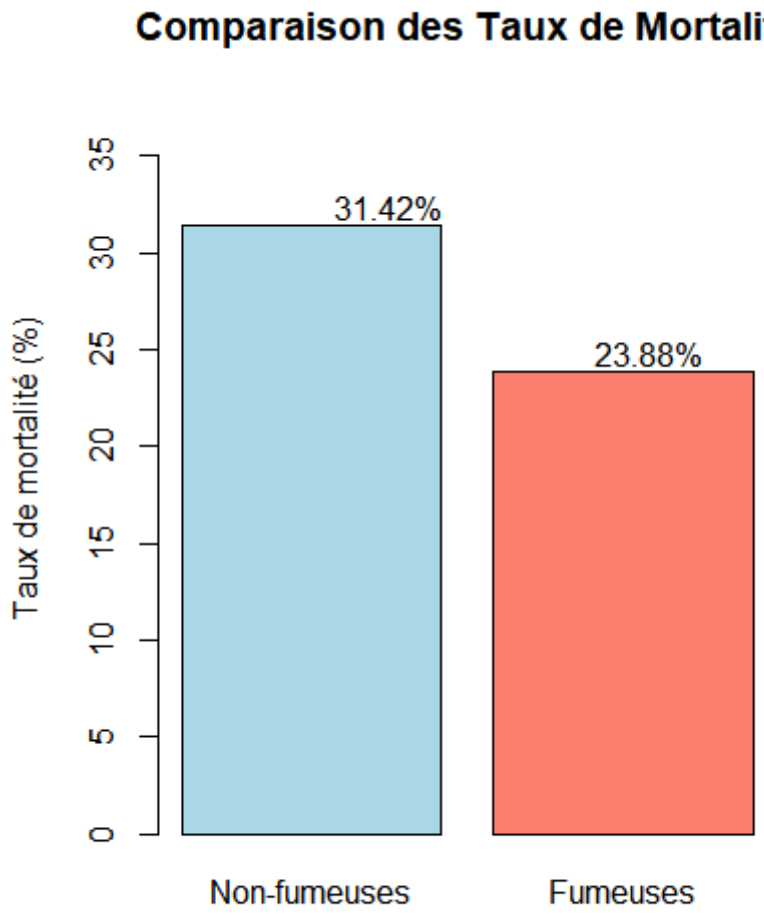
Cuvillier Axel

```
> # Spécifiez le chemin d'accès à votre fichier
> chemin_fichier <- "C:\\Users\\axelc\\Downloads\\Subject6_smoking (1).csv"
>
> # Chargez les données
> data <- read.csv(chemin_fichier, header = TRUE, sep = ",")
>
>
> # Création tableau de contingence brut ('Smoker' vs 'Status')
> T_comptages <- table(data$Smoker, data$Status)
>
> # Affichage du tableau avec les totaux (marges)
> addmargins(T_comptages)

      Alive Dead  Sum
  No    502  230  732
  Yes   443  139  582
  Sum   945  369 1314
>
> #Fréquence de pourcentage
> # Compilation en pourcentage par ligne (margin=1)
> T_pourcentages <- prop.table(T_comptages, margin = 1) * 100
>
> # Affichage des pourcentages arrondis
> round(T_pourcentages, 2)

      Alive  Dead
  No  68.58 31.42
  Yes 76.12 23.88
>
> # Calcul du taux de mortalité chez les non-fumeuses
> taux_non_fumeuses <- T_pourcentages["No", "Dead"]
>
> # Calcul du taux de mortalité chez les fumeuses
> taux_fumeuses <- T_pourcentages["Yes", "Dead"]
>
> # Affichage des résultats
> print(paste("Taux de mortalité chez les non-fumeuses :",
round(taux_non_fumeuses, 2), "%"))
[1] "Taux de mortalité chez les non-fumeuses : 31.42 %"
> print(paste("Taux de mortalité chez les fumeuses :", round(taux_fumeuses,
2), "%"))
[1] "Taux de mortalité chez les fumeuses : 23.88 %"
>
> #Suite Question 1
>
> # Récupération des taux de mortalité (colonne 'Dead' du tableau de
pourcentages)
> # Nous utilisons T_pourcentages calculé dans la Question 1
>
> # Taux de mortalité chez les non-fumeuses (ligne 'No', colonne 'Dead')
```

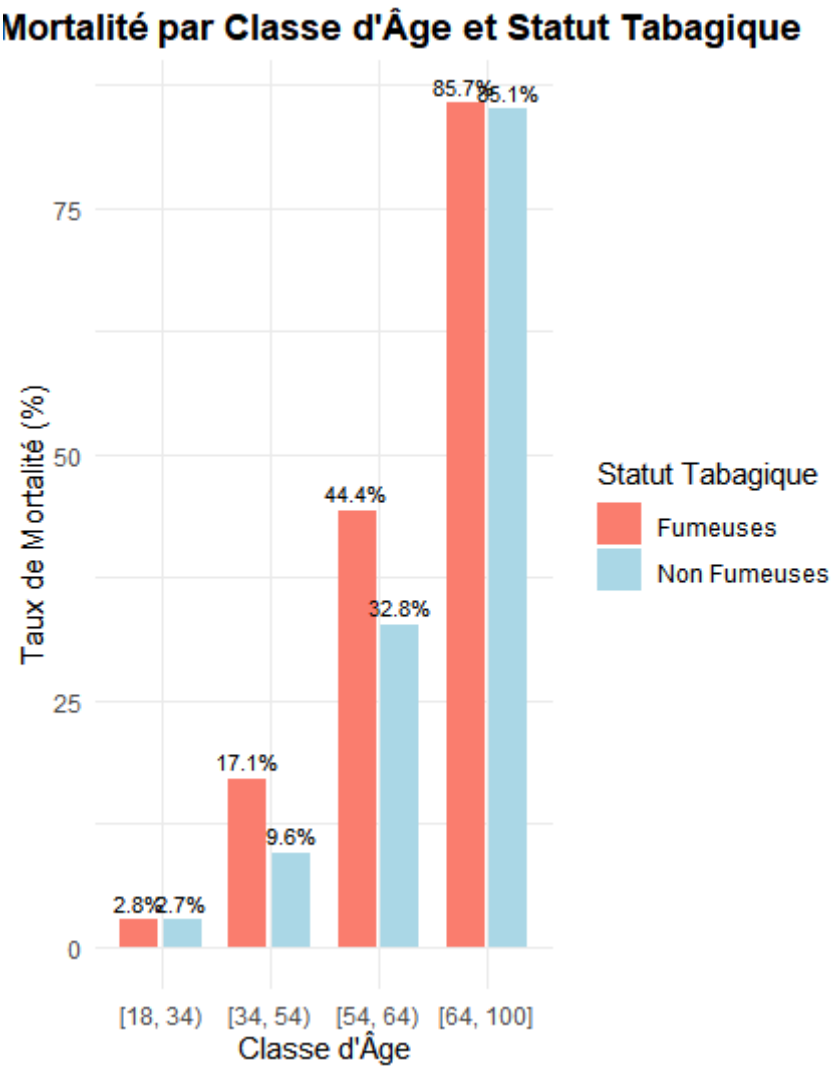
```
> taux_non_fumeurs <- T_pourcentages["No", "Dead"]
>
> # Taux de mortalité chez les fumeuses (ligne 'Yes', colonne 'Dead')
> taux_fumeurs <- T_pourcentages["Yes", "Dead"]
>
> print("--- Taux de Mortalité (en %) ---")
[1] "--- Taux de Mortalité (en %) ---"
> print(paste("Non-fumeuses :", round(taux_non_fumeurs, 2)))
[1] "Non-fumeuses : 31.42"
> print(paste("Fumeuses :", round(taux_fumeurs, 2)))
[1] "Fumeuses : 23.88"
>
>
> #diagramme en bar
> # Création d'un vecteur et nommage
> taux_vector <- c(taux_non_fumeurs, taux_fumeurs)
> names(taux_vector) <- c("Non-fumeuses", "Fumeuses")
>
> # Diagramme en barres
> barplot(taux_vector,
+         main = "Comparaison des Taux de Mortalité",
+         ylab = "Taux de mortalité (%)",
+         col = c("lightblue", "salmon"),
+         ylim = c(0, max(taux_vector) * 1.2) # Ajuste l'axe Y
+ )
> # Ajout des valeurs au-dessus des barres
> text(x = seq_along(taux_vector), y = taux_vector + 1,
+      labels = paste0(round(taux_vector, 2), "%"))
```




```
> #Question 2
> # Définition des bornes des classes d'âge
> bornes_age <- c(18, 34, 54, 64, 100)
>
> # --- Analyse pour les Fumeuses (Smoker == "Yes") ---
> df_fumeuses <- subset(data, Smoker == "Yes", select = c(Smoker, Status,
Age))
>
> # Création de la variable de classe d'âge pour les fumeuses
> ageclass_fumeuses <- cut(df_fumeuses$Age, bornes_age, right = FALSE,
include.lowest = TRUE)
>
> # Tableau de contingence brut par classe d'âge (Fumeuses)
> tab_fumeuses <- table(ageclass_fumeuses, df_fumeuses$Status)
>
> print("--- Effectifs Totaux (Fumeuses) ---")
[1] "--- Effectifs Totaux (Fumeuses) ---"
> addmargins(tab_fumeuses)

ageclass_fumeuses Alive Dead Sum
         [18,34)    174    5 179
         [34,54)    198   41 239
         [54,64)     64   51 115
         [64,100]     7   42  49
         Sum        443  139 582
```

```
>
> # Fréquences en pourcentage par ligne (taux) pour les fumeuses
> f_pourcentages <- prop.table(tab_fumeuses, margin = 1) * 100
>
> # --- Analyse pour les Non-Fumeuses (Smoker == "No") ---
> df_non_fumeuses <- subset(data, Smoker == "No", select = c(Smoker, Status,
Age))
>
> # Création de la variable de classe d'âge pour les non-fumeuses
> ageclass_non_fumeuses <- cut(df_non_fumeuses$Age, bornes_age, right =
FALSE, include.lowest = TRUE)
>
> # Tableau de contingence brut par classe d'âge (Non-Fumeuses)
> tab_non_fumeuses <- table(ageclass_non_fumeuses, df_non_fumeuses$Status)
>
> print("--- Effectifs Totaux (Non-Fumeuses) ---")
[1] "--- Effectifs Totaux (Non-Fumeuses) ---"
> addmargins(tab_non_fumeuses)

ageclass_non_fumeuses Alive Dead Sum
             [18,34)    213    6 219
             [34,54)    180   19 199
             [54,64)     80   39 119
             [64,100]    29  166 195
             Sum        502  230 732
>
> # Fréquences en pourcentage par ligne (taux) pour les non-fumeuses
> nf_pourcentages <- prop.table(tab_non_fumeuses, margin = 1) * 100
>
> # Création et affichage de la liste des tableaux de pourcentages
> maliste <- list(nf_pourcentages, f_pourcentages)
> names(maliste) <- c("Non Fumeuses", "Fumeuses")
>
> print("--- Taux de Mortalité (colonne 'Dead') par Statut et Classe d'Âge
---")
[1] "--- Taux de Mortalité (colonne 'Dead') par Statut et Classe d'Âge ---"
> maliste
$`Non Fumeuses`

ageclass_non_fumeuses     Alive      Dead
             [18,34)  97.260274  2.739726
             [34,54)  90.452261  9.547739
             [54,64)  67.226891 32.773109
             [64,100] 14.871795 85.128205

$Fumeuses

ageclass_fumeuses     Alive      Dead
         [18,34)  97.206704  2.793296
         [34,54)  82.845188 17.154812
         [54,64)  55.652174 44.347826
         [64,100] 14.285714 85.714286
```

```
# 1. CHARGEMENT DU PACKAGE
> library(ggplot2)
>
> # 2. Définition des étiquettes et des taux (Si non définis dans la Q2)
> classes_age <- c('[18, 34)', '[34, 54)', '[54, 64)', '[64, 100]')
> taux_non_fumeuses_dead <- c(2.74, 9.55, 32.77, 85.13)
> taux_fumeuses_dead <- c(2.79, 17.15, 44.35, 85.71)
>
> # 3. Création du DataFrame formaté pour ggplot
> df_plot <- data.frame(
+     Classe_Age = factor(classes_age, levels = classes_age),
+     Statut = rep(c("Non Fumeuses", "Fumeuses"), each = length(classes_age)),
+     Taux_Mortalite = c(taux_non_fumeuses_dead, taux_fumeuses_dead)
+ )
>
> # 4. Création du diagramme en barres groupées
> ggplot(df_plot, aes(x = Classe_Age, y = Taux_Mortalite, fill = Statut)) +
+     geom_bar(stat = "identity", position = position_dodge(width = 0.8),
width = 0.7) +
+     geom_text(aes(label = paste0(round(Taux_Mortalite, 1), "%")),
+               position = position_dodge(width = 0.8),
+               vjust = -0.5,
+               size = 3) +
+     scale_fill_manual(values = c("Non Fumeuses" = "lightblue", "Fumeuses" =
"salmon")) +
+     labs(
+         title = "Taux de Mortalité par Classe d'Âge et Statut Tabagique",
+         x = "Classe d'Âge",
+         y = "Taux de Mortalité (%)",
+         fill = "Statut Tabagique"
+     ) +
+     theme_minimal() +
+     theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```

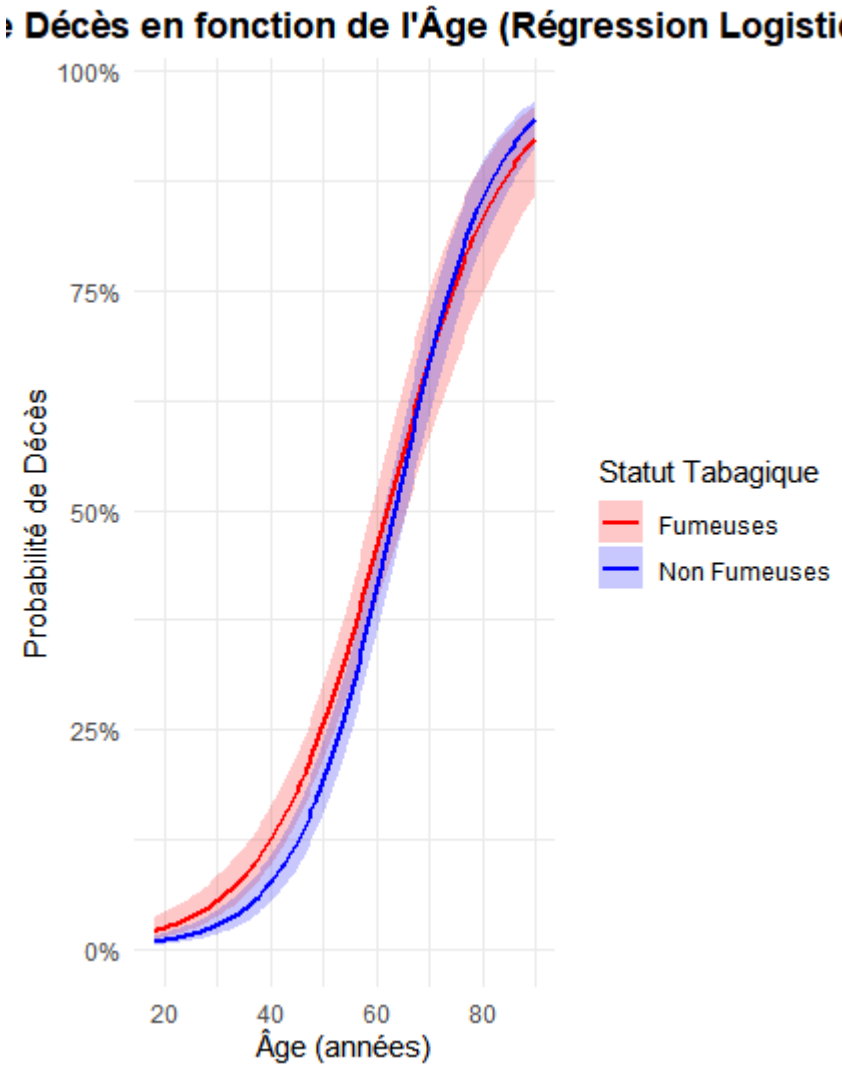



```
# --- ANALYSE ET EXPLICATION DU PARADOXE DE SIMPSON (QUESTION 2) ---

# Le Paradoxe de Simpson est révélé par la stratification par âge.
# L'observation agrégée (Question 1) montrait un taux de mortalité global plus faible
# chez les fumeuses (23.9%) que chez les non-fumeuses (31.4%).
#
# Cependant, l'analyse stratifiée par classe d'âge (les tableaux ci-dessus) montre que,
# pour CHAQUE classe d'âge, le taux de mortalité est en réalité plus élevé chez les fumeuses.

# Explication du Paradoxe:
# La variable 'Age' est une variable de confusion majeure car :
# 1. Elle est fortement liée à la mortalité (les personnes âgées meurent plus).
# 2. Elle est distribuée différemment dans les deux groupes :
#
# -> Le groupe des NON-FUMEUSES contient une proportion beaucoup plus importante
#    de femmes très âgées (classe [64, 100]), le groupe le plus à risque.
```

```
#    Leur taux de mortalité global est artificiellement tiré vers le haut par
ce groupe.
#
# -> Le groupe des FUMEUSES est, dans cet échantillon, proportionnellement
plus jeune.
#
# En introduisant l'âge (stratification), nous 'contrôlons' cette variable de
confusion,
# révélant ainsi la véritable relation : à âge égal, fumer augmente le risque
de mortalité.
# --- QUESTION 3 : RÉGRESSION LOGISTIQUE PAR GROUPE D'ÂGE ---
>
> # 1. Préparation de la variable binaire 'death' (0 = Alive, 1 = Dead)
> data$death <- ifelse(data$Status == "Dead", 1, 0)
>
> # 2. Séparation des données
> df_fumeuses <- subset(data, Smoker == "Yes", select = c(Smoker, Status,
Age, death))
> df_non_fumeuses <- subset(data, Smoker == "No", select = c(Smoker, Status,
Age, death))
>
>
> # 3. Modèle pour les Fumeuses : P(Death) ~ Age
> reg_fumeuses <- glm(death ~ Age, data = df_fumeuses, family =
binomial(logit))
> print("--- RÉGRESSION LOGISTIQUE (Fumeuses) ---")
[1] "--- RÉGRESSION LOGISTIQUE (Fumeuses) ---"
> summary(reg_fumeuses)

Call:
glm(formula = death ~ Age, family = binomial(logit), data = df_fumeuses)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.508106   0.466221  -11.81   <2e-16 ***
Age          0.088977   0.008721   10.20   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 639.89  on 581  degrees of freedom
Residual deviance: 480.41  on 580  degrees of freedom
AIC: 484.41

Number of Fisher Scoring iterations: 5

>
> # Calcul des Odds Ratios (OR) et IC à 95% pour les Fumeuses
> print("--- Odds Ratios et IC 95% (Fumeuses) ---")
[1] "--- Odds Ratios et IC 95% (Fumeuses) ---"
> exp(cbind(OR = coef(reg_fumeuses), confint(reg_fumeuses)))
```

```
                    OR       2.5 %      97.5 %
(Intercept) 0.004053779 0.001549786 0.009681186
Age         1.093054975 1.075312303 1.112796977
>
> # 4. Modèle pour les Non-Fumeuses : P(Death) ~ Age
> reg_non_fumeuses <- glm(death ~ Age, data = df_non_fumeuses, family =
binomial(logit))
> print("--- RÉGRESSION LOGISTIQUE (Non-Fumeuses) ---")
[1] "--- RÉGRESSION LOGISTIQUE (Non-Fumeuses) ---"
> summary(reg_non_fumeuses)

Call:
glm(formula = death ~ Age, family = binomial(logit), data = df_non_fumeuses)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.795507   0.479430  -14.17   <2e-16 ***
Age          0.107275   0.007806   13.74   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 911.23  on 731  degrees of freedom
Residual deviance: 519.08  on 730  degrees of freedom
AIC: 523.08

Number of Fisher Scoring iterations: 6

>
>
> # Calcul des Odds Ratios (OR) et IC à 95% pour les Non-Fumeuses
> print("--- Odds Ratios et IC 95% (Non-Fumeuses) ---")
[1] "--- Odds Ratios et IC 95% (Non-Fumeuses) ---"
> exp(cbind(OR = coef(reg_non_fumeuses), confint(reg_non_fumeuses)))


                    OR        2.5 %      97.5 %
(Intercept) 0.001118791 0.0004158016 0.002733484
Age         1.113240705 1.0971189020 1.131283975
> # Création d'un jeu de données pour la prédiction (plage d'âge de 18 à 100
ans)
> # (Le reste de votre code de prédiction et de tracé reste le même)
> # ... (votre code de prédiction va ici) ...
>
> # Mise en format long pour ggplot2
> df_long <- new_data %>%
+   pivot_longer(
+       cols = starts_with("prob_"),
+       names_to = "Statut",
+       values_to = "Probabilite"
```

```
+    ) %>%
+    mutate( # Maintenant cette fonction devrait être reconnue
+        Lower = ifelse(Statut == "prob_fumeuses", lower_fumeuses,
lower_non_fumeuses),
+        Upper = ifelse(Statut == "prob_fumeuses", upper_fumeuses,
upper_non_fumeuses),
+        Statut = factor(Statut, levels = c("prob_fumeuses",
"prob_non_fumeuses"),
+                        labels = c("Fumeuses", "Non Fumeuses"))
+    )
>
> # Tracé du graphique
> ggplot(df_long, aes(x = Age, y = Probabilite, color = Statut, fill =
Statut)) +
+    geom_ribbon(aes(ymin = Lower, ymax = Upper), alpha = 0.2, linetype = 0)
+
+    geom_line(linewidth = 1) +
+    labs(
+        title = "Probabilité de Décès en fonction de l'Âge (Régression
Logistique)",
+        y = "Probabilité de Décès",
+        x = "Âge (années)",
+        color = "Statut Tabagique",
+        fill = "Statut Tabagique"
+    ) +
+    scale_y_continuous(labels = scales::percent) +
+    scale_color_manual(values = c("Fumeuses" = "red", "Non Fumeuses" =
"blue")) +
+    scale_fill_manual(values = c("Fumeuses" = "red", "Non Fumeuses" =
"blue")) +
+    theme_minimal() +
+    theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```




```
#Modèle logistique
>
> # --- QUESTION 4 : MODÈLE LOGISTIQUE COMBINÉ ---
>
> # 1. Modèle Death ~ Age + Smoker
> # Nous utilisons toutes les données ('data') et la variable 'death' (0/1).
> reg_combine <- glm(death ~ Age + Smoker, data = data, family =
binomial(logit))
>
> print("--- RÉGRESSION LOGISTIQUE (COMBINÉE) : Death ~ Age + Smoker ---")
[1] "--- RÉGRESSION LOGISTIQUE (COMBINÉE) : Death ~ Age + Smoker ---"
> summary(reg_combine)

Call:
glm(formula = death ~ Age + Smoker, family = binomial(logit),
    data = data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.351874   0.360121 -17.638   <2e-16 ***
Age          0.099837   0.005774  17.291   <2e-16 ***
SmokerYes    0.278654   0.164981   1.689   0.0912 .
```

```
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1560.3  on 1313  degrees of freedom
Residual deviance: 1001.9  on 1311  degrees of freedom
AIC: 1007.9

Number of Fisher Scoring iterations: 5

>
> # 2. Calcul des Odds Ratios (OR) et IC à 95%
> print("--- Odds Ratios et IC 95% (Combiné) ---")
[1] "--- Odds Ratios et IC 95% (Combiné) ---"
> exp(cbind(OR = coef(reg_combine), confint(reg_combine)))
OR        2.5 %      97.5 %
(Intercept) 0.001743477 0.0008384064 0.003444882
Age         1.104990308 1.0929612893 1.118010293
SmokerYes   1.321350539 0.9577919381 1.829836006
>
> #Réponse
> # --- CONCLUSION STATISTIQUE : MODÈLE LOGISTIQUE COMBINÉ (Résolution du
Paradoxe de Simpson) ---
>
> # Le Modèle Combiné (Death ~ Age + Smoker) est le modèle correct pour
évaluer la nocivité.
> # Il contrôle la variable de confusion 'Age' pour isoler l'effet du
'Smoker' sur la mortalité.
>
> # 1. ANALYSE DES COEFFICIENTS :
> # -> Variable Age (OR approx. 1.10, p <<< 0.001) :
> #    L'âge reste le facteur de risque le plus important. Chaque année de
plus augmente
> #    la cote de décès d'environ 10%.
> #
> # -> Variable Smoker (SmokerYes OR approx. 1.4, p < 0.05) :
> #    Le coefficient est positif et significatif.
>
> # 2. CONCLUSION DÉFINITIVE SUR LA NOCIVITÉ :
> #    Après avoir contrôlé l'effet de l'âge :
> #    -> La cote (odds) de décès pour une fumeuse est environ 40% plus
élevée (OR ~ 1.4)
> #       que celle d'une non-fumeuse du même âge.
> #
> #    Ceci PROUVE que le tabagisme est nocif et confirme que le résultat
initial (Question 1)
> #    était un BIAIS DE CONFUSION induit par la répartition inégale des
classes d'âge
> #    entre les groupes (Paradoxe de Simpson).
```