biostat

camille mathilde

22/12/2020

Introduction

De nos jours, les maladies sont de plus en plus étudiées et de mieux en mieux comprises. De nombreux organismes et instituts recherchent des solutions pour vaincre ces maladies en trouvant des traitements. De nombreuses diciplines sont impliquées, notamment la biostatistique, qui permet d'étudier différentes situations avec des données. Dans notre cas, nous allons étudier l'occurence des maladies coronarienne du coeur. Une étude préalable à déja été faite sur une cohorte d'individus. Ces individus ont répondu à une date donnée à une enquête sur les habitudes alimentaires. Les réponses ont été recueilli dans une base de donnée que l'on nomme **Coeur**.
L'objectif de cette étude sera de connaître les bonnes habitudes alimentaires à adopter pour se prevenir d'une maladie coronarienne.

Contents

Présentation de la base de donnée

Notre table de données contient 337 individus et 15 variables qui sont

- **Id** : l'identifiant du sujet.
- **DateEntrée** et **Date de sortie** : les dates d'entrée et de sortie de l'étude.
- **Date Naissance** : la date de naissance.
- **Statut** : si la sortie de l'enquête est due à une maladie coronarienne de cœur, alors le type de maladie est indiqué (la signification du code n'est pas précisée ici). Si l'individu est sain à la sortie de l'enquête, alors le code vaut 0.
- **Emploi** : le type d'emploi.
- **MoisEnquête** : le mois (1= Janvier, 12= Décembre) où l'individu a répondu à l'enquête sur ses pratiques alimentaires.
- **Taille/Poids** : la taille et le poids de l'individu (en cm et en kg).
- **Graisse** : la quantité moyenne de graisse ingérée par jour (g/jour).
- **Fibres** : la quantité moyenne de fibres ingérée par jour (g/jour).
- **Consommation** : la quantité de calories(/100) ingérée par jour.
- **hauteConsomation** : une variable binaire, recodage de la variable consommation.
- **MCC** : une variable binaire, recodage de la variable statut (1=MCC, 0=pas de MCC).

Nous avons également rajouté la variable **IMC** en divisant le poids par la taille au carrée pour faire un lien entre la condition physique et la maladie du coeur car le poids ou la taille tout seul ne suffisent pas pour savoir si une personne est en bonne santé ou en surpoids.

Nous nous sommes également rendu compte qu'il y avait des erreurs dans le recodage de la variables **statut**. En effet, certains individus avaient contracté une maladie du coeur, mais la variable recodage **MCC** ne l'avait pas pris en compte nous avons donc rectifié ça. Nous avons aussi remarqué que certaines variables qualitatives étaient en **numeric**, ce qui posera problème pour notre étude. Nous les recodons donc en **factor**. Nous observons enfin que la table de données contient des valeurs manquantes. Nous enlevons donc chaque ligne qui contient au moins une valeur manquante. Nous passons donc de 337 à 328 individus.

Regardons les 5 premières lignes de notre table de données :

```
coeur <- readRDS('data/my_data_frame.rds')
coeur <-coeur%>%drop_na()
coeur <-coeur%>%dplyr::select(-X1)
coeur$statut<-as.factor(coeur$statut)
coeur$emploi<-as.factor(coeur$emploi)
coeur$moisEnqu_e<-as.factor(coeur$moisEnqu_e)
coeur$hauteConsomation<-as.factor(coeur$hauteConsomation)

coeur<-mutate(coeur,imc =poids /(taille/100)^2)
coeur<-coeur %>% dplyr::select(-MCC)
coeur <-mutate(coeur,MCC =case_when(
  statut!=0~1,
  TRUE~0))

pander(head(coeur))
```

Table 1: Table continues below

id	dateEntree	dateSortie	dateNaissance	statut	emploi
102	17/01/76	02/12/86	02/03/39	0	Driver
59	16/07/73	05/07/82	05/07/12	0	Driver
126	17/03/70	20/03/84	24/12/19	13	Conductor
16	16/05/69	31/12/69	17/09/06	3	Driver
247	16/03/68	25/06/79	10/07/18	13	Bank worker
272	16/03/69	13/12/73	06/03/20	3	Bank worker

Table 2: Table continues below

moisEnqu_e	consommation	taille	poids	graisse	fibre
1	22.86	181.6	88.18	9.168	1.4
7	23.88	166	58.74	9.651	0.935
3	24.95	152.4	49.9	11.25	1.248
5	22.24	171.2	89.4	7.578	1.557
3	18.54	177.8	97.07	9.147	0.991
3	20.31	175.3	61.01	8.536	0.765

hauteConsomation	imc	MCC
<=2750 KCals	26.74	0
<=2750 KCals	21.32	0
<=2750 KCals	21.48	1
<=2750 KCals	30.51	1
<=2750 KCals	30.71	1
<=2750 KCals	19.86	1

Remarquons que cet ensemble d'individu est bien une cohorte car nous avons relevé certaines covariables et les trois données fondamentales qui sont, la date d'entrée dans l'étude, la date de sortie dans l'étude et le cause de sortie dans l'étude. Nous pouvons ajouter que les covariables utilisées dans l'étude sont fixe.

Statistique descriptive

Notre jeu de donnée présente 5 variables quantitatives et 9 variables qualitatives.

Faisons un sommaire des variables quantitatives :

consommation	fibre	graisse	taille	poids
Min. :17.48	Min. :0.605	Min. : 7.26	Min. :152.4	Min. : 46.72
1st Qu.:25.46	1st Qu.:1.367	1st Qu.:11.15	1st Qu.:168.9	1st Qu.: 64.64
Median :28.11	Median :1.679	Median :12.60	Median :173.0	Median : 72.80
Mean :28.35	Mean :1.723	Mean :12.76	Mean :173.4	Mean : 72.40
3rd Qu.:31.10	3rd Qu.:1.939	3rd Qu.:14.02	3rd Qu.:177.8	3rd Qu.: 79.44
Max. :43.96	Max. :5.351	Max. :21.63	Max. :190.5	Max. :106.14

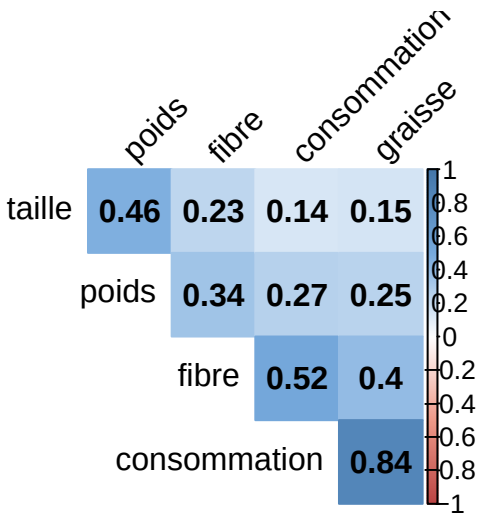
Les individus mangent en moyenne 2835 calories par jours. Ils ingèrent de plus 12.76 grammes de gras par jour, mesurent 173 cm et pèsent 72.40 kilo-gramme en moyenne.

Faisons maintenant un sommaire des variables quantitatives :

statut	emploi	moisEnqu_e	hauteConsomation
0 :252	Bank worker:147	11 : 39	<=2750 KCals:149
3 : 18	Conductor : 83	1 : 37	>2750 KCals :179
13 : 18	Driver : 98	3 : 37	NA
12 : 12	NA	2 : 34	NA
5 : 10	NA	5 : 34	NA
1 : 8	NA	12 : 33	NA
(Other): 10	NA	(Other):114	NA

Il y a 2 fois plus de **Bank worker** que de **Conductor** ou de **Driver**. Il y a 149 personnes qui mangent moins de 2750 calories par jours et 179 personnes qui en mangent plus.

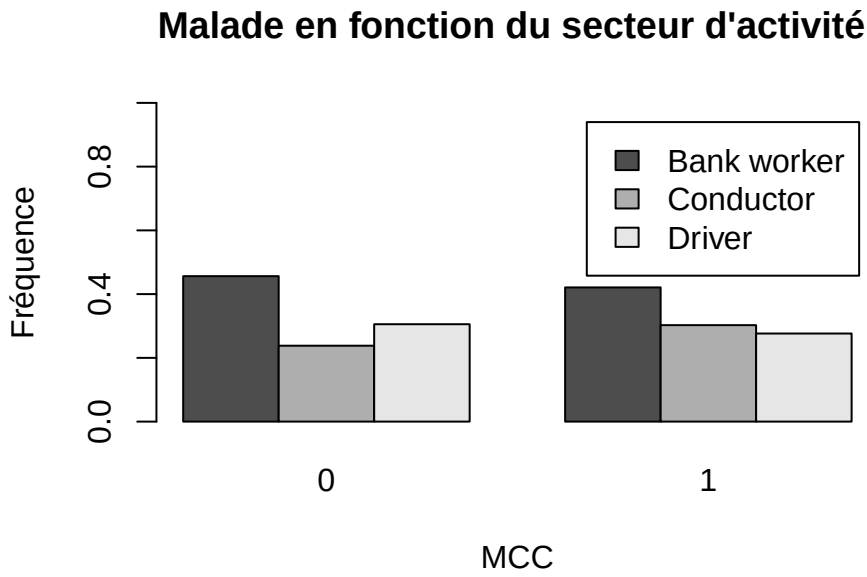
Regardons désormais la corrélation entre les variables quantitatives :



Nous observons que toutes les variables sont corrélées positivement.

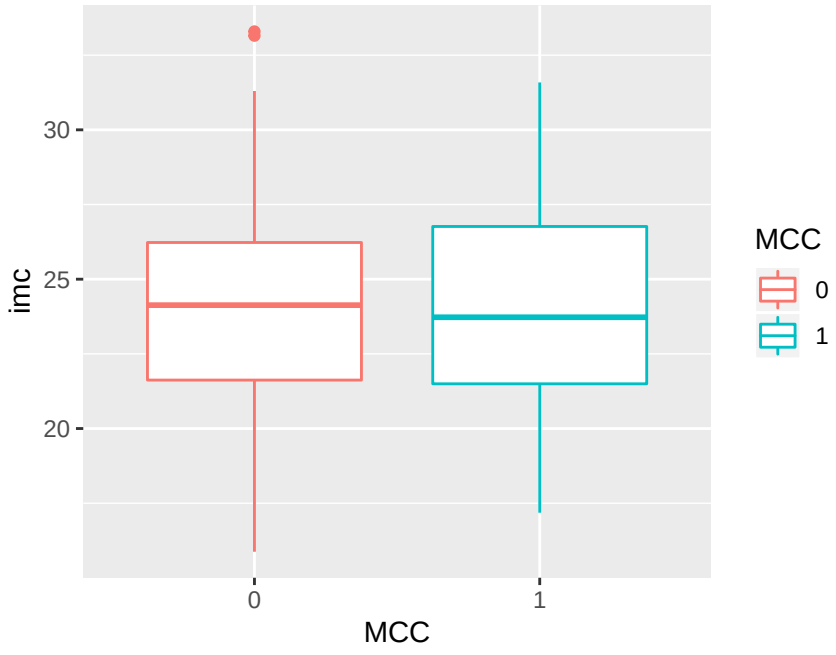
La variable **consommation** est très corrélée avec la variable **graisse** mais elle est très peu corrélée avec la variable **taille**. La variable **taille** est peu corrélée avec la variable **graisse**.

Nous allons maintenant regarder le lien entre la maladie et le secteur d'activité.



Nous observons que la proportion de conducteur est plus élevée chez les malades que chez les non malades.

Nous allons maintenant regarder le lien entre l'IMC et les maladies coronariennes.



Nous remarquons que la mediane de l'IMC du groupe de personnes malades est à peine plus basse que la médiane pour le groupe de personnes non malade. Une simple analyse descriptive ne suffit pas pour obtenir des résultats bien concluants, nous allons continuer avec des méthodes plus poussées. Nous commencerons par des régréssions logistiques.

Etude de la consommation de fibre

Régression logistique binaire

Nous allons commencer par une régréssion logistique qui nous permettera d'expliquer la variable **MCC** en fonction de certaines covariables. Rappelons que dans la régréssion logistique ce n'est pas la réponse binaire qui est modélisé mais la probabilité de réalisation d'une des deux modalités (avoir une maladie coronarienne ou non).

Nous allons commencer par regarder le meilleur modèle. Nous enlevons la **taille** et le **poids** pour laisser **imc** car elles sont très fortement corrélées. Nous avons tout d'abord voulu comparer en comparant les AIC avec stepAIC.

```
coeurlogbin<-coeur%>%dplyr::select(-id,-dateEntree,-dateSortie,-dateNaissance,-statut,
                                   -poids,-taille)
res<-glm(MCC~.,family = binomial(logit),data=coeurlogbin)
stepAIC(res)
```

Le modèle ayant le plus faible AIC est le suivant :

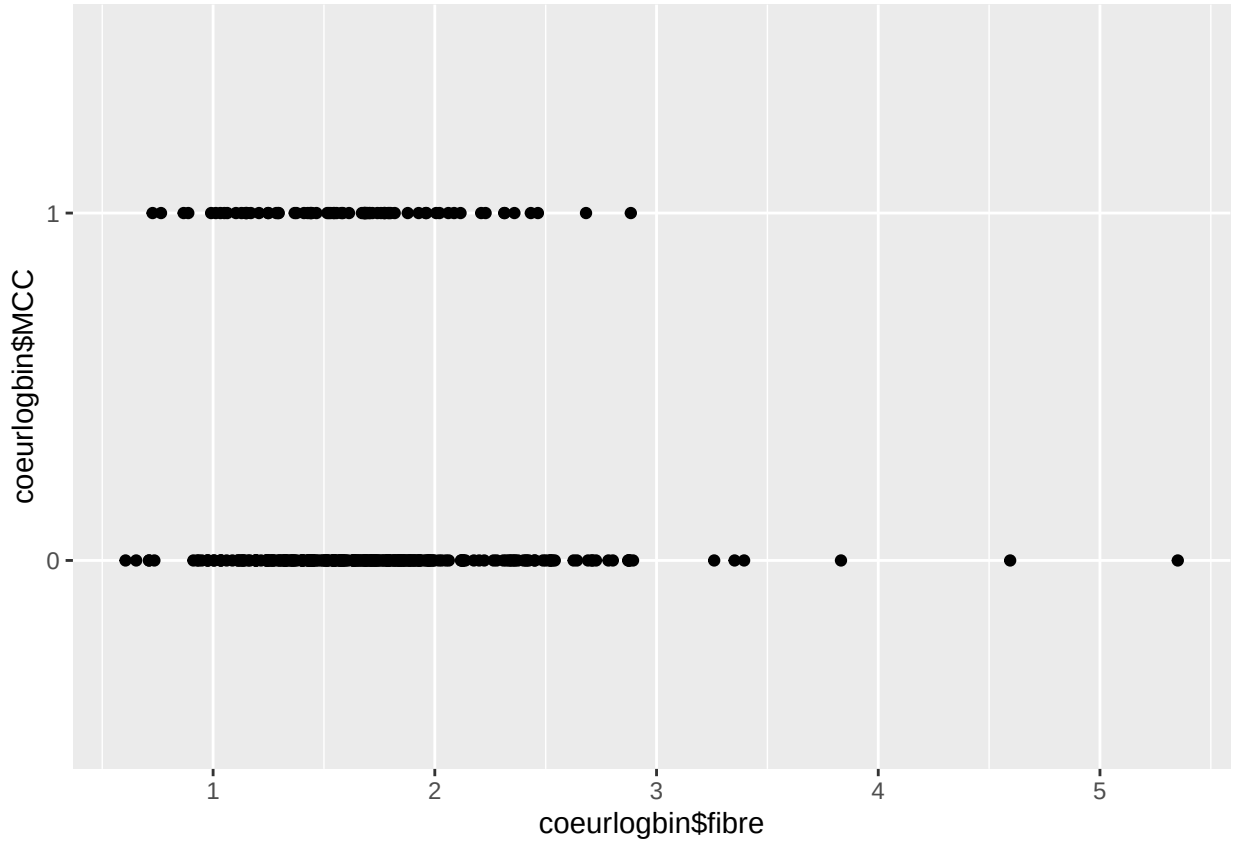
$$y = \mu + \beta_1 fibre$$

Nous avons également utilisé les p-values pour trouver le meilleur modèle en enlevant à la main petit à petit les variables les moins significatives et nous obtenons encore le même modèle.

```
coeurlogbin<-coeur%>%dplyr::select(-id,-dateEntree,-dateSortie,-dateNaissance,-statut,
                                   -poids,-taille)
res<-glm(MCC~.-emploi-moisEnqu_e-consommation-1-hauteConsomation-imc-graisse,
         family = binomial(logit),data=coeurlogbin)
summary(res)
```

Regardons donc la relation entre les maladies coronariennes et la consommation de fibre.

```
ggplot(coeurlogbin,aes(x=coeurlogbin$fibre,y=coeurlogbin$MCC))+geom_point()
```



Nous observons que les personnes mangeant plus de 2g de fibre par jour sont moins malades.

Nous voulons faire une régression logistique, nous allons donc vérifier les conditions d'application. Il est recommandé d'avoir en pratique 10 fois plus d'évenements que de paramètres dans le modèle. Nous allons utiliser ici 2 paramétres (en comptant l'intercept) nous devrions donc avoir au moins 20 malades.

```
table(coeurlogbin$MCC)
```

```
##
##   0   1
## 252  76
```

Nous avons 76 malades nous pouvons donc continuer.

Il faut maintenant vérifier que nous ne sommes pas dans le cas de surdipersion c'est à dire qu'il ne faut pas que la dispersion réelle des données soit supérieure à celle prévue par la théorie car dans ce cas l'erreur standard des paramètres est sous-estimée ce qui peut conduire à des p-valeurs très faible et donner des conclusions erronnées. Evaluons donc s'il y a ou non une surdispersion :

```
reslog<-glm(MCC~fibre,family = binomial(logit),data=coeurlogbin)
summary(reslog)
```

```
##
## Call:
## glm(formula = MCC ~ fibre, family = binomial(logit), data = coeurlogbin)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -0.9166  -0.7624  -0.7002  -0.5372   1.9817
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.3413     0.4588  -0.744   0.4569
## fibre        -0.5101     0.2678  -1.905   0.0568 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 355.11  on 327  degrees of freedom
## Residual deviance: 351.12  on 326  degrees of freedom
## AIC: 355.12
##
## Number of Fisher Scoring iterations: 4
```

$$\frac{devianceresiduelle}{nddl} = \frac{351.12}{326} = 1.08$$

nous pouvons ainsi considérer qu'il n'y a pas surdispersion.

Nous pouvons donc maintenant utiliser la régression logistique et faire des interprétations :

```
reslog<-glm(MCC~fibre-1,family = binomial(logit),data=coeurlogbin)
summary(reslog)
```

```
##
## Call:
## glm(formula = MCC ~ fibre - 1, family = binomial(logit), data = coeurlogbin)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0030  -0.7823  -0.6959  -0.4816   2.0741
##
## Coefficients:
##       Estimate Std. Error z value Pr(>|z|)
## fibre -0.70288    0.07799  -9.013   <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: 454.70  on 328  degrees of freedom
## Residual deviance: 351.66  on 327  degrees of freedom
## AIC: 353.66
##
## Number of Fisher Scoring iterations: 4
```

```
confint(reslog)
```

```
##      2.5 %     97.5 %
## -0.8604690 -0.5542772
```

$$ln\left(\frac{odds(Y=1|fibre=x_1+1)}{odds(Y=1|fibre=x_1)}\right)=\hat{\beta}_{fibre}=-0.70288$$

Augmenter la consommation de fibre de 1g par jour va multiplier la chance d'avoir une maladie coronarienne par au moins $\exp(-0.8604690)=0.43$ et au plus $\exp(-0.5542772)=0.57$. Autrement dit, si nous augmentons notre consommation de fibre de 1g par jour nous divisons par au moins (1/0.43)=2.32 l'odds de contracter une maladie coronarienne.
Nous pouvons alors conseiller de manger des fibres si nous voulons nous prévenir des maladies coronariennes.

Etude de la consommation

Régression polytomique ordonné

Nous voulons maintenant étudier la consommation à l'aide d'une régression polytomique ordonné. En effet nous aimerions voir ce qui influence la consommation de calories des personnes. Nous Nous voulons au moins 3 modalités, nous recodons donc la variable **consommation** pour qu'elle soit qualitative à 3 modalités. Nous classons ainsi la consommation en "faible", "moyen" et "élevée". Nous avons alors bien une relation d'ordre $"faible" < "moyen" < "élevée"$.

```
coeur <-mutate(coeur,consom_dec =case_when(
  consommation<23~"faible",
  consommation>22 & consommation<30~"moyen",
  consommation>29 ~"élevée"))
coeur$consom_dec<-as.factor(coeur$consom_dec)
```

Nous faisons un stepAIC pour choisir le meilleur modèle au sens du critère **AIC**.

```
coeurlogbin1<-coeur%>%dplyr::select(-id,-dateEntree,-dateSortie,-dateNaissance,-statut,
                                    -hauteConsomation,-consommation)
modele<-polr(consom_dec~.,data=coeurlogbin1)
stepAIC(modele)
```

Nous obtenons qu'il faut garder **graisse** et **fibre**, nous faisons donc la régréssion polytomique ordonné selon ces variables.

```
modele2<-polr(consom_dec~graisse+fibre,data=coeurlogbin1)
modele2
```

```
## Call:
## polr(formula = consom_dec ~ graisse + fibre, data = coeurlogbin1)
##
## Coefficients:
##    graisse      fibre
## -0.4308596 -0.5959642
##
## Intercepts:
## élevée|faible  faible|moyen
##     -7.240439     -6.698937
##
## Residual Deviance: 504.4718
## AIC: 512.4718
```

```
confint(modele2)
```

```
##              2.5 %     97.5 %
## graisse -0.559942 -0.3106530
## fibre   -1.099625 -0.1207172
```

Notons que la commande *polyr* utilisée renvoie l'opposé du coefficient β considéré donc nous obtenons les "vrais" résultats suivants : $\hat{\beta}_{graisse} = 0.43$ et $\hat{\beta_{fibre}} = 0.59$ avec les intervalles de confiance suivant : $\hat{\beta}_{graisse} \in [0.31, 0.56]$ et $\hat{\beta}_{fibre} \in [0.12, 1.10]$

Nous pouvons regarder noter que :

$$ln\left(\frac{odds(Y \leq moyen|graisse = x_1 + 1, fibre = x_2)}{odds(Y \leq moyen|graisse = x_1, fibre = x_2)}\right) = \hat{\beta}_{graisse}$$

A consommation de fibre fixée, augmenter la consommation de graisse de 1 g/jour va multiplier l'odds de $Y \leq moyen$ par au moins $exp(0.31) = 1.36$ et au plus $exp(0.56) = 1.75$.

De plus :

$$ln\left(\frac{odds(Y \leq moyen|fibre = x_1 + 1, graisse = x_2)}{odds(Y \leq moyen|fibre = x_1, graisse = x_2)}\right) = \hat{\beta}_{fibre}$$

A consommation de graisse fixée, augmenter la consommation de fibre de 1 g/jour va multiplier l'odds de $Y \leq moyen$ par au moins $exp(0.12) = 1.12$ et au plus $exp(1.10) = 3$. Ainsi manger plus de fibre va augmenter les chances que la personne mange moins de calorie dans la journée. Ceci n'est pas surpenant car les fibres sont connus pour être rassasiant.

Nous avons vu précedemment qu'augmenter sa consommation de fibre était une bonne chose pour se prévenir des maladies coronariennes, nous savons maintenant que à consommation de graisse fixée la consommation de fibre permet également au personne de ne pas manger trop de calories.

Étude de l'emploi

Régression polytomique non ordonné

Nous voulons maintenant étudier le l'emploi à l'aide d'une régression polytomique non ordonné. En effet nous aimerions voir ce qui influence la préférence de choisir un type d'emploi par apport aux autres. La variable emploi est composée de 3 modalités, qui sont : *conductor*, *bankworker* et *driver*.

Nous faisons un stepAIC pour choisir le meilleur modèle au sens du critère AIC

Nous obtenons qu'il faut garder taille et poids. Nous faisons donc la régréssion polytomique non ordonné selon ces variables.

```
modele4<-multinom(emploi~poids+taille,data=coeurlogbin1)
```

```
## # weights:  12 (6 variable)
## initial  value 360.344831
## iter  10 value 302.773746
## iter  20 value 301.538107
## final  value 301.526738
## converged
```

```
modele4
```

```
## Call:
## multinom(formula = emploi ~ poids + taille, data = coeurlogbin1)
##
## Coefficients:
##           (Intercept)       poids     taille
## Conductor    34.13104 -0.07620753 -0.1694273
## Driver       22.11401 -0.01424113 -0.1231291
##
## Residual Deviance: 603.0535
## AIC: 615.0535
```

```
confint(modele4)
```

```
## , , Conductor
##
##                   2.5 %      97.5 %
## (Intercept) 30.3445677 37.91750699
## poids       -0.1114239 -0.04099117
## taille      -0.1959546 -0.14290007
##
## , , Driver
##
##                    2.5 %      97.5 %
## (Intercept) 16.95239383 27.27563010
## poids       -0.04357888  0.01509662
## taille      -0.15673368 -0.08952445
```

Bank worker est la modalité de référence pour emploi.

$$\hat{\beta}_{poids|conductor} = ln\left(\frac{\frac{P(conductor|poids=x_1+1,taille=x_2)}{P(Bankworker|poids=x_1+1,taille=x_2)}}{\frac{P(conductor|poids=x_1,taille=x_2)}{P(Bankworker|poids=x_1,taille=x_2)}}\right) = -0.07620753$$

À taille fixé, un poids de 1kg en plus va multiplier par au moins $exp(-0.1114239) = 0.8945595$ et au plus$exp(-0.04099117) = 0.9598376$ la préférence de conductor par rapport à Bank worker.

$$\hat{\beta}_{taille|conductor} = ln\left(\frac{\frac{P(conductor|taille=x_1+1,poids=x_2)}{P(Bankworker|taille=x_1+1,poids=x_2)}}{\frac{P(conductor|taille=x_1,poids=x_2)}{P(Bankworker|taille=x_1,poids=x_2)}}\right) = -0.1694273$$

À poids fixé, une taille qui augmente de 1cm va multiplier par au moins $exp(-0.1959546) =$ et au plus $exp(-0.14290007) = 0.8668407$ la préférence de conductor par rapport à Bank worker.

$$\hat{\beta}_{taille|driver} = ln\left(\frac{\frac{P(driver|taille=x_1+1,poids=x_2)}{P(Bankworker|taille=x_1+1,poids=x_2)}}{\frac{P(driver|taille=x_1,poids=x_2)}{P(Bankworker|taille=x_1,poids=x_2)}}\right) = -0.1231291$$

À poids fixé, une taille qui augmente de 1cm va multiplier par au moins $exp(-0.15673368) = 0.8549317$ et au plus $exp(-0.08952445) = 0.9143659$ la préférence de driver par rapport à Bank worker.

Ainsi, la taille et le poids ont un effet sur le métier que l'on va choisir.

Étude de l'incidence instantannée de la maladie

Dans cette partie, nous allons chercher à répondre à la problématique suivante : à une date donnée, quelle sera le taux de nouveaux malades dans la population étudiée ?

Nous allons désormais nettoyer la base de données et faire des transformations de format des variables pour pouvoir utiliser le modèle de Cox. Les variables qui contiennent une date (date entrée, date sortie et date de naissance) sont de type "charactère" nous allons donc dans un premier temps les convertir en type "Date".

Ensuite, la fonction qui exécute le modèle de cox à besoin de valeurs numériques représentant les dates. Sous R, chaque date est repésentée par un nombre de jour à partir d'une date d'origine : le 1 janvier 1970. Nous allons donc créer une variable qui récupérera ce nombre pour chaque date correspondant à chaques individus.

Nous pouvons désormais faire nos analyse avec le modèle de Cox. Pour une première analyse, nous estimerons notre modèle en prenant en compte toute les covariables possible présentent dans la table de données. Mais le modèle de risques proportionnels de Cox fait plusieurs hypothèses. Ainsi, il est important d'évaluer si un tel modèle ajusté décrit correctement les données.

Ici, nous allons discuter de trois types de diagonostiques pour le modèle de Cox:

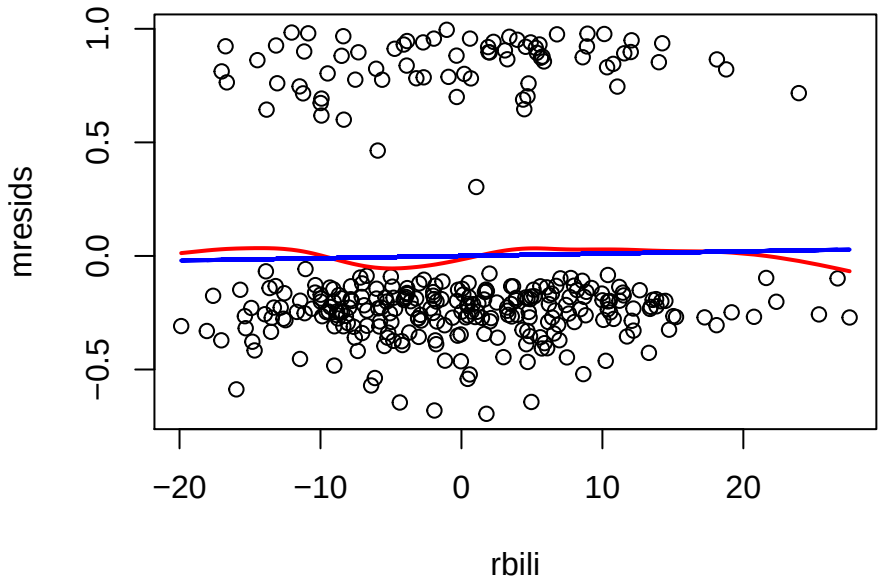
- Test de l'hypothèse des risques proportionnels.
- Examiner les observations influentes (ou les valeurs aberrantes).
- Détection de la non-linéarité des variables.

Détection de la non-linéarité des variables

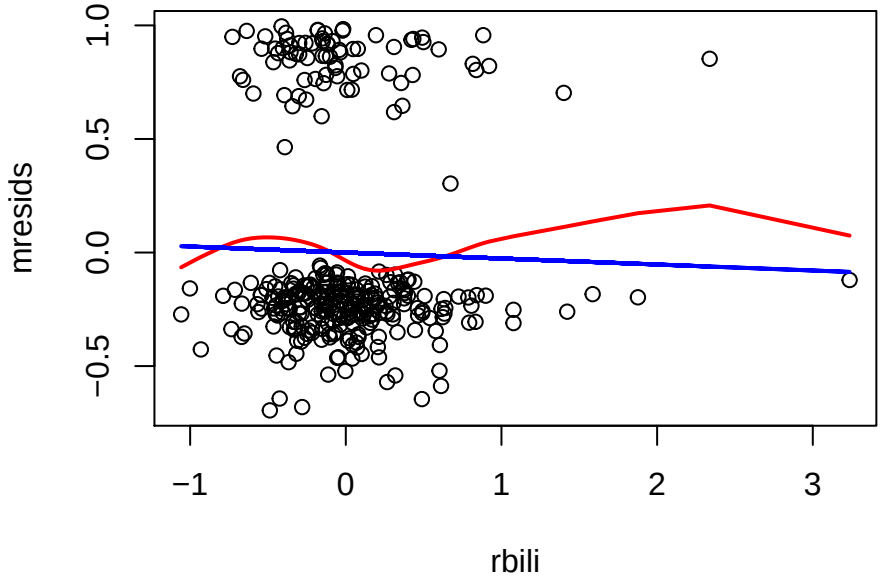
Souvent, nous supposons que les covariables continues ont une forme linéaire. Cependant, cette hypothèse doit être vérifiée. Le traçage des résidus de Martingale par rapport à des covariables continues est une approche courante utilisée pour détecter la non - linéarité ou, en d'autres termes, pour évaluer la forme fonctionnelle d'une covariable. Pour une covariable continue donnée, les modèles du graphique peuvent suggérer que la variable n'est pas correctement ajustée.

La non-linéarité n'est pas un problème pour les variables catégorielles, nous n'examinons donc que les graphiques des résidus de martingale par rapport à une variable continue. Testons avec le poids :

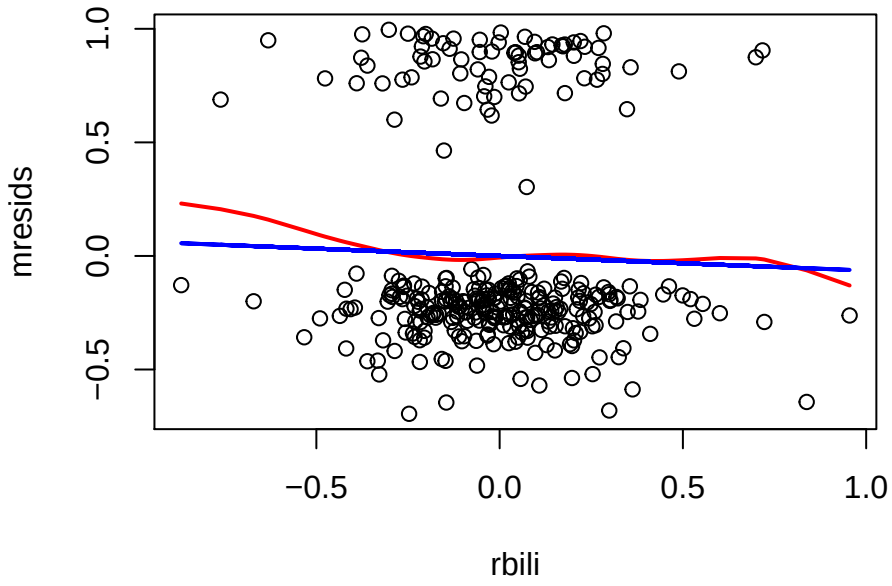
```
survie=Surv(coeur2$date_entree_num,coeur2$date_sortie_num,coeur2$MCC)
res=coxph(survie~fibre+taille+poids+consommation+
            emploi+graisse,id=id,data=coeur2)
mresids <- residuals( res, type="martingale" )
lmfit <- lm(poids~taille+consommation+graisse+fibre,data=coeur2 )
rbili <- lmfit$resid
ord <- order( rbili )
mresids <- mresids[ ord ]
plot( rbili, mresids )
lines( smooth.spline( rbili, mresids, df=6 ), col="red", lwd=2 )
lines( rbili, fitted(lm( mresids ~ rbili )), col="blue", lwd=2 )
```



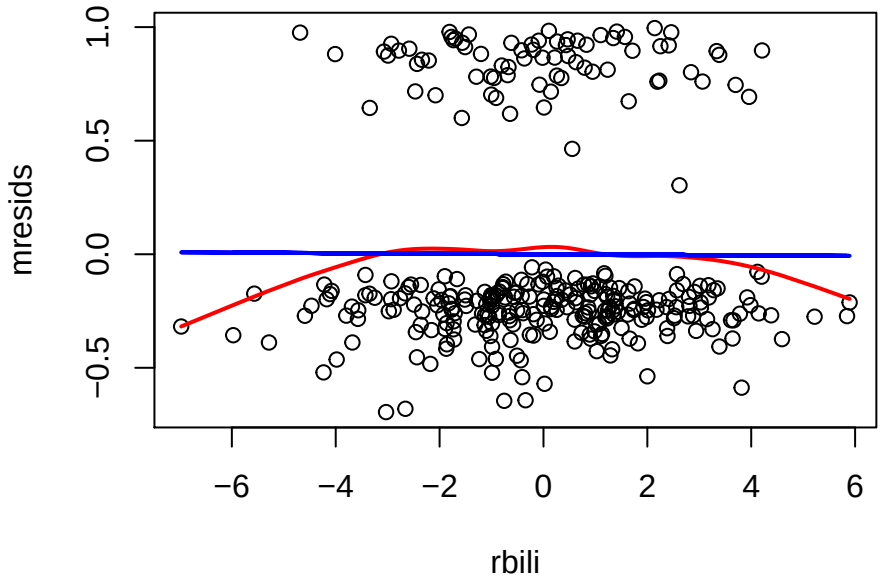
Il semble que le poids soit linéaire. Maintenant testons avec la variable fibre



On distingue clairement 2 groupes de résidus et la courbe varie un peu. On va appliquer une transformation logarithme à cette variable :

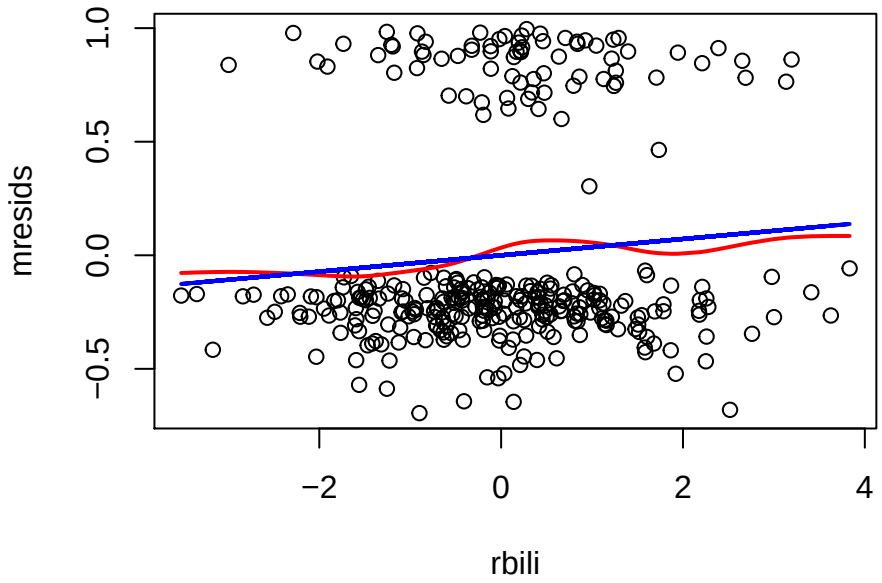


On voit que c'est déja un peu mieux. On appliquera le logarithme à fibre dans nos modèles. Testons la consommation

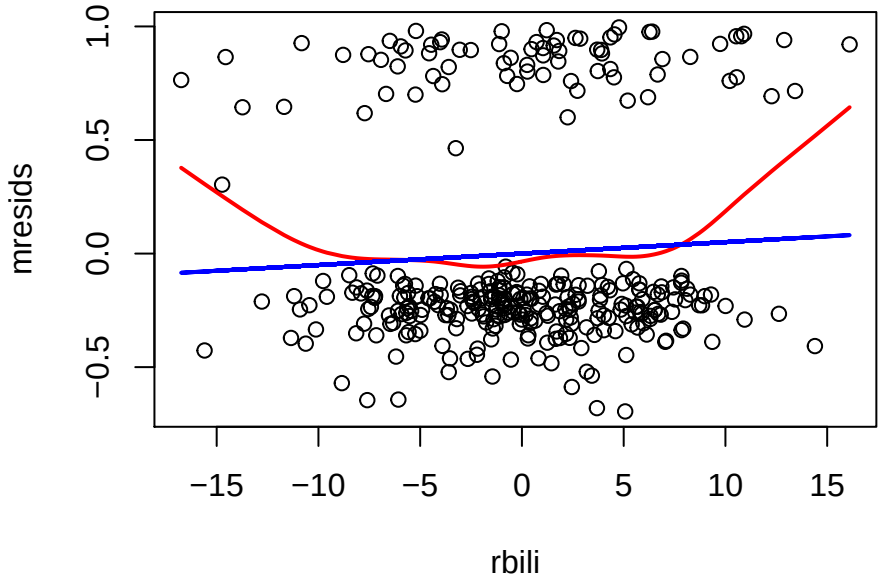


Il n'y a pas besoin d'appliquer de transformations, les résidus ne sont pas clairement entassé comme l'exemple fibre.

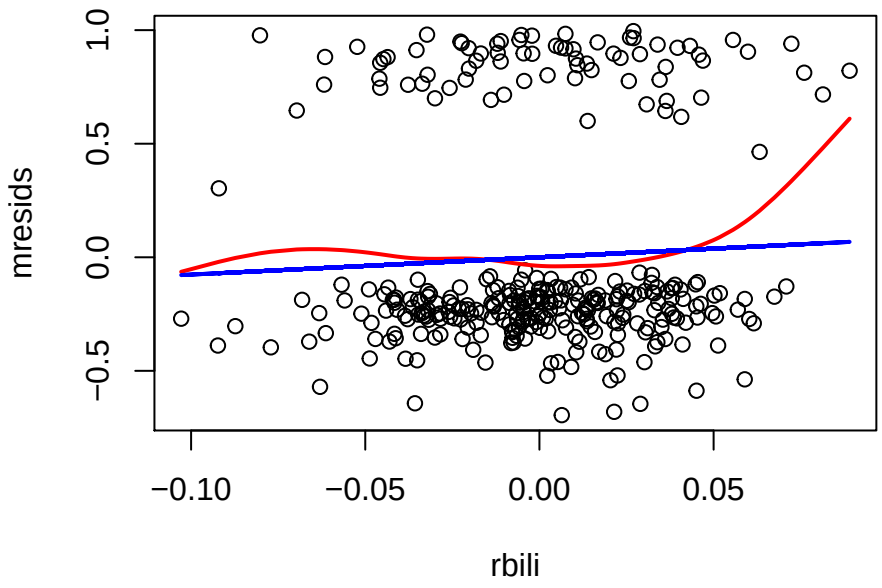
Testons pour graisse :



Comme précédemment, nous n'avons pas besoin d'appliquer de transformation.
Testons désormais la taille :



Essayons de voir ce qu'il se passe lorsque nous appliquons le logarithme :

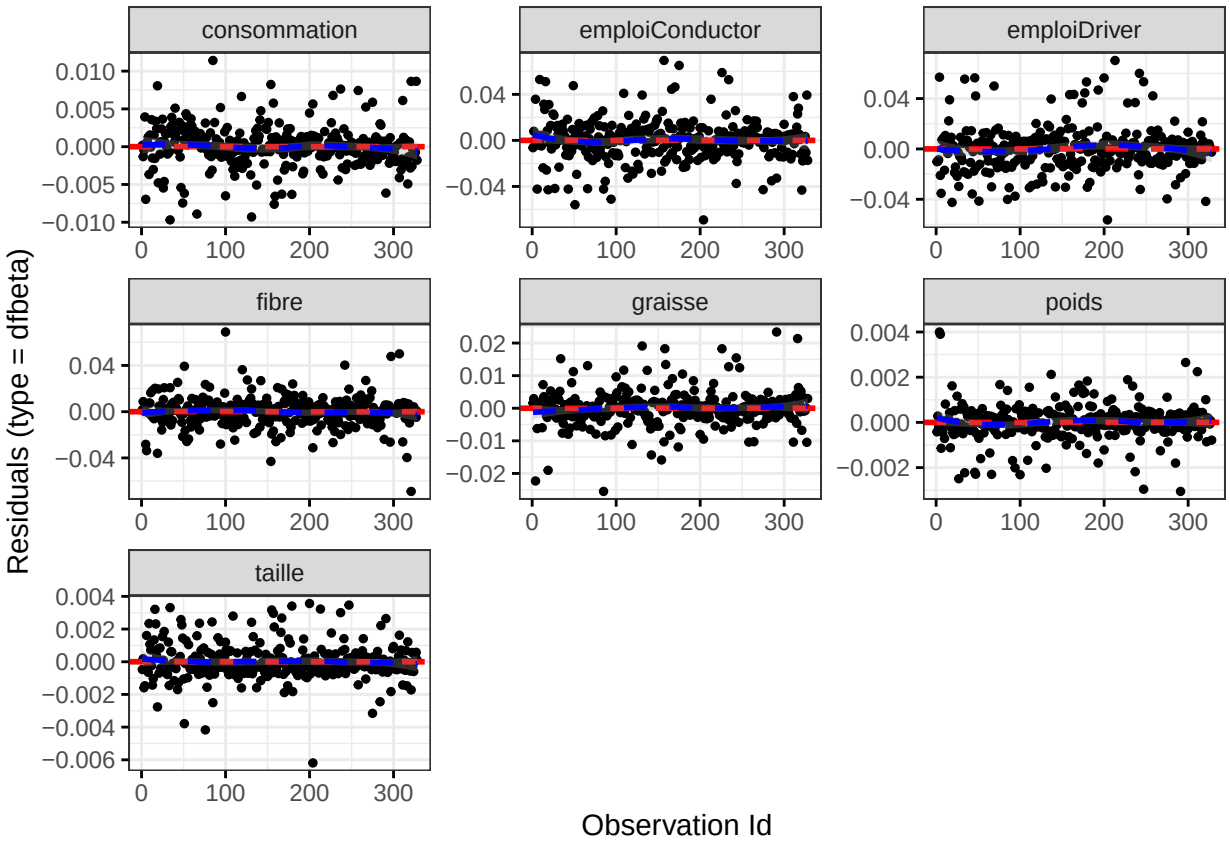


On la courbe parait plus stable. On fera donc une transformation logarithme sur la variable taille.

Examiner les observations influentes (ou les valeurs aberrantes).

Pour tester des observations influentes ou des valeurs aberrantes, nous pouvons utiliser les valeurs *dfbeta*. Le principe est de comparer le coefficient de une variable donnée lorsque un point donnée participe ou pas à la régression. On mesure l'influence d'un point sur le coefficient estimé.

```
ggcoxdiagnostics(res, type = "dfbeta",
                 linear.predictions = FALSE, ggtheme = theme_bw())
```



Nous voyons que que la comparaison des valeurs les plus élevées aux coefficients de régression suggère qu'aucune des observations n'est terriblement influente individuellement.

Test de l'hypothèse des risques proportionnels.

Regardons désormais si l'hypothèse d'indépendance du temps des formes multiplicatives et des covariables est vérifiée. L'hypothèse des risques proportionnels peut être vérifiée à l'aide de tests statistiques et de diagnostics graphiques basés sur les résidus de Schoenfeld.

```
res=coxph(survie~log(fibre)+log(taille)+poids+consommation+
            emploi+graisse,id=id,data=coeur2)
res.c=cox.zph(res)
res.c
```

```
##                chisq df     p
## log(fibre)    0.0159  1 0.900
## log(taille)   0.0155  1 0.901
## poids         0.0730  1 0.787
## consommation 2.9173  1 0.088
## emploi       4.2252  2 0.121
## graisse      1.5216  1 0.217
## GLOBAL       9.3482  7 0.229
```

Le test conduit à ne pas rejeter cette hypothèse au seuil de 5% : aucune covariable n'a un effet dépendant du temps.

Modèle de Cox

Maintenant que nous avons vérifié ces trois hypothèses, nous allons analyser le modèle de Cox. Commençons par introduire toutes les variables dans le modèle.

```
res=coxph(survie~log(fibre)+log(taille)+poids+emploi+
            graisse,id=id,data=coeur2)
summary(res)
```

```
## Call:
## coxph(formula = survie ~ log(fibre) + log(taille) + poids + emploi +
##     graisse, data = coeur2, id = id)
##
##   n= 328, number of events= 76
##
##                       coef  exp(coef)   se(coef)      z Pr(>|z|)
## log(fibre)      -0.3890266  0.6777163  0.4414874 -0.881   0.3782
## log(taille)     -8.1322981  0.0002939  3.5218163 -2.309   0.0209 *
## poids            0.0002774  1.0002775  0.0139152  0.020   0.9841
## emploiConductor  0.0279616  1.0283562  0.3214823  0.087   0.9307
## emploiDriver    -0.0623337  0.9395693  0.2985412 -0.209   0.8346
## graisse         -0.0484210  0.9527326  0.0571941 -0.847   0.3972
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## log(fibre)      0.6777163     1.4755 2.853e-01    1.6101
## log(taille)     0.0002939  3402.6100 2.954e-07    0.2924
## poids           1.0002775     0.9997 9.734e-01    1.0279
## emploiConductor 1.0283562     0.9724 5.476e-01    1.9310
## emploiDriver    0.9395693     1.0643 5.234e-01    1.6867
## graisse         0.9527326     1.0496 8.517e-01    1.0657
##
## Concordance= 0.617  (se = 0.031 )
## Likelihood ratio test= 13.15  on 6 df,   p=0.04
## Wald test            = 13.62  on 6 df,   p=0.03
## Score (logrank) test = 13.68  on 6 df,   p=0.03
```

Le test de Wald teste l'effet d'une covariable avec les autres covariable dans le modèle. S'il n'est pas significatif, cela ne veut pas dire qu'il ne le serait pas dans le modèle constitué uniquement de cette covariable.

Le test de Wald pour la covariable « taille » montre que le coefficient correspondant est fortement significatifs au seuil 5% $p_{values} < 0.05$. Les autres covariables ne modifient pas significativement cette incidence lorsque taille est dans le modèle.

Nous sélectionnons les variables pas à pas c'est à dire que nous enlèvons celle dont la p_{valeur} est la plus élevée. Ensuite nous refaisons tourner le modèle et nous recommençons jusqu'à obtention de toutes les variables significatives.

Nous obtenons le modèle suivant :

```
res=coxph(survie~log(taille),id=id,data=coeur2)
summary(res)
```

```
## Call:
## coxph(formula = survie ~ log(taille), data = coeur2, id = id)
##
##   n= 328, number of events= 76
```

```
##
##                   coef  exp(coef)   se(coef)      z Pr(>|z|)
## log(taille) -9.474e+00  7.685e-05  2.904e+00 -3.262  0.00111 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##             exp(coef) exp(-coef) lower .95 upper .95
## log(taille) 7.685e-05      13013 2.592e-07   0.02279
##
## Concordance= 0.588  (se = 0.03 )
## Likelihood ratio test= 10.35  on 1 df,   p=0.001
## Wald test            = 10.64  on 1 df,   p=0.001
## Score (logrank) test = 10.66  on 1 df,   p=0.001
```

```
confint(res)
```

```
##                 2.5 %    97.5 %
## log(taille) -15.16573 -3.781619
```

En supposant que toutes les autres covariables sont fixées, l'augmentation de la taille de 1cm multiplie l'incidence instentannée de la maladie par au moins, $exp(-15.16573) = 2.591834e-07$ et au plus $exp(-3.781619) = 0.02278577$.

Auparavant, nous avont constater que dans le modèle constitué de toutes les covariables, tous les autres coefficients n'étaient pas significatifs au seuil de 5 pourcents. Mais qu'en est t-il si on test avec un modèle avec seulement 1 covariable. Regardons ce qui se passe pour fibre :

```
res=coxph(survie~log(fibre),id=id, data=coeur2)
summary(res)
```

```
## Call:
## coxph(formula = survie ~ log(fibre), data = coeur2, id = id)
##
##   n= 328, number of events= 76
##
##               coef exp(coef) se(coef)      z Pr(>|z|)
## log(fibre) -0.8061    0.4466   0.3699 -2.179   0.0293 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
## log(fibre)    0.4466      2.239    0.2163    0.9221
##
## Concordance= 0.569  (se = 0.032 )
## Likelihood ratio test= 4.71  on 1 df,   p=0.03
## Wald test            = 4.75  on 1 df,   p=0.03
## Score (logrank) test = 4.7  on 1 df,   p=0.03
```

```
confint(res)
```

```
##                2.5 %      97.5 %
## log(fibre) -1.531084 -0.08111641
```

« fibre » a un coefficient significatif au seuil 5% car $p_{values} = 0.0293 < 0.05$.

En supposant que toutes les autres covariables sont fixées, l'augmentation de fibre de 1 unité multiplie l'incidence instantannée de la maladie par au moins $exp(-1.531084) = 0.2163011$.

Testons maintenant pour la variable graisse

```
 res=coxph(survie~coeur2$graisse,id=coeur2$id)
summary(res)
```

```
## Call:
## coxph(formula = survie ~ coeur2$graisse, id = coeur2$id)
##
##   n= 328, number of events= 76
##
##                     coef exp(coef) se(coef)      z Pr(>|z|)
## coeur2$graisse -0.09396   0.91032  0.05125 -1.833   0.0668 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                exp(coef) exp(-coef) lower .95 upper .95
## coeur2$graisse    0.9103      1.099    0.8233     1.007
##
## Concordance= 0.564  (se = 0.033 )
## Likelihood ratio test= 3.5  on 1 df,   p=0.06
## Wald test            = 3.36  on 1 df,   p=0.07
## Score (logrank) test = 3.35  on 1 df,   p=0.07
```

```
confint(res)
```

```
##                      2.5 %      97.5 %
## coeur2$graisse -0.1944183 0.006496239
```

« graisse » a une $p_{values} = 0.0668 > 0.05$. De plus, l'intervalle de confiance n'est pas interpretable car il contient 0. Nous ne pouvons pas faire d'estimation avec ce modèle.

```
res=coxph(survie~coeur2$emploi,id=coeur2$id)
summary(res)
```

```
## Call:
## coxph(formula = survie ~ coeur2$emploi, id = coeur2$id)
##
##   n= 328, number of events= 76
##
##                           coef exp(coef) se(coef)     z Pr(>|z|)
## coeur2$emploiConductor 0.4637    1.5899   0.2786 1.664   0.0961 .
## coeur2$emploiDriver    0.2240    1.2511   0.2841 0.788   0.4305
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                        exp(coef) exp(-coef) lower .95 upper .95
## coeur2$emploiConductor     1.590     0.6290    0.9209     2.745
## coeur2$emploiDriver        1.251     0.7993    0.7168     2.183
##
## Concordance= 0.548  (se = 0.031 )
## Likelihood ratio test= 2.73  on 2 df,   p=0.3
## Wald test            = 2.78  on 2 df,   p=0.2
## Score (logrank) test = 2.82  on 2 df,   p=0.2
```

```
confint(res)
```

```
##                               2.5 %    97.5 %
## coeur2$emploiConductor -0.08243578 1.0097674
## coeur2$emploiDriver    -0.33288997 0.7808997
```

Les intervalles de confiance comprennent 0, rien n'est interpretable.

Conclusion

Nous avons dans ce projet, appliqué toutes les méthodes que nous avons étudié en cours, sur cette table de donnée, ou du moins, celles qui étaient possible d'appliquer. Nous avons compris globalement compris les hypotheses du modèle de Cox qu'il fallait validé. En les appliquant, nous nous sommes rendu compte qu'il était assez compliquer de faire des interprétations précises sur la base de graphiques seulement et toutes les sources que nous avons trouvé utilisait ces méthodes graphiques. Malheureusement avec le peu de temps que nous avons, il nous est difficile d'approfonfire plus nos recherches pour cette partie, mais l'essentielle à été compris.

Au vu de nous résultat, nous pouvons dire que manger plus de fibre pourrais contribuer à la diminution de maladie.

Bibliographie

Proportional Hazards Regression Diagnostics, Dan Gillen, 2016

STHDA $http://www.sthda.com/english/wiki/cox-model-assumptions$

Cours7- Exemple et programmation $http://iml.univ-mrs.fr/\ reboul/duree62011.pdf$

Tests dans les modèles de durée $http://iml.univ-mrs.fr/\ reboul/duree52013.pptx.pdf$

Le modèle de Cox $https://perso.univ-rennes1.fr/valerie.monbet/ExposesM2/2013/cox_model.pdf$