---
title: "devoir MAN 2016"
author: "sarah FELDMAN"
date: "21 octobre 2016"
output: html_document

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

Q1/Dans un premier temps décrivez vos variables.
```{r , echo=FALSE,warning=FALSE,message=FALSE}
library(psy)
library(corrplot)
library(gridExtra)
library(ggplot2)
#library(car)
library(dplyr)
library(grid)
library(knitr)

rt <- read.csv2("data/presentationTPretinol.csv")
```

Nombre de variables

```{r, echo=FALSE, eval= FALSE}
dim(rt)
nrow(rt)
ncol(rt)
```
Le jeu de données présentation retinol fait `r nrow(rt)` lignes et `r ncol(rt)` colonnes ou variables.

Type de variable

```{r, echo=FALSE}
rt$sexe <-as.numeric(recode(rt$sexe,0,1)) #1->0=masc 2->1=fem ; recode from package dplyr. Je veux garder la structure numérique.
rt$vitamine <- factor(rt$vitamine, levels=c(1,2,3), labels=c("souvent","parfois","non"))
#rt$vitamine <- relevel(rt$vitamine, ref="non")

rt$tabac <- factor (rt$tabac, levels=c(1,2,3), labels=c("jamais","autrefois","actuellement"))
```

Les variables sont toutes reconnues comme quantitatives
mais en regardant plus attentivement les min et max on peut voir que certaines variables ne sont pas quantitatives:

- sexe : variable binaire

sexe a un minimum de 1 et un maximum de 2 c'est donc une variables qualitative binaire. Je la recode en 0(homme)/1(femme) pour une interprétation plus facile par la suite. Je m'assure que c'est bien une valeur numérique pour pouvoir l'intégrer dans ma matrice de corrélation en question 2.

- tabac et vitamine : variables qualitatives ordinales

Les variables tabac et vitamine sont des variables qualitatives ordinales, il y a bien un ordre dans les classes mais je ne peux pas dire si l'espace entre 2 classes est le même. Pour tabac par exemple : ne pas fumer versus fumer anciennement aura-t-elle la même différence d'effet que fumer anciennement versus fumer actuellement? Je préfère donc les coder comme des variables qualitatives en attendant d'en savoir plus. Pour cela je les transforme simplement en facteur.
Je garde comme classe de référence "souvent" pour vitaline et "jamais" pour tabac car c'est ce qui a pour moi le plus de sens médicalement parlant. Je n'indique pas qu'elles sont ordonnées car je veux les analyser d'abord comme étant non ordonnées.

- Les autres variables sont quantitatives continues:
age, bmi, calories, graisses, fibres, alcool, cholestérol, betadiet,retdiet,betaplasma,retplasma


```{r , echo=FALSE}
# quantcont<- c("age", "bmi", "calories", "graisses", "fibres", "alcool",
"cholestérol","betadiet","retdiet","betaplasma","retplasma")
#rt[,quantcont] ne marche pas : je recherche si je n'ai pas mal écrit une des variables :
# quantcont[!quantcont %in% names(rt)] : cholestérol au lieu de cholesterol
quantcont<- c("age", "bmi", "calories", "graisses", "fibres", "alcool",
"cholesterol","betadiet","retdiet","betaplasma","retplasma")
qual.ord <- c("tabac","vitamine")
binaire <- c("sexe")
#grid.table(cbind(variables.quantitatives.continues=quantcont,variables.quantitatives.discretes=quantdis,variable.binaire=binaire))

kable(cbind(variables.quantitatives.continues=quantcont,variables.qualitatives.orinales=c(qual.ord,
rep("",9)),variable.binaire=c(binaire,rep("",10)))) #je dois rajouter des "" à la fin des vecteurs trop courts pour pouvoir créer un tableau
```


Données abberrantes

Observer ses variables sert également à dépister les "bizarreries".
Par exemple, j'ai un doute concernant le max d'alcool, je vais donc regarder plus attentivement la variable, par exemple avec une table.
```{r , echo = FALSE, eval=FALSE}
table(rt$alcool)
```

```
Je passe de 35 à 203 verres par semaine...203/7=29 verres par jour, il y a probablement un erreur de codage.

La valeur extrême d'alcool influe beaucoup sur la symétrie de ma distribution. On peut le voir notamment en comparant la moyenne (`r round(mean(rt$alcool),2)`) et la médiane (`r median(rt$alcool)`). Cependant je ne sais pas si je peux la retirer sans conséquence, je préfère donc la garder pour la suite du devoir.


## Distribution des variables

```{r , message=FALSE,echo=FALSE}

non_cont<- c("sexe.bin","tabac.f","vitamine.f")
names_rt <- names(rt)[!names(rt) %in% sapply(c("sexe","tabac","vit"),grep,names(rt),value=T)] #grep est utile si je donne à ces 3 variables des noms avec suffixe

# qage <- qplot(rt$age,binwidth=1)
# qbmi <- qplot(rt$bmi,binwidth=1)
# qcalories <- qplot(rt$calories,binwidth=50)
# qgraisses <- qplot(rt$graisses,binwidth=2)
# qfibres <- qplot(rt$fibres,binwidth=0.5)
# qalcool <- qplot(rt$alcool,binwidth=3)
# qcholesterol <- qplot(rt$cholesterol,binwidth=50)
# qbetadiet <- qplot(rt$betadiet,binwidth=40)
# qretdiet <- qplot(rt$retdiet,binwidth=40)
# qbetaplasma <- qplot(rt$betaplasma,binwidth=30)
# qretplasma <- qplot(rt$retplasma,binwidth=30)
#pl <- lapply(names_rt, function(x)get(paste0("q",x)))

pl <- lapply(names_rt, function(.x) qplot(rt[,.x],xlab=NULL, main=paste("distribution de ", .x),fill=I("navajowhite3"), col=I("pink4")))

ml <- marrangeGrob(pl,ncol=2,nrow=3,top = NULL)
print(ml)
```

Lecture des histogrammes :

- bmi, graisses, cholestérol, betadiet, retdiet et betaplasme on des distribution à peu près normales mais asymétriques.
- calories, fibres, ret plasma ont de distributions d'allure normales
- age a une distribution irrégulière avec 2 cloches à 40 et 75 ans.
- alcool a une distribution très asymétrique qui ne semble pas normale

```{r, echo=FALSE}

varnorm<- c("retplasma", "bmi", "cholesterol", "retdiet", "graisses", "betadiet", "betaplasma", "calories", "fibres")
varnnorm<-c("age", "sexe", "tabac", "vitamine", "alcool","", "", "", "")

#grid.table(cbind(variables.normales=varnorm,variables.non.normales=varnnorm))
kable(cbind(variables.normales=varnorm,variables.non.normales=varnnorm))

```


Voici un résumé des différentes variables :

- pour les variables quantitatives :

```{r , echo=FALSE}
kable(data.frame (
  #var=names(rt[quantcont]),
  n=apply(rt[quantcont],2,function(x)sum(!is.na(x))),
  missing=apply(rt[quantcont],2,function(x)sum(is.na(x))),
  moyenne =round(apply(rt[quantcont],2,mean,na.rm=T),1),
  median=round(apply(rt[quantcont],2,median,na.rm=T),1),
  q1= round(apply(rt[quantcont],2,quantile,na.rm=T),1)[2,],
  q3= round(apply(rt[quantcont],2,quantile,na.rm=T),1)[4,],
  rmin=round(apply(rt[quantcont],2,min,na.rm=T),1),
  max=round(apply(rt[quantcont],2,max,na.rm=T),1),
  distribution.normale=sapply(names(rt[quantcont]),function(x)ifelse(x %in% varnorm, "oui",ifelse(x %in% varnnorm, "non",NA)))
),rownames="variables")
```

- pour les variables qualitatives:

```{r, echo=FALSE}
t1 <- cbind(table(rt$sexe,useNA="a"),round(prop.table(table(rt$sexe,useNA="a"))*100,0))
tot1 <- cbind(sum(table(rt$sexe)),sum(prop.table(table(rt$sexe,useNA="a"))*100))
t1 <- rbind(t1,tot1)
t1 <- t1[c(1,2,4,3),]
rownames(t1) <- c("0 (homme)","1 (femme)","N", "missing")
colnames(t1) <- c("Fréquence","Pourcentage")
t1 <- data.frame(t1)
t1$sexe <- rownames(t1)
rownames(t1) <- NULL
t1 <- t1[,c(3,1,2)]
kable(t1,rownames="sexe")

t2 <- cbind(table(rt$vitamine,useNA="a"),round(prop.table(table(rt$vitamine,useNA="a"))*100,0))
tot2 <- cbind(sum(table(rt$vitamine)),sum(prop.table(table(rt$vitamine,useNA="a"))*100))
t2 <- rbind(t2,tot2)
t2 <- t2[c(1,2,3,5,4),]
#rownames(t2) <- c(1,2,3,"N", "missing")
rownames(t2) <- c(rownames(t2)[1:3],"N", "missing")
```
```

```
colnames(t2) <- c("Fréquence","Pourcentage")
t2 <- data.frame(t2)
t2$vitamine <- rownames(t2)
rownames(t2) <- NULL
t2 <- t2[,c(3,1,2)]
kable(t2,rownames="vitamine")

t3 <- cbind(table(rt$tabac,useNA="a"),round(prop.table(table(rt$tabac,useNA="a"))*100,0))
tot3 <- cbind(sum(table(rt$tabac)),sum(prop.table(table(rt$tabac,useNA="a"))*100))
t3 <- rbind(t3,tot3)
t3 <- t3[c(1,2,3,5,4),]
#rownames(t3) <- c(1,2,3,"N", "missing")
rownames(t3) <- c(rownames(t3)[1:3],"N", "missing")
colnames(t3) <- c("Fréquence","Pourcentage")
t3 <- data.frame(t3)
t3$tabac <- rownames(t3)
rownames(t3) <- NULL
t3 <- t3[,c(3,1,2)]
kable(t3,rownames="tabac")

#grid.arrange(tableGrob(t2,rows=NULL),tableGrob(t3,rows=NULL),tableGrob(t1,rows=NULL),ncol=2)
```


Exemples de représentation graphique des variables :

Diagrammes pour tabac et vitamine
Je veux représenter des pourcentages, un barplot est pour moi plus parlant qu'un camembert (pie).

```{r, echo=FALSE}
#pl <- lapply(c("tabac","vitamine"), function(.x) qplot(rt[,.x],xlab=NULL, main=paste("plot", .x),fill=I("navajowhite3"),
asp="",col=I("pink4"))) #je narrive pas à changer l'échelle des ordonnées avec qplot, je passe à ggplot

pl <- lapply(c("tabac","vitamine"), function(.x) {
  ggplot(rt, aes(x = rt[,.x])) +
    geom_bar(aes(y = (..count..)/sum(..count..)),fill=I("lightsteelblue2"), col=I("paleturquoise4")) +
    xlab(.x) + ylab("percent")+ ggtitle (paste0("Répartition de ",.x)) +
    scale_y_continuous(labels = scales::percent)
})

ml <- marrangeGrob(pl,nrow=1,ncol=2,top = NULL)
print(ml)
```


Boîtes à moustache pour les variables quantitatives comme age ou ret plasma

```{r, echo=FALSE}
pl <- lapply(c("age","retplasma"), function(.y) {
  ggplot(rt, aes(y = rt[,.y], x=1)) +
    stat_boxplot(geom = "errorbar", width = 0.5, color="lightsteelblue4") +
    geom_boxplot(fill = "lightsteelblue2", color = "lightsteelblue4", width=1) +
    scale_x_discrete() + xlab(NULL) +ylab(ifelse(.y=="age","age (an)","rétinol plasmatique (ng/ml)")) + ggtitle
(paste0("Dispersion de ", .y))
})
ml <- marrangeGrob(pl,nrow=1,ncol=2,top = NULL)
print(ml)

```


Q2/Etudiez les relations existant entre toutes les paires possibles de variables.

NB : ce ne sont que les 9 variables de la régression demandées en question 3 qui sont concernées.

```{r,echo=FALSE}
#var <- c("retplasma","age","sexe","bmi","tabac","vitamine","cholesterol","alcool","retdiet") #version lorsque tabc et vitamines
sont quantitatives
var <- c("retplasma","age","sexe","bmi","cholesterol","alcool","retdiet")
#var %in% names(rt)#pour voir quelle variable j'ai mal recopié
```


Matrice de corrélation

Je peux faire une matrice de corrélation. Je n'inclue ni vitamine ni tabac dans cette matrice car je les ai considérées comme des variables qualitatives à plusieurs classes. Je garde sexe qui est binaire.

Aucune condition n'est nécessaire pour faire des coefficients de corrélation. C'est pour les tester que nous avons besoin de vérifier les confidions de validité.

```{r, echo = FALSE}
mat <- round(cor(rt[,var]),3)
#print(mat)
kable(mat)
```

Il faut ensuite interprêter la matrice. Pour cela je peux faire des schémas:

```{r , echo=FALSE}
corrplot(cor(rt[,var]),method="circle")
```

Dans cette visualisation de la matrice de corrélation, on voit surtout la corrélation entre cholestérol et retdiet car les cercles ont une densité élevée et on voit que la corrélation est positive car le cercle est de couleur bleu.

Je trouve que la représentation graphique de l'analyse en composante principale s'interprête plus facilement :

```{r, echo=FALSE}
mdspca(rt[,var])
```

- Lecture de l'acp :
    + bmi est très proches du centre du cercle donc non interprétable.
    + les paires retdiet-cholestérol, retplasma-age sont fortement associés (corrélation positive), et ces deux groupes de variables sont indépendants l'un de l'autre car forme un angle droit avec le centre.
    + sexe est également indépendant de retdiet et cholestérol.
    + alcool-sexe est négativement corrélée.

- Je peux aussi sélectionner dans la matrice les valeurs absolues supérieures ou égales à certain niveau de corrélation :

```{r,echo=FALSE}
couples<-lapply(c(0.2,0.4),function(w){
  #pour supprimer les doublons
  mat2<- lower.tri(mat,diag=FALSE)
  rownames(mat2)<-rownames(mat)
  colnames(mat2) <- colnames(mat)
  mat2 <- ifelse(mat2==TRUE,mat,0)
  #pour chercher les coefficients de corrélation superieur à w
  w_r <- which(abs(mat2)>=w )
  #pour trouver les noms de ligne et colonne de ces coefficients
  which_couple <- lapply(w_r,function(x){
    k <- arrayInd(x, dim(mat2))
    #paste(rownames(mat2)[k[,1]],"-", colnames(mat2)[k[,2]])
    d<-data.frame(var1=rownames(mat2)[k[,1]], var2=colnames(mat2)[k[,2]],r=mat2[x])
   # colnames(d)<- c("variable 1","variable 2", "coefficient de corrélation")
    return(d)
  })
  #Je colle les listes
  #browser()
  which_couple <- data.frame(do.call(rbind,which_couple))
  #Je nomme les 2 listes de niveau supérieur selon la valeur du coefficient
  #colnames(which_couple) <- c(paste0("var1.r",w),paste0("var2.r",w))
  return(which_couple)
})
couples_rename <- couples
colnames(couples_rename[[1]])<- c("variable 1","variable 2", "coefficient de corrélation")
kable(couples_rename[[1]])

```

4 couples ont un coefficient de corrélation entre 0.2 et 0.4 (en nombre absolu).
age-retplasma, sexe-age, cholesterol-sexe, alcool-sexe.

Il n'y a qu'un couple avec un coefficient de corrélation supérieur ou égal à 0.4 :
retdiet-cholestérol

Tests de corrélation

Pertinence
- Est-ce pertinent de faire de tests de corrélation pour chaque variable?
On peut se poser la question, en effet je n'avais aucune hypothèse de départ quant à ces corrélations et multiplier le nombre de tests augmente le risque alpha. Cependant pour les besoins du devoir je le fais quand même, mais je ne testerai que les corrélations supérieur à 0.2. En effet je ne sais pas quel sens je donnerai à une corrélation significative de 0.01 par exemple...

Conditions de validité
Avant de faire une test de corrélation, il faut tester les conditions de validité :

- Une des 2 variables du couple testé doit suivre une loi normale.

Je considère comme normale une variable dont l'histogramme montre une distribution en cloche. Lorsque la distribution ne semble pas normale, il faut interpréter les tests avec prudence. Je ne préfère pas faire des tests non paramétriques type test de corrélation de spearman, ni sur certaines variables pour garder une cohérence, ni sur toutes les variables pour ne pas m'empêcher de faire des tests paramétriques par la suite.

Résultats

```{r, echo=FALSE}
couplesb <- couples[[1]]
couplesb$testvalid <- ifelse( couplesb$var1 %in% varnorm | couplesb$var2 %in% varnorm,TRUE,FALSE)
couplesb$test<-sapply(1:nrow(couplesb), function(x) {
  .var1 <- couplesb[x,1]
  .var2 <- couplesb[x,2]
  #browser()
  .var1rt <- rt[,as.character(.var1)]
  .var2rt <- rt[,as.character(.var2)]
  testcouple<-cor.test(.var1rt,.var2rt)
  # if couplesb$testvalid[x] testcouple<-cor.test(.var1rt,.var2rt)
  # else testcouple <- cor.test(.var1rt,.var2rt, method="spearman") #je préfère ne pas utiliser du non paramétrique car m'empêcherait de faire des analyses plus poussées ensuite
  ##browser()
  pcor<-round(testcouple$p.value,5)
  return(pcor)
})
```

```
kable(couplesb)
#couples$test <- cor.test(get(paste0("rt$",var1)),get(paste0("rt$",var2)))
```

Toutes les corrélations supérieures ou égales à 0.2 sont significatives car p<= 0.05. Il faut cependant prêter attention au fait que 2 couples ont des conditions de validité probablement non remplies : sexe-age et sexe-alcool car age et alcool n'ont pas une allure normale (et sexe est binaire donc ne peut pas être normale).

étude des liens entre une variable qualitative à plusieurs classe et des variables quantitatives:

Graphiquement: je regarde la dispersion des variables quantitatives dans les différentes classes grâce à des boxplots

```{r, echo=FALSE}
sapply(c("tabac","vitamine"),function(.i){
  pl <- lapply (var[! var %in% "sexe"], function(.x) {
    ggplot(rt, aes(x = get(.i), y = get(.x))) +
      geom_boxplot(fill = "lightsteelblue2", color = "lightsteelblue4") +
      scale_x_discrete() + xlab(.i) + ylab(.x) +ggtitle (paste0("dispersion de ",.x," selon ",.i))
  })

  ml <- marrangeGrob(pl,ncol=2,nrow=3,top = NULL)

  print(ml)

})
```

Comparaison de moyenne entre plusieurs groupes (plus de 2) : ANOVA

Conditions de validité :

- Variance du même ordre de grandeur dans tous les sous groupes :

```{r, echo=FALSE}
var_tab <- data.frame(round(sapply(var[! var %in% "sexe"], function(.x) by(rt[,.x],rt$tabac,sd,na.rm=T)),0))
var_tab$tabac <- levels(rt$tabac)
var_tab <- var_tab[,c(7,1:6)]
rownames(var_tab) <- NULL
kable(var_tab)

var_vit <- data.frame(round(sapply(var[! var %in% "sexe"], function(.x) by(rt[,.x],rt$vitamine,sd,na.rm=T)),0))
var_vit$vitamine <- levels (rt$vitamine)
var_vit <- var_vit[,c(7,1:6)]
rownames (var_vit) <- NULL
kable(var_vit)

#grid.arrange(tableGrob(var_tab,rows=NULL),tableGrob(var_vit,rows=NULL),nrow=2)
```

C'est le cas partout sauf pour la variable alcool. Je préfère donc ne pas faire de test avec alcool.

- Distribution normale :
Je fais l'approximation que toutes les variables suivent une loi normale, d'autant plus que l'anova est un test qui résiste bien à des ditribution qui s'éloignent un peu de la normale.

Je fais donc des ANOVA entre ma variable qualitative (vitamine ou tabac) et mes variables quantitatives (alcool exclue)
```{r, echo=FALSE}
names.vit.an <- c("tabac","vitamine")
tab.vit.an<-lapply(names.vit.an,function(.i){
anov<- data.frame(t(sapply(var[! var %in% c("sexe","alcool")], function(.x){
  #browser()
 res <- lm(get(.x) ~ get(.i), rt )
 dp <- drop1(res, test="F")
 pv <- round(dp$`Pr(>F)`[!is.na(dp$`Pr(>F)`)],3)
 #names(pv)<-.i
})))
})
tab.vit <- do.call(rbind,tab.vit.an)
tab.vit$variable_qualitative <- names.vit.an
tab.vit <- tab.vit[,c(6,1:5)]

kable(tab.vit)
#grid.arrange(tableGrob(tab.vit,rows=NULL))
```

interprétation :

- tabac est significativement associée à retplasma et age.
- vitamine est significativement associée à age.

Etude du lien entre variables qualitatives

Je fais un test du chi 2 pour les couples :

- tabac-vitamine
- tabac-sexe
- sexe-vitamine

Condition de validité : Effectifs théoriques supérieurs à 5
```{r, echo=FALSE,eval=FALSE}
table(rt$tabac,rt$vitamine) #De tête je multiplie les plus petits totaux ligne et colonne/total
```

```
table(rt$tabac,rt$sexe)
table(rt$sexe,rt$vitamine)
```

Tous les effectifs théoriques sont supérieurs à 5 pour les 3 tableaux de contingence, je peux faire un chi2.

Test du Chi2

Le seuil de significativité est p<=0.05 :

- tabac et vitamine ne sont pas significativement liées : `r round(chisq.test(rt$tabac,rt$vitamine, correct=FALSE)$p.value,3)`
- tabac et sexe sont significativement liées : p value = `r round(chisq.test(rt$tabac,rt$sexe, correct=FALSE)$p.value,3)`
- sexe et vitamine sont significativement liées : p value = `r round(chisq.test(rt$sexe,rt$vitamine, correct=FALSE)$p.value,3)`

Q3/Effectuez ensuite une régression linéaire où la variable à expliquer sera la concentration en rétinol plasmatique, les autres variables étant explicatives. Recherchez des interactions entre les variables explicatives.

Analyse en composante principale focalisée

Je peux commencer par regarder les interaction enter retplasma et les variables explicatives avec une acp focalisée. Je retire cependant vitamine et tabac car fpca se base sur une matrice de corrélation, il faut donc faire attention à l'interprêtation du schéma car les interactions seront peutêtre modifiées en rajoutant vitamine et tabac:

```{r , echo=FALSE}
fpca(retplasma ~ age + sexe + bmi + cholesterol + alcool + retdiet,data=rt)
```

2 variables semblent significativement liée à retplasma : age et sexe.

- age-retplasma est un couple qui ressortait dans la matrice de corrélation avec un r=0.212 significativement différent de 0.
- sexe-retplasma ne ressortait pas car inférieur à 0.2, comme on peut le voir sur l'acp focalisée.

régression linéaire multiple

La variable à expliquer retplasma étant une variable Quantitative, je peux faire une régression linéaire. Et dans la mesure où j'introduis plusieurs variables explicatives, ce sera une régression linéaire multiple. Vitamine et tabac feront parti du modèle, je me suis bien assurée auparavent de les coder en facteur afin qu'elles soit recodées automatiquement en (nclasses-1) variables binaires.

```{r}
mod <- lm(retplasma ~ age + sexe + bmi + tabac + vitamine + cholesterol + alcool + retdiet,rt)
```

Vérification des conditions de validité:
Il y a 3 conditions de validité aux modèles :

- la normalité de bruit
- la variance du bruit ne doit dépendre ni de la variable à expliquer ni de la variable explicative
- le bruit doit être du vrai bruit

En pratique on ne teste que la première des conditions :

```{r , echo=FALSE}
#hist(resid(mod)) #Je n'ai pas réussi à le transformer en ggplot
qplot(resid(mod),binwidth=100, fill=I("navajowhite3"), col=I("pink4"), main="distribution des résidus du modèle")
```

La distribution des résidus a une allure normale, mon modèle est donc valide.

Interprêtation des coefficients beta du modèle

```{r , echo=FALSE}
s <- summary(mod)
coef<-s$coefficients
estim <- data.frame(estimateur=round(as.numeric(coef[,1]),2))
ciinf <- as.numeric(round(confint(mod),2)[,1])
cisup <- as.numeric(round(confint(mod),2)[,2])
estim$CI95 <- paste0("[", ciinf," - ", cisup,"]")
estim$pval <- round(as.numeric(coef[,4]),3)
estim$signif <- ifelse (estim$pval<=0.05, "*","")
rownames(estim)<- rownames(coef)
kable(estim)
```

Légende : * p value <= 0.05 soit coefficient beta significativement différent de 0.

Interprétation :

- L'age et le sexe sont significativement associés à la concentration plasmatique en retinol (p value<=0.05 et un intervalle de confiance à 95% ne contenant pas 0). Quand l'âge du sujet augmente d'un an, la concentration augmente de 2.26 unités. J'avais recodé les femmes en 1 et homme en 0, donc les femmes ont une concentration plasmatique en rétinol plus faible que les hommes d'environ 100 unités.

- Le coefficient de tabac "autrefois" est significativement différent de "jamais". ceux qui fumaient autrefois ont donc une concentration plasmatique en rétinol significativement plus élevée de 58 unités par rapport à ceux qui n'ont jamais fumé. Par contre on ne retrouve pas cette augmentation chez ceux qui fument actuellement.

- Les autres estimateurs ne sont pas significatifs.

Je peux regarder si les variables tabac et vitamine sont globalement associées à retplasma:

Données obtenues avec la fonction drop1.

```{r, echo=FALSE}
dr <- drop1(mod,.~.,test="F")
coef <- data.frame(p.Value=round(dr$`Pr(>F)`[rownames(dr)%in% c("vitamine","tabac")],3)) #Je ne prends que les lignes vitamine et tabac
rownames(coef) <- rownames(dr)[rownames(dr)%in% c("vitamine","tabac")]
kable(coef)
```

Ni tabac ni vitamine ne sont globablement associées à retplasma.

Recherche d'interactions entre les variables explicatives:

```{r, echo=FALSE}
mod_inter <- data.frame(add1(mod,~.^2,test="Chisq")) #je ne sais pas si on peut utiliser F ici alors je préfère utiliser chisq qui marche pour lm aussi si j'ai bien compris.
isort <- order (mod_inter$Pr..Chi.)
mod_order <- mod_inter[isort,]
mod_order$Pr..Chi. <- round(mod_order$Pr..Chi.,4)
mod_order$signif <- ifelse(mod_order$Pr..Chi.<= 0.05,"*","")

.df<-mod_order
kable(.df)
```

Légende : * p value <= 0.05 soit interaction significative

Interprétation
J'ai 9 interactions significatives (p value<=0.05) : vitamine-alcool, cholestérol-alcool, bmi-alcool, age-alcool,sexe-bmi,alcool-retdiet, tabac-alcool,sexe-alcool,sexe-tabac.

Je trace la représentation graphique des pvalue des 28 interactions possibles :
```{r, echo=FALSE}
x <- 1-na.omit(mod_order$Pr..Chi.)
y <- length(x):1
g<-ggplot(na.omit(mod_order),aes(x,y))+
  geom_point(size=2)+
  xlab("1-p")+ylab("Np")
g+  geom_abline(slope=coefficients(lm(1:length(x) ~ -1 + x[length(x):1])),color="lightblue4", size=1)
```

J'ai bien 9 points au dessus de la ligne.

Q4/Transformez la variable "rétinol plasmatique" en une variable binaire (en la coupant en deux au niveau de la médiane). Refaites les calculs précédents en ayant recours cette fois à une régression logistique.

régression logistique : construction du modèle
La variable à expliquer est binaire, il faut donc faire une régression logistique

```{r, echo=FALSE}
rt$retplasma.bin <- ifelse(rt$retplasma < median(rt$retplasma,na.rm=TRUE), 0,1)
```

```{r}
mod.bin <- glm(retplasma.bin ~ age + sexe + bmi + tabac + vitamine + cholesterol + alcool + retdiet, data = rt, family = "binomial")
```

Vérification des conditions de validité

Pour que le modèle soit valide, il faut au moins 5 à 10 évènements par variable explicative.
J'ai 6 variables explicatives comptant comme une variable, et 2 variables comptant comme n(classe) - 1 variables soit 2 pour vitamines et 2 pour tabac (sexe est binaire et ne compte donc que pour une variable).
J'ai donc un total de 10 variables.
10*5=50
10*10=100

```{r, echo=FALSE, eval=FALSE}
table(rt$^retplasma.bin)
```

retplasma.bin : 157 sujets n'ont pas la variable, 158 l'ont. Donc 158 sujets ont des variables explicatives.
100 est inférieur à 158 donc le modèle est valable.

Interprêtation des coefficients beta du modèle

```{r , echo=FALSE}

s <- summary(mod.bin)
coef<-s$coefficients
estim <- data.frame(estimate=round(as.numeric(coef[,1]),4))
ciinf <- as.numeric(round(confint(mod.bin),4)[,1])
cisup <- as.numeric(round(confint(mod.bin),4)[,2])
estim$CI95 <- paste0("[", ciinf," - ", cisup,"]")
estim$pval <- round(as.numeric(coef[,4]),3)
```

```
estim$signif <- ifelse (estim$pval<=0.05, "*","")
rownames(estim)<- rownames(coef)
kable(estim)

```

Seul le coefficient beta de age est significativement différent de 0 (c'est également le seul à avoir un intervalle de confiance à 95% ne contenant pas 0).

Je ne peux pas interpréter les coefficients tels quels, j'interprète uniquement l'exponentiel de ces coefficients qui sont égales aux Odds ratio.

```{r, echo=FALSE}
e<- round(exp(coefficients(mod.bin)),3) #exponentiel du coefficient donne l'OR
e<-data.frame(OR=e)
e$CI95 <- paste0("[",round(exp(ciinf),3)," - ",round(exp(cisup),3),"]")
e$signif <- estim$signif
kable(e)
```

légende : * pvalue <= 0.05 soit OR significativement différent de 1
NB: pour retdiet l'intervalle de confiance est [1-1] à cause de l'arrondi mais contient 1 en réalité.

retplasma.bin n'est pas rare(car j'ai pris la médiane pour le construire...) donc les OR ne peuvent pas être interprétés comme des risques relatifs.
On voit cependant que le coefficient de l'âge a beau être significatif, le rapport des odds est très proche de 1 donc l'âge est significativement différent mais de très très peu...

Je peux regarder si les variables tabac et vitamine sont globalement associées à retplasma.bin:

Données obtenues avec la fonction drop1.

```{r, echo=FALSE}
dr <- drop1(mod.bin,.~.,test="Chisq") #Attention : test= Chisq pour la régression logistique
coef <- data.frame(p.Value=round(dr$`Pr(>Chi)`[rownames(dr)%in% c("vitamine","tabac")],3)) #Je ne prends que les lignes vitamine et tabac
rownames(coef) <- rownames(dr)[rownames(dr)%in% c("vitamine","tabac")]
kable(coef)
```

Ni tabac ni vitamine ne sont globablement associées à retplasma.

Recherche d'interactions entre les variables explicatives:

```{r, echo=FALSE}
mod_inter.bin <- data.frame(add1(mod.bin,~.^2,test="Chisq"))
isort.bin <- order (mod_inter.bin$Pr..Chi.)
mod_order.bin <- mod_inter.bin[isort.bin,]
mod_order.bin$Pr..Chi. <- round(mod_order.bin$Pr..Chi.,4)
mod_order.bin$signif <- ifelse(mod_order.bin$Pr..Chi.<= 0.05,"*","")

mytheme <- ttheme_default(base_size=10)

a <-tableGrob(mod_order.bin[1:14,], rows = rownames(mod_order.bin)[1:14],theme = mytheme)
b<-tableGrob(mod_order.bin[15:nrow(mod_order.bin),], rows = rownames(mod_order.bin)[15:nrow(mod_order.bin)],theme = mytheme)

.df.bin <- mod_order.bin
kable(.df.bin)

#ne pas utiliser grid car coupe les tableaux ou les superpose...
# grid.draw(a)
# grid.newpage()
# grid.draw(b)

#grid.arrange(a,b,ncol=2)

length(rownames(.df.bin)[-29]) #nombre d'interaction (le 29e est none)
```

Légende : * p value <= 0.05 soit coefficient beta significativement différent de 0

Interprétation :
J'ai 5 interactions significatives (p value<=0.05) : age-alcool, vitamine-alcool, sexe-alcool, cholesterol-alcool, tabac-alcool.

Je trace la représentation graphique des pvalue des 28 interactions possibles

```{r, echo=FALSE}
x <- 1-na.omit(mod_order.bin$Pr..Chi.)
y <- length(x):1

g<-ggplot(na.omit(mod_order),aes(x,y))+
  geom_point(size=2)+
  xlab("1-p")+ylab("Np")
g+  geom_abline(slope=coefficients(lm(1:length(x) ~ -1 + x[length(x):1])),color="lightblue4", size=1)
```

J'ai 9 points au dessus de la ligne mais 5 points qui se détachent du groupe. L'interprétation n'est pas aisée.