Projet de Biostatistiques : Cancer du poumon

Clara Gamberdello, Karla Pem

2026-01-08

Introduction

Dans le cadre d'un projet de biostatistique nous avons étudié un jeu de données provenant d'une étude cas témoins portant sur le cancer bronchique primitif, ou cancer du poumon. Nous considérons, les cas : patients présentant un cancer du poumon confirmé et les témoins : sujet sans cancer recrutés dans les mêmes établissements de soins. Nous cherchons a établir des liens entre le risque de cancer du poumon et différents facteurs tel que le tabagisme l'exposition a la fumée domestique ou professionnelle le sexe ou la bronchopneumopathie chronique entre autres.

Dans toute cette étude les intervalles de confiance et les conclusions statistique se feront avec une confiance de 95%.

Description du jeu de donnée

Le jeu de donnée contient 550 individus dont 230 cas et 320 témoins avec des informations sur 12 variables, à noté que les informations concernants les expositions a la fumée dans le milieu professionnel ou domestique ont été recueillis à l'aide d'un questionnaire standardisé complété lors d'un entretien préalable.

Nous observons les variables suivantes :

Les variables sont les suivantes :

- "statut_cas_temoin" : Précise le groupe de la personne Cas ou Témoin.
- "age", "sexe", "imc" : Donne l'age, le sexe et l'imc.
- "region" : Donne la région parmis ("Centre-Val-de-Loire", "Hauts-de-France", "Ile-de-France").
- "niveau_etudes" : Donne le niveau d'étude selon 3 modalités ("Primaire","Secondaire", "Supérieur").
- "tabagisme" : Donne le statut de l'individu sur le tabac selon 3 modalités ("Ancien fumeur", "Fumeur actuel", "Jamais").
- "exposition_professionnelle" : Donne l'exposition au tabac des individus sur leur lieu de travail selon 3 modalités ("Aucune exposition", "Exposition faible", "Exposition importante").
- "bronchopneumopathie_chronique" : Présence ou non d'une BPCO ("Oui", "Non").
- "exposition_domestique_fumee" : Donne l'exposition domestique à la fumée de tabac ("Oui", "Non").

Nous en visualisons les premières lignes :

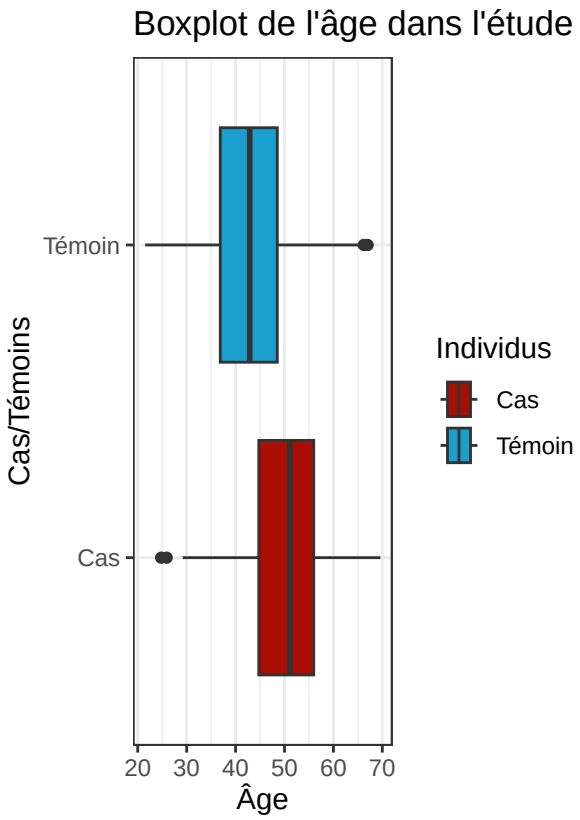
```r
CancerPoumon <- read.csv("CancerPoumongood.csv", sep=";")
head(CancerPoumon)
```

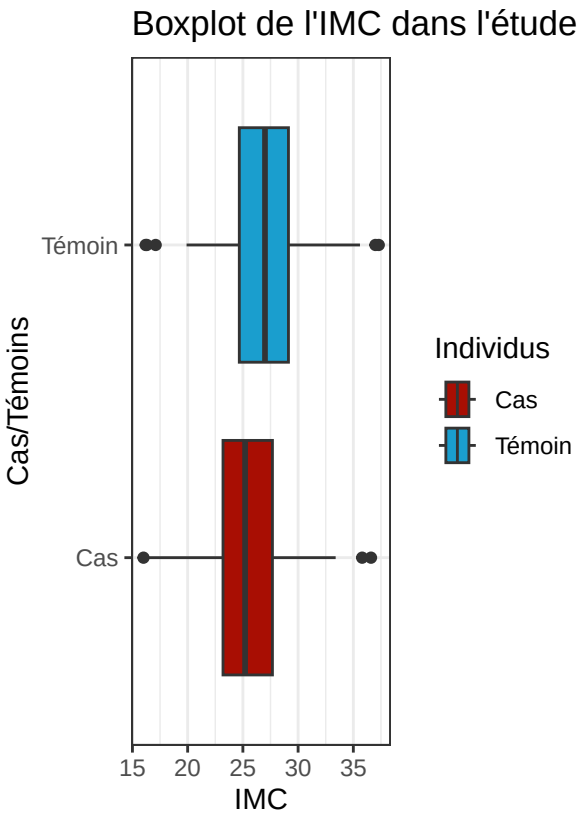
```
##   X id_sujet statut_cas_temoin  age  sexe     tabagisme
## 1 1        1               Cas 39,7 Homme Fumeur actuel
## 2 2        2               Cas 69,6 Femme Fumeur actuel
## 3 3        3               Cas 49,6 Homme        Jamais
## 4 4        4               Cas 56,4 Homme Fumeur actuel
## 5 5        5               Cas 53,7 Homme Fumeur actuel
## 6 6        6               Cas 65,5 Homme Fumeur actuel
##   exposition_professionnelle  imc bronchopneumopathie_chronique niveau_etudes
## 1      Exposition importante 19.8                           Oui      Primaire
## 2      Exposition importante 29.9                           Non    Secondaire
## 3          Aucune exposition 23.3                           Non      Primaire
## 4      Exposition importante 24.8                           Oui      Primaire
## 5          Exposition faible 31.4                           Non    Secondaire
## 6      Exposition importante 24.8                           Non     Supérieur
##   exposition_domestique_fumee             region
## 1                         Oui       Ile-de-France
## 2                         Oui Centre-Val-de-Loire
## 3                         Non Centre-Val-de-Loire
## 4                         Non     Hauts-de-France
## 5                         Non     Hauts-de-France
## 6                         Oui       Ile-de-France
```

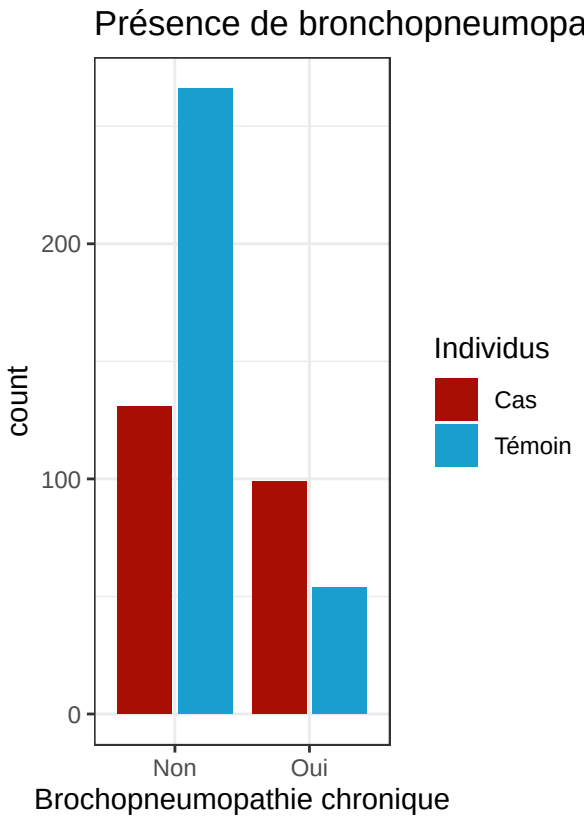
Statistiques descriptives

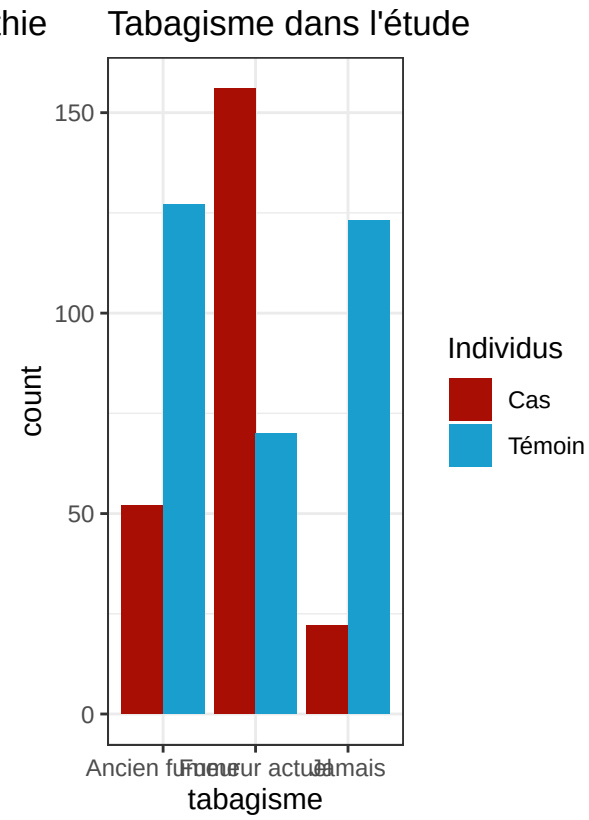
Nous prenons soin avant de décrire notre jeu de données de passer notre variable d'âge en numérique.

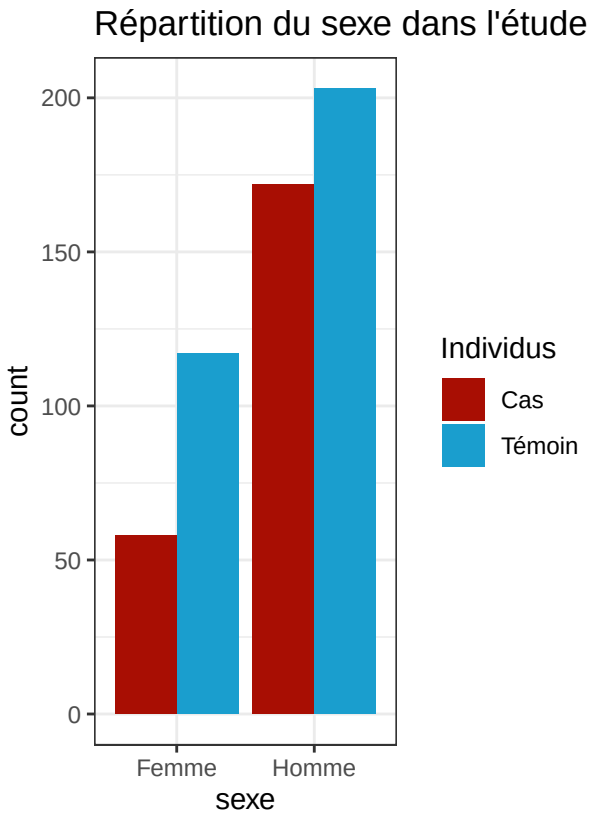
```r
CancerPoumon <- CancerPoumon %>%
  mutate(age=as.numeric(str_replace(age, ",", ".")))
```

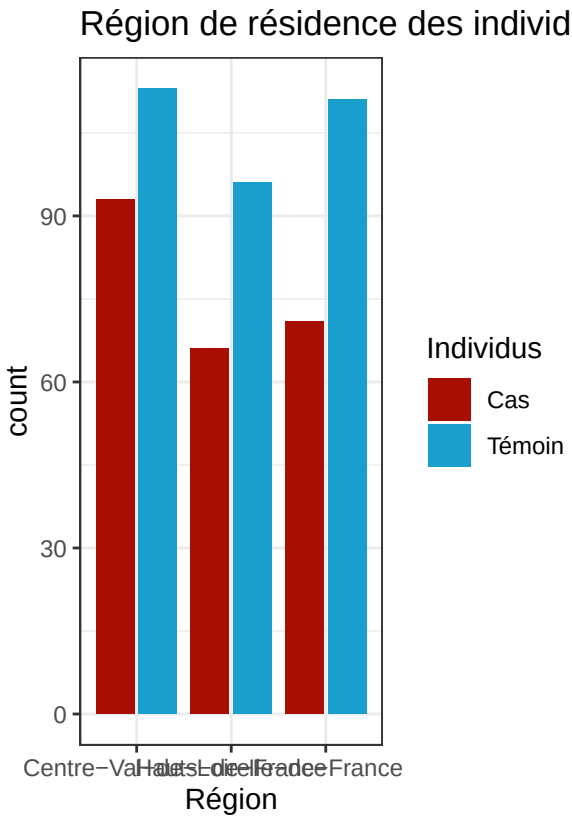

Boxplot de l'âge dans l'étude
Témoin
Cas
Cas/Témoins
20
30
40
50
60
70
Âge
Individus
Cas
Témoin

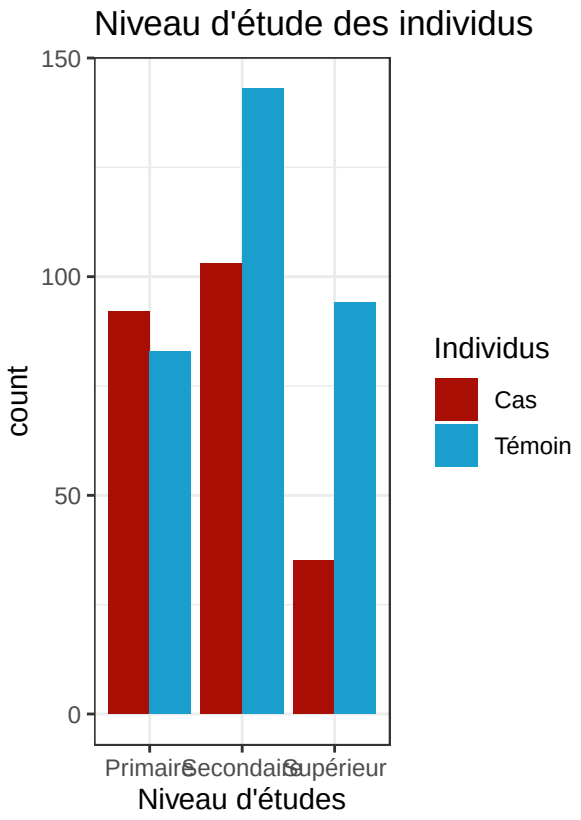

Boxplot de l'IMC dans l'étude
Témoin
Cas
Cas/Témoins
15
20
25
30
35
IMC
Individus
Cas
Témoin

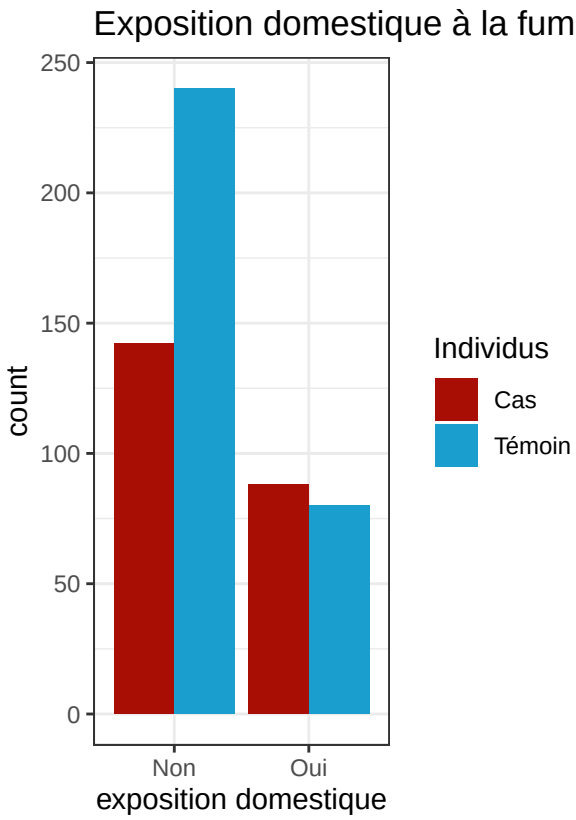

Présence de bronchopneumopathie
count
200
100
0
Non
Oui
Brochopneumopathie chronique
Individus
Cas
Témoin

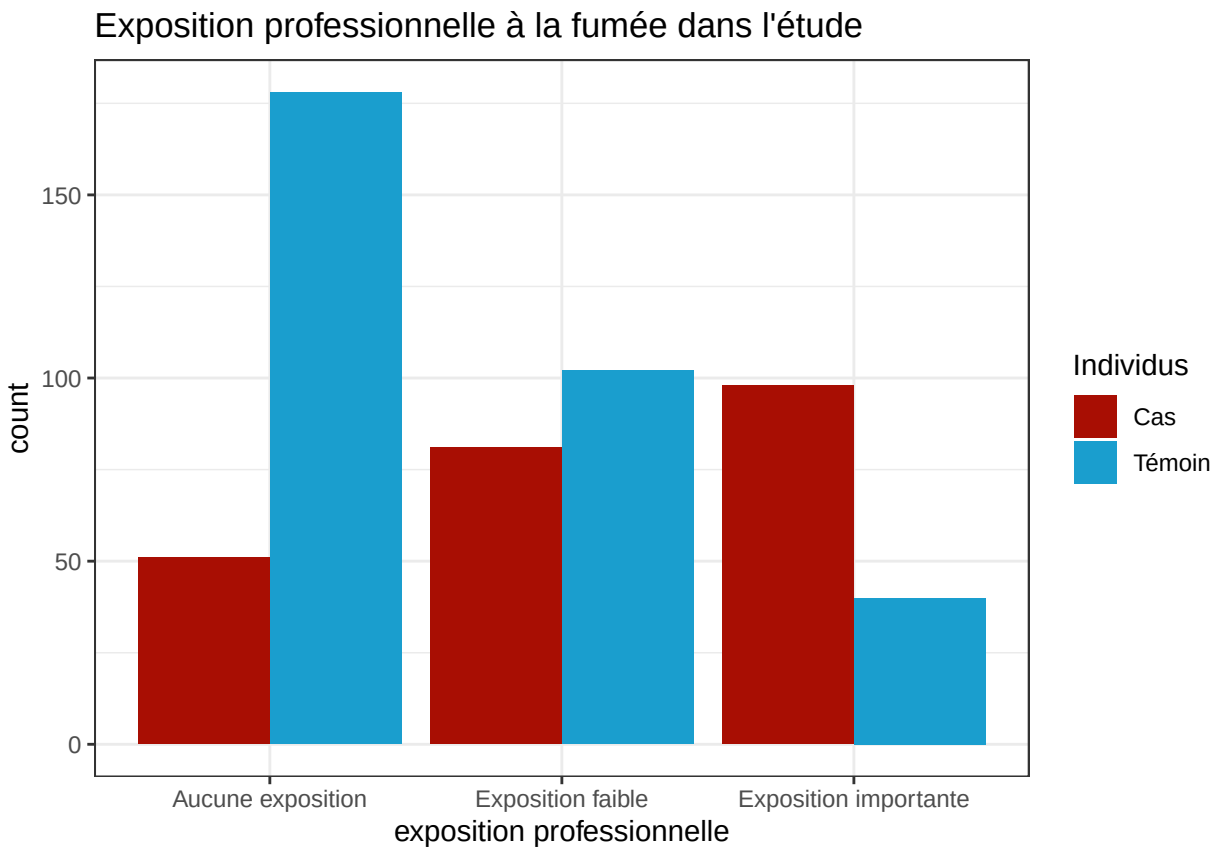

Tabagisme dans l'étude
count
150
100
50
0
Ancien fumeur
Fumeur actuel
Jamais
tabagisme
Individus
Cas
Témoin


Répartition du sexe dans l'étude
count
200
150
100
50
0
Femme
Homme
sexe
Individus
Cas
Témoin


Région de résidence des individus
count
90
60
30
0
Région
Individus
Cas
Témoin


Niveau d'étude des individus
150
100
50
0
count
Primaire
Secondaire
Supérieur
Niveau d'études
Individus
Cas
Témoin


Exposition domestique à la fumée
250
200
150
100
50
0
count
Non
Oui
exposition domestique
Individus
Cas
Témoin



Nous réalisons donc différents barplots et boxplots pour nos variables. Nous constatons que les individus "cas" sont en moyenne plus âgés que les témoins, en revanche leur IMC est légèrement plus faible en moyenne que celle des témoins. Nous consatons également que le nombre de cas et de témoins dans les différentes régions semble suivre une proportion à peu près équivalente. En ce qui concerne la bronchopneumopathie, on peut constater que les personnes atteintes de cette pathologie sont plus généralement des cas et à l'inverse les personnes ne souffrant pas de cette maladie sont plus nombreux dans le groupe des témoins. Cet étude contient plus d'homme que de femme, on pense remarquer qu'il y a plus d'homme atteint d'un cancer que de femme par rapport a leur totaux. Nous nous interessons également à la variable tabagisme, on constate ici très nettement que le nombre de cas chez les personnes n'ayant jamais fumés est bien plus faible que pour les anciens fumeurs et les fumeurs actuel. Il semble y avoir plus de cas chez les fumeurs actuels que chez les anciens fumeurs. Nous observons que les individus ayant une faible exposition professionnelle à la fumée ou pas d'exposition domestique sont plus nombreux chez les témoins, en revanche parmi les individus qui ont un cancer plus nombreux sont ceux qui ont une exposition importante à la fumée. De manière générale, pour les cas et les témoins les personnes n'ayant pas d'exposition domestique à la fumée sont plus nombreuse. Enfin le nombre de cas ayant un niveau d'étude supérieur est plus faible que les individus avec niveau d'études primaires ou secondaires.

Analyse

Régression logistique

On appelle modèle logistique le modèle défini par :

$Y \in \{0,1\}, \quad X \in \mathbb{R}^p, \quad \exists \beta \in \mathbb{R}^p$, la vraisemblance conditionnelle de Y sachant X=x est :

$$L(Y \mid X = x) = \mathcal{B}\Big(F(\langle \beta, x \rangle)\Big),$$

et on a :

$$f_{\beta,x}(y) = F(\langle \beta, x \rangle)^y \left(1 - F(\langle \beta, x \rangle)\right)^{1-y}, \quad y \in \{0, 1\},$$

où F est la fonction logistique :

$$F(t) = \frac{e^t}{1 + e^t} = \frac{1}{1 + e^{-t}}.$$

Interprétation des coefficients :

On définit le log-odds comme :

$$\ln\Big(\text{odds}\big(\mathbb{P}(Y = 1 \mid X = x)\big)\Big) = \langle \beta, x \rangle$$

ou, $odds(A) = \frac{\mathbb{P}(A)}{1-\mathbb{P}(A)}$.

Chaque coefficient β_j quantifie la variation du log-odds de Y=1 lorsque la variable x_j augmente d'une unité, toutes les autres variables étant maintenues constantes :

Variation du log-odds : $\ln(\text{odds}(Y = 1 \mid x_j + 1)) - \ln(\text{odds}(Y = 1 \mid x_j)) = \beta_j$

Pour une interprétation plus intuitive, on peut calculer e^{β_j} qui correspond à l'odds ratio, il représente le facteur multiplicatif appliqué aux odds de Y=1 pour une augmentation de x_j.

Pour pouvoir réaliser nos modèles nous changeons la variable statut_cas_temoins pour la recoder en variable binaire :

```
CancerPoumon$statut_cas_temoin <- ifelse(CancerPoumon$statut_cas_temoin=="Cas", 1, 0)
```

Modèle univarié

Tabagisme

Nous réalisons un modèle univarié avec la variable qualitative tabagisme, en effet c'est la variable qui nous parait être la plus logique dans le risque d'appartition du cancer.

```
# Logistique cancer/tabac
reslt <- glm(statut_cas_temoin ~ tabagisme, family=binomial, data=CancerPoumon)
summary(reslt)
```

```
##
## Call:
## glm(formula = statut_cas_temoin ~ tabagisme, family = binomial,
##     data = CancerPoumon)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -0.8929     0.1646  -5.424 5.84e-08 ***
## tabagismeFumeur actuel    1.6943     0.2186   7.750 9.22e-15 ***
## tabagismeJamais          -0.8282     0.2841  -2.916  0.00355 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 747.67  on 549  degrees of freedom
## Residual deviance: 618.92  on 547  degrees of freedom
## AIC: 624.92
##
## Number of Fisher Scoring iterations: 4
```

```
confint(reslt)
```

```
## Waiting for profiling to be done...
```

```
##                              2.5 %     97.5 %
## (Intercept)            -1.223665 -0.5768201
## tabagismeFumeur actuel  1.271652  2.1296074
## tabagismeJamais        -1.400675 -0.2828631
```

Nous observons nos résultats avec la modalité ancen fumeur comme modalité de référence.

Modalité "Fumeur actuel" :

$$1.272 \leq ln\Big(\frac{odds(statut\_cas\_temoin = 1 \mid tabagisme = Fumeur\_actuel)}{odds(statut\_cas\_temoin = 1 \mid tabagisme = Ancien\_fumeur)}\Big) \leq 2.130$$

Etre fumeur actuel, plutôt qu'ancien fumeur, multiplie l'odds d'avoir un cancer du poumon par au moins $exp(1.272) = 3.57$ et au plus $exp(2.130) = 8.41$.

Modalité "Jamais" :

$$-1.401 \leq ln\Big(\frac{odds(statut\_cas\_temoin = 1 \mid tabagisme = Jamais)}{odds(statut\_cas\_temoin = 1 \mid tabagisme = Ancien\_fumeur)}\Big) \leq -0.283$$

N'avoir jamais fumé, plutôt qu'être ancien fumeur, divise l'odds d'avoir un cancer du poumon par au moins $exp(-1.401) = 0.25$ et au plus $exp(-0.283) = 0.75$.

Nous pouvons donc en conclure que le tabagisme est un facteur de risque d'apparition d'un cancer du poumon.

IMC

Au vu du boxplot obtenue nous réalisons un modèle univarié avec IMC comme variable explicative, nous affichons également l'intervalle de confiance associé.

```
# Logistique cancer/imc
reslt_imc <- glm(statut_cas_temoin~imc, family= binomial, data=CancerPoumon)
summary(reslt_imc)
```

```
##
## Call:
## glm(formula = statut_cas_temoin ~ imc, family = binomial, data = CancerPoumon)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.22728    0.69280   4.658 3.19e-06 ***
```

```
## imc         -0.13577    0.02636  -5.151 2.59e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 747.67  on 549  degrees of freedom
## Residual deviance: 718.75  on 548  degrees of freedom
## AIC: 722.75
##
## Number of Fisher Scoring iterations: 4
```

```
confint(reslt_imc)
```

```
## Waiting for profiling to be done...
```

```
##                  2.5 %     97.5 %
## (Intercept)  1.8918118  4.6120720
## imc         -0.1885505 -0.0850612
```

$$-0.189 \leq ln\Big(\frac{odds(statut\_cas\_temoin = 1 \mid imc = x+1)}{odds(statut\_cas\_temoin = 1 | imc = x)}\Big) \leq -0.085$$

Une unité de plus sur l'imc, divise l'odds d'avoir un cancer du poumon par au moins $exp(-0.189) = 0.83$ et au plus $exp(-0.085) = 0.92$.

Ainsi le fait d'avoir un imc plus élevé semble avoir un effet protecteur sur le risque d'apparition de la maladie.

Exposition professionnelle à la fumée

Nous observons ensuite le modèle univarié avec la variable exposition professionnelle.

```
# Logistique cancer/exposition pro
reslt_exp <- glm(statut_cas_temoin~exposition_professionnelle, family= binomial, data=CancerPoumon)
summary(reslt_exp)
```

```
##
## Call:
## glm(formula = statut_cas_temoin ~ exposition_professionnelle,
##     family = binomial, data = CancerPoumon)
##
## Coefficients:
##                                                Estimate Std. Error z value
## (Intercept)                                     -1.2500     0.1588  -7.870
## exposition_professionnelleExposition faible      1.0194     0.2177   4.684
## exposition_professionnelleExposition importante  2.1460     0.2458   8.730
##                                                Pr(>|z|)
## (Intercept)                                    3.55e-15 ***
## exposition_professionnelleExposition faible    2.82e-06 ***
## exposition_professionnelleExposition importante  < 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: 747.67  on 549  degrees of freedom
## Residual deviance: 660.32  on 547  degrees of freedom
## AIC: 666.32
##
## Number of Fisher Scoring iterations: 4
```

```
confint(reslt_exp)
```

```
## Waiting for profiling to be done...
```

```
##                                                 2.5 %     97.5 %
## (Intercept)                                -1.5709339 -0.9469576
## exposition_professionnelleExposition faible  0.5962861  1.4506327
## exposition_professionnelleExposition importante  1.6728463  2.6378173
```

Nous analysons selon la variable de référence “aucune exposition”.

Modalité : Exposition faible

$$0.596 \leq ln(\frac{odds(statut\_cas\_temoin = 1 \mid Exposition\_pro = faible)}{odds(statut\_cas\_temoin = 1 \mid exposition\_pro = aucune)}) \leq 1.450$$

Avoir une exposition professionnelle faible à la fumée plutôt que d’être exposée à aucune fumée, multiplie l’odds d’avoir un cancer du poumon par au moins $exp(0.596) = 1.81$ et au plus $exp(1.450) = 4.26$.

Modalité : Exposition importante

$$1.672 \leq ln\Big(\frac{odds(statut\_cas\_temoin = 1 \mid Exposition\_pro = importante)}{odds(statut\_cas\_temoin = 1 \mid exposition\_pro = aucune)}\Big) \leq 2.637$$

Avoir une exposition professionnelle importante à la fumée plutôt que d’être exposée à aucune fumée, multiplie l’odds d’avoir un cancer du poumon par au moins $exp(1.672) = 5.32$ et au plus $exp(2.637) = 13.97$.

Ce que nous attendions se vérifie, l’exposition professionnelle à la fumée semble être un facteur de risque de cancer du poumon.

Exposition domestique à la fumée

Enfin nous nous proposons de faire un modèle avec la variable exposition domestique à la fumée.

```
# Logistique cancer/exposition domestique
reslt_exd <- glm(statut_cas_temoin~exposition_domestique_fumee, family= binomial, data=CancerPoumon)
summary(reslt_exd)
```

```
##
## Call:
## glm(formula = statut_cas_temoin ~ exposition_domestique_fumee,
##     family = binomial, data = CancerPoumon)
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)                      -0.5248     0.1059  -4.957 7.16e-07 ***
## exposition_domestique_fumeeOui    0.6201     0.1873   3.311 0.000929 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 747.67  on 549  degrees of freedom
## Residual deviance: 736.66  on 548  degrees of freedom
## AIC: 740.66
##
## Number of Fisher Scoring iterations: 4
```

```
confint(reslt_exd)
```

```
## Waiting for profiling to be done...
```

```
##                                     2.5 %     97.5 %
## (Intercept)                    -0.7343546 -0.3189611
## exposition_domestique_fumeeOui  0.2537883  0.9886250
```

Modalité : Exposition importante

$$0.253 \leq ln\Big(\frac{odds(statut\_cas\_temoin = 1 \mid Exposition\_domestique = importante)}{odds(statut\_cas\_temoin = 1 \mid exposition\_domestique = aucune)}\Big) \leq 0.988$$

Avoir une exposition domestique importante à la fumée plutôt que d'être exposée à aucune fumée, multiplie l'odds d'avoir un cancer du poumon par au moins $exp(0.253) = 1.287$ et au plus $exp(0.988) = 2.68$.

Comme pour l'exposition professionnelle, l'exposition domestique semble être un facteur de risque de cancer du poumon mais de manière moins importante.

```
reslt_etu <- glm(statut_cas_temoin ~ niveau_etudes, family= binomial, data=CancerPoumon)
summary(reslt_etu)
```

Niveau d'étude

```
##
## Call:
## glm(formula = statut_cas_temoin ~ niveau_etudes, family = binomial,
##     data = CancerPoumon)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 0.1029     0.1514   0.680   0.4965
## niveau_etudesSecondaire    -0.4311     0.1990  -2.166   0.0303 *
## niveau_etudesSupérieur     -1.0909     0.2493  -4.377 1.21e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 747.67  on 549  degrees of freedom
## Residual deviance: 727.45  on 547  degrees of freedom
## AIC: 733.45
##
## Number of Fisher Scoring iterations: 4
```

```
confint(reslt_etu)
```

```
## Waiting for profiling to be done...
```

```
##                               2.5 %      97.5 %
## (Intercept)             -0.1935535  0.40096430
## niveau_etudesSecondaire -0.8226953 -0.04177522
## niveau_etudesSupérieur  -1.5878876 -0.60904900
```

En interprétant comme précédemment, on observe que le fait d'avoir un niveau d'étude supérieur par rapport à un niveau d'étude primaire semble diminuer significativement l'odds d'avoir un cancer, de même, avoir un niveau d'étude secondaire par rapport à un niveau primaire diminue également l'odds, mais de façon moins significative.

```
reslt_age <- glm(statut_cas_temoin~age, family= binomial, data=CancerPoumon)
summary(reslt_age)
```

Age

```
##
## Call:
## glm(formula = statut_cas_temoin ~ age, family = binomial, data = CancerPoumon)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.44919    0.58659  -9.290   <2e-16 ***
## age          0.10954    0.01226   8.935   <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: 747.67  on 549  degrees of freedom
## Residual deviance: 645.15  on 548  degrees of freedom
## AIC: 649.15
##
## Number of Fisher Scoring iterations: 3
```

```
confint(reslt_age)
```

```
## Waiting for profiling to be done...
```

```
##                  2.5 %     97.5 %
## (Intercept) -6.6373664 -4.3345574
## age          0.0862185  0.1343482
```

Ici de nouveau avec les mêmes interprétations que plus haut, nous pouvons conclure qu'augementer l'age de 1 ans multiplie l'odds d'avoir un cancer du poumon d'au moins $e^{0.08} = 1.08$ et d'au plus $e^{0.13} = 1.14$.

```
reslt_bron <- glm(statut_cas_temoin~bronchopneumopathie_chronique, family= binomial, data=CancerPoumon)
summary(reslt_bron)
```

Bronchopneumopathie chronique

```
##
## Call:
## glm(formula = statut_cas_temoin ~ bronchopneumopathie_chronique,
##     family = binomial, data = CancerPoumon)
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -0.7083     0.1067  -6.636 3.23e-11 ***
## bronchopneumopathie_chroniqueOui   1.3144     0.2000   6.571 4.99e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 747.67  on 549  degrees of freedom
## Residual deviance: 702.19  on 548  degrees of freedom
## AIC: 706.19
##
## Number of Fisher Scoring iterations: 4
```

```
confint(reslt_bron)
```

```
## Waiting for profiling to be done...
```

```
##                                     2.5 %     97.5 %
## (Intercept)                     -0.9201874 -0.5013846
## bronchopneumopathie_chroniqueOui  0.9264469  1.7115845
```

Nous obtenons des p-value inférieure à 0.05 de plus 0 n'est pas contenu dans l'intervalle de confiance la bronchopneumopathie semble être un facteur de risque du cancer du poumon.

```
reslt_reg <- glm(statut_cas_temoin~region, family= binomial, data=CancerPoumon)
summary(reslt_reg)
```

Région

```
##
## Call:
## glm(formula = statut_cas_temoin ~ region, family = binomial,
##     data = CancerPoumon)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -0.1948     0.1400  -1.391    0.164
## regionHauts-de-France  -0.1799     0.2125  -0.846    0.397
## regionIle-de-France    -0.2521     0.2066  -1.220    0.223
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 747.67  on 549  degrees of freedom
## Residual deviance: 746.07  on 547  degrees of freedom
## AIC: 752.07
##
## Number of Fisher Scoring iterations: 4
```

```
confint(reslt_reg)
```

```
## Waiting for profiling to be done...
```

```
##                            2.5 %     97.5 %
## (Intercept)           -0.4708557 0.07883374
## regionHauts-de-France -0.5980802 0.23588192
## regionIle-de-France   -0.6586355 0.15210418
```

Ici nous obtenons des p-value élevées et 0 est contenu dans les intervalles de confiance.

Nous avons réalisés des modèles univariés qui nous ont permis d'étudier séparément la potentielle association entre les facteurs de risques possibles et l'apparition d'un cancer du poumon. Cependant, il est possible que ces associations soit influencés par des facteurs dit confondant, autrement dit des variables pouvant influencer les variables explicatives et la variable à expliquer. Nous allons donc créer des modèles multivariés.

Modèles multivariés

Avec les variables quantitatives

On va maintenant tester le modèle additif avec les variables age et imc. On testera aussi le modèle avec un effet d'interaction entre age et imc.

```
# Modèle age + imc :
res_log_quanti3 <- glm(statut_cas_temoin ~ age + imc, family=binomial, data=CancerPoumon)
summary(res_log_quanti3)
```

```
##
## Call:
## glm(formula = statut_cas_temoin ~ age + imc, family = binomial,
##     data = CancerPoumon)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.06930    0.93583  -2.211    0.027 *
## age          0.10748    0.01246   8.627  < 2e-16 ***
## imc         -0.12509    0.02838  -4.407 1.05e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 747.67  on 549  degrees of freedom
## Residual deviance: 624.42  on 547  degrees of freedom
## AIC: 630.42
##
## Number of Fisher Scoring iterations: 4
```

```
# Modèle avec age*imc :
res_log_quanti4 <- glm(statut_cas_temoin ~ age*imc, family=binomial, data=CancerPoumon)
summary(res_log_quanti4)
```

```
##
## Call:
## glm(formula = statut_cas_temoin ~ age * imc, family = binomial,
##     data = CancerPoumon)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.885968   4.713664  -1.249   0.2118
## age          0.188554   0.099182   1.901   0.0573 .
## imc          0.021130   0.178431   0.118   0.9057
## age:imc     -0.003104   0.003751  -0.828   0.4079
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 747.67  on 549  degrees of freedom
## Residual deviance: 623.73  on 546  degrees of freedom
## AIC: 631.73
##
## Number of Fisher Scoring iterations: 4
```

```
# Intervalles pour age + imc :
confint(res_log_quanti3)
```

```
## Waiting for profiling to be done...

##                   2.5 %      97.5 %
```

```
## (Intercept) -3.91956682 -0.24388292
## age          0.08376821  0.13267951
## imc         -0.18184085 -0.07036777
```

$ln(odds(statut\_cas\_temoin = 1 \mid age = x_1,\, imc = x_2)) = \mu + \beta_{age} \times x_1 + \beta_{imc} \times x_2$

Le modèle avec interraction entre age et imc donne des résultats non significatifs ce n'est pas le cas du modèle additif que nous allons donc interpréter :

Les intervalles de confiance obtenus sont les suivant : $\beta_{age} \in [0.084, 0.133]$ $\beta_{imc} \in [-0.182, -0.070]$

Interpretation :

$$0.084 \leq ln\Big(\frac{odds(statut\_cas\_temoin = 1 \mid age = x_1 + 1,\, imc = x_2)}{odds(statut\_cas\_temoin = 1 \mid age = x_1,\, imc = x_2)}\Big) \leq 0.133$$

A imc fixé, avoir un âge plus élevé de 1 unité (1 ans), va multiplier l'odds d'avoir un cancer du poumon par au moins $e^{0.084} = 1.088$ et au plus $e^{0.133} = 1.142$.

$$-0.182 \leq ln\Big(\frac{odds(statut\_cas\_temoin = 1 \mid age = x_1,\, imc = x_2 + 1)}{odds(statut\_cas\_temoin = 1 \mid age = x_1,\, imc = x_2)}\Big) \leq -0.070$$

A age fixé, avoir un imc plus élevé de 1 unité, va diviser l'odds d'avoir un cancer du poumon par au moins $e^{0.070} = 1.072$ et au plus $e^{0.182} = 1.20$.

Un patient avec un âge plus élevé augmente significativement l'odds d'avoir un cancer du poumon, tandis qu'un patient avec un IMC plus élevé diminue significativement cet odds.

Avec toutes les variables

Nous souhaitons trouvé avec le critère AIC le meilleur modèle. Nous réalisons donc un modèle complet avec toutes les variables.

```
modele_complet <- glm(statut_cas_temoin ~ age + sexe + tabagisme + exposition_professionnelle + broncho

stepAIC(modele_complet, direction="backward")
```

```
## Start:  AIC=457.47
## statut_cas_temoin ~ age + sexe + tabagisme + exposition_professionnelle +
##     bronchopneumopathie_chronique + niveau_etudes + exposition_domestique_fumee +
##     region + imc
##
##                                 Df Deviance    AIC
## - region                         2   432.81 456.81
## <none>                               429.47 457.47
## - exposition_domestique_fumee    1   432.93 458.93
## - sexe                           1   432.97 458.97
## - niveau_etudes                  2   437.52 461.52
## - imc                            1   444.60 470.60
## - bronchopneumopathie_chronique  1   456.61 482.61
## - exposition_professionnelle     2   477.54 501.54
## - age                            1   476.57 502.57
## - tabagisme                      2   505.91 529.91
##
## Step:  AIC=456.81
```

```
## statut_cas_temoin ~ age + sexe + tabagisme + exposition_professionnelle +
##     bronchopneumopathie_chronique + niveau_etudes + exposition_domestique_fumee +
##     imc
##
##                                 Df Deviance    AIC
## <none>                               432.81 456.81
## - exposition_domestique_fumee    1   435.55 457.55
## - sexe                           1   436.46 458.46
## - niveau_etudes                  2   442.45 462.45
## - imc                            1   448.01 470.01
## - bronchopneumopathie_chronique  1   459.68 481.68
## - exposition_professionnelle     2   480.29 500.29
## - age                            1   480.85 502.85
## - tabagisme                      2   508.07 528.07
```

```
##
## Call:  glm(formula = statut_cas_temoin ~ age + sexe + tabagisme + exposition_professionnelle +
##     bronchopneumopathie_chronique + niveau_etudes + exposition_domestique_fumee +
##     imc, family = "binomial", data = CancerPoumon)
##
## Coefficients:
##                                     (Intercept)
##                                        -3.68867
##                                             age
##                                         0.09543
##                                       sexeHomme
##                                         0.49932
##                           tabagismeFumeur actuel
##                                         1.66833
##                                 tabagismeJamais
##                                        -0.79361
##     exposition_professionnelleExposition faible
##                                         1.28121
## exposition_professionnelleExposition importante
##                                         2.02335
##                bronchopneumopathie_chroniqueOui
##                                         1.37923
##                          niveau_etudesSecondaire
##                                         0.16310
##                           niveau_etudesSupérieur
##                                        -0.85726
##                   exposition_domestique_fumeeOui
##                                         0.42858
##                                             imc
##                                        -0.13065
##
## Degrees of Freedom: 549 Total (i.e. Null);  538 Residual
## Null Deviance:       747.7
## Residual Deviance: 432.8     AIC: 456.8
```

La fonction stepAIC ne retire que la variable région avant de s'arrêter. Nous observons donc le modèle avec toutes les variables sans région.

```r
modele_complet_2 <- glm(statut_cas_temoin ~ age + sexe + tabagisme + exposition_professionnelle + bronch
summary(modele_complet_2)
```

```
##
## Call:
## glm(formula = statut_cas_temoin ~ age + sexe + tabagisme + exposition_professionnelle +
##     bronchopneumopathie_chronique + niveau_etudes + exposition_domestique_fumee +
##     imc, family = "binomial", data = CancerPoumon)
##
## Coefficients:
##                                              Estimate Std. Error z value
## (Intercept)                                 -3.68867    1.22275  -3.017
## age                                          0.09543    0.01492   6.398
## sexeHomme                                    0.49932    0.26290   1.899
## tabagismeFumeur actuel                       1.66833    0.27973   5.964
## tabagismeJamais                             -0.79361    0.34607  -2.293
## exposition_professionnelleExposition faible  1.28121    0.28989   4.420
## exposition_professionnelleExposition importante  2.02335    0.31953   6.332
## bronchopneumopathie_chroniqueOui             1.37923    0.27473   5.020
## niveau_etudesSecondaire                      0.16310    0.27413   0.595
## niveau_etudesSupérieur                      -0.85726    0.35894  -2.388
## exposition_domestique_fumeeOui               0.42858    0.25920   1.653
## imc                                         -0.13065    0.03448  -3.790
##                                              Pr(>|z|)
## (Intercept)                                  0.002555 **
## age                                          1.57e-10 ***
## sexeHomme                                    0.057529 .
## tabagismeFumeur actuel                       2.46e-09 ***
## tabagismeJamais                              0.021836 *
## exposition_professionnelleExposition faible  9.89e-06 ***
## exposition_professionnelleExposition importante 2.42e-10 ***
## bronchopneumopathie_chroniqueOui             5.16e-07 ***
## niveau_etudesSecondaire                      0.551872
## niveau_etudesSupérieur                       0.016927 *
## exposition_domestique_fumeeOui               0.098235 .
## imc                                          0.000151 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 747.67  on 549  degrees of freedom
## Residual deviance: 432.81  on 538  degrees of freedom
## AIC: 456.81
##
## Number of Fisher Scoring iterations: 5
```

```r
confint(modele_complet_2)
```

```
## Waiting for profiling to be done...
```

```
##                                                 2.5 %      97.5 %
```

```
## (Intercept)                                        -6.11919460 -1.31404899
## age                                                 0.06700752  0.12561510
## sexeHomme                                          -0.01254765  1.02035803
## tabagismeFumeur actuel                              1.12948744  2.22846876
## tabagismeJamais                                    -1.48730911 -0.12596300
## exposition_professionnelleExposition faible         0.72216884  1.86134111
## exposition_professionnelleExposition importante     1.40956446  2.66504811
## bronchopneumopathie_chroniqueOui                    0.84924308  1.92881302
## niveau_etudesSecondaire                            -0.37205528  0.70473675
## niveau_etudesSupérieur                             -1.57176734 -0.16094188
## exposition_domestique_fumeeOui                     -0.07891130  0.93927949
## imc                                                -0.19973623 -0.06421259
```

Nous constatons ici que toutes choses égales par ailleurs les variables age, tabagisme exposition professionnelle niveau d'étude imc et bronchopneumopathie chronique sont séparément significatives nous avons :

$$0.883 \leq ln\Big(\frac{odds(statut\_cas\_temoin = 1 \mid brochopneumo = oui,\, age,\, sexe,\, tabagisme,\, expo\_pro,\ldots\ )}{odds(statut\_cas\_temoin = 1 \mid bronchopneumo = non,\, age,\, sexe,\, tabagisme,\, expo\_pro,\ldots\ )}\Big) \leq 1.942$$

Toutes choses égales par ailleurs, souffrir de bronchopneumopathie chronique multiplie l'odds d'avoir un cancer du poumon par au moins $exp(0.883) = 2.41$ et au plus $exp(1.942) = 6.97$.

Les autres intervalles de confiances peuvent être interprétés comme précédemment.

Dans ce modèle, l'âge est significativement associé à une augmentation du risque, indiquant que la probabilité de survenue de l'événement augmente avec l'avancée en âge. Le tabagisme actuel apparaît comme un facteur de risque important, de même que l'exposition professionnelle, qu'elle soit faible ou importante, cette dernière montrant un effet particulièrement marqué. La présence d'une bronchopneumopathie chronique est également associée à une augmentation significative du risque. À l'inverse, n'avoir jamais fumé est associé à une diminution significative du risque, suggérant un effet protecteur, tout comme le fait d'avoir un niveau d'études supérieur. L'imc est également associé à une diminution significative du risque dans ce modèle, indiquant un effet protecteur. En revanche, le sexe masculin, le niveau d'études secondaire et l'exposition domestique à la fumée ne sont pas significativement associés au risque, leurs intervalles de confiance incluant la valeur nulle. On remarque donc au vu des modèles univariés que l'exposition domestique ou le niveau d'étude secondaire sont des facteurs confondants.

Régression polytomique ordonné

Dans cette partie, nous nous sommes permis d'inverser le modèle, c'est-à-dire de considérer la variable statut_cas_temoin comme variable explicative, afin de pouvoir traiter des variables catégorielles ordinales comme variables à expliquer. Cela est possible ici, car l'estimation des paramètres dans le modèle $Y \sim E, F$ est la même que dans le modèle $E \sim Y, F$.

On appelle modèle polytomique ordonnée le modèle défini par :

$X \in \mathbb{R}^p$ et Y est qualitative ordonnée de modalités $m_0 < m_1 < \ldots < m_q$, on a $(c_1, \ldots, c_{q-1}) \in \mathbb{R}^{q-1}$ tq $\forall x \in \mathbb{R}^p$ et $\forall k = 0, \ldots, q-1$:

$$\mathbb{P}(Y = m_k \mid X = x) = F(c_{k+1} - < \beta, x >) - F(c_k - < \beta, x >)$$

ou F est toujours la fonction logistique.

Ce modèle est obtenu par une modèlisation par variable latente c'est-à-dire $Y = m_k \Leftrightarrow Y^* \in]c_k, c_{k+1}]$, avec $c_0 = -\infty$ et $c_{q+1} = +\infty$. Une conséquence importante d'une modèlisation par variable latente est que peut importe la modalité m_k choisit, l'interpretation sera la même !

Car on a pour odds-ratio :

$$\frac{\text{odds}(Y \leq m_k \mid X = x_j + 1)}{\text{odds}(Y \leq m_k \mid X = x_j)} = e^{-\beta_j}$$

et $e^{-\beta_j}$ ne dépend pas de m_k.

Il faut également faire attention lors de l'interprétation des coefficients a ne pas oublié le "-" devant le logarithme de l'odds ratio.

Meilleur modèle avec l'AIC sur Exposition professionnelle

On va maintenant regarder un modèle avec toute les variables de la table, sauf les variables "X" et "id_sujet" qui sont les 2 même et qui ne sont pas intéressante. Pour choisir le meilleur modèle on utilise la fonction "stepAIC", pour prédire "exposition_professionnelle" qui a pour modalités "Aucune exposition" < "Exposition faible" < "Exposition importante".

```
CancerPoumon$statut_cas_temoin <- ifelse(CancerPoumon$statut_cas_temoin=="1", "Cas", "temoin")
res_poly <- polr(factor(exposition_professionnelle) ~ statut_cas_temoin + age + sexe + tabagisme + imc +

stepAIC(res_poly, direction="backward")
```

```
## Start:  AIC=1118.13
## factor(exposition_professionnelle) ~ statut_cas_temoin + age +
##     sexe + tabagisme + imc + bronchopneumopathie_chronique +
##     niveau_etudes + exposition_domestique_fumee + region
##
##                                 Df    AIC
## - tabagisme                      2 1115.0
## - niveau_etudes                  2 1115.0
## - sexe                           1 1116.2
## - age                            1 1116.5
## - imc                            1 1116.5
## - bronchopneumopathie_chronique  1 1116.9
## - region                         2 1117.8
## <none>                             1118.1
## - exposition_domestique_fumee    1 1118.4
## - statut_cas_temoin              1 1162.8
##
## Step:  AIC=1114.99
## factor(exposition_professionnelle) ~ statut_cas_temoin + age +
##     sexe + imc + bronchopneumopathie_chronique + niveau_etudes +
##     exposition_domestique_fumee + region
##
##                                 Df    AIC
## - niveau_etudes                  2 1112.1
## - sexe                           1 1113.1
## - age                            1 1113.3
## - imc                            1 1113.3
## - bronchopneumopathie_chronique  1 1113.8
## - region                         2 1114.9
## <none>                             1115.0
## - exposition_domestique_fumee    1 1115.0
## - statut_cas_temoin              1 1177.1
```

```
##
## Step:  AIC=1112.14
## factor(exposition_professionnelle) ~ statut_cas_temoin + age +
##     sexe + imc + bronchopneumopathie_chronique + exposition_domestique_fumee +
##     region
##
##                                Df    AIC
## - sexe                          1 1110.2
## - imc                           1 1110.5
## - age                           1 1110.6
## - bronchopneumopathie_chronique  1 1110.9
## - region                        2 1111.7
## - exposition_domestique_fumee    1 1112.0
## <none>                            1112.1
## - statut_cas_temoin             1 1176.2
##
## Step:  AIC=1110.24
## factor(exposition_professionnelle) ~ statut_cas_temoin + age +
##     imc + bronchopneumopathie_chronique + exposition_domestique_fumee +
##     region
##
##                                Df    AIC
## - imc                           1 1108.6
## - age                           1 1108.7
## - bronchopneumopathie_chronique  1 1109.0
## - region                        2 1109.8
## - exposition_domestique_fumee    1 1110.0
## <none>                            1110.2
## - statut_cas_temoin             1 1174.8
##
## Step:  AIC=1108.59
## factor(exposition_professionnelle) ~ statut_cas_temoin + age +
##     bronchopneumopathie_chronique + exposition_domestique_fumee +
##     region
##
##                                Df    AIC
## - age                           1 1107.0
## - bronchopneumopathie_chronique  1 1107.3
## - region                        2 1108.1
## - exposition_domestique_fumee    1 1108.4
## <none>                            1108.6
## - statut_cas_temoin             1 1176.4
##
## Step:  AIC=1107.05
## factor(exposition_professionnelle) ~ statut_cas_temoin + bronchopneumopathie_chronique +
##     exposition_domestique_fumee + region
##
##                                Df    AIC
## - bronchopneumopathie_chronique  1 1105.7
## - region                        2 1106.6
## - exposition_domestique_fumee    1 1106.8
## <none>                            1107.0
## - statut_cas_temoin             1 1191.1
##
```

```
## Step:  AIC=1105.69
## factor(exposition_professionnelle) ~ statut_cas_temoin + exposition_domestique_fumee +
##     region
##
##                               Df    AIC
## - region                       2 1105.2
## - exposition_domestique_fumee  1 1105.6
## <none>                           1105.7
## - statut_cas_temoin            1 1192.6
##
## Step:  AIC=1105.17
## factor(exposition_professionnelle) ~ statut_cas_temoin + exposition_domestique_fumee
##
##                               Df    AIC
## - exposition_domestique_fumee  1 1104.9
## <none>                           1105.2
## - statut_cas_temoin            1 1191.7
##
## Step:  AIC=1104.89
## factor(exposition_professionnelle) ~ statut_cas_temoin
##
##                     Df    AIC
## <none>                 1104.9
## - statut_cas_temoin  1 1189.7

## Call:
## polr(formula = factor(exposition_professionnelle) ~ statut_cas_temoin,
##     data = CancerPoumon)
##
## Coefficients:
## statut_cas_temointemoin
##               -1.552647
##
## Intercepts:
##     Aucune exposition|Exposition faible Exposition faible|Exposition importante
##                             -1.3112805                               0.3264455
##
## Residual Deviance: 1098.893
## AIC: 1104.893
```

Le sous modèle avec la plus faible AIC est celui qui a pour variables explicatives statut_cas_temoin.

```
res_poly2 <- polr(factor(exposition_professionnelle) ~ statut_cas_temoin, data=CancerPoumon)
summary(res_poly2)
```

```
##
## Réajustement pour obtenir le Hessien

## Call:
## polr(formula = factor(exposition_professionnelle) ~ statut_cas_temoin,
##     data = CancerPoumon)
##
## Coefficients:
```

```
##                           Value Std. Error t value
## statut_cas_temointemoin -1.553     0.1722  -9.016
##
## Intercepts:
##                                          Value   Std. Error t value
## Aucune exposition|Exposition faible     -1.3113  0.1434    -9.1446
## Exposition faible|Exposition importante  0.3264  0.1288     2.5338
##
## Residual Deviance: 1098.893
## AIC: 1104.893
```

```
confint(res_poly2)
```

```
## Attente de la réalisation du profilage...
##
## Réajustement pour obtenir le Hessien

##     2.5 %    97.5 %
## -1.893574 -1.218140
```

Interpretation :

$$-1.893 \leq -ln\Big(\frac{odds(expo\_pro \leq Exposition\_faible \mid statut = Temoin)}{odds(tabagisme \leq Fumeur\_actuel \mid statut = Cas)}\Big) \leq -1.218$$

$$1.218 \leq ln\Big(\frac{odds(expo\_pro \leq Exposition\_faible \mid statut = Temoin)}{odds(tabagisme \leq Fumeur\_actuel \mid statut = Cas)}\Big) \leq 1.893$$

Par rapport aux cas, les témoins ont un odds d'être dans une catégorie d'exposition professionnelle inférieure ou égale à Exposition faible multiplié par au moins $exp(1.218) = 3.380$ et au plus$exp(1.893) = 6.639$.

Ainsi les personnes atteintes de cancer du poumon sont plus significativement plus exposées à la fumée dans le milieu professionnelle. Nous pouvons donc confirmer l'interpétation réalisée précédemment.

Meilleur modèle avec l'AIC sur tabagisme

On va mainatenant essayer de prédire la variable Y="tabagisme" qui a trois modalités ordonnées : "Ancien fumeur" < "Fumeur actuel" < "Jamais" : ou "Jamais" < "Ancien fumeur" < "Fumeur actuel".

```
res_poly3 <- polr(factor(tabagisme) ~ statut_cas_temoin + age + sexe + exposition_professionnelle + imc

stepAIC(res_poly3, direction="backward")
```

```
## Start:  AIC=1201.28
## factor(tabagisme) ~ statut_cas_temoin + age + sexe + exposition_professionnelle +
##     imc + bronchopneumopathie_chronique + niveau_etudes + exposition_domestique_fumee +
##     region
##
##                                Df    AIC
## - niveau_etudes                 2 1197.3
## - exposition_professionnelle    2 1197.4
## - sexe                          1 1199.3
## - bronchopneumopathie_chronique 1 1199.3
```

```
## - exposition_domestique_fumee    1 1199.9
## - age                            1 1200.1
## <none>                             1201.3
## - statut_cas_temoin              1 1201.3
## - region                         2 1205.0
## - imc                            1 1205.4
##
## Step:  AIC=1197.32
## factor(tabagisme) ~ statut_cas_temoin + age + sexe + exposition_professionnelle +
##     imc + bronchopneumopathie_chronique + exposition_domestique_fumee +
##     region
##
##                                 Df    AIC
## - exposition_professionnelle     2 1193.4
## - bronchopneumopathie_chronique  1 1195.3
## - sexe                           1 1195.3
## - exposition_domestique_fumee    1 1195.9
## - age                            1 1196.1
## <none>                             1197.3
## - statut_cas_temoin              1 1197.4
## - region                         2 1201.1
## - imc                            1 1201.6
##
## Step:  AIC=1193.4
## factor(tabagisme) ~ statut_cas_temoin + age + sexe + imc + bronchopneumopathie_chronique +
##     exposition_domestique_fumee + region
##
##                                 Df    AIC
## - bronchopneumopathie_chronique  1 1191.4
## - sexe                           1 1191.4
## - exposition_domestique_fumee    1 1192.0
## - age                            1 1192.2
## <none>                             1193.4
## - statut_cas_temoin              1 1194.1
## - region                         2 1197.4
## - imc                            1 1197.7
##
## Step:  AIC=1191.41
## factor(tabagisme) ~ statut_cas_temoin + age + sexe + imc + exposition_domestique_fumee +
##     region
##
##                               Df    AIC
## - sexe                         1 1189.4
## - exposition_domestique_fumee  1 1190.0
## - age                          1 1190.2
## <none>                           1191.4
## - statut_cas_temoin            1 1192.3
## - region                       2 1195.4
## - imc                          1 1195.7
##
## Step:  AIC=1189.42
## factor(tabagisme) ~ statut_cas_temoin + age + imc + exposition_domestique_fumee +
##     region
##
```

```
##                               Df    AIC
## - exposition_domestique_fumee  1 1188.0
## - age                          1 1188.2
## <none>                           1189.4
## - statut_cas_temoin            1 1190.3
## - region                       2 1193.4
## - imc                          1 1193.7
##
## Step:  AIC=1188.03
## factor(tabagisme) ~ statut_cas_temoin + age + imc + region
##
##                     Df    AIC
## - age                1 1186.9
## <none>                 1188.0
## - statut_cas_temoin  1 1188.6
## - region             2 1191.7
## - imc                1 1192.4
##
## Step:  AIC=1186.87
## factor(tabagisme) ~ statut_cas_temoin + imc + region
##
##                     Df    AIC
## <none>                 1186.9
## - statut_cas_temoin  1 1189.3
## - region             2 1190.5
## - imc                1 1191.0


## Call:
## polr(formula = factor(tabagisme) ~ statut_cas_temoin + imc +
##     region, data = CancerPoumon)
##
## Coefficients:
## statut_cas_temointemoin                     imc   regionHauts-de-France
##              0.34470974            -0.05604152             -0.47424623
##     regionIle-de-France
##             -0.42975985
##
## Intercepts:
## Ancien fumeur|Fumeur actuel        Fumeur actuel|Jamais
##                  -2.3254793                  -0.5292985
##
## Residual Deviance: 1174.869
## AIC: 1186.869
```

Le sous modèle avec le plus faible AIC est celui qui a pour variables explicatives “statut_cas_temoin”, “imc” et “region”.

```
res_poly4 <- polr(factor(tabagisme) ~ statut_cas_temoin + imc + region, data=CancerPoumon)
summary(res_poly4)
```

```
##
## Réajustement pour obtenir le Hessien
```

```
## Call:
## polr(formula = factor(tabagisme) ~ statut_cas_temoin + imc +
##     region, data = CancerPoumon)
##
## Coefficients:
##                             Value Std. Error t value
## statut_cas_temointemoin  0.34471    0.16332   2.111
## imc                     -0.05604    0.02263  -2.476
## regionHauts-de-France   -0.47425    0.19553  -2.425
## regionIle-de-France     -0.42976    0.18999  -2.262
##
## Intercepts:
##                              Value   Std. Error t value
## Ancien fumeur|Fumeur actuel -2.3255  0.6024    -3.8600
## Fumeur actuel|Jamais        -0.5293  0.5939    -0.8912
##
## Residual Deviance: 1174.869
## AIC: 1186.869
```

```
# Intervalles de confiance :
confint(res_poly4)
```

```
## Attente de la réalisation du profilage...
##
## Réajustement pour obtenir le Hessien
```

```
##                               2.5 %      97.5 %
## statut_cas_temointemoin  0.02526546  0.66578775
## imc                     -0.10063188 -0.01182337
## regionHauts-de-France   -0.85882124 -0.09190443
## regionIle-de-France     -0.80332027 -0.05813098
```

Interpretation des coefficients :

Pour la variable "statut_cas_temoin" :

$$0.025 \leq -ln\Big(\frac{odds(tabagisme \leq Fumeur\_actuel \mid imc = x,\, region,\, statut = Temoin)}{odds(tabagisme \leq Fumeur\_actuel \mid imc = x,\, region,\, statut = Cas)}\Big) \leq 0.666$$

$$-0.666 \leq ln\Big(\frac{odds(tabagisme \leq Fumeur\_actuel \mid imc = x,\, region,\, statut = Temoin)}{odds(tabagisme \leq Fumeur\_actuel \mid imc = x,\, region,\, statut = Cas)}\Big) \leq -0.025$$

À imc et région fixés fixé, appartenir aux témoins, plutôt qu'aux cas divise l'odds d'être dans une catégorie de tabagisme inférieure ou égale à Fumeur actuel par au moins $exp(-0.666) = 0.514$ et au plus $exp(-0.025) = 0.975$.

Pour la variable "imc" :

$$-0.101 \leq -ln\Big(\frac{odds(tabagisme \leq Fumeur\_actuel \mid imc = x+1,\, region,\, statut)}{odds(tabagisme \leq Fumeur\_actuel \mid imc = x,\, region,\, statut)}\Big) \leq -0.012$$

$$0.012 \leq ln\Big(\frac{odds(tabagisme \leq Fumeur\_actuel \mid imc = x+1,\, region,\, statut)}{odds(tabagisme \leq Fumeur\_actuel \mid imc = x,\, region,\, statut)}\Big) \leq 0.101$$

A région et statut_cas_temoin fixé, avoir un imc d'une unité en plus, multiplie l'odds d'être dans une catégorie de tabagisme inférieure ou égale à Fumeur actuel par au moins $exp(0.012) = 1.012$ et au plus $exp(0.101) = 1.106$.

Pour la variable "region" :

Pour la modalité "Haut-de-france" :

$$-0.859 \leq -ln\Big(\frac{odds(tabagisme \leq Fumeur\_actuel \mid imc = x,\, region = Haut\_de\_france,\, statut)}{odds(tabagisme \leq Fumeur\_actuel \mid imc = x,\, region = Centre\_Val\_de - Loire,\, statut)}\Big) \leq -0.092$$

$$0.092 \leq ln\Big(\frac{odds(tabagisme \leq Fumeur\_actuel \mid imc = x,\, region = Haut\_de\_france,\, statut)}{odds(tabagisme \leq Fumeur\_actuel \mid imc = x,\, region = Centre\_Val\_de\_Loire,\, statut)}\Big) \leq 0.859$$

A imc et statut_cas_temoin fixé, habité dans la région Haut-de-france, plutôt que dans la région Centre-Val-de-Loire multiplie l'odds d'être dans une catégorie de tabagisme inférieure ou égale à Fumeur actuel par au moins $exp(0.092) = 1.096$ et au plus $exp(0.859) = 2.360$.

Pour la modalité "Ile-de-France" :

$$-0.803 \leq -ln\Big(\frac{odds(tabagisme \leq Fumeur\_actuel \mid imc = x,\, region = Ile\_de\_France,\, statut)}{odds(tabagisme \leq Fumeur\_actuel \mid imc = x,\, region = Centre\_Val\_de\_Loire,\, statut)}\Big) \leq -0.058$$

$$0.058 \leq ln\Big(\frac{odds(tabagisme \leq Fumeur\_actuel \mid imc = x,\, region = Ile\_de\_France,\, statut)}{odds(tabagisme \leq Fumeur\_actuel \mid imc = x,\, region = Centre\_Val\_de\_Loire,\, statut)}\Big) \leq 0.803$$

A imc et statut_cas_temoin fixé, habité dans la région Ile-de-France, plutôt que dans la région Centre-Val-de-Loire multiplie l'odds d'être dans une catégorie de tabagisme inférieure ou égale à Fumeur actuel par au moins $exp(0.058) = 1.060$ et au plus $exp(0.803) = 2.232$.

En conclusion, le meilleur modèle au sens de l'AIC montre que le niveau de tabagisme est significativement associé au statut cas/témoin, à l'IMC et à la région de résidence. Toutes choses égales par ailleurs, le statut cas/témoin influence la répartition dans les catégories de tabagisme, l'IMC est lié au niveau de tabagisme, et il existe des différences régionales marquées. Ces résultats indiquent que le tabagisme est qu'il est lié au cancer mais également à d'autres facteurs.

Durant ce projet nous avons également essayé de faire différents modèles avec interractions mais les résultats n'étant jamais concluant et significatifs nous ne les avons pas inclus dans ce rapport.

```
AIC(reslt, reslt_imc, reslt_exp, reslt_exd, reslt_etu, reslt_reg, reslt_age, reslt_bron, res_log_quanti3
```

```
##                  df       AIC
## reslt             3  624.9182
## reslt_imc         2  722.7505
## reslt_exp         3  666.3183
## reslt_exd         2  740.6561
## reslt_etu         3  733.4516
## reslt_reg         3  752.0659
## reslt_age         2  649.1469
## reslt_bron        2  706.1937
## res_log_quanti3   3  630.4151
## modele_complet_2 12  456.8072
## res_poly2         3 1104.8929
## res_poly4         6 1186.8685
```

Nous comparons tous les modèles étudiés avec l'AIC, nous obtenons les résultats les plus faibles avec le modèle univérié avec le tabagisme et l'exposition professionnelle ainsi que le modèle multivarié avec toutes les variables explicatives sauf région mais celui ici à un nombre plus important de degré de liberté.

Conclusion

Cette étude met en évidence une association forte entre le cancer du poumon et plusieurs facteurs de risque majeurs, en particulier le tabagisme, l'âge, l'exposition professionnelle à la fumée et la bronchopneumopathie chronique. Les modèles multivariés confirment que ces facteurs restent significativement associés au cancer après ajustement, tandis qu'un IMC plus élevé apparaît comme un facteur protecteur. Les analyses complémentaires montrent également que le niveau de tabagisme et l'exposition à la fumée varient selon des caractéristiques individuelles. Globalement, ces résultats confirment le rôle central du tabac et des expositions à la fumée dans le cancer du poumon.