estimés chez les patients à risque
params_age     <- c(loc = 66.00, scale = 9.84, shape = -0.55)
params_glucose <- c(loc = 92.38, scale = 33.60, shape = 0.42)
params_imc     <- c(loc = 27.89, scale = 5.30, shape = -0.083)

#Niveaux de quantiles
probs <- c(0.90, 0.95, 0.99)

# Calcul des quantiles
quant_age     <- qgev(probs, loc = params_age["loc"],     scale = params_age[
"scale"],     shape = params_age["shape"])
quant_glucose <- qgev(probs, loc = params_glucose["loc"], scale = params_gluc
ose["scale"], shape = params_glucose["shape"])
quant_imc     <- qgev(probs, loc = params_imc["loc"],     scale = params_imc[
"scale"],     shape = params_imc["shape"])

# Affichage
quant_table <- data.frame(
  Quantile = paste0(probs * 100, "%"),
```

```
  Age = round(quant_age, 2),
  Glucose = round(quant_glucose, 2),
  IMC = round(quant_imc, 2)
)

print(quant_table)
```

```
##   Quantile   Age Glucose   IMC
## 1      90% 78.70  218.24 38.77
## 2      95% 80.40  290.91 41.84
## 3      99% 82.47  564.68 48.16
```

Interprétation des résultats : Âge ≥ 80.4 ans (95e percentile) : les patients à partir de cet âge sont dans les 5% les plus âgés parmi ceux à risque ce qui implique un cas très critiques.

Glucose ≥ 290.9 mg/dL (95e percentile) : niveau très élevé, proche ou au-delà de seuils de diabète grave ce qui est une forte alerte métabolique.

IMC ≥ 41.8 (95e percentile) : correspond à un cas d'obésité sévère, peu fréquente mais associée à des comorbidités importantes.

Seuil optimal des variables Age, IMC, glycémie obtenus à partir de l'indice de Youden pour séparer les cas d'AVC des non-AVC

```
# ROC et seuil optimal pour Age
roc_age <- roc(df_selection$AVC, df_selection$Age)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
coords_age <- coords(roc_age, "best", best.method = "youden", transpose = FAL
SE)
seuil_age_opt <- coords_age$threshold


# ROC et seuil optimal pour Taux_glucose_moyen
roc_glucose <- roc(df_selection$AVC, df_selection$Taux_glucose_moyen)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

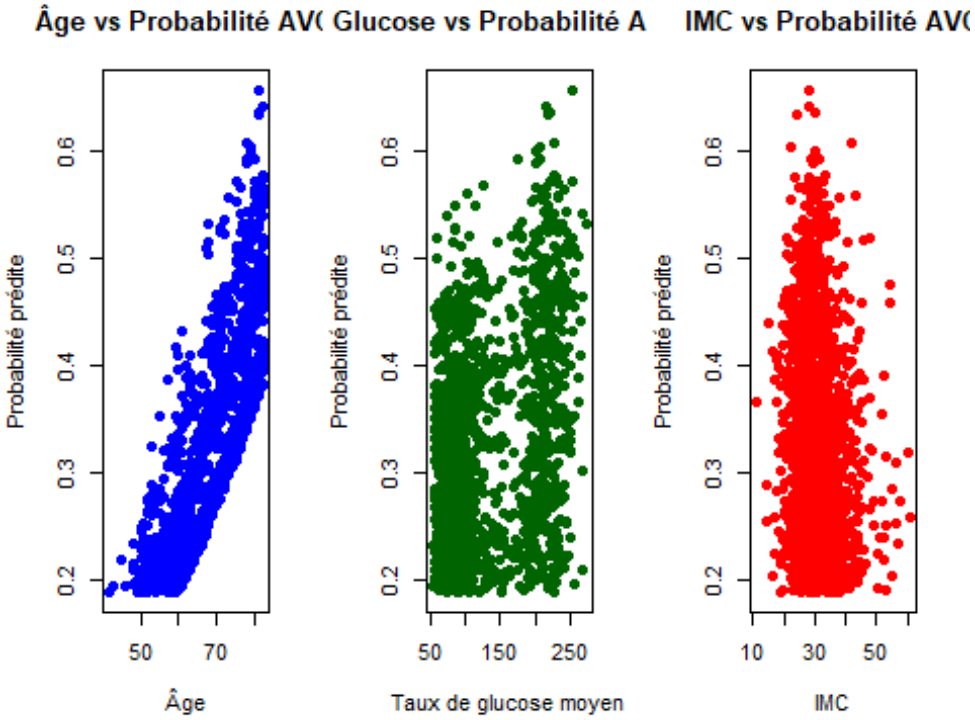
```
coords_glucose <- coords(roc_glucose, "best", best.method = "youden", transpo
se = FALSE)
seuil_glucose_opt <- coords_glucose$threshold


# ROC et seuil optimal pour IMC
roc_imc <- roc(df_selection$AVC, df_selection$IMC)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls > cases

coords_imc <- coords(roc_imc, "best", best.method = "youden", transpose = FAL
SE)
seuil_imc_opt <- coords_imc$threshold


# Affichage des seuils optimaux arrondis
cat("Seuil optimal pour l'âge :", round(seuil_age_opt, 1), "\n")

## Seuil optimal pour l'âge : 67.5

cat("Seuil optimal pour la glucose :", round(seuil_glucose_opt, 2), "\n")

## Seuil optimal pour la glucose : 104.46

cat("Seuil optimal pour l'IMC :", round(seuil_imc_opt, 2), "\n")

## Seuil optimal pour l'IMC : 28.9
```

tendance entre chaque variable continue et la probabilité prédite d'AVC.

```
par(mfrow = c(1, 3))

plot(patients_a_risque$Age, patients_a_risque$prob_pred,
     main = "Âge vs Probabilité AVC",
     xlab = "Âge", ylab = "Probabilité prédite", pch = 19, col = "blue")

plot(patients_a_risque$Taux_glucose_moyen, patients_a_risque$prob_pred,
     main = "Glucose vs Probabilité AVC",
     xlab = "Taux de glucose moyen", ylab = "Probabilité prédite", pch = 19,
col = "darkgreen")

plot(patients_a_risque$IMC, patients_a_risque$prob_pred,
     main = "IMC vs Probabilité AVC",
     xlab = "IMC", ylab = "Probabilité prédite", pch = 19, col = "red")
```



Âge vs Probabilité prédite d'AVC

Il existe une relation croissante entre l'âge et la probabilité prédite d'AVC, plus l'âge augmente, plus la probabilité prédite d'AVC est élevée

Taux de glucose moyen vs Probabilité prédite d'AVC

il y a une légère tendance, mais beaucoup de dispersion. On peut conclure que les patients ayant un taux de glucose élevé peuvent avoir une probabilité prédite d'AVC plus importante, mais ce n'est pas systématique.Donc le taux de glucose moyen contribue au risque, mais son effet est moins marqué que celui de l'âge

IMC vs Probabilité prédite d'AVC

la probabilité prédite d'AVC est répartie de façon homogène pour toutes les valeurs d'IMC.On observe aussi une grande dispersion verticale, quelle que soit la valeur de l'IMC. Alors on peut conclure que l'IMC n'a pas d'effet linéaire évident la probabilité prédite d'AVC dans ce sous-groupe de patients à risque.

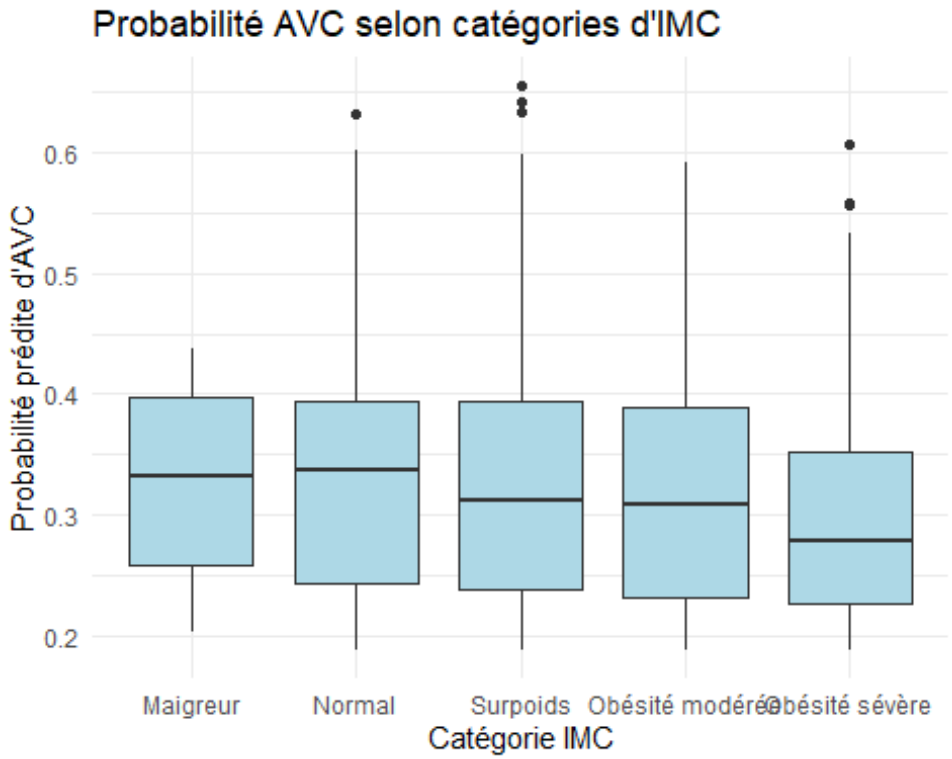
Decomposition en categorie

```
# IMC : catégories
patients_a_risque$IMC_cat <- cut(df_selection$IMC,
                  breaks = c(-Inf, 18.5, 25, 30, 35, Inf),
                  labels = c("Maigreur", "Normal", "Surpoids", "Obésité modér
ée", "Obésité sévère"))
```

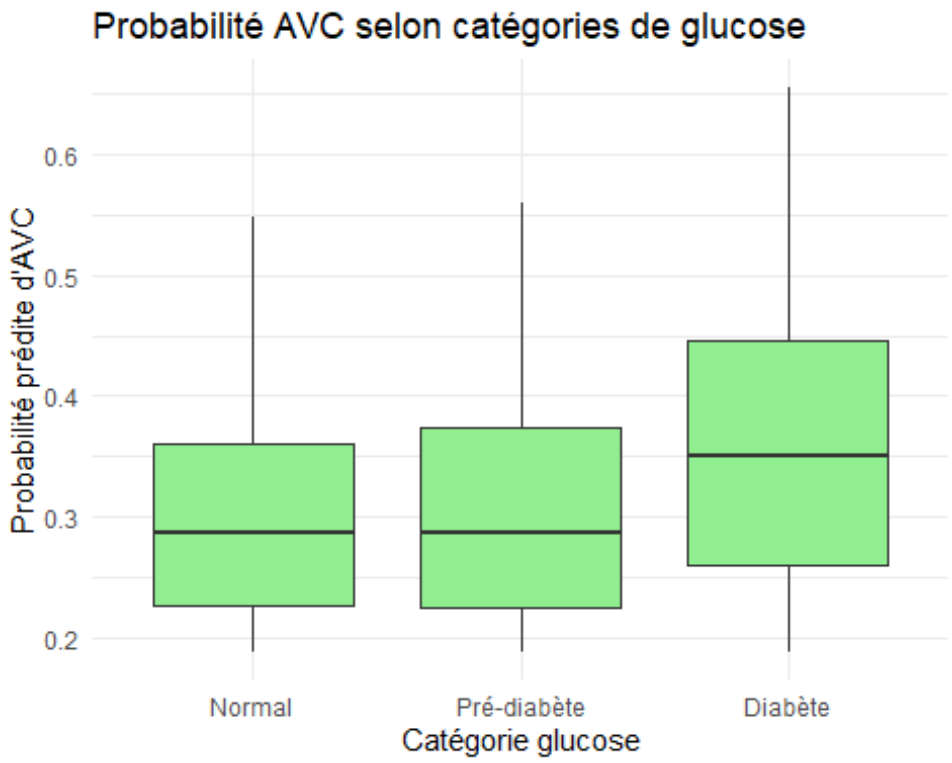
```r
# Taux_glucose_moyen : catégories basées sur seuils cliniques (en g/L)
patients_a_risque$Glucose_cat <- cut(df_selection$Taux_glucose_moyen,
                      breaks = c(-Inf, 100, 125, Inf),
                      labels = c("Normal", "Pré-diabète", "Diabète"))

# Age : catégories arbitraires (exemple)
patients_a_risque$Age_cat <- cut(df_selection$Age,
                  breaks = c(40, 65, 75, 80, Inf),
                  labels = c("40-65", "65-75", "75-80", "80+"))
```

Visualisation

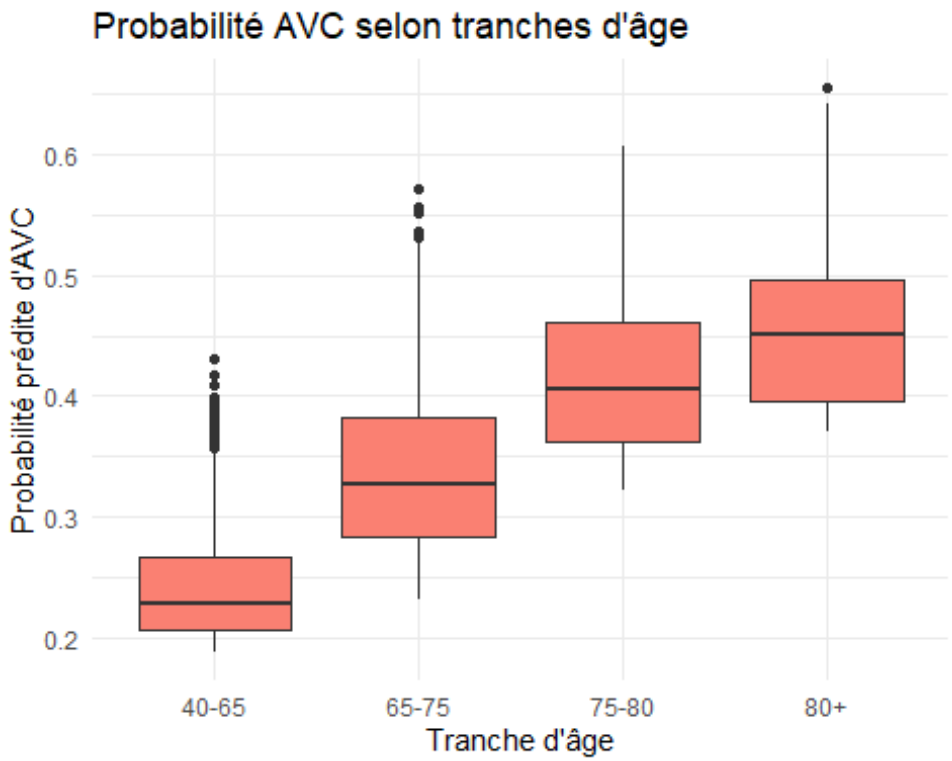
```r
# IMC
ggplot(patients_a_risque, aes(x = IMC_cat, y = prob_patien_risque)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Probabilité AVC selon catégories d'IMC",
       x = "Catégorie IMC", y = "Probabilité prédite d'AVC") +
  theme_minimal()
```




```r
# Taux de glucose
ggplot(patients_a_risque, aes(x = Glucose_cat, y = prob_patien_risque)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Probabilité AVC selon catégories de glucose",
       x = "Catégorie glucose", y = "Probabilité prédite d'AVC") +
  theme_minimal()
```




```r
# Age
ggplot(patients_a_risque, aes(x = Age_cat, y = prob_patien_risque)) +
  geom_boxplot(fill = "salmon") +
  labs(title = "Probabilité AVC selon tranches d'âge",
       x = "Tranche d'âge", y = "Probabilité prédite d'AVC") +
  theme_minimal()
```



Test de kruskall wall

```
# Test de Kruskal-Wallis pour IMC_cat
kruskal_imc <- kruskal.test(prob_patien_risque ~ IMC_cat, data = patients_a_r
isque)
cat("Test de Kruskal-Wallis pour IMC_cat :\n")

## Test de Kruskal-Wallis pour IMC_cat :

print(kruskal_imc)

##
##  Kruskal-Wallis rank sum test
##
## data:  prob_patien_risque by IMC_cat
## Kruskal-Wallis chi-squared = 22.886, df = 4, p-value = 0.0001334

# Test de Kruskal-Wallis pour Glucose_cat
kruskal_glucose <- kruskal.test(prob_patien_risque ~ Glucose_cat, data = pati
ents_a_risque)
cat("\nTest de Kruskal-Wallis pour Glucose_cat :\n")

##
## Test de Kruskal-Wallis pour Glucose_cat :

print(kruskal_glucose)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  prob_patien_risque by Glucose_cat
## Kruskal-Wallis chi-squared = 109.95, df = 2, p-value < 2.2e-16
```

```
# Test de Kruskal-Wallis pour Age_cat
kruskal_age <- kruskal.test(prob_patien_risque ~ Age_cat, data = patients_a_r
isque)
cat("\nTest de Kruskal-Wallis pour Age_cat :\n")
```

```
##
## Test de Kruskal-Wallis pour Age_cat :
```

```
print(kruskal_age)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  prob_patien_risque by Age_cat
## Kruskal-Wallis chi-squared = 1128.4, df = 3, p-value < 2.2e-16
```

Analyse

Le test de Kruskal-Wallis montre les p-values très faibles (< 0.05) signifient que l'hypothèse nulle selon laquelle laquelle les médianes des probabilités prédites sont égales dans tous les groupes est rejetée pour chaque variable.

On constate aussi sur les graphe que La probabilité d'AVC augmente clairement avec l'âge,

40-59 ans : La probabilité est basse, autour de 0.2.

60-69 ans : On observe une augmentation significative, avec une probabilité d'environ 0.3.

70-79 ans : La probabilité continue de croître et atteint environ 0.5.

80 ans et plus : C'est la tranche d'âge avec le risque le plus élevé, dépassant légèrement 0.5

Taux de glucose

La probabilité d'AVC augmente avec le niveau de glucose, elle est plus basse chez les patients normals. On note une augmentation modérée par rapport à la normale chez les patients pré-diabetes. Cette probabilité devient plus élevée chez les patients diabétiques et atteint environ 0.35.

IMC

La médiane de la probabilité prédite d'AVC est presque similaire dans toutes les catégories, autour de 0.3 à 0.35.

On note la présence de quelques valeurs extrêmes (patients avec une probabilité prédite > 0.6) sont visibles dans chaque groupe, notamment en surpoids et obésité sévère.

la tendance de probabilité n'est pas clairement croissante l'IMC seul n'est pas un facteur discriminant du risque d'AVC,