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 intituts 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 enquete sur les habitudes alimentaires. Les réponses ont été recueilli dans une base de donnée que l'on nomme **Coeur**.

Présentation de la base de donnée

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

- **Id** : identifiant du sujet
- **DateEntrée** et **Date de sortie** : les dates d'netrée 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 (on code dont la signification n'est pas précisée ici). Si l'individu est sain à la sortie de l'enqûete, 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** (en cm et en kg)
- **Graisse** : quantité moyenne de graisse ingérée par jour (g/jour).
- **Fibres** : 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 consomation
- **MCC** : une variable binaire, recodage de la variable statut(1=MCC, 0= pas de MCC)

Voici les 5 premières ligne de notre table de donnée :

```
coeur <- readRDS('data/my_data_frame.rds')
coeur <-coeur%>%drop_na()
coeur <-coeur%>%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)



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

id	dateEntree	dateSortie	dateNaissance	statut	emploi
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	MCC
<=2750 KCals	0
<=2750 KCals	0
<=2750 KCals	1
<=2750 KCals	1
<=2750 KCals	1
<=2750 KCals	1

Cet ensemble d'individu est bien une cohorte car on a relever certaines covariables et les trois données fondammentales qui sont, la date d'entrée dans l'étude, la date de sortie dans l'étude et le cause de sortie dans l'étude.

On peut ajouter que les covariables utilisées dans l'étude sont fixe.

Nous observons aussi que la table de données contient des valeurs manquantes. Nous enleverons donc chaque ligne qui contient au moins une valeur manquante.

On passe donc de 337 à 328 individus.

statistique descriptive

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

Faisons un sommaire des variables quantitatives

```
df_quanti<-coeur%>%select(consommation,fibre,graisse,taille,poids)%>%summary()
pander(df_quanti)
```

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

En moyenne, les individus mangent 2835 calories par jours, ils ingèrent 12.76 gramme de gras par jour en moyenne, ils mesurent 173 cm et pèsent 72.40 kilo.

Faisons un sommaire des variables quantitatives

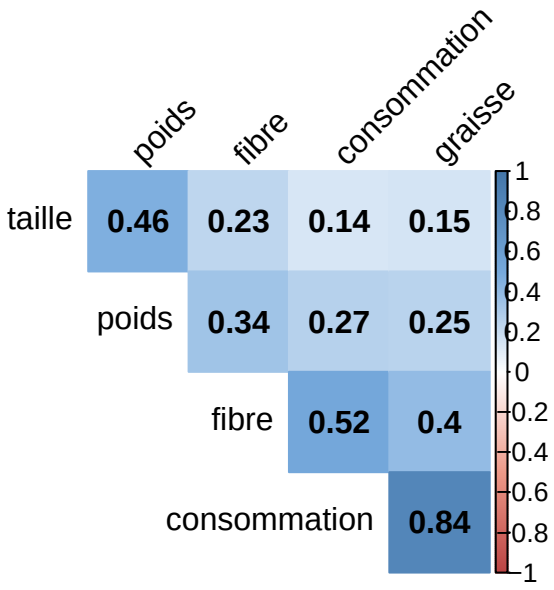
```
df_quali<-coeur%>%select(statut,emploi,moisEnqu_e,hauteConsomation)%>%summary()
pander(df_quali)
```

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.

Désormais, regardons la corrélation entre les variables quantitatives :

```
df_quanti1<-coeur%>%select(consommation,fibre,graisse,taille,poids)
col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA"))
corrplot(cor(df_quanti1), method="color", col=col(200), type="upper", order="hclust",
addCoef.col = "black", tl.col="black", tl.srt=45, diag=FALSE)
```



On observe que toutes les variables sont corrélées positivement.

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

La variable **taille** est peu corrélée avec la variable **graisse**

regression logistique

modele de cox

test