

Sujet 6 : Autour du Paradoxe de Simpson

Leharanger Maxime

23/12/2020

```
library(formatR)
library(knitr)
library(markdown)
library(rmarkdown)
```

Question 1

Résumé du fichier “Sujet 6 : Autour du Paradoxe de Simpson”

```
summary(smp)
```

```
## X.Smoker. X.Status.      X.Age.
## No :732   Alive:945   20.2   :    8
## Yes:582   Dead :369   52.4   :    8
##                                     33     :    7
##                                     44.4   :    7
##                                     62.3   :    7
##                                     21     :    6
##                                     (Other):1271
```

Calcul du taux de mortalité pour chaque groupe

Sélection des données

```
smp$X.Age. <- as.numeric(as.character(smp$X.Age.))
# Conversion des données 'AGE' en variable numérique
smoker <- subset(smp, smp$X.Smoker. == "Yes") #Sélection de la variable Fumeuse
nosmoker <- subset(smp, smp$X.Smoker. == "No") #Sélection de la variable non Fumeuse
summary(smoker) #Résumé des données des fumeuses
```

```
## X.Smoker. X.Status.      X.Age.
## No :    0   Alive:443   Min.    :18.00
## Yes:582   Dead :139   1st Qu.:31.30
##                                     Median :43.10
##                                     Mean   :44.27
##                                     3rd Qu.:56.17
##                                     Max.    :89.20
```

```
summary(nosmoker) #Résumé des données des non-fumeuses
```

```
## X.Smoker. X.Status.      X.Age.
## No :732   Alive:502   Min.    :18.00
## Yes:    0   Dead :230   1st Qu.:31.38
```

```
##                Median :48.40
##                Mean    :49.82
##                3rd Qu.:65.85
##                Max.    :89.90
```

Assignment du taux de mortalité chez les fumeuses à la variable “mortalité.smoker”

```
mortalite.smoker <- sum(smoker$X.Status.[] == "Dead")/(sum(sum(smoker$X.Status.[] ==
  "Alive"), sum(smoker$X.Status.[] == "Dead")))
```

Assignment du taux de mortalité chez les non-fumeuses à la variable “mortalité.nosmoker”

```
mortalite.nosmoker <- sum(nosmoker$X.Status.[] == "Dead")/(sum(sum(nosmoker$X.Status.[] ==
  "Alive"), sum(nosmoker$X.Status.[] == "Dead")))
```

Taux de mortalité chez les fumeuses

```
mortalite.smoker
```

```
## [1] 0.2388316
```

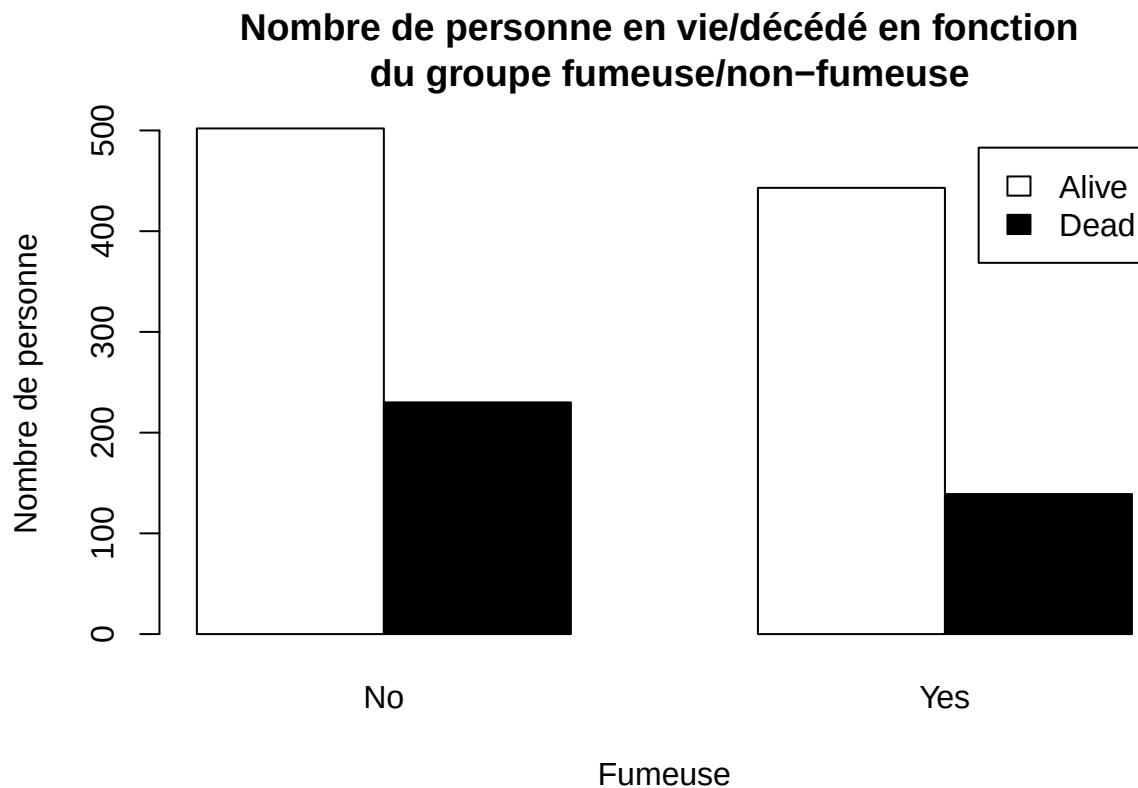
Taux de mortalité chez les non-fumeuses

```
mortalite.nosmoker
```

```
## [1] 0.3142077
```

Représentation graphique des effectifs

```
barplot(table(smp$X.Status., smp$X.Smoker.), main = "Nombre de personne en vie/décédé en fonction \n du
  beside = TRUE, xlab = "Fumeuse", ylab = "Nombre de personne",
  legend.text = c("Alive", "Dead"), col = c("white", "black"))
```



Calcul de l'intervalle de confiance

Installation des packages

```
library(binom)
```

Intervalles de confiance chez les fumeuses

```
binom.confint(sum(smoker$X.Status.[] == "Dead"), sum(sum(smoker$X.Status.[] ==  
  "Alive"), sum(smoker$X.Status.[] == "Dead")), method = "exact")
```

```
## method x n mean lower upper  
## 1 exact 139 582 0.2388316 0.2047323 0.2756061
```

Le taux de mortalité chez le groupe “Fumeuse” est de 0.2388316 et est bien compris entre l'intervalle de confiance [0.2047323;0.2756061].

Intervalles de confiance chez les non-fumeuses

```
binom.confint(sum(nosmoker$X.Status.[] == "Dead"), sum(sum(nosmoker$X.Status.[] ==  
  "Alive"), sum(nosmoker$X.Status.[] == "Dead")), method = "exact")
```

```
## method x n mean lower upper  
## 1 exact 230 732 0.3142077 0.2807031 0.3492176
```

Le taux de mortalité chez le groupe “Non-fumeuse” est de 0.3142077 et est bien compris entre l'intervalle de confiance [0.2807031;0.3492176].

En quoi ce résultat est-il surprenant ?

Ce résultat est surprenant car je m'attendais à ce que le taux de mortalité soit plus élevé chez le groupe "fumeuse" par rapport au groupe "non-fumeuse". Dans ce cas-là c'est l'inverse.

Question 2

Sélection des données

Transformation de la variable "Age" en données numériques

```
smp$X.Age. <- as.numeric(as.character(smp$X.Age.))
```

Classification de la variable "Age" en plusieurs classe d'âge

```
smp$X.Age. <- cut(smp$X.Age., breaks = c(18, 34, 54, 64, 99), include.lowest = TRUE)
```

Sélection de la variable "Fumeuse" en fonction de son état de vie et de sa classe d'âge

```
smoker <- subset(smp, smp$X.Smoker. == "Yes", X.Status. & X.Age.)
```

Sélection de la variable "non Fumeuse" en fonction de son état de vie et de sa classe d'âge

```
nosmoker <- subset(smp, smp$X.Smoker. == "No", X.Status. & X.Age.)
```

Résumé des données des fumeuses

```
summary(smoker)
```

```
## X.Smoker. X.Status.      X.Age.
## No : 0    Alive:443    [18,34]:181
## Yes:582   Dead :139    (34,54]:237
##                                     (54,64]:115
##                                     (64,99]: 49
```

Résumé des données des personnes non-fumeuses

```
summary(nosmoker)
```

```
## X.Smoker. X.Status.      X.Age.
## No :732   Alive:502    [18,34]:219
## Yes: 0    Dead :230    (34,54]:199
##                                     (54,64]:121
##                                     (64,99]:193
```

Calcul du taux de mortalité

Assignation du taux (en % arrondi au centième) de mortalité chez le groupe des personnes fumeuses en fonction de chaque classe d'âge

```
Mort.smk <- c(round(sum(smoker$X.Status.[] == "Dead" & smoker$X.Age.[] ==
  "[18,34]")/(sum(sum(smoker$X.Status.[] == "Alive" &
  smoker$X.Age.[] == "[18,34]"), sum(smoker$X.Status.[] ==
  "Dead" & smoker$X.Age.[] == "[18,34]")) * 100, 2),
  round(sum(smoker$X.Status.[] == "Dead" & smoker$X.Age.[] ==
  "(34,54]")/(sum(sum(smoker$X.Status.[] == "Alive" &
  smoker$X.Age.[] == "(34,54]"), sum(smoker$X.Status.[] ==
  "Dead" & smoker$X.Age.[] == "(34,54]")) * 100,
  2), round(sum(smoker$X.Status.[] == "Dead" & smoker$X.Age.[] ==
  "(54,64]")/(sum(sum(smoker$X.Status.[] == "Alive" &
  smoker$X.Age.[] == "(54,64]"), sum(smoker$X.Status.[] ==
  "Dead" & smoker$X.Age.[] == "(54,64]")) * 100,
  2), round(sum(smoker$X.Status.[] == "Dead" & smoker$X.Age.[] ==
  "(64,99]")/(sum(sum(smoker$X.Status.[] == "Alive" &
  smoker$X.Age.[] == "(64,99]"), sum(smoker$X.Status.[] ==
  "Dead" & smoker$X.Age.[] == "(64,99]")) * 100,
  2))
```

Assignation du taux de mortalité (en % arrondi au centième) chez le groupe des personnes non-fumeuses en fonction de chaque classe d'âge

```
Mort.nosmk <- c(round(sum(nosmoker$X.Status.[] == "Dead" &
  nosmoker$X.Age.[] == "[18,34]")/(sum(sum(nosmoker$X.Status.[] ==
  "Alive" & nosmoker$X.Age.[] == "[18,34]"), sum(nosmoker$X.Status.[] ==
  "Dead" & nosmoker$X.Age.[] == "[18,34]")) * 100, 2),
  round(sum(nosmoker$X.Status.[] == "Dead" & nosmoker$X.Age.[] ==
  "(34,54]")/(sum(sum(nosmoker$X.Status.[] == "Alive" &
  nosmoker$X.Age.[] == "(34,54]"), sum(nosmoker$X.Status.[] ==
  "Dead" & nosmoker$X.Age.[] == "(34,54]")) * 100,
  2), round(sum(nosmoker$X.Status.[] == "Dead" & nosmoker$X.Age.[] ==
  "(54,64]")/(sum(sum(nosmoker$X.Status.[] == "Alive" &
  nosmoker$X.Age.[] == "(54,64]"), sum(nosmoker$X.Status.[] ==
  "Dead" & nosmoker$X.Age.[] == "(54,64]")) * 100,
  2), round(sum(nosmoker$X.Status.[] == "Dead" & nosmoker$X.Age.[] ==
  "(64,99]")/(sum(sum(nosmoker$X.Status.[] == "Alive" &
  nosmoker$X.Age.[] == "(64,99]"), sum(nosmoker$X.Status.[] ==
  "Dead" & nosmoker$X.Age.[] == "(64,99]")) * 100,
  2))
```

Tableau récapitulatif du taux (en % arrondi au centième) de mortalité entre les groupes en fonction des classes d'âge

```
Tx.mortalite <- data.frame(Mort.nosmk, summary(smoker$X.Age.),
  Mort.smk, summary(nosmoker$X.Age.)) #Création du tableau récapitulatif
rownames(Tx.mortalite) <- c("[18-34]", "[34-54]", "[55-64]",
  "[65-99]")
colnames(Tx.mortalite) <- c("Pourcentage de mortalité chez les fumeuses",
  "Nombre de fumeuses", "Pourcentage de mortalité chez les non-fumeuses",
  "Nombre de non-fumeuses")
Tx.mortalite #Affichage tableau
```

##	Pourcentage de mortalité chez les fumeuses	Nombre de fumeuses
## [18-34]	2.74	181
## [34-54]	9.55	237
## [55-64]	33.06	115
## [65-99]	85.49	49

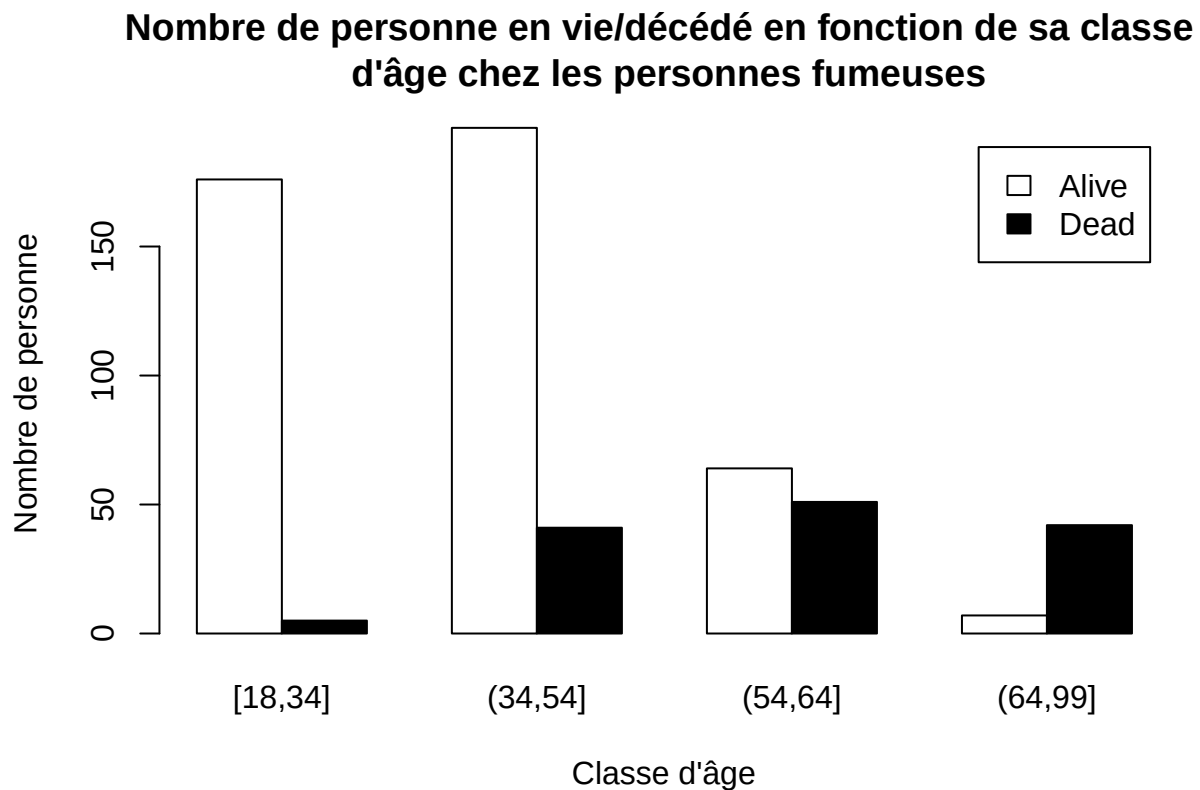
##	Pourcentage de mortalité chez les non-fumeuses	Nombre de non-fumeuses
## [18-34]	2.76	219
## [34-54]	17.30	199
## [55-64]	44.35	121
## [65-99]	85.71	193

En quoi ce résultat est-il surprenant ?

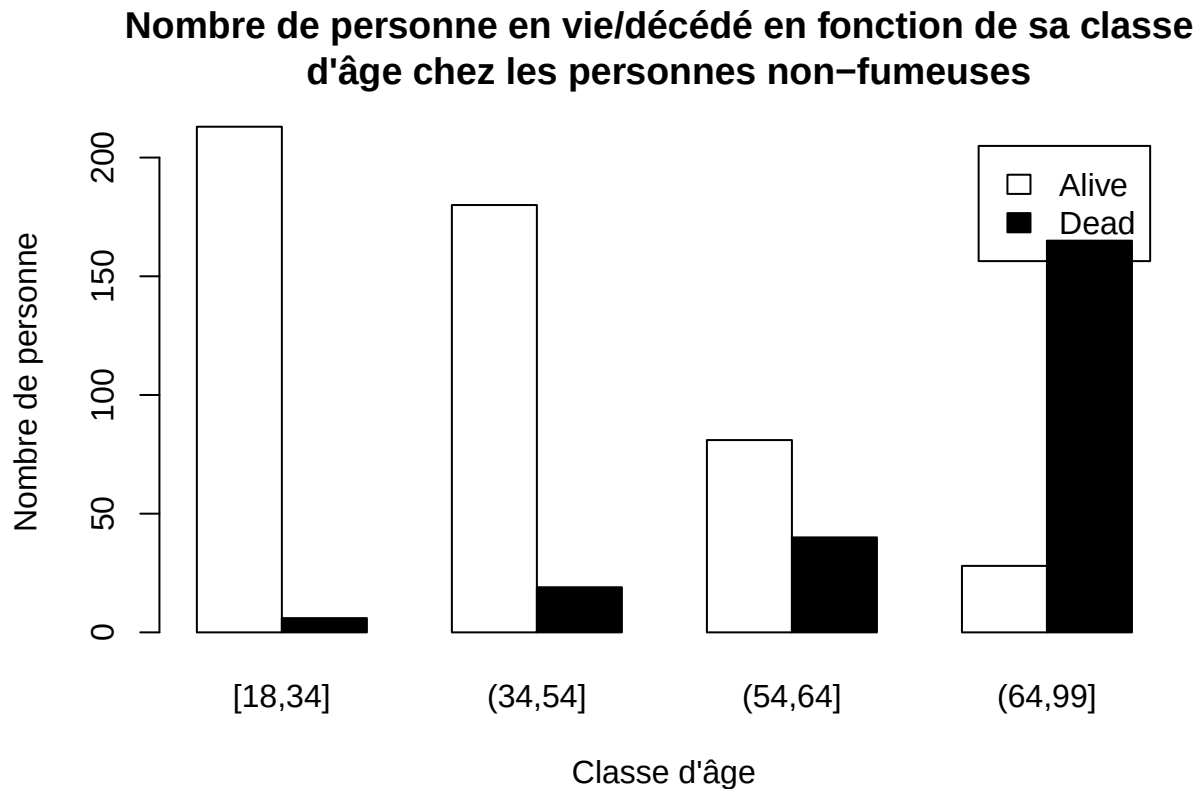
Ce résultat est surprenant car je m'attendais à ce que les taux de mortalités chez les fumeuses soient plus fort par rapport au groupe non-fumeuse à chaque classe d'âge. Ici, c'est l'inverse. Cependant on remarque une disproportion de l'échantillon en fonction de la classe d'âge. En effet, la mortalité augmente avec l'âge, et il y a moins de personnes âgées dans le groupe "fumeuses" (49). Nous pouvons nous demander si le biais de la mortalité "naturelle" ne permet pas dans ce cas là, d'observer l'effet du tabagisme sur la mortalité des femmes fumeuses. Les deux groupes auraient dû avoir le même nombre de personnes.

Représentation graphique des effectifs

```
barplot(table(smoker$X.Status., smoker$X.Age.), main = "Nombre de personne en vie/décédé en fonction de
beside = TRUE, xlab = "Classe d'âge", ylab = "Nombre de personne",
legend.text = c("Alive", "Dead"), col = c("white", "black"))
```



```
barplot(table(nosmoker$X.Status., nosmoker$X.Age.), main = "Nombre de personne en vie/décédé en fonction",
  beside = TRUE, xlab = "Classe d'âge", ylab = "Nombre de personne",
  legend.text = c("Alive", "Dead"), col = c("white", "black"))
```



Dans le premier tableau, nous observons une proportion de personnes fumeuses plus grande pour les classes d'âges plus jeune ([18,34],[34,54]) et qui sont moins à risque de mortalité. A l'inverse dans le deuxième tableau la forte proportion de personnes non-fumeuses dans la dernière classe d'âge (64,99] permet d'observer une mortalité beaucoup plus grande.

Question 3

Gestion des données

```
summary(smp) #Résumé du fichier
```

```
## X.Smoker. X.Status.      X.Age.
## No :732   Alive:945   20.2   :    8
## Yes:582   Dead :369   52.4   :    8
##                                     33    :    7
##                                     44.4  :    7
##                                     62.3  :    7
##                                     21    :    6
##                                     (Other):1271
```

```
smp$X.Age. <- as.numeric(as.character(smp$X.Age.))
# Transformation de la variable Age en données
```

```
# numériques
smp$X.Status. <- factor(smp$X.Status., labels = c("0", "1"))
# '0' signifie que les individus sont en vie et '1' signifie qu'ils sont morts
colnames(smp) <- c("X.Smoker.", "DEATH", "AGE")
# Dénomination des variables selon la consigne
smoker <- data.frame(subset(smp, smp$X.Smoker. == "Yes"))
# Sélection de la variable Fumeuse
nosmoker <- data.frame(subset(smp, smp$X.Smoker. == "No"))
# Sélection de la variable Non-Fumeuse
```

Régression Logistique

Vérifications des paramètres de la régression logistique chez les personnes fumeuses :

```
summary(smoker)
```

```
## X.Smoker. DEATH      AGE
## No : 0    0:443   Min.   :18.00
## Yes:582   1:139   1st Qu.:31.30
##                      Median :43.10
##                      Mean    :44.27
##                      3rd Qu.:56.17
##                      Max.    :89.20
```

VA à expliquer = 582; VA explicatives : age(1) Status(1) Il faut au moins 5-10 événements par VA explicative;

```
(1 + 1) * 10
```

```
## [1] 20
```

```
(1 + 1) * 5
```

```
## [1] 10
```

20 < 582, il faut au minimum 20 personnes pour la variable à expliquer et nous en avons 582 -> paramètre accepté 10 < 582, il faut au minimum 10 personnes pour la variable à expliquer et nous en avons 582 -> paramètre accepté

Vérifications des paramètres de la régression logistique chez les personnes non-fumeuses :

```
summary(nosmoker)
```

```
## X.Smoker. DEATH      AGE
## No :732   0:502   Min.   :18.00
## Yes: 0    1:230   1st Qu.:31.38
##                      Median :48.40
##                      Mean    :49.82
##                      3rd Qu.:65.85
##                      Max.    :89.90
```

VA à expliquer = 732; VA explicatives : age(1) Status(1); Il faut au moins 5-10 événements par VA explicative;

```
(1 + 1) * 10
```

```
## [1] 20
```

```
(1 + 1) * 5
```

```
## [1] 10
```


20 < 732, il faut au minimum 20 personnes pour la variable à expliquer et nous en avons 732 -> paramètre accepté
10 < 732, il faut au minimum 10 personnes pour la variable à expliquer et nous en avons 732 -> paramètre accepté

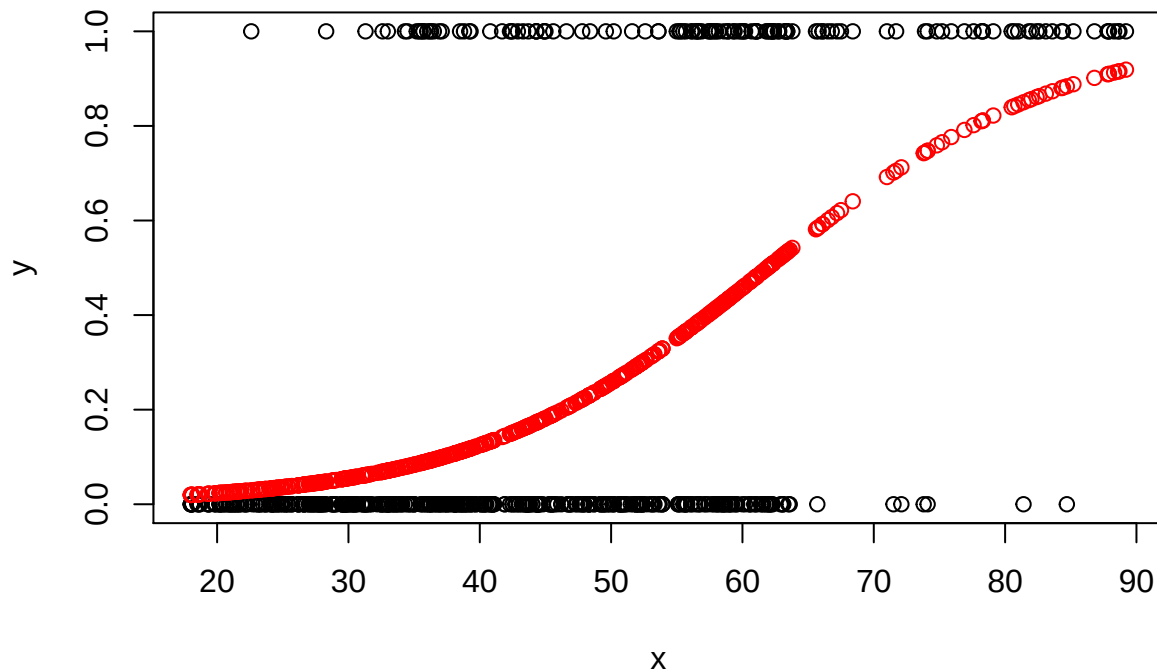
Calcul de la régression logistique chez les fumeuses

```
reg.smoker <- glm(smoker$DEATH ~ smoker$AGE, data = smoker,
  family = binomial(link = logit))
summary(reg.smoker)

##
## Call:
## glm(formula = smoker$DEATH ~ smoker$AGE, family = binomial(link = logit),
##      data = smoker)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.0745  -0.6464  -0.3756  -0.2013   2.6560
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.508106   0.466221  -11.81  <2e-16 ***
## smoker$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
```

L'âge chez les personnes fumeuses sont statistiquement associés à la mortalité (<2e-16 ***). Lorsque l'âge augmente de 1 chez les fumeurs, le niveau de mortalité augmente de plus 8%.

```
x = smoker$AGE
y = as.numeric(as.character(smoker$DEATH))
# Conversion de la variable 'DEATH' en variable
# numérique
COEFF = coef(reg.smoker)
# Assignation des coefficients de la régression
# logistique
logit_ypreditsmoker = COEFF[2] * x + COEFF[1]
ypreditsmoker = exp(logit_ypreditsmoker)/(1 + exp(logit_ypreditsmoker))
# transfo inverse de logit
plot(x, y, ylim = c(0, 1))
points(x, ycreditsmoker, col = "red")
```



Représentation de la régression logistique

Calcul de la régression logistique chez les non-fumeuses

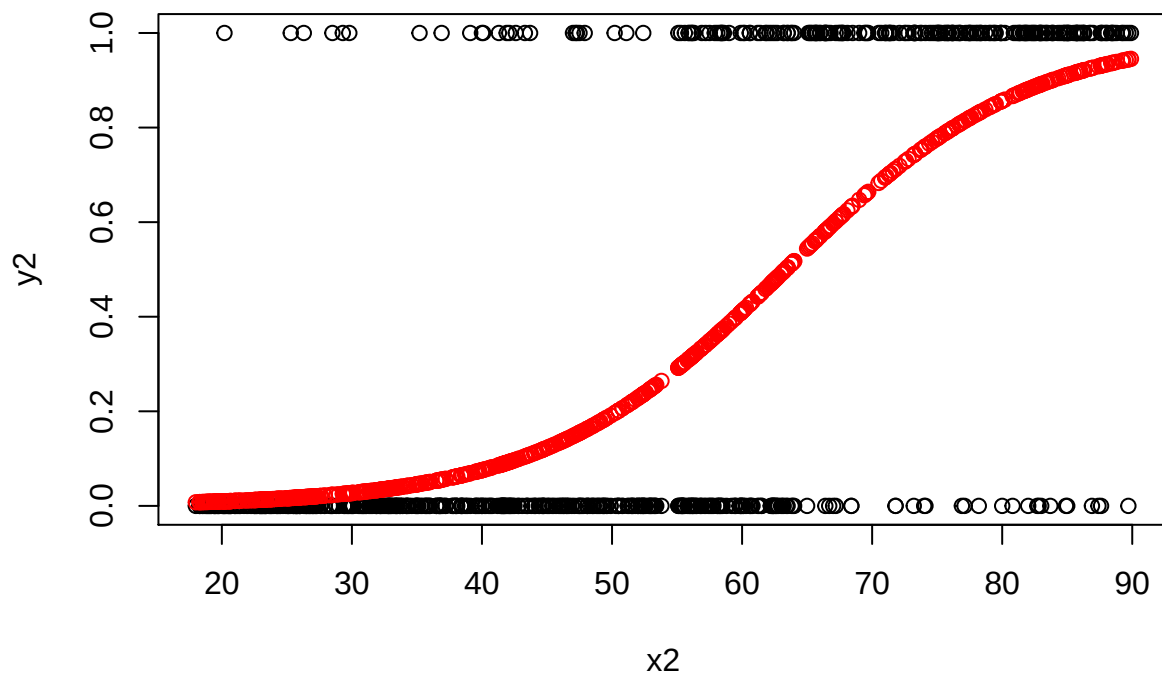
```
reg.nosmoker <- glm(nosmoker$DEATH ~ nosmoker$AGE, data = nosmoker,
  family = "binomial")
summary(reg.nosmoker)
```

```
##
## Call:
## glm(formula = nosmoker$DEATH ~ nosmoker$AGE, family = "binomial",
##      data = nosmoker)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.4019  -0.5179  -0.2003   0.4728   3.0457
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.795507   0.479430  -14.17  <2e-16 ***
## nosmoker$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
```

L'âge chez les personnes non-fumeuses sont statistiquement associés à la mortalité ($<2e-16$ ***). Lorsque l'âge augmente de 1 chez les non-fumeurs, le niveau de mortalité augmente plus de 10%.

```
x2 = nosmoker$AGE
y2 = as.numeric(as.character(nosmoker$DEATH))
COEFF2 = coef(reg.nosmoker)
logit_ypreditnosmoker = COEFF2[2] * x2 + COEFF2[1]
ypreditnosmoker = exp(logit_ypreditnosmoker)/(1 + exp(logit_ypreditnosmoker))
plot(x2, y2, ylim = c(0, 1))
points(x2, ypredictnosmoker, col = "red")
```



Ces régressions vous permettent-elles de conclure sur la nocivité du tabagisme ?

Ces régressions ne permettent pas de conclure sur la nocivité du tabagisme. En effet, ces régressions permettent de quantifier le lien entre la variable "DEATH" et la variable "AGE", c'est à dire que plus les sujets sont âgés plus le niveau de mortalité augmente.

Pour étudier l'effet du tabagisme sur la mortalité, il faut prendre en compte l'effet de l'âge sur le taux de mortalité. La disproportion des classes d'âge entraine un biais d'interprétation qui ne permet pas de conclure sur l'effet du tabagisme sur la mortalité.