Autour du Paradoxe de Simpson

Brahima DIARRA

22 juin 2020

l'objectif de ce document est de reproduire les résultats d'un sondage sur un sixième des électeurs de la ville de Wickham, une ville au nord-est de l'Angleterre dans les années 1972-1974. Cette étude se voulait d'éclairer des travaux sur les maladies thyroïdiennes et cardiaques (Tunbridge et al. 1977). Une suite de cette étude a été menée vingt ans plus tard (Vanderpump et al. 1995). Certains des résultats avaient trait au tabagisme et cherchaient à savoir si les individus étaient toujours en vie lors de la seconde étude. Par simplicité, nous nous restreindrons aux femmes et parmi celles-ci aux 1314 qui ont été catégorisées comme "fumant actuellement" ou "n'ayant jamais fumé". Il y avait relativement peu de femmes dans le sondage initial ayant fumé et ayant arrêté depuis (162) et très peu pour lesquelles l'information n'était pas disponible (18). La survie à 20 ans a été déterminée pour l'ensemble des femmes du premier sondage.

Les données utilisées sont téléchargeables en format CSV à cette adresse. chaque ligne renseigne si la personne fume ou non, si elle est vivante ou décédée au moment de la seconde étude, et son âge lors du premier sondage.

Tâches préliminaires

Cette section va consister à faire les travaux préliminaires avant d'aborder les tâches mentionnées dans l'étude de cas. Ces tâches sont :

1. Représenter dans un tableau le nombre total de femmes vivantes et décédées sur la période en fonction de leur habitude de tabagisme ;
2. Reprendre la question 1 (effectifs et taux de mortalité) en rajoutant une nouvelle catégorie liée à la classe d'âge ;
3. Envisager d'essayer de réaliser une régression logistique afin d'éviter un biais induit par des regroupements en tranches d'âges arbitraires et non régulières.

Téléchargement des données

Nous allons télécharger les données, en utilisant l'adresse du lien indiqué ci-dessus, pour les stocker sur notre machine pour ne pas avoir à les charger à chaque fois qu'on veut relancer l'analyse.

```
#Url de téléchargement
data_url <- "https://gitlab.inria.fr/learninglab/mooc-rr/mooc-rr-ressources/-/raw/master/module3/Practi

#Test d'existence du fichiers de données
data_file <- "subject6_smoking.csv"

if (!file.exists(data_file)){
  download.file(data_url, data_file, method = "auto")
}
```

Lecture des données

Après avoir télécharger les données et les stocker sur le disque local, nous pouvons les charger dans R et procéder à l'inspection.

```
data <- read.csv(data_file)
head(data)
```

```
##   Smoker Status  Age
## 1    Yes  Alive 21.0
## 2    Yes  Alive 19.3
## 3     No   Dead 57.5
## 4     No  Alive 47.1
## 5    Yes  Alive 81.4
## 6     No  Alive 36.8
```

```
tail(data)
```

```
##      Smoker Status  Age
## 1309     No  Alive 42.1
## 1310    Yes  Alive 35.9
## 1311     No  Alive 22.3
## 1312    Yes   Dead 62.1
## 1313     No   Dead 88.6
## 1314     No  Alive 39.1
```

On peut aussi vérifier la présence d'observations manquantes

```
lignes_na <- apply(data, 1, function(x)any(is.na(x)))
data[lignes_na,]
```

```
## [1] Smoker Status Age
## <0 rows> (or 0-length row.names)
```

La sortie nous indique qu'il n'y a pas d'observations manquantes dans les données. On peut procéder aux analyses

Installation et chargement des packages utilisées

Nous utilisons une série de package pour pouvoir faire cette analyse, nous allons d'abord tester si un package parmi la liste n'est pas installé, on l'installe puis on charge l'ensemble des packages nécessaires.

```
## list()
```

```
## [[1]]
## [1] "dplyr"     "stats"     "graphics"  "grDevices" "utils"     "datasets"
## [7] "methods"   "base"
##
## [[2]]
## [1] "ggplot2"   "dplyr"     "stats"     "graphics"  "grDevices" "utils"
## [7] "datasets"  "methods"   "base"
##
## [[3]]
##  [1] "janitor"   "ggplot2"   "dplyr"     "stats"     "graphics"  "grDevices"
##  [7] "utils"     "datasets"  "methods"   "base"
##
## [[4]]
```

Table 1: Nombre total de femmes vivantes et décédées sur la période en fonction de leur habitude de tabagisme

Status/Smoker	No	Yes	Total
Alive	502	443	945
Dead	230	139	369
Total	732	582	1314

Source: Sondage auprès d'un sixième des électeurs afin d'éclairer des travaux sur les maladies t

```
##  [1] "knitr"     "janitor"   "ggplot2"   "dplyr"     "stats"     "graphics"
##  [7] "grDevices" "utils"     "datasets"  "methods"   "base"
##
## [[5]]
##  [1] "kableExtra" "knitr"      "janitor"    "ggplot2"    "dplyr"
##  [6] "stats"      "graphics"   "grDevices"  "utils"      "datasets"
## [11] "methods"    "base"
```

Mission dévolues par le MOOC

1. Représentation dans un tableau du nombre total de femmes vivantes et décédées sur la période en fonction de leur habitude de tabagisme.

Nombre total

Pour avoir ce tableau, nous utilisons la fonction *tabyl* du package janitor qui lui-même se sert de dplyr.

```
data %>%
  tabyl(Status, Smoker) %>%
  adorn_totals(c("row", "col")) %>%
  adorn_title(placement = "combined") %>%
  kable(align = c("l",rep("c",3)),caption = "Nombre total de femmes vivantes et décédées sur la période
  kable_styling(bootstrap_options = c("striped", "hover","responsive"), fixed_thead = T, font_size = 13
  #row_spec(seq(1,nlevels(data$Status)+2, by=1), background = "#D5E4EB")
footnote(general = "Sondage auprès d'un sixième des électeurs afin d'éclairer des travaux sur les maladi
           general_title = "Source: ",
           footnote_as_chunk = T)
```

Il ressort du tableau précédent que sur les 1314 femmes, 369 sont décédées sur la période et 945 étaient toujours en vie. Parmi les 369 femmes décédées, 230 ne fumaient pas alors que 130 fumaient.

Calcul du taux de mortalité suivant le statut de tabagisme

Le taux de mortalité est le rapport entre le nombre de femmes décédées dans un groupe et le nombre total de femmes dans ce groupe.

```
data %>%
  tabyl(Status, Smoker) %>%
  adorn_totals(c("row", "col")) %>%
  adorn_title(placement = "combined") %>%
  adorn_percentages("col") %>%
  adorn_pct_formatting(digits = 1, affix_sign = F) %>%
  kable(align = c("l",rep("c",3)),caption = "Taux de mortalité en fonction de l'habitude de tabagisme")
```

Table 2: Taux de mortalité en fonction de l'habitude de tabagisme

Status/Smoker	No	Yes	Total
Alive	68.6	76.1	71.9
Dead	31.4	23.9	28.1
Total	100.0	100.0	100.0

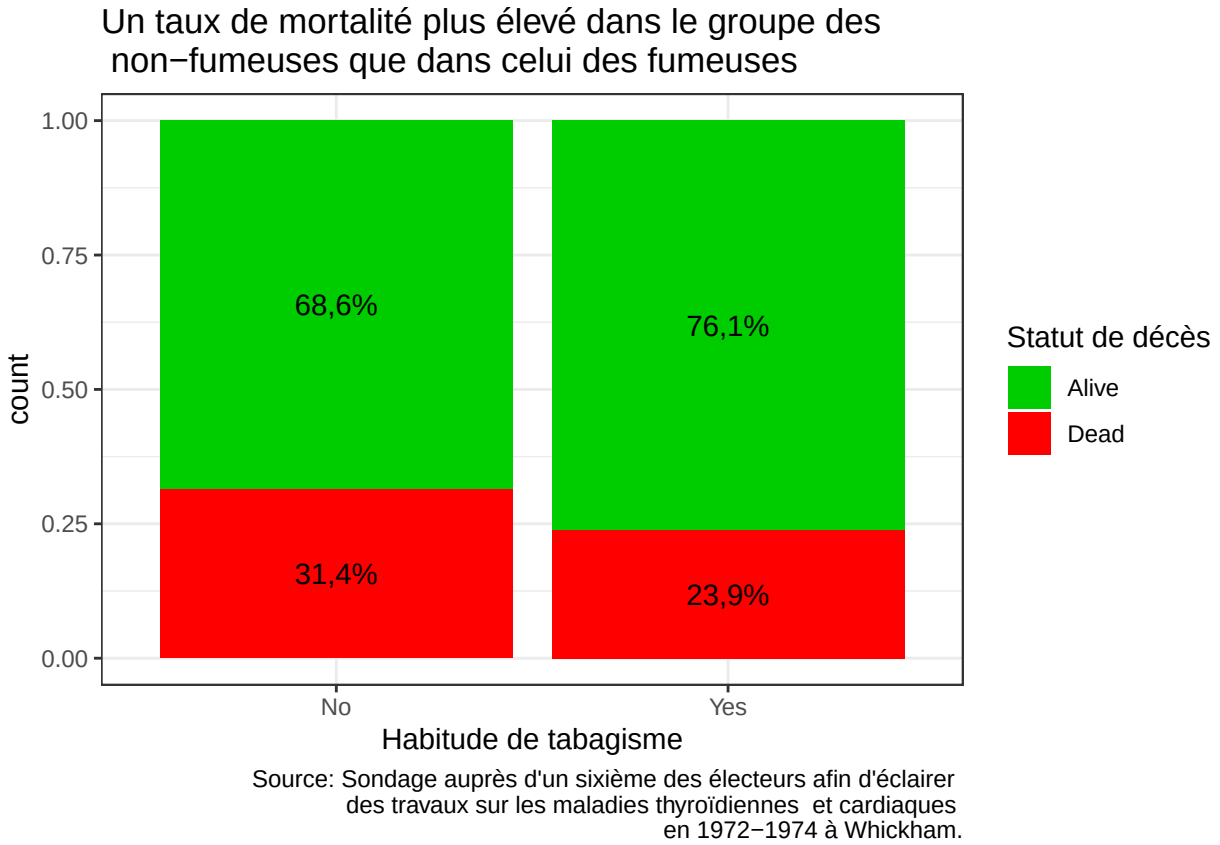
Source: Sondage auprès d'un sixième des électeurs afin d'éclairer des travaux sur les maladies t

```
  kable_styling(bootstrap_options = c("striped", "hover","responsive"), fixed_thead = T, font_size = 13
  #row_spec(seq(1,nlevels(data$Status)+2, by=1), background = "#D5E4EB")
footnote(general = "Sondage auprès d'un sixième des électeurs afin d'éclairer des travaux sur les maladi
           general_title = "Source: ",
           footnote_as_chunk = T)
```

Ce tableau montre que le taux de mortalité est de 31,4% pour les femmes qui ne fumaient pas et 23,9% pour celles qui fumaient. Le graphique suivant permet de mieux cerner les différences entre les deux groupes.

```
data_tx <- data %>% group_by(Smoker) %>% count(Status) %>%
  mutate(ratio=scales::percent(n/sum(n), decimal.mark = ",", accuracy = 0.1))

ggplot(data,aes(x=Smoker,fill=Status))+
  geom_bar(position="fill")+
  geom_text(data=data_tx, aes(y=n,label=ratio),
            position=position_fill(vjust=0.5))+
  scale_fill_manual(values = c("green3", "red1"))+
  labs(title = "Un taux de mortalité plus élevé dans le groupe des \n non-fumeuses que dans celui des fu
       x = "Habitude de tabagisme",
       fill = "Statut de décès",
       caption = "Source: Sondage auprès d'un sixième des électeurs afin d'éclairer \n des travaux sur l
  theme_bw()
```



Le test de Khi2 sur ces deux variables permet de conclure que ces différences sont significatives comme le montre la sortie suivante:

```
chisq.test(data$Smoker, data$Status)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  data$Smoker and data$Status
## X-squared = 8.7515, df = 1, p-value = 0.003093
```

De l'analyse des tableaux qui précède, il est constaté que le taux de mortalité est plus importante dans le groupe des femmes qui ne fumaient pas (31,4%) que dans celui des femmes qui fumaient (23,9%). Ce résultat est étonnant dans la mesure où le tabac a, de manière générale, une incidence négative sur la santé des fumeurs, en conséquence, le taux de mortalité devrait être plus important dans le groupe des femmes qui fumaient.

2. Prise en compte de la classe d'âge dans l'analyse

Création des classes d'âges

On considérera les classes suivantes dans l'analyse: 18-34 ans, 34-54 ans, 55-64 ans, plus de 65 ans. Pour ce faire, on peut utiliser la fonction *cut* de base ou la fonction *icut* du package *questionner*

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   31.30   44.80   47.36   60.60   89.90
```

Table 3: Nombre de décès en fonction de la tranche d'âge

Status/agecl	18-34 ans	35-54 ans	55-64 ans	plus de 65 ans	Total
Alive	389	376	145	35	945
Dead	11	60	91	207	369
Total	400	436	236	242	1314

Source: Sondage auprès d'un sixième des électeurs afin d'éclairer des travaux sur les maladies t

```
##      data$agecl    n percent
##       18-34 ans  400   30.4%
##       35-54 ans  436   33.2%
##       55-64 ans  236   18.0%
##  plus de 65 ans  242   18.4%
##           Total 1314  100.0%
```

La répartition en classe d'âge montre que 30,4% des femmes de l'étude sont dans la tranche d'âge 18-34 ans, le tiers d'entre elles ont entre 35 et 54 ans et 36,4% ont 55 ans ou plus.

Croisement de la mortalité avec les classes d'âges

Le tableau suivant permet de croiser la mortalité en fonction de la tranche d'âge.

```
data %>%
  tabyl(Status, agecl) %>%
  adorn_totals(c("row", "col")) %>%
  adorn_title(placement = "combined") %>%
  kable(align = c("l",rep("c",3)),caption = "Nombre de décès en fonction de la tranche d'âge") %>%
  kable_styling(bootstrap_options = c("striped", "hover","responsive"), fixed_thead = T, font_size = 13
  #row_spec(seq(1,nlevels(data$Status)+2, by=1), background = "#D5E4EB")
footnote(general = "Sondage auprès d'un sixième des électeurs afin d'éclairer des travaux sur les maladi
           general_title = "Source: ",
           footnote_as_chunk = T)
```

Il ressort de ce tableau que l'effectif du nombre de mort a tendance en augmenté en fonction de l'âge des enquêtés.

```
data %>%
  tabyl(Status, agecl) %>%
  adorn_totals(c("row", "col")) %>%
  adorn_title(placement = "combined") %>%
  adorn_percentages("col") %>%
  adorn_pct_formatting(digits = 1, affix_sign = F) %>%
  kable(align = c("l",rep("c",3)),caption = "Taux de mortalité en fonction de la tranche d'âge") %>%
  kable_styling(bootstrap_options = c("striped", "hover","responsive"), fixed_thead = T, font_size = 13
  #row_spec(seq(1,nlevels(data$Status)+2, by=1), background = "#D5E4EB")
footnote(general = "Sondage auprès d'un sixième des électeurs afin d'éclairer des travaux sur les maladi
           general_title = "Source: ",
           footnote_as_chunk = T)
```

Les fréquences relatives confirment le constat précédent. En effet, le taux de mortalité global est 28,1%; ce taux est de 2,8% pour les femmes ayant moins de 35 ans, 13,8% pour celles ayant entre 35 et 54 ans et plus de 85% pour les femmes qui ont 65 ans ou plus. Voyons à présent l'association eventuelle entre l'âge et le tabagisme.

Table 4: Taux de mortalité en fonction de la tranche d'âge

Status/agecl	18-34 ans	35-54 ans	55-64 ans	plus de 65 ans	Total
Alive	97.2	86.2	61.4	14.5	71.9
Dead	2.8	13.8	38.6	85.5	28.1
Total	100.0	100.0	100.0	100.0	100.0

Source: Sondage auprès d'un sixième des électeurs afin d'éclairer des travaux sur les maladies t

Table 5: Habitude de tabagisme en fonction de la tranche d'âge

Smoker/agecl	18-34 ans	35-54 ans	55-64 ans	plus de 65 ans	Total
No	54.8	45.6	51.3	79.8	55.7
Yes	45.2	54.4	48.7	20.2	44.3
Total	100.0	100.0	100.0	100.0	100.0

Source: Sondage auprès d'un sixième des électeurs afin d'éclairer des travaux sur les maladies t

```
data %>%
  tabyl(Smoker, agecl) %>%
  adorn_totals(c("row", "col")) %>%
  adorn_title(placement = "combined") %>%
  adorn_percentages("col") %>%
  adorn_pct_formatting(digits = 1, affix_sign = F) %>%
  kable(align = c("l",rep("c",3)),caption = "Habitude de tabagisme en fonction de la tranche d'âge") %>%
  kable_styling(bootstrap_options = c("striped", "hover","responsive"), fixed_thead = T, font_size = 13
  #row_spec(seq(1,nlevels(data$Status)+2, by=1), background = "#D5E4EB")
footnote(general = "Sondage auprès d'un sixième des électeurs afin d'éclairer des travaux sur les malad
           general_title = "Source: ",
           footnote_as_chunk = T)
```

Aucun lien ne semble se dégager entre ces deux variables au vu du tableau, par contre le test du khi2 conclue au rejet d'absence de liaisons entre elles.

```
chisq.test(data$Smoker, data$agecl)
```

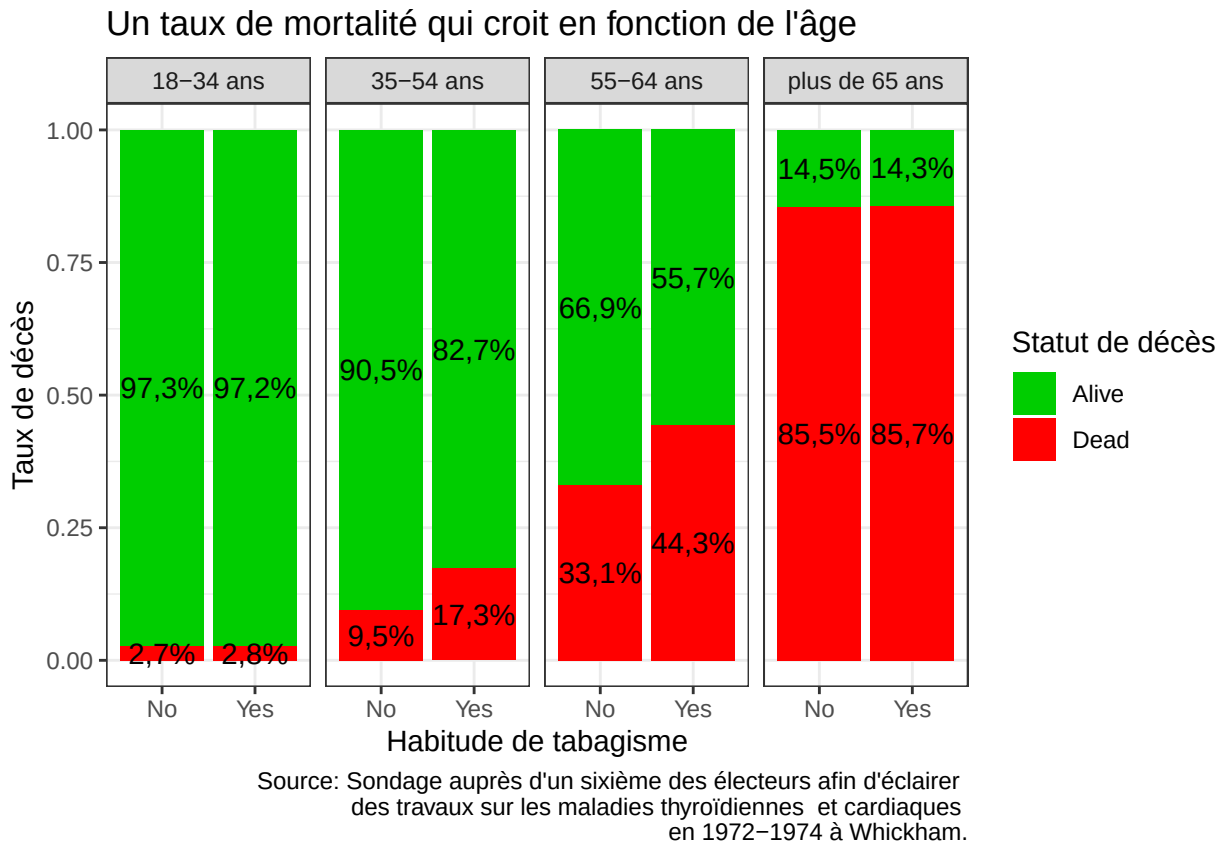
```
##
##  Pearson's Chi-squared test
##
## data:  data$Smoker and data$agecl
## X-squared = 76.636, df = 3, p-value < 2.2e-16
```

Le graphique suivant donne la réprésentation graphique de la mortalité en fonction de l'âge et du tabagisme.

```
data_tx <- data %>%
  group_by(Smoker, agecl) %>%
  count(Status) %>%
  group_by(Smoker, agecl) %>%
  mutate(n_smok = sum(n)) %>%
  mutate(ratio=scales::percent(n/n_smok, decimal.mark = ",", accuracy = 0.1))

ggplot(data,aes(x=Smoker,fill=Status))+
  geom_bar(position="fill")+
```

```
geom_text(data=data_tx, aes(y=n,label=ratio),
          position=position_fill(vjust=0.5))+
scale_fill_manual(values = c("green3", "red1"))+
labs(title = "Un taux de mortalité qui croit en fonction de l'âge",
     x = "Habitude de tabagisme",
     y = "Taux de décès",
     fill = "Statut de décès",
     caption = "Source: Sondage auprès d'un sixième des électeurs afin d'éclairer \n des travaux sur l
facet_grid(~agecl)+
theme_bw()
```



Ce graphique montre que le taux de mortalité est affecté par l'âge pour les femmes se sitaunt dans les tranches d'âges 35-54 et 55-64 ans. En effet, le taux de mortalité pour les femmes qui fumaient de la tranche d'âge 35-54 ans est plus important (17,3%) que pour celles qui ne fumaient pas (9,5%). Ce constat est le même pour la tranche d'âge 55-64 ans avec 44,3% de taux de mortalité pour les femmes qui fumaient contre 33,1% pour celles qui ne fumaient pas.

Il ressort ainsi que ces deux tranches d'âges contredisent la conclusion générale tiré à la première question selon laquelle, le taux de mortalité était plus important dans le groupe des femmes qui ne fumaient pas que dans celui des femmes qui fumaient. Cela pourrait s'expliquer par l'importance de la mortalité dans la tranche d'âge *plus de 65 ans* dans laquelle plus de 56% de la mortalité est concentrée. Le tabagisme n'ayant pas d'effet sur le taux de mortalité dans cette tranche (le taux de mortalité sont sensiblement égaux pour les fumeuses ainsi que pour les non-fumeuses), l'effet du tabagisme sur la mortalité se trouve attenuer par ce taux important de mortalité de cette tranche d'âge. Ce paradoxe est appelé *paradoxe de Simpson*

3. Utilisation de la régression logistique

Afin d'éviter un biais induit par des regroupements en tranches d'âges arbitraires et non régulières, nous envisageasons de réaliser une régression logistique. Si on introduit une variable Death valant 1 ou 0 pour indiquer si l'individu est décédé durant la période de 20 ans, on peut étudier le modèle Death ~ Age pour étudier la probabilité de décès en fonction de l'âge selon que l'on considère le groupe des fumeuses ou des non fumeuses.

Création de la variable Death

Pour créer cette variable, on utilise la fonction *if_else* du package de base.

```
data$Death <- ifelse(data$Status == "Dead", 1, 0)

#Vérification du codage
with(data, table(Status, Death))
```

```
##        Death
## Status    0   1
##   Alive 945   0
##   Dead    0 369
```

Estimation du modèle et représentation graphique

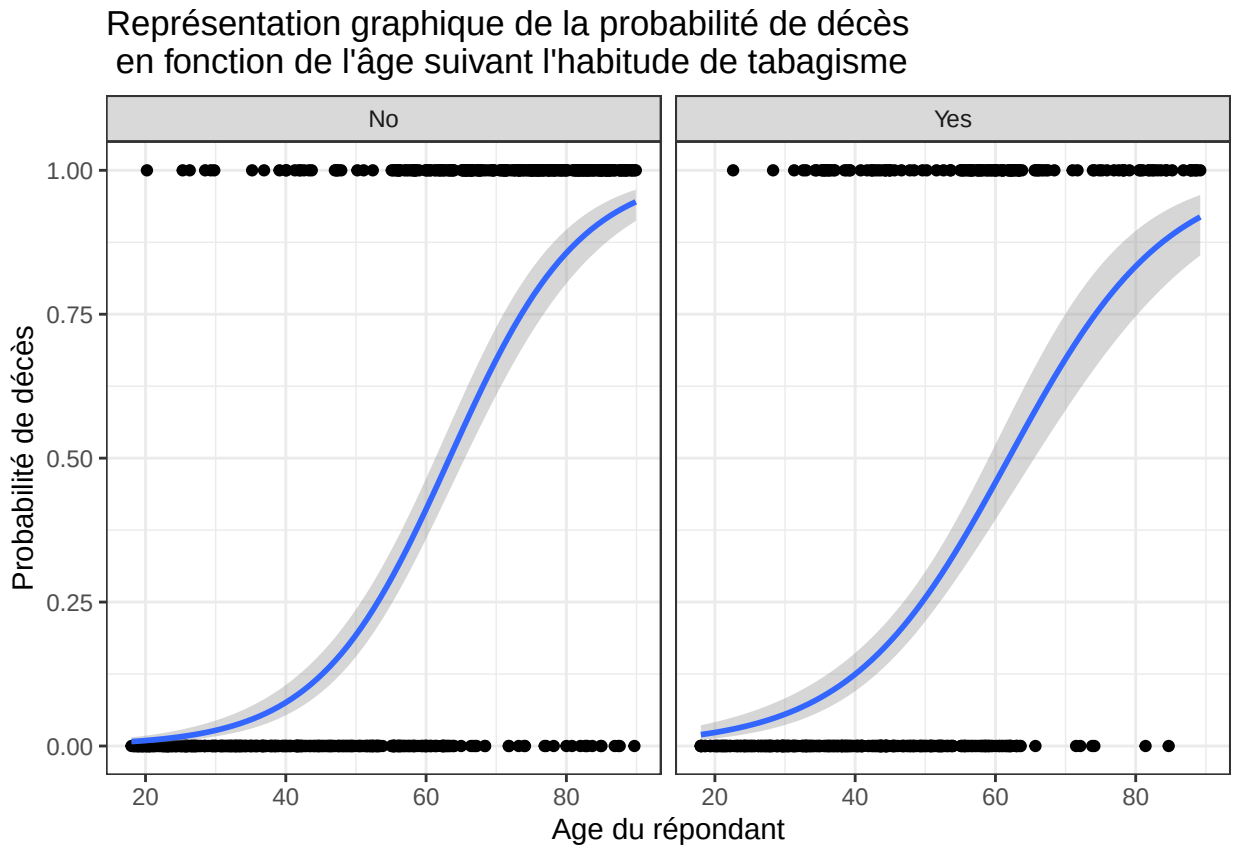
```
logistic_reg = glm(data=data, Death ~ Age,
                   family=binomial(link='logit'))
summary(logistic_reg)
```

```
##
## Call:
## glm(formula = Death ~ Age, family = binomial(link = "logit"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3335  -0.5897  -0.2848   0.4551   2.8803
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.104537   0.321414  -18.99   <2e-16 ***
## Age          0.097651   0.005555   17.58   <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: 1560.3  on 1313  degrees of freedom
## Residual deviance: 1004.8  on 1312  degrees of freedom
## AIC: 1008.8
##
## Number of Fisher Scoring iterations: 5
```

L'estimation du modèle montre que l'âge a effectivement un impact sur la probabilité de décès. L'augmentation de l'âge d'une unité influence positivement la probabilité de décès des femmes enquêtés indépendamment de

l'habitude de tabagisme. Le graphique suivant est relatif à la probabilité de décès en fonction de l'âge selon que l'on considère le groupe des fumeuses ou des non fumeuses.

```
ggplot(data, aes(x=Age, y=Death)) + geom_point() +
  stat_smooth(method="glm", method.args=list(family="binomial"), se=TRUE)+
  labs(title = "Représentation graphique de la probabilité de décès\n en fonction de l'âge suivant l'hab
       x = "Age du répondant",
       y = "Probabilité de décès")+
  facet_grid(~Smoker)+
  theme_bw()
```



Ce graphique montre que la probabilité de décès augmente en fonction de l'âge quel que soit l'habitude de tabagisme du répondant. La region de confiance est plus étroite pour les non-fumeuses que pour les fumeuses.